Judgment and Decision Making, Vol. 3, No. 4, April 2008, pp. 304–316

New tests of cumulative prospect theory and the priority heuristic: Probability-outcome tradeoff with branch splitting

Michael H. Birnbaum*
Decision Research Center
California State University at Fullerton

Abstract

Previous tests of cumulative prospect theory (CPT) and of the priority heuristic (PH) found evidence contradicting these two models of risky decision making. However, those tests were criticized because they had characteristics that might "trigger" use of other heuristics. This paper presents new tests that avoid those characteristics. Expected values of the gambles are nearly equal in each choice. In addition, if a person followed expected value (EV), expected utility (EU), CPT, or PH in these tests, she would shift her preferences in the same direction as shifts in EV or EU. In contrast, the transfer of attention exchange model (TAX) and a similarity model predict that people will reverse preferences in the opposite direction. Results contradict the PH, even when PH is modified to include a preliminary similarity evaluation using the PH parameters. New tests of probability-consequence interaction were also conducted. Strong interactions were observed, contrary to PH. These results add to the growing bodies of evidence showing that neither CPT nor PH is an accurate description of risky decision making.

Keywords: choice, cumulative prospect theory, decision making, lexicographic semiorder, priority heuristic, prospect theory, utility.

1 Introduction

This paper compares three models that attempt to describe risky decision making. These models are cumulative prospect theory (CPT) (Tversky & Kahneman, 1992), Birnbaum's (1999) transfer of attention exchange model (TAX), and the priority heuristic (PH) of Brandstätter, Gigerenzer, and Hertwig (2006). The PH model is based on the idea that people compare one attribute at a time, such as the minimum prizes. In addition, the similarity model of Rubinstein (1988) as modified by Leland (1994) is also relevant to these studies, although these studies were not designed to test that model.

Birnbaum (1999; 2004b; 2008b) reviewed a number of critical tests that refute any rank dependent utility (RDU) model (Quiggin, 1993) including rank and sign-dependent utility (Luce & Fishburn, 1991; 1995; Luce, 2000), CPT, and Expected utility (EU). Birnbaum (2008a; 2008b) noted that many of the same tests that refute CPT also contradict the priority heuristic. For example, the priority heuristic predicted fewer than half of the modal choices analyzed by Birnbaum (1999), by Birnbaum (2004a), and by Birnbaum and Navarrete (1998).

Some of these choices included cases where 90% or more of the participants satisfied stochastic dominance but the priority heuristic predicts indifference. In other choices, significantly more than half of the participants (about 70% of undergraduates) violated stochastic dominance, but the priority heuristic predicts that people should satisfy it.

Brandstätter, Gigerenzer, and Hertwig (2008a) responded that properties of these choices may have induced people to use other heuristics drawn from a person's "adaptive toolbox." Presumably, decision makers first decide what rule to use, then they either apply that rule or choose to use another rule. The mechanism that decides what rule to use has not yet been specified; it is described instead with lists of "triggering conditions," which are estimated from data like parameters. Brandstätter et al. (2008a) concluded that the priority heuristic does not apply when there is a stochastic dominance relation in the choice.

In addition, Brandstätter et al. (2008a) argued that certain choices reviewed by Birnbaum (2008a) used gambles that differed in expected value (EV). Brandstätter et al. (2008a) presented a figure to show that the priority heuristic is not accurate when expected values (EVs) differ, which led them to suppose that two strategies are at

*I thank Jeffrey P. Bahra for assistance in testing the participants. Thanks are due to Eduard Brandstätter and Jonathan Leland for useful debate and discussion of these ideas. Address: Prof. Michael H. Birnbaum, Department of Psychology, CSUF H-830M, P.O. Box 6846, Fullerton, CA 92834–6846 Email: mbirnbaum@fullerton.edu

work, one for "easy" choices (that differ in EV) and one for "harder" choices where EVs are nearly equal. From the data, Brandstätter et al. (2008a) estimated that, when the ratio of EV exceeds 2, people act as if they choose the gamble with the higher EV. Brandstätter et al. consider EV ratio as a proxy for the "difficulty" of a choice, but do not necessarily hold that people actually compute ratios of EV. They argued that the priority heuristic is accurate for "difficult" choices in which EVs are nearly equal.

However, Birnbaum (2008c) noted that EV ratios in Birnbaum and Navarrete (1998) had been inside the region where PH is supposed to apply; in that study, the priority heuristic failed to reproduce even half of the modal choices correctly. Brandstätter et al. (2008a) replicated part of that study and their results confirmed that the priority heuristic reproduced fewer than half of the modal choices that they chose for replication (Birnbaum, 2008c). To account for the results, Brandstätter et al. noted that Birnbaum and Navarrete (1998) used many choices in which both gambles of a choice had the same probability distribution and in some choices two branches had the same probability. PH was not accurate for such choices, so Brandstätter, et al. (2008a) theorized that people use a "toting up" heuristic for choices in which two branches had the same probability. In some of the choices in Birnbaum and Navarrete (1998), there was a common probability-consequence branch in both choices, which was theorized to trigger editing rules and other heuristics that were called up to account for the failures of the priority heuristic.

The arguments of Brandstätter et al. (2008a) might also provide excuses for previous failures of CPT as well.

This paper devises a new type of test that avoids the exceptions stated above. In these tests, one alternative does not stochastically dominate the other, there are no common probability-consequence branches, probabilities of the consequences are not equal, and expected values are nearly equal. In addition, unlike previous tests, the new tests use shifts in expected value and expected utility to "help" predictions of PH and CPT. That is, expected value and expected utility are both manipulated such that, if a person shifts his or her judgments in the same direction as the changes in EU or EV, his or her choices will appear consistent with PH and CPT. However, the choices are designed so that the TAX model with parameters typical of previous research predicts that people will shift their choices in the opposite direction of EV, EU, CPT, and PH. To understand how the new test was devised, see Appendix A, which shows that the property tested can be deduced as a theorem from any rank dependent utility model, including CPT, and that the property also follows from the PH; and see Appendix B, which shows that the TAX model can systematically violate the property.

1.1 New test: Probability-outcome tradeoff and coalescing

The following choices illustrate the new test. In each case, the participant chooses the urn from which a ticket will be drawn at random from 100 otherwise identical tickets. The value printed on the ticket drawn determines the prize.

Choice 1: Would you prefer S or R?

S:	80 tickets to win \$66	R:	60 tickets to win \$92
	10 tickets to win \$8		10 tickets to win \$90
	10 tickets to win \$7		30 tickets to win \$7

Choice 2: Would you prefer S' or R'?

S':	70 tickets to win \$66	R':	70 tickets to win \$92
	10 tickets to win \$63		20 tickets to win \$8
	20 tickets to win \$7		10 tickets to win \$7

According to any RSDU, CPT, RDU, EU, or EV model, if a person prefers R over S, then that person should prefer R' over S', apart from random error. (In Appendix A, it is proved that, under any model in this class,

$$\begin{aligned} R = (x, p-r; x^-, r; z, 1-p) \succ \\ S = (y, q; z^+, s; z, 1-q-s) \Rightarrow \\ R' = (x, p; z', r'; z, 1-p-r') \succ \\ S' = (y, q-s'; y', s', z, 1-q), \end{aligned}$$

with $x > x^- > y > y' > z^+ > z \geq 0$. and all probabilities between 0 and 1.)

Three data patterns are compatible with the property, $R \succ S \Rightarrow R' \succ S'$: SS', SR', and RR'; however, no one should show the pattern RS' (except by error), according to any of these models. With parameters estimated by Tversky and Kahneman (1992), the CPT model predicts that people should prefer R over S and R' over S': the RR' pattern. The certainty equivalents according to that model are CE(R) = 50.9, CE(S) = 42.7, CE(R') = 51.2, and CE(S') = 42.3. However, it is important to keep in mind that CPT with any functions and parameters for utility and weighting implies the general property that rules out RS'.

According to the priority heuristic (PH), a person first compares lowest consequences of a gamble and chooses the gamble with the higher lowest consequence if they differ by more than 10% of the largest consequence in either gamble, rounded to the nearest prominent number (\$10 in this case). But the lowest consequences are equal in both choices (\$7). When the lowest consequences are not sufficiently different, the person supposedly chooses the gamble with the smaller probability to get the lowest consequence, if these differ by 0.1 or more. A person should therefore choose S and R' in these two choices

Table 1: Predicted choice combinations of different models in the new test of probability-consequence tradeoff with branch splitting. PH = priority heuristic; EV = expected value; CPT = cumulative prospect theory; LS = lexicographic semiorder.

		Choice 2:	
		Choose S'	Choose R'
Choice 1:	Choose S	Editing (rounding) + PH	Prior PH , "as if" EV + PH, & Modified similarity + PH
	Choose R	Predicted by prior TAX and by Leland's similarity model. Refutes any CPT.	Prior CPT; Several LS models

because they have the lowest probabilities of getting the worst prize; that is, the PH implies the pattern SR'.

If the probabilities of the lowest consequences differed by less than 0.1, the person is theorized to next compare the highest prizes and choose by that criterion. When there are more than two branches and the first three comparisons yield no decision, the person next compares the probabilities to win the highest prize and decides on that basis alone, if there is any difference. And if all four criteria yield no choice, the person chooses randomly. When gambles have three or more branches, the PH assumes that people never examine intermediate branches.

To account for previous failures of PH, Brandstätter et al. (2008a) postulated that people might edit choices using the editing rules of prospect theory (Kahneman & Tversky, 1979). Suppose people first rounded off consequences in the example above that differ by less than $5 and combined these branches with approximately equal consequences, and then applied the priority heuristic. If so, then people would choose S over R and S' over R'; i.e., the SS' pattern.

One can easily construct other lexicographic semiorder models in which the difference threshold for probability is greater than 0.1 or where highest prizes are compared before probabilities. In either of these models, a person would choose R over S and R' over S' (RR') because of the difference in the highest consequences of the two gambles. Because they argued against these alternatives, Brandstätter et al. (2006, 2008a) would not find these variations of their model to be attractive. Nevertheless, this study will investigate these possibilities as well.

Brandstätter et al. (2008b) cited the similarity heuristic of Rubinstein (1988) and Leland (1994) as another heuristic people might use. However, the similarity heuristic of Leland (1998) assumes that people first choose by EU, an assumption criticized by Brandstätter et al. (2006).

Brandstätter et al. (2008a; 2008b) modified the similarity model of Rubinstein (1988) and Leland (1994), and proposed that this modified similarity evaluation might precede the application of the priority heuristic. Besides excluding Leland's use of EU as the first step, they apparently reject the idea in Rubinstein (1988) and Leland (1994) that similarity may involve a nonlinear transformation between objective payoffs and a subjective scale of similarity. Instead, they theorized that people skip Leland's (1994) first step and apply the second and third steps using the same parameters as in the priority heuristic to the objective cash values and probabilities.

According to this modified similarity plus priority heuristic (described more precisely in Appendix C), people first check for transparent dominance and take the dominant alternative if there is one (Steps 1–10, Appendix C). Next, they compare consequences and probabilities to see if one gamble is favored by a cash difference of at least 10% of the maximal prize, rounded to nearest prominent number ($10 in this case) and the other gamble is not favored by a cash difference as large as $10 nor favored by a probability difference greater than or equal to 0.1. They also check to see if one gamble is favored by a probability difference greater than or equal to 0.1 and the other is not favored by such a difference or a difference in consequences as large as $10 (Steps 11–20, Appendix C). If this modified similarity heuristic determines no preference, it is theorized that people next use the priority heuristic (Steps 21–29, Appendix C).

In Choice 1, we see that the risky gamble has a higher best prize but the safe gamble has a higher probability to win the best prize. Because these conflicting differences exceed $10 and 0.1, this choice cannot be resolved by the similarity evaluation, so people should use the priority heuristic (last nine steps in Appendix C). In Choice 2, note that the risky gamble has a higher best prize but the safe gamble has the higher middle consequence (and both differences exceed $10), so the modified similarity heuristic also implies that people should apply the priority heuristic in Choice 2 as well. Thus, the modified similarity heuristic plus priority heuristic makes the same predictions as the priority heuristic in these choices because of the absence of a "no-conflict" solution.

Table 1 presents a summary of these predictions, PH implies SR'; editing (rounding and combining) + PH implies SS'; certain other lexicographic semiorders imply RR' (such as the priority heuristic with the assumption that probability differences must exceed 0.2 to be decisive). In addition, if people used the same parameters to examine similarity before applying the PH, they should also follow the pattern, SR'. But no version of the PH has yet been proposed that predicts RS'; indeed, the PH assumes that people never examine middle branches of gambles with more than two branches. It is the consequences on the middle branches that would justify choosing R and S' in this case.

According to the TAX model with parameters taken from prior research, people should prefer R over S and S' over R'; i.e., the pattern RS'. There are also reasonable parameters such that TAX could imply other patterns such as SS' or RR'. TAX thus differs from the CPT model and the heuristic models above in that it can predict RS' instead of SR'. Intuitively, the reason that TAX makes these predictions is that when a probability-consequence branch is split, the splinter branches have greater total weight than when they are combined. By splitting branches leading to higher (or lower) consequences, one can improve (or diminish) a gamble, respectively.

Although TAX and CPT have the same number of parameters, TAX can also handle the data pattern, SR', because EU is a special case of TAX. A "control" test is therefore included such that TAX with the same parameters implies SR'. The model and parameters should be able to predict cases in which we should observe SR' and predict cases in which we should find RS' (Appendix B).

The similarity model of Leland (1998) begins with an evaluation of EU, followed by a check that tests for transparent dominance. Next, a person chooses one alternative if it is superior and dissimilar on one or more attributes and if the other alternative is not noticeably better on any attribute. If we theorize that $90 is superior and dissimilar to $8 but that $92 is similar to $66 (and assuming .8 is similar to .6 and .1 is similar to .3 in probability), people might choose R in the first choice; and, if $63 is dissimilar and better than $8 and if $92 is similar to $66, people might choose S' in the second choice. This model makes similar predictions to TAX, except for the "control" choices.

1.2 Interactive independence

This study also includes new tests of a second property, called *interactive independence* proposed by Birnbaum (submitted) as a test of the family of lexicographic semiorders. In a lexicographic semiorder, any attribute that is the same in both gambles can be changed in both gambles without reversing preference. For example, consider Choices 3 and 4:

Choice 3: Do you prefer S'' or R''?
S'': 90 tickets to win $50 R'': 90 tickets to win $100
10 tickets to win $20 10 tickets to win $5

Choice 4: Do you prefer S''' or R'''?
S''': 10 tickets to win $50 R''': 10 tickets to win $100
90 tickets to win $20 90 tickets to win $5

Assuming no interaction between probability and consequences, $S'' \succ R'' \Leftrightarrow S''' \succ R'''$

According to PH, people should choose the "safe" gamble in both cases because the lowest consequence of the "safe" gambles always exceeds the lowest consequence of the "risky" gamble by $15, and this exceeds 10% of the highest consequence ($10). Thus, PH implies the data pattern, $S''S'''$. Other heuristic models in which the difference threshold exceeds $15 or in which people examine the highest consequences first could imply the pattern, $R''R'''$, but no lexicographic semiorder implies the interactive pattern, $R''S'''$.

Previous tests of interactive independence employed cases in which the ratio of EVs exceeded 2 (Birnbaum, submitted; Birnbaum & Gutierrez, 2007). In the new tests presented here, the EV ratio is always less than 2, so the priority heuristic is supposed to apply, according to Brandstätter et al. (2006; 2008a).

More generally, interactive independence of probability and consequences holds that

$$S'' = (x, p;\, y, 1-p) \succ R'' = (x', p;\, y', 1-p) \Leftrightarrow$$
$$S''' = (x, q;\, y, 1-q) \succ R''' = (x', q;\, y', 1-q)$$

where $x' > x > y > y' \geq 0$. According to TAX or CPT, people should violate interactive independence by showing interactions between probability and consequences. With parameters from previous research, both of these models imply the pattern, $R''S'''$, as do many other models, including EU. Interactive independence should be satisfied by any lexicographic semiorder (Birnbaum, submitted) as well as the priority heuristic and the similarity models of Rubinstein (1988) and Leland (1994), apart from the step in which EU is compared.

According to the similarity model (Rubinstein, 1988; Leland, 1994), if there is no transparent dominance relation, as is the case in Choices 3 and 4, people next look for a contrast favoring one gamble that is great enough to be dissimilar. Let $x' \succ_{\$} x$ indicate that the cash value x' is preferred to x and dissimilar, and let $x' \sim_{\$} x$ indicate that the difference in cash is not great enough to be dissimilar. Because the differences are the same in both choices, people either choose the "risky" gamble in both cases (assuming $100 $\succ_{\$}$ $50 but $20 $\sim_{\$}$ $5 in the example), or people should choose the "safe" gamble in both cases (if $100 $\sim_{\$}$ $50 but $20 $\succ_{\$}$ $5), or people should be indifferent in both choices (if \$100 $\succ_{\$}$ \$50 and \$20 $\succ_{\$}$ \$5 or if \$100 $\sim_{\$}$ \$50 and \$20 $\sim_{\$}$ \$5).

However, the Leland (1994) model includes EU theory, so it can violate interactive independence, given suitable parameters for the utility function and EU threshold. In addition, with free parameters for the similarity relations, the Leland (1994) model can imply the pattern RS' for Choice Problems 1 and 2. For example, suppose that \$66 $\sim_{\$}$ \$92 but \$8 $\prec_{\$}$ \$90, 0.8 $\sim_P$ 0.6, and 0.1 $\sim_P$ 0.3, where $\sim_P$ indicates that the probabilities are not different enough to be dissimilar. If so, people should prefer R over S in Choice 1. Similarly, with the assumptions above and \$63 $\succ_{\$}$ \$8, people should choose S' over R' in Choice 2. Leland (personal communication, February 26, 2008) predicted this pattern of results correctly before seeing the results.

2 Method

The method was similar to Birnbaum (1999). Embedded among a series of decision tasks were 30 choices of this study. These choices were presented twice, separated by other tasks that required about 10 minutes intervening between repetitions. Each choice was displayed via computer as follows:

First Gamble:
80 tickets to win \$66
10 tickets to win \$8
10 tickets to win \$7
OR
Second Gamble:
60 tickets to win \$92
10 tickets to win \$90
30 tickets to win \$7

Participants clicked a button beside the gamble they would rather play in each choice. They were informed that 3 participants (about 1 per fifty) would play one of their chosen gambles for real cash, so they should choose carefully.

Main design. The main design included eight choices (four pairs of choices) constructed from the following recipe:

$R = (x, p - r; x^-, r; z, 1 - p)$ or
$S = (y, q; z^+, s; z, 1 - q - s)$
and
$R' = (x, p; z', r'; z, 1 - p - r')$ or
$S' = (y, q - s'; y^-, s'; z, 1 - q)$

Levels of consequences $(x, x^-, y, y^-, z', z^+, z)$ were (\$95, \$90, \$55, \$52, \$12, \$12, \$11), (\$100, \$98, \$52, \$50, \$12, \$12, \$10), (\$100, \$98, \$50, \$48, \$4, \$3, \$2), and (\$92, \$90, \$66, \$63, \$8, \$8, \$7) in the four tests, respectively. Levels of probabilities $(p,\ q,\ r,\ s,\ r',\ s')$ were (0.65, 0.90, 0.05, 0.05, 0.30, 0.10), (0.20, 0.30, 0.10, 0.10, 0.10, 0.30), (0.10, 0.20, 0.05, 0.40, 0.30, 0.10), and (0.70, 0.80, 0.10, 0.10, 0.10, 0.20). These levels ($x > x^- > y > y^- > z' \geq z^+ > z \geq$ 0) ensure that R' dominates R and S dominates S' by first order stochastic dominance.

Control Choices: Two additional "control" choices were constructed from the same splitting manipulation as follows: S = (\$50, 0.2; \$12, 0.1; \$10, 0.7) versus R = (\$100, 0.05; \$90, 0.05; \$0, 0.9) and S' = (\$50, 0.15; \$40, 0.05; \$0, 0.8) versus R' = (\$100, 0.1; \$12, 0.2; \$10, 0.7). For these choices, however, TAX with parameters taken from previous research implies that people should choose S and R', as do CPT, EU, EV, and PH.

Interactive independence design. The second design, testing interactive independence, included choices of the following type:

R'' = $(x', p; y', 1 - p)$ or S'' = $(x, p; y, 1 - p)$

where $x' > x > y > y'$, and p is manipulated between choices. All ten choices used $(x',\ y')$ = (\$100, \$5). Five trials used $(x,\ y)$ = (\$50, \$20), where p = 0.1, 0.3, 0.5, 0.7, and 0.9. Five others used $(x,\ y)$ = (\$70, \$20), with the same 5 levels of probability.

In addition to 20 trials comprising the main designs, there were four warmup choices and six filler choices. These were presented in random order, restricted so that no two trials from the same design appear on successive trials. Complete materials are available from the URL: `http://psych.fullerton.edu/mbirnbaum/SPR_07/choice_gambles_F_07_01v2.htm`

Participants were 167 undergraduates who participated as one option toward an assignment in Introductory Psychology; 69% were female and 96% were 21 years of age or younger.

3 Results

Table 2 reports the number of people who showed each choice pattern for two replications of eight choices of the main design. From these data, the rates of "error" and the "true" probabilities of choosing the "risky" gamble can be estimated for each choice. The error rate for a choice is estimated from preference reversals when the same choice is repeated, using the model described in Appendix D, which allows each person to have a different true pattern of preferences and each item to have a different error rate.

The first row of Table 2 shows that 120 of the 167 participants chose the "risky" gamble on both presentations of the choice between S = (\$66, 0.8; \$8, 0.1; \$7, 0.1) and R = (\$92, 0.6; \$90, 0.1; \$7, 0.3), and only 15 chose S both times. According to PH, most people should have chosen S because it has a lower probability of yielding the smallest prize, and the difference in probability is 0.2,

Table 2: Replication data used to estimate true probability and error rates for each choice. Entries under SS, SR, RS, and RR are the observed numbers of people who showed each combination of choices on the two replications. For example, 120 people chose the risky gamble in both replicates of Choice 12 (first row of the table). The chi-squares in the right-most column evaluate the fit of the true and error model to these frequencies. All are acceptable fits.

No	Safe Gambles, S	Risky Gambles, R	p_R	e	SS	SR	RS	RR	$\chi^2(1)$
12	($66, 0.8; $8, 0.1; $7, 0.1)	($92, 0.6; $90, 0.1; $7, 0.3)	0.90	0.11	15	16	16	120	0
20	($66, 0.7; $63, 0.1; $7, 0.2)	($92, 0.7; $8, 0.2; $7, 0.1)	0.34	0.16	80	20	24	43	0.4
26*	($55, 0.9; $12, 0.05; $11, 0.05)	($95, 0.6; $90, 0.05; $11, 0.35)	0.87	0.20	20	32	21	94	2.3
15*	($55, 0.8; $52, 0.1; $11, 0.1)	($95, 0.65; $12, 0.3; $11, 0.05)	0.40	0.22	64	28	29	46	0.0
6	($52, 0.3; $12, 0.1; $10, 0.6)	($100, 0.1; $98, 0.1; $10, 0.8)	0.79	0.23	28	34	24	81	1.7
14	($52, 0.2; $50, 0.1; $10, 0.7)	($100, 0.2; $12, 0.3; $10, 0.5)	0.52	0.24	51	29	32	55	0.1
18*	($50, 0.2; $3, 0.4; $2, 0.4)	($100, 0.05; $98, 0.05; $2, 0.9)	0.59	0.18	50	21	27	69	0.7
10*	($50, 0.1; $48, 0.1; $2, 0.8)	($100, 0.1; $4, 0.3; $2, 0.6)	0.36	0.23	66	32	28	41	0.3
*28**	($50, 0.2; $12, 0.1; $10, 0.7)	($100, 0.05; $90, 0.05; $0, 0.9)	0.17	0.14	103	26	14	24	3.5
*30**	($50, 0.15; $40, 0.05; $0, 0.8)	($100, 0.1; $12, 0.2; $10, 0.7)	0.95	0.07	8	10	12	137	0.2

* Choices marked with asterisk had the risky gamble presented in the first position, with the safe gamble in the second position. Unmarked choices had the opposite arrangement. Choices 28 and 30 are the "control" choices in which all models agree.

which exceeds 0.1. Instead, 90% of the participants are estimated to prefer R, and estimated "error" rate is 11% on this choice.

The second row of Table 2 shows that 80 people chose S' = ($66, 0.7; $63, 0.1; $7, 0.2) over R' = ($92, 0.7; $8, 0.2; $7, 0.1) on both presentations compared to 43 who chose R' both times. The PH predicts that most people should prefer R' because of the 0.1 lower probability to get the lowest prize. However, 66% of the participants are estimated to truly prefer S', with an "error" rate of 16%.

Of the eight modal choices in the main design, PH is correct in only one case, Choice 14, where 52% are estimated to prefer the "risky" choice. PH correctly predicts the modal choices of the two "control" choices, as do TAX and CPT with their prior parameters.

The last column of Table 2 contains statistical tests of the true and error model; none is significant (α = 0.05), indicating that the model of error can be retained as descriptive.

Table 3 shows the results of the true and error model extended to choice combinations (Appendix D). This model assumes that each participant has one and only one of the four possible preference patterns for each test of the main design. Each participant exhibited one of sixteen possible observed data patterns (one of four possible patterns in each replication). Parameters were estimated by fitting the frequencies of these sixteen observed data frequencies so as to minimize, $G = 2\sum_{i=1}^{16} O_i \cdot ln(O_i/E_i)$, where O_i and E_i are respectively the obtained and predicted frequencies of the sixteen possible data patterns.

According to CPT or PH, no one should show the choice pattern RS', except by "error"; i.e., $p_{RS'} = 0$. According to TAX, however, people should indeed show this pattern. For the first test in Table 3, the observed frequencies of the 16 data patterns, $SS'SS'$, $SS'SR'$, $\dots$, $RR'RR'$ are 4, 3, 11, 2, 2, 6, 2, 1, 9, 4, 56, 11, 1, 2, 19, and 34. The modal pattern was $RS'RS'$, shown by 56 of the 167 participants. The first row in Table 3 displays the estimate of $p_{RS'}$ = 0.62; that is, most people (62%) are estimated to violate the predictions of CPT and PH on this test.

In all four tests in the main design, the estimated probability of RS' (bold font), which is inconsistent with any RDU or CPT or PH model is higher than that of the SR' pattern, which is predicted by PH and compatible with RDU and CPT/RSDU. For example, in the first test, 62% are estimated to have RS' as their true data pattern compared to 5% with the opposite reversal. A fair proportion of cases (29% in the first row) are consistent with the pattern RR', which is predicted by CPT with its prior parameters and which is also compatible with TAX (with other parameters) and with certain lexicographic semiorders, such as the one in which people compare highest consequences first.

Suppose people edited choices by rounding off consequences differing by less than $5 and combined branches leading to rounded consequences by adding their probabilities. If a person were to edit first and then apply PH, she or he would show the SS' pattern. However, Table 3

Table 3: Estimated True probabilities of each response pattern in the new tests of probability-outcome tradeoff with coalescing, monotonicity, and transitivity. The true and error model is tested by $\chi^2(10)$; all four show acceptable fits. The CPT and PH models imply that no one should show the RS' pattern of reversal, except by error. The $\chi^2(1)$ statistics in the right-most column test this hypothesis (that $p_{RS'} = 0$); all are large and significant, indicating systematic evidence against CPT and PH.

No	Safe Gambles, S and S'	Risky Gambles, R and R'	e	$p_{SS'}$	$p_{SR'}$	$p_{RS'}$	$p_{RR'}$	$\chi^2_{(10)}$	$\chi^2_{(1)}$
12	($66, 0.8; $8, 0.1; $7, 0.1)	($92, 0.6; $90, 0.1; $7, 0.3)	0.11	0.04	0.05	**0.62**	0.29	12.4	312.3
20	($66, 0.7; $63, 0.1; $7, 0.2)	($92, 0.7; $8, 0.2; $7, 0.1)	0.16						
26*	($55, 0.9; $12, 0.05; $11, 0.05)	($95, 0.6; $90, 0.05; $11, 0.35)	0.20	0.14	0.00	**0.44**	0.42	10.9	86.1
15*	($55, 0.8; $52, 0.1; $11, 0.1)	($95, 0.65; $12, 0.3; $11, 0.05)	0.22						
6	($52, 0.3; $12, 0.1; $10, 0.6)	($100, 0.1; $98, 0.1; $10, 0.8)	0.23	0.10	0.12	**0.38**	0.41	4.3	52.6
14	($52, 0.2; $50, 0.1; $10, 0.7)	($100, 0.2; $12, 0.3; $10, 0.5)	0.24						
18*	($50, 0.2; $3, 0.4; $2, 0.4)	($100, 0.05; $98, 0.05; $2, 0.9)	0.18	0.28	0.13	**0.36**	0.23	7.9	60.8
10*	($50, 0.1; $48, 0.1; $2, 0.8)	($100, 0.1; $4, 0.3; $2, 0.6)	0.23						
	Predictions of models with prior parameters:			Editing +PH	PH	TAX	CPT		

* Choices marked with asterisk have the risky gamble presented first, with the safe gamble in the second position. Entries in bold show estimates of $p_{RS'}$ these should be zero according to PH and CPT.

shows that only 4% were estimated to show this pattern in the first row, and in no case did the sum of SR' and SS' reach a majority in any row.

These analyses were conducted for each individual separately by adding the four tests of the main design (with two repetitions each) for each person. It was found that 117 of the 167 people (70%) showed more reversals of the type RS' than of SR'; 32 (19%) had more of SR', and the rest (11%) split evenly or showed no reversals. Thus, significantly more than half of individuals show the pattern predicted by TAX and which contradicts both the PH and CPT.

The last column in Table 3 presents statistical tests of the hypothesis that the true probability of the RS' pattern is 0, as implied by both PH and CPT. In all four cases, there is significant, systematic evidence violating those models. The magnitude of the violations appears to be larger in the first two tests in Table 3, where probability to win the highest consequence in S is high (0.8 and 0.9) and relatively smaller in the second two tests in Table 3.

With "control" choices, in contrast, the estimate is that 83% truly prefer S = ($50, 0.2; $12, 0.1; $10, 0.7) over R = ($100, 0.05; $90, 0.05; $0, 0.9), with an error rate of 14%, whereas only 5% truly prefer S' = ($50, 0.15; $40, 0.05; $0, 0.8) over R' = ($100, 0.1; $12, 0.2; $10, 0.7), with an error rate of 7%. These are consistent with PH, CPT, and TAX. This choice apparently creates difficulty for the Leland (1994) model because people should choose R over S because $90 is presumed superior and dissimilar to $12. (See Appendix E).

Table 4 presents the results of choices testing interactive independence, with the information arranged as in Table 2. For example, in the first row, it is shown that 99 people chose the "risky" gamble in both replications of Choice 7, and that 33 chose the "safe" gamble in both replications of this choice. The estimated proportion of people who truly prefer S = ($50, 0.9; $20, 0.1) over R = ($100, 0.9; $5, 0.1) was estimated by the true and error model to be 0.24. According to the priority heuristic, most people should have chosen the safe gamble, S, in every row in Table 4 because S always has a lowest prize of $20, which exceeds the lowest prize of R ($5) by more than $10. Instead, most people chose R in Choice 7 (first row). Similar results are observed in Choice 25 in the same table.

The probability to win the higher prize (which is the same in both gambles of each choice in Table 4) changes from 0.9 to 0.1 in each set of five successive choices in the table. According to any lexicographic semiorder, probability should have no effect in Table 4, because probability is the same in both gambles of each choice. Instead, estimated choice probabilities range from 0.24 to 0.87 in the first series (first five rows of Table 4) and from 0.29 to 0.84 in the second series (last five rows), contrary to this prediction. Tests of the true and error model are presented in the last column. Only the first one is significant, making one significant test out of the twenty tests in Tables 2 and 4. When the same type of error model is fit to these data as in Table 3, it is estimated that 63% of participants truly switched from R in Choice 7 to S in Choice 9, con-

Table 4: Replication data used to estimate true probability and error rate for each choice in the tests of interactive independence. Entries under RR, RS, SR, and SS show observed frequencies of each combination of choices on the two replications. According to EV + PH, the estimated choice percentages should be the same in all rows. According to either TAX or CPT, the probabilities of choosing the "safe" gamble should increase within each series, showing evidence of interaction between probability and prizes. The chi-squares in the right-most column evaluate the fit of the true and error model; only the first is significant.

No	Risky Gambles, R	Safe Gambles, S	p_S	e	RR	RS	SR	SS	$\chi^2(1)$
7	($100, 0.9; $5, 0.1)	($50, 0.9; $20, 0.1)	0.24	0.13	99	26	9	33	7.91
17	($100, 0.7; $5, 0.3)	($50, 0.7; $20, 0.3)	0.41	0.17	71	29	16	51	3.70
11	($100, 0.5; $5, 0.5)	($50, 0.5; $20, 0.5)	0.66	0.11	46	21	12	88	2.42
23	($100, 0.3; $5, 0.7)	($50, 0.3; $20, 0.7)	0.83	0.13	24	20	18	104	0.11
9	($100, 0.1; $5, 0.9)	($50, 0.1; $20, 0.9)	0.87	0.12	19	19	17	112	0.11
25	($100, 0.9; $5, 0.1)	($70, 0.9; $20, 0.1)	0.29	0.16	86	25	19	37	0.82
5	($100, 0.7; $5, 0.3)	($70, 0.7; $20, 0.3)	0.59	0.14	53	16	23	75	1.25
21	($100, 0.5; $5, 0.5)	($70, 0.5; $20, 0.5)	0.70	0.08	44	16	7	100	3.41
13	($100, 0.3; $5, 0.7)	($70, 0.3; $20, 0.7)	0.85	0.14	21	21	19	106	0.10
19	($100, 0.1; $5, 0.9)	($70, 0.1; $20, 0.9)	0.84	0.12	23	21	13	110	1.86

trary to the family of lexicographic semiorders, as well as the modified similarity heuristic plus priority heuristic.

Because EV ratios never exceed 2, these data refute the "as if EV" + PH model analyzed by Brandstätter et al. (2006, 2008a). Instead, the data show an interaction between probability and consequences: when probability to win the better consequence is high, people tend to choose the risky gamble with the higher best consequence. When probability to win is lower, people tend to choose the safe gamble with the better lowest consequence.

4 Discussion

Neither PH nor CPT provides an accurate description of the data in the main design (Tables 2 and 3). The new test does not use common branches, equal probabilities, large ratios of expected value, or changes in EU that oppose PH and CPT. Therefore, no model using EV, EU or CPT can account for the shifts in preference because the reversals of preference observed in Table 3 are opposite the changes in EV, EU, and CPT. The results are compatible with the TAX model in which splitting of the branch leading to the highest consequence improves the gamble and splitting of the lower branch lowers the evaluation of the gamble.

Tests of interactive independence (Table 4) show evidence of interaction that refutes PH and all lexicographic semiorder models. Evidence of interaction persists in these new tests despite the fact that EV ratios are less than 2 in all cases. Both CPT and TAX can account for this interaction, as do many other models, including EU.

Among the models considered here, PH is least accurate because it fails to describe results in Tables 2, 3, or 4. CPT correctly predicts the pattern in Table 4, but it fails to predict the results in Tables 2 and 3. The similarity model of Leland (1994, 1998) is more accurate than CPT because it can account for the data in the main design, and by means of its EU feature it could potentially handle the results in Table 4. TAX is most accurate because it correctly predicts the patterns in all three tables. In addition, the TAX model with previous parameters correctly predicted the "control" choices, where quantitatively stronger manipulations reversed the pattern observed in Tables 2 and 3.

Marley and Luce (2005) have shown that a family of gains decomposition utility (GDU) models has properties similar to those of TAX. Models in this family that violate coalescing remain consistent with results of Tables 2 and 3, and GDU can also account for Table 4. See Birnbaum (2007) for tests of certain of those models.

After seeing these results, a reviewer proposed a "striking difference" heuristic for Choices 12 and 20 of Table 3. This new heuristic introduces a new parameter defining a "striking difference," but as shown in Appendix E, no single value of the new parameter reconciles all of the data in Tables 2, 3, and 4, nor does it account for the "control" choices.

This study was not designed to test the similarity model of Leland (1998); however, that model remains compatible with these data, if we allow different parameters for different participants and allow EU to handle the results of Table 4. Unlike the striking difference model, the Leland model allows a nonlinear transformation to utility, so that similarity need not be a function of objective differences in consequences.

The simplest interpretation of these results combined with previous evidence is that neither CPT nor PH, with or without their editing rules of rounding and combination, is an accurate model of risky decision making. Nor is the PH plus the modified similarity heuristic compatible with these data. It seems possible that someone might construct yet another excuse consisting of a new "heuristic" that is "triggered" by the conditions of this experiment in order to explain why those theories failed to predict these results. But at this point, given the cumulative evidence against these models, the number of post hoc excuses seems to have grown too large, in my opinion, to retain those models as plausible descriptions.

These results do not disprove the "adaptive toolbox" approach because that approach cannot in principle be refuted. The "adaptive toolbox" approach assumes that people have different mental tools for handling different kinds of problems. The idea that people can evaluate a combination of stimuli by more than one operation is plausible and consistent with evidence (e.g., Birnbaum, 1982; Mellers, Ordóñez, & Birnbaum, 1992); the theory that people have only one way to compare or combine stimuli is not at issue. Instead, the issue appears to be that proponents of the "adaptive toolbox" view have restricted what is permitted to go in the toolbox. They have argued that people lack tools that have simple mathematical descriptions and allowed that the only tools available to humans and animals are those that have simple verbal descriptions. So far, the heuristic models have been successful in describing data that were already available previous to the construction of the heuristics, but the constructed models have not yet had success in predicting the results of a new experiment designed to test their implications.

For example, lexicographic semiorder models imply that once a decision is reached for any reason, no amount of difference in other attributes should have any effect. This property, termed *priority dominance* (Birnbaum, submitted), has not had success in handling real data. The lexicographic models also imply that small changes in two attributes, each of which is too small to reverse a decision, cannot combine to reverse a decision reached for another reason. This property, termed *integrative independence*, has also been disproved (Birnbaum & LaCroix, 2008). As shown in Birnbaum (submitted), Birnbaum and Gutierrez (2008), and in this study (Table 4), data violate the implication of interactive independence, which is implied by the priority heuristic, lexicographic semiorders, and the similarity evaluation model (apart from the use of EU as in Leland, 1994; 1998). Furthermore, the priority heuristic, lexicographic semiorders, and the similarity model can predict violations of transitivity, but such violations have been rare when the data are analyzed by the true and error model (Birnbaum & Gutierrez, 2007).

References

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Eds.), *Social attitudes and psychophysical measurement*, pp. 401–485). Hillsdale, N. J.: Erlbaum.

Birnbaum, M. H. (1997). Violations of monotonicity in judgment and decision making. In A. A. J. Marley (Eds.), *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. (2004a). Causes of Allais common consequence paradoxes: An experimental dissection. *Journal of Mathematical Psychology, 48*, 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. (2007). Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes, 102*, 154–173.

Birnbaum, M. H. (2008a). Evaluation of the priority heuristic as a descriptive model of risky decision making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*, 253–260.

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

Birnbaum, M. H. (2008c). Postscript: Rejoinder to Brandstätter et al. (2008). *Psychological Review, 115*, 260–262.

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

Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes, 71*, 161–194.

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

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

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

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

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008a). Risky Choice with Heuristics: Reply to Birnbaum (2008), Johnson, Schulte-Mecklenbeck, & Willemsen (2008) and Rieger & Wang (2008). *Psychological Review, 115*, 281–289.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008b). Postscript: Rejoinder to Johnson et al. (2008) and Birnbaum (2008). *Psychological Review, 115*, 289–290.

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.

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

Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first order gambles. *Journal of Risk and Uncertainty, 4*, 29–59.

Luce, R. D., & Fishburn, P. C. (1995). A note on deriving rank-dependent utility using additive joint receipts. *Journal of Risk and Uncertainty*, 11, 5–16.

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

Mellers, B. A., Ordóñez, L., & Birnbaum, M. H. (1992). A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes, 52*, 331–369.

Quiggin, J. (1993). *Generalized expected utility theory: The rank-dependent model.* Boston: Kluwer.

Rubinstein, A. (1988). Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?). *Journal of Economic Theory, 46*, 145–153.

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

Appendix A

Let $(x,\ p;\ y,\ q;\ z,\ 1\ {-p}-q)$ represent a gamble with probabilities, p, q, and $1\ {-p}-q$ to win x, y, or z, respectively, where $x > y > z \geq 0$. Any RDU or RSDU model, including CPT, implies the following property:

$$\begin{aligned} R = (x, p-r; x^{-}, r; z, 1\ {-p}) &\succ \\ S = (y, q; z^{+}, s; z, 1\ {-q}-s) &\Rightarrow \\ R' = (x, p; z', r'; z, 1-p-r') &\succ \\ S' = (y, q-s'; y^{-}, s'; z, 1-q) & \end{aligned}$$

where $x > x^{-} > y > y^{-} > z^{+}$, $z' > z \geq 0$, and all of the branch probabilities are less than 1 and greater than 0.

Proof: These models imply transitivity, coalescing, and consequence monotonicity (Birnbaum & Navarrete, 1998). *Transitivity* assumes that if $R \succ Q$ and $Q \succ S$ then $R \succ S$. *Coalescing* assumes that if two branches yield the same consequence, they can be combined; therefore, $A = (x,\ p;\ x,\ q;\ z,\ 1-p-q) \sim A' = (x,\ p+q;\ z,\ 1-p-q)$, and B = $(x, p; z, q; z, 1-p\ {-q}) \sim B' = (x,\ p; z, 1-p)$. *Monotonicity* assumes that increasing a consequence, holding everything else constant, improves the gamble. Thus, $A^{+} = (x^{+}, p; y, q; z, r) \succ A = (x, p; y, q; z, r) \Leftrightarrow x^{+} \succ x$; similarly, improving y or z in Gamble A would also improve A.

By monotonicity, $R \succ S$ implies $(x,\ p-r;\ x,\ r;\ z,\ 1-p) \succ R \succ S \succ (y, q; z, s; z, 1-q\ {-s})$. By coalescing and transitivity, we have $(x, p; z, 1-p) \succ R \succ S \succ (y, q; z, 1-q)$. Applying coalescing and transitivity again, we can split the gambles as follows: $(x, p; z, r'; z, 1-p-r') \succ R \succ S \succ (y,\ q-s';\ y,\ s';\ z,\ 1-q)$. By consequence monotonicity, $R' \succ (x, p; z, r'; z, 1-p-r') \succ R \succ S \succ (y,\ q-s';\ y,\ s';\ z,\ 1-q) \succ S'$. By transitivity, $R' \succ S'$. This construction is similar to that of Birnbaum (1997) in that it uses the properties of transitivity, coalescing, and consequence monotonicity. However, although R' dominates R and S dominates S', there is no stochastic dominance relation between R and S or between R' and S'.

Because EU and EV are special cases of CPT, EU and EV also imply this same property.

Priority Heuristic: In the same recipe, choose x, p, y, q, and z, such that $x{\cdot}p = y{\cdot}q$ and $z = 0$. With these constraints, expected values are equal; therefore, PH is supposed to be applicable to such choices. To make EU more nearly equal, we can increase z in both gambles and decrease the prize in the "safer" gamble (y) a bit to adjust for risk aversion. Next, select levels (of p, q, r, s, r', and s') so that all probabilities are greater than 0 and less than 1, and so that $(1-p)-(1-q-s) \geq 0.1$ and $(1-q)-(1-p-r') \geq 0.1$. According to PH, people should choose $S \succ R$ and $R' \succ S'$.

Appendix B: The TAX model

The "special TAX model" can be written for three-branch gambles, $G = (x_1, p_1; x_2, p_2; x_3, p_3)$, where $x_1 \geq x_2 \geq x_3 > 0$, when $\delta > 0$, as follows:

$$U(G) = \frac{Au(x_1) + Bu(x_2) + Cu(x_3)}{A + B + C} \qquad (1)$$

where

$A = t(p_1) - 2\delta t(p_1)/4$
$B = t(p_2) - \delta t(p_2)/4 + \delta t(p_1)/4$
$C = t(p_3) + \delta t(p_1)/4 + \delta t(p_2)/4$

Eq. 1 is a weighted average of consequence utilities, where weights depend on probabilities of the consequences and ranks of the consequences on discrete branches. In practice, the weighting function is approximated by $t(p) = p^\gamma$, where $0 < \gamma < 1$ (a typical value is 0.7), and $u(x) = x^\beta$, where $0 < \beta \leq 1$. Birnbaum and Chavez (1997) reported that the median estimated value of $\beta = 0.61$, whereas Birnbaum and Navarrete (1998) reported that a median estimate of $\beta = 0.41$. Many data can be roughly approximated with $\beta = 1$ when consequences are small (e.g., Birnbaum, 1999); however, optimal estimates of β for the data fit by Brandstätter et al. (2006) are also less than 1 (Birnbaum, 2008a). This TAX model is the same as in Birnbaum (1999), except a notational convention has been changed so that $\delta > 0$ here corresponds to $\delta < 0$ in Birnbaum (1999) and earlier papers.

According to TAX, with $\gamma < 1$ and $\delta > 0$, splitting the branch leading to the best consequence increases the relative weight of that consequence, making the gamble better. Splitting the branch leading to the worst consequence increases its weight, making the gamble worse. Start with $R_0 = (x, p; z, 1 - p)$ and $S_0 = (y, q; z, 1 - q)$, with levels chosen as in Appendix A. Create R by splitting the branch leading to x, and create S by splitting the branch leading to z. Next, reduce the consequence of the splinter in R (i.e., x^- is slightly less than x) and slightly increase z in S ($z^+ > z$); if these adjustments in consequences are small, TAX still predicts that people should choose R over S. Next, create R' and S' by splitting branch leading to z in R_0 and splitting the branch of S_0 leading to y. Again, adjust consequences on the splinters slightly so that TAX predicts $S' \succ R'$. For example, with parameters of $\beta = 0.6$, $\gamma = 0.7$, and $\delta = 1$, TAX values for S and R in Choice 12 of Table 2 (Choice 1 of the Introduction) are 22.1 and 40.3, respectively. The predictions for S' and R' in Choice 20 of Table 2 (Choice 2 of the Introduction) are 33.8 and 24.7, respectively. Based on the same parameters, TAX predicts the same pattern, RS', in the other three tests in Table 2 (Choices 26 and 15, 6 and 14, and 18 and 10) as well. However, these same parameters predict the opposite reversal in the "control" choices, where TAX values are 15.3 for S = ($50, 0.2; $12, 0.1; $10, 0.7) and 14.6 for R = ($100, 0.05; $90, 0.05; $0, 0.7), respectively, 10.4 for S' = ($50, 0.15; $40, 0.05; $0, 0.8), and 17.34 for R' = ($100, 0.1; $12, 0.2; $10, 0.7). In summary, TAX with the same parameters predicts SR' for the "control" choices in which larger changes in consequences are pitted against the splitting manipulation, but it predicts the pattern RS' for the four tests in the main design.

Appendix C: Modified similarity plus priority heuristic

This model was suggested by Brandstätter et al. (2008a) as a way to account for data that violate the priority heuristic. It applies when there are exactly three branches in both gambles: $G = (x,\ p;\ y,\ q;\ z,\ 1 - p - q)$; $F = (x',\ p';\ y',\ q';\ z',\ 1 - p' - q')$; where $x > y > z \geq 0$, and $x' > y' > z' \geq 0$. Unlike the priority heuristic, this model assumes that people attend to the middle branch of a three-branch gamble. It is assumed that $\Delta = \delta$ = rounded value of $\max(x, x')/10$ and that $\Delta_p = \delta_p = 0.1$. All branch probabilities are between 0 and 1 and sum to 1 (none are zero). Steps 1–10 test for transparent dominance; Steps 11–20 are similar to Steps 1–10 but test for similarity; Steps 21–29 are the priority heuristic, and must be done in the order specified.

1. If $x - x' > 0$, $y' - y \leq 0$, $z' - z \leq 0$, $p' - p \leq 0$, and $q' - q \leq 0$, choose G
2. Else if $x' - x > 0$, $y - y' \leq 0$, $z - z' \leq 0$, $p - p' \leq 0$, and $q - q' \leq 0$, choose F
3. Else if $y - y' > 0$, $x' - x \leq 0$, $z' - z \leq 0$, $p' - p \leq 0$, and $q' - q \leq 0$, choose G
4. Else if $y' - y > 0$, $x - x' \leq 0$, $z - z' \leq 0$, $p - p' \leq 0$, and $q - q' \leq 0$, choose F
5. Else if $z - z' > 0$, $x' - x \leq 0$, $y' - y \leq 0$, $p' - p \leq 0$, and $q' - q \leq 0$, choose G
6. Else if $z' - z > 0$, $x - x' \leq 0$, $y - y' \leq 0$, $p - p' \leq 0$, and $q - q' \leq 0$, choose F
7. Else if $p - p' > 0$, $x' - x \leq 0$, $y' - y \leq 0$, $z' - z \leq 0$, and $q' - q \leq 0$, choose G
8. Else if $p' - p > 0$, $x - x' \leq 0$, $y - y' \leq 0$, $z - z' \leq 0$ and $q - q' \leq 0$, choose F
9. Else if $q - q' > 0$, $x' - x \leq 0$, $y' - y \leq 0$, $z' - z \leq 0$, and $p' - p \leq 0$, choose G
10. Else if $q' - q > 0$, $x - x' \leq 0$, $y - y' \leq 0$, $z - z' \leq 0$, and $p - p' \leq 0$, choose F

11. Else if $x - x' \geq \Delta$, $y' - y < \Delta$, $z' - z < \Delta$, $p' - p < \Delta_p$, and $q' - q < \Delta_p$, choose G
12. Else if $x' - x \geq \Delta$, $y - y' < \Delta$, $z - z' < \Delta$, $p - p' < \Delta_p$, and $q - q' < \Delta_p$, choose F
13. Else if $y - y' \geq \Delta$, $x' - x < \Delta$, $z' - z < \Delta$, $p' - p < \Delta_p$, and $q' - q < \Delta_p$, choose G
14. Else if $y' - y \geq \Delta$, $x - x' < \Delta$, $z - z' < \Delta$, $p - p' < \Delta_p$, and $q - q' < \Delta_p$, choose F
15. Else if $z - z' \geq \Delta$, $x' - x < \Delta$, $y' - y < \Delta$, $p' - p < \Delta_p$, and $q' - q < \Delta_p$, choose G
16. Else if $z' - z \geq \Delta$, $x - x' < \Delta$, $y - y' < \Delta$, $p - p' < \Delta_p$, and $q - q' < \Delta_p$, choose F
17. Else if $p - p' \geq \Delta_p$, $x' - x < \Delta$, $y' - y < \Delta$, $z' - z < \Delta$, and $q' - q < \Delta_p$, choose G
18. Else if $p' - p \geq \Delta_p$, $x - x' < \Delta$, $y - y' < \Delta$, $z - z' < \Delta$, and $q - q' < \Delta_p$, choose F
19. Else if $q - q' \geq \Delta_p$, $x' - x < \Delta$, $y' - y < \Delta$, $z' - z < \Delta$, and $p' - p < \Delta_p$, choose G
20. Else if $q' - q \geq \Delta_p$, $x - x' < \Delta$, $y - y' < \Delta$, $z - z' < \Delta$ and $p - p' < \Delta_p$, choose F
21. Else if $z - z' \geq \delta$, choose G
22. Else if $z' - z \geq \delta$, choose F
23. Else if $(1 - p' - q') - (1 - p - q) \geq \delta_p$, choose G
24. Else if $(1 - p - q) - (1 - p' - q') \geq \delta_p$, choose F
25. Else if $x - x' \geq \delta$, choose G
26. Else if $x' - x \geq \delta$, choose F
27. Else if $p - p' \geq \delta_p$, choose G
28. Else if $p' - p \geq \delta_p$, choose F
29. Else choose randomly.

The parameters of Brandstätter et al. (2006, 2008a, 2008b) are $\Delta_p = \delta_p = 0.1$ and $\Delta = \delta = \$10$ (in this study), which are derived from the base 10 number system. This model does not account for the data in Tables 2, 3, and 4. The striking difference heuristic (Appendix E) might be interpreted as repetition of Steps 11–20 with new values of $\Delta \neq \delta$ and $\Delta_p \neq \delta_p$.

Appendix D: True and error model

The "true and error" model provides a null hypothesis to test if violations of a theory might be due to random "error" (Birnbaum & Gutierrez, 2007). Suppose that different people may have different "true" preferences, but each person may make an "error" in discovering or reporting her or his true preference on any given trial. Presenting the choice between S and R twice, the probability that a person will reverse from S to R is as follows:

$$p(SR) = p_S(1 - e)e + p_R e(1 - e) = e(1 - e)$$

where $p(SR)$ is the probability of this reversal, p_S, is the probability that the person truly prefers S, $p_R = 1 - p_S$ is the probability that the person truly prefers R, and e is the "error" rate for this choice. It is assumed that $0 \leq e < 1/2$ and that errors are independent. The person who truly prefers S has made a correct decision on the first presentation of the choice and has made an "error" on the second presentation. Similarly, $P(RS) = (1 - e)e$. This expression shows that one can estimate error rates strictly from preference reversals between repeated presentations of the same choices.

The probability that a person chooses R both times is as follows:

$$p(RR) = p_S e^2 + p_R(1 - e)^2;$$

those who truly prefer S made two errors and those who truly prefer R made two correct reports. This model has two parameters, p_S and e, to fit the frequencies of four data patterns, SS, SR, RS, and RR, which sum to the number of participants. There is one degree of freedom remaining to test the fit of this model.

This model can be extended to a behavioral property with two choices (e.g., S versus R and S' versus R' as in Choices 1 and 2 of the introduction), in which there are two replications of each choice. This model is used to estimate "true" probabilities of the four possible preference patterns, fitting the frequencies of the sixteen possible observed data patterns for two replications of two choices, from $SS'SS'$, $SS'SR'$, $SS'RS'$, $\ldots$, $RR'RR'$.

Each of these 16 predicted probabilities is the sum of four terms, representing the four possible true patterns. For example, the probability of the observed data pattern $RS'RS'$ is as follows:

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

where e is the error rate for the choice between S and R, and e' is the error rate on the choice between S' and R'; $p_{SS'}$, $p_{SR'}$, $p_{RS'}$, and $p_{RR'}$ are "true" probabilities of the four patterns. There are fifteen other equations like the above for the other 15 possible data patterns.

Because frequencies of the sixteen data patterns sum to n, there are 15 degrees of freedom in the data. Three degrees of freedom are used to estimate $p_{SS'}$, $p_{SR'}$, $p_{RS'}$, with the fourth determined (because the four sum to 1).

Two degrees of freedom are used for error rates for the two choices, leaving 10 degrees of freedom to test fit of the true and error model.

According to the priority heuristic and the family of CPT/RSDU/RDU/EU/EV Models, $p_{RS'} = 0$, and therefore, the only occurrences of RS' result from "errors." This special case leaves one additional degree of freedom, yielding a Chi-Square test with 1 df comparing the fit of this special case against that of the general model. According to the PH, $p_{SS'}$ = $p_{RR'}$ = $p_{RS'}$ = 0. If we assume that some people might use editing ($p_{SS'} > 0$) and if others use a lexicographic semiorder with a different priority order or probability threshold ($p_{RR'} > 0$), this larger collection of models would make the same predictions as CPT with all parameters free; namely, $p_{RS'} = 0$. These χ^2 values are presented in Table 3: all four tests are large and significant.

5 Appendix E: Striking difference heuristic and similarity

A reviewer proposed the following new heuristic (or new parameters) to reconcile the results in Choices 12 and 20 of Table 2. In Choice 12, there is a "striking difference" between \$90 and \$8. In Choice 20, the "striking difference" is between \$63 and \$8. Presumably, when there is a striking difference, the person chooses the gamble with the better value on the striking difference. If we theorize that this heuristic precedes the priority heuristic, it might provide an excuse for the failure of the priority heuristic in the main design. For Choices 26 and 15, the difference between \$52 and \$12 (\$40) is striking but the difference between \$95 and \$55 (also \$40) is not. For Choices 6 and 14, the choice between \$50 and \$12 (\$38) is striking but the difference between \$100 and \$52 (\$48) is not.

Next, examine Choice 7 in Table 4. This choice has a difference of \$100 versus \$50 in the highest consequences. If this difference is "striking," (and \$20 versus \$5 is not), it would explain why people choose the risky gamble in this choice, instead of the "safe" gamble. The priority heuristic predicts that people should choose the "safe" gamble because its lowest consequence is more than \$10 higher than that of the risky gamble. But if the difference between \$100 and \$50 is striking, the risky gamble in Choice 9 should also have been chosen, which also contains this same difference. Instead only 13% do so, contradicting the striking difference heuristic as the explanation of the failure of the priority heuristic. Similarly if the difference between \$100 and \$50 is striking, people should have chosen the risky gamble in Choice 10, whereas only 36% do so. In addition, people should have chosen R = (\$100, 0.05; \$90, 0.05; \$0, 0.9) over S = (\$50, 0.2; \$12, 0.1; \$10, 0.7) because of two striking differences favoring R (\$90 versus \$12 and \$100 versus \$50); instead, only 17% made this choice.

The model of Brandstätter et al. (2006, 2008a, 2008b) assumes that a difference of \$40 cannot be both striking and not striking, because it assumes that people use cash differences rather than differences in utility. In the model of Leland (1994), however, similarity need not be a function of differences in cash value.

If we reject the priority heuristic with or without these variations, we can reconcile these results with Leland's (1998) similarity heuristic, which includes an evaluation of EU. To do this, we estimate parameters from these data, which gives a fairly good account of these results except for the "control" choices.

According to the similarity model, we could use the following parameters to fit the modal choices: from Table 2: \$90 $\succ_\$$ \$8; \$63 $\succ_\$$ \$8, \$92 $\sim_\$$ \$66, 0.8 $\sim_P$ 0.6, 0.3 $\sim_P$ 0.1, 0.2 $\sim_P$ 0.1) (\$90 $\succ_\$$ \$12, \$52 $\succ_\$$ \$12, \$95 $\sim_\$$ \$55, .9 $\sim_P$ 0.6, 0.35 $\sim_P$ 0.05), (\$98 $\succ_\$$ \$12, 0.8 $\sim_P$ 0.6, 0.3 $\sim_P$ 0.1), (\$98 $\succ_\$$ \$3, \$48 $\succ_\$$ \$4, \$100 $\sim_\$$ \$50) in Table 2. While this similarity model is more accurate than the priority heuristic for Tables 2 and 3, it does not provide a consistent account of the results in Table 4. From Table 4, we require that \$100 $\succ_\$$ \$50, \$100 $\succ_\$$ \$70, but \$20 $\sim_\$$ \$5. The Leland (1994) model, by means of EU, however, could potentially reconcile the main trend in Table 4.

In Choice 26, the difference between \$90 and \$12 must be dissimilar to explain why most people chose the risky gamble. If so, then in the first "control" choice, the majority should have preferred R = (\$100, 0.05; \$90, 0.05; \$0, 0.9) over S = (\$50, 0.2; \$12, 0.1; \$10, 0.7) because it has the dissimilar difference (\$90 versus \$12) on its middle branch. However, only 17% conformed to this preference, contradicting the hypothesis that \$90 versus \$12 is a dissimilar difference. Except for this choice, the Leland (1994; 1998) model, like the TAX model, cannot be rejected as a potential description of these results.