Journal of Experimental Psychology:
Learning, Memory, and Cognition
1997, Vol. 23, No. 2, 406-426

0278-7393/97/$3.00

Tests of Consequence Monotonicity in Decision Making Under Uncertainty

Detlof von Winterfeldt and Ngar-Kok Chung
University of Southern California

R. Duncan Luce and Younghee Cho
University of California, Irvine

Consequence monotonicity means that if 2 gambles differ only in 1 consequence, the one having the better consequence is preferred. This property has been sustained in direct comparisons but apparently fails for some gamble pairs when they are ordered in terms of judged monetary certainty equivalents. In Experiments 1 and 3 a judgment procedure was compared with 2 variants of a choice procedure. Slightly fewer nonmonotonicities were found in one of the choice procedures, and, overall, fewer violations occurred than in previous studies. Experiments 2 and 3 showed that the difference was not due to procedural or stimulus presentation differences. Experiment 4 tested a noise model that suggested that the observed violations were due primarily to noise in estimating certainty equivalents, and so, despite some proportion of observed violations, consequence monotonicity cannot be rejected in that case.

A fundamental principle of almost all formal theories of decision making under uncertainty or risk (i.e., prescribed probabilities) is that when one alternative dominates another, then the dominated one can be eliminated from consideration in making choices. Dominance can take several distinct forms, and we focus on the one—called *consequence monotonicity*—that is both the most crucial for theories of choice as well as the most controversial empirically. It says that if a gamble is altered by replacing one consequence by a more preferred consequence, then the modified gamble clearly dominates the original one and so it is also chosen over the original one. Certainly this assumption seems compelling. Figure 1 illustrates a specific example. Indeed, most people, when faced with their violations of consequence monotonicity, treat them as misjudgments to be corrected (Birnbaum, 1992; Brothers, 1990; Kahneman & Tversky, 1979). So what is the problem?

The validity of consequence monotonicity was first brought into question by what is now called the Allais paradox (Allais, 1953; Allais & Hagen, 1979) and later by the closely related common ratio effect (Kahneman & Tversky, 1979). The Allais paradox is shown in Figure 2. It has been pointed out (Kahneman & Tversky, 1979; Luce, 1992) that the significance of these paradoxes is ambiguous because, in addition to monotonicity, they embody an assumption that a compound gamble is seen as indifferent to its formally equivalent reduced form. It is unclear which assumption has failed. When direct choices are made without any reduction, the number of violations is markedly reduced (Brothers, 1990). This result implies that a failure in reducing compound gambles to first-order ones, rather than a failure of consequence monotonicity, is the underlying source of the paradox. More recently, however, new evidence of a different character has cast doubt directly on consequence monotonicity. Rather than asking for a choice between pairs of gambles, these newer studies are based on the simple observation that it is much simpler to find for each gamble the sum of money that is considered indifferent to the gamble—such a sum is called a *certainty equivalent* of the gamble—and then to compare these numerical amounts. The simplest estimation procedure is to obtain a judgment of the dollar estimate, which is called a *judged certainty equivalent*. For n gambles this entails n judgments rather than the $n(n-1)/2$ binary comparisons, which is a major savings for large n. However, for certain classes of gambles, these judged certainty equivalents appear to violate consequence monotonicity (Birnbaum, Coffey, Mellers, & Weiss, 1992; Mellers, Weiss, & Birnbaum, 1992). We discuss these experimental findings in greater detail further on. The question is whether these results should be interpreted as violations of consequence monotonicity or whether there is something of a procedural problem in evaluating preferences.

Assuming that decision makers have some sense of preference among alternatives, what types of questions elicit that attribute? There really is no a priori way to decide this. The major part of the theoretical literature on decision making, which is mostly written by statisticians and economists with a smattering of contributions from psychologists, has cast matters in terms of choice: If g and h are two alternatives, decision makers are asked to choose between them, and when g is chosen over h we infer that the decision maker prefers g to h. These theories are cast in terms of a binary relation of "preferred or indifferent to." Of course, each of us continually chooses when purchasing goods. An

Detlof von Winterfeldt and Ngar-Kok Chung, Institute of Safety and Systems Management, University of Southern California; R. Duncan Luce and Younghee Cho, Institute of Mathematical Behavioral Sciences, University of California, Irvine.

This work was supported by the National Science Foundation under Grant SES-8921494 and Grant SES-9308915. We would like to thank Barbara Mellers and Michael Birnbaum for letting us use their stimulus materials and for many constructive discussions on the topic of this article.

Correspondence concerning this article should be addressed to Detlof von Winterfeldt, 2062 Business Center Drive, Suite 110, Irvine, California 92612.

Figure 1. A chance experiment in which with probability .80 one receives $96 and with probability .20 one receives $6 in the left gamble and $0 in the right gamble. Thus, the left gamble dominates the right one because $6 > $0.

alternative approach, favored by some psychologists, is that the evaluation of alternatives in terms of, for example, money is a (if not the) basic way of evaluating alternatives. This is what a storekeeper does in setting prices, although we usually suppose that the goods are actually worth less to the seller than the price he or she has set on them.

Our bias is that choices are the more fundamental of the two. Given that bias, then interpreting apparent violations of consequence monotonicity arising from judged certainty equivalents becomes problematic. If, perhaps by suitable instructions, we can induce decision makers to report the "true worth" of each alternative in the sense meant by choice theorists [namely, $CE(g)$ is the certainty equivalent of g if, in a choice between g and $CE(g)$, the decision maker is indifferent to which is received], then such judged certainty equivalents should establish the same ordering as do choices. However, this consistency does not appear to hold using any of the instructions so far formulated for eliciting judged certainty equivalents (Bostic, Herrnstein, & Luce, 1990; Tversky, Sattah, & Slovic, 1988). That being the case, is there a method of estimating certainty equivalents that is order preserving and, if so, do these estimates satisfy consequence monotonicity?

Bostic et al. (1990) adapted from psychophysics a choice method that, within the noise levels of choices, appears to be order preserving. We explore whether the estimates obtained using this method satisfy consequence monotonicity. After probing these issues, we ultimately conclude that in our experiments, at least, there is little effect of procedure, but

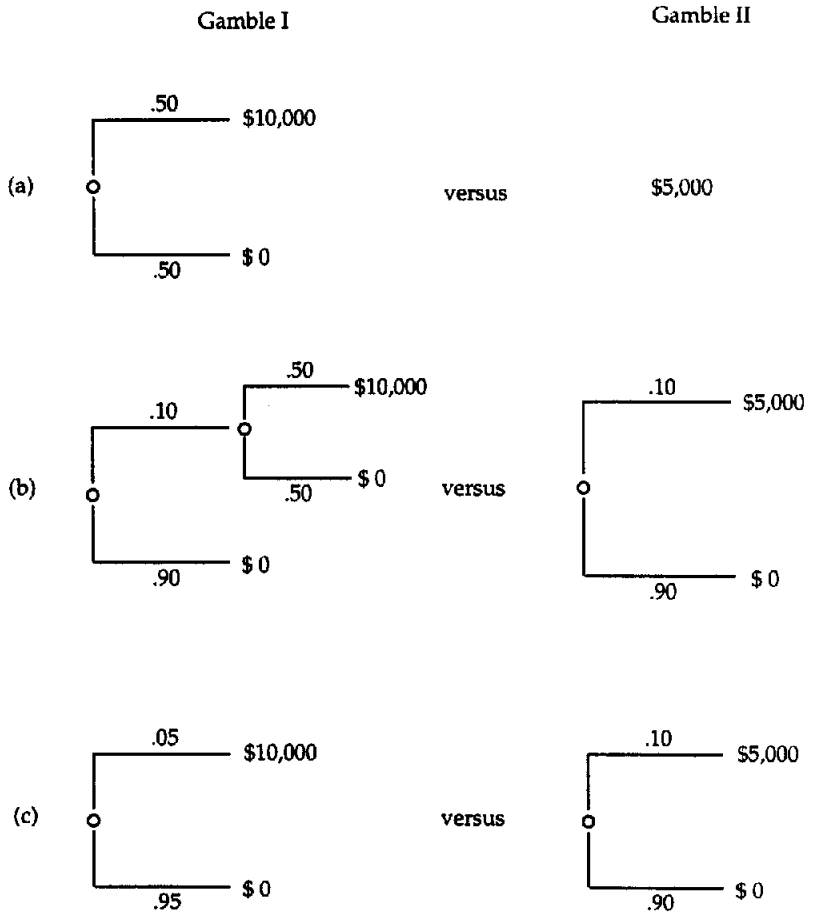

Figure 2. Illustration of the Allais paradox. The notation is as in Figure 1. Many participants prefer Gamble II to Gamble I in Conditions (a) and (b), which response is consistent with consequence monotonicity, but prefer Gamble I to Gamble II in Condition (c). The left side of (b) reduces under ordinary probability to the left side of (c). This reversal of preference between Conditions (a) and (c) is known as the Allais paradox.

rather that much of the apparent problem arises from variability of the estimates.

Consequence Monotonicity in Binary Gambles

By a *sure consequence* we mean a concrete object about which there is no uncertainty. In our experiments, sure consequences will be sums of money. Let **C** denote a set of sure consequences with typical elements x and y. A gamble is constructed by conducting some chance experiment—such as a toss of dice or a drawing from an urn of colored balls—the possible outcomes of which are associated with sums of money. In the simplest cases, as in Figure 1, let $(x,E;y)$ denote the binary gamble in which, when the chance experiment is realized, the sure consequence x is received if the event E occurs, or the sure consequence y is received if E does not occur. Further, let $\succsim$ be a preference relation over all binary gambles generated from the set **E** of all events and the set **C** of consequences.

Suppose E is an event that is neither the null nor the universal one and x, y, and z are sure consequences; then, (binary) *consequence monotonicity* is defined to mean

$x \succsim y$ if and only if

$$(x,E;z) \succsim (y,E;z) \text{ if and only if } (z,E;x) \succsim (z,E;y). \quad (1)$$

If in Equation 1 any of x, y, and z are gambles rather than sure consequences, then because two independent chance experiments have to be carried out to decide which consequence is received, the alternatives are called *second-order compound gambles* and we speak of *compound consequence monotonicity*. Throughout the article, we use the latter term only if the compound gamble is not reduced to its equivalent first-order form, that is, to a gamble in which a single chance experiment determines which consequence is received and does so under the same event conditions as in the second-order gamble. This is most easily illustrated when the only thing that is known about E is its probability p of occurring, in which case we write $(x,p;y)$. In that case, the first-order equivalent to the compound gamble $[(x,p;y),q;z]$ is $[x,pq; y,(1-p)q;z,(1-q)]$, with the meaning that x arises with probability pq, y with probability $(1-p)q$, and z with probability $(1-q)$.

Almost all contemporary formal models of decision making under uncertainty have a numerical representation in which the utility of a gamble is calculated as a sum over events of the utilities of the associated consequences times weights (somewhat similar to probabilities) that are attached to the events. Included in this general framework are all versions of subjective expected utility theory (Savage, 1954), prospect theory (Kahneman & Tversky, 1979) and its generalizations (Luce, 1991; Luce & Fishburn, 1991, 1995; Tversky & Kahneman, 1992; Wakker & Tversky, 1993), and several rank-dependent expected utility theories (for example, Chew, Karni, & Safra, 1987; Gilboa, 1987; Luce, 1988; Quiggin, 1982, 1993; Schmeidler, 1986; Segal, 1987, 1989; Wakker, 1989; Yaari, 1987). All of these theories imply Equation 1.

Previous Tests of Consequence Monotonicity

Direct tests of Equation 1 entail comparisons of such pairs as $(x,E;z)$ and $(y,E;z)$ or $(z,E;x)$ and $(z,E;y)$ when x is larger than y. Brothers (1990), working with money consequences and prescribed probabilities, created 18 pairs of gambles of this form. For example, a typical pair of gambles was (\$50,.67;\$100) vs. (−\$50,.67;\$100). Both gambles were presented in the form of pie charts on one page, and participants were asked to indicate their strength of preference for one of the two gambles on a rating scale that ranged from 0 (*no preference* = indifference) to 8 (*very strong preference*). In this situation, a participant merely had to recognize that the two gambles varied only in the first consequence, and his or her choices should follow clearly from which gamble had the higher first consequence. It would be very unusual to find violations of consequence monotonicity in such direct comparisons, and, indeed, Brothers found virtually none: Only 5 out of 540 judgments showed a violation, and the few participants (4 out of 30) who showed any violation classed them as "mistakes" in the debriefing session.

Brothers (1990) also tested compound consequence monotonicity. First, he established a preference or indifference between two first-order gambles that had identical or close-to-identical expected values but otherwise were dissimilar. For example, he asked his participants to compare the gambles g = (\$720,.33;−\$280) and h = (\$250,.40;−\$80). Having established a preference or indifference in this first pair of gambles, he then created two second-order gambles in which one outcome was one of the first-order gambles and the other outcome was a common monetary amount. For example, the test pair with the above first-order gambles was [\$50,.33;(\$720,.33;−\$280)] vs. [\$50,.33;(\$250,.33;−\$80)]. In these second-order gambles, \$50 was the consequence with probability .33 and either g or h, respectively, with probability .67. According to compound consequence monotonicity, the preference between the first-order gamble pair should persist when comparing the second-order gamble pair. The test stimuli were again presented side by side on a page, which thus allowed participants to recognize the common elements in the two second-order gambles. About 35% of the participants violated compound monotonicity. However, 34% of participants also reversed their choice of preference when the same pair of first-order gambles was presented on two occasions. Thus, the violations appeared to be mainly due to the participants' unreliability when comparing gambles with equal or similar expected values. In fact, most participants indicated that they recognized the common elements in the compound monotonicity tests and ignored them.

In contrast to Brothers's (1990) choice procedure and compound gambles, Birnbaum et al. (1992) used judged certainty equivalents and only first-order gambles, and they found a striking pattern of violations of consequence monotonicity to certain pairs of binary first-order gambles. Their gambles, which had hypothetical consequences ranging from \$0 to \$96, were represented as pie diagrams and were presented in a booklet. For each gamble, the participants wrote down their judged certainty equivalent under one of

three distinct points of view: Some were asked to provide "buying" prices, some "selling" prices, and some "neutral" prices. The neutral prices were described as meaning that the subjects are neither the buyer nor seller of the lottery, but rather a neutral judge who will decide the fair price or true value of the lottery so that neither the buyer nor the seller receives any advantage from participants' judgment. For most stimuli, consequence monotonicity held up quite well. However, for several special stimuli, the dominated gamble, on average, received a higher judged median certainty equivalent than did the dominating one. For example, in the pair ($96,.95;$0) and ($96,.95;$24), the former gamble, although dominated by the latter, received a higher median certainty equivalent than did the dominating gamble. This was also true when the amount of $96 was replaced by $72 and when the probability of .95 was replaced by any probability higher than about .80. Birnbaum et al. found this result to hold for about 50% of the participants for all three different points of view.

Birnbaum et al. (1992) interpreted this violation of monotonicity in terms of their configural-weighting model. In contrast to the standard weighted-average utility models, the configural-weighting model multiplies the utility of each consequence by a weight that is a function not only of the probability of receiving that consequence but also of the actual value of the consequence. Summed over all of the events in a gamble, the weights add to unity. The observed nonmonotonicity is easily accommodated in the configural-weighting model by assigning to low-probability events smaller weights when the consequence is 0 than when it is positive. Fitting weights to a large set of data, Birnbaum et al. found this effect for low-probability events. In this theory, weights can also depend on the point of view of the contemplated transaction (e.g., buying vs. selling), thus accommodating the observed differences in the preference order of the gambles arising from assuming different pricing strategies.

Subsequent replications and modifications of this experiment established that the effect is robust. For example, Mellers, Weiss, et al. (1992) used pie charts and peculiar dollar amounts in an attempt to reduce participants' tendency to do numerical calculations when estimating certainty equivalents. They also increased the hypothetical stakes to hundreds of dollars and used negative amounts. In all cases similar to those of the first study, they found violations except for gambles with mixed gains and losses and except when relatively few gambles were evaluated. In accord with the configural-weighting theory account, Mellers, Weiss, et al. hypothesized that participants are more likely to ignore, or put very little weight on, a zero outcome. In contrast, when both outcomes are nonzero, the configural weighting leads to some weighted average of the two utilities.

Further replications by Birnbaum and Sutton (1992) using judged certainty equivalents confirmed these findings. However, they also clearly showed that the effect disappeared when participants were presented with the pairs of gambles and were asked to choose between them directly.

Judged and Choice Certainty Equivalents

Our concern is whether the result of Birnbaum and colleagues is typical when certainty equivalents of any type are used or whether it disappears when elicited by some other procedure. Because the property of consequence monotonicity is stated in terms of choices, we suspect that some form of choice-based procedure for estimating certainty equivalents may yield a different result. One reason for entertaining this belief is the well-known preference-reversal phenomenon in which judged certainty equivalents of certain gambles exhibit an order that is the reverse of that determined by choices (Grether & Plott, 1979; Lichtenstein & Slovic, 1971; Mellers, Chang, Birnbaum, & Ordóñez, 1992; Slovic & Lichtenstein, 1983; Tversky et al., 1988; Tversky, Slovic, & Kahneman, 1990). Specifically, this reversal occurs when for two gambles with equal expected values, one has a moderately large probability of winning a small amount of money (called a *P-bet*) and the other has a small probability of winning a fairly large amount of money (called a *$-bet*). Typically the judged certainty equivalent for the $-bet is larger than that for the P-bet, whereas participants choose the P-bet over the $-bet when making a direct choice. The fact that judged certainty equivalents reverse the order of choices makes abundantly clear that they do not always yield the same information as choices.

In an attempt to estimate certainty equivalents that do exhibit the same ordering as choice, Bostic et al. (1990) adopted a psychophysical up–down method called parameter estimation by sequential testing (commonly referred to as PEST) in which participants engage only in choices between gambles and monetary amounts. Bostic et al. investigated the impact on the preference-reversal phenomenon of this choice-based method for estimating certainty equivalents as compared with judged estimates. A particular gamble–money pair is presented, a choice is made, and after many intervening trials that gamble is again presented with an amount of money. This time the money amount presented by the experimenter is (a) smaller than it was the first time if the money had been the choice in the earlier presentation or (b) larger if the gamble had been the choice. Such presentations are made many times, with the adjustments in the money amount gradually being reduced in size until the amounts are oscillating back and forth in a narrow region. An average of these last values is taken as an estimate of the certainty equivalent. The exact procedure is described in detail further on. Because such certainty equivalents are derived using choices, they are called *choice certainty equivalents*. Bostic et al. found that when choice certainty equivalents were used, the proportion of observed reversals was about what one would expect given the estimated inconsistency, or noise, associated with the choice procedure. In other words, they found no evidence against assuming that choice certainty equivalents agree with the choice-defined preference order. Thus, a natural question to consider is whether the violations of consequence monotonicity also tend to vanish when a choice procedure is used to establish certainty equivalents.

Birnbaum (1992) examined an alternative response mode

in which participants compared a gamble and a monetary amount on an ordered list of 27 monetary amounts and circled all of the monetary amounts on the list that they preferred to the gamble. Tversky and Kahneman (1992) used a similar approach in estimating certainty equivalents. Violations of consequence monotonicity persisted in the Birnbaum study. Although this procedure is formally a choice procedure, we are not at all convinced that it is likely to provide the same certainty equivalents as does an up–down choice procedure. In particular, it may be that participants follow the strategy of first establishing a judged certainty equivalent of the gamble and then circling the monetary amounts larger than that judged value. In that case, the procedure, although nominally one of choices, is virtually the same as a judged certainty equivalent procedure. Alternatively, even if a participant compares a gamble and each monetary amount individually, the estimated certainty equivalent may be affected by the distribution of money amounts in the list.

Our aim was to clarify whether or not the violations of monotonicity found using judged certainty equivalents could be significantly reduced by using choice-based certainty equivalents. In the first experiment, we used a fairly close analog to the Birnbaum et al. (1992) and Mellers, Weiss, et al. (1992) experiments. Our stimuli were very similar to theirs but were displayed on a computer screen individually. The certainty equivalents were estimated using three response modes: one direct judgment and two indirect choice procedures. Although this experiment used similar stimuli and elicited judged certainty equivalents as in the previous studies, we found substantially fewer violations than in the earlier experiments and, indeed, were left in serious doubt about whether the judgment procedure actually produces violations of consequence monotonicity. To find out whether our computer display of an individual gamble, which is different from their booklet display of multiple gambles, may have led to this difference, we conducted a second experiment that exactly replicated the Mellers, Weiss, et al. experiment and used both types of displays. The third experiment was an attempt to reduce further the noise levels found in Experiments 1 and 2. To that end we attempted to increase the realism of the first experiment, which we hoped would increase the participants' seriousness in evaluating the gambles, by providing a scenario in which the gambles were described as hypothetical college stipend applications with large potential dollar awards. Although the numbers of violations dropped, as we had hoped, considerable noise remained in the group data, making the test less sensitive than one would wish. The fourth, and final, experiment was aimed at modeling and estimating the magnitude and nature of the noise in individuals to see if noise is sufficient to account for the observed violations of monotonicity or whether the evidence really suggests underlying violations of monotonicity.

Experiment 1

As was just noted, the major purpose of Experiment 1 was to collect both judged and choice-based certainty equivalents to gambles and to check consequence monotonicity for each procedure. To that end, we used three different response modes to estimate certainty equivalents. In one, participants were asked to judge directly the certainty equivalents of gambles. The previous studies that used this mode had, as noted earlier, found substantial evidence for violations of consequence monotonicity. To investigate whether consequence monotonicity holds when certainty equivalents are derived from choice responses, we used the PEST procedure much as in the study by Bostic et al. (1990) in which participants were asked to choose between playing a gamble and taking a sure amount of money; the certainty equivalent of a gamble was derived by systematically changing the monetary amount paired with the gamble. Although we chose the PEST procedure to generate independent choice trials, it is lengthy and cumbersome. We therefore introduced a second choice procedure, similar to one sometimes used by decision analysts, which we called QUICKINDIFF for its presumed speed. Participants were asked to indicate a strength of preference for a given alternative until they reached a point of indifference between playing a given gamble and taking a sure amount of money.

To make our results comparable to the previous studies, we adopted the same stimuli used in Birnbaum et al. (1992) and Mellers, Weiss, et al. (1992), but we elected to display them on a computer screen individually rather than as drawings in a booklet.

Method

Participants. Thirty-three undergraduate students at the University of California, Irvine participated in this experiment for partial credit for courses.

Stimuli and design. The 15 stimuli are shown in Table 1. They were adopted from the study by Birnbaum et al. (1992) and closely

Table 1
Stimuli Used in Experiment 1

Number	Gamble	Number	Gamble	Number	Gamble
1	($96,.05;$0)	6	($96,.05;$6)	11	($96,.05;$24)
2	($96,.20;$0)	7	($96,.20;$6)	12	($96,.20;$24)
3	($96,.50;$0)	8	($96,.50;$6)	13	($96,.50;$24)
4	($96,.80;$0)	9	($96,.80;$6)	14	($96,.80;$24)
5	($96,.95;$0)	10	($96,.95;$6)	15	($96,.95;$24)

Note. A gamble denoted as ($x,$p$;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$.

matched their design. The independent variables were the lowest outcome ($0, $6, or $24), the probability to win $96 (.05, .20, .50, .80, or .95), and the response mode (judged certainty equivalent, or JCE; QUICKINDIFF; PEST). The highest consequence was always $96. All of these variables were manipulated in a within-subject design; each participant provided certainty equivalents for all 15 stimuli under all three response modes.

Stimulus presentation and response modes. All gambles and sure consequences were presented on a computer screen as shown in Figure 3. The sure consequence was always on the right side and the gamble was on the left side of the balance beam. A "blown up" version of the gamble was also given on the left side. The pie segment representing the risky event in the gamble was scaled to correspond in size to the numerical probability. The bar at the bottom of the display was used only with the QUICKINDIFF procedure to indicate strength of preference. The display was presented in color on a 13-in. (33-cm) Macintosh screen, with the pie segments in blue and red and with the areas showing the dollar amounts shaded in green.

For the JCE response mode, the sure amount box on the right of the display had a question mark rather than a monetary amount, and the bar at the bottom of Figure 3 was not displayed. Participants were instructed to type in the monetary amount that reflected their judged certainty equivalent to the gamble in the other box. The participants were told that this value was the dollar amount to which they would be exactly indifferent in a choice between taking it for sure and playing the gamble. This was the same instruction used for the neutral point of view in Birnbaum et al.'s (1992) study. After the amount was typed in, the computer screen opened a dialogue box which asked, "Are you really indifferent?" If they were not indifferent, they could go back and reenter the certainty equivalent. If they were satisfied with that value, it was stored as the judged certainty equivalent for the gamble in question.

One choice procedure, called PEST, was similar to that of Bostic et al. (1990). Again, this display was as in Figure 3 except that the strength-of-preference bar was omitted. On the first presentation of any gamble, a sure amount was selected from a uniform distribution between the minimum and the maximum amount of the gamble. Participants were instructed to click the OK button under their choice, either playing the gamble or receiving the sure amount. Once their response was entered, the computer algorithm calculated a second sure amount by taking a fixed step size (⅓ of the range between the minimum and maximum consequences of the gamble) in the direction of indifference. For example, with the stimulus in Figure 3, the step is approximately $19, and so if the participant indicated a preference for the sure amount of $50, the next sure amount presented would be $31 = $50 − $19. Alternatively, if the participant indicated a preference for the gamble, the second sure amount would be $69 = $50 + $19. The computer stored the modified money amount for the next presenta-

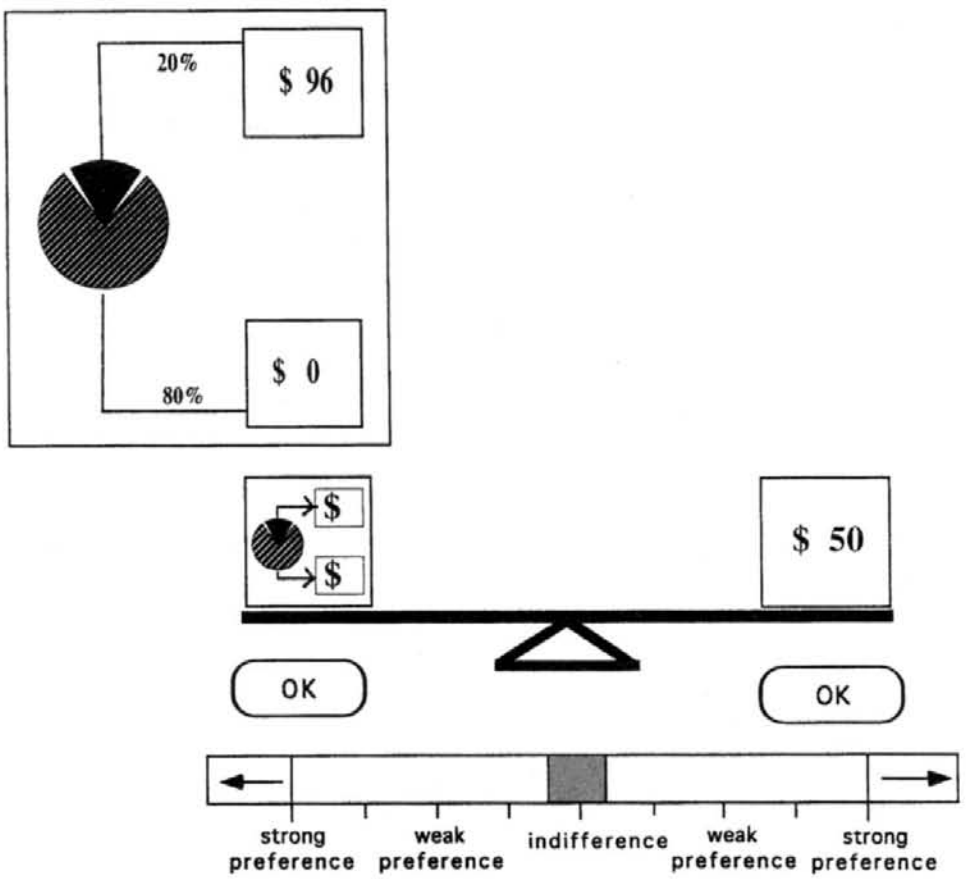

Figure 3. Computer display of gamble stimulus. The bottom preference bar appeared only in the QUICKINDIFF task, not in the JCE and PEST tasks. QUICKINDIFF was the quick indifference procedure (based on sequential strength-of-preference judgments); JCE = judged certainty equivalent; PEST = parameter estimation by sequential testing.

tion of the gamble, which would occur after some intervening trials consisting of other gambles.

On the second presentation of the gamble in question and the modified sure amount, the participant again entered his or her preference, and the computer calculated the next sure amount. If the participant preferred either the sure amount or the gamble three times in a row, the step size was doubled. If the participant changed the preference either from the sure amount to the gamble or vice versa, the step size was halved. If a participant made a "wrong" choice (in the sense of choosing the sure amount when it was smaller than the worst consequence of the gamble, or choosing the gamble when the sure amount was larger than the best consequence of the gamble), the computer alerted the participant to the mistake and asked for a revision.

The procedure was terminated when the step size became smaller than $\frac{1}{50}$ of the range of consequences. The certainty equivalent of the gamble was then estimated to be the average of the lowest accepted and the highest rejected sure amounts among the last three responses.

To try to keep participants from recalling previous responses to particular gambles, we conducted the PEST trials as follows. First, there were 15 test stimuli, and each was presented once before the second round of presentations was begun. Second, 20 filler stimuli were interspersed randomly among the test stimuli. These were similar to the test gambles but had different consequences. After each trial, whether a test or filler stimulus, there was a 50% chance that the next item would be a filler. If it was, one of the 20 filler stimuli was selected at random. If it was a test stimulus, the next one in line on the test stimulus list was selected. Because of this procedure, the number of filler stimuli between tests varied and, more important, even after most of the test stimuli were completed and so eliminated from further presentation, there were still enough fillers to mask the nature of the sequencing.

The QUICKINDIFF procedure began, as did the PEST procedure, by presenting a gamble and a sure amount that was randomly chosen from the consequence range of the gamble. However, instead of being asked for a simple choice, the participant was asked to provide a strength-of-preference judgment for the chosen alternative. To assist participants in expressing their strength of preference, we showed a horizontal bar below the balance beam with markers indicating varying strengths ranging from *indifference* to *strong preference* (see Figure 3). Participants simply dragged the shaded square marker in the direction of the preferred item, stopping at the point that best reflected their strength of preference.

After each response, the QUICKINDIFF algorithm calculated the next sure amount, taking into account both the direction and strength of preference. It did so by using a linear interpolation between the sure amount and the maximum consequence of the gamble when the preference was for the gamble and between the sure amount and the minimum of the gamble when the preference was for the sure amount. The step size in the interpolation was determined by the strength of preference. If that strength was large, the step size was large; if it was small, the step size was small. The algorithm also reset the ranges for the possible certainty equivalent. When the sure amount was chosen, that value became a new upper bound for the certainty equivalent, and when the gamble was chosen, the sure amount was a new lower bound for the certainty equivalent.

After several such trials, participants tended to report indifference in choosing between the gamble and the sure amount and pressed one of the OK buttons, leaving the square marker at the indifference point. The program again checked by asking "Are you really indifferent?" thus allowing the participant to reenter the process or exit it. After exiting, the final sure amount was recorded as the estimated certainty equivalent for the gamble.

Procedure. Participants were introduced to the experiment, the display, and the first response mode. Several trial items were presented and the experimenter supervised the participants' responses to these items, responded to questions, and corrected obvious errors. The experimenter instructed participants that they should make their judgments independently on each trial and assume that they could play the gamble but once. They were encouraged to think of the amounts of money as real outcomes of real gambles. To help the participant understand the stimuli presented on the computer screen, the experimenter demonstrated the displayed gamble with an actual spinner device—a physical disk with adjustable regions and a pointer that could spin and whose chance location in a region would determine the consequence received.

All data in a response mode were collected before the next one was started. The JCE response mode was presented either first or last, and the other two modes were counterbalanced. The initial order of stimuli was randomized separately for each participant and each response mode.

To motivate the participants, we placed them into a competition where the outcome was based on their responses. After the experiment, 10 trials were chosen randomly for each participant, and scores were calculated according to the participant's responses to these 10 trials as follows: For each of the chosen trials, if the participant had selected the sure amount, the score was that amount; otherwise, the gamble was played on the spinner and its dollar outcome was the score. The total score for each individual was the sum of the scores in these 10 selected trials. The 3 participants who received the highest scores were paid off with a dollar amount equaling one tenth of their scores. (The actual payoffs determined after the experiment were \$66.10, \$64.10, and \$58.50 for these 3 participants.)

The experiment was run in a single session with breaks. It lasted between 1 and 2 hr.

Results

Fifteen pairs of gambles exhibited a strict consequence dominance relationship as shown in Table 2. We operationally defined, as did Mellers, Weiss, et al. (1992), that a violation of monotonicity had occurred when a participant gave a higher certainty equivalent to the dominated gamble. Table 2 shows the percentages of participants who violated consequence monotonicity, so defined, for each of the 15 stimulus pairs and for each of the three response modes. Overall, the percentages of violations of consequence monotonicity were 19% with the PEST procedure, 30% with the QUICKINDIFF procedure, and 31% with the JCE procedure. There were occasional ties, primarily with the JCE method, when the probability of winning the \$96 amount was .95. The ties were caused by participants either choosing the maximum amount of \$96 or rounding the responses for these gambles to either \$95 or \$90.

The JCE and QUICKINDIFF data show violations of consequence monotonicity for pairs of stimuli at all values of p, whereas the PEST data show a proportion increasing from about 0% when $p = .05$ to as much as 39% when $p = .95$. The low p values are, of course, the gamble pairs with larger differences in expected values. At the highest p value,

Table 2
Percentages of Participants Violating Consequence Monotonicity in Experiment 1 (N = 31)

Stimulus pair	Procedure		
	PEST	QUICKINDIFF	JCE
($96,.05;$0) vs. ($96,.05;$6)	0	23	26
($96,.20;$0) vs. ($96,.20;$6)	16	39	52
($96,.50;$0) vs. ($96,.50;$6)	35	58	45
($96,.80;$0) vs. ($96,.80;$6)	39	39	42
($96,.95;$0) vs. ($96,.95;$6)	32	39	32
($96,.05;$6) vs. ($96,.05;$24)	3	3	6
($96,.20;$6) vs. ($96,.20;$24)	0	3	13
($96,.50;$6) vs. ($96,.50;$24)	16	32	32
($96,.80;$6) vs. ($96,.80;$24)	26	42	42
($96,.95;$6) vs. ($96,.95;$24)	39	45	39
($96,.05;$0) vs. ($96,.05;$24)	0	0	16
($96,.20;$0) vs. ($96,.20;$24)	0	10	23
($96,.50;$0) vs. ($96,.50;$24)	10	26	19
($96,.80;$0) vs. ($96,.80;$24)	32	35	48
($96,.95;$0) vs. ($96,.95;$24)	39	52	35
Average	19	30	31

Note. A gamble denoted as ($x,p;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. PEST = parameter estimation by sequential testing; QUICKINDIFF = quick indifference procedure (based on sequential strength-of-preference judgments); JCE = judged certainty equivalent.

the PEST and JCE data have comparable proportions of violations, and the QUICKINDIFF data are somewhat worse.

Because the estimated certainty equivalents were quite variable among participants, we chose the median rather than the mean responses to represent aggregate findings. As in Mellers, Weiss, et al. (1992), we plot the median certainty equivalents against the probability of winning the larger amount ($96) in the gamble for the three response modes separately. The results are shown in the three panels of Figure 4. Each set of connected points shows the median certainty equivalents as a function of the probability of receiving the common value of $96. The curves are identified by whether the complementary outcome is $0, $6, or $24. According to consequence monotonicity, the curve for the $24 gambles should lie wholly above that for the $6 gamble, which in turn should lie wholly above that for the $0 gamble. In all methods, we find a slight tendency for crossovers between the $0 and $6 curves and very few crossovers between the $24 curve and other curves.

We applied a two-tailed Wilcoxon test to compare the certainty equivalents of each pair across participants for each value of the gamble probability p. The results are shown in Table 3. None of the pairs at $p = .95$ was significantly different for any of the response modes. Some pairs, notably those for $p = .05$, .20, or .50, showed significant differences (i.e., at $p < .05$ or $p < .01$). Because the Wilcoxon test does not indicate whether a significant difference is in the direction of supporting or violating consequence monotonicity, we inferred this information by directly comparing the median certainty equivalents. All instances of significant Wilcoxon tests in Table 3 support consequence monotonicity.

Discussion

Our main concern in this experiment was to compare the response-mode effects in testing consequence monotonicity, especially for gambles with a high probability of receiving $96. For all response modes, we found some proportions of

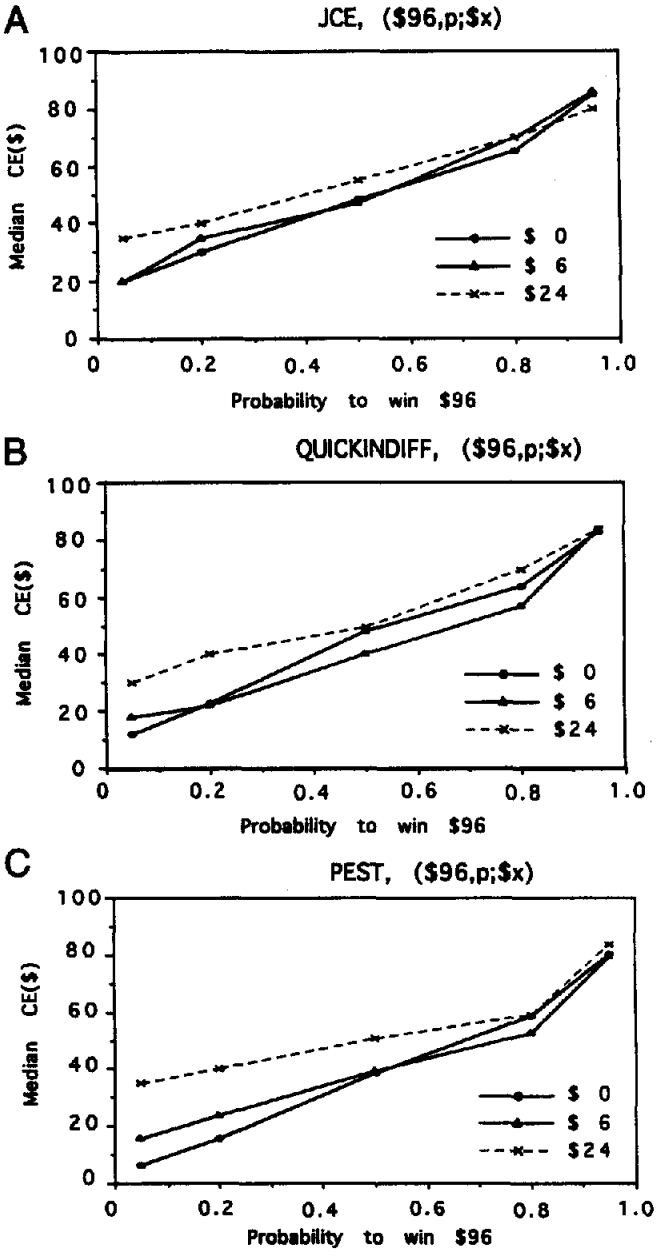

Figure 4. Median certainty equivalents (CEs) for the gambles ($96,$p$;$x), $x being one of $0, $6, or $24 ($N = 31$). JCE = judged uncertainty equivalent; QUICKINDIFF = quick indifference procedure (based on sequential strength of preference judgments); PEST = parameter estimation by sequential testing.

Table 3
Z Scores for the Wilcoxon Tests of the Certainty Equivalents (CEs) of Gamble Pairs in Experiment 1

Stimulus pair	$p = .05$	$p = .20$	$p = .50$	$p = .80$	$p = .95$
			JCE		
CE($96,*p*;$0) vs. CE($96,*p*;$6)	−0.95	−1.45	−1.29	−0.05	−0.71
CE($96,*p*;$6) vs. CE($96,*p*;$24)	−3.77**	−3.27**	−1.91	−0.22	−1.09
CE($96,*p*;$0) vs. CE($96,*p*;$24)	−2.84**	−2.29*	−2.91**	−0.13	−0.77
			QUICKINDIFF		
CE($96,*p*;$0) vs. CE($96,*p*;$6)	−2.99**	−0.65	−0.56	−0.06	−0.10
CE($96,*p*;$6) vs. CE($96,*p*;$24)	−4.59**	−4.33**	−2.54*	−1.71	−0.22
CE($96,*p*;$0) vs. CE($96,*p*;$24)	−4.78**	−4.56**	−2.25*	−1.96*	−0.14
			PEST		
CE($96,*p*;$0) vs. CE($96,*p*;$6)	−4.86**	−2.63**	−1.84	−1.12	−1.76
CE($96,*p*;$6) vs. CE($96,*p*;$24)	−4.84**	−4.86**	−4.29**	−2.13*	−1.31
CE($96,*p*;$0) vs. CE($96,*p*;$24)	−4.86**	−4.78**	−4.46**	−2.52*	−1.85

Note. The asterisks indicate significant differences that are in support of the monotonicity assumption. A gamble denoted as ($x,*p*;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. JCE = judged certainty equivalent; QUICKINDIFF = quick indifference procedure (based on sequential strength-of-preference judgments); PEST = parameter estimation by sequential testing.
*$p < .05$. **$p < .01$.

violations for these gambles; on the whole, they were slightly lower for the PEST response mode than for the JCE response mode, and both of these were substantially lower than that for the QUICKINDIFF response mode. This was mainly due to the clear differences in the observed proportions of violations for the gambles of the form ($96,*p*;$0) and ($96,*p*;$6) when p was low. For example, for $p \leq .50$, these violations were less for the PEST procedure (ranging from 0% to 35%) than for the QUICKINDIFF (23% to 58%) and JCE (26% to 52%) procedures.

The smaller proportion of violations found for the PEST procedure might have arisen because the first sure amount was chosen randomly from the range of outcomes. This means that the first sure amount for ($96,*p*;$24) is on average larger than that for ($96,*p*;$6), which, in turn, is on average larger than that for ($96,*p*;$0). Beginning the PEST sequence with a larger sure amount might lead to higher certainty equivalents, as reported by Brothers (1990), which in turn would lead to a smaller proportion of violations than with the JCE procedure. To check for this, we reanalyzed the PEST data by eliminating all certainty equivalents whose first sure amounts for the ($96,*p*;$0) were below $6 when the gamble was compared with ($96,*p*;$6) or below $24 when the gamble was compared with ($96,*p*;$24). There still was no clear pattern of violation of monotonicity.

The major conclusion from this experiment is that according to overall violations of consequence monotonicity, the methods are ordered PEST < JCE < QUICKINDIFF. However, for large values of p, where the most reversals occur, there is little difference between the PEST and JCE methods. It may be pertinent to note that for gambles with a high proportion (.80 and .95) of receiving $96, the proportions of violations we observed under the JCE procedure (32% to 48%) were substantially less than those found in the studies by Birnbaum, et al. (1992; 60%) and Mellers, Weiss, et al. (1992; 54%). Although we adopted the same stimulus and response mode as in these previous studies, there were some procedural differences. In Experiment 2 we attempted to make a closer comparison of our procedures and those of the earlier studies.

The largest proportion of violations occurred under the QUICKINDIFF procedure. Even though this procedure was designed to mimic an elicitation method often used by decision analysts, by no means did it appear immune to violations of consequence monotonicity. Perhaps this procedure allows, or even invites, participants to exit the program prematurely by stating indifference between a displayed sure amount and the gamble when it is only approximately determined. If a participant accepts indifference without much thought, the results would show a stronger component of random error and thereby increase apparent violations of monotonicity.

Across all response modes, most violations occur when the gambles have a high probability of winning $96 and so have very similar expected values. Thus it is possible that some of these violations were due to random errors in estimating certainty equivalents. The median pattern shows no crossovers for the $24 stimulus and only a minor crossover for the $6 stimulus at the .80 probability of winning $96. Furthermore, the Wilcoxon tests showed that the certainty equivalents of any pair of gambles at $p = .95$ were not significantly different. Without further data, we cannot be sure whether the obtained proportions of violations reflect true violations of monotonicity or result simply from random fluctuations in participants' responses.

To gain additional understanding of the relation between these methods and consequence monotonicity, we needed to take several directions. The first was to try to determine

whether our somewhat lower—although still high—proportion of violations than was found in earlier experiments was due to changing from a paper display of the stimuli to one on a computer monitor. We did this in Experiment 2 (although it was actually run after Experiment 3). A second concern was whether we could figure out a better way to capture the participants' attention and thereby reduce further the noise level. We did this in Experiment 3 by attempting to provide money amounts and a scenario that would be quite meaningful in the real world. And third, there was the serious problem of noise in the data, both between and within subjects. In Experiment 4 we attempted to estimate the nature and amount of the noise we were encountering and to determine if it was sufficient to explain, under the hypothesis that consequence monotonicity holds, the observed violations.

Experiment 2

Of the several procedures used in Experiment 1 (and also Experiment 3; see below), the JCE procedure most closely resembles that used in earlier experiments. Although this procedure produced medium to high percentages of violation (between 32% and 48% for the most susceptible gamble pairs), they are somewhat less than those found by Birnbaum et al. (1992) and Mellers, Weiss, et al. (1992) for similar stimuli and conditions (neutral point of view). Several procedural differences might have caused the difference in obtained violations. We displayed the stimuli on the computer screen as opposed to in a booklet, and the number of stimuli was smaller (15 stimuli) than in the study by Mellers, Weiss, et al. Although we are not sure whether the display affected the results, it is also possible that, instead, the smaller number of stimuli could have affected the results. With a larger number of stimuli, participants might have developed simplifying estimation strategies. Such strategies might be expected to involve calculations that are more prone to response biases. In fact, Mellers, Weiss, et al. reported a smaller proportion of violations for fewer stimuli (30 or fewer). In addition, the financial incentive instructions—that 10 gambles would be played to determine the participant's score, and those who won the three highest scores would become winners—might have led participants to be less risk averse or even risk seeking, because the payoff is the average and only the participants with the top three averages would win money. Therefore, in the present experiment, we attempted to replicate the studies of Birnbaum et al. and Mellers, Weiss, et al. both in their original booklet presentation and in our computer display.

Method

Participants. Thirty-three undergraduate students from the University of California, Irvine and 10 undergraduate students from the University of Southern California participated in this experiment for partial credit for psychology courses.

Stimuli and design. The stimuli were those previously used and provided to us by Mellers, Weiss, et al. (1992). In total, there were 77 gambles, 45 of which were test gambles and the remaining 32 of which were fillers. Table 4 shows the complete list of the 45 test gambles. The 45 test gambles consisted of 3 sets of 15 gambles. For a gamble $(x,p;y)$, p was either .05, .20, .50, .80, or .95. In the first set, x was \$96 and y was either \$0, \$6, or \$24. In the second set, x and y were 5 times their values in the first set, and in the third set they were 10 times their values in the first set.

The main factor under investigation was possible differences between the booklet and computer task in testing consequence monotonicity. All of the conditions were compared within subjects. Each participant received all 77 gambles and performed both tasks.

Stimuli presentation. In the replication of the Mellers, Weiss, et al. (1992) and Birnbaum et al. (1992) experiments, the materials were presented in a seven-page booklet in which all gambles were presented as pie charts without numerical probabilities. Dollar amounts were attached to the two segments of the pie chart. The booklet contained an instruction page, which was followed by a page with 10 warm-up stimuli. The test and filler stimuli were presented on five subsequent sheets with up to 18 stimuli on each sheet. The order of the stimulus sheets in the booklet task was

Table 4
Stimuli Used in Experiment 2

Number	Gamble	Number	Gamble	Number	Gamble
1	(\$96,.05;\$0)	16	(\$480,.05;\$0)	31	(\$960,.05;\$0)
2	(\$96,.05;\$6)	17	(\$480,.05;\$30)	32	(\$960,.05;\$60)
3	(\$96,.05;\$24)	18	(\$480,.05;\$120)	33	(\$960,.05;\$240)
4	(\$96,.20;\$0)	19	(\$480,.20;\$0)	34	(\$960,.20;\$0)
5	(\$96,.20;\$6)	20	(\$480,.20;\$30)	35	(\$960,.20;\$60)
6	(\$96,.20;\$24)	21	(\$480,.20;\$120)	36	(\$960,.20;\$240)
7	(\$96,.50;\$0)	22	(\$480,.50;\$0)	37	(\$960,.50;\$0)
8	(\$96,.50;\$6)	23	(\$480,.50 \$30)	38	(\$960,.50;\$60)
9	(\$96,.50;\$24)	24	(\$480,.50;\$120)	39	(\$960,.50;\$240)
10	(\$96,.80;\$0)	25	(\$480,.80;\$0)	40	(\$960,.80;\$0)
11	(\$96,.80;\$6)	26	(\$480,.80;\$30)	41	(\$960,.80;\$60)
12	(\$96,.80;\$24)	27	(\$480,.80 \$120)	42	(\$960,.80;\$240)
13	(\$96 .95;\$0)	28	(\$480,.95;\$0)	43	(\$960,.95;\$0)
14	(\$96,.95;\$6)	29	(\$480,.95;\$30)	44	(\$960,.95;\$60)
15	(\$96 .96;\$24)	30	(\$480,.95 \$120)	45	(\$960,.95;\$240)

Note. A gamble denoted as $(\$x,p;\$y)$ means the following: Obtain \$$x$ with probability p; otherwise \$$y$ with probability $1 - p$.

varied according to a Latin square design. Each gamble had a number associated with it. The response sheet had numbered lines for the 10 warm-up trials and numbered lines for all the gambles. The columns were labeled "Sell" and "Buy." The participants were instructed to judge the certainty equivalents under either the seller's point of view or the buyer's point of view as follows:

> Imagine that you own all the gambles in the experiment. . . . If you think the gamble is desirable, write the minimum amount (dollars and cents) you would accept to sell the gamble in the column labeled "Sell." If you think the gamble is undesirable, write the maximum amount (dollars and cents) you would pay to avoid playing it in the column labeled "Pay."

Participants wrote the appropriate price for each gamble on the corresponding number on the response sheets.

The computer-generated stimulus display was identical to that described in Experiment 1, but only the JCE response mode was used. The gambles were displayed as shown in Figure 3, both with pie charts and probabilities. The instructions were similar to those described in Experiment 1 except they omitted the motivation for actually playing some gambles and winning real money. All 77 stimuli were presented in random order.

Procedure. The tasks were run in separate sessions, separated by at least 1 week. Each session lasted for about 45 min. The order of the two tasks was counterbalanced across participants.

Results

Because we detected no notable differences between the students from the two universities, we pooled their data. Table 5 shows the percentages of monotonicity violations for the nine pairs of stimuli for which Mellers, Weiss, et al. (1992) and Birnbaum et al. (1992) had found the strongest violations. The overall percentages of violations were about the same for the booklet task (36%) as for the computer task (37%). They are about the same as those in Experiment 1 and again are somewhat lower than those reported by Mellers, Weiss, et al. and Birnbaum et al. Table 5 also shows the percentages of ties, which were surprisingly high in both tasks. Most ties seemed to occur because participants rounded responses, for example, to \$95 or \$90 in the (\$96,.95;\$0) gamble.

The panels in Figure 5 show the median certainty equivalents of the test gambles in Table 5 as a function of the probability of winning the higher amount. Although there are some crossovers at the higher probability end, none is striking. Table 6 shows the results of Wilcoxon tests for the nine gamble pairs with $p = .95$ for receiving the largest outcome. Only one test shows a significant difference in the booklet task, and this test supports monotonicity. In the computer task, two tests show significant differences in the direction of violating monotonicity, and two show a significant difference in support of monotonicity. All other tests have nonsignificant p values.

Discussion

With both the computer and booklet tasks, we found some evidence of a violation of consequence monotonicity for the nine stimulus pairs for which previous experiments showed the strongest violations. Apparently, the introduction of the computer procedure neither decreased nor increased the number of violations. The observed proportion of violations in the booklet task (36%) was again somewhat smaller than that obtained by Mellers, Weiss, et al. (1992) and Birnbaum et al. (1992). The crossovers at high probability in the plot of median certainty equivalents against the probability to receive the highest outcome were less pronounced, but three pairs for the computer task were significantly different in the direction of violations of monotonicity. The percentage of ties (38%) was larger than that found by Mellers, Weiss, et al. (14%) and by Birnbaum et al. (28%) for equivalent conditions. When ties are ignored, the ratio of violations to nonviolations was, however, similar to the ratios from the earlier studies.

Table 5
Percentages of Participants Violating Consequence Monotonicity or Assigning Equal Certainty Equivalents (Ties) in Experiment 2 (N = 43)

	Booklet		Computer	
Stimulus pair	Percent violation	Percent ties	Percent violation	Percent ties
(\$96,.95;\$0) vs. (\$96,.95;\$6)	30	33	47	37
(\$96,.95;\$6) vs. (\$96,.95;\$24)	28	44	44	35
(\$96,.95;\$0) vs. (\$96,.95;\$24)	35	35	49	40
(\$480,.95;\$0) vs. (\$96,.95;\$30)	26	60	28	40
(\$480,.95;\$30) vs. (\$96,.95;\$120)	35	33	21	44
(\$480,.95;\$0) vs. (\$96,.95;\$120)	42	33	23	28
(\$960,.95;\$0) vs. (\$96,.95;\$60)	47	33	40	35
(\$960,.95;\$60) vs. (\$96,.95;\$240)	30	35	40	37
(\$960,.95;\$0) vs. (\$96,.95;\$240)	51	33	44	42
Average	36	38	37	38

Note. A gamble denoted as (\$$x$,$p$;\$$y$) means the following: Obtain \$$x$ with probability p; otherwise \$$y$ with probability $1 - p$.

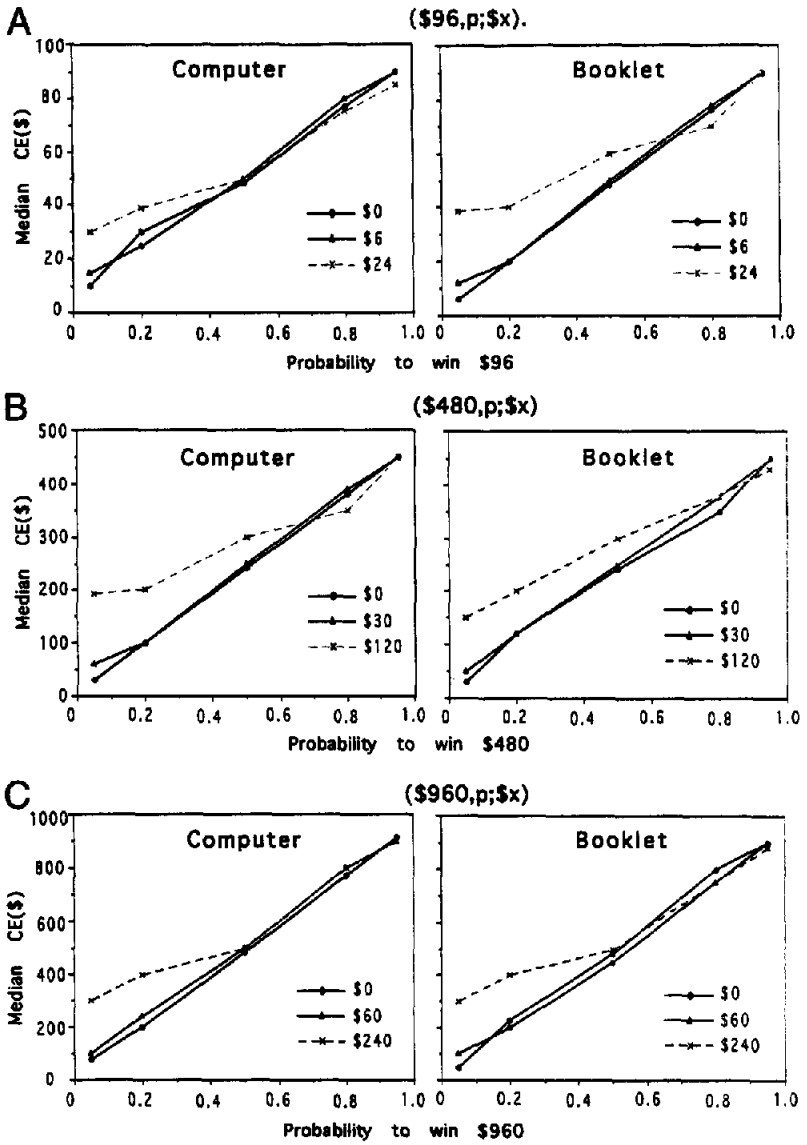

Figure 5. Median judged certainty equivalents (CEs) for the computer task versus the booklet task for the gambles ($96,*p*;$*x*), ($480,*p*;$*x*), and ($960,*p*;$*x*), with $*x* as given in the legends (N = 43).

In summary, this experiment did not show clear differences between the computer and booklet displays in observed proportions of violations; however, we did observe consistently fewer violations than in the earlier experiments. These persistent differences in our experiments, whatever their cause, cannot be attributed to such procedural differences as mode of stimulus display or number of stimuli.

Experiment 3

As noted earlier, the purpose of this experiment was to increase somewhat the realism of the situation. This was achievable in at least two ways. The first way, used in Experiment 1, was to play out for money some of the gambles after completion of the experiment, thereby providing the participants an opportunity to win more or less money depending on their choices and on chance. Our intention was to create monetary incentives, but the stakes were necessarily rather modest. The other way, pursued in the present experiment, was to create a somewhat realistic scenario for the gambles but without an actual opportunity to play. The stakes can be increased under this approach but at the expense of not providing any direct actual financial incentive. For gambles and stakes we presented hypothetical stipend offers for amounts between $0 and $9,600. In all other regards, Experiment 3 was identical to Experiment 1.

Method

Participants. Twenty-four undergraduate students from the University of Southern California participated in the experiment, and they received a flat fee of $6 per hour for their participation.

Table 6
Z Scores for the Wilcoxon Tests of the Two Certainty Equivalents (CEs) in Experiment 2

Stimulus pair	Booklet task	Computer task
CE($96,.95;$0) vs. CE($96,.95;$6)	−1.26	−2.02*
CE($96,.95;$6) vs. CE($96,.95;$24)	−0.94	−2.40††
CE($96,.95;$0) vs. CE($96,.95;$24)	−1.87*	−3.25††
CE($480,.95;$0) vs. CE($480,.95;$30)	−0.69	−0.32
CE($480,.95;$30) vs. CE($480,.95;$120)	−1.46	−0.74
CE($480,.95;$0) vs. CE($480,.95;$120)	−0.75	−1.20
CE($960,.95;$0) vs. CE($960,.95;$60)	−0.94	−1.26
CE($960,.95;$60) vs. CE($960,.95;$240)	−0.55	−0.94
CE($960,.95;$0) vs. CE($960,.95;$240)	−1.37	−1.87*

Note. The asterisks indicate significant differences that are in support of the monotonicity assumption. The daggers indicate significant differences that are in violation of the monotonicity assumption. A gamble denoted as ($x,p;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$.
*$p < .05$. ††$p < .01$.

Stimulus presentation and response modes. These were identical to those of Experiment 1 except that all dollar amounts were multiplied by 100.

Design. The design was identical to that of Experiment 1.

Procedure. Participants were introduced to the task by the experimenter. They were told to imagine that they had to choose between two options for obtaining a stipend for the next semester. The first was to accept a sure offer right now, for example, $2,400. The second was to forgo the sure offer and to apply for a second stipend that carried a larger amount ($9,600); however, there was a chance (which varied over .05, .20, .50, .80, and .95) that the second stipend would be lowered to either $0, $600, or $2,400. Participants were told to imagine that the uncertainty would be resolved quickly but that once they gave up the sure stipend, it would be given to someone else and they could not revert to it.

For each response mode, the experimenter guided the participants through some learning trials and then set them up to complete all of the trials of that response mode before they started the next response mode. The experiment lasted from 1 ½ to 2 hr.

Results

Table 7 shows the overall pattern of violations of consequence monotonicity for all stimulus pairs and the three response modes. The overall percentages of violations were 14% for PEST, 25% for QUICKINDIFF, and 17% for JCE, which are smaller than those obtained in Experiment 1 and are ordered in the same way. The percentages of violations were also lower for the gamble pairs with low probabilities

Table 7
Percentages of Participants Violating Consequence Monotonicity in Experiment 3 (N = 24)

Stimulus pair	Procedure		
	PEST	QUICKINDIFF	JCE
($9,600,.05;$0) vs. ($9,600,.05;$600)	4	25	4
($9,600,.20;$0) vs. ($9,600,.20;$600)	8	29	21
($9,600,.50;$0) vs. ($9,600,.50;$600)	29	17	42
($9,600,.80;$0) vs. ($9,600,.80;$600)	29	29	29
($9,600,.95;$0) vs. ($9,600,.95;$600)	33	63	33
($9,600,.05;$600) vs. ($9,600,.05;$2,400)	4	8	8
($9,600,.20;$600) vs. ($9,600,.20;$2,400)	4	17	8
($9,600,.50;$600) vs. ($9,600,.50;$2,400)	4	25	8
($9,600,.80;$600) vs. ($9,600,.80;$2,400)	17	24	29
($9,600,.95;$600) vs. ($9,600,.95; $2,400)	29	33	21
($9,600,.05;$0) vs. ($9,600,.05;$2,400)	0	8	4
($9,600,.20;$0) vs. ($9,600,.20;$2,400)	0	21	4
($9,600,.50;$0) vs. ($9,600,.50;$2,400)	4	13	4
($9,600,.80;$0) vs. ($9,600 .80;$2,400)	17	24	21
($9,600,.95;$0) vs. ($9,600,.95;$2,400)	25	46	33
Average	14	25	17

Note. A gamble denoted ($x,p;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. PEST = parameter estimation by sequential testing; QUICKINDIFF = quick indifference procedure (based on sequential strength-of-preference judgments); JCE = judged certainty equivalent.

of winning $9,600, which were the gamble pairs with the largest expected-value differences. For the three stimuli that most directly tested Mellers, Weiss, et al.'s (1992) violation pattern (i.e., those that involved a .95 probability of winning $9,600), the percentages were largest, but again not as large as in previous studies: The highest percentages of violations were found in the QUICKINDIFF procedure (33%, 46%, and 63%), followed by the closely similar JCE (21%, 33%, and 33%), and PEST (29%, 33%, and 25%) procedures.

The three panels of Figure 6 show the median certainty equivalents as a function of the probability of receiving $9,600. When $p = .95$, we find some crossover in the JCE mode and to a lesser degree in the QUICKINDIFF mode. The crossovers do not occur at all in the PEST procedure.

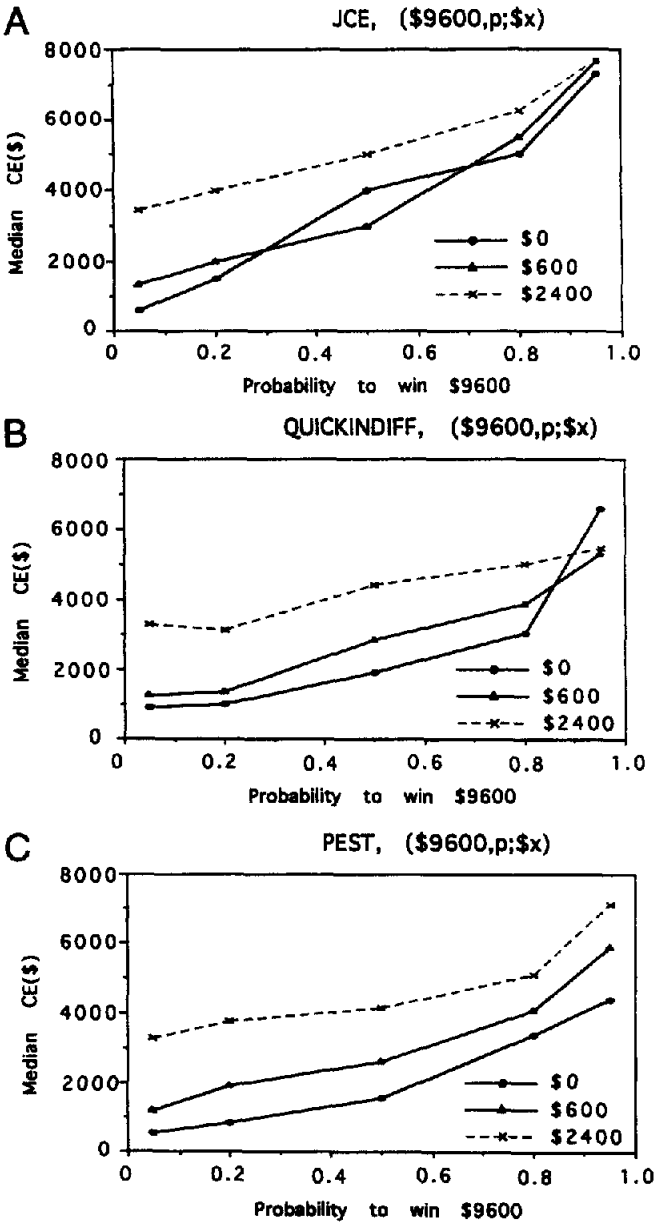

Figure 6. Median certainty equivalents (CEs) for the gambles ($9,600,*p*;$*x*), $*x* being one of $0, $600, or $2,400 ($N = 24$). JCE = judged certainty equivalent; QUICKINDIFF = quick indifference procedure (based on sequential strength-of-preference judgments); PEST = parameter estimation by sequential testing.

The results of Wilcoxon tests are shown in Table 8. The certainty equivalents of any pair of gambles with $p = .95$ were not significantly different for the JCE and QUICKINDIFF procedures. For the PEST procedure, the certainty equivalents for two pairs of gambles at $p = .95$ were significantly different ($p < .01$) but again in the direction of supporting monotonicity. The other results were fairly similar to those of Experiment 1, with all significant results supporting rather than violating monotonicity. As with the data in Experiment 1, we reanalyzed the PEST data by eliminating all first sure things below $600 or $2,400, respectively, for the $0 stimulus. There were no changes in the response pattern.

Discussion

All three response modes in Experiment 3 produced fewer violations of consequence monotonicity than in Experiment 1. For example, violations in the PEST procedure were reduced from 19% to 14%; in the QUICKINDIFF procedure, from 30% to 25%; and in the JCE procedure, from 31% to 17%. The significant reduction of violations in the JCE mode was observed in ($9,600,*p*;$0) vs. ($9,600,*p*; $600) when $p = .05$ or .20. The QUICKINDIFF procedure showed somewhat higher proportions of violations than the PEST and JCE procedures, which were nearly the same. The crossovers did not appear with the PEST procedure but were still present with the QUICKINDIFF and JCE procedures. However, none of the certainty equivalent differences in the direction of violations of monotonicity was significant. Overall, the number of violations of monotonicity was smaller than in the studies by Mellers, Weiss, et al. (1992) and Birnbaum et al. (1992) except for the QUICKINDIFF procedure. Because this procedure exhibited in both Experiments 1 and 3 appreciably more violations than the JCE and PEST procedures, which may have arisen from a tendency to "early exit," it was dropped from Experiment 4.

Experiment 4

One general problem in studying decision making under risk is the relatively large amount of noise in the data—both inconsistencies on repeated presentations within a subject and differences among subjects. Although estimating certainty equivalents of gambles and constructing the preference relation by comparing them within subjects may appear to circumvent this specific problem, the fact is that the estimated certainty equivalents are also quite noisy and subject to similar inconsistencies. These difficulties call into question the precision of inferences we can make from both JCE and PEST data.

In reporting the degree of violations of monotonicity, we have focused, as have others, mainly on results from the within-subject data analysis: For each pair of gambles, we have compared the strict numeric value of the estimated

Table 8
Z Scores for the Wilcoxon Tests of the Certainty Equivalents (CEs) of Gamble Pairs in Experiment 3

Stimulus pair	$p = .05$	$p = .20$	$p = .50$	$p = .80$	$p = .95$
		JCE			
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$600)	−3.29**	−2.63**	−0.42	−1.39	−0.64
CE($9,600,$p$;$600) vs. CE($9,600,$p$;$2,400)	−4.06**	−3.70**	−3.64**	−1.57	−1.48
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$2,400)	−4.26**	−4.16**	−3.51**	−2.84**	−1.70
		QUICKINDIFF			
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$600)	−2.54*	−1.51	−2.80**	−1.68	−0.57
CE($9,600,$p$;$6) vs. CE($9,600,$p$;$2,400)	−3.80**	−3.34**	−3.14**	−2.72**	−1.49
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$2,400)	−3.11**	−3.60**	−3.49**	−2.82**	−0.69
		PEST			
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$600)	−3.71**	−4.01**	−2.29*	−2.40*	−1.13
CE($9,600,$p$;$600) vs. CE($9,600,$p$;$2,400)	−4.26**	−4.26**	−4.11**	−3.57**	−2.86**
CE($9,600,$p$;$0) vs. CE($9,600,$p$;$2,400)	−4.20**	−4.20**	−4.26**	−3.71**	−3.20**

Note. The asterisks indicate significant differences that are in support of the monotonicity assumption. A gamble denoted ($x,$p$;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. JCE = judged certainty equivalent; QUICKINDIFF = quick indifference procedure (based on sequential strength-of-preference judgments); PEST = parameter estimation by sequential testing.
*$p < .05$. **$p < .01$.

certainty equivalents for each individual and reported the proportion of participants who exhibited violations of monotonicity. (Of course, the plots of medians are across subjects.) The three previous experiments showed a substantial proportion of violations (29%–42%) for the gamble pairs consisting of ($96,.95;$0) versus ($96,.95;$6) and versus ($96,.95;$24). However, because the estimated certainty equivalents were quite noisy, comparing the strict numerical value of certainty equivalents may not correctly reveal the underlying trend, especially when the difference of expected values of two gambles is small, as it obviously is in these cases. When the median certainty equivalents of two gambles were compared as a function of probability to win $96 (Figures 4–6), there were no consistent crossovers, and the Wilcoxon tests showed very few significant differences of the two certainty equivalents across subjects.

Thus, we need some means of estimating and taking into account the noise in individual certainty equivalents. This can be done in two ways. First, one can obtain repeated measures of certainty equivalents for each gamble for each individual and apply some statistical analysis. Second, in comparing the observed certainty equivalents, one can devise a noise model and estimate the proportion of violations to be expected under that model and compare those predictions with the observed violations. Because the first approach is fairly time-consuming, especially using the PEST procedure, and is subject to other confounding effects, such as the development of response strategies, we attempted the second approach.

Statistical Model of Noise

To this end, we somewhat mimicked a method used by Slovic and Lichtenstein (1968) in studying the importance of outcome variances. They compared the certainty equivalents of two gambles jointly received—called duplex gambles—with the certainty equivalent of a single first-order gamble that was the convolution of the two. For that purpose they used the test–retest certainty equivalent differences to establish a benchmark of statistical variability. So, we conducted a fourth experiment in which each certainty equivalence was estimated twice, at least 1 week apart, for each gamble, and we used these data to construct a simple model to estimate the magnitude of noise-generated violations of monotonicity.

For any pair of test gambles g and h where h dominates g, let $EV(g)$ and $EV(h)$ denote their expected values, $CE_1(g)$ and $CE_1(h)$ their certainty equivalents estimated from the first session, and $CE_2(g)$ and $CE_1(h)$ those estimated from the second session. Because h dominates g, $EV(h) > EV(g)$. A violation of monotonicity occurs when the observed certainty equivalents satisfy $CE(h) < CE(g)$. The question is whether the proportion of these violations is to be expected

under the null hypothesis that consequence monotonicity holds but while the data are as noisy as they are.

The goal is to estimate the distribution corresponding to that null hypothesis. In doing so, we encounter three major issues. First, what family of distributions should we assume? The distributions over the certainty equivalents are almost certainty not normal for these gambles. For example, the distribution of the certainty equivalent of (\$96,.95,\$$x$) with $x \ll$ \$96 probably has a mean of about \$80 to \$90, is truncated to the right by \$96, and has a long left-hand tail. However, we are not interested in this distribution, but in the difference distribution of two such similarly skewed distributions. Assuming two such distributions, skewed as described, we simulated the difference distributions and found them to be approximately symmetrical and very near to normal. We therefore assume that the noise distribution of the certainty equivalent difference is normal.

Second, what do we take to be the mean difference in certainty equivalents under the null hypothesis of consequence monotonicity? One approach is to assume that the participants are close to expected-value maximizers and thus that the mean certainty equivalent difference is close to the difference in expected values. However, for most participants, the observed certainty equivalents were substantially less than the expected value. So an alternate assumption is that they are expected-utility maximizers with a sharply risk-averse utility function. In this case, the certainty equivalence difference is roughly three times the expected value difference. We therefore conducted two noise analyses: one using as the mean the difference of the two expected values and the other using as the mean three times that.

Third, what do we take as the standard deviation of the difference? Here we estimate it from the observed differences in test–retest values of the same gamble, $\Delta_{1,2}(g,g) = CE_1(g) - CE_2(g)$ and $\Delta_{1,2}(h,h) = CE_1(h) - CE_2(h)$, and then use that estimate to predict the proportion of noise-induced violations.

There are two complications in estimating the variance. First, as we shall see, the CE_1s are systematically somewhat larger than the CE_2s. That being so, plus some concern about memory effects within a single session, we therefore pooled all of the differences $\Delta_{2,1}(g,g) = -\Delta_{1,2} = CE_2(g) - CE_1(g)$ and $\Delta_{1,2}(h,h) = CE_1(h) - CE_2(h)$ and calculated the usual variance estimate for these pooled observations. Second, the judgment data exhibited a considerable number of tied responses. These ties seemed to occur because participants often restricted their responses to either the highest consequence (\$96) or to nearby multiples of 5 (e.g., \$95 or \$90). Because any simple continuous noise model does not allow ties to occur, we need to estimate the proportions of ties and exclude them in the analysis of noise.

To predict the fraction q of ties, we assumed that the proportion of ties between $CE(g)$ and $CE(h)$ would be equal to the pooled proportion of ties in the test–retest certainty equivalents of each gamble; that is,

$$\tfrac{1}{2}[\text{ties of}[CE_1(g) \text{ and } CE_2(g)] + \text{ties of}[CE_1(h) \text{ and } CE_2(h)]].$$

Because of differences between the JCE and PEST procedures, we applied different criteria in predicting the ties in test and retest certainty equivalents. For the JCE method, *equal* was defined to be exactly equal dollar amounts. For the PEST procedure, *equal* was defined to be an interval based on the termination rule that we imposed in estimating the certainty equivalent of each gamble. The reason for doing this is that it is impossible, given the termination rule of the PEST procedure, to tell the actual sign of the difference closer than within this interval. Specifically, let R_g and R_h denote the ranges of the test gambles g and h, respectively. Then, we treated any difference $|CE_1(g) - CE_2(g)| < R_g/50$ and $|CE_1(h) - CE_2(h)| < R_h/50$ as a tie.

The obtained proportions of ties between $CE(g)$ and $CE(h)$ were computed as the pooled proportion of ties in the test and retest certainty equivalents:

$$\tfrac{1}{2}[\text{ties of}[CE_1(g) \text{ and } CE_2(h)] + \text{ties of}[CE_1(h) \text{ and } CE_2(g)]].$$

Again, we necessarily applied a different criterion for the two response modes. As above, for the JCE mode, *equal* was defined to be exactly equal dollar amounts. For the PEST mode, a *tie* was defined in terms of the following interval:

$$|CE_1(g) - CE_2(h)| \text{ or } |CE_1(h) - CE_2(g)| < \tfrac{1}{2}(R_g + R_h)/50.$$

Excluding the ties from consideration, we assume for each gamble pair and each participant that the differences $\Delta_{1,2}(h,g) = CE_1(h) - CE_2(g)$ and $\Delta_{1,2} = CE_1(h) - CE_2(g)$ are both distributed normally with mean $EV(h) - EV(g)$ (or $3 \times [EV(h) - EV(g)]$) and variance σ^2 estimated as the variance of the pooled across-subject values of the two test–retest differences $\Delta_{2,1}(g,g)$ and $\Delta_{1,2}(h,h)$. The predicted proportion of violations of monotonicity, namely $Pr[CE(h) - CE(g) < 0]$, was calculated based on these assumptions for each participant and each gamble pair, and these proportions were then averaged over the participants.

If we let q denote the estimated proportion of ties, then the predicted proportion of monotonicity violations, excluding the predicted ties from consideration, is $(1 - q)Pr[CE(h) - CE(g) < 0]$.

Method

Participants. Thirty-two undergraduate students from the University of Southern California participated in this experiment. They were paid \$7.50 per hour for their service.

Stimulus presentation and response modes. The stimuli were nine of the test stimuli used in Experiment 2 (see Table 9) plus 20 filler items for the PEST procedure. All stimuli were presented on a computer monitor that generated stimulus displays as described in Experiment 1 (see Figure 3); they were presented in random order.

Procedure. Except for omitting the selection of 10 gambles and playing them for money, we followed the same procedure as in Experiment 1 for both the JCE and PEST procedures. Their order was counterbalanced across participants.

The experiment was run in two sessions, each lasting about 45 min, that were separated by at least 1 week. Each session was the same except for independent randomization.

Table 9
Descriptive Statistics of the Test–Retest Certainty Equivalents (CEs) in Experiment 4 (N = 32)

Gamble	EV	Mean CEs		Median CEs		d		t(31)
		Time 1	Time 2	Time 1	Time 2	M	SD	
				JCE				
($96,.95;$0)	$91.2	$74.0	$72.8	$80.0	$80.0	1	13	0.50
($96,.95;$6)	$91.5	$70.3	$72.3	$77.5	$80.0	−2	17	−0.65
($96,.95;$24)	$91.4	$72.3	$74.0	$75.0	$80.0	−2	19	−0.52
($480,.95;$0)	$456	$361	$349	$400	$350	13	87	0.82
($480,.95;$30)	$458	$363	$343	$400	$350	20	94	1.21
($480,.95;$120)	$462	$369	$358	$390	$360	11	72	0.88
($960,.95;$0)	$912	$752	$720	$800	$750	32	175	1.03
($960,.95;$60)	$915	$750	$695	$812	$700	55	167	1.87
($960,.95;$240)	$924	$750	$725	$800	$738	25	183	0.76
				PEST				
($96,.95;$0)	$91	$72	$71	$79.0	$80.5	1	11	0.62
($96,.95;$6)	$92	$74	$72	$80.5	$80.0	2	14	0.77
($96,.95;$24)	$92	$76	$73	$85.8	$79.5	3	13	1.15
($480,.95;$0)	$456	$343	$325	$377	$371	17	95	1.03
($480,.95;$30)	$458	$356	$333	$396	$370	23	101	1.30
($480,.95;$120)	$462	$373	$351	$394	$386	22	86	1.43
($960,.95;$0)	$912	$685	$609	$757	$653	76	251	1.71
($960,.95;$60)	$915	$690	$652	$801	$787	38	210	1.02
($960,.95;$240)	$924	$740	$685	$829	$729	56	209	1.51

Note. A gamble denoted ($x,p;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. EV = expected value; d denotes test–retest difference score; JCE = judged certainty equivalent; PEST = parameter estimation by sequential testing.

Results

Table 9 summarizes statistics of the test–retest assessments (expected values, the means and medians of the certainty equivalents, and the means and standard deviations of the differences of certainty equivalents over the two sessions). Note that the estimated certainty equivalents are typically smaller than the corresponding expected values, and as was noted earlier, the differences in expected value for the crucial monotonicity tests are very small. Overall, both the mean and median certainty equivalents are slightly higher in the first test than in the second one, although none of the differences in each gamble pair is significant. The certainty equivalents assessed using the JCE procedure are similar to those assessed by the PEST procedure. The Wilcoxon test (Table 10) showed no significant violation of consequence monotonicity for the JCE procedure and only one significant violation for the PEST procedure.

Table 11 shows the predicted proportion of ties estimated from the test–retest data, the predicted violations and nonviolations estimated from the noise model, and the corresponding observed proportions for the JCE method. In this table, the mean of the noise distribution was assumed to be the difference between the expected values of the two gambles. The expected and observed proportions are very close. The chi-square values testing differences between observed and predicted violations, nonviolations, and ties are not significant for any gamble. Table 12 shows the same results under the assumption that the mean of the noise distribution is three times as large as the expected value difference. As expected, the predicted proportions of violations are reduced and those of nonviolations correspondingly increased. Nonetheless, the results vis-à-vis consequence monotonicity are essentially identical to those of Table 11.

Tables 13 and 14 show the same analyses for the PEST procedure. The proportion of ties is somewhat larger here

Table 10
Z Scores for the Wilcoxon Tests of the Two Certainty Equivalents (CEs) in Experiment 4

Stimulus pair	z score	
	JCE	PEST
CE($96,.95;$0) vs. CE($96,.95;$6)	−1.73	−2.95**
CE($96,.95;$6) vs. CE($96,.95;$24)	−0.21	−1.51
CE($96,.95;$0) vs. CE($96,.95;$24)	−0.72	−2.92**
CE($480,.95;$0) vs. CE($480,.95;$30)	−0.73	−2.48*
CE($480,.95;$30) vs. CE($480,.95;$120)	−0.53	−2.25†
CE($480,.95;$0) vs. CE($480,.95;$120)	−1.03	−2.99**
CE($960,.95;$0) vs. CE($960,.95;$60)	−0.24	−1.36
CE($960,.95;$60) vs. CE($960,.95;$240)	−0.80	−1.89
CE($960,.95;$0) vs. CE($960,.95;$240)	−0.59	−2.52*

Note. The asterisks indicate significant differences that are in support of the monotonicity assumption. The dagger indicates a significant difference in violation of the monotonicity assumption. A gamble denoted ($x,p;$y) means the following: Obtain $x with probability p; otherwise $y with probability $1 - p$. JCE = judged certainty equivalent; PEST = parameter estimation by sequential testing.
*,†$p < .05$. **$p < .01$.

Table 11
Expected and Observed Percentages of Nonviolation, Violation, and Ties in Monotonicity Tests for the JCE Task Using [EV(h) − EV(g)] as the Mean of the Noise Distribution (N = 32)

Gamble pair (*g* vs. *h*)	Nonviolation		Violation		Tie		$\chi^2(2)$
	Expected	Observed	Expected	Observed	Expected	Observed	
($96,.95;$0) vs. ($96,.95;$6)	43	38	42	48	16	14	0.65
($96,.95;$6) vs. ($96,.95;$24)	46	50	42	36	13	14	0.48
($96,.95;$0) vs. ($96,.95;$24)	46	45	41	39	13	16	0.29
($480,.95;$0) vs. ($480,.95;$30)	44	39	43	45	13	16	0.48
($480,.95;$30) vs. ($480,.95;$120)	44	47	40	38	16	16	0.13
($480,.95;$0) vs. ($480,.95;$120)	45	52	40	39	16	9	1.14
($960,.95;$0) vs. ($960,.95;$60)	45	38	44	53	11	9	1.11
($960,.95;$60) vs. ($960,.95;$240)	47	56	43	39	9	5	1.45
($960,.95;$0) vs. ($960,.95;$240)	47	50	42	38	11	13	0.30
Average	45	46	42	42	13	12	

Note. A chi-square of 5.99 is significant at the .05 level. A gamble denoted ($*x*,*p*;$*y*) means the following: Obtain $*x* with probability *p*; otherwise $*y* with probability 1 − *p*. JCE = judged certainty equivalent.

than for the JCE procedure, which is due to the less stringent definition of a tie. However, the overall conclusion is the same: The noise model provides an excellent fit to the observed proportions of nonviolations, violations, and ties.

Because we elicited certainty equivalents twice for each gamble, the participants had two opportunities to violate monotonicity. Table 15 tabulates the observed proportion of 0, 1, and 2 violations for the two procedures. The most striking result of this tabulation is how rarely participants violated monotonicity twice, especially with the PEST procedure.

Discussion

Experiment 4 yielded violations of monotonicity in about the same proportions as we found in our earlier experiments. For example, there were 37% violations for the JCE procedure and 34% for the PEST procedure at $p = .95$ in Experiment 1 versus 42% and 31%, respectively, in Experiment 4. Again, both JCE values are somewhat smaller than those reported by Mellers, Weiss, et al. (1992) and Birnbaum et al. (1992). We still are not sure why this is so, although a speculation is offered below. A noise model based on test–retest certainty equivalent estimates for the same gambles predicted 12% ties for the JCE procedure and 20% for the PEST procedure. If we exclude ties from consideration, the observed proportions of violations and nonviolations were not significantly different from those predicted on the basis of our noise model. Moreover, participants rarely violated monotonicity twice. Our conclusion is that given the unfortunately noisy character of the data, the observed violations of consequence monotonicity are, to a large extent, consistent with the underlying hypothesis of consequence monotonicity.

Conclusions

Earlier tests of consequence monotonicity—one of the most fundamental assumptions of numerous theories of decision making under uncertainty—provided mixed results. The earliest, based on a conjecture of Allais (1953), was ambiguous because both monotonicity and reduction of compound gambles were involved and it was not clear which was the source of trouble. When two gambles in the

Table 12
Expected and Observed Percentages of Nonviolation, Violation, and Ties in Monotonicity Tests for the JCE Task Using 3 × [EV(h) − EV(g)] as the Mean of the Noise Distribution (N = 32)

Gamble pair (*g* vs. *h*)	Nonviolation		Violation		Tie		$\chi^2(2)$
	Expected	Observed	Expected	Observed	Expected	Observed	
($96,.95;$0) vs. ($96,.95;$6)	44	38	40	48	16	14	0.92
($96,.95;$6) vs. ($96,.95;$24)	49	50	38	36	13	14	0.12
($96,.95;$0) vs. ($96,.95;$24)	52	45	36	39	13	16	0.58
($480,.95;$0) vs. ($480,.95;$30)	45	39	42	45	13	16	0.62
($480,.95;$30) vs. ($480,.95;$120)	48	47	37	38	16	16	0.01
($480,.95;$0) vs. ($480,.95;$120)	50	52	35	39	16	9	1.00
($960,.95;$0) vs. ($960,.95;$60)	46	38	43	53	11	9	1.44
($960,.95;$60) vs. ($960,.95;$240)	51	56	40	39	9	5	0.04
($960,.95;$0) vs. ($960,.95;$240)	52	50	37	38	11	13	0.09
Average	49	46	39	42	13	12	

Note. A chi-square of 5.99 is significant at the .05 level. A gamble denoted ($*x*,*p*;$*y*) means the following: Obtain $*x* with probability *p*; otherwise $*y* with probability 1 − *p*. JCE = judged certainty equivalent; *EV* = expected value.

Table 13
Expected and Observed Percentages of Nonviolation, Violation, and Ties in Monotonicity Tests for the PEST Task Using [EV(h) − EV(g)] as the Mean of the Noise Distribution (N = 32)

Gamble pair (*g* vs. *h*)	Nonviolation		Violation		Tie		$\chi^2(2)$
	Expected	Observed	Expected	Observed	Expected	Observed	
($96,.95;$0) vs. ($96,.95;$6)	37	41	36	33	27	27	0.18
($96,.95;$6) vs. ($96,.95;$24)	39	42	36	28	25	30	0.84
($96,.95;$0) vs. ($96,.95;$24)	40	48	34	28	27	23	1.05
($480,.95;$0) vs. ($480,.95;$30)	43	47	42	39	14	14	0.17
($480,.95;$30) vs. ($480,.95;$120)	43	55	40	34	17	11	1.99
($480,.95;$0) vs. ($480,.95;$120)	46	58	41	34	13	8	1.91
($960,.95;$0) vs. ($960,.95;$60)	36	47	36	31	28	22	1.60
($960,.95;$60) vs. ($960,.95;$240)	36	48	34	27	30	25	2.04
($960,.95;$0) vs. ($960,.95;$240)	40	59	37	22	23	19	5.27
Average	40	49	37	31	23	20	

Note. A chi-square of 5.99 is significant at the .05 level. PEST = parameter estimation by sequential testing. A gamble denoted ($*x*,*p*;$*y*) means the following: Obtain $*x* with probability *p*; otherwise $*y* with probability 1 − *p*. *EV* = expected value.

consequence monotonicity condition were compared directly, the monotonicity assumption was not rejected (Brothers, 1990; Birnbaum & Sutton, 1992; Kahneman & Tversky, 1979). However, when research participants directly judged the certainty equivalents of gambles, and the preference relations of gamble pairs were constructed indirectly by comparing the estimated certainty equivalents, consequence monotonicity appeared to be consistently violated, especially for gambles involving very small probabilities of receiving zero or near-zero outcomes (Birnbaum, 1992; Birnbaum & Sutton, 1992; Birnbaum et al., 1992; Mellers, Weiss, et al., 1992). The present series of four experiments was conducted in an attempt to clarify the relation between the choice and the judged certainty equivalent results.

We conjectured that monotonicity violations may be more prevalent in judged certainty data than in choice data because judged certainty equivalents possibly involve simplifying calculations. We were motivated, then, to test the dependence of violations on response modes. We studied three: judged certainty equivalents, or JCE, and two types of choice-based ones, PEST and QUICKINDIFF. Although the JCE response mode in Experiment 1 produced more violations than the other two for the gambles with low probabilities of receiving $96, this pattern was not replicated in Experiment 3, where the stakes were increased by a factor of 100 and a somewhat more realistic scenario was used. In both experiments, all response modes produced some fraction of violations for the gambles when the probability of receiving the maximum outcome was large. There were no noticeable differences in the proportion of observed violations between the JCE and PEST response modes. When the median estimated certainty equivalents of two gambles were compared as a function of probability to win $96 (or $9,600), there were no consistent crossovers, and the results of Wilcoxon tests did not reject consequence monotonicity between the two certainty equivalents for any of the gamble pairs in Experiments 1 and 3. Indeed, for the PEST procedure, the Wilcoxon tests provided significant support for consequence monotonicity in 10 of 15 tests in Experiment 1 and in 14 of 15 tests in Experiment 3, with all other tests not being significant.

The QUICKINDIFF response mode was an attempt to

Table 14
Expected and Observed Percentages of Nonviolation, Violation, and Ties in Monotonicity Tests for the PEST Task Using 3 × [EV(h) − EV(g)] as the Mean of the Noise Distribution (N = 32)

Gamble pair (*g* vs. *h*)	Nonviolation		Violation		Tie		$\chi^2(2)$
	Expected	Observed	Expected	Observed	Expected	Observed	
($96,.95;$0) vs. ($96,.95;$6)	39	41	35	33	27	27	0.05
($96,.95;$6) vs. ($96,.95;$24)	43	42	32	28	25	30	0.42
($96,.95;$0) vs. ($96,.95;$24)	45	48	28	28	27	23	0.19
($480,.95;$0) vs. ($480,.95;$30)	45	47	41	39	14	14	0.08
($480,.95;$30) vs. ($480,.95;$120)	46	55	37	34	17	11	1.28
($480,.95;$0) vs. ($480,.95;$120)	51	58	37	34	13	8	0.94
($960,.95;$0) vs. ($960,.95;$60)	37	47	35	31	28	22	1.40
($960,.95;$60) vs. ($960,.95;$240)	39	48	32	27	30	25	1.26
($960,.95;$0) vs. ($960,.95;$240)	43	59	34	22	23	19	3.58
Average	43	49	35	31	23	20	

Note. A chi-square of 5.99 is significant at the .05 level. A gamble denoted ($*x*,*p*;$*y*) means the following: Obtain $*x* with probability *p*; otherwise $*y* with probability 1 − *p*. PEST = parameter estimation by sequential testing. *EV* = expected value.

Table 15
Observed Distribution of 0, 1, and 2 Violations of the Monotonicity Assumption (N = 32)

Gamble pair	Observed violation count 0	1	2
JCE			
($96,.95;$0) vs. ($96,.95;$6)	5	23	4
($96,.95;$6) vs. ($96,.95;$24)	11	19	2
($96,.95;$0) vs. ($96,.95;$24)	11	17	4
($480,.95;$0) vs. ($480,.95;$30)	7	23	2
($480,.95;$30) vs. ($480,.95;$120)	13	15	4
($480,.95;$0) vs. ($480,.95;$120)	10	19	3
($960,.95;$0) vs. ($960,.95;$60)	5	22	5
($960,.95;$60) vs. ($960,.95;$240)	10	20	2
($960,.95;$0) vs. ($960,.95;$240)	10	20	2
Average	9	20	3
PEST			
($96,.95;$0) vs. ($96,.95;$6)	9	23	0
($96,.95;$6) vs. ($96,.95;$24)	12	19	1
($96,.95;$0) vs. ($96,.95;$24)	14	16	2
($480,.95;$0) vs. ($480,.95;$30)	4	28	0
($480,.95;$30) vs. ($480,.95;$120)	9	22	1
($480,.95;$0) vs. ($480,.95;$120)	10	21	1
($960,.95;$0) vs. ($960,.95;$60)	10	21	1
($960,.95;$60) vs. ($960,.95;$240)	14	17	1
($960,.95;$0) vs. ($960,.95;$240)	15	17	0
Average	11	20	1

Note. A gamble denoted ($x,p;$y) means: Obtain $x with probability p; otherwise $y with probability $1 - p$. JCE = judged certainty equivalent; PEST = parameter estimation by sequential testing.

develop a faster procedure than that provided by PEST for estimating the choice-induced certainty equivalents. However, QUICKINDIFF exhibited appreciably more violations of consequence monotonicity than the other two procedures. Perhaps this is because it easily allows participants to exit the program before actual indifference is established. We cannot recommend its use as now formulated, but further investigation of procedures faster than PEST is needed.

Another fast technique that is sometimes used (Birnbaum et al., 1992; Tversky & Kahneman, 1992) to estimate certainty equivalents involves participants' marking which of a list of sure consequences are better than a gamble. Although this is nominally choice based, we are suspicious that it is a judgment procedure in disguise.

Unexpectedly, in both Experiments 1 and 3 under the JCE procedure a smaller proportion of violations was exhibited than was reported in the earlier studies by Birnbaum et al. (1992) and Mellers, Weiss, et al. (1992). To investigate the possibility that our using a computer monitor presentation rather than a booklet of drawings of pie diagrams might have been the source of this difference, we conducted Experiment 2 to compare the two JCE presentations. We again found lower proportions of violations in both presentations than in the earlier studies. In the booklet task, the Wilcoxon tests showed no significant violations of monotonicity for any of the gamble pairs that were suspected to produce violations. In the computer task, two of the nine tests showed significant violations of monotonicity. So, the question remains: Why are we getting smaller, nonsignificant proportions of violations? One possibility, which we did not investigate, is that the earlier data were collected in classroom situations whereas ours were in individual laboratory settings. Perhaps the personal attention given to our participants to ensure that they understood what was to be done may have made them more careful.

To test whether the observed proportions of violations of both JCE and PEST response modes really can be explained simply by the (apparently inherent) noise in estimating certainty equivalents, we proposed a noise model utilizing the test–retest data collected in Experiment 4. When ties were excluded from consideration, the predicted proportions of violations from the noise model were not significantly different from those observed in both the JCE and PEST procedures. In addition, when participants had two opportunities to violate consequence monotonicity with the same gamble pair, they very rarely did so. Our conclusion, therefore, is that when the noisy nature of the data are taken into account, we cannot reject the hypothesis of consequence monotonicity.

Although we were unable to reject consequence monotonicity for either the judged or the PEST-based certainty equivalents using our noise model, that model has at least two rather severe limitations. First, we used test–retest data to estimate the number of ties to be added to the continuous noise model. These same data were also used to estimate the variance of the underlying noise distribution across subjects. The ideal way to analyze the noise underlying our estimates of certainty equivalents would be to repeatedly measure it and to apply statistical tests for each individual. Although ideal, it is impractical in this context, especially the PEST one, to obtain a sufficiently large sample of individual test–retest data. One likely difficulty in pursuing this approach lies in how to reduce memory effects without using many filler gambles. In the PEST procedure, it is impossible to test a large set of gambles in one or two successive sessions.

Second, the certainty equivalents estimated in the retest session, which was performed at least 1 week later than the test session, were consistently less than those of the test ones. It would be interesting to obtain the test–retest certainty equivalents in the same session or, if that is not feasible, in sessions separated by less time than a week.

Despite a fairly large proportion of observed violations of consequence monotonicity, which seems to be largely due to the level of noise in estimating them, we conclude from the several analyses that consequence monotonicity cannot be rejected. This is true for both judged certainty equivalents and PEST-determined certainty equivalents. The fact that there is little difference in the proportions of violations of consequence monotonicity between the two procedures and the greater ease of collecting JCE judgments seem to recommend the JCE procedure over the PEST procedure. Nonetheless, we urge caution in using judged certainty equivalents to test choice-based theories of decision making. For one thing, the median data slightly favor the PEST

procedure for reducing violations of consequence monotonicity; for a second, the judged estimates have a serious restriction in the numbers used, and so many ties occur; and for a third, we know from the preference-reversal experiments that judged certainty equivalents do not always agree with choices.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Americaine. *Econometrica, 21,* 503–546.

Allais, M., & Hagen, O. (Eds.). (1979). *Expected utility hypothesis and the Allais' paradox.* Dordrecht, The Netherlands: Reidel.

Birnbaum, M. H. (1992). Violations of monotonicity and contextual effects in choice-based certainty equivalents. *Psychological Science, 3,* 310–314.

Birnbaum, M. H., Coffey, G., Mellers, B. A., & Weiss, R. (1992). Utility measurement: Configural-weight theory and the judge's point of view. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 331–346.

Birnbaum, M. H., & Sutton, S. E. (1992). Scale convergence and utility measurement. *Organizational Behavior and Human Decision Processes, 52,* 183–215.

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior and Organization, 13,* 193–212.

Brothers, A. J. (1990). *An empirical investigation of some properties that are relevant to generalized expected-utility theory.* Unpublished doctoral dissertation, University of California, Irvine.

Chew, S. H., Karni, E., & Safra, Z. (1987). Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic Theory, 42,* 370–381.

Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics, 16,* 65–88.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69,* 623–638.

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

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions, *Journal of Experimental Psychology, 89,* 46–55.

Luce, R. D. (1988). Rank-dependent, subjective expected-utility representations. *Journal of Risk and Uncertainty, 1,* 305–332.

Luce, R. D. (1991). Rank- and sign-dependent linear utility models for binary gambles. *Journal of Economic Theory, 53,* 75–100.

Luce, R. D. (1992). Where does subjective-expected utility fail descriptively? *Journal of Risk and Uncertainty, 5,* 5–27.

Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first-order gambles. *Journal of Risk and Uncertainty, 5,* 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.

Mellers, B. A., Chang, S., Birnbaum, M. H., & Ordóñez, L. D. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 347–361.

Mellers, B., 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 Organizations, 3,* 324–343.

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

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Schmeidler, D. (1986). Subjective probability and expected utility without additivity. *Econometrica, 57,* 571–587.

Segal, U. (1987). Some remarks on Quiggin's anticipated utility. *Journal of Economic Behavior and Organization, 8,* 145–154.

Segal, U. (1989). Anticipated utility: A measure representation approach. *Annals of Operations Research, 19,* 359–373.

Slovic, P., & Lichtenstein, S. (1968). Importance of variance preferences in gambling decision. *Journal of Experimental Psychology, 78,* 646–654.

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *American Economic Review, 73,* 596–605.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 204–217.

Tversky, A., Sattah, S., & Slovic, P. (1988). Contingent weighting judgment and choice. *Psychological Review, 95,* 371–384.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review, 80,* 204–217.

Wakker, P. P. (1989). *Additive representations of preferences: A new foundation of decision analysis.* Dordrecht, The Netherlands: Kluwer Academic.

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

Yaari, M. E. (1987). The dual theory of choice under risk, *Econometrica, 55,* 95–115.

Received April 7, 1993
Revision received July 29, 1996
Accepted July 29, 1996 ■