3 Generalization Across People, Procedures, and Predictions: Violations of Stochastic Dominance and Coalescing

Michael H. Birnbaum and Teresa Martin

ABSTRACT

Stochastic dominance is implied by certain normative and descriptive theories of decision making. However, significantly more than half of participants in laboratory studies chose dominated gambles over dominant gambles, despite knowing that some participants would play their chosen gambles for real money. Systematic event-splitting effects were also observed, as significantly more than half of the participants reversed preferences when choosing between the split versions of the same choices. Similar violations were found with five different ways of displaying the choices. Studies conducted via the Web show that the effects generalize beyond the laboratory, even to highly educated people who have studied decision making. Results are consistent with configural weight models, which predict violations of stochastic dominance and coalescing, but not with rank- and sign-dependent utility theories, including cumulative prospect theory, which must satisfy these properties. This research program illustrates three directions for testing the generality of theories – generality across people, procedures, and new predictions.

CONTENTS

Support was received from National Science Foundation Grants SBR-9410572 and SES 99-86436. Experiment 3 is from a master's thesis by the second author (Martin, 1998) under the direction of the first. We thank Sandra Schneider for suggesting the test of graphic displays (Experiments 4 and 5), Michele Underwood for her assistance during pilot testing of pie chart displays, and Christof Tatka for suggesting a reversal of the branch order (Experiment 5).

In choosing between risky gambles, it is eminently rational to obey stochastic dominance. If gamble *A* always gives at least as high a prize as gamble *B* and sometimes better, gamble *A* is said to dominate gamble *B*. Few deny that one should choose the dominant gamble over the dominated gamble once one comprehends dominance.

Stochastic dominance is not only a rational principle of decision making, it is also imposed by descriptive theories that are supposed to predict the empirical choices that people make. Kahneman and Tversky (1979) proposed that people detect and reject dominated gambles in an editing phase that precedes evaluation. Rank-dependent expected utility (RDEU) theory (Quiggin, 1982, 1993), rank- and sign-dependent utility (RSDU) theory (Luce, 2000; Luce & Fishburn, 1991, 1995), cumulative prospect theory (CPT) (Tversky & Kahneman, 1992; Wakker & Tversky, 1993), lottery-dependent utility theory (Becker & Sarin, 1987), and others (e.g., Camerer, 1992; Lopes & Oden, 1999; Machina, 1982) assume or imply that people obey stochastic dominance.

Therefore, the finding by Birnbaum and Navarrete (1998) that there are choices in which 70% of the people tested violate stochastic dominance is not only upsetting to the view that people are rational, but also disproves descriptive theories that retain stochastic dominance. This finding was not the result of happenstance. The choices tested by Birnbaum and Navarrete had been designed by Birnbaum (1997) to violate stochastic dominance, according to configural weight models and parameters fit to previous data involving choices that tested other properties.

Recipe for Violations of Stochastic Dominance

Birnbaum (1997) noted that configural weight models known as the *rank-affected multiplicative* (RAM) and *transfer of attention exchange* (TAX) models imply violations of stochastic dominance in a recipe that can be illustrated by the following example. Start with $G_0 = (\$12, .1; \$96, .9)$, a gamble with a .1 probability of winning \$12 and a .9 probability of winning \$96. Now split the lower branch of G_0 (.1 to win \$12) to create a slightly better gamble, $G+ = (\$12, .05; \$14, .05; \$96, .9)$. Next, split the higher branch of G_0 to create a slightly worse gamble, $G- = (\$12, .1; \$90, .05; \$96, .85)$. Clearly, $G+$ dominates G_0, which dominates $G-$.

The RAM model with parameters of Birnbaum and McIntosh (1996) and the TAX model with parameters of Birnbaum and Chavez (1997) both predict that people should prefer $G-$ to $G+$, violating stochastic dominance. The RAM and TAX models are configural weight models that represent the subjective values of gambles by their weighted averages, with weights that are affected by ranks. The theories allow a subjective function of prizes, $u(x)$, and a weighting function of probability, $S(p)$; in addition, they allow configural weighting that is affected by the ranks of the branches' consequences.

In the RAM and TAX models, each branch (each distinct probability–payoff combination) of a gamble carries some weight. When a fixed probability is split to create two branches from one, the sum of the weights of the two separate branches can exceed the weight of the coalesced branch, unlike the RDEU models. These configural weight models have some similarity to the RDEU models in that weights are affected by ranks, but the definition of *ranks* differs between the two approaches. In the RDEU models, cumulative weight is a monotonic function of cumulative probability (rank); however, in RAM and TAX, it is the distinct probability-consequence *branches* in the display that have "ranks" and are carriers of weight.

To illustrate the TAX model, assume that a gamble's utility is a weighted average of the utilities of its consequences. Suppose, for simplicity, that subjective probability is proportional to objective probability, and suppose that utilities are proportional to monetary consequences. So far, we have expected value. We now add the key idea: Suppose in three-branch gambles that any branch with a lower-valued consequence "taxes" (or "takes") one-fourth of the weight of any distinct branch with a higher-valued consequence. The configural weights of the lowest, middle, and highest outcomes of $G+ = (\$12, .05; \$14, .05; \$96, .9)$ are then, respectively, $w_{\mathrm{L}} = .05 + (1/4)(.05) + (1/4)(.9) = .2875$; $w_{\mathrm{M}} = .05 - (1/4)(.05) + (1/4)(.9) = .2625$; and $w_{\mathrm{H}} = .9 - (1/4)(.9) - (1/4)(.9) = .45$. The average value of $G+$ is therefore \$50.32. Similarly, for $G- = (\$12, .1; \$90, .05; \$96, .85)$, $w_{\mathrm{L}} = .325$, $w_{\mathrm{M}} = .25$, and $w_{\mathrm{H}} = .425$, for an average of \$67.2, which exceeds \$50.32 for $G+$, violating dominance.

It is worth noting that this pattern of weighing was not fit to violations of stochastic dominance post hoc, but rather was estimated from violations of branch independence. Violations of branch independence are compatible with the RAM, TAX, and RDEU models. Thus, data that are compatible with all three models were used to make a new prediction that distinguishes the class of configural models (that violate stochastic dominance) from the class of models that satisfy this property.

The class of rank-dependent RDEU/RSDU/CPT models must satisfy stochastic dominance for any choices (Birnbaum & Navarrete, 1998, pp. 57–58; Luce, 1998, 2000). For example, with the CPT model and the parameters of Tversky and Kahneman (1992), the corresponding certainty equivalents of the gambles are \$70.26 for $G+$ against 65.17 for $G-$.

Equations for the CPT, RAM, and TAX models are presented in Birnbaum and Navarrete (1998, pp. 54–57). Calculations for the CPT, RAM, and TAX models can be made in URL http://psych.fullerton.edu/mbirnbaum/taxcalculator.htm, and http://psych.fullerton.edu/mbirnbaum/cwtcalculator.htm, which are described in Birnbaum et al. (1999). These on-line, Netscape-compatible JavaScript calculators can be used to compute certainty equivalents according to the CPT model and parameters fit to Tversky and Kahneman (1992) and to the RAM and TAX models and parameters of Birnbaum (1997, 1999a). The calculators allow the user to compute certainty equivalents of gambles with two to five nonnegative consequences. The user can also change parameter values to explore their effects on predictions.

Birnbaum and Navarrete (1998) tested four variations of this recipe for $G-$ versus $G+$ with 100 undergraduates and found that about 70% violated dominance, averaged over the four variations.

Birnbaum, Patton, and Lott (1999) tested a new sample of 110 students with five new variations of the same recipe and found an average of 73% violations. These studies also tested two properties derived by Birnbaum (1997), which he named *lower cumulative independence* and *upper cumulative independence*. These properties are also implied by RSDU/RDEU/CPT theories, and they were also violated systematically.[1]

[1] Birnbaum (1997, p. 96) derived the following conditions (where $0 < z < x' < x < y < y' < z'$ and $p + q + r = 1$):

Lower cumulative independence: If $S = (z, r; x, p; y, q)$ is preferred to $R = (z, r; x', p; y', q)$, then $S'' = (x', r; y, p + q)$ preferred to $R'' = (x', r + p; y', q)$.

Upper cumulative independence: If $S' = (x, p; y, q; z', r)$ is not preferred to $R' = (x', p; y', q; z', r)$, then $S''' = (x, p + q; y', r)$ is not preferred to $R''' = (x', p; y', q + r)$.

Any theory that satisfies comonotonic independence, monotonicity, transitivity, and coalescing must satisfy both lower and upper cumulative independence (Birnbaum &

Because violations of stochastic dominance contradict so many proposed descriptive theories, it is important to determine if the results are unique to the particular procedures used in previous research. Can the conclusions of these laboratory studies be generalized to predict the results with procedures and people other than those tested?

The experiments of Birnbaum and Navarrete (1998) and Birnbaum et al. (1999) required undergraduates to make more than 100 choices between gambles. Participants were not paid, so they had no financial incentive to choose wisely. In addition, people were asked not only to choose the gamble they preferred, but also to state the amount they would pay to receive their chosen gamble rather than the other gamble. If the results are unique to these procedures, such as the method of display of the gambles, the lack of financial incentives, or the instruction to judge strength of preference, then perhaps the RDEU class of models could be retained at least for certain types of experiments.

The purpose of this chapter is to review experiments that followed those of Birnbaum and Navarrete (1998) and Birnbaum et al. (1999) in order to examine more closely the conditions under which people violate stochastic dominance. Two of the studies were conducted via the World Wide Web in order to recruit participants who were demographically diverse in order to check the generality of the results to groups other than college students. Five studies featured here have not been previously published.

Changes in Procedures

The following changes in procedure were made: (1) Offer financial incentives; perhaps with financial incentives, people might conform to stochastic dominance. (2) Collect fewer choices per person; perhaps with many trials, people get bored, careless, or adopt simple strategies that have systematic errors. (3) Try other formats for displaying the gambles; if the violations of stochastic dominance are due to processes of choice (as opposed to evaluation of the gambles), changing the juxtaposition of the branches might affect the incidence of violations. (4) Put related

Navarrete, 1998, pp. 53–54), including the class of RDEU/RSDU/CPT models. Both cumulative independence properties were systematically violated, as predicted by the configural weight RAM (Birnbaum, 1999b) and TAX models (Birnbaum & Navarrete, 1998).

choices on the same page to allow judges to see the consistency of their choices more easily. (5) Remove instructions or feedback concerning violations of transparent dominance in the warm-ups used in previous research; perhaps this procedure somehow affects the strategies adopted. (6) Omit the procedure whereby judges were asked to evaluate the difference between the two gambles; perhaps the task of judging strength of preference alters the choice process.

In Experiments 1 and 2, all six of these variations of procedure were made, using two variations of the format of Kahneman and Tversky (1979) for presentation of each choice. In Experiment 3, we used the procedure of Birnbaum and Navarrete (1998), with extensions to include a greater variation of the recipe for stochastic dominance. Experiments 4 and 5 recruited participants via the World Wide Web. Such samples are demographically diverse and allow the investigator to check the generality of results across demographic groups. In Experiments 4 and 5, two other variations for presentation of the gambles were tried. In Experiment 4, either text or pie charts were used to display the probabilities (perhaps with pie charts, judges can "see" stochastic dominance more easily). In Experiment 5, the order of the consequences was reversed in order to see if this reversal would produce different results from those obtained with pie charts.

Test of Event Splitting/Coalescing

Coalescing is the assumption that if a gamble has two equal consequences, one can combine them by adding their probabilities without changing the utility of the gamble. For example, $GS = (\$12, .1; \$12, .1; \$96, .8)$ should be indifferent to $G = (\$12, .2; \$96, .8)$. Coalescing was assumed as an editing principle, *combination*, in the original prospect theory (Kahneman & Tversky, 1979). Coalescing is implied by RDEU/RSDU/CPT theories but not by configural weight theories (Birnbaum & Navarrete, 1998; Luce, 1998). Luce (1998) showed that coalescing also distinguishes other decision-making theories and that coalescing and rank-dependent additivity can be used to deduce rank-dependent expected utility theory. Birnbaum (1999a) hypothesized that violations of coalescing might account for violations of stochastic dominance, cumulative independence, and also upper-tail independence (studied by Wu, 1994).

Note that event splitting was used as one ingredient of the recipe creating violations of stochastic dominance. Our present studies tested if event splitting can be used to also eliminate violations of stochastic dominance within the same gambles. Although there is no asymmetry in the mathematics between coalescing and event splitting, intuitively the two ways to convert gambles are different. There is only one way to coalesce equal consequences in a gamble (converting from gamble *GS* to *G*), but there are many ways to split events to convert a gamble into equivalent gambles (converting *G* to *GS*).

Luce (1998, p. 91) noted that previous tests of event splitting (Humphrey, 1995; Starmer & Sugden, 1993) were not optimal. He remarked, "data from the coalesced and uncoalesced cases were from two nonoverlapping sets of subjects, so it is not possible to do a two-by-two cross tabulation. . . . Given . . . [that] there are substantial individual preference differences among people, I view this as a decidedly weak test of the property." The present studies all use designs that support strong tests.

To test coalescing more directly, we split consequences in $G+$ and $G-$ to create four-outcome gambles, $GS+$ and $GS-$. The split versions of these examples are $GS+ =$ ($12, .05; $14, .05; $96, .05; $96, .85) versus $GS- =$ ($12, .05; $12, .05; $90, .05; $96, .85). The choice, $GS+$ versus $GS-$ is really the same choice as $G+$ versus $G-$, except for coalescing. In Table 3.1, $G+$ and $G-$ are I and J in row 5, respectively, and $GS+$ and $GS-$ are *U* and *V* in row 11.

The configural weight RAM model of Birnbaum and McIntosh (1996), with parameters estimated in previous research, predicts that judges should violate stochastic dominance by preferring $G-$ to $G+$. The TAX model of Birnbaum and Chavez (1997) makes the same prediction. These configural weight models also predict that judges should show an event-splitting effect by preferring $GS+$ to $GS-$.

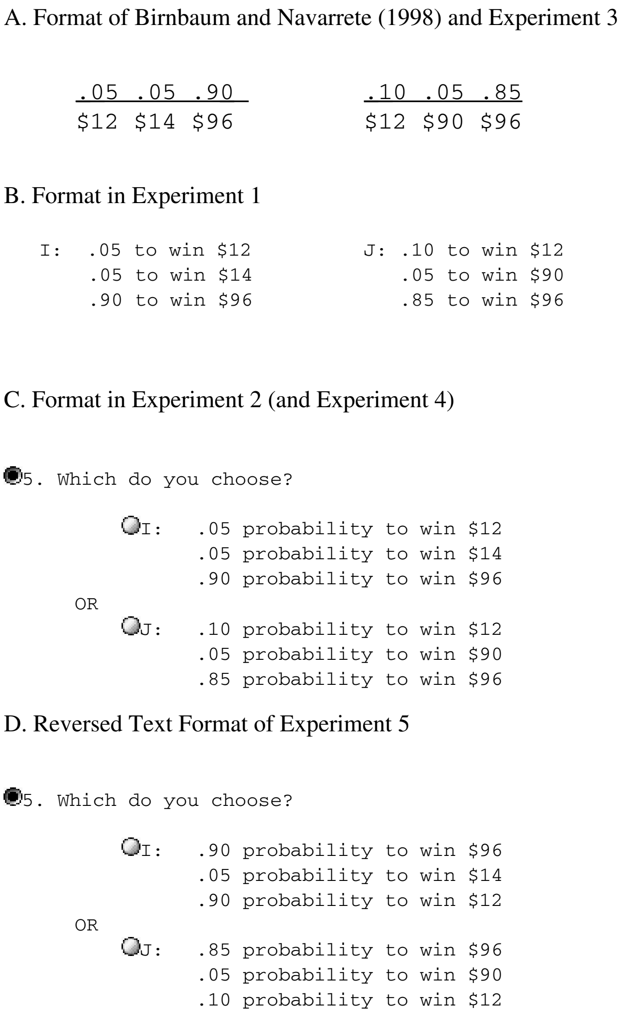
Methods

In Experiment 1, 31 undergraduates (10 male and 21 female) were told that they would have a chance to play one gamble for real money. They were told that 2 people (out of 31) would be randomly selected to play one gamble for either the face amount, half the face amount, or twice the face amount, which might yield cash prizes as high as $220. This instruction appeared to produce considerable excitement among the participants. One trial would be selected, and each selected judge would play the gamble that she or he chose on that trial during the study. Judges were reminded that the trial selected might be any one of the

Table 3.1. *Choices, Predictions, Modal Choices, and Percentage Choices (Experiments 1 and 2)*

			Predictions			Results % Choice	
Row	Choice		TAX	CPT	Mode	Exp 1	Exp 2
1	A: .50 to win $0 .50 to win $100	B: .50 to win $25 .50 to win $35	A	A	B	52	52
2	C: .50 to win $0 .50 to win $100	D: .50 to win $45 .50 to win $50	D	D	D	81*	61
3	E: .50 to win $4 .30 to win $96 .20 to win $100	F: .50 to win $4 .30 to win $12 .20 to win $100	E	E	E	100*	88*
4	G: .40 to win $2 .50 to win $12 .10 to win $108	H: .40 to win $2 .50 to win $96 .10 to win $108	H	H	H	97*	100*
5	I: .05 to win $12 .05 to win $14 .90 to win $96	J: .10 to win $12 .05 to win $90 .85 to win $96	J	I	J	74*	82*
6	K: .80 to win $2 .10 to win $40 .10 to win $44	L: .80 to win $2 .10 to win $10 .10 to win $98	K	L	L	65	58
7	M: .06 to win $6 .03 to win $96 .91 to win $99	N: .03 to win $6 .03 to win $8 .94 to win $99	M	N	M	77*	55
8	O: .80 to win $10 .20 to win $44	P: .90 to win $10 .10 to win $98	P	P	P	77*	70*
9	Q. .20 to win $40 .80 to win $98	R: .10 to win $10 .90 to win $98	Q	Q	Q	65	46
10	S: .10 to win $40 .10 to win $44 .80 to win $110	T: .10 to win $10 .10 to win $98 .80 to win $110	T	S	T	74*	79*
11	U: .05 to win $12 .05 to win $14 .05 to win $96 .85 to win $96	V: .05 to win $12 .05 to win $12 .05 to win $90 .85 to win $96	U	U	U	94*	79*
12	W: .05 to win $10 .05 to win $98 .90 to win $106	X: .05 to win $44 .05 to win $49 .90 to win $106	W	X	W	68	
13	Y: .03 to win $6 .03 to win $6 .03 to win $96 .91 to win $99	Z: .03 to win $6 .03 to win $8 .03 to win $99 .91 to win $99	Z	Z	Z	90*	97*
14	A′: .05 to win $10 .95 to win $98	B′: .10 to win $44 .90 to win $98	B′	B′	B′	71*	

* Denotes a percentage significantly different from 50%. Choice percentages indicate agreement with the modal choice in Experiment 1, and are therefore all above 50%, except in row 9, where only 46% in Experiment 2 chose Q.



A. Format of Birnbaum and Navarrete (1998) and Experiment 3

```
 .05 .05 .90             .10 .05 .85
 $12 $14 $96             $12 $90 $96
```

B. Format in Experiment 1

```
I:  .05 to win $12        J: .10 to win $12
    .05 to win $14           .05 to win $90
    .90 to win $96           .85 to win $96
```

C. Format in Experiment 2 (and Experiment 4)

```
●5. Which do you choose?

         ○I:   .05 probability to win $12
               .05 probability to win $14
               .90 probability to win $96
     OR
         ○J:   .10 probability to win $12
               .05 probability to win $90
               .85 probability to win $96
```

D. Reversed Text Format of Experiment 5

```
●5. Which do you choose?

         ○I:   .90 probability to win $96
               .05 probability to win $14
               .90 probability to win $12
     OR
         ○J:   .85 probability to win $96
               .05 probability to win $90
               .10 probability to win $12
```

Figure 3.1 Four formats for presentation of a choice between gambles.

choices, so they should make each choice carefully. Judges circled their preferred gamble in each choice. [Immediately after the choices were completed, games were played publicly as promised, and two students (who seemed quite animated) won cash by drawing slips randomly from urns.]

The 14 choices in Table 3.1 were printed on a single side of a page, using the format in Figure 3.1B. Figure 3.1A shows the format of Birnbaum and Navarrete (1998) and of Birnbaum et al. (1999), also used in Experiment 3, for comparison.

Two pairs of choices tested stochastic dominance and coalescing. Rows 5 and 11 included $G+ =$ (\$12, .05; \$14, .05; \$96, .90) $= I$ versus $G- =$ (\$12, .10; \$90, .05; \$96, .85) $= J$, and $GS+ =$ (\$12, .05; \$14, .05; \$96, .05; \$96, .85) $= U$ versus $GS- =$ (\$12, .05; \$12, .05; \$90, .05; \$96, .85) $= V$. $GS+$ versus $GS-$ is really the same choice as $G+$ versus $G-$, except for coalescing. Rows 7 and 13 created another variation of the same recipe, with the dominant gamble counterbalanced in the left–right position.

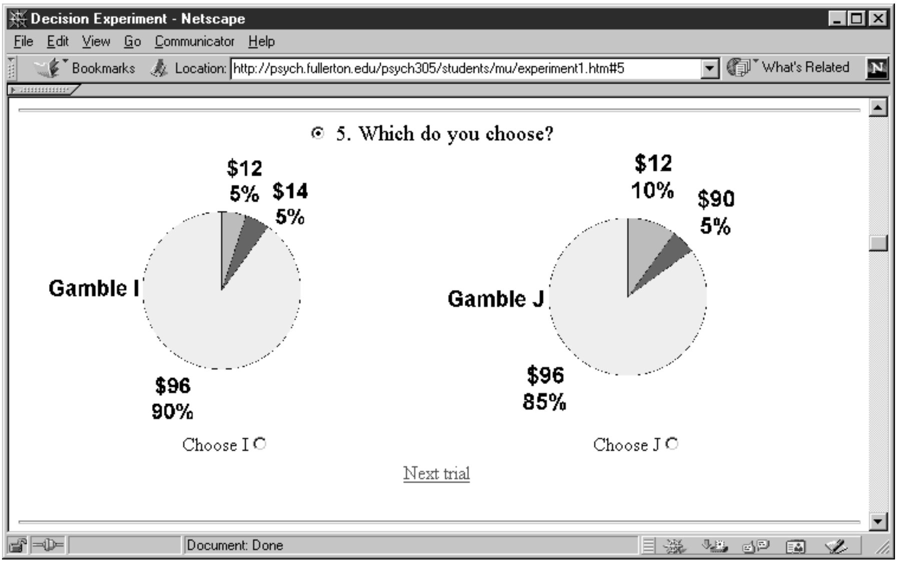
In Experiment 2, undergraduates (8 male and 24 female) viewed the gambles presented on a computer screen instead of on paper. Twelve of the choices in Table 3.1 were included (all except rows 12 and 14). The gamble printed on the left in the paper version was placed above the other gamble in the computer version, as illustrated in Figure 3.1C. Formats in Figures 3.1B and 3.1C differ in the juxtaposition of the branches.

In Experiment 2, choices were displayed on a computer screen by a browser, displaying a HyperText Markup Language form, which collected the data. Judges used the mouse to click a radio button next to the gamble they preferred. They were told that two participants would be selected randomly to play one of their chosen gambles for the face value. Experiment 2 was a laboratory pilot test of procedures later used by Birnbaum (1999b, 2000) and in Experiments 4 and 5.

Experiment 3 used 100 undergraduates, tested with procedures of Birnbaum and Navarrete (1998), in which choices were displayed as in Figure 3.1A.

Two choices for each set compared $G+$ versus $G-$ and $GS+$ versus $GS-$, where $GS+ = (x, s; x+, r; y, q; y, p)$, $GS- = (x, s; x, r; y^-, q; y, p)$, $G+ = (x, s; x+, r; y, p+q)$, and $G- = (x, s+r; y^-, q; y, p)$, where $s = 1 - p - q - r$. There were five sets with $(x, x+, y^-, y)$ = (\$4, \$10, \$92, \$98), (\$6, \$12, \$93, \$99), (\$6, \$9, \$91, \$97), (\$7, \$10, \$84, \$90), and (\$3, \$9, \$91, \$97), for rows 1 to 5 of Table 3.3, respectively. The values of (r, q, p) are given in Table 3.3. Note that choices in the first two rows resemble those used in previous research, but choices in the last three rows use smaller values of p and larger values of q. These 10 choices were included among 103 others (Martin, 1998). Instructions made it clear that in Experiment 3 the "financial incentives" were strictly hypothetical (as in Birnbaum and Navarrete, 1998).

Experiments 4 and 5 were conducted via the Web, using the general instructions and the 20 choices of Birnbaum (1999b). Experiment 4 compared choices displayed as in Figure 3.1C against those displayed by means of pie charts, as in Figure 3.2. Experiment 5 tested a reversal of the order of consequences, as shown in Figure 3.1D, against pie charts.



Decision Experiment - Netscape
File Edit View Go Communicator Help
Bookmarks Location: http://psych.fullerton.edu/psych305/students/mu/experiment1.htm#5
What's Related
$12
5%
$14
5%
Gamble I
$96
90%
$12
10%
$90
5%
Gamble J
$96
85%
Next trial
Document: Done

5. Which do you choose?

Choose I ○ Choose J ○

Figure 3.2 Use of pie charts to display gambles in Experiments 4 and 5.

The 999 participants (304 male and 689 female) were recruited via search engines, links in Web sites listing contests and games with prizes, and links in sites listing psychology experiments. This recruitment procedure is similar to that used in the Internet B study of Birnbaum (2000). Participants were assigned to conditions by clicking their birth months, which linked them to different variations of display procedures. Every third birth month was assigned to text and other months to the pie chart display. During the run of the study, the association of birth months with conditions was counterbalanced by Latin Square.

In text display formats, probability was described in terms of the number of equally likely tickets out of 100 in an urn, from which 1 ticket would be drawn to determine the prize. In the pie chart display format, the probability mechanism was described as a spinner device whose pointer was equally likely to stop in any equal-sized slice of the pie. The percentage of each slice was also displayed, as illustrated in Figure 3.2. Web participants were informed that three participants (in each study) would be selected to play one of their chosen gambles.

Results

Choices 1 and 2 of Table 3.1 form an indirect test of consequence monotonicity: $A = (\$0, .5; \$100, .5)$ versus $B = (\$25, .5; \$35, .5)$ and $C = A$ versus $D = (\$45, .5; \$55, .5)$. If a subject prefers B to A, then the same judge should prefer D to A because D strictly dominates B. There were no violations in Experiments 1 or 2. Rows 3 and 4 were transparent tests of consequence monotonicity in three-outcome gambles. They are termed *transparent* because everything is the same except the value of one consequence. There was 1 violation out of 62 choices in Experiment 1 and 4 out of 66 in Experiment 2. The overall rate of 3.9% violations of consequence monotonicity is similar to the figure of 4% reported by Birnbaum and Navarrete (1998) for similar tests.

The last three columns of Table 3.1 show modal choice and the percentages of judges who made the modal choice in Experiments 1 and 2, respectively. Modal choices are the same in both studies, except in choice 9; none of the choice percentages differ significantly between Experiments 1 and 2. Asterisks indicate choice percentages that deviate significantly from 50% by a two-tailed binomial sign test with $\alpha = .05$, tested separately in each experiment. Predicted choices in Table 3.1 were calculated for CPT and TAX models using previously published parameters.

We can ask how well these models predict the results. The prior CPT and TAX models make different predictions for 5 of the 14 choices (Choices 5, 6, 7, 10, and 12). In four of these five choices (except row 6), the majorities were consistent with the TAX model's predictions, using previously estimated parameters. Each judge was scored for these differential predictions, and it was found that 43* of the judges in Experiments 1 and 2 had more choices consistent with the prior TAX model than with the prior CPT model, compared to only 12 whose choices had a majority consistent with the CPT model. (Throughout this chapter, asterisks indicate statistical significance, tested against the binomial null hypothesis, with $p = 1/2$ and $\alpha = .05$. In this case, 43* indicates that significantly more than half of the 43 + 12 subjects had more choices correctly predicted by TAX.)

Although one might improve fits by estimating parameters from the data, violations of stochastic dominance in rows 5 and 7 refute CPT (RDEU/RSDU) with any parameters.

Cross-tabulations for the two tests of stochastic dominance and coalescing are shown in Table 3.2 for Experiments 1 and 2. In all four tests, the majority of participants violated stochastic dominance in the coalesced form *and* satisfied stochastic dominance in the split form.

In Experiment 1, 23* and 24* judges violated stochastic dominance in the choices of rows 5 and 7 of Table 3.1, respectively. In Experiment 1, 20 judges violated stochastic dominance on both of these choices, 7 violated it once, and 4 had no violations (76% violations overall).

Table 3.2. *Tests of Stochastic Dominance and Coalescing*[a]

	Rows 7 and 13		Rows 5 and 11	
Split	$G+ = N$	$G- = M$	$G+ = I$	$G- = J$
Experiment 1				
$GS+$	6	22*	8	21*
$GS-$	1	2	0	2
TOTAL	7	24*	8	23*
	Rows 7 and 13		Rows 5 and 11	
Split	$G+ = N$	$G- = M$	$G+ = I$	$G- = J$
Experiment 2				
$GS+$	14	18*	5	21*
$GS-$	1	0	1	6
TOTAL	15	18	6	27*

[a] Each entry shows the number of judges who had each conjunction of preferences in each pair of choice problems.

Note: Asterisks indicate frequencies significantly greater than 50% of participants.

In Experiment 2, 27 and 18 judges violated stochastic dominance in rows 5 and 7, respectively; there were 17 with two violations, 11 with one, and only 5 with no violations (68.2% violations overall). Averaged over studies and rows, the average rate of violation is 71.9%. These rates of violation are significantly greater than 50% but not significantly different from the rate of 70% violations reported by Birnbaum and Navarrete (1998).

In Experiment 1, there were 28* and 29* judges who satisfied stochastic dominance in the split form (rows 11 and 13 of Table 3.1) of the same two choices, which represents 92% satisfactions overall. According to any theory that satisfies coalescing and transitivity, there should have been no changes in preference due to event splitting, except by chance. Instead, the off-diagonal reversals of 21 to 0 and 22 to 1 each has a probability of less than 3 in 1 million, given the null hypothesis that either switch of preference is equally likely. In addition, the majority (19, or 61%) violated stochastic dominance on *both* comparisons of three-outcome gambles *and* satisfied it on *both* choices between four-outcome split variations of the same gambles. Results for Experiment 2 are similar, with 21–1 and 18–1 counts for preference reversals produced by event splitting and 88% satisfaction of stochastic dominance when split.

Table 3.3. *Tests of Event Splitting and Stochastic Dominance in Experiment 3*

Gambles				Choice Patterns			
r	q	p		$-+$	$--$	$+-$	z
.01	.02	.96	*a*	74*	3	3	8.09
.03	.03	.92		61*	3	1	7.62
.03	.74	.14	*a*	49	2	3	6.38
.30	.30	.30	*a*	43	1	1	6.33
.04	.59	.36		32*	3	6	4.22

Notes: + indicates satisfaction of dominance, i.e., $G+$ preferred to $G-$ and $GS+$ preferred to $GS-$. The most common pattern, $-+$, represents a violation in the coalesced form and a satisfaction in the split form. In rows marked *a*, dominant gambles were presented on the right. Values of z test event splitting (all are significant). Asterisks indicate frequencies significantly different from 50 in $-+$ column only (n = 100).

Results for Experiment 3 are shown in Table 3.3. In all five variations, event-splitting effects are significant, because the $-+$ pattern ($G-$ preferred to $G+$ and $GS+$ preferred to $GS-$) occurs significantly more often than the opposite pattern of preferences ($+-$) in each row (z values are all significant). Summing over rows for each person, 91* of the 100 judges had more preference reversals of $-+$ than the opposite switch, 2 had more of the opposite, and 7 were tied.

The choices in the lower rows of Table 3.3 were designed on the basis of intuition to produce fewer violations (more satisfactions) of stochastic dominance. Table 3.3 shows that the frequency of violations of stochastic dominance (the sum of $-+$ and $--$) varies with the pattern of probabilities (p, q, r) used to construct the choices. The variations with large p and small q and r, like those used in previous research, give more violations of stochastic dominance (70.5% for the first two rows) than those with smaller values of p and larger q (43% for the last three rows). The frequencies of $-+$ in the first two rows are each significantly greater than 50%; however, the frequency in the last row is significantly less than 50%. (These row differences are also significant by tests of correlated proportions.) The prior TAX model does not predict this reversal of preference between rows in Table 3.3. It is not yet clear if this effect reveals a structural flaw in the TAX model or merely represents an error in specification of its functions or parameters.

Internet Research: Lab and Web

The Internet provides a new way to conduct judgment and decision-making research. It has several advantages, the most important being the reduced cost and effort required to test large numbers of participants. Because the study can run day or night without a laboratory assistant, and because the data are directly entered and organized by the software that controls the study, one can do in weeks what used to take half a year to complete (Birnbaum, 2000, 2001).

There are two obvious causes for concern with Web studies. The first is that there is less control in a Web study compared to that in the laboratory. For example, one can make sure that laboratory participants do not use calculators, but one could only instruct Web participants and ask them if they followed instructions; one cannot control conditions via the Web in the same way as in the laboratory. The second concern is that Web participants may have quite different demographic characteristics from those recruited from the usual college subject pool. Although heterogeneity is a concern for studies with small samples, the demographic variations in large samples of Web studies presents an opportunity to check the generality of laboratory results to a wider variety of people than are usually studied.

Birnbaum (1999b) wanted to recruit people highly expert in decision making. Members of the Society for Mathematical Psychology and the Society for Judgment and Decision Making were recruited to complete an experiment on-line. The recruitment method appears to have been successful, because there were 95 "experts," participants with doctorates who also indicated that they had read a scientific article or book on the theory of decision making. These 95 were among the 1,224 people from 44 nations who participated.

In order to compare the Web with the laboratory studies, a sample of 124 students from the usual subject pool were also tested in the laboratory using the same materials. It was found that the rate of violation of stochastic dominance varied significantly with demographic characteristics of gender, education, and experience reading an article on decision making. However, even among the 95 "experts" in the study, 46 violated stochastic dominance on the choice in row 5 of Table 3.1 *and* satisfied it in the split form, compared to only 7 with the opposite reversal of preference.

Following the recruitment of the highly educated sample, a "B" sample was recruited by methods designed to reach the general public

Table 3.4. *Violations of Stochastic Dominance (S. Dom.), Monotonicity (Mono.), and Coalescing (Experiments 4–5)*

Group	Sample *n*	S. Dom. (%)	Mono (%)
Exp 4 and 5	999	63.5	9.8
Text	172	59.6	9.9
Text *Rev*	169	62.4	10.7
Pies	353	65.2	8.5
Pies*	305	64.4	8.2
Internet B	737	58.8	7.9

Notes: Experiment 4 compared Text with pies; Experiment 5 compared reversed text (Text *Rev*) with pies*. Internet B refers to the sample reported by Birnbaum (2000) recruited by the same methods as in Experiments 4–5. The procedures of Internet B matched those of the Text condition of Experiment 4.

Table 3.5. *Violations of Stochastic Dominance and Monotonicity in Experiments 4–5, by Demographic Characteristics*

Group	Sample *n*	S. Dom. (%)	Mono (%)
Females	689	65.8	9.4
Males	304	58.4	10.9
Read *DM*	219	56.2	9.8
Canada	41	58.5	7.3
U.K.	57	64.0	12.2
N. Europe	47	44.7	12.8

Note: N. Europe = Belgium, Switzerland, Estonia, Finland, Germany, Hungary, the Netherlands, Norway, and Sweden. Read *DM* indicates participants who claim to have read a scientific work on decision making.

(Birnbaum, 2000). These participants were recruited by links in sites advertising "free" games and contests with prizes.

The findings on the Web and in the laboratory yield much the same results: Violations of stochastic dominance can be markedly reduced by event splitting. In the Internet B sample ($n = 737$), it was found that 59% violated stochastic dominance in the coalesced form compared to only 8% in the split form.

Experiments 4 and 5 used samples recruited via the Web to assess two other formats for presentation of the gambles. Results of these studies are shown in Tables 3.4, 3.5, 3.6. Table 3.4 shows that gambles presented

Table 3.6. *Violations by Educational Level in Experiments 4–5*

Group	Sample *n*	S. Dom. (%)	Mono (%)
<12	49	61.2	12.2
12	214	68.2	11.2
13–15	364	63.6	8.1
16	244	62.7	11.5
17–19	73	60.3	12.3
20	55	54.5	1.8

Notes: Group education levels by year-equivalents of education; <12 = non–high school graduate; 12 = high school graduate, 13–15 = some college; 16 = college degree; 17–19 = graduate studies; 20 = doctorate. Violations of stochastic dominance (S. Dom.) and monotonicity (Mono) are given in percentages, averaged over two variations of $G+$ versus $G-$ and $GS+$ versus $GS-$.

as pie charts or with reversed order of branches yielded results quite comparable with those obtained previously. The text condition ($n = 172$) corresponds to that used in the Internet B in stimulus format, method of recruiting the sample, and rates of violation in coalesced and split forms. The reversed text (Figure 3.1D) and pie chart (Figure 3.2) formats of presentation yield very similar results that show slightly higher rates of violation of stochastic dominance in the coalesced form. Overall rates of violation are lower in the Internet than in the laboratory studies, but most Web participants are also better educated than the subject pool laboratory samples.

Tables 3.5 and 3.6 illustrate how one can partition the data from a Web study to analyze the results within different demographic groups. Consistent with previous results (Birnbaum, 1999b, 2000), we found that education, male gender, and having read a scientific work on decision making were correlated with lower rates of violation of stochastic dominance. Overall, the rate of violation in the coalesced form is 63.5%, but among those 219 who report having read a work on decision making, it is 56.2%. The rate among females is 65.8% and among males it is 58.4%, a gender difference observed in previous studies (Birnbaum, 1999b, 2000). The rate is lower among Northern and Central European participants compared to Americans, but this group is also more highly educated than the average U.S. participant. Table 3.6 shows that those with doctorates are less likely to violate stochastic dominance than those having only high school degrees.

Although there are differences between methods of presentation and among groups in Tables 3.4 to 3.6, the overall findings are that rates of violation are still quite high in all conditions and demographic groups, and that rates are much higher in the coalesced form than in the split form. Thus, the conclusions of the research regarding decision-making models are very much the same in the laboratory and on the Web.

The present studies found significant violations of stochastic dominance and coalescing despite changes in procedure that were hypothesized to possibly reduce or eliminate the effects. From a lack of difference between procedures, one cannot conclude that all variations of procedure have no effect. However, one can conclude from Experiments 1 and 2 that violations of stochastic dominance reported by Birnbaum and Navarrete (1998) are robust enough that they can be replicated with statistically significant results in small samples, even with six changes in procedure hypothesized to reduce or eliminate the effects. One can reject the hypothesis that these changes reduced violations of stochastic dominance to the minority. The rates of stochastic dominance violation are similar to those reported previously, when the participants and choices resemble those used in previous research.

Financial Incentives

The effects of financial incentives have been studied in a number of papers (see the review in Camerer & Hogarth, 1999). However, there does not yet seem to be a body of evidence showing that preferences among gambles with positive consequences are systematically affected by the difference between real and hypothetical financial incentives (Camerer, 1989; Camerer & Weber, 1992; Mellers, Weiss, & Birnbaum, 1992; Tversky & Kahneman, 1992). Experiment 3, like previous studies with strictly hypothetical incentives, found rates of violation similar to those of Experiments 1 and 2, which had real incentives. Experiments 4 and 5 also used real incentives, yet rates of violation significantly exceeded 50% in all experiments.

Perhaps the stakes were too low in these studies to produce majority conformance to stochastic dominance. Although some theorize that behavior would change if the stakes were high enough, it would be useful to have a theory that specifies exactly how and why behavior depends on the magnitude of the stakes. If a person systematically violates dominance on many small, repeated decisions, the global effect could be quite large, so it is hard to see why systematic violations in small decisions are compatible with global rationality. Furthermore, it is not clear that people who gamble with real and large stakes (e.g., those who play blackjack in Las Vegas casinos) are behaving any more rationally than those who make judgments of their likely behavior in hypothetical tasks.

Although it is possible that the modest prizes (actual prizes won varied from $50 to $120) caused people to violate stochastic dominance intentionally, the behavior of the students in Experiments 1 and 2 gave every appearance of enthusiastic interest. Because the same people who violate stochastic dominance in rows 5 and 7 largely satisfy consequence monotonicity in rows 3, 4, 11, and 13, we think it is more likely that violations of stochastic dominance are due to a lack of understanding rather than to a lack of motivation.

Framing of Choices Is Not Necessary

Tversky and Kahneman (1986) reported a violation of stochastic dominance that was produced by "masking" the dominance relation with framing. The framing was accomplished by making it seem that the dominated gamble always gave the same prize or a higher one for every possible "event" (color of a marble drawn from an urn). Because the numbers of marbles of a given color differed in the two gambles, the events were not really the same. Their manipulation produced 58% violations of stochastic dominance, which was not significantly greater than 50% in their study but was quite different from the results in another framing of the same choice, where the same color marble always gave the same or a higher prize for the dominant gamble. Results such as those of the present studies suggest that the event framing used by Tversky and Kahneman is not required to produce large violations.

Birnbaum, Yeary, Luce, and Zhou (submitted) asked participants to judge the buying and selling prices of the same gambles presented for choice by Birnbaum and Navarrete (1998) ($G+$ and $G-$ were judged on separate trials, separated by many others). They found that judgments showed the same violations of stochastic dominance as reviewed here for choices. Thus, choice is not required to produce the violations. Explanations that hinge on comparisons, contrasts, frames, or regrets between the components of gambles do not appear to explain why violations are observed in judgment as well as choice. Instead, it seems more plausible to attribute the effect to the evaluation of the gambles rather than to choice processes.

Event-Splitting Effects and Distribution Effects

All five experiments reported here find evidence of powerful event-splitting effects. Most judges (92%, 88%, 95%, and 90% in Experiments 1, 2, 3, and 4–5, respectively) satisfied stochastic dominance in the four-outcome split versions gambles *GS*+ and *GS*−. However, most judges (76%, 68%, 70%, and 64% in Experiments 1, 2, 3, and 4–5, respectively) violated stochastic dominance with choices between three-outcome gambles that resemble the original recipe for *G*+ and *G*−. The RAM and TAX models with prior parameters correctly predict this reversal between the coalesced and split forms of the same choices.

Experiment 3 (Table 3.3) also shows that rates of violation of stochastic dominance and coalescing depend on how the probability is split in the recipe. This finding indicates that the results cannot be explained by the idea that judges ignore probabilities and choose on the basis of consequences alone (the consequences are nearly identical in all rows of Table 3.3). The majority reversal between the first and last rows is not predicted by the prior parameters in the TAX model. This phenomenon deserves further theoretical and empirical investigation.

Martin (1998) tested for event-splitting effects in another design in which choices between two-outcome gambles, *R* and *S*, which were presented with either the higher or lower consequences of either *R* or *S* split or coalesced. This design was intended to test event-splitting independence, the assumption that event splitting should have the same directional effect whenever the same probability paired with a positive consequence is split (Birnbaum & Navarrete, 1998, p. 71). She found that splitting the higher branch produced a significant improvement of either gamble, but splitting the lower branch had nonsignificant effects. Independence would have required that splitting the lower (but positively valued) branch should also have improved the gamble, whereas configural weight averaging models with prior parameters predict that splitting the lower-valued branch makes the gamble worse. Thus, Martin's (1998) data with that design added evidence of event-splitting effects (for higher-valued consequences), but they neither clearly refuted nor conformed to event-splitting independence.

It is interesting to consider that event splitting can be used both to induce violations of stochastic dominance and to reduce them. Event splitting was used to produce $G-$ and $G+$ from G_0, which created the violations of stochastic dominance; and splitting again created $GS-$ and $GS+$, which reversed preferences by producing satisfaction of stochastic dominance. These results strengthen the case made by Birnbaum and Navarrete (1998) and Birnbaum et al. (1999) that coalescing is the property whose failure explains violations of RDEU/RSDU/CPT models.

Three Types of Generalization: Participants, Procedures, and Predictions

This program of research illustrates important directions for generalizing psychological research: generalization across participants, generalization across procedures, and generalization across novel predictions.

Generalization across groups of people can be facilitated greatly by recruiting participants from the Internet. By recruiting large and heterogeneous samples of participants, Internet research allows one to check if the results found with undergraduates generalize to those obtained in other groups. The ability to obtain large samples of high-quality data quickly, conveniently, and at low cost via the Web will likely accelerate the pace of empirical research. In the Internet studies reviewed here, violations of stochastic dominance were observed in the majority of all samples except those highly educated in decision making, among whom violations were still substantial. Although rates of violation depend on gender, education, and experience reading an article on decision making, the conclusions regarding theory would be essentially the same for all sub-samples tested.

Similarly, these results show that results can be generalized across a variety of different procedures. Although rates of violation appear slightly different with different methods for displaying the gambles, the same conclusions are reached with or without financial incentives and with five different ways to display the choices. The findings do not appear to hinge on certain other particulars of procedure used in the early laboratory studies of Birnbaum and Navarrete (1998) and Birnbaum et al. (1999), such as the stimulus display or the instruction used in those studies to judge the strength of preference.

Event splitting or coalescing might also be considered a variable of procedure because both represent different ways of displaying the same (objective) choices. This variable makes a very large, significant difference in every group studied and within every variation of display procedure tested so far. Event-splitting effects are not compatible with the RDEU/RSDU/CPT class of models, but the configural weight RAM and TAX models predict these effects. Clearly, this variable needs to be represented by theory.

The third type of generalization, extrapolation to new predictions, is probably the most important for the study of theory. This approach can hardly be considered a new direction in our field, because the derivation and empirical testing of new implications that distinguish classes of theories has long been recognized as a classic technique in empirical science. However, this concept may not have had as much success in psychology as in other fields of science because of the many variables of context and individual differences that affect small laboratory experiments.

Perhaps as a consequence of seeing results vary from one laboratory to another, there has been a tendency to look at behavior as the result of conflicting principles that are so numerous and complex that no theory will ever account for all data. Tversky and Kahneman (1992), for example, gave a "pessimistic assessment" that neither their CPT model nor any model will suffice to account for all choices between gambles. I do not share their view, as I think that as good theory is developed, what seemed exceptions and complexities fall out as implications of the new theory. What is an anomaly or a paradox to one theory is the prediction of another.

I think that many failures to replicate are due to imprecise descriptions of experimental paradigms in scientific publications. The use of Internet experiments has the advantage that all of the details of procedure are available to other scientists, facilitating clean replication and variation of procedure.

This chapter presents a case in which this classical approach of testing differential predictions to new situations leads to a clean result. One class of theories (including RDEU/RSDU/CPT models) cannot account for violations of stochastic dominance and coalescing, and another class (including TAX and RAM models) predicted them in advance of the experimental tests. In addition to these successful predictions, configural weight models correctly predicted violations of lower and upper cumulative independence, and they account for traditional Allais paradoxes with the same set of parameters (Birnbaum, 1999a). When a theory accounts for old data and continues to make successful new predictions, one begins by induction to believe that the next new prediction from the theory will hold. At the same time, induction applied to the history of science shows that all theories eventually fall as new implications are derived and tested. But as each new result becomes well established, it restricts possible theoretical representations and brings us closer to scientific understanding.

References

Becker, J., & Sarin, R. (1987). Lottery dependent utility. *Management Science*, *33*, 1367–1382.

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. (1999a). Paradoxes of Allais, stochastic dominance, and decision weights. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards* (pp. 27–52). Norwell, MA: Kluwer.

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

Birnbaum, M. H. (2000). Decision making in the lab and on the Web. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 3–34). San Diego, CA: Academic Press.

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Upper Saddle River, NJ: Prentice Hall.

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*(2), 161–194.

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes*, *67*, 91–110.

Birnbaum, M. H., Luce, R. D., Yeary, S., & Zhou, L. (submitted). Contingent Valuation, Endowment, or Viewpoint Effects: Testing properties in judgments of buying and selling prices of lotteries. *Manuscript*.

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.

Birnbaum, M. H., Patton, J. N., & Lott, M. K. (1999). Evidence against rank-dependent utility theories: Violations of cumulative independence, interval independence, stochastic dominance, and transitivity. *Organizational Behavior and Human Decision Processes*, *77*, 44–83.

Camerer, C. F. (1989). An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty*, *2*, 61–104.

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Eds.), *Utility theories: Measurements and applications* (pp. 207–251). Boston: Kluwer.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production theory. *Journal of Risk and Uncertainty*, *19*, 7–42.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*, 325–370.

Humphrey, S. J. (1995). Regret aversion or event-splitting effects? More evidence under risk and uncertainty. *Journal of Risk and Uncertainty*, *11*, 263–274.

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

Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, *43*, 286–313.

Luce, R. D. (1998). Coalescing, event commutativity, and theories of utility. *Journal of Risk and Uncertainty*, *16*, 87–113.

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

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.

Machina, M. J. (1982). Expected utility analysis without the independence axiom. *Econometrica*, *50*, 277–323.

Martin, T. (1998) *Comparing rank dependent, subjective weight, and configural weight utility models: Transitivity, monotonicity, coalescing, stochastic dominance, and event splitting independence*. Master's thesis, California State University, Fullerton.

Mellers, B. A., Weiss, R., & Birnbaum, M. H. (1992). Violations of dominance in pricing judgments. *Journal of Risk and Uncertainty*, *5*, 73–90.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3, 324–345.

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

Starmer, C., & Sugden, R. (1993). Testing for juxtaposition and event-splitting effects. *Journal of Risk and Uncertainty*, *6*, 235–254.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, *59*, S251–S278.

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

Wakker, P., & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, *7*, 147–176.

Wu, G. (1994). An empirical test of ordinal independence. *Journal of Risk and Uncertainty*, *9*, 39–60.