Empirical Economics (1994) 19:223–253

EMPIRICAL ECONOMICS

Expected Utility Theory and the Experimentalists

GLENN W. HARRISON[1]

Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC29208, USA

Abstract: The experimental evidence against expected utility theory is, on balance, either uninformative or unconvincing. When one modifies the experiments to mitigate these criticisms the evidence tends to support traditional theory.

JEL-Classification System-Numbers: D81, C91, B41

Expected Utility Theory has been under severe attack in recent years. The source of this attack has been the observation of apparently robust behavioral anomalies in decisions that individual experimental subjects make in controlled environments. The popular implication of these observations is that Expected Utility Theory (EUT) is systematically misleading as a descriptive model of human behavior under uncertainty. Many alternative models of individual choice behavior have been proposed that can account for some or all of the alleged anomalies. We argue that the experiments in question do not meet widely accepted sufficient conditions for a valid controlled experiment proposed by Smith [1982; pp. 930 ff.]. Moreover, when some of those experiments are modified to satisfy those conditions the anomalies in question either vanish or significantly diminish.

Four conditions are proposed by Smith [1982]. The first is *Nonsatiation* in the reward medium v: utility $U(v)$ is a monotone increasing function of v. Thus we can expect that a choice that generates v' will be preferred to a choice that generates v'' whenever $v' > v''$.

The second condition is *Saliency* of the reward medium: there is a mapping, perhaps only implicit in the rules of the experiment, between rewards and the messages m of the institution under study. Thus we can expect that a message (e.g., a bid, or a stated preference) m' that maps into v' will be observed, instead of a message m'' that maps into v'', whenever $v' > v''$.

The third conditions is *Dominance* of the rewards over the subjective costs (or benefits) to a subject from participating in the experimental task. In practice this

[1] Dewey H. Johnson Professor of Economics, Department of Economics, College of Business Administration, University of South Carolina. I am grateful for comments from seminar participants at the University of Melbourne, University of South Carolina, University of Stockholm, and the University of Western Ontario. John Hey provided a firm, but sympathetic, editorial hand.

0377-7332/94/2/223-253 $2.50

condition requires that the rewards corresponding to the null hypothesis are "perceptibly and motivationally greater" than the rewards corresponding to the alternative hypothesis. Assume for the moment that each hypothesis refers to a specific message: m_o for the null and m_a for the alternative. In this case we require that the rewards associated with these two messages, v_o and v_a respectively, be such that $v_0 > v_a + \delta$, where δ is the subjective cost to the agent from sending m_o rather than m_a.[2]

Figure 1 illustrates these concepts for two experimental designs, *A* and *B*. The payoff function for each experiment is drawn so that they each have the same payoff v_o for the same message m_o. However, they differ greatly with respect to the payoff under the alternative hypothesis, message m_a. For the value of δ shown Experiment *A* fails to satisfy the Dominance precept whereas experiment *B* does satisfy it. The implication is that one should avoid an experimental design such as *A* in favor of one such as *B* whenever possible.

Experimentalists do not have any objective notion of the value of δ for a wide range of tasks, and it is unlikely that any "objective" measures of δ for a given subject pool in a given task environment are likely to gain wide acceptance. Nonetheless, it is always possible to report the *design* of a given experiment *conditional on a range of values for* δ that can be widely accepted as plausible. Assume that $\hat{\delta}$ is "the" value of δ that is accepted ($\hat{\delta} > 0$). Then any experimental design such that $v_0 \leq v_a + \hat{\delta}$ will be said to fail the Dominance requirement conditional on $\hat{\delta}$ as the assumed perceptive and motivational threshold. Much like Good Bayesians[3] report posterior-based inferences for a wide class of

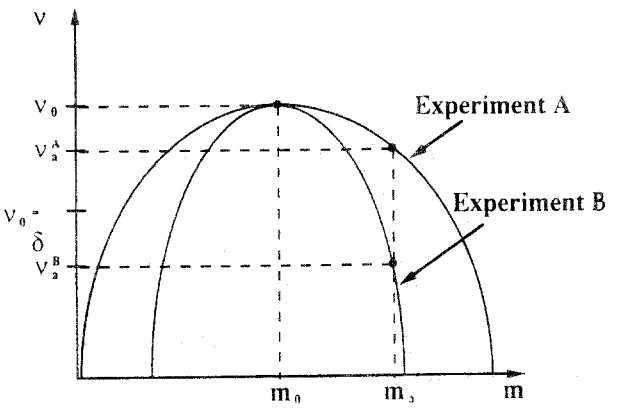

Fig. 1. The payoff dominance problem

[2] To be accurate one should probably write δ as a function of m_o and m_a, as well as noting that it will be agent-specific.

[3] Such as Leamer and Leonard [1983].

priors, so should Good Experimentalists report the range of values for δ such that their experimental design ensures that $v_0 > v_a + \delta$ for given v_o and v_a. In such cases the reader of the study can draw his own inferences from the data as a function of his own priors over $\hat{\delta}$.

An important practical extension of the concept of Dominance is required when one or more of the hypotheses is not simple. Assume that we have a simple point-null hypothesis (e.g., $m = m_o$) and a composite alternative hypothesis (e.g., $m \neq m_o$). In this case there typically exists, for a given δ, a set of messages $\overline{m}$ arbitrarily close to m_o, such that $v(\overline{m}) \geq v_0(m_0) - \delta$. Conversely, there exists a set of messages $\hat{m}$ such that $v(\hat{m}) < v_0(m_0) - \delta$. We may then say that no observations in the set $\overline{m}$ can be claimed to satisfy the Dominance requirement; one can only make such claims about observations in the set $\hat{m}$. Figure 2 illustrates these concepts. Harrison [1989] [1992] demonstrates that many experiments have been inadvertently designed such that virtually all of the observations fall in the set $\overline{m}$, conditional on plausibly small values of δ. Thus one must nihilistically insist that the subjects have a sufficiently low threshold δ, perhaps even claiming $\delta = 0$, in order to maintain the conclusion that such observations allow one to reject the null hypothesis.[4]

We argue that many of the experimental anomalies that are claimed to violate EUT do not satisfy the Saliency requirement or, if they do, generally fail to satisfy the Dominance requirement for plausible (perceptual or motivational) threshold values. The anomalies examined here that *do* survive the Dominance requirement turn out to be the clear exception rather than the "rule". On

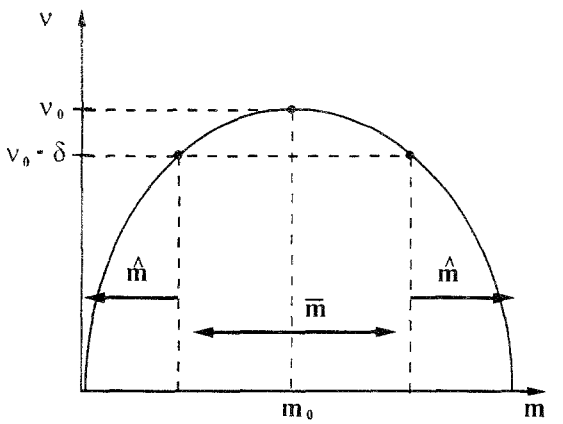

Fig. 2. Region of payoff dominant messages

[4] If δ is strictly positive then there *always* exists *some* sub-optimal message that fails the Dominance requirement. In practice, however, the discrete nature of individual behavioral responses will often provide a natural solution to this nihilist's dilemma.

balance, then, the experimental evidence against EUT evaluated here is either uninformative or unconvincing.

It is appropriate to focus on four groups of experiments that have played what I would argue are an infamous historical role in the critique of EUT and the received theory of choice under uncertainty.[5] The first is the Allais Paradox experiment devised by Allais [1953]. The second set is the Preference Reversal experiment developed by Grether and Plott [1979]. The third set is a group of simple choice experiments underlying the heuristics that are rationalized by the Prospect Theory of Kahneman and Tversky [1979]. The final set is the test of Bayes Rule devised by Grether [1981]; although not formally a component of EUT, Bayes Rule lies at the core of most interesting economic models of behavior under uncertainty.

1 The Allais Paradox

1.1 The Experiment

Consider the "probability triangle" in Figure 3 defined over the probabilities of three monetary prizes: a low prize P_L, an intermediate prize P_I, and a high prize P_H, with $P_L < P_I < P_H$. Let π_L, π_I and π_H be the probabilities that each of P_L, P_I and P_H will eventuate in a given lottery. Because $\pi_L + \pi_I + \pi_H = 1$, we can express π_I implicitly in terms of π_L and π_H (i.e., given values of π_L and π_H, π_I is uniquely determined).

The Allais Paradox, or AP, can be conveniently represented in a probability triangle.[6] Consider four separate lotteries shown in Figure 3 as A, A^*, B and B^*. Note that A is related to A^* in the same way as B is related to B^*: just increase π_H by a certain amount (e.g., 0.10) and increase π_L by a certain amount (e.g., 0.01), thereby implicitly decreasing π_I by a certain amount (e.g., 0.11). From the perspective of EUT the crucial feature of this configuration is that the chords AA^* and BB^* are parallel straight lines.

---

[5] Many other experiments have tested aspects of EUT, but none have been quite as influential as the three considered here. Harrison [1989] [1992] and Drago and Heywood [1989] consider a wide range of experiments and argue that they suffer from the payoff Dominance problem in one way or another. In many cases a cursory reading of the experimental procedures of an experiment are enough for one to decide that the decisions could not possibly satisfy the Dominance criteria. For example, Camerer [1989] uses a random-selection rule in which 1-in-N subjects are actually paid. As N becomes large a lack of Dominance is assured. (Starmer and Sugden [1990] study the effectiveness of this random-selection procedure, with mixed results: see Davis and Holt [1993; p. 452–455] for further discussion).

[6] Davis and Holt [1993; p. 438–442] provide a nice exposition of this pedagogic device.

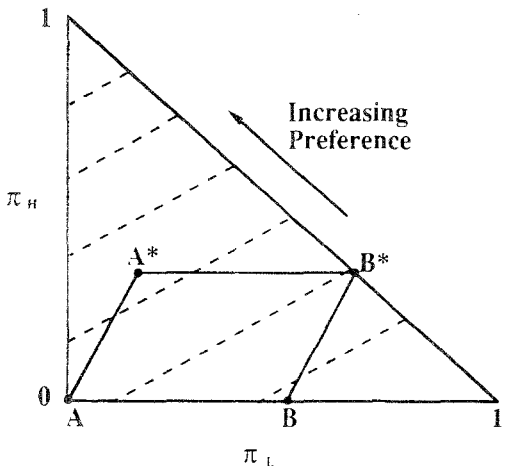

Fig. 3. The Allais paradox

EUT assumes that the expected utility of any bundle of uncertain prizes is equal to the utility of each prize weighted simply by the probability of that prize. Thus the indifference curves of a non-satiated agent are straight lines, as shown by dashed lines in Figure 3. The agent whose tastes are depicted in Figure 3 would choose A^* over A and B^* over B when asked to make these two bilateral choices.

An important property of the slopes of the indifference curves of an agent obeying EUT is that they depend on his attitudes to risk (e.g., see Machina [1986; p. 126–7]). A risk-neutral agent will have indifference curves with slope $\tau = (P_I - P_L)/(P_H - P_I)$. A risk-preferring agent will have indifference curves that are relatively flatter than τ. Crucially, a risk-averse (RA) agent will have indifference curves that are steeper than τ. Figure 4 illustrates why this is important: for any AP configuration of lotteries, there exists an attitude to risk such that the agent is indifferent between the lotteries he is asked to choose between.[7] Typically the experiments testing the AP do not allow subjects to report indifference.[8]

With this background, what is the AP? The AP refers to two observed patterns of choice: (1) subjects who choose A over A^* but choose B^* over B, and (2) subjects who choose A^* over A but choose B over B^*. Conlisk [1989] found (in

[7] In Figure 4 we draw the case of a RA agent and the normal AP configuration with the chord AA^* steeper than τ (for simplicity we assume here that $\tau = 1$, although this is not the case for the prizes considered later). Clearly our claim is general. If the chord were flatter than τ, we need only depict a sufficiently risk-loving agent's preferences. It is similarly obvious that a RA agent could well have indifference curves even steeper than the slope of the AA^* chord.

[8] Exceptions include Battalio, Kagel and Jiranyakul [1990] (who observed few subjects reporting indifference) and Hey and DiCagno [1990].

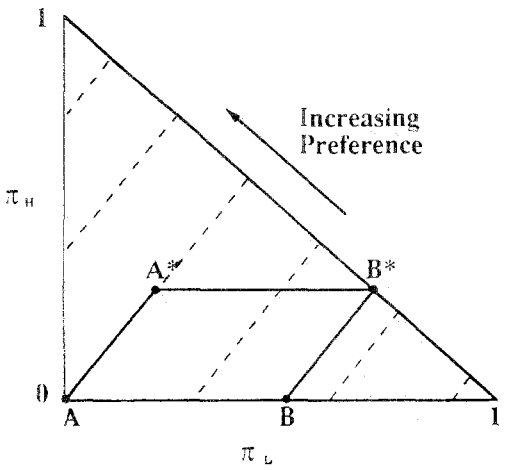

Fig. 4. The Allais paradox with particular risk-neutral preferences

his "Basic Version") that 50.4% of 236 University of California at San Diego economics undergraduates chose either (1) or (2) when their payoffs were hypothetical and the prizes expressed in terms of "millions of dollars". Moreover, the vast majority (86.5%) of those 119 students chose pattern (1), implying that there is a systematic pattern of responses.

In what sense should we call the observed behavior "paradoxical"? For the agent depicted in Figure 3, either of pattern (1) or (2) violates EUT. In general, violations occur true for any agent whose indifference curves do not overlap the chord AA^* (and hence also do not overlap BB^*). What about the agent depicted in Figure 4? Forced by experimental design to make a choice between A and A^*, he would no doubt "flip a coin". Similarly for his forced choice between B and B^*. The result, with a large enough sample of such agents, would match exactly the 50.4% rate of AP behavior observed by Conlisk [1988]. In this, deliberately extreme, case the AP behavior is purely an artifact of the experimenter requiring the subject to report a strict preference. The subject foregoes zero expected income or utility

Convenient as this argument may be for EUT, it is clearly too strong. First, it requires that the agent in question have attitudes to risk that imply indifference curves that precisely match the chord AA^*. Although the degree of risk aversion that might be required in Figure 4 is not implausible, the requirement that it be a specific degree of risk aversion is implausible. Second, we require that all agents in the sample have the same attitude to risk. This further stretches the plausibility of our argument, since AP behavior has been observed in hypothetical questionnaires using a wide range of subjects who plausibly have a wide range of attitudes to risk.

What our argument does identify, especially in the polar form of Figure 4, are two potentially important treatments in the design of AP experiments. One is to

allow subjects to report indifference, given that "as if" AP anomalies could be observed under the null hypothesis of EUT if one does not permit such reports. The other treatment is to control for the risk attitudes of agents, so that we can ensure (under the null hypothesis) that we have a design such as in Figure 3 instead of a design such as in Figure 4.

To control for risk attitudes we adopt the lottery procedure of Roth and Malouf [1979] and Berg, Daley, Dickhaut and O'Brien [1986]. This procedure allows the experimenter to induce "as if" risk-neutral behavior by conducting an experiment in which the rewards are units of probability that a subject will win some monetary prize. A convenient way to operationalize this is to reward subjects with "lottery tickets", with every extra ticket earned increasing the probability that the monetary prize will be won. Irrespective of the subject's attitudes to risk defined over monetary prizes, this procedure induces risk neutrality with respect to the reward medium of the experiment (lottery tickets). Thus we redefine P_L, P_I and P_H in terms of numbers of lottery tickets, and denote the monetary lottery prize M; the subject receives zero cash payoff if he loses the lottery, and $M > 0$ if he wins the lottery.

There has been some debate over the ability of this procedure to induce risk-neutral behavior in certain experimental contexts (e.g., the first-price sealed-bid auctions of Cox, Smith and Walker [1985]). However, Berg, Daley, Dickhaut and O'Brien [1986] have tested the procedure in a binary choice context such as employed here, an they found that the procedure indeed induced the predicted choices.[9] Moreover, it can be shown that this particular test of the lottery procedure does *not* fail the Dominance requirement (see Harrison [1992; p.1430, fn.5]). This point is especially important in terms of our present design objectives: to construct a Salient and Dominant test of the AP.

1.2 Saliency: The Conlisk Experiments

Despite the voluminous literature on the AP, Conlisk [1989] appears to have been the first to adopt salient rewards. He gave 53 subjects, drawn from the same population as his non-salient experiments discussed above, the standard binary choices over the following prospects: *A*: certainty of \$5; *A**: 1/100 chance of \$0, 89/100 chance of \$5, and 10/100 chance of \$25; *B*: 89/100 chance of \$0, and 11/100 chance of \$5; and *B**: 90/100 chance of \$0, and 10/100 chance of \$25. He found that only 3 of his subjects, or 5.7% of his sample, chose prospects that violate EUT. His results are presented in Table 1.

---

[9] As a slight qualification, note that their test compared the choices of "as if" risk-averse agents with "as if" risk-preferring agents. They did not study the behavior of "as if" risk-neutral agents. However, their test does demonstrate that there is nothing in the binary choice context that might invalidate the lottery procedure, which is the issue in debate.

Table 1. Results from selected allais paradox experiments

Experiment	Payoffs	Number of Subjects	Percentage of Outcomes				
			EUT-consistent		EUT violations		
			AB	*A*B**	*A*B*	*AB**	Total
AP0	Hypothetical	20	0	65	0	35	100%
Conlisk	Hypothetical	236	8	7	42	44	100%
AP1	Salient	20	0	85	0	15	100%
Conlisk	Salient	53	0	94	0	6	100%
AP2	Salient & Dominant	20	0	95	0	5	100%

Reassuring as these results might seem for EUT, two doubts remain. First, Conlisk [1989; p. 34] also observed that the AP anomalies disappeared almost completely in a practice round using the same subjects but with hypothetical payoffs. Thus it is not at all obvious that Saliency, as distinct perhaps from the absence of dizziness due to payoffs not having the word *million* attached, is the savior of EUT.

The second doubt concerns the splitting of the sample: about half of the subjects chose between *A* and *A**, while the other half chose between *B* and *B**. The reported results refer to the pooling of the choices of these sub-samples as if they had all been reported by the full sample. Sample size is not the issue here, since the results are so overwhelmingly consistent with one or other hypothesis (happily, EUT). Rather the problem is that one must implicitly assume that the tastes of the two sub-samples are "sufficiently homogeneous" that how one sub-sample behaves on the *AA** choice can be validly compared to how the other sub-sample behaves on the *BB** choice. Requiring an auxiliary assumption such as this biases the design against EUT. The evidence strongly supports EUT, of course, so this is not a decisive problem with the design.

1.3 Saliency and Dominance

Without controlling for the risk attitudes of subjects, and ensuring Dominance, it is not possible to claim that observed AP anomalies are due to a failure of EUT. AP anomalies appear to disappear completely when one merely makes the rewards Salient, as just noted, but it is not obvious that Saliency is the key treatment. Therefore we designed and implemented a new experiment to focus on these issues.[10]

[10] The sample sizes for many of the new experiments reported here are quite small, so the *precise* quantitative results deserve to be interpreted with a grain of salt.

Each subject was asked to make two binary choices. The prospects employed were: *A*: 500 cents with certainty; *A**: 1/100 chance of 0 cents, 89/100 chance of 500 cents, and 10/100 chance of 2000 cents; *B*: 89/100 chance of 0 cents, and 11/100 chance of 500 cents; and *B**: 90/100 chance of 0 cents, and 10/100 chance of 2000 cents. Three experimental conditions were employed. In condition *AP0* all payoffs were hypothetical. In condition *AP1* all payoffs were as stated above. A risk-neutral subject could expect to earn $8.50 in each of these two conditions if he made the choices predicted by EUT.

In condition *AP2* all payoffs were as stated above except that the word "cents" was replaced with "points" and it was explained that every point earned would equal one extra "lottery ticket". Each extra lottery ticket gave the subject an extra 1% probability of winning C$40 in the final lottery. The value $40 was chosen because that implies the same expected value of the experiment to a risk-neutral agent in condition *AP2* as in *AP1*: we wanted to ensure that expected rewards were comparable under the null hypotheses in condition *AP2* (viz., risk-neutral EUT-consistent choices) and condition *AP1* (viz., EUT-consistent choices).[11] Note also that the risk-neutral subject in condition *AP2* foregoes $1.45 in *each* binary choice if he does not obey EUT. The same values apply in condition *AP1* if one makes the auxiliary assumption that the subject is risk-neutral. The entire session lasted about 15 minutes in each condition.[12] Thus we would argue that the incentives that subjects faced satisfied the Dominance requirement.

The results of our experiments are reported in Table 1. We find that all of the observed choices that violate EUT are of one kind: *AB**, in which the prospect with no uncertainty (prospect *A*) seduces subjects to violate EUT. When payoffs are not Salient EUT is violated by 35% of the subjects. When payoffs are Salient, however, the percentage of subjects violating EUT drops to 15% (condition *AP1*) and 5% (condition *AP2*). We therefore observe a very powerful effect simply from ensuring Salient payoffs.

The strength of the effect of Salience makes it difficult to observe any strong effect from adding Dominance: although the drop in EUT violations from 15% to 5% is in the predicted direction, the sample size does not allow us to claim that this is a statistically significant decline.[13]

---

[11] A subject behaving in accord with EUT could expect to earn 850 tickets in condition *AP2*. The maximum number of lottery tickets that could be earned by certain choices and some good luck is 4000. Thus the subject would have an expected probability of $850 \div 4000 = 0.2125$ of winning the monetary prize *M*. For the expected payoff to be $8.50 we therefore set $M = \$8.50 \div 0.2125 = \$40$.

[12] These sessions were conducted after the subjects had participated in first-price sealed-bid auction experiments such as those reported in Harrison [1989]. In each case the previous experiment had employed exactly the same payment condition. Thus the subjects in condition *AP2* were already familiar with the lottery procedure (although it was explained to them afresh). All subjects were drawn from the undergraduate economics student population at the University of Western Ontario.

[13] To see this point, consider a simple χ^2 test of the null hypothesis that the observed choice frequency under *AP2* is the same as the expected choice frequency found in *AP1* (i.e., that we only expect to see an effect from Saliency). We then have $\chi^2 = 1.568$ with 3 degrees of freedom, and are unable to reject the null hypothesis at any standard significance level.

2 Preference Reversals

2.1 The Experiment

Grether and Plott [1979], hereafter GP, design a beautiful experiment to test a phenomenon discovered by psychologists known as "preference reversals". These reversals involve the same subject stating that he prefers one wager to another, and separately reporting "selling prices" for those wagers that imply a different preference. It is as if his preference over the two wagers reverses as we move from one preference-elicitation procedure to another. What is so nice about the GP experiment is that it was designed with economists in mind, and addresses many of the sneaking suspicions that economists have about such results from experimental psychology. What is so disturbing about the GP experiment is that, notwithstanding these design features, preference reversals persist.

The basic experiment of GP (their Experiment 1) consists of three Parts. In Part 1 the subject is asked to state a preference for one wager over another (indifference was allowed). Three pairs of wagers are evaluated in this manner. In Part 3 the subject is likewise asked to state his preference for one of the two wagers in another three pairs. Thus in Parts 1 and 3 we have stated preferences for six distinct pairs of wagers. In Part 2 the subject is asked to state selling prices for each of the 12 wagers involved in Parts 1 and 3. The Becker, DeGroot and Marschak [1964] (BDM) procedure is employed to elicit these selling prices, with buying prices drawn at random from a uniform distribution defined over the interval 0 to 999. This procedure is designed to elicit the certainty-equivalent of an arbitrary risky prospect.[14]

Two methods of financial reward were employed by GP. In the first method 44 subjects received \$7 for simply completing the experiment, irrespective of their decisions. In the second method 46 subjects were given a \$7 credit and then told that at the end of the experiment one of their 18 decisions (6 reported preferences and 12 valuations) would be selected and used to determine if they earn more or less money.[15] It was possible for the subjects to lose up to \$2 if

---

[14] The task is to devise a procedure that gives subjects a financial incentive to truthfully state their certainty-equivalent for a given wager. BDM solve this problem by asking the subject to state a minimum "selling price" for the wager and then randomly choosing a "buying price" for the wager. If the buying price exceeds the stated selling price the subject receives the former price; if not, he gets to play the wager. Providing that this buying price is randomly determined and is independent of the stated selling price at any stage, it can be easily shown that the subject maximizes the expected utility of the reward by setting his selling price equal to this monetary equivalent of the basic wager. Harrison [1992; pp. 1428–9] explains how one evaluates the strength of the financial incentives under this procedure.

[15] Holt [1986a] has shown that this procedure, of selecting one decision to reward subjects, may induce "as if" preference reversal behavior when the Independence Axiom of standard expected

they had some bad luck, leaving them assured of a net take home payoff of at least $5. We will focus here on subjects that were financially motivated to make better decisions.

There were two early attempts to modify the GP design to determine if the reversal phenomenon is robust. The first is by Pommerehne, Schneider and Zweifel [1982]. They argue that the subjects in the GP experiment faced very little incentive to make correct decisions. However, it is not at all clear that they have in mind anything with the word "incentives" that is, or should be, of interest to economists. In order to "create a more stimulating situation" (p. 569) for the subjects they increased the face value of the prizes in the wagers. They did this by denominating the prizes in Swiss Francs, so that a prize of US$3.86 in the GP experiment would now be shown as Fr.386. Clearly increasing the *face value* alone will not help matters: if the exchange rate between Swiss Francs and dollars were 0.01, the prizes would be identical in terms of purchasing power (give or take epsilon due to violations of PPP).[16] Simply denominating the currency units more finely cannot possibly help matters. If it does induce more careful decision making then the subject must be suffering from some "numeraire illusion" which is not good for experimental design or control.[17]

In any case, the subjects did not even receive the stated face value after all. In fact the experiments used "play money", with subject earnings converted into real money at the end of the experiment at a conversion rate that was unknown to them at the time of their decisions. This rate was determined, ex post the subjects' decisions, so as to divide up a fixed total of Fr.2000 between 84 subjects. This implies that each subject would have received around US$13.6 over 12 decisions on average, or US$1.09 per decision on average. We therefore agree with Grether and Plott [1982; p. 575] that "it is not obvious that the incentives were greater" than in the original GP experiments.

The second attempt to modify the design of the GP experiments is by Reilly [1982]. He ran two series of experiments. The first series, his Stage 1, is virtually

---

utility theory is violated. This is a comforting result, since that is arguably the one axiom of expected utility theory that is the least damaging to dispose of (see Machina [1982]). It should be noted that GP use this procedure, following BDM, to avoid "wealth effects" from early earnings or losses from contaminating responses in later decisions. Given that truthful revelation is a strongly dominant strategy in the BDM elicitation procedure, it is not clear why "wealth effects' should matter at all.

[16] As it turns out, the exchange rate was about Fr.1.82 over 1980–81, so the face value of Fr.386 would have been about US $212.

[17] Unfortunately this procedure of increasing the face value of experimental rewards in order to "improve subject motivation" or to "improve statistical inference" is all too common. Many experimentalists denominate their payoffs in "francs" or "pesos" rather than "dollars or cents", without any clear notion as to why this should encourage better decisions: e.g., see Forsythe, Palfrey and Plott [1982; p. 542]. The fact that this design feature has not (yet) been associated with odd experimental results does not justify regarding it as a "successful" design innovation (Plott and Sunder [1982, p. 667]). This argument has nothing to do with the possible benefits of conducting experiments in some currency other than cash: see Roth and Malouf [1979] and Berg, Daley, Dickhaut and O'Brien [1986].

a replication of GP[18] but with slightly different prizes in the basic wagers. He replicated their main results.

Contemplation of subject behavior during Stage 1 led Reilly [1982] to adopt a number of design features in Stage 2 that were "... intended to strengthen monetary incentives and reduce confusion among the subjects" (p. 580). These new design features were: (i) the use of smaller groups of subjects in any one session, to encourage the asking of clarifying questions; (ii) an increase in the running time of the experiment, so as not to press any subjects into unduly hasty decisions; (iii) clarification that the rewards were not hypothetical; (iv) the provision of the expected value of the wager to each subject; and (v) the use of the interval (−400,1400) rather than (0,999) in order to generate buyout prices in the BDM procedure. These changes in design were associated with modest but significant reductions in the extent of observed preference reversals. Nonetheless, the phenomenon persists.

2.2 *Inferences from the Experiment*

As we have noted above, the preference reversal phenomenon discovered in financially motivated subjects by GP has been replicated and found to be robust. One implication is that we should embrace tractable modifications of standard expected utility theory that allow "as if" preference reversals (see Holt [1986a], Karni and Safra [1987], Loomes and Sugden [1983], Machina [1982] [1987] and Segal [1988] for careful arguments along these lines, and Cox and Epstein [1989] for experiment that are inconsistent with one of these modifications). Another implication, far less attractive to economists, is that we must pay theoretical and empirical attention to the way in which valuations or preferences are assumed to be elicited or communicated in decision-making environments.

[18] Actually there is one disparity between the procedures that Reilly [1982] used and those of GP. In conversation Reilly has pointed out that he randomized over the six lottery pairs when deciding on the payoff to subjects. If a given lottery pair was selected for payoff, the subjects direct preferences would be used to determine which lottery of that pair would be played out using the BDM procedure and the subjects reported valuation for that lottery. In conversation Grether confirms that GP actually randomized over the eighteen decisions. In each experiment the subjects were told only that one decision would be used to determine the payoff. We shall assume that subjects interpreted this as implying what was actually done in the GP experiment. Subjects in our replication of GP were explicitly told at the outset that one of their 18 decisions would be selected.

2.3 *The Cost of Misbehavior*

One question that is typically not addressed in the three experimental studies reviewed above is *how* inconsistent are the reported decisions of subjects? Vague statements by Pommerehne, Schneider and Zweifel [1982; p. 569–570] about subjects not being motivated, or sensing small differences in payoffs, can readily be made explicit.

The opportunity cost of an inconsistent report is the expected income that is foregone by not reporting consistently. There are many ways to make consistent reports, and we make the simplest possible assumptions. Specifically, assume that the subject reports his ordinal preferences in Parts 1 and 3 according to his true (consistent) preferences. Then the entire "cost" of any misbehavior is borne in the Part 2 decisions, the selling price reports. Without any difference to the results reported below, assume that the valuations elicited for one type of wager (the "P-bet", in the parlance used in this literature) is in accord with the subject's true preferences. Thus the "false report" that may imply a preference reversal is the valuation elicited for the other wager (the paired "$-bet"). We evaluate the foregone expected income of this false report. If the ordinal ranking implied by the reported valuations is the same as the directly-elicited ordinal ranking of Parts 1 and 3, then no preference reversal is implied and the foregone expected income is of course zero.

It should also be remembered that any single decision by the subjects in these experiments had only a 1-in-18 chance of being used by the experimenter to decide the final payoff. Thus the cost of an inconsistent decision at any stage is to be divided by 18 to obtain the appropriate cost on a "per decision" basis. We use these procedures to evaluate the opportunity cost of inconsistency in the experiments conducted by Reilly [1982] and a replication of the GP experiment.[19]

A *The Reilly Experiments*

In his Stage 2 sessions Reilly [1982] employed 45 subjects in a design in which expected values of each wager were not provided (his Group 1). Pooling over all 265 cases in which an ordinal preference was expressed in Parts 1 and 3, he reports (Table 2, p. 581) a preference reversal frequency of 41.9 percent.

---

[19] After my replication was conducted and analyzed, Charles Plott kindly provided the original GP data. Qualitatively identical results obtain from using the GP data as in my replication. I am also grateful to Robert Reilly for providing complete access to the (collated) data from Group 1 of his Stage 2 experiment.

We find that the foregone expected income of all of these decisions is a mere 0.644 cents per decision, averaging over all subjects. Of these costs, 80%, 93%, 96% and 100% of the subjects had average costs no greater than 1, 2, 3 and 4 cents, respectively.

There is one feature of the Reilly [1982] design that might bias the results in favour of observing low costs of misbehavior. Specifically, the increase in the buyout range for the BDM procedure from (0,999) to (−400,1400) has the unfortunate feature of reducing the monetary incentives that the subjects faced. Harrison [1992; Table 1, p. 1429] demonstrates that a subject with the GP buyout interval (0,999) would have to report selling prices of plus or minus 45, 64, 78, 90 or 100 cents in order to generate expected foregone incomes of 1, 2, 3, 4 or 5 cents, respectively. With the *expanded* buyout range the foregone expected income of these false reports *drops* to 0.550, 1.120, 1.668, 2.224 and 2.749 cents, respectively.

We can hypothetically evaluate the effect of expanding the buyout range on the cost of the observed subject behavior. This exercise provides some check that the non-dominance of payoffs in this experiment is not solely attributable to the expansion of the buyout range. Specifically, we simply set all reported valuations less than zero cents to zero and all reports greater than 999 cents to 999. The effect of this is to slightly increase the average cost of inconsistent valuations from 0.644 cents to 0.788 cents, pooling over all 45 subjects. Clearly this feature of the Reilly [1982] design, although unfortunate from the perspective of enhancing subject incentives, makes little difference to our conclusions.[20]

B The GP Replications

We have replicated the design of Experiment 1 of GP. Our procedures differ slightly, primarily due to the use of a microcomputer laboratory to undertake

---

[20] One feature of these preference reversal experiments that makes a major difference is the use of only one decision to reward subjects. As noted earlier, this implies that each costly decision has only a 1-in-18 chance of actually being costly. We can hypothetically evaluate the effect of this design feature by assuming that subjects were paid according to all of their decisions. This raises the average cost of "as if" preference reversals from 0.6444 cents to a serious 11.596 cents. We emphasize, however, that this thought-experiment is not the empirical experiment that these subjects were faced with. It is therefore a properly open behavioral question if they would report the same degree of "inconsistent" valuations if forced to live with the full costs of same.

Moreover, the earnings of subjects leap as one pays them for every decision. By stating selling prices of zero for all of the Part 2 decisions, for example, a subject could expect to earn $60 from the demand buyout (an average buyout of $5 time 12 decisions). On top of any earnings from Parts 1 and 2, this would make a foregone expected income of 11 cents seem trifling. This result raises an intriguing issue – what is the "appropriate" metric with which to evaluate the opportunity cost of an observed decision? Implicitly we have adopted throughout a metric of foregone income, measured simply in dollars and cents. However, in this setting it seems more appropriate to consider *percent* foregone income relative to the expected income at the optimal decision.

the experiment. All subjects were economics undergraduates at the University of Western Ontario.

Experiment *R1* employed 20 subjects with no monetary incentives to make good decisions: all subjects were paid a fixed amount just for participating. One variant of the GP procedure is that we played out the results of each decision after it was made, rather than selecting just one decision at the end of the experiment. The rate of observed preference reversals was 49%, somewhat higher than the 33% observed by GP (Table 5, p. 632). The hypothetical foregone expected income from these inconsistent choices was an average 18.3 cents per decision. These subjects hypothetically earned an average of $111 for the experiment, however.

Experiment *R2* employed 14 subjects with payoffs depending on just one of their decisions. The preference reversal rate was 45%, very close to the rate we observed in R1 for completely unmotivated subjects drawn form the same population. *The foregone expected income per inconsistent decision averaged only 0.6 cents!* We conclude that the subjects in these preference reversal experiments had virtually no incentive to behave any more consistently than they did.

C *Controlling for Payoff Dominance*

In an effort to increase the motivation of subjects we conducted an experiment with 13 subject facing the same design as *R2* but with a buyout range of (0,500) instead of (0,999). This constrains the selling prices that can be reported, but doubles the expected cost of a false report in this region. Moreover, the expected actuarial value of the GP wagers varies between $1.35 and $3.86, which is well within the truncated buyout interval. We observe a preference reversal rate of 46%, roughly as before. We also observe an average foregone expected income of only 0.5 cents, again as in experiment *R2*.

The Dominance problem arises in these experiments because of two factors. The first is the use of the BDM procedure for the elicitation of selling prices. For any given lottery, the elicited prices can be widely dispersed around the "true" certainty equivalent and the subject not forego significant amounts of expected income. The second factor is that the expected value of the paired lotteries are deliberately very similar. The expected values in the six pairs of GP [1979; Table 2, p. 629] differed by only 1, 7, 3, 6, 6, and 8 cents, respectively. The rationale for having lottery pairs with similar expected values, although not made explicit in GP, is to focus solely on the disparate effects of the elicitation devices (direct preference reports versus selling prices).

We designed an experiment to test the effect of controlling for payoff Dominance by varying the increment in selling prices that could be reported. The idea is to require the subject to state a selling price in increments of 25 cents, 50 cents, 100 cents or 200 cents, rather than the default increment of 1 cent. By doing this

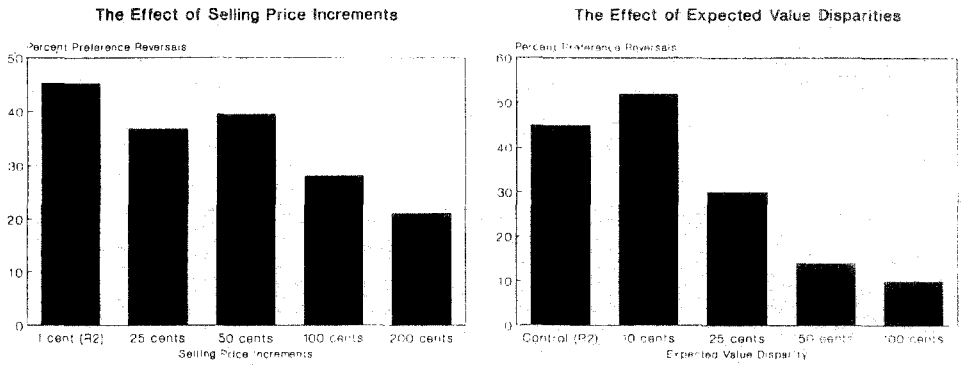

Fig. 5. Preference reversals

we reduce the accuracy of the report, but increase the opportunity cost of false reports. For example, an increment of 25 cents implies that a non-truthful report 25 cents either side of the true certainty-equivalent will have up to 25 times the probability of being costly for the subject. In each of these experiments we retained the original buyout range of (0,999).

The results presented in Figure 5 indicate a noticeable reduction in the rate of preference reversals using this treatment. Nonetheless, the payoff Dominance problem is not easily eradicated, even with this treatment. The average cost of "as if" preference reversals varies from 0.591 cents in our Control experiment *R2* to 0.777 cents, 0.716 cents, 0.417 cents, and 0.351 cents, respectively, for the experiments with increments of 25, 50, 100 and 200 cents. These experiments employed 19, 19, 19 and 15 subjects, respectively.

We also designed an experiment to test the effect of controlling for payoff Dominance by varying the expected value of the lottery pairs. If we make the disparity in expected values large enough for a given pair, then presumably the flatness of the payoff function with respect to selling prices will matter less and less. For example, assume that the subject is effectively indifferent between reports that are 25 cents either side of his true certainty-equivalent but is sensitive to the cost of large deviations. If the expected values of the lotteries are more than 50 cents apart then the selling prices elicited for each will generate the true ordinal preference even if they are "off" in terms of accurately representing the certainty-equivalent value. Moreover, if we assume that the opportunity cost of a non-truthful selling price increases monotonically from the true selling price, then as we steadily increase the disparity in expected values we should expect to see monotonically fewer preference reversals.

Admirable as this logic is, the calculations in Harrison [1992, Table 1, p. 1429] suggest that we may need to induce potentially huge differences in expected value in order to get a noticeable reduction in preference reversals. Using the example presented there, the subject could state selling pries in a $2.00 range around the true certainty-equivalent and still only forego an expected income of 5 cents.

Notwithstanding this pessimistic outlook, it seemed appropriate to venture into the laboratory and see if human subjects perceived things as outlined above – after all, experimental economics *is* an observational discipline. From Figure 5 it is apparent that they did not! The first treatment employed was to ensure that the expected values of each of the paired lotteries differed by 100 cents. In each of these experiments we employed a buyout range of (0,999), again wanting to focus solely on the effect of changing the expected values.[21] The preference reversal rate fell to only 10% over 25 subjects, compared to the 45% observed in experiment *R2* with 14 subjects drawn from the same population. Drunk with the brute simplicity of this observation, it seemed natural to see if smaller disparities in expected value would have a similar effect. Figure 5 shows that there is indeed a marked reduction in the rate of preference reversals for expected value disparities of 50 cents (over 13 subjects). The rate is somewhat higher when the disparity is only 25 cents (12 subjects), and is actually higher than in *R2* when the disparity is just 10 cents (12 subjects).

Nonetheless, the monotonic decline in the rate of preference reversals as the expected value disparity is increased from 25 cents up to 100 cents is consistent with our underlying hypothesis that the experimental evidence against EUT relies, in part, on a failure to control for payoff Dominance. The fact that a disparity of only 100 cents was sufficient to induce consistent behavior is somewhat surprising, but serves as a useful reminder that we are necessarily dealing with subjective and environment-specific costs of decision-making in experiments.

3 Prospect Theory

Kahneman and Tversky [1979] and Tversky and Kahneman [1986], hereafter generically KT and TK, catalogue a wide range of experimental anomalies from the perspective of EUT. To the extent that virtually all of these anomalies involve non-salient hypothetical questionnaires, one is tempted to dismiss them out of hand as not meeting the sufficient conditions for a valid microeconomics experiments. Two counter-arguments cause us to consider these anomalies more carefully, however. The first is the claim that these anomalies "persist even in the presence of significant monetary payoffs" (KT [1986; p. 90]; see also Thaler [1986; p. 96]). The second counter-argument is the fact that

[21] The expected values were changed by increasing the "win prize" in the P-bet of GP [1979; Table 2], except for pair 4 in which we lowered the value of "win prize". Consider pair 1, for example. The P-bet has a probability of winning of 35/36. To increase the expected value of the bet by ζ we simply increase the "win prize" by $\zeta/(35/36)$. Thus an increase in expected value of 100 cents requires an increase in the "win prize" of $103 \simeq 102.857$ cents.

these anomalies, and the Prospect Theory constructed to account for them, have achieved wide recognition and some respect amongst economists.

We evaluate the claim about the unimportance of monetary payoffs in two stages. First we re-examine those few experiments of KT and TK that did indeed use financial incentives, and that are heavily and repeatedly cited as support for the claim.[22] We find that the majority of these experiments do not satisfy the Dominance precept. Then we report the results of some new experiments designed to avoid our criticisms.

3.1 Some Old, But Salient, Experiments

A Salient Experiment 1

KT [1972] conducted one experiment with 97 Stanford undergraduates in which there were 3 questions. Each question defined a stochastic process (e.g., number of days in a year in which more boys than girls are born in a given hospital). The mean of this process is also specified (e.g., one-half of the days). A critical value above the mean is then introduced (e.g., number of days in which more than 60% of the babies born are boys). The subject is finally asked if this critical value is more likely to be observed in a small sample or a large sample (e.g., in a small hospital which has about 15 births per day or a larger hospital which has about 45 births per day).

The subjects were paid $1 if their answer to one of the three problems was correct. Thus the incentive to provide a correct answer on any given problem was 33 cents. Pooling over all three problems, 71% of the answers were incorrect. Viewing this proportion as representative of the likelihood that a given subject would give an incorrect answer, we see that such misbehavior leads to the subject foregoing an expected payoff of 0.71 (33 cents) = 23.6 cents per decision.

It should also be noted that KT gave the subject three possible answers. Two of these allowed an unambiguous choice of one or other of the samples (e.g., "The larger hospital", or "The smaller hospital"). The third allowed for a rough indifference to be expressed: viz., "About the same (i.e., within 5% of each other)". This approximate indifference option was chosen in 47% of all answers, and may indicate subject confusion or disinterest. If we disregard such choices, the foregone expected payoff from unambiguous errors drops to (70/152) 33 = 15.2 cents per decision. KT do not say how long this experimental session lasted.

---

[22] Undoubtedly there are several other experiments conducted with financial incentives that we have overlooked. Nonetheless, the general message should be quite clear from the sample we do consider. Harrsion [1992;p.1430] evaluates one further salient experiment by TK [1986; pp. 80–81].

B *Salient Experiment 2*

TK [1973] conducted one experiment with 50 Stanford undergraduates which involves them considering the following diagram:

x x 0 x x x
x x x x 0 x
x 0 x x x x
x x x 0 x x
x x x x x 0
0 x x x x x

A path was defined for the subject as "any descending line which starts at the top row, ends at the bottom row, and passes through exactly one symbol (x or 0) in each row". Subjects were simply asked if they thought that there are more paths connecting six x's and no 0 or more paths connecting five x's and one 0. Each subject received \$1 for a correct answer and zero payoff for an incorrect answer.

The distribution of paths is "well-known", to combinatorial theorists presumably, as being binominal with $p = 5/6$ and $n = 6$. Thus about 33% of the paths consists of six x's, and about 40% of the paths consists of exactly five x's. Hence the correct answer is that there are more paths consisting of exactly five x's. The proportion of paths that consists of exactly 4, 3, 2, 1 or 0 x's is each significantly lower than either of these two (around 20% for exactly 4 x's, and below 6% for each of the other paths).

Each subject was rewarded with \$1 if he gave a correct answer. Nonetheless, 38 of the subjects gave the wrong answer. Viewing these as responses from a homogenous sample, a subject is therefore likely to make a mistake with 0.76 probability, implying a foregone expected payoff (due to this likelihood of mistake) of 76 cents. Relative to the extraordinary computational difficulty of the task this is not a large payoff.[23] Again, TK do not say how much time the subjects were given to respond to the question.

C *Salient Experiment 3*

TK [1981] posed 3 simple stochastic choice questions to a single group of subjects and gave one-tenth of each group the opportunity to play the gamble

[23] Smith [1982; p. 934, fn. 17] correctly points out that the Dominance requirement for a given (cognitive or computational) task should be defined relative to the availability of (perceptual or calculating) aids. Did these subjects have pen and paper available? Did they have pocket calculators? Were there any implicit or explicit time limits on their answers?

they chose. These questions were as follows, with the percent of subjects choosing each option reported in square brackets:

1) Choose between (*A*) a sure gain of \$30 [78%], and (*B*) 80% chance to win \$45 and 20% chance to win nothing [22%];
2) Choose between (*C*) 25% chance to win \$30 and 75% to win nothing [42%], and (*D*) 20% chance to win \$45 and 80% chance to win nothing [58%]; and
3) Consider the following two stage game. In the first stage, there is a 75% chance to end the game without winning anything, and a 25% chance to move into the second stage. If you reach the second stage you have a choice between (*E*) a sure win of \$30 [74%], and (*F*) 80% chance to win \$45 and 20% chance to win nothing [26%].

Problems (2) and (3) are identical in EUT terms, and problem (2) is a simple positive affine transformation of problem (1). Hence a preference for *A* over *B* implies a preference for *C* over *D* and for *E* over *F*, according to EUT.

Assume that the observed choices refer to a homogenous pool of risk-neutral agents, so that we can think of the responses as coming from one representative agent rather than the sample of 85 subjects. The expected value of each lottery is readily calculated as: (*A*) \$30, (*B*) \$36, (*C*) \$7.50, (*D*) \$9, (*E*) \$7.50 and (*F*) \$9.

How costly was it for our representative subject to behave as observed? By only reporting a preference for *B* over *A* with probability 0.22 (instead of with probability 1), he foregoes $(1 - 0.22)((36 - 30) \div 10) = \$0.468$. By only reporting a preference for *D* over *C* with probability 0.58, he foregoes $(1 - 0.58)((9.00 - 7.50) \div 10) = \$0.063$. Similarly, he foregoes \$0.111 in problem (3). Thus the total response pattern for all three questions implies a foregone expected income to a risk neutral agent of $0.468 + 0.063 + 0.111 = \$0.642$.

Given the clear pattern of preferences in questions (1) and (3), and the lack of a clear preference pattern in question (2), it seems more natural to assume that our representative agent is risk averse. Let us assume further that the certainty-equivalent income required to compensate for the riskiness of prospect *B* is \$6 (plus ε). Thus our agent foregoes $0.22(6 \div 10) = \$0.132$ by reporting a preference for B over A with probability 0.22 (instead of with zero probability). Similarly, he foregoes $0.58(1.50 \div 10) = \$0.087$ and $0.26(1.50 \div 10) = \$0.039$ in problems (2) and (3) by stating a preference for the prospect that is more risky with probability 0.58 and 0.26, respectively. Thus a risk averse agent foregoes as little as $0.132 + 0.087 + 0.039 = \$0.258$. This is clearly a lower bound, since we assume the least possible certainty-equivalent such that the preference under EUT would not be for *B* over *A*.

We conclude that a representative subject who behaved in the probabilistic pattern observed would, under plausible enough assumptions, have foregone between 26 cents and 47 cents.

D Salient Experiment 4

TK [1983; p. 303] devise an experiment to illustrate "the conjunction fallacy". This is a phenomenon involving two probabilistic events, A and B. The subject is asked, in effect, whether "A" or "A and B" are more likely. Sadly for EUT they tend to select the latter. Consider the following instructions:

> Consider a regular six-sided die with four green faces and two red faces. The die will be rolled 20 times and the sequence of greens (G) and reds (R) will be recorded. You are asked to select one sequence, from a set of three, and you will win $25 if the sequence you chose appears on successive rolls of the die. Please check the sequence of greens and reds on which you prefer to bet.
>
> 1. RGRRR
> 2. GRGRRR
> 3. GRRRRR

The subjects were 260 undergraduates at UBC and Stanford, with 125 of them playing out the gamble as stated in the instructions (one presumes that an oral addition to these instructions alert subjects in the hypothetical payment to the fact that their decisions are not salient). Sequence 1 corresponds to event A, and "a G at the beginning of the sequence" corresponds to event B, hence sequence 2 corresponds to the conjunction of A and B. The subjects facing monetary payoffs chose sequence 1, 2 and 3 in 33%, 65% and 29% of the observed choices. The choice pattern for the other subjects was virtually identical.

The exact probability of each sequence occurring in a 20-trial sequence is messy to calculate. A close approximation can be obtained by asking what the probability is of observing sequence 1 in any 5 rolls of the die ($= \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = 0.0082304$), of observing sequence 2 in any 6 rolls ($= 0.0082304 \cdot \frac{2}{3} = 0.0054869$), and of observing sequence 3 in any 6 rolls ($= 0.0027434$). With a $25 prize, the expected value of choosing sequence 1, 2 or 3 is 20.576 cents, 13.717 cents and 6.858 cents, respectively. By choosing sequence 2 over sequence 1 a risk neutral subject forgoes only 6.859 cents. If we again think of the observed frequency of choices as representing the probability of a given (representative) subject choosing sequence 2 or 3 over sequence 1, we find that the opportunity cost of the sub-optimal behavior is $0.65(6.859) + 0.02(13.718) = 4.458 + 0.274 = 4.732$. Again, relative to the extraordinary computational difficulty of the decision facing the subject, this is not a particularly costly error.

3.2 Some New Experiments

A large number of the choice anomalies presented by KT rest on a failure of the axiom of EUT that states that if $B \succ A$ then $(B; p) \succ (A; p)$ for any probability

p, where $(X; q)$ denotes a (possibly compound) lottery in which X is received with probability q. Consider, for example, Problems 7 and 8 of KT [1979; p. 267]. The subject is asked to choose between A: (6000;0.45) and B: (3000;0.90), and 86% of their subjects chose prospect B. The subject is separately asked to choose between C: (6000;0.001) and D: (3000;0.002), and 73% of the subjects chose C. Clearly C is just the compound lottery (A;1/450) and D the compound lottery (B;1/450). EUT predicts that the subjects should choose D over C, but the modal choice pattern of this subject pool is the reverse of this.[24]

In order to focus on the question of payoff Dominance we have conducted a series of Salient experiments in which subjects were presented with a number of

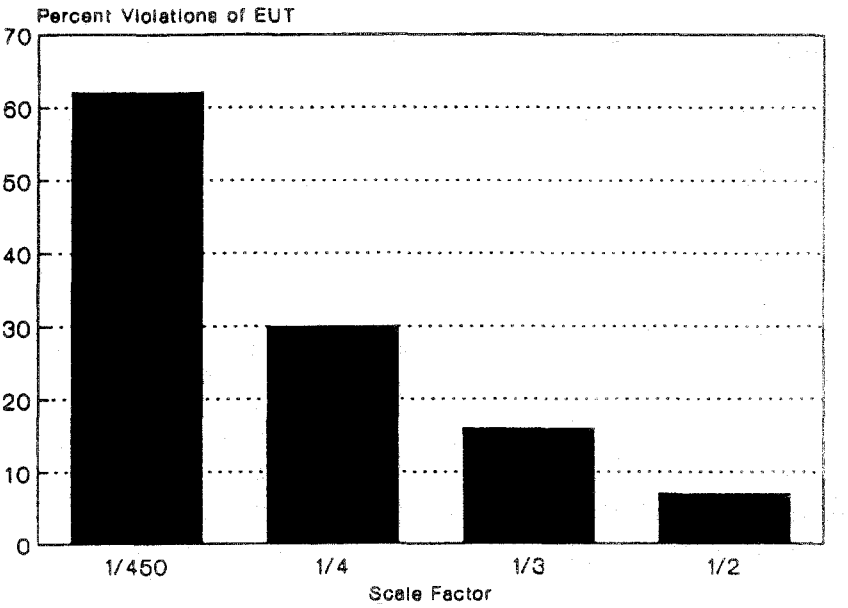

Fig. 6. Tests of EUT in simple pairwise lottery choices

[24] One frustrating aspect of the reporting style of KT is that they do not typically report the pattern of choices in paired decision problems. In this example we would like to know the percentage of subjects that chose B over A and *then* chose C over D, violating EUT. Given that 14% of the subjects chose A over B and 73% chose C over D, it is possible that all of the subjects in this 14% behaved consistently with EUT and chose C, so that the percentage of subjects violating EUT is only 73% − 14% = 59%. This is still, of course, likely to be significantly greater than zero, but it is a more meaningful measure of the performance of EUT. It should also be noted that although KT [1979; p. 266] report when a preference is "significant at the .01 level" they do not tell us the null or alternative hypotheses, let alone the statistical procedure used. Their classification of the significance of their results is consistent with a χ^2 test of a null of equi-probable choices against a two-sided alternative. Note that these tests appear to have been performed for the pairwise choice, not for the pattern of choices predicted by EUT.

pairwise choice problems of this kind. We vary the "scale factor" p in these problems so as to evaluate behaviorally the trade-off between payoff Dominance and violations of EUT. To see why this is important, consider the expected value of the above prospects for a risk neutral subject, presuming that "6000" refers to "6000 cents". Lotteries A and B have an expected value of \$2.70, whereas C and D only have an expected value of \$0.01! Thus if a strict preference over A and B is based on some attitude to risk, so must any preference defined over C and D. However, any differences in the certainty-equivalents of the lottery pairs A and B must be miniscule for the lottery pairs C and D for plausible attitudes to risk. The upshot of this is that if we consider a range of values of the "scale factor" p we should be able to capture an effect from varying the extent of payoff Dominance.

Figure 6 displays the results of our experiments, conducted with 50 undergraduate economics students from the University of Western Ontario. The monetary prizes used were C\$6.00 and C\$3.00 as above. The "scale factor" p was set equal to a range of values, 1/450, 1/4, 1/3 and 1/2. Each subject faced the five choices in a random sequence. The results are very clear. As the payoff Dominance problem is reduced, by increasing the value of p, the extent of violations of EUT declines dramatically. Note that we observe 62% of all choices for the $p = 1/450$ problems were inconsistent with EUT, which is a result comparable to KT [1979]. This indicates that, unlike the Allais Paradox considered earlier, the pure effect of Salient payoffs is not sufficient to ensure consistent choice behavior. Rather, one must also ensure that the payoffs for the decision problem satisfy Dominance.

4. Bayes Rule

4.1 The Experiment

The most influential study on the extent to which agents use Bayes Rule in making statistical inferences has been Grether [1980]. He devised a classic experiment to test the hypothesis that subjects behave as if they follow Bayes Rule in a simple statistical inference problem.[25] One alternative hypothesis, advocated by KT [1972], is that agents use a Representativeness Heuristic in such problems which will tend to bias their inferences relative to the prediction derived from following Bayes Rule. This Heuristic essentially holds that subjects

[25] A number of studies have extended this test of Bayes Rule to market environments: see Anderson and Sunder [1988], Camerer [1987] and Duh and Sunder [1986].

will evaluate the probability of a sample according to the population process that it most closely "represents". From a Bayesian perspective it is as if such subjects attach too little weight to prior probabilities and too much weight to sample evidence. With non-diffuse priors this will lead to systematic departures from Bayes Rule.

Grether [1980; p. 540] correctly notes three major problems with the experiments that the psychologists claim support this Heuristic. First, the inferential problem itself involved highly subjective verbal situations which subjects might not know how to operationalize. Second, there may have been some incredibility with respect to the random processes that were described to the subjects: the Peterson and Ulehla [1965] binary choice experiments are a celebrated example of this effect. Finally, the psychology experiments were not Salient.

The experimental design proposed by Grether [1980] nicely eliminates these problems. The instructions that our subjects received, in a replication of his design, succinctly describe the task:

> The experimenters are trying to determine how people make decisions. We have designed a simple choice experiment, and we shall ask you to make decisions at various times. The amount of money you make will depend on how good your decisions are. During the experiment you will be asked to make ten decisions. At the end of the experiment we shall randomly choose one of your ten decisions and give you $10 if it is correct. If it is incorrect you will receive $1. [In treatment B2 the previous 2 sentences were replaced with: If a decision is correct you will receive $1.00. If it is incorrect you will only receive $0.10. Therefore, if you make ten correct decisions you will earn $10.00.] We shall then repeat the experiment for another ten decisions, with the same rules. [In treatment B0 the previous 5 sentences were replaced with: You will receive $2.50 for making these decisions. During the experiment you will be asked to make ten decisions. We shall then repeat the experiment for another ten decisions, with the same rules. You will receive $2.50 for each set of ten decisions ($5 for all twenty decisions).]
>
> The computer is going to use three randomizing devices in this experiment, which you can think of as bingo cages designated as CAGE A, CAGE B, and CAGE X. Inside both CAGE A and CAGE B are six balls, some of which are marked with an N and some with a G. CAGE A has four N's and 2 G's and CAGE B has three N's and 3 G's. Inside CAGE X there are six balls numbered one, two, three, four, five, and six.
>
> The experiment will proceed as follows. First, the computer will spin CAGE X. Before each run we will tell you that if certain numbers come up, we will choose CAGE A, and otherwise we will choose CAGE B. For example, if 1, 2, 3 or 4 are drawn, we will pick CAGE A; if a 5 or 6 is drawn, we will pick CAGE B. After drawing from CAGE X, the computer will choose the appropriate CAGE (A or B). Then the computer will make six draws from the CAGE that has been selected, replacing the drawn ball each time. You will be told the results of the draws by the computer. You will then be asked to indicate (to the computer, and in writing on your Record Sheet) whether you think the draws come from CAGE A or CAGE B.
>
> The only talking allowed during the course of the experiment will be to clarify questions you may have about the procedure, and these questions should be directed to the experimenter. The computer will now take you through a trial decision to make sure that you understand the procedure. This decision (Decision 0) will not be used to determine your payoffs.

These instructions vary slightly from those used by Grether [1980; p. 555–6], primarily with respect to the use of a microcomputer to conduct the experiment. The only problem that this variation may have generated is the credibility of the

Table 2. Parameters and expected payoffs in Bayes rule experiment

Prior Probability of Cage *A*	Number of N's in sample...						
	0	1	2	3	4	5	6
	(a) Posterior Probability of Cage *A*						
.67	.149	.260	.413**	.584**	.737	.849	.918
.5	.081	.149	.260	.413***	.584***	.737	.849
.33	.042	.081	.149	.260	.413*	.584**	.737
	(b) Correct Decision						
.67	B	B	B**	A*	A	A	A
.5	B	B	B	B***	A***	A	A
.33	B	B	B	B	B*	A**	A
	(c) Expected Payoffs for Correct Decision (Grether Design; in U.S. cents)						
.67	42.7	29.2	10.6**	10.2*	28.8	42.4	50.8
.5	50.9	42.7	29.2	10.6***	10.2***	28.8	42.4
.33	55.7	50.9	42.7	29.2	10.6*	10.2*	28.8
	(d) Expected Payoffs for Correct Decision (Current Design; in Canadian cents)						
.67	63.2	43.2	15.7**	15.1*	42.7	62.8	75.2
.5	75.4	63.2	43.2	15.7***	15.1***	42.7	62.8
.33	82.4	75.4	63.2	43.2	15.7*	15.1**	42.7

Note: Recall that Cage *A* has 4 N's and Cage *B* has 3 N's.

random process[26], although if this effect is systematic it will bias results away from the null hypothesis (Bayes Rule).

Table 2 shows the parameters and expected payoffs in our experiment. Panels (a) and (b) compute the posterior probability of Cage A generating each possible sample as a function of the prior and the observed sample. Those entries with one or more asterisks all have approximately the same posterior probability favouring the correct decision according to Bayes Rule. However, those cells with exactly one asterisk are instances in which the Representativeness Heuristic would lead a subject to make the incorrect decision: with 3 N's the Heuristic implies that the population would be Cage B, and with 4 N's it would be Cage A. In each case there are specific priors that would overshadow this sample evidence. Cells with two asterisks are instances in which the Heuristic would favor neither choice, and three asterisks denote cells in which the Heuristic favors the correct decision. The attractive feature of this design is that the predictions of the two hypotheses differ sharply for a set of possible out-

[26] Having the computer draw the sample may also have made it difficult for subjects to operationalize the random process. Hence the instructions contain liberal references to the idea that the subject could think of these drawings as if they had come from an urn (the physical process employed by Grether [1980]).

comes that hold constant the strength of the incentives for one of the hypotheses (the null, Bayes Rule).

Three financial incentive structures were employed. The first was to pay subjects a fixed amount for making a certain number of decisions (C$3 in our experiment). The second was to pay subjects for one of their decisions. The third was to pay subjects for each decision. In each case we revealed the true Cage generating the sample after each decision.

Each subject was asked to make 20 decisions. These were divided into two groups of ten, with any earnings tallied at the end of each set of ten decisions. In terms of the information about total earnings revealed to subjects the first and second set of ten decisions correspond to the usual notions of "inexperienced subject behavior" and "experienced subject behavior", respectively. Grether [1980] conducted one session with his subjects, lasting an average of 16.45 periods.[27] He employs a non-standard definition of "experience" when analysing his results: he calls a given subject's decision an experienced one if the subject had previously been exposed to the same prior-sample cell in panel (a) of Table 2. There is nothing "wrong" with this definition, but it does differ from the standard experimental notion of an experienced subject as one who has participated in the entire experiment before.

The reason that the precise notion of "subject experience" is of concern here is that one of the most important and often cited conclusions from Grether [1980] is that subjects do tend to follow the Heuristic instead of Bayes Rule, and that "monetary incentives do not appear to affect the behavior of inexperienced subjects (...) This is not true, however, for the experienced subjects." (p. 552). His results suggest that a combination of experience and monetary incentives only makes the case for Bayes Rule having a strong effect on the probability that a subject will make the correct choice (see his Table 4).

Finally, consider the question of the level of payoffs. Grether [1980] paid the subjects in his Fixed incentives sessions US$7; ours were paid C$3 for each set of 10 decisions in treatment B0. In his One-Shot sessions Grether [1980] rewarded subjects with US$15 if they chose the correct cage in the decision selected for payment, and US$5 if they chose incorrectly. These incentives imply the expected payoffs for a correct decision in Panel (c) of Table 2. In order to have roughly comparable payoffs, after allowing for the U.S.-Canada exchange rate (approximately US$1 = C$1.3 at the time of the experiments), our One-Shot treatment B1 effectively used C$10 and C$1 for rewards. In the Multiple Reinforcement treatment B2 our subjects received C$1 per correct decision and C$0.10 (in effect) per incorrect decision. Thus the overall expected payoff for a correct decision was the same between treatments B1 and B2 (ignoring income effects).

[27] Grether [1980] does not say how many decisions were conducted in each of his 7 sessions. The average number is derived from the reported total number of subjects (341, on p. 541) and decisions (5608, in Table V on p. 550).

The substantive questions that motivated this 3 × 2 design were twofold. First, does "experience" as normally defined in the experimental literature influence the extent of the bias in decision-making predicted by the Heuristic? Answers to this question will come from an orthogonal comparison of behavior for a given payment method. The second question, prompted by the Saliency and Dominance precepts, is whether the provision of financial incentives in (three) different ways lowers the opportunity cost (i.e., foregone expected income) of sub-optimal decisions. That is, do we observe subjects making less costly errors if they face financial incentives not to do so?[28] Prompted by some remarks by Holt [1986b], the present design operationalizes "financial incentives" in three distinct ways: Fixed (non-marginal) incentives, One-Shot incentives and Multiple Reinforcement incentives.

All subjects were drawn from the economics undergraduate program at the University of Western Ontario. Each subject received C$2 for just showing up at the appointed time. This $2 payment served the double function of providing the reward for incorrect decisions referred to above.[29] Fourteen, twenty and sixteen subjects participated in treatments B0, B1 and B2, respectively.

4.2 *Inferences from the Experiment*

The results on the direct comparison of Bayes Rule and the Representativeness Heuristic are shown in Table 3. We observe clear evidence that the Heuristic strongly influences the propensity of unmotivated and inexperienced subjects to make correct decisions. The percent of such decisions that were correct is only 32% when the Heuristic favors the wrong choice, but it is 65% when the Heuristic reinforces the correct choice and it is precisely 50% when the Heuristic provides no guidance.[30] However, when these unmotivated and experienced subjects tackle the same class of problems[31] the Heuristic has no noticeable

---

[28] In a trivial sense no subject can make any costly errors in treatment B0! We evaluate the cost of errors in these experiments as if the subject had foregone payoffs such as applied in treatments B1 and B2. We will simply refer to the "cost of errors" in B0, rather than the more accurate "cost of hypothetical errors".

[29] Thus a subject in session B1 was told that a correct decision would earn him an additional C$10 and that he would not earn any additional payoff if his decision was incorrect. Whether or not one interprets the $2 as a show-up reward or a reward for incorrect decisions has no effect on the marginal incentive for correct decisions.

[30] All of our claims concerning the results from treatment B0 are statistically significant at standard confidence levels using non-parametric Kolmogorov-Smirnov test procedures (as implemented numerically from Press, Flannery, Teukolsky and Vetterling [1986]).

[31] The subjects did not face the same random sequence of prior-outcome combinations in their "inexperienced" and "experienced" sets of decisions. The random processes were re-seeded for each separate decision.

Table 3. Percent correct decisions in Bayes rule experiments

Experiment	Payment Method	Experienced Subjects?	Representativeness Heuristic Favors...		
			Wrong Choice	Neither Choice	Correct Choice
Grether	*Fixed*	No	54	74	84
	One-Shot	No	61	80	82
B0	*Fixed*	No	32	50	65
		Yes	63	62	59
B1	*One-Shot*	No	43	54	50
		Yes	42	33	43
B2	*Multiple*	No	50	44	55
		Yes	39	38	39

Note: These percentages only refer to the six decisions which have roughly the same posterior odds in favor of the correct decision and which test the representativeness heuristic (those with one or more asterisks in Table 2).

influence at all. The propensity of such subjects to make correct decisions is about 60% irrespective of the disparate predictions based on the Heuristic.

The influence of the Heuristic is weak when inexperienced subjects facing One-Shot motivation make their decisions, and nonexistent when these subjects are experienced. Finally, there is no evidence whatsoever for the Heuristic when subjects with Multiple Reinforcement motivation make decisions. This final conclusion is true irrespective of the experience level of the subjects.

4.3 The Cost of Misbehavior

Table 4 shows the percentage distribution of the costs of sub-optimal decisions for each experiment and experience level. In all six cases the distribution is heavily skewed with a median cost of zero cents. The average costs are 11.2, 11.1 and 11.0 cents for inexperienced subjects in B0, B1 and B2, respectively, and 11.5, 10.0 and 8.6 cents for experienced subjects. Given the skewness of these distributions, it should be emphasized that the median is a better descriptive statistic than the mean as a measure of central tendency.

We therefore conclude that our experiments show that there is no evidence against Bayes Rule when subjects are experienced or face repetitive financial motivation. Moreover, any deviations from Bayes Rule tended to be virtually costless ones, irrespective of the experience level or type of financial incentives used.

Table 4. Costliness of decisions in Bayes rule experiments

Expected Cost in Cents	Experience Level	*B0*	*B1*	*B2*
0	Inexperienced	62%	65%	63%
	Experienced	61	68	72
15–16	Inexperienced	23	19	20
	Experienced	20	19	16
42–44	Inexperienced	9	13	13
	Experienced	15	9	8
62–64	Inexperienced	6	4	4
	Experienced	3	5	4
≥75	Inexperienced	0	1	1
	Experienced	0	1	0

5 Conclusions

It may appear that our tone has been overly defensive, even polemical, in support of the received theory of choice under uncertainty. That may be so, but some perspective on the monolithic acceptance of the reported experimental anomalies may justify such a position. For example, Zeckhauser [1986; p. 254] proposes as an "obvious" and "untestable" axiom that for "... any rational choice, the behavioralists (e.g., Amos Tversky) can produce a laboratory counterexample." We disagree with this view. Our review of several of the most widely cited pieces of experimental evidence contrary to EUT and Bayes Rule leads to the conclusion that they do not satisfy the accepted precepts of experimental economics. Moreover, modifications to the experiments to remedy these design weaknesses results in observed choice behavior consistent with the predictions of economic theory.

References

Allais M (1979) Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'Ecole Americaine. Econometrica 21:1953, 503–546; English translation in Allais M, Hagen O (eds) Expected utility hypotheses and the allais paradox

Battalio RC, Kagel JH, Jiranyakul K (1990) Testing between alternative models of choice under uncertainty: Some initial results. Journal of Risk and Uncertainty 3:25–50

Becker GM, DeGroot MH, Marschak J (1964) Measuring utility by a single-response sequential method. Behavioral Science 9:226–232
Berg JE, Daley LA, Dickhaut JW, O'Brien JR (1986) Controlling preferences for lotteries on units of experimental exchange. Quarterly Journal of Economics 101:281–306
Camerer CF (1987) Do biases in probability judgement matter in markets? Experimental evidence. American Economic Review 77:981–997
Camerer CF (1989) An experimental test of several generalized expected utility theories. Journal of Risk and Uncertainty 2:61–104
Chew SH, Waller WS (1986) Empirical tests of weighted utility theory. Journal of Mathematical Psychology 30:55–72
Conlisk J (1989) Three variants on the allais example. American Economic Review 79:392–407
Conover WJ (1980) Practical nonparametric statistics. New York Wiley, Second Edition
Cox JC, Epstein S (1989) Preference reversals without the independence axiom. American Economic Review 79:408–426
Davis DD, Holt CA (1993) Experimental economics. Princeton Princeton University Press
Duh RR, Sunder S (1986) Incentives, learning, and processing of information in a market environment: An ecaminatio of the base rate fallacy. In: Moriarty S (ed) Laboratory Market Research. Norman OK University of Oklahoma Press.
Forsythe R, Palfrey TR, Plott CR (1982) Asset Valuation in an experimental market. Econometrica 50:537–582
Grether DM (1980) Bayes rule as a descriptive model: The representativeness heuristic. Quarterly Journal of Economics 95:537–557
Grether DM, Plott CR (1979) Economic theory of choice and the preference reversal phenomenon. American Economic Review 69:623–638
Grether DM, Plott CR (1982) Economic theory of choice and the preference reversal phenomenon: Reply. American Economic Review 72:575
Harrison GW Theory and misbehavior of first-price auctions. American Economic Review 79: 749–762
Harrison GW (1992) Theory and misbehavior of first-price auctions: Reply. American Economic Review 82:1426–1443
Harrison GW, Morgan PB (1990) Search intensity in experiments. Economic Journal 100:478–486
Hey JD, Di Cagno D (1990) Circles and triangles: An experimental estimation of indifference lines in the marschak-machina triangle. Journal of Behavioral Decision Making 3:279–306
Holt, CA (1986a) The independence axiom and preference reversals. American Economic Review 76:508–513
Holt CA (1986b) Discussant's comments In: Moriarty S (ed.). Laboratory Market Research. Norman OK University of Oklahoma Press
Kahneman D, Tversky (1972) Subjective probability: A judgement of representitiveness. Cognitive Psychology 3:430–454
Kahneman D, Tversky A (1973) On the psychology of prediction. Psychological Review 84:237–251
Kahneman D, Tversky A (1978) Prospect theory: An analysis of decision under risk. Econometrica 47:263–291
Karni E, Safra Z (1987) Preference reversals' and the observability of preferences by experimental methods. Econometrica 55:675–685
Leamer EE, Leonard HB (1983) Reporting the fragility of regression estimates. Review of Economics and Statistics 65:306–317
Loomes G, Sugden R (1983) A rationale for preference reversal. American Economic Review 73: 428–432
Machina MJ (1982) Expected utility analysis without the independence axiom. Econometrica 50: 277–323
Machina MJ (1987) Choice under uncertainty: Problems solved and unsolved. Journal of Economic Perspectives 1:121–154

Mosteller F, Nogee P (1951) An experimental measurement of utility. Journal of Political Economy 59:371–404

Peterson CR, Ulehla ZJ (1965) Sequential patterns and maximizing. Journal of Experimental Psychology 69:1–4

Plott CR, Sunder S (1982) Efficiency of experimental security markets with insider information: An application of rational-expectations models. Journal of Political Economy 90:663–698

Pommerehne WW, Schneider F, Zweifel P, (1982) Economic theory of choice and the preference reversal phenomenon: A reexamination. American Economic Review 72:569–574

Press WH, Flannery BP, Teukolsky SA, Vetterling WA (1986) Numerical recipes: The art of scientific computing. Cambridge University Press

Reilly RJ (1982) Preference reversal: Further evidence and some suggested modifications in experimental design. American Economic Review 72:576–584

Roth AE (1988) Laboratory experimentation in economics: A methodological overview. Economic Journal 89:974–1031

Roth AE, Malouf MWK (1979) Game-theoretic models and the role of information in bargaining. Psychological Review 86:574–594

Segal U (1988) Does the preference reversal phenomenon necessarily contradict the independence axiom? American Economic Review 78:233–236

Starmer C, Sugden R (1991) Does the random-lottery incentive system elicit true preferences? An experimental investigation. American Economic Review 81:971–978

Thaler RH (1986) The psychology and economics conference handbook: Comments on simon, on einhorn and hogarth, and on tversky and kahneman. In: Hogarth RM, Reder MW (eds) Rational Choice. Chicago University of Chicago Press

Tversky A, Kahneman D (1971) Belief in the law of small numbers. Psychological Bulletin 76: 1105–1110

Tversky A, Kahneman D (1973) Availability: A heuristic for judging frequency and probability. Cognitive Psychology 5:207–232

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–458

Tversky A, Kahneman D (1983) Extensional versus intutive reasoning: The conjunction fallacy in probability judgement. Psychological Review 90:293–315

Tversky A, Kahneman D (1986) Rational choice and the framing of decisions. In: Hogarth RM, Reder MW (eds) Rational Choice. Chicago University of Chicage Press

Zeckhauser R (1986) Comments: Behavioral versus rational economics: What you see is what you conquer In: Hogarth RM, Reder MW (eds) Rational Choice. Chicago University of Chicago Press