Testing for intransitivity of preferences predicted by a lexicographic semi-order

Michael H. Birnbaum *, Roman J. Gutierrez

Department of Psychology, California State University, Fullerton, H-830M, CA 92834-6846, USA

Abstract

A classic paper by Tversky (Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76, 31–48.) reported that some people systematically violated transitivity of preference when choosing between specially constructed risky gambles. This conclusion remained controversial because his statistical analysis did not allow each participant to have a different true preference order. Recently, however, it has been argued that an inherently intransitive process governs risky decision making. This paper uses a relatively new statistical technique for testing transitivity and analyzes two new experiments in which hundreds of participants made choices among the same gambles studied by Tversky. It was found that very few people repeated intransitive patterns. The incidence of violation of transitivity was slightly higher when probability was displayed graphically without numerical information, but even in this condition few participants were intransitive. Furthermore, even among those few who appeared to be intransitive, most showed dimension interaction, contrary to implications of the lexicographic semi-order. These results cast doubt on a lexicographic semi-order as a descriptive theory of risky decision making.
© 2007 Elsevier Inc. All rights reserved.

Transitivity of preference is the assumption that if a person prefers A to C and C to E, then that person prefers A to E. This principle is not only assumed by most theoreticians to be a rational principle to obey, but it is also assumed by most, but not all, descriptive theories of risky decision making. It was therefore surprising when May (1954), and later Tversky (1969) reported that some people are systematically intransitive when presented with specially constructed choices. They noted that if real, such findings ruled out many popular descriptive theories.

The gambles used by Tversky (1969) are listed in Table 1. Each was a two-branch gamble with probability *p* to win cash prize *x*, otherwise receive nothing. These were designed so that if people used a lexicographic semi-order, their preferences would be intransitive. A semi-order

(Luce, 1956) has the property that small differences are responded to as if there were no difference. A lexicographic order is illustrated by the task of alphabetizing a list of words. First one examines the first letters in each word, and if they differ, the ranking is based on this letter alone; only if the first letters are the same does one check the second letters, which if different determine the order, and so on (Fishburn, 1971; Hausner, 1954).

According to the lexicographic semi-order (LS) interpretation of Tversky's (1969) results, people first compared probabilities to win in the two gambles. When this difference was large enough (greater than Δ), they chose the gamble with the higher probability to win. However, when this difference was small, they compared prizes and chose the gamble with the larger prize. Suppose the difference in probability, Δ, must be greater than 0.1 to be decisive. In that case, people would prefer A to C and C to E, but prefer E to A.

---

* Corresponding author. Fax: +1 714 278 7134.
E-mail address: mbirnbaum@fullerton.edu (M.H. Birnbaum).

Table 1
Gambles used by Tversky (1969)

Gamble	Probability (p)	Prize (x)	Expected value
A	.29	5.00	1.46
B	.33	4.75	1.58
C	.38	4.50	1.69
D	.42	4.25	1.77
E	.46	4.00	1.83

All gambles are of the form p to win x, otherwise 0.

González-Vallejo (2002) proposed a stochastic difference model to account for intransitivity as in Tversky (1969) as well as other phenomena in choice. In this model, people integrate relative differences in dimensions and compare this integrated value to a difference limen. This model includes a random error component that allows the model to predict choice proportions.

Brandstätter, Gigerenzer, and Hertwig (2006) presented the priority heuristic (PH), a variant of the LS model. Unlike the stochastic difference model, this model assumes that people make a decision for one reason only and do not integrate dimensions. They argued that this model provides a more accurate description of certain risky decision making data than transitive utility models such as cumulative prospect theory (CPT) (Tversky & Kahneman, 1992) or the transfer of attention exchange (TAX) models (Birnbaum, 1999). According to the PH, the majority should be intransitive when asked to choose between the Tversky (1969) gambles.

Tversky's (1969) conclusions regarding transitivity, however, remained controversial. Although some replicated Tversky's results (Budescu & Weiss, 1987; Lindman & Lyons, 1978; Montgomery, 1977), other investigators did not concede that the apparent violations reported were "real". Iverson and Falmagne (1985) reanalyzed Tversky's (1969) data and concluded that if each participant were allowed to have a different transitive order, the data would not require rejection of the null hypothesis of transitivity. In addition, Tversky (1969) had selected participants by means of a pre-test. Of 22 participants who were pre-tested, Tversky selected only those eight people who were most intransitive in the pre-test for use in his main experiment, and not all of these eight appeared intransitive in the main experiment.

Reviewers of this literature concluded that the purported evidence of intransitivity was not convincing enough that theoreticians should abandon transitive models (Luce, 2000; Luce & von Winterfeldt, 1994; Rieskamp, Busemeyer, & Mellers, 2006; Stevenson, Busemeyer, & Naylor, 1991). Indeed, Tversky went on to publish transitive theories of risky decision making (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992).

Iverson, Myung, and Karabatsos (2006) recently reanalyzed Tversky's (1969) data by means of a Neo-Bayesian approach (Myung, Karabatsos, & Iverson, 2005) and concluded that his data indeed violated weak stochastic transitivity. Weak stochastic transitivity (WST) is the principle that if $P(A, C) > 1/2$, $P(C, E) > 1/2$, then $P(A, E) > 1/2$. But the property of WST is itself a complex one because it confounds both transitive and intransitive triads. It is possible that every participant is transitive, yet group data can violate weak stochastic transitivity (An example will be described in the section after the next section).

Morrison (1963) analyzed the implications of a family of 42 models like those of Thurstone (1927) and Luce (1959), and argued that one should test not only weak stochastic transitivity but also the triangle inequality. The triangle inequality holds that, $1 \leqslant P(A, B) + P(B, C) + P(C, A) \leqslant 2$. Although Tversky (1969) cited Morrison (1963), he did not report tests of the triangle inequality. Regenwetter, Stober, Dana, and Kim (2006) argued transitivity is better analyzed via the triangle inequality, which they found was satisfied by Tversky's (1969) data. But even the triangle inequality, by combining data from different response patterns could give an erroneous answer to the question of transitivity.

A similar debate concerning transitivity of preference for risky gambles has also occurred with respect to violations of transitivity predicted from regret theory. In this literature, investigators analyze frequencies of response patterns, which is a better way to present the data. Studies found that violations of transitivity predicted by regret theory are more frequent than violations of the opposite pattern (Humphrey, 2001; Loomes, Starmer, & Sugden, 1989, 1991; Loomes & Taylor, 1992; Starmer & Sugden, 1998). Starmer (1999) also used this criterion of asymmetry to argue for another type of intransitivity. However, these studies also remain controversial because a purely transitive model with random errors can produce asymmetry (Luce, 2000; Sopher & Gigliotti, 1993). In addition, new experiments analyzed by more precise methods do not find evidence supporting regret theory (Birnbaum & Schmidt, submitted for publication).

The purpose of this paper is to examine the issue of transitivity using a relatively new method for analyzing errors in such data, and with a large sample of participants. We use a variant of a "true and error" model that allows each person to have a different preference pattern that may or may not be transitive, and each choice is allowed to have a different rate of error (Sopher & Gigliotti, 1993). In addition, the error rate on each item is estimated from preference reversals that occur when the same choice is presented repeatedly to the same participants (Birnbaum, 2005b). This constraint improves the power of the test of transitivity. Two new experiments are conducted using the Tversky gambles with hundreds of participants in order to determine if the

violations are "real" and to determine if they are consistent with the PH. If there are systematic violations of transitivity, a large class of theories must be either rejected or revised to account for the violations.

The class of transitive models includes expected utility theory (EU), cumulative prospect theory (CPT), prospective reference theory (PRT), transfer of attention exchange (TAX), gains decomposition utility (GDU) and many others (Birnbaum, 1999; Luce, 2000; Marley & Luce, 2005; Tversky & Kahneman, 1992; Wu, Zhang, & Gonzalez, 2004; Viscusi, 1989). Any theory that assumes that people evaluate gambles independently and choose (or tend to choose) the gamble with the higher overall evaluation will satisfy transitivity. Although these models can be compared to each other by means of special experiments testing properties that distinguish them (Birnbaum, 1999, 2004a, 2004b, 2005a, 2005c; Marley & Luce, 2005), they all share in common the property of transitivity.

Models that violate transitivity include the lexicographic semi-order (Fishburn, 1971; Hausner, 1954; Tversky, 1969), the additive difference model [including regret theory of Loomes and Sugden (1982), similarity theory (Leland, 1994, 1998) and Fishburn's (1982, 1991, 1992) Skew-symmetric bilinear utility], Bordley's (Bordley, 1992; Bordley & Hazen, 1991) expectations-based Bayesian variant of Viscusi's PRT model, the stochastic difference model (González-Vallejo, 2002), the priority heuristic (Brandstätter et al., 2006), context-dependent model of the gambling effect (CDG, Bleichrodt & Schmidt, 2005) and context- and reference-dependent utility (CRU, Bleichrodt & Schmidt, 2002).

In order to devise tests of the transitive models, one uses models from the second class to design experiments that are predicted to yield violations. In this study, we use the PH, which implies that overall choice probabilities should violate weak stochastic transitivity with the Tversky (1969) gambles.

The rest of this paper is organized as follows. The next section describes the PH and shows its predicted violations of transitivity. A parameter is introduced to expand the model's predictions. The third section analyzes transitivity in terms of response patterns, the fourth describes the error model; the fifth section describes the first experiment with two studies; the sixth section presents results showing that violations of transitivity are rare and do not show the majority preferences predicted by PH. The discussion of the first experiment attempts to account for the difference in results between our results and Tversky's in terms of differences in the methods. The eighth and ninth sections describe the second experiment and its results, which show that even when probability is displayed as pie charts without numerical information, violations of transitivity are not characteristic of the majority of individuals. The last section discusses implications for the priority heuristic and related models as descriptions of risky decision making.

Theoretical predictions

According to the priority heuristic (PH) model of Brandstätter et al. (2006), people choose between two-branch gambles by first considering the lowest consequences. If the difference in the lowest consequences is greater than or equal to 10% of the largest consequence in the two gambles, the decision is based on lowest consequence only. If the difference in lowest consequences is smaller than this criterion, people consider probability. If the difference in probability exceeds a critical difference (Δ), people choose based on probability only. Only if both of these differences are not decisive does the person consider the highest prizes to win, in which case the decision is based on this dimension alone. Brandstätter et al. (2006) show that with $\Delta = 0.1$, their model gives a fairly accurate description of the aggregate data of Kahneman and Tversky (1979), Tversky and Kahneman (1992), Tversky (1969), and several other studies.

Because all of the gambles in Table 1 have the same lowest consequence ($0), the dimension with next highest priority is probability. But each one-step difference in probability is less than 0.1, so the majority should choose the gamble with higher prizes in such comparisons; that is, $A \succ B$, $B \succ C$, $C \succ D$, $D \succ E$. Every two-step difference in Table 1 also differs by less than 0.1 in probability so $A \succ C$, $B \succ D$, $C \succ E$. However, intransitivity is predicted because the model implies that $D \succ A$, $E \succ B$, and $E \succ A$ because probability differences in these choices exceed 0.1.

Although Brandstätter et al. (2006) argue against estimating parameters from the data, it seems reasonable to allow that individual differences or different experimental manipulations might produce different values of Δ. Intransitivity will occur if $.05 < \Delta < .17$, in which case $E \succ A$ will contradict other choices. When $\Delta > .08$, it follows that $A \succ C$, $B \succ D$, $C \succ E$; and for $\Delta < .13$, we also have $D \succ A$ and $E \succ B$. A completely transitive order can be observed if Δ is small enough, in which case the gambles will be ordered strictly by probability to win; alternately, if Δ is large enough, gambles would be transitively ordered by their cash prizes.

Analysis of transitivity

The properties of weak stochastic transitivity and the triangle inequality are complex relations that confound transitive and intransitive data patterns. Therefore, testing only these properties is not the best way to analyze transitivity. It is best to analyze the data in terms of more detailed response patterns, as in Table 2. When

Table 2
Patterns of choice for three gambles, A, C, and E

Notation	Preference pattern	Preference order
000	$A \succ C$; $C \succ E$; $E \succ A$	Intransitive
001	$A \succ C$; $C \succ E$; $E \prec A$	$A \succ C \succ E$
010	$A \succ C$; $C \prec E$; $E \succ A$	$E \succ A \succ C$
011	$A \succ C$; $C \prec E$; $E \prec A$	$A \succ E \succ C$
100	$A \prec C$; $C \succ E$; $E \succ A$	$C \succ E \succ A$
101	$A \prec C$; $C \succ E$; $E \prec A$	$C \succ E \succ A$
110	$A \prec C$; $C \prec E$; $E \succ A$	$E \succ C \succ A$
111	$A \prec C$; $C \prec E$; $E \prec A$	Intransitive

The pattern predicted by the priority heuristic (with $\Delta = .1$) is 000.

there are three gambles, for example A, C, and E from Table 1, there are eight possible response patterns for paired choices, shown in Table 2. Two of these response patterns are intransitive (000 and 111), and the other six are transitive. We assume that each person can have a different "true" preference pattern, which may or may not be transitive, as listed in Table 2.

Note that even if everyone is perfectly consistent with transitivity, it can happen that weak stochastic transitivity will be violated. For example, suppose $P(100) = P(010) = P(001) = 1/3$, it follows that: $P(A, C) = P(000) + P(001) + P(010) + P(011) = 2/3$, $P(C, E) = P(000) + P(001) + P(100) + P(101) = 2/3$, and $P(E, A) = P(000) + P(010) + P(100) + P(110) = 2/3$. Hence aggregated data can violate weak stochastic transitivity even though every individual person had a transitive order. It could also happen that many people systematically violate transitivity and yet weak stochastic transitivity and the triangle inequality can be satisfied. For these reasons, one should analyze individual response patterns as in Table 2 in addition to overall choice proportions.

When there are five stimuli, there are 10 paired comparisons. For ten choices, there are $2^{10} = 1024$ possible data patterns, of which $5! = 120$ are transitive. If each choice is presented twice, there are more than a million possible data patterns (1024×1024). To further complicate matters, the observed data pattern may differ from a person's "true" data pattern because of random "errors" of choice. To simplify matters and to provide a statistical null hypothesis for transitivity, we need a model that allows random error but does not itself entail transitivity.

True and error model

Models of error have been discussed by Birnbaum (2004b, 2005b), Carbone and Hey (2000), González-Vallejo (2002), Harless and Camerer (1994), Hey and Orme (1994), Luce (1959, 1994), Morrison (1963), Sopher and Gigliotti (1993), Thurstone (1927), and others. Probabilistic choice models of Thurstone (1927), Luce (1959), Busemeyer and Townsend (1993), Hey and Orme (1994) implicitly assume transitivity in the absence of error; that

of González-Vallejo (2002) assumes intransitivity. Because we plan to test transitivity, it seems best to use a model of error that is neutral with respect to transitivity.

We assume that if people made no errors, the same person would make the same decision every time when presented the same choice. The proportion of preference reversals between replications of the same choice should therefore allow us to estimate the error rate for that choice. Different people, however, might make different decisions given the same choice because they have different true preferences. Once the errors are estimated, our approach is similar to that of Harless and Camerer (1994) and Sopher and Gigliotti (1993). Whereas Harless and Camerer (1994) assumed that choices between different gambles had the same error rates, which seems unrealistic, Sopher & Gigliotti allowed error rates in choices between different gambles to be unequal. We use Birnbaum's (2005b) improvement, which uses repeated presentations of the same choices in order to unambiguously estimate the error rate for each distinct choice. An advantage of this approach is that error terms are not constructed to conveniently "explain away" violations of transitivity [which was an objection to Sopher and Gigliotti (1993)], but are estimated by a method that is neutral with respect to the issue under investigation. This approach increases the power of the test of transitivity.

Consider a choice between A and C that is presented twice to the same participants. Some people will choose A both times, some will choose C both times, some will switch from A to C and some switch from C to A. The probability of switching from A to C is given as follows:

$$P(AC) = p(1-e)e + (1-p)(1-e)e = e(1-e) \qquad (1)$$

where p is the probability that a person "truly" prefers A over C and e is the error rate for this choice. It is assumed that errors are independent and that $e < 1/2$. Presumably, the more "difficult" a choice, the higher the error rate on that choice, and the more likely a person would be to reverse preferences when the choice is presented again. Those who "truly" prefer A over C have correctly reported their preference the first time and made an error the second time, whereas those who "truly" prefer C have also made one error and one correct response. This model implies that the probability of switching from A to C equals the probability of switching from C to A. The proportion of either type of preference reversals to a given choice, therefore, is $2e(1-e)$, which can be used to estimate the error rate for that choice. For example, if we observed 32% reversals of preference when a given choice was presented twice, we would infer that $e = 0.2$.

We can also estimate the "true" probability of preferring A to C from the probability that a person chooses A both times, which is given as follows:

$$P(AA) = p(1-e)^2 + (1-p)e^2 \qquad (2)$$

Those who truly prefer A have made two correct reports and those who truly prefer C have made two errors. Eqs. (1) and (2) permit estimation of both p, the true probability of preferring A, and e, the rate of error. This model is called the "true and error" model after these interpretations of p and e. Suppose there were 32% preference reversals on this choice, if we observed that only 4% chose A both times, we would infer from Eq. (2) that $p = 0$; i.e., no one truly prefers A. With the same percentage of reversals, if we observed that 34% chose A both times, we would infer that $p = .5$. Except when $p = 0$ or $p = 1$, this model implies that choices on two replications of the same choice will not be independent (even though the errors are), because different people have different true preferences. Given a 2×2 table of frequencies for the four choice patterns, we can perform a χ^2 test of independence, which in general will be significant. We can also perform a $\chi^2(1)$ test of the fit of the true and error model to the same data, which should not be significant, if the model is an accurate description of the data.

The next step is to extend this model to analyze an experiment with three choices, AC, CE, and AE, designed to test transitivity. The model should predict the observed frequencies of the eight possible patterns of response in Table 2. The probability of exhibiting the intransitive pattern 000 is as follows:

$$
\begin{aligned}
P(000) = {} & p_{000}(1-e_1)(1-e_2)(1-e_3) + p_{001}(1-e_1)(1-e_2)e_3 \\
& + p_{010}(1-e_1)e_2(1-e_3) + p_{011}(1-e_1)e_2e_3 \\
& + p_{100}e_1(1-e_2)(1-e_3) + p_{101}e_1(1-e_2)e_3 \\
& + p_{110}e_1e_2(1-e_3) + p_{111}e_1e_2e_3
\end{aligned}
$$

$$(3)$$

where $P(000)$ is the probability of the observed intransitive data pattern; p_{000} is the probability that a person has 000 as her or his "true" pattern; and e_1, e_2, and e_3 are the probabilities of making an "error" in expressing preference on the three respective choices. There are seven other equations like Eq. (3) for the other seven observable patterns of Table 2. There are eight response frequencies (which sum to the number of participants), to be fit by eight true probabilities (which sum to 1) and three error rates. If we only had the eight frequencies of response patterns as data, the parameters would not be well-defined by the data (10 parameters to be estimated from eight frequencies). This is where use of repeated measures allows us to fit the most general form of the model and still have degrees of freedom for a test.

In this study, each choice is repeated with positions of the two gambles counterbalanced. With two such repetitions, there are 64 possible response patterns (8×8), and the equations are expanded to allow for up to six errors or correct reports. The equation for repeating the 000 pattern on both presentations, for example, is the same as Eq. (3), except that each e or $(1 - e)$ term is squared.

We have thus enriched the data without increasing the number of parameters.

Because many of these 64 cells have small frequencies, data are partitioned as follows. We count the number of people who repeat each data pattern (show the same pattern on both replications) and the average number who show each data pattern on either the first or second repetitions but not both. This partition results in 16 mutually exclusive frequencies that sum to the number of participants. There are thus 15 degrees of freedom in the partitioned data, from which we estimate seven probabilities of true patterns (the eighth is determined because they sum to 1), and three error terms, leaving five degrees of freedom.

This "true and error" model with replications is neutral with respect to the issue of transitivity. It allows us to estimate the "true" probability of violating transitivity from the observed frequencies of the data patterns. This probability of intransitivity is given by $p_{000} + p_{111}$. This model thus provides a statistical null hypothesis for transitivity; the transitive model is a special case of this general "true and error" model where $p_{000} + p_{111} = 0$.

Method of Experiment 1

Participants served in a series of judgment and decision making studies that included this one. They viewed choices between gambles via the Internet and clicked a button to indicate the gamble in each choice they would rather play. Gambles were described in terms of containers holding 100 tickets from which one ticket would be chosen at random to determine the prize. In the first study, choices were displayed as in the following example:

First gamble:
46% chance to win $4.00
54% chance to win $0
OR
Second gamble:
33% chance to win $4.75
67% chance to win $0

There were 20 choices consisting of all 10 possible pairs of gambles A, B, C, D, and E of Table 1 with position (first or second gamble) counterbalanced. Expected values, displayed in Table 1, were not presented to the participants.

Two versions of the experiment were tested, with different prize values. Because the prizes were worth more (corrected for inflation) in 1969 than the same dollar values would be today, we also tested the same probabilities with prizes 100 times as large. Choices in this second version were also based on drawing a ticket from an urn containing 100 otherwise identical tickets; each choice appeared as in the following example:

First Gamble:
    46 tickets to win $400
    54 tickets to win $0
OR
Second Gamble:
    33 tickets to win $475
    67 tickets to win $0

Within each version, the order of 20 choices was randomized. There were 417 college students in the first variation with small prizes. They participated as one option toward an assignment in lower division psychology. Of these, 60% were female; 90% were 22 years or younger. These participants were asked if they played poker, and 177 said that they did.

At the time this study was conducted, poker was experiencing renewed popularity, and sports channels were broadcasting poker tournaments with expert commentary evaluating probabilities and strategies. We thought that poker players might be more likely to obey transitivity than would non-players, since players presumably have experience with risky decision making, perhaps for real cash consequences. However, this poker-playing variable was not found to correlate in any meaningful way with the data, so results of both groups are combined in the analyses presented.

There were 327 participants from the same subject pool who completed the second version with larger prizes; of these, 57% were female and 93% were 22 years or younger. The two versions were embedded among a dozen other judgment and decision-making tasks; they were always separated by at least two other intervening tasks that required about 10 min.

About half in each variation were tested in laboratories containing Internet-connected computers and others participated via the Internet at times and places chosen by the participant. The variable of laboratory versus self-testing did not produce any discernable differences, so data are also combined over these groups in the analyses that follow.

Complete materials can be viewed from the following URLs:

    http://psych.fullerton.edu/mbirnbaum/psych466/exps/choicepagetaversky.htm

    http://psych.fullerton.edu/mbirnbaum/decisions/rep_tversky.htm

Results of Experiment 1

Tables 3 and 4 show the proportion that chose the column gamble over the row gamble in the first and second variations of the study, respectively. Note that in both tables, all values above the diagonal exceed 1/2 and all values below the diagonal are less than 1/2. Both matrices are therefore perfectly consistent with weak

Table 3
Choice proportions in tests of weak stochastic transitivity ($n = 417$)

Row gamble	Proportion choosing column (second) gamble over row (first) gamble				
	A	B	C	D	E
A		0.729	0.770	0.801	0.854
B	0.302		0.683	0.791	0.787
C	0.161	0.290		0.736	0.779
D	0.113	0.161	0.237		0.631
E	0.130	0.173	0.146	0.331	

These proportions are consistent with the transitive order $E \succ D \succ C \succ B \succ A$, which agrees with probability to win, expected value, TAX, and CPT.

Table 4
Choice proportions, as in Table 3, for second variation in which prizes were 100 times as large ($n = 327$)

Row gamble	Proportion choosing column (second) gamble over row (first) gamble				
	A	B	C	D	E
A		0.676	0.789	0.798	0.780
B	0.352		0.676	0.801	0.789
C	0.177	0.309		0.661	0.740
D	0.177	0.177	0.327		0.584
E	0.180	0.211	0.245	0.365	

Although some proportions differ between Tables 3 and 4, these data are also consistent with WST.

stochastic transitivity. The majority order agrees with probability to win: $E \succ D \succ C \succ B \succ A$. This order, denoted as *EDCBA*, also agrees with expected value (EV), CPT, TAX, and other models, despite the design in which expected values are nearly equal.

These results are quite different from the pattern predicted by the PH of Brandstätter et al. (2006) or observed by Tversky (1969) for pre-selected participants. According to the PH, those 14 cells in Table 3 (and 14 in Table 4) in which the gambles differ by one or two steps should have favored the gamble with the higher prize. Instead, in all 28 such cases, the majority of participants chose the gamble with the higher probability, even though the probability differences are less than .1. Consequently, the PH is correct in predicting only 30% of the majority choices.

Tables 5 and 6 show the analysis of response patterns for gambles *A*, *C*, and *E*, as in Table 2 for the first and second variations, respectively. Table 5 shows that only five of 417 people repeated the intransitive pattern predicted by the PH on both presentations of these three choices, and five people repeated the opposite pattern. The most frequent pattern was the transitive pattern 110, which was repeated by 220 participants; this pattern is consistent with the order *ECA* (i.e., $E \succ C \succ A$). Table 6 shows that of the 327 participants in the second variation, only one person repeated the predicted pattern of intransitivity and one repeated the opposite pattern for this triad.

Table 5
Observed frequencies of each choice pattern for gambles *A*, *C*, and *E*

Choice pattern	Observed frequencies				Model calculations		
	Rep 1	Rep 2	Both	Rep 1 or 2 not both	$\hat{p}$	Predicted both	Predicted 1 or 2 not both
000	14	28	5	16	.03	8.1	8.6
001	18	25	15	6.5	.07	15.3	6.5
010	23	38	1	29.5	.00	4.7	37.2
011	12	5	3	5.5	.01	2.8	5.9
100	24	33	7	21.5	.03	7.8	26.0
101	5	6	1	4.5	.00	0.9	5.5
110	301	256	220	58.5	.85	196.6	67.6
111	19	25	5	17	.02	4.6	17.9
Sum	416	416	257	159	1	240.9	175.1

Error terms, estimated from preference reversals between counterbalanced positions are 0.13, 0.08, and 0.06 for $\hat{e}_1$, $\hat{e}_2$, and $\hat{e}_3$, respectively. According to this model, only about 5% of participants are intransitive.

Table 6
Observed frequencies of each choice pattern for gambles *A*, *C*, and *E* for gambles with prizes 100 times as large

Choice pattern	Observed frequencies				Model calculations		
	Rep 1	Rep 2	Both	Rep 1 or 2 not both	$\hat{p}$	Predicted both	Predicted 1 or 2 not both
000	10	13	1	10.5	.01	1.5	5.3
001	31	33	21	11	.15	31.1	9.3
010	8	10	2	7	.01	2.0	12.5
011	9	13	3	8	.01	3.1	5.8
100	26	25	7	18.5	.03	7.8	21.8
101	13	14	2	11.5	.02	3.4	5.5
110	224	207	189	26.5	.77	156.9	43.4
111	6	12	1	8	.00	1.1	16.5
Sum	327	327	226	101	1	206.9	120.1

Error terms, estimated from preference reversals between counterbalanced positions are 0.06, 0.10, and 0.08 for $\hat{e}_1$, $\hat{e}_2$, and $\hat{e}_3$, respectively. According to this model, only about 1% of participants are intransitive.

Before describing complete response patterns in the data, we first report self-consistency. Between the repetitions, the average rates of agreement over all 10 choices were 81.5% and 83.9% in the first and second variations with 417 and 327 participants, respectively. It was found that in the first version 115, 85, and 79 people of 417 were perfectly consistent on all 10 choices, consistent on 9 of 10, and 8 of 10 choices respectively. Of the 327 in the second variation, there were 115 who were perfectly self-consistent between repetitions, with 67 and 43 who were consistent on 9 or 8 of 10, respectively.

Of the 115 who were perfectly reliable in the first variation, 93 (81%) were perfectly consistent with the transitive order *EDCBA*, 8 were perfectly consistent with the opposite transitive order, *ABCDE*, which is consistent with magnitude of cash prize. Only one person showed an intransitive pattern consistent with the PH. Of the 115 perfectly consistent participants in the second variation, 96 (83%) were perfectly consistent with the order *EDCBA*, and 12 were perfectly consistent with the opposite order. Of the remaining seven participants, four had patterns at least partially consistent with the LS including one who conformed to the PH.

For those who were not perfectly consistent, we treated inconsistency for a given choice as missing data for that choice. This reduces the number of possible data patterns from more than a million to a maximum of 1024. Including all 417 participants in the first variation, there were 238 people (57%) whose consistent choices were perfectly compatible with the transitive order, *EDCBA*, and 21 who were perfectly consistent with the transitive order *ABCDE*. In the second variation, there were 174 of 327 (53%) whose consistent choices were perfectly consistent with *EDCBA* and 31 completely consistent with the opposite order.

According to the PH of Brandstätter et al. (2006), people should choose $A \succ B$, $B \succ C$, $C \succ D$, and $D \succ E$. To search for individuals who might agree with PH, therefore, we selected data for people who satisfied these predicted preferences on one-step comparisons. We found only 18 of 417 participants in the first variation. Of these 18, 12 were consistent with the transitive order, *ABCDE*. Five of the remaining six showed $E \succ A$, as predicted by an LS model in which $.05 < \Delta < .17$.

With larger prizes, there were 29 of 327 people who showed the pattern predicted by the PH for one-step comparisons. Of these, 16 showed the transitive order $ABCDE$, and 10 showed $E \succ A$. These incidences are much lower than predicted by PH or observed by Tversky.

Lowering the criterion for intransitivity to match Tversky's (1969) lenient standard, we counted those participants who made the predicted one-step choices at least 75% of the time and who preferred $E \succ A$. There were 32 and 30 people who satisfied this criterion, amounting to 7.6% and 9.3% in the first and second variation, respectively. Thus, even by this lower standard, we observe few cases of intransitive behavior. To determine if these cases might be "real" or due entirely to chance, we fit the true and error model to the data.

The error terms (e_1, e_2, e_3 for AB, BC, and CA choices, respectively) were estimated directly from the number of preference reversals between repetitions of each choice, using Eq. (1). This method may not minimize a test statistic, but it has the advantage that the estimated error terms are independent of assumptions concerning transitivity. In the first version, there were 61 of 417 who switched from choosing A over C when A was in the second position to choosing C over A, when C was in the second position and 32 who made the opposite switch. This corresponds to an error rate of $\hat{e}_1 = 0.13$. The model assumes that people should be equally likely to make either switch. Instead, all three choices in both the first and second variations showed the same asymmetry; that is, people tended to favor slightly the gamble in the second position. The other estimated error rates were $\hat{e}_2 = 0.08$, $\hat{e}_3 = 0.06$ in the first variation, and they were 0.06, 0.10, and 0.08 in the second variation, respectively.

When the true and error model is fit to the data of each version of the experiment separately, with error rates determined as above, it is found that intransitive patterns are estimated to be low in probability, which is a consequence of the small numbers who repeated these patterns. The sixth columns in Tables 5 and 6 show the estimated true probabilities of the eight response patterns. In Table 5, it is estimated that about 3% and 2% were "truly" intransitive on A, C, and E, with the patterns 000 and 111, respectively. In Table 6, both types of intransitivity sum to only 1%.

The predicted frequencies, shown in the right-most columns of Tables 5 and 6 do a reasonable job of fitting the observed frequencies in Tables 5 and 6, but the deviations are statistically significant. The χ^2 comparing the 16 observed and predicted frequencies in Tables 5 and 6 are 17.2 and 37.3, indicating significant deviations from the true and error model.

In addition to the above-mentioned asymmetry of preference reversals (reflecting a slight tendency to prefer the gamble in the second position), the data suggest that some people have higher rates of error than others. For example, the number of people who actually repeated the 000 pattern of intransitivity is 5 but the predicted number is 8.1 in Table 5. In contrast, the observed number who showed this pattern on one replicate or the other (but not both) is 16 and the predicted number is only 8.6.

A more accurate version of the model would allow each subject to have a different noise amplifier as follows:

$$e_{ij} = \sigma_j \varepsilon_i \tag{4}$$

where e_{ij} is the error rate for choice i and person j; σ_j is the error amplifier for person j, and ε_i is the error rate for choice i. In this model, some choices are more difficult than others, and some participants are noisier than others. Each participant has a different "true" pattern, which may or may not be transitive.

The model in Eq. (4) would require estimation of hundreds of parameters reflecting each subject's noise parameter. A simpler version of this model was fit in which the 417 participants were divided into a "reliable" group (those 200 who agreed with their own choices 9 or 10 times) and a "less reliable" group (178 whose agreement ranged from 6 to 8). The "true" probabilities of each response pattern were assumed independent of reliability. This model uses 11 parameters to fit 32 frequencies (16 in each group), and it achieved a $\chi^2 = 18.7$, an acceptable fit. The estimated error rates in the reliable group were 0.023, 0.022, and 0.014. The error rates in the less reliable group were estimated to be 8.15 times as large. The true probabilities were estimated to be 0.02, 0.06, 0.00, 0.01, 0.01, 0.00, 0.88, and 0.01, for response patterns 000 to 111, respectively. Note that two transitive orders, ACE and ECA, account for $88 + 6 = 94\%$ of the participants in this model, with 3% intransitive. Similar results were found in the second variation.

Discussion of Experiment 1

Our data show that the vast majority of participants tested were transitive for choices among the Tversky (1969) gambles. According to the PH of Brandstätter et al. (2006), the majority should have shown intransitivity. Instead, only two people showed the pattern predicted by PH (one in each study). The true and error model estimates that perhaps 3% of participants were "truly" intransitive for the triad of A, C, and E. This incidence is much lower than reported by Tversky (1969), even when we take into account that he selected subjects from a pretest.

Tversky (1969) pre-tested 18 Harvard undergraduates from which 8 (44%) were judged intransitive by Tversky's lenient criterion on the pretest. Of these 8, Tversky found 4 who appeared to show the intransitive pattern predicted by the PH for A, C, and E. From his data,

we would thus expect that perhaps 22% (4 of 18) to 44% (8 of 18) of our unselected participants should have shown this pattern. In a supplementary study, Tversky (1969) found 13 of 36 unselected participants (36%) who were judged intransitive by the same lenient criterion. Lindman and Lyons (1978) used a Bayesian analysis and concluded that 22 of 42 (52%) participants were more likely intransitive than transitively ordered by either probability or prize, and 8 others were plausibly intransitive, making an overall incidence of perhaps 70%. Budescu and Weiss (1987) found that 36% of their participants were classified as intransitive by Tversky's criterion. Montgomery (1977) reported that 7 of 21 (33%) were intransitive by the same standard. Instead, we observed fewer than 10% were intransitive by that same standard, and the estimated incidence is still lower when errors are taken into account.

We suspected that procedural differences accounted for the differences between our results and those of Tversky (1969). We revised our methods in Experiment 2 in an attempt to find higher rates of intransitivity. Experiment 1 differed from Tversky's in at least six ways.

First, we did not pre-test and select participants who appeared to be intransitive.

Second, our participants were tested via computer rather than individually tested by a person. When testing face-to-face, an experimenter might provide subtle reinforcements to participants that might affect behavior. Experimenters who know the hypothesis might also miscode or make data entry errors that conform to the anticipated theory or introduce other biases that may cause data to confirm the expected result (Rosenthal & Fode, 1961). But Budescu and Weiss (1987) found 8 of 22 participants were intransitive by Tversky's criterion, despite testing via computer.

Third, we used students from a different generation and a different type of university. Fourth, our participants performed this task as one among several studies of choices among gambles. The context provided by the other gambles may have been different from that produced by the experimental and "filler" trials used by Tversky and others. Fifth, Tversky used a greater number of replications in his main study.

Sixth, we used text (numeric) displays of both prizes and probabilities, whereas Tversky (1969) used graphic representation of probability. Tversky (1969), Lindman and Lyons (1978), and Montgomery (1977) also used pie charts to display probability, and actual probabilities were not presented numerically.

Of these six factors, we thought that the graphical display was the most important difference. Therefore, in Experiment 2, we used pie charts to represent probability, either with or without text presentation of probability. In addition, we conducted a follow-up with an older and better-educated group of people who had not served in our other decision tasks.

Hypotheses of Experiment 2

Intransitivity might arise from a non-compensatory model such as a lexicographic semi-order such as PH, from an additive difference model, or it might arise instead from a very simple illusion in a graphical display, even if people use a transitive, integrative model like EU, CPT or TAX. Suppose the subjective values of probability are altered by a perceptual illusion that depends on the two pies presented on each trial. In particular, suppose that when two pies differ by only 1/24, that participants perceive that these two represent the same probability. Assuming that people used these assimilated values in EU, CPT or TAX, for example, they would show intransitivity, because differing payoffs would be integrated with the (equal) common probability weight.

This theory is similar to the editing strategy of "simplification" (Kahneman & Tversky, 1979), in which "non-essential" differences are discarded; however, as an editing strategy, it should have applied in our first experiment. If people rounded numerical probabilities differing by less than .1 to the same value, for example, they would have shown intransitivity in Experiment 1, even if they substituted these rounded values into a transitive model like EU, CPT or TAX. As a perceptual illusion, however, intransitivity should be found only when probabilities are presented as graphs and not accompanied by text giving numerical probability information. Experiment 2 therefore used random assignment to two conditions: some participants received probability information graphically only [without numerical probabilities as in Tversky (1969)], and others received graphs with text providing numerical probabilities.

In order to test the interpretation of inherently intransitive decision processes (such as the additive difference model or PH's lexicographic semi-order) against this perceptual discrimination hypothesis, we included two tests of dimension interaction in Experiment 2. Dimension interaction is implied by models like EU, CPT, and TAX, but not by additive difference models like the stochastic difference model, PH, or LS models, which imply no interaction (Birnbaum, submitted for publication-b). According to any version of the PH, dimensions are considered one at a time, so any dimension that takes the same value in both gambles of a choice should have no effect on the choice. For example, consider the following two choices:

R:	.05 to win $7.25	*S*:	.05 to win $4.25
	.95 to win $1.25		.95 to win $3.25
R':	.95 to win $7.25	*S'*:	.95 to win $4.25
	.05 to win $1.25		.05 to win $3.25

Note that probability to win the larger prize is the same in both gambles within each choice, but it differs between choices. According to PH, people should

choose S and S' because the difference in the lowest consequences ($2) is greater than 10% of $7.25, the maximum consequence. Suppose, however, that this difference is not large enough to be decisive; in that case, the probabilities are the same in both cases, so they would be ignored, and decision-makers would choose R and R', since these have the higher best consequence.

Similarly, a decision-maker following a lexicographic semi-order who considers the dimensions in any order should prefer either both "safe" gambles or both "risky" gambles, but he or she would not switch from "risky" to "safe" or vice versa, except by random error. According to the TAX model (CPT, EU, and others), however, people should switch from S to R' because of the multiplicative interaction between functions of probability and prizes.

Experiment 2 thus allows a comparison of four types of theories: First, purely transitive, interactive models (like EU, CPT, TAX, etc.) imply that we should observe transitive preferences in both conditions of Experiment 2, and we should observe interactions between probability and prizes. Second, the perceptual illusion theory (added to an otherwise transitive model) implies that we should observe intransitive preferences only in the graphical condition; we should find transitivity in the graph with text condition, and we should see evidence of dimension interaction. Third, the editing strategy of simplification followed by a model like EU, CPT or TAX implies intransitivity in both conditions of Experiment 2 and it allows dimension interaction. Four, the lexicographic semi-order and PH as well as the stochastic difference model imply that we should observe intransitivity in both conditions of Experiment 2 and we should not observe evidence of dimension interaction.

Method of Experiment 2

Participants in Experiment 2 were randomly assigned to two conditions. In both conditions, probabilities were represented by pie charts. In the *graph only* condition, no numerical information was available on the probabilities; in the *graph and text* condition, pie charts were displayed and numerical probabilities (percentages) were displayed as text on the display. An example trial from the graph only condition is shown in Fig. 1. Participants were instructed to imagine a spinner that is equally likely to stop in any equal-sized wedge of the pie. Participants indicated their choices by clicking a button under the gamble they preferred. Clicking a "next trial" link would display the next trial.

The choices were the same as in Experiment 1 except that four choices were added in order to test dimension interaction. These were $R = (\$7.25, 0.05; \$1.25, 0.95)$ versus $S = (\$4.25, 0.05; \$3.25, 0.95)$, and $R' = (\$7.25, 0.95; \$1.25, 0.05)$ versus $S' = (\$4.25, 0.95; \$3.25, 0.05)$, which were repeated with position (first or second) counterbalanced. These trials were intermixed with the other choices and separated by at least three intervening trials from one another. Other aspects of method were the same as in Experiment 1. Complete materials can be viewed at the following URLs:

http://psych.fullerton.edu/mbirnbaum/decisions/new_pies/new_tversky_.htm
http://psych.fullerton.edu/mbirnbaum/decisions/new_pies/new_tversky_np_.htm

There were 321 new undergraduate participants, of whom 160 were randomly assigned to the graph only condition and 161 to the graph and text condition. Of these, 61% were female and 95% were 22 years of age
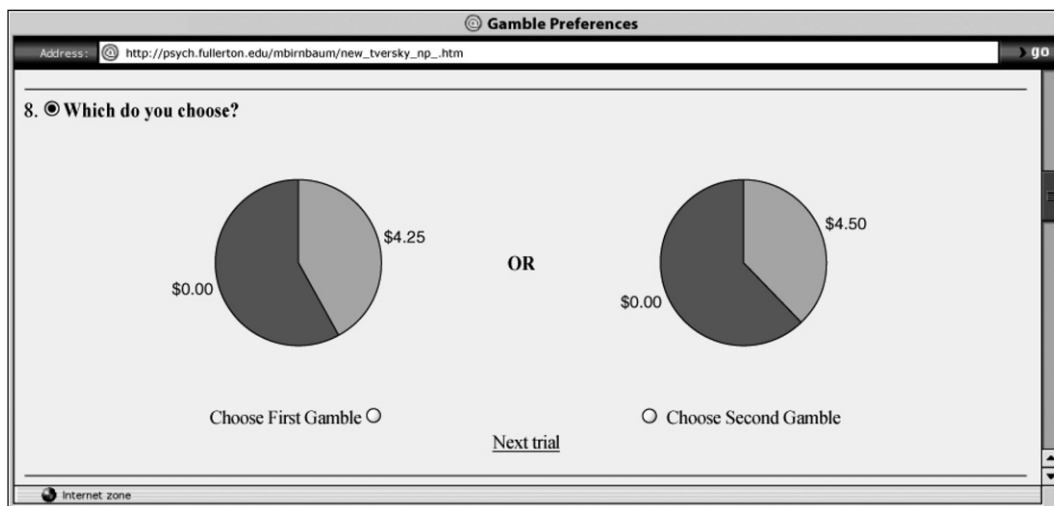

Fig. 1. Graph only format for presentation of a choice (Experiment 2). This figure shows the choice between gambles D (first gamble) and C (second gamble). In the graph and text condition, the percentages of each pie slice were printed; in this case, 42% and 38% to win the higher prize in D and C, respectively.

or younger. These participants also served in at least three other studies of judgment and decision making (requiring about 15 min) before this task.

Follow-up study to Experiment 2

In order to examine if different results would be obtained with people who had not served in our preliminary studies, an additional group of 340 participants were recruited via the WWW who served in this task without performing other decision tasks first. This group did not receive the choices testing dimension interaction, in case these might alter their decision rules. Of these, 64% were female; 44% were 22 years of age or younger, and 14% were 40 years of age or older. Whereas none of the participants in the main study were college graduates, 59% of the follow-up group held bachelor's degrees and 8% held doctorates. These were also randomly assigned to the two conditions with probability displayed graphically as pies only, or as pies with probability information presented as text.

Results of Experiment 2

Table 7 shows how participants responded to the counterbalanced presentations of AC, CE, and AE,

respectively. The table shows that for the AC choice, there were 16 who switched from choosing A to C, 18 switched from C to A, and 225 chose C on both versions of the choice. The table shows χ^2 tests of independence, all of which are significant. This violation of independence is consistent with the true and error model, because there are true individual differences. The table also shows the estimated values of p_{AB}, p_{BC}, and p_{AE}, which are 0.79, 0.78, and 0.84, respectively. The error rates and χ^2 tests of fit for the true and error model are also shown in the table. All three of these values are much smaller than the corresponding tests of independence, and one is significant.

Table 8 shows the observed frequencies of response patterns to the A, C, and E triad, displayed as in Tables 5 and 6. The most common response pattern, 110, was shown by 208, 213, and 186 people on the first replicate, second, and both replicates, respectively. This pattern agrees with the modal pattern of Experiment 1, that is, it agrees with probability to win (as well as EV, TAX or CPT), as in Tables 5 and 6. There were six people who repeated the intransitive pattern predicted by PH and two who repeated the opposite pattern. Of these six, four were in the graph only condition.

When the true and error model is fit to these data, the estimated probability of the intransitive pattern predicted by PH is 3%, and the opposite pattern is esti-

Table 7
Preference reversals between repetitions, used to estimate error probabilities (Experiment 2)

Choice	Response combination				$\chi^2(1)$ indep	True and error estimates		$\chi^2(1)$ True and error
	XX	XY	YX	YY		$\hat{p}$	$\hat{e}$	
AB	62	16	18	225	164.0	.79	0.056	0.12
BC	64	24	11	222	165.0	.78	0.062	4.69
AC	48	10	13	250	186.9	.84	0.038	0.39

The χ^2 tests of independence show that we cannot reproduce the data with the assumption that the probability to show a choice combination is the product of the separate probabilities of each choice. The true and error model also uses two parameters to fit the same four frequencies of response combinations.

Table 8
Observed frequencies of each choice pattern for gambles A, C, and E with probability displayed as pies (Experiment 2, $n = 321$)

Choice pattern	Observed frequencies				Model calculations		
	Rep 1	Rep 2	Both	Rep 1 or 2 not both	$\hat{p}$	Predicted both	Predicted 1 or 2 not both
000	19	11	6	9	.03	8.1	4.6
001	40	43	32	9.5	.17	38.6	7.2
010	13	17	4	11	.02	4.4	12.9
011	6	9	1	6.5	.00	1.2	3.8
100	23	19	7	14	.03	7.3	14.4
101	6	2	1	3	.00	1.0	3.7
110	208	213	186	24.5	.74	173.0	30.8
111	6	7	2	4.5	.01	1.9	8.3
Sum	321	321	239	82	1	235.5	85.5

Error terms, estimated from preference reversals between counterbalanced positions are 0.06, 0.06, and 0.04 for $\hat{e}_1$, $\hat{e}_2$, and $\hat{e}_3$, respectively. According to this model, about 4% of participants are intransitive.

mated to have a true incidence of 1%. The χ^2 of fit for the true and error model is 13.1, and the model appears to give a reasonable approximation to the data. However, when the probabilities of intransitive patterns are fixed to zero, the χ^2 jumps to 188.8, a difference of 175.7; the statistical test rejects the hypothesis that all participants were transitive in favor of the hypothesis that 3% were intransitive.

Examining complete response patterns of individual data, the most frequent response pattern was again the one perfectly consistent with probability (same as EV, TAX or CPT), shown by 84 participants. Thirteen were perfectly consistent with the opposite order, consistent with magnitudes of the prizes. No one was perfectly consistent with the PH model.

To search for participants who might show a tendency toward the predicted pattern of intransitivity, we again used Tversky's lenient criterion. This criterion selects participants whose choices agreed with prize value for the one-step differences, *AB*, *BC*, *CD*, *DE* at least 75% of the time and who also preferred *E* over *A*. There were 37 who satisfied this criterion (11.5%). Of these 37, 27 were from the graph only condition and 10 were from the graph and text condition. This difference is statistically significant, $\chi^2 = 8.95$, indicating that this data pattern is more frequent in the graph only condition than in the graph and text condition. This result is consistent with the perceptual illusion hypothesis. Among the 160 participants in the graph only condition, 17% are judged intransitive by Tversky's lenient criterion. This incidence is closer to but still smaller than previously published results.

Results of the dimension interaction tests are presented in Table 9. According to PH, people should choose the "safe" gamble in both cases. The cell entries show the number of participants who showed each combination of preferences on two presentations of each type of choice. The tests of independence are again significant, but in this table, tests of the true and error model were not significant, shown in the last column. The estimated "true" probability of choosing *S* or *S'* in the first choice is 0.95 in the first choice and 0.07 in the second type of choice. Of all 321 participants, 190 (59%) showed the same interactive pattern on both tests (four choices). These results violate the predictions of PH, LS, and additive difference models.

Among the 37 participants who met Tversky's criterion of intransitivity, 26 (70%) had all four choices in the test of dimension interaction indicating interaction. None of these 37 chose the "safe" gamble on all four choices, the pattern predicted by PH. Because most of the people who appeared intransitive by Tversky's criterion also showed evidence of dimension interaction, the LS and PH theories are not as accurate as the perceptual illusion hypothesis for this small group of participants who appear intransitive.

The follow-up study to Experiment 2 was conducted to investigate if the context of the laboratory situation may have induced greater conformity to transitivity than might be obtained if participants did not perform other decision tasks or experience the tests of dimension interaction. Table 10 shows observed frequencies for Graph only and graph and text conditions for the follow-up study. There were seven people who repeated the intransitive pattern predicted in the graph only condition compared to only one in the graph and text condition. It was found that 80 of 340 participants met Tversky's lenient criterion (23.5%), with 54 of these in the graph only condition.

The true and error model was fit to response frequencies for the *A*, *C*, *E*, comparisons in four groups of data: graph only or graphs with text for undergraduates tested in the laboratory or for Web recruits. In the graphs with text conditions, analyses showed estimated rates of "true" violation of transitivity was 1.8% for undergraduates

Table 10
Observed frequencies of each choice pattern for gambles *A*, *C*, and *E* with probability displayed as pies without and with probability text (follow-up to Experiment 2, *n* = 340)

Choice pattern	Graphs only			Graphs and text		
	Rep 1	Rep 2	Both	Rep 1	Rep 2	Both
000	16	12	7	3	3	1
001	12	12	10	11	13	10
010	9	5	0	4	5	1
011	3	3	2	1	0	0
100	38	30	19	15	15	8
101	2	8	1	2	1	0
110	118	125	108	102	101	93
111	3	6	0	1	1	0
Sum	201	201	147	139	139	113

Table 9
Test of dimension interaction (Experiment 2, *n* = 321)

Choice		Choice combination				Indep χ^2	True and error model		
R	*S*	*RR*	*RS*	*SR*	*SS*		$\hat{p}$	$\hat{e}$	χ^2
($7.25, 0.05; $1.25, 0.95)	($4.25, 0.05; $3.25, 0.95)	19	42	46	212	5.4	.95	0.17	0.2
($7.25, 0.95; $1.25, 0.05)	($4.25, 0.95; $3.25, 0.05)	220	40	37	23	16.2	.07	0.14	0.1

PH implies that people should choose "safe" gamble (*S*) in both cases; instead, nearly everyone chose the safe gamble in the first choice and the risky gamble (*R*) in the second, consistent with interactive models.

tested in the laboratory and 1.1% for Web recruits, respectively. With the graph only presentation format, the estimates are 5.8% and 5.8%, for undergraduates and Web recruits, respectively. We had conjectured that Web participants who did not perform other decision studies might be more likely to show intransitive behavior than undergraduates who were tested in the laboratory, but that hunch was not supported by the data. Nor was there any evidence that the two types of participants, undergraduates tested in the laboratory or more highly educated people recruited via the WWW, showed any systematic difference for transitivity.

Combining both groups in the graph only condition ($n = 361$), the fit of the general true and error model was $\chi^2(5) = 10.76$, an acceptable fit. The estimated true probabilities for the patterns 000 through 111 are, .06, .14, .01, .01, .09, .00, .69, and .00, respectively. However, we can reject the hypothesis that no one was intransitive in favor of the hypothesis that 5.8% were intransitive because the model with intransitive patterns set to zero yielded significant deviations, $\chi^2(7) = 79.55$; therefore, $\chi^2(2) = 68.79$. The estimate from this analysis indicated that 5.6% in the graphs only conditions showed the systematic pattern of intransitivity of the type observed by Tversky and compatible with the perceptual illusion hypothesis.

In sum, the best description of our data is that the vast majority of participants exhibit transitivity and dimension interaction. In addition, however, about 6% of participants in the graph only condition appear to show a perceptual illusion in which two pies that are similar are treated as the same, and these (equal) values are then used to compute utility by a model that has an interaction between probability and value.

Discussion

Although there is some evidence that a small percentage of participants may be systematically intransitive, our studies find that this percentage is much smaller than one would require for a model that is supposed to describe the majority, as the priority heuristic is supposed to do. In the first experiment where probability and prize information were both given numerically, the estimated percentage of intransitive participants was quite small, as was the estimated percentage in the second study when probability was displayed by pies accompanied by numerical information about probability. When probability was displayed by pies without numerical information, the rate of intransitivity was significant, but still accounted for only 6% of participants.

The success of transitivity in our data is compatible with findings of Birnbaum and Schmidt (submitted for publication), who tested for intransitivity predicted by the CRU and CDG models of Bleichrodt and Schmidt (2002, 2005). The true and error model gave a statistically acceptable fit to their data. They also found considerable individual differences among people in their transitive orders; there were four transitive orders that were used by significant subgroups of the sample. That study included "sure things" among the gambles compared, and perhaps that feature contributed to greater disagreement among participants in how to order such alternatives. In that study, so few people repeated intransitive orders that it was possible to retain the hypothesis that everyone was transitive in that study. Birnbaum and Schmidt also tested for intransitivity predicted by additive difference models (regret theory and majority rule), also finding that very few people repeated intransitive patterns. They found transitive behavior for both American and German samples.

The finding that participants in several experiments satisfy transitivity cannot prove that transitivity holds everywhere; there may be other gambles or other procedures that lead to systematic violations. However, the finding that most people are transitive with gambles that are predicted to show intransitivity by certain theories shifts the burden of proof to those who would argue for those theories; they should demonstrate that those theories are descriptive.

Interestingly, we found that the vast majority of the data that were internally consistent (cases where a person agreed with his or her own choices) also followed the transitive order of probability to win, EV, TAX, and CPT, all of which agreed in this study. The percentages were 81% and 83% in the two conditions of Experiment 1, and 80% in Experiment 2, respectively. Interestingly, this consensus among people occurred despite the fact that Tversky's gambles were designed to have nearly identical expected values. No one was observed in the laboratory making any calculations, and it seems doubtful that people used EV because in other tasks completed by the same laboratory participants, we found that most of the same people showed systematic violations of EV. For example, 70% of these same participants chose $S = (\$45, 0.5; \$35, 0.5)$ over $R = (\$100, 0.5; \$0, 0.5)$, despite the lower EV of S ($\$40$) compared to that of R ($\$50$).

The data do not agree with the predictions of PH, which implies that the majority should violate transitivity. That heuristic fails to predict the majority choice in seven of the ten choices among the gambles in Table 1 in all five groups (two conditions of Experiment 1, two conditions of Experiment 2 and the follow-up to Experiment 2).

Although the PH is supposed to be accurate in describing selected results of risky decision making (Brandstätter et al., 2006), it does not account for the "new paradoxes" described by (Birnbaum, 1999, 2004a, 2004b, 2005a, 2005b, 2005c; Birnbaum & Navarrete, 1998). These "new paradoxes" are stronger

than mere failures of prediction that might be saved by different parameters or different functions within CPT. They are cases where CPT leads to self-contradiction when it attempts to account for the results.

The new paradoxes were not designed to test PH, but PH fails to predict them. For example, the majority of participants violate stochastic dominance in the recipe devised by Birnbaum (1997) and tested by Birnbaum and Navarrete (1998). In a new variation of that test, for example, 71% of 270 participants chose $F = (\$89, 0.7; \$88, 0.1; \$11, 0.2)$ over $G = (\$90, 0.8; \$13, 0.1; \$12, 0.1)$. According to the priority heuristic, people should instead choose G over F because it has a lower probability of the worst consequence. In this choice, G stochastically dominates F; it has a higher lowest outcome, a higher best outcome, lower probability to get the worst consequence, a higher probability to get the best consequence, and a higher EV. The priority heuristic does not use the consequence on the middle branch, which is the only way in which F is better than G in this case.

Nor does PH predict the results of branch splitting in the Allais paradoxes (Birnbaum, 2004a). PH predicts the opposite pattern of violation of restricted branch independence from the observed pattern (Birnbaum, 2004a, 2004b; Birnbaum & Navarrete, 1998), and PH cannot explain violations of distribution independence (Birnbaum, 2005c). Of 32 choices analyzed by Birnbaum (submitted for publication-a) for the purpose of comparing TAX and CPT, CPT made wrong predictions in 17 cases, and the PH was wrong in 19 cases. Although these choices were devised to compare CPT and TAX, they also systematically violate PH. Brandstätter et al. (2006) concede that their model does not account for the new paradoxes and that it does not predict when people will satisfy or violate stochastic dominance.

Birnbaum (submitted for publication-b) has developed and tested three properties, priority dominance, dimension integration, and dimension interaction that are implied by a wider family of LS models. He found evidence in direct violation of LS models and also failed to find violations of intransitivity where predicted by PH.

Our second experiment tested dimension interaction. According to PH and additive difference models like the stochastic difference model, there should be no interaction between dimensions. Of 320 participants, 190 showed dimension interaction on both tests (four choices). Even among those 37 identified as possibly intransitive in Experiment 2, 27 showed systematic dimension interaction, contrary to LS models including PH.

Had the predicted pattern of intransitivity been prevalent and had the tests of dimension interaction in Experiment 2 conformed to predictions of PH, we would have had reasons to favor PH over TAX. Instead, our data show so few cases of intransitivity predicted by PH and such strong evidence of dimension interaction, that the model as stated by Brandstätter et al. (2006) must be seriously questioned as a descriptive model of risky decision making.

Brandstätter et al. (2006) reported that the PH is not accurate for choices in which the ratio of expected values within a choice falls outside the interval between 1/2 and 2. But the gambles used by Tversky (1969) differ in EV at most by 25%, so all choices in Tables 3 and 4 fall well inside the domain where PH is supposed to be accurate. Indeed, the 28 cases in Tables 3 and 4 where PH makes the wrong predictions are those with the smallest ratios of EV. In the tests of dimension interaction, the EVs in the first choice in Table 9 are $1.55 and $3.30 in the first choice (R versus S) and they are $6.95 and $4.20 ($R'$ versus S'). Ironically, PH fails in the second choice with the less extreme ratio (1.65) rather than the case with the more extreme ratio. Thus, the incorporation of EV into PH does not account for these data unless we retain EV and drop PH entirely.

Brandstätter et al. (2006, p. 425) conjectured that with low stakes, people might use the highest consequence as their first priority. According to this conjecture, if people considered all the consequences in these studies to be "low", they should have ordered the gambles $ABCDE$ instead of $EDCBA$. This conjecture therefore also fails to describe our majority results.

It seems reasonable to review the procedures used in this study, since our findings show less intransitivity compared to those reported by Tversky and others. The procedures used in Experiment 1 have been used previously to replicate standard violations of EV and EU such as risk aversion, the Allais common consequence paradox, Allais common ratio paradox, and other commonly reported results. They yield similar results to those reported by other investigators with other techniques (Birnbaum, 2001). Indeed, a number of variations of formats for displaying gambles and choices have been found to yield very similar and stable results (Birnbaum, 2004b, 2006).

Based on these results, we suspect that the intransitive behavior observed by Tversky (1969) and others holds only in very narrow circumstances that we have been unable to fully reproduce. Among those remaining differences, we suspect the following three are most important.

First, in Tversky (1969) each gamble was displayed on a card containing a pie chart without probability displayed numerically. Any choice required two cards, which were probably not aligned in a fixed position. In contrast, our pie chart displays were presented in an aligned arrangement on an unmoving computer screen (Fig. 1). Perhaps this better alignment allowed our participants to make sharper discriminations of probability from the pie charts than was possible for Tversky's

participants with pies printed on cards. Indeed, it is in the graph only format that we find the most evidence of intransitivity. However, Budescu and Weiss (1987) reported that four of eleven participants who viewed choices described in text on a computer screen also showed evidence of intransitivity by Tversky's lenient criterion.

Second, our pies were aligned with one side vertical, as shown in Fig. 1. Tversky's display was instead aligned with the bisector of the wedge representing probability aligned vertically. In either arrangement, the participant need only compare one edge of the pie slices; however, Tversky's orientation may make it more difficult to discriminate. Consider a clock set to 4:00 and another clock set to 3:00. The participant can ignore the hand facing 12:00, and see a 30 degree difference in the right-most ("hour") hands. However, in Tversky's arrangement, the right-most "hands" for this same 1/12 difference in probability would differ by only 15 degrees. This perceptual illusion interpretation suggests that if we reduced the probability differences in our stimuli or used Tversky's arrangement, we should observe a greater incidence in intransitivity than we observed in Experiment 2.

Third, our participants did not see as many repeated choices as in Tversky. Perhaps the incidence of intransitive behavior is greater in longer experiments.

To explain the small but significant incidence of intransitivity in our graph only condition, we think that when participants are presented a choice in which they are unable to discern any difference in probability given the graphic displays, they behave as though they substitute the same subjective probability into their interactive, transitive model. This interpretation would explain why people might show intransitivity and the same people show evidence of dimension interaction.

Thus, we conclude that violations of intransitivity in this situation are like evidence of friction in a test of the law of falling bodies. If one drops a feather and coin from a tower, the coin hits the ground well before the feather. But in a vacuum, both hit the ground at nearly the same time. We suspect that transitivity is like the law of falling bodies in which small objects fall with acceleration independent of their mass, and that factors that create small violations are like frictional forces in physics—these are forces to be reckoned with, but they are not the main story. We think the "friction" is created by the perceptual display in which participants have difficulty discriminating probability.

Transitivity does not entail nor does it necessarily presume rationality. Any theory in which people evaluate the gambles independently and tend (apart from error) to choose the gamble with the higher evaluation will conform (apart from error) to transitivity. The TAX model violates stochastic dominance in ways that we do not regard as rational and yet this model satisfies

transitivity. Conversely, it has been argued that some intransitive preferences can be compatible with normative principles (Fishburn, 1991, 1992).

We have identified conditions that lead to small, but significant violations of transitivity, but we have not been able to replicate higher rates such as reported by some others. These results lead to several new questions. First, most participants ordered the gambles by a transitive order that matched expected value, probability to win, TAX, and CPT. Is it possible to devise an experiment in which there is consensus for a transitive order that is different from that of EV and predicted by TAX? Second, would greater incidence of intransitivity be observed if prizes were of a greater magnitude? Third, would people become more intransitive if they served in a longer experiment? Fourth, will transitivity continue to hold up in studies with other recipes that imply intransitivity according to intransitive theories?

In summary, we tested for violations of transitivity where previously reported by Tversky and predicted from the PH. When data are analyzed using an error model in which people can have different "true" choice patterns, but vary in their responses to the same choices because of "errors", we find few exceptions to the hypothesis that everyone had a transitive preference order. Although there may be a few people in some situations where transitivity of preference fails, we do not think that our data merit rejection of transitive theories for studies in which both probability and prize information are presented numerically. Tests of dimension interaction refute three models that have been proposed to account for intransitive preferences: PH, LS, and additive difference models. Because our procedures have been used in other studies of risky decision making where people clearly violate implications of EU and CPT, we find it hard to believe that the major phenomena of risky decision making can be best explained by inherently intransitive choice processes.

Acknowledgments

References

Birnbaum, M. H. (1997). Violations of monotonicity in judgment and decision making. In A. A. J. Marley (Ed.), *Choice, decision, and*

measurement: Essays in honor of R. Duncan Luce (pp. 73–100). Mahwah, NJ: Erlbaum.

Birnbaum, M. H. (1999). Testing critical properties of decision making on the Internet. *Psychological Science, 10*, 399–407.

Birnbaum, M. H. (2001). A Web-based program of research on decision making. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 23–55). Lengerich, Germany: Pabst Science Publishers.

Birnbaum, M. H. (2004a). Causes of Allais common consequence paradoxes: an experimental dissection. *Journal of Mathematical Psychology, 48*(2), 87–106.

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

Birnbaum, M. H. (2005a). A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *Journal of Risk and Uncertainty, 31*, 263–287.

Birnbaum, M.H. (2005b). Testing properties of decision making models with fallible data. Working manuscript, available from M. Birnbaum.

Birnbaum, M. H. (2005c). Three new tests of independence that differentiate models of risky decision making. *Management Science, 51*, 1346–1358.

Birnbaum, M. H. (2006). Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *Journal of Economic Psychology, 27*, 737–761.

Birnbaum, M.H. (submitted for publication-a). New paradoxes of risky decision making. submitted for publication.

Birnbaum, M.H. (submitted for publication-b). Testing lexicographic semi-orders as models of decision making: priority dominance, dimension integration, dimension interaction, and transitivity.

Birnbaum, M. H., & Navarrete, J. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty, 17*, 49–78.

Birnbaum, M.H., & Schmidt, U. (submitted for publication). Empirical tests of intransitivity predicted by models of risky choice.

Bleichrodt, H., & Schmidt, U. (2005). Context- and reference-dependent utility. Working paper, available from U. Schmidt.

Bleichrodt, H., & Schmidt, U. (2002). A context-dependent model of the gambling effect. *Management Science, 48*, 802–812.

Bordley, R. F. (1992). An intransitive expectations-based Bayesian variant of prospect theory. *Journal of Risk and Uncertainty, 5*, 127–144.

Bordley, R., & Hazen, G. B. (1991). SSB and weighted linear utility as expected utility with suspicion. *Management Science, 37*, 396–408.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: choices without tradeoffs. *Psychological Review, 113*, 409–432.

Budescu, D. V., & Weiss, W. (1987). Reflection of transitive and intransitive preferences: a test of prospect theory. *Organizational Behavior and Human Decision Processes, 39*, 184–202.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision Field Theory: a dynamic cognition approach to decision making. *Psychological Review, 100*, 432–459.

Carbone, E., & Hey, J. D. (2000). Which error story is best? *Journal of Risk and Uncertainty, 20*(2), 161–176.

Fishburn, P. C. (1971). Lexicographic orders, utilities, and decision rules: a survey. *Management Science, 17*, 1442–1471.

Fishburn, P. C. (1982). Nontransitive measurable utility. *Journal of Mathematical Psychology, 26*, 31–67.

Fishburn, P. C. (1991). Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty, 4*, 113–134.

Fishburn, P. (1992). Nontransitive preferences and normative decision theory. In J. Geweke (Ed.), *Decision making under risk and uncertainty: New models and empirical findings* (pp. 3–10). Dordrecht: Kluwer Academic Publishers.

González-Vallejo, C. (2002). Making trade-offs: a probabilistic and context-sensitive model of choice behavior. *Psychological Review, 109*, 137–155.

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

Hausner, M. (1954). Multidimensional utilities. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes* (pp. 167–180). New York: Wiley.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica, 62*, 1291–1326.

Humphrey, S. J. (2001). Non-transitive choice: event-splitting effects or framing effects? *Economica, 68*, 77–96.

Iverson, G.J., Myung, J.I., & Karabatsos, G. (2006). Intransitivity of preference: Revisited. Working paper, available from Geoffrey J. Iverson, Institute for Mathematical Behavioral Sciences, UCI, Irvine, CA 92697.

Iverson, G., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences, 10*, 131–153.

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

Leland, J. W. (1994). Generalized similarity judgments: an alternative explanation for choice anomalies. *Journal of Risk and Uncertainty, 9*, 151–172.

Leland, J. W. (1998). Similarity judgments in choice under uncertainty: a re-interpretation of the predictions of regret theory. *Management Science, 44*, 659–672.

Lindman, H. R., & Lyons, J. (1978). Stimulus complexity and choice inconsistency among gambles. *Organizational Behavior and Human Performance, 21*, 146–159.

Loomes, G., Starmer, C., & Sugden, R. (1989). Preference reversal: information-processing effect or rational non-transitive choice? *Economic Journal, 99*, 140–151.

Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica, 59*, 425–440.

Loomes, G., & Sugden, R. (1982). Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal, 92*, 805–824.

Loomes, G., & Taylor, C. (1992). Non-transitive preferences over gains and losses. *The Economic Journal, 102*, 357–365.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica, 24*, 178–191.

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

Luce, R. D. (1994). Thurstone and sensory scaling: then and now. *Psychological Review, 101*, 271–277.

Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Mahwah, NJ: Lawrence Erlbaum Associates.

Luce, R. D., & von Winterfeldt, D. (1994). What common ground exists for descriptive, prescriptive, and normative utility theories? *Management Science, 40*(2), 263–279.

Marley, A. A. J., & Luce, R. D. (2005). Independence properties vis-à-vis several utility representations. *Theory and Decision, 58*, 77–143.

May, K. O. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica, 22*, 1–13.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In H. Jungerman & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 347–362). Dordrecht-Holland: D. Reidel.

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

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology, 49*, 205–225.

Regenwetter, M., Stober, C.D., Dana, J., & Kim, A. (2006). Tversky's intransitivity of preference revisited. Paper presented to Society for Mathematical Psychology, Vancouver, July 2006.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: a review of research on preferential choice. *Journal of Economic Literature, 35*, 631–661.

Rosenthal, R., & Fode, K. L. (1961). The problem of experimenter outcome-bias. In D. P. Ray (Ed.), *Series research in social psychology*. Washington, DC: National Institute of Social and Behavioral Science.

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

Starmer, C. (1999). Cycling with rules of thumb: an experimental test for a new form of non-transitive behavior. *Theory and Decision, 46*, 141–158.

Starmer, C., & Sugden, R. (1998). Testing alternative explanations of cyclical choices. *Economica, 65*, 347–361.

Stevenson, M. K., Busemeyer, J. R., & Naylor, J. C. (1991). Judgment and decision-making theory. In M. Dunnette & L. M. Hough (Eds.), *New handbook of industrial–organizational psychology* (pp. 283–374). Palo Alto, CA: Consulting Psychologist Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286 (Reprinted 1994, 101, 266–270).

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

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.

Viscusi, W. K. (1989). Prospective reference theory: toward an explanation of the paradoxes. *Journal of Risk and Uncertainty, 2*, 235–264.

Wu, G., Zhang, J., & Gonzalez, R. (2004). Decision under risk. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 399–423). Oxford: Oxford University Press.