RISK AVERSION IN THE LABORATORY

Glenn W. Harrison and E. Elisabet Rutström

ABSTRACT

We review the experimental evidence on risk aversion in controlled laboratory settings. We review the strengths and weaknesses of alternative elicitation procedures, the strengths and weaknesses of alternative estimation procedures, and finally the effect of controlling for risk attitudes on inferences in experiments.

Attitudes to risk are one of the primitives of economics. Individual preferences over risky prospects are taken as given and subjective in all standard economic theory. Turning to the characterization of risk in applied work, however, one observes many restrictive assumptions being used. In many cases individuals are simply assumed to be risk neutral;[1] or perhaps to have the same constant absolute or relative aversion to risk.[2] Assumptions buy tractability, of course, but at a cost. How plausible are the restrictive assumptions about risk attitudes that are popularly used? If they are not plausible, perhaps there is some way in which one can characterize the distribution of risk attitudes so that it can be used to analyze the implications of relaxing these assumptions. If so, such characterizations will condition inferences about choice behavior under uncertainty, bidding in auctions, and behavior in games.

Risk Aversion in Experiments
Research in Experimental Economics, Volume 12, 41–196

ISSN: 0193-2306/doi:10.1016/S0193-2306(08)00003-3

We examine the design of experimental procedures that can be used to estimate risk attitudes of individuals. We also investigate how the data generated by these procedures should be analyzed. We focus on procedures that allow "direct" estimation of risk preferences by eliciting choices in non-interactive settings, since we want to minimize the role of auxiliary or joint hypotheses about Nash Equilibrium (NE) behavior in games. It is important to try to get estimates that are independent of such joint assumptions, in order that the characterizations that emerge can be used to provide tighter tests of those joint assumptions.[3] Nevertheless, we also include designs that rely on subjects recognizing a dominant strategy response in a game against the experimenter, although we will note settings in which the presumption that subjects actually use these might be suspect.[4]

In Section 1 we consider the major procedures used to elicit risk attitudes. In Section 2 we review the alternative ways in which risk attitudes have been estimated from observed behavior using these procedures. In Section 3 we examine the manner in which measures of risk attitudes are used to draw inferences about lab behavior. Section 4 offers some thoughts on several open and closed issues, and Section 5 draws some grand conclusions.

Our review is intended to complement the review by Cox and Sadiraj (2008) of theoretical issues in the use of concepts of risk aversion in experiments, as well as the review by Wilcox (2008a) of econometric issues involved in identifying risk attitudes when there is allowance for unobserved heterogeneity and "mistakes" by subjects. We take some positions on these theoretical and econometric issues, but leave detailed discussion to their surveys.

We default to thinking of risk attitudes as synonymous with the properties of the utility function, consistent with traditional expected utility theory (EUT) representations. When we consider rank-dependent and sign-dependent specifications, particularly in Sections 2 and 3, the term "risk attitudes" will be viewed more broadly to take into account the effects of more than just the curvature of the utility function.

Appendix A descriptively reviews the manner in which the humble "lottery" has been represented in laboratory experiments. Although we do not focus on the behavioral effects that may arise from the framing of the lotteries, we need to be aware that the stimulus provided to subjects often varies significantly from experiment to experiment. In effect, we experimenters are assuming that the subject views the lottery the way we view the lottery; the validity of this assumption of common knowledge between subject and observer rests, in large part, on the representation chosen by the experimenter. Some day a systematic comparison of the effects of these alternatives on risk attitudes should be undertaken, but here we simply want to provide a reminder that alternative representations exist and are used.[5] We return to this issue much later, since it relates to the manner in which laboratory experiments might provide artifactual representations of the uncertainty subjects face in the field.

In Appendices B, C, D, and E we examine in some depth the data and inferences drawn from four heavily cited studies of risk attitudes. The objective is to be very clear as to what these studies find, and what they do not find, since references to the literature are often casual and sometimes even inaccurate.

Appendices B and C focus on two bona fide classics in the area. Hey and Orme (1994) (HO) introduced a robust experimental design to test EUT, a maximum likelihood (ML) estimation procedure that does not impose parametric functional forms, and a careful discussion of the role of "errors" when making inferences about risk attitudes. Holt and Laury (2002) (HL) introduced a justifiably popular method for eliciting risk attitudes for an individual, as well as important innovations in the ML estimation of risk aversion that go beyond simplistic functional forms.

Appendices D and E focus on two studies that illustrate the problems that arise when experiments suffer from design issues or draw general inferences from restrictive models. Kachelmeier and Shehata (1992) (KS) apply an elicitation procedure that is popular, but which generates so much noise that reliable inferences cannot be drawn. Gneezy and Potters (1997) (GP) consider the important issue of "evaluation periods" on risk attitudes, but confound that valuable objective with extremely restrictive specifications of risk attitudes, leading them to incorrectly conclude that risk attitudes change with evaluation periods. In each of these studies there is an important objective; in the one case, examining risk attitudes among very poor subjects for whom the stakes are huge, and in the other case, considering the framing of the choice in a fundamental manner. But the problems with each study show why one has to pay proper attention to design and inferential issues before drawing reliable conclusions.

We conclude that there is systematic evidence that subjects in laboratory experiments behave as if they are risk averse. Some subjects tend towards a mode of risk neutrality (RN), but very few exhibit risk-loving behavior. The degree of risk aversion is modest, but does exhibit heterogeneity that is correlated with observable individual characteristics.

Some risk elicitation methods are expected to provide more reliable estimates than others, due to the simplicity of the task and the transparency of the incentives to respond truthfully. Limited evidence exists on the stability of risk attitudes across elicitation instruments, but there is some evidence to indicate that roughly equal measures of risk aversion can be obtained in the laboratory using a variety of procedures that are *a priori* attractive. There are also several methods for eliciting risk that we do not recommend.

Inferences about risk attitudes can be undertaken using several empirical approaches. One approach is to infer bounds on parameters for a limited class of (one-parameter) utility functions, but a preferable approach is to estimate a latent structural model of choice. Developments in statistical software now allow experimenters to undertake such structural estimation using ML methods. In addition, inferences about risk attitudes depend on whether the data generating process is viewed from the lens of a single model of choice behavior: there is striking evidence that two or more models may have support from different subjects or different task domains. Appropriate statistical tools exist that allow one to model the extent to which one model or another is favored by the data, and for which subjects and task domains. We review evidence that subjects exhibit some modest amounts of probability weighting, and some controversial evidence concerning the extent of loss aversion. Much of the behavioral folklore on probability weighting and loss aversion has employed elicitation procedures and/or statistical methods, which are piecemeal or have *ad hoc* properties.

Our final topic for discussion is how the characterization of behavior in a wide range of experimental tasks is affected by the treatment of risk attitudes, or confounded by the lack of such a treatment. Examples reviewed here include tests of EUT, estimates of discount rates, and evaluations of alternative models of bidding behavior in auctions. One open issue, with the potential to undermine many inferences in experimental economics, is the extent to which sample selection is driven by risk attitudes. A related concern is the reliability of measurements of treatment effects when subjects have some choice as to which treatment to participate in.

In brief, risk attitudes play a central role in experimental economics, and the nuances of measuring and controlling them demand the attention of every experimenter.

1. ELICITATION PROCEDURES

Five general elicitation procedures have been used to ascertain risk attitudes from individuals in the experimental laboratory using non-interactive settings. The first is the *Multiple Price List* (MPL), which entails giving the subject an ordered array of binary lottery choices to make all at once. The MPL requires the subject to pick one of the lotteries on offer, and then the experimenter plays that lottery out for the subject to be rewarded. The second is a series of *Random Lottery Pairs* (RLP), in which the subject picks one of the lotteries in each pair, and faces multiple pairs in sequence. Typically one of the pairs is randomly selected for payoff, and the subject's preferred lottery is then played out as the reward. The third is an *Ordered Lottery Selection* (OLS) procedure in which the subject picks one lottery from an ordered set. The fourth method is a *Becker–DeGroot–Marschak* (BDM) auction in which the subject is asked to state a minimum certainty-equivalent (CE) selling price to give up the lottery he has been endowed with. The fifth method is a hybrid of the others: the *Trade-Off* (TO) design, in which the subject is given lotteries whose prizes (or probabilities) are endogenously defined in real-time by prior responses of the same subject, and some CE elicited. We also review several miscellaneous elicitation procedures that have been proposed.

1.1. The Multiple Price List Design

The earliest use of the MPL design in the context of elicitation of risk attitudes is, we believe, Miller, Meyer, and Lanzetta (1969). Their design confronted each subject with five alternatives that constitute an MPL, although the alternatives were presented individually over 100 trials. The method was later used by Schubert, Brown, Gysler, and Brachinger (1999), Barr and Packard (2002), and Holt and Laury (2002). Appendix C reviews the HL experiments in detail.

The HL instrument provides a simple test for risk aversion using an MPL design. Each subject is presented with a choice between two lotteries, which we can call A or B. Panel A of Table 1 illustrates the basic payoff matrix presented to subjects. The first row shows that lottery A offered a 10% chance of receiving \$2 and a 90% chance of receiving \$1.60. The expected value of this lottery, EV^A, is shown in the third-last column as \$1.64, although the EV columns were not presented to subjects.[6] Similarly, lottery B in the first row has chances of payoffs of \$3.85 and \$0.10, for an expected value of \$0.48. Thus, the two lotteries have a relatively large difference in expected values, in this case \$1.17. As one proceeds down the matrix, the expected value of both lotteries increases, but the expected value of lottery B becomes greater than the expected value of lottery A.

The subject chooses A or B in each row, and one row is later selected at random for payout for that subject. The logic behind this test for risk

Table 1. Lottery Choices in the Holt/Laury and Binswanger Risk Aversion Instruments.

Lottery A				Lottery B				EV^{A}	EV^{B}	Difference
p ($2)		*p* ($1.60)		*p* ($3.85)		*p* ($0.10)				
A. Holt and Laury (2002) instrument with payoffs at the 1× level[a]										
0.1	$2	0.9	$1.60	0.1	$3.85	0.9	$0.10	$1.64	$0.48	$1.17
0.2	$2	0.8	$1.60	0.2	$3.85	0.8	$0.10	$1.68	$0.85	$0.83
0.3	$2	0.7	$1.60	0.3	$3.85	0.7	$0.10	$1.72	$1.23	$0.49
0.4	$2	0.6	$1.60	0.4	$3.85	0.6	$0.10	$1.76	$1.60	$0.16
0.5	$2	0.5	$1.60	0.5	$3.85	0.5	$0.10	$1.80	$1.98	– $0.17
0.6	$2	0.4	$1.60	0.6	$3.85	0.4	$0.10	$1.84	$2.35	– $0.51
0.7	$2	0.3	$1.60	0.7	$3.85	0.3	$0.10	$1.88	$2.73	– $0.84
0.8	$2	0.2	$1.60	0.8	$3.85	0.2	$0.10	$1.92	$3.10	– $1.18
0.9	$2	0.1	$1.60	0.9	$3.85	0.1	$0.10	$1.96	$3.48	– $1.52
1	$2	0	$1.60	1	$3.85	0	$0.10	$2.00	$3.85	– $1.85

Alternative	Probability of Bad Outcome	Bad Outcome (Indian Rupees)	Probability of Good Outcome	Good Outcome (Indian Rupees)	Expected Value
B. Binswanger (1980, 1981) instrument with payoffs at the rupees 50 level[b]					
O	1/2	50	1/2	50	50
A	1/2	45	1/2	95	70
B	1/2	40	1/2	120	80
B*	1/2	35	1/2	125	80
C	1/2	30	1/2	150	90
C*	1/2	20	1/2	160	90
E	1/2	10	1/2	190	100
F	1/2	0	1/2	200	100

[a] Experiments were also conducted at the 20×, 50×, and 90× level.

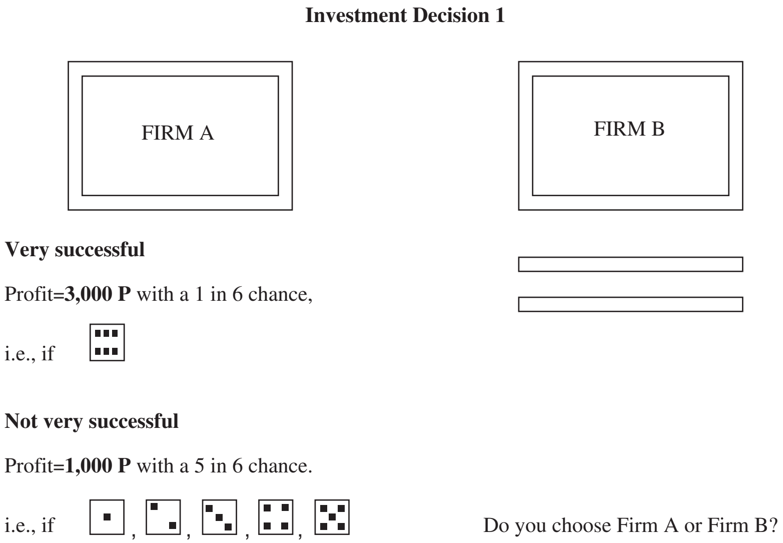
[b] Experiments were also conducted at the rupees 0.5 level (compared to alternative O) and at the rupees 5 level, with roughly 2 weeks interval.

aversion is that only risk loving subjects would take lottery B in the first row, and only risk-averse subjects would take lottery A in the second last row. Arguably, the last row is simply a test that the subject understood the instructions, and has no relevance for risk aversion at all.[7] A risk-neutral subject should switch from choosing A to B when the EV of each is about the same, so a risk-neutral subject would choose A for the first four rows and B thereafter.

The HL instrument is typically applied using a random lottery incentive procedure in which one row is selected to be played out according to the choices of the subjects, rather than all rows being played out. But that is not an essential component of the instrument, even if it is popular and widely used in many experiments to save scarce experimental funds. We discuss the random lottery incentive procedure in detail in Section 3.8.

The MPL instrument has one apparent weakness as an elicitation procedure: it might suggest a frame that encourages subjects to select the middle row, contrary to their unframed risk preferences. The antidote for this potential problem is to devise various ‘‘skewed’’ frames in which the middle row implies different risk attitudes, and see if there are differences across frames. Simple procedures to detect such framing effects, and correcting them statistically if present, have been developed (e.g., Harrison, Lau, Rutström, & Sullivan, 2005; Andersen, Harrison, Lau, & Rutström, 2006; Harrison, List, & Towe, 2007). The evidence suggests that there may be some slight framing effect, but it is not systematic and can be easily allowed for in the estimation of risk attitudes.

A variant of the MPL instrument was developed in the laboratory by Schubert et al. (1999).[8] Figs. 1 and 2 illustrate the interface provided to subjects by Barr and Packard (2002), in a sequential field implementation of this variant used in Chile. Respondents were confronted with a series of gambles framed first as an investment. The experiment then elicited their CE for an uncertain lottery. Trained experimenters asked the respondents to imagine themselves as investors choosing whether to invest in Firm A, whose profits were determined by its chances of success or failure, or Firm B, whose profits were fixed regardless of how well it fared. The experimenter explained the probabilities of Firm A’s success, the payoffs from Firm A in each state, and the fixed payoff from Firm B. The respondents were then asked to decide in which firm to invest. After registering their answer, the experimenter would raise the amount of the secure payoff, and ask the respondents to choose between the two firms again. As the amount of the secure payoff grew, investing in Firm A looked less attractive to a risk-averse respondent. In this way a CE, the point at which respondents would



Investment Decision 1

FIRM A

FIRM B

Very successful

Profit=**3,000 P** with a 1 in 6 chance,

i.e., if

Not very successful

Profit=**1,000 P** with a 5 in 6 chance.

i.e., if , , , ,

Do you choose Firm A or Firm B?

Fig. 1. Primary MPL Instrument of Barr and Packard (2002).

no longer risk investing in Firm A, was elicited for each gamble. The probability of Firm A's failure was altered three times while keeping the state-specific payoffs constant, and in the fourth investment gamble the payoffs were altered. A risk-averse subject would state a value for Firm B below the expected value of Firm A, and a risk-loving subject would state a value for Firm A above the expected value of Firm A. The subject knew that the CE "price list" would span the range shown in Fig. 2 before the sequence began.

Two variants of the MPL instrument were developed by Harrison et al. (2005d; Section 3.1), and studied at length by Andersen et al. (2006a). One is called the Switching MPL method, or sMPL for short, and simply changes the MPL to ask the subject to pick the switch point from one lottery to the other. Thus, it enforces monotonicity, but still allows subjects to express indifference at the "switch" point, akin to a "fat switch point." The subject was then paid in the same manner as with MPL, but with the non-switch choices filled in automatically. The other variant is the Iterative MPL method, or iMPL for short. The iMPL extends the sMPL to allow the individual to make choices from refined options within the option last chosen. That is, if someone decides at some stage to switch from option A to option B between values of $10 and $20, the next stage of an iMPL would

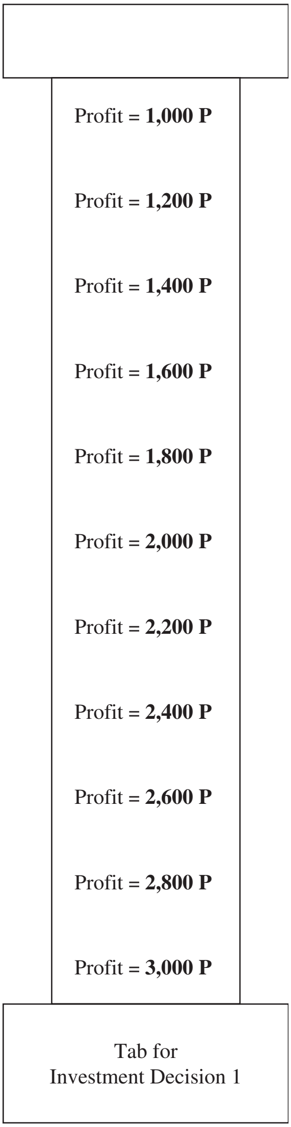


Fig. 2. Slider in MPL Instrument of Barr and Packard (2002).

then prompt the subject to make more choices within this interval, to refine the values elicited.[9] The computer implementation of the iMPL restricts the number of stages to ensure that the intervals exceed some *a priori* cognitive threshold (e.g., probability increments of 0.001). The iMPL uses the same incentive logic as the MPL and sMPL.[10]

Another feature of the MPL should be noted, although it is not obvious that it is a weakness or a strength: the fact that subjects see all choices in one (ordered) table. One alternative is to have the subjects make each binary lottery choice in a sequence, embedding them into the RLP design of Section 1.2. It is possible that allowing the subject to see all choices in one frame might lead some subjects to make more consistent choices than they would otherwise. Which approach, then, is the correct one to use? The answer depends on the inferential objective of the design, and the external context that the implied measure of risk aversion is to be applied to. We view the MPL and RLP as two different elicitation procedures: their effect on behavior should be studied systematically, in the manner we illustrate later in Section 2.5. We do not believe that consistency should always be the primary criterion for selection across elicitation procedures, particularly when one allows formally for the stochastic choice process (Section 2.3 and Wilcox (2008a)) and the possibility that it could interact with the elicitation procedure in some manner. Evidence for different risk attitudes across procedures is, by definition, a sign of a procedural artifact. But that evidence needs to be documented with formal statistical models and, if present, recognized as a behavioral corollary of using that procedure.

In summary, the set of MPL instruments provides a relatively transparent procedure to elicit risk attitudes. Subjects rarely get confused about the incentives to respond truthfully, particularly when the randomizing devices are physical die that they know that they will toss themselves.[11] As we demonstrate later, it is also possible to infer a risk attitude interval for the specific subject, at least under some reasonable assumptions.

1.2. The Random Lottery Pair Design

The RLP design has not been used directly to infer risk attitudes, but has been generally used to test the predictions of EUT. Hey and Orme (1994) used an extensive RLP design to estimate utility functionals over lotteries for individuals non-parametrically. The use of the random lottery design, coupled with treating each pairwise choice as independent, implicitly means that the estimates they provide rely on the EUT specification.

Related experimental data, from the earlier "preference reversal" debate, provide comparable evidence of risk aversion for smaller samples (see Grether and Plott, 1979 and Reilly, 1982). Additionally, many prominent experiments testing EUT provide observations based on a rich array of lotteries that vary in terms of probabilities and monetary prizes; for example, see Camerer (1989), Battalio, Kagel, and Jiranyakul (1990), Kagel, MacDonald, and Battalio (1990), Loomes, Starmer, and Sugden (1991), Harless (1992), and Harless and Camerer (1994). In most cases the published study only reports patterns of choices, with no information on individual characteristics, but they can be used to obtain general characterizations of risk attitudes for that subject pool.

Hey and Orme (1994) asked subjects to make direct preference choices over 100 pairs of lotteries, in which the probabilities varied for four fixed monetary prizes of £0, £10, £20, and £30. Subjects could express direct preference for one lottery over the other, or indifference. One of the pairs was actually chosen at random at the end of the session for payout for each subject, and the subject's preferences over that pair applied. Some days later the same subjects were asked back to essentially repeat the task, facing the same lottery combinations in different presentation order.

HO used pie charts to display the probabilities of the lotteries they presented to subjects. A sample display from their computer display to subjects is shown in Fig. 3. There is no numerical referent for the probabilities, which must be judged from the pie chart. As a check, what fraction would you guess that each slice is on the left-hand lottery? In fact, this lottery offers £10 with probability 0.625, and £30 with probability 0.385. The right-hand lottery offers the same probabilities, as it happens, but with prizes of £10 and £20, respectively. Fig. 4 illustrates a modest extension of this display to include information on the probabilities of each pie slice, and was used in a replication and extension of the HO experiments by Harrison and Rutström (2005).

HO used their data to estimate a series of utility functionals over lotteries, one for each subject since there were 100 observations for each subject in each task. This is a unique data set since most other studies rely on pooled data over individuals and the presumption that unobserved heterogeneity (after conditioning on any collected individual characteristics, such as sex and race and income) is random.

The EUT functional that HO estimated was non-parametric, in the sense that they directly estimated the utility of the two intermediate outcomes, normalizing the lowest and highest to 0 and 1, respectively. This attractive approach works well when there are a small number of final outcomes

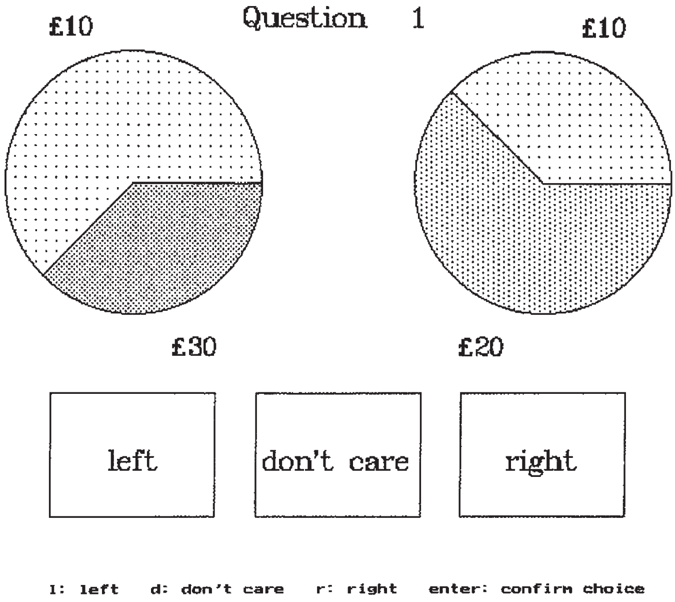


Fig. 3. Lottery Display Used by Hey and Orme (1994).

across many choices, as here, but would not be statistically efficient if there had been many outcomes. In that case it would be appropriate to use some parametric functional form for utility, and estimate the parameters of that function. We illustrate these points later.

The RLP instrument is typically used in conjunction with the random lottery payment procedure in which one choice is picked to be played out, but this is again not essential to the logical validity of the instrument.

The great advantage of the RLP instrument is that it is extremely easy to explain to subjects, and the incentive compatibility of truthful responses apparent. Contrary to the MPL, it is generally not possible to directly infer a risk attitude from the pattern of responses, and some form of estimation is needed. We illustrate such estimations later.

1.3. The Ordered Lottery Selection Design

The OLS design was developed by Binswanger (1980, 1981) in an early attempt to identify risk attitudes using experimental procedures with real

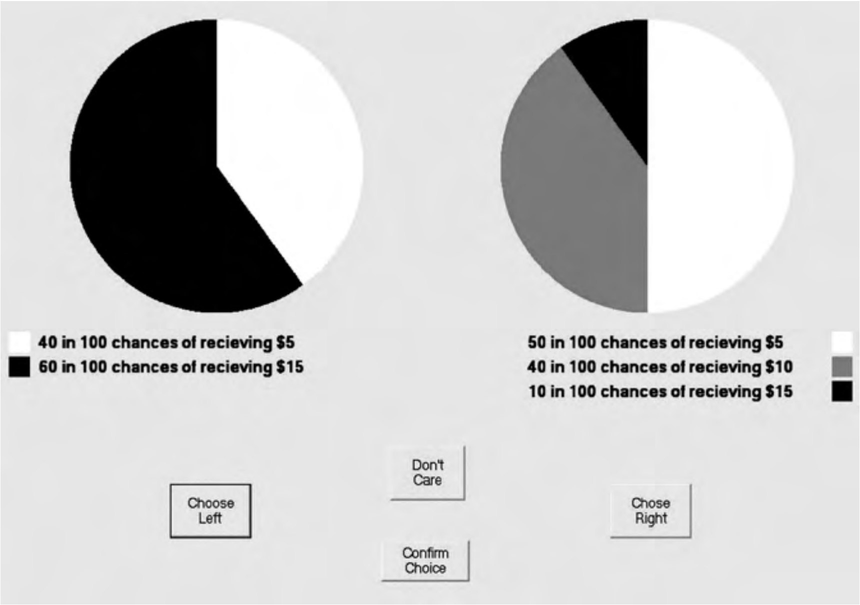

40 in 100 chances of recieving $5
60 in 100 chances of recieving $15
50 in 100 chances of recieving $5
40 in 100 chances of recieving $10
10 in 100 chances of recieving $15
Don't Care
Choose Left
Chose Right
Confirm Choice

Fig. 4. Lottery Display for Hey and Orme (1994) Replication.

payoffs. Each subject is presented with a choice of eight lotteries, shown in each row of panel B of Table 1, and asked to pick one. Alternative O is the safe option, offering a certain amount. All other alternatives increase the average actuarial payoff but with increasing variance around that payoff.

The lotteries were actually presented to subjects in the form of photographs of piles of money, to assist illiterate subjects. Each lottery had a generic label, such as the ones shown in the left column of panel B of Table 1. Fig. 5 shows the display used by Barr (2003) in a field replication of the basic Binswanger OLS instrument in Zimbabwe, and essentially matches the graphical display used in the original experiments (Hans Binswanger; personal communication). Because the probabilities for each lottery outcome are 1/2, this instrument can be presented relatively simply to subjects.[12]

The OLS instrument was first used in laboratory experiments by Murnighan, Roth, and Shoumaker (1987, 1988), although they only used the results to sort subjects into one group that was less risk averse than the other. Beck (1994) utilized it to identify risk aversion in laboratory subjects,

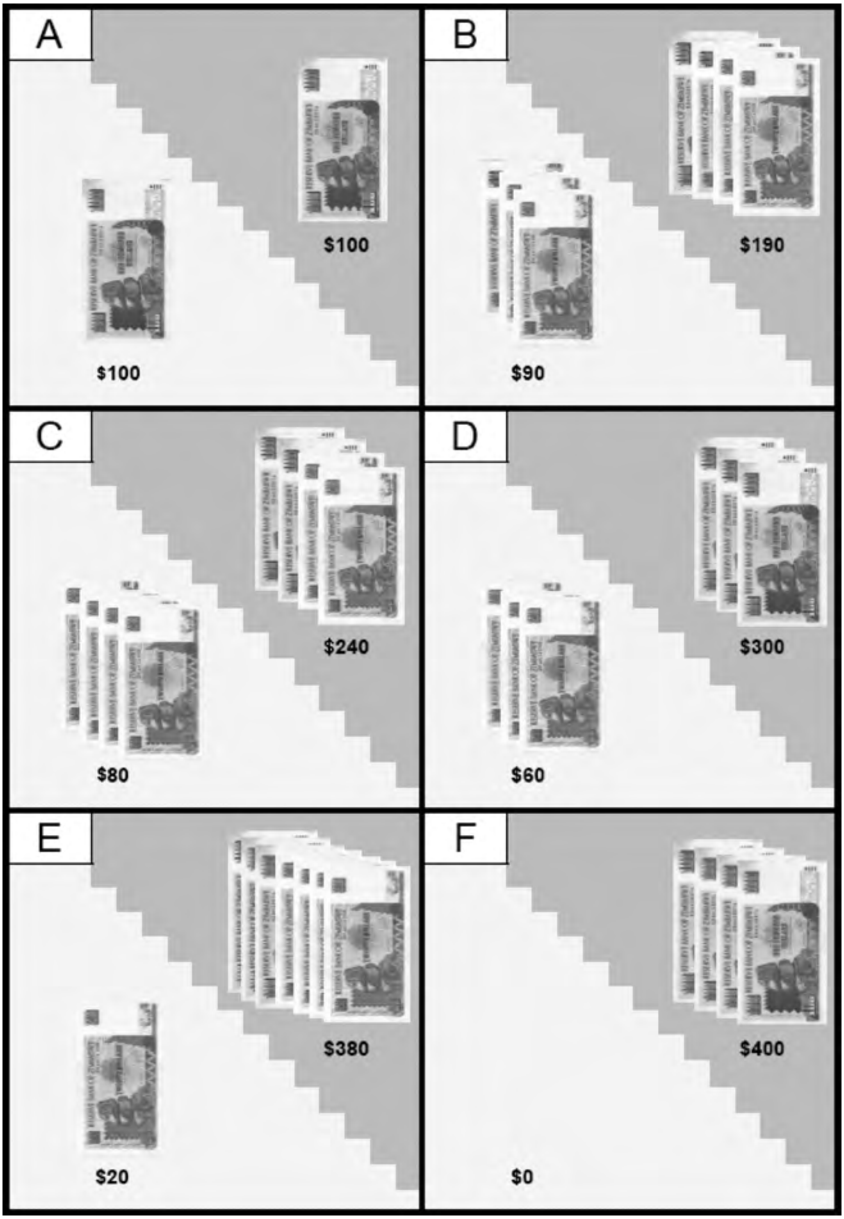


Fig. *5*. Lottery Display of Binswanger Replication by Barr (2003).

prior to them making group decisions about the dispersion of everyone else's potential income. This allowed an assessment of the extent to which subjects in the second stage chose more egalitarian outcomes because they were individually averse to risk or because they cared about the distribution of income. Eckel and Grossman (2002, 2008) used the OLS instrument to directly measure risk attitudes, as well as an innovative application in which subjects guessed the risk attitudes of other subjects. They found that subjects did appear to use sexual stereotypes in guessing the risk attitudes.

The OLS instrument is easy to present to subjects, but has two problems when used to make inferences about non-EUT models of choice behavior. The versions that restrict probabilities to 1/2 make it virtually impossible to use these responses to make inferences about probability weighting, which play a major role in rank-dependent alternatives to EUT. Of course, there is nothing in the instrument itself that restricts the probabilities to 1/2, although that has been common. The second problem is that the use of the certain amount may frame the choices that subjects make in a manner that makes them "sign-dependent," such that the certain amount provides a reference point to identify gains and losses. This concern applies more broadly, of course, but in the OLS instrument there is a natural and striking reference point for (some) subjects to use. We consider both of these issues later when we consider inferences from observed choices.

Engle-Warnick, Escobal, and Laszlo (2006) undertake laboratory experiments with the OLS instrument to test the effect of presenting the choices in different ways. The baseline mimics the procedures of Binswanger (1980, 1981) and Barr (2003), shown in Fig. 5, except that five lotteries were arrayed in a circle in an ordered counter-clockwise fashion, with the certain amount at 12 o'clock. The first treatment then presents the ordered pairs of lotteries in a binary choice fashion, so that the subject makes four binary choices. The second treatment extends these binary choices by including a lottery that is dominated by one of the original binary pairs. The dominated lottery is always presented in between the non-dominated lotteries, so it appears to be physically intermediate. Each subject made 13 decisions, which were randomized in order and left–right presentation (for the undominated lotteries). The statistical analysis of the results is unfortunately couched in terms of ordinal measures of the degree of risk aversion, such as the number of safe choices, and it would be valuable to see the effect of these treatments on estimated measures of relative risk aversion (RRA) using more explicit statistical methods (e.g., per Section 2.2, and particularly Sections 2.5 and 2.6). But there is evidence that the instruments are positively correlated, although the correlation is significantly less than one.

In particular, the correlation between the baseline OLS instrument and the transformed binary choice version for Canadian university students is 0.63, but it is only 0.31 for Peruvian farmers. Moreover, the introduction of a dominated lottery appeared to have no significant effect on the correlation of risk attitudes for the Canadian university students, but considerable effects on the correlation for Peruvian farmers.

1.4. The Becker–DeGroot–Marschak Design

The original BDM design developed by Becker, DeGroot, and Marschak (1964) was modified by Harrison (1986, 1990) and Loomes (1988) for use as a test for risk aversion.[13] This design was later used by McKee (1989), Kachelmeier and Shehata (1992) and James (2007) in similar exercises. The basic idea is to endow the subject with a series of lotteries, and to ask for the "selling price" of the lottery. The subject is told that a "buying price" will be picked at random, and if the buying price that is picked exceeds the stated selling price, the lottery will be sold at that price and the subject will receive that buying price. If the buying price equals or is lower than the selling price, the subject keeps the lottery and plays it out.

It is relatively transparent to *economists* that this auction procedure provides a formal incentive for the subject to truthfully reveal the CE of the lottery. However, it is not clear that subjects always understand this logic, and responses may be sensitive to the exact nature of the instructions given. For the instrument to elicit truthful responses, the experimenter must ensure that the subject realizes that the choice of a buying price does not depend on the stated selling price.[14] If there is reason to suspect that subjects do not understand this independence, the use of physical randomizing devices (e.g., die or bingo cages) may mitigate such strategic thinking. Of course, the BDM procedure is formally identical to a two-person Vickrey sealed-bid auction, with the same concerns about subjects not understanding dominant strategies without considerable training (Harstad, 2000; Rutström, 1998).

A major concern when choosing elicitation formats is the strength of the incentives provided at the margin, that is, the magnitude of the losses generated by misrepresenting true preferences. While the BDM is known to have weak incentives around the optimum (Harrison, 1992), the same is also true for other elicitation formats.[15] Comparing the incentive properties of the BDM to the MPL in a pairwise evaluation of a safer and a riskier lottery, we find that the expected loss from errors in the latter is a weighted average of the losses implied for the safe and the risky evaluations respectively in the BDM. The incentives in the BDM can be strengthened through a careful choice of the range of the buying prices and are generally stronger the higher is the variance of the lottery being valued.[16]

Plott and Zeiler (2005) express a concern with the way that the BDM mechanism is popularly implemented. Appendix D reviews in detail an application of the BDM mechanism for eliciting risk attitudes by Kachelmeier and Shehata (1992) and illustrates some possible problems. It may be possible to re-design the BDM mechanism to avoid some of these problems,[17] but more attractive elicitation procedures are available.

1.5. The Trade-Off Design

Wakker and Deneffe (1996) propose a TO method to elicit utility values which does not make any assumption about whether the subject weighs probabilities. This is an advantage compared to the methods widely used in the "judgement and decision-making literature," such as the CE or probability-equivalent methods,[18] since those methods assume that there is no probability weighting. The TO method proceeds by asking the subject to consider two lotteries defined over prizes x_0, x_1, r, and R and probabilities p and $1-p$: $(x_1, p; r, 1-p)$ and $(x_0, p; R, 1-p)$. It is assumed that $R > r$, p is some fixed probability of receiving the first outcome, and that x_0 is some fixed and small amount such as \$0. The subject is asked to tell the experimenter what x_1 would make him indifferent between these two lotteries. Call this stage 1 of the TO method. Then the subject is asked the same question about the lotteries $(x_2, p; r, 1-p)$ and $(x_1, p; R, 1-p)$ and asked to state the x_2 that makes him indifferent between these two. Call this stage 2 of the TO method.

If the subject responds truthfully to these questions, it is possible to infer that $u(x_2)-u(x_1)=u(x_1)-u(x_0)$ using the logic explained by Wakker and Deneffe (1996; p. 1134). Setting $u(x_0)=0$, we can then infer that $u(x_2)=2\times u(x_1)$. A similar argument leads to an elicited x_3 such that $u(x_3)=3\times u(x_1)$, and so on. If we wanted to stop at x_3, we could then renormalize $u(x_1)$ to 1/3, so that the we have elicited utility over the unit interval.

The obvious problem with the TO method as implemented by Wakker and Deneffe (1996) is that it is not incentive compatible: subjects have a transparent incentive to overstate the value of x_1, and indeed all other elicited amounts. Assume that subjects are to be incentivized in the obvious manner by one of the lotteries in each task being picked by a coin toss to be played out (or by just one such lottery being picked at random over all three stages). First, by overstating x_1 in stage 1 the subject increases the final outcome received if a lottery in stage 1 is used to pay him because x_1 is one of the outcomes in one of the lotteries in stage 1. Second, by overstating x_1 in stage 1 the subject increases the final outcome received if a lottery in stage 2 is used to pay him, since x_1 is used to define one of the lotteries in stage 2. Thus, we would expect some subject to ask us, sheepishly in stage 1, "how large an x_1 am I allowed to state?"

It is surprising that the issue of incentive compatibility was not even discussed in Wakker and Deneffe (1996), but since the actual experiments they report were hypothetical, even an otherwise incentive compatible mechanism could have problems generating truthful answers. There is a recognition that the "chaining" of old responses into new lotteries might lead to error propagation (p. 1148), but that is an entirely separate matter than strategic misrepresentation.

The TO method was extended by Fennema and van Assen (1999) to consider losses as well as gains. The experiments were all hypothetical, primarily to avoid the ethical problems of exposing subjects to real losses. The TO method was extended by Abdellaoui (2000) to elicit probability weights after utilities have been elicited. Real rewards were provided for one randomly selected binary choice in the gain domain from one randomly selected subject out of the 46 present, but the issue of incentive compatibility is not discussed. There is an attempt to elicit utility values in a non-sequential manner, which might make the chaining effect less transparent to inexperienced subjects, but again this only mitigates the second of the sources of incentive incompatibility.[19] Bleichrodt and Pinto (2000) proposed a different way of extending the TO method to elicit probability weights, but only applied their method to hypothetical utility elicitation in the health domain. They provide a discussion (p. 1495) of "error propagation" that points to some of the literature on stochastic error specifications considered in Section 2.3, but in each case assume that the error has mean zero, which misses the point of the incentive incompatibility of the basic TO method. Abdellaoui, Bleichrodt, and Paraschiv (2007b) extend the TO method to elicit measures of loss aversion. Their experiments were for hypothetical rewards, and they do not discuss incentive compatibility.[20]

1.6. Miscellaneous Designs

There are several experimental designs that attempt to elicit risk attitudes that do not easily fit into one of the five major designs considered above.

We again ignore any designs that do not claim to elicit risk attitudes in any conceptual sense that an economist would recognize, even if those designs might elicit some measure which is empirically correlated in some settings with the measures of interest to economists.

Fehr and Goette (2007) estimate a loss aversion parameter using a Blind Loss Aversion model of behavior, "extending" the Myopic Loss Aversion model of Benartzi and Thaler (1995); we review the latter model in detail in Section 3.5. They ask subjects to consider two lotteries, expressed here in equivalent dollars instead of Swiss Francs:

Lottery A: Win $4.50 with probability 1/2, lose $2.80 with probability 1/2. Otherwise get $0.
Lottery B: Play six independent repetitions of lottery A. Otherwise get $0.

Subjects could participate in both lotteries, neither, or either. Fehr and Goette (2007) assume that subjects have a linear utility function for stakes that are this small, relying on the theoretical arguments of Rabin (2000) rather than the data of Holt and Laury (2002) and others. They also assume that there is no probability weighting: even though Quiggin (1982; Section 4) viewed 1/2 as a plausible fixed point in probability weighting, most others have assumed or found otherwise. If one is blind to the effects of curvature of the utility function and probability weighting then the only thing left to explain choices over these lotteries is loss aversion. On the other hand, it becomes "heroic" to then extrapolate those estimates to explain behavior that one has elsewhere (p. 304) assumed to be characterized by stakes large enough that strictly concave utility is plausible *a priori*. Of course, the preferred model (p. 306) assumes away concavity and only uses the loss aversion parameter, but without explanation for why behavior over such stakes should be driven solely by loss aversion instead of risk attitudes more generally.[21]

Tanaka, Camerer, and Nguyen (2007) (TCN) propose a method to elicit risk and time preferences from individuals. They assume a certain parametric structure in their risk elicitation procedure, assuming Cumulative Prospect Theory (CPT): specifically, power Constant Relative Risk Aversion (CRRA) utility functions for gains and losses, and the one-parameter version of the Prelec (1998) probability weighting function. They further assume that the CRRA coefficient for gains and losses is the same. We consider these functional forms in detail in Sections 3.1 and 3.2. The upshot is they seek to elicit one parameter σ that controls the concavity or convexity of the utility function, one parameter α that controls the curvature of the probability weighting function, and one parameter λ that determines the degree of loss aversion. Their elicitation procedure for time preferences is completely separate conceptually from their elicitation procedure for risk attitudes, and is not used to infer anything about risk preferences.[22]

To elicit the first two parameters, σ and α, TCN ask subjects to consider three MPL sheets. The first sheet contains 14 options akin to those used in the Holt and Laury (2002) MPL procedure, shown in panel A of Table 1. The difference is that the probabilities of the high or low outcomes in each lottery stay constant from row to row, but the high prize in the "risky" lottery get larger and larger: the risky lottery start off in row 1 as "relatively risky" but with a relatively low expected value, and changes so that in the last row it becomes "extremely risky" but with a substantially higher expected value. The specific, fixed probabilities used are 0.3 for the high prize in the safe lottery and 0.1 for the high prize in the risky lottery. Subjects are asked to pick a switch point in this sheet, akin to the sMPL procedure of Andersen et al. (2006a); of course, this is just a monotonicity-enforcing variant of the basic MPL procedure of Holt and Laury (2002). So we can see that behavior in the first sheet elicits an interval for σ if we had ignored probability weighting, just as it elicited an interval for the CRRA coefficient in Holt and Laury (2002; Table 3, p. 1649). But with probability weighting allowed, all we can infer from this choice are combinations of intervals for σ and α. TCN indicate (p. 8, fn. 11) that the values of σ and α they report are actually "rounded mid-points" of the intervals. For example, one interval they infer is $0.65 < \sigma < 0.74$ and $0.66 < \alpha < 0.74$, and they round this to the values $\sigma = 0.7$ and $\alpha = 0.7$. They note in a footnote to Table A1 (p. 33) that the boundaries of the intervals are approximated to the nearest 0.05 increments. If subjects do not switch they use the approximate values at the last possible interval; in fact, the implied interval should have a finite value for a lower bound and ∞ for the upper bound, as noted by Coller and Williams (1999).[23] For their particular parameters there are seven such combinations of interval pairs.

The second sheet in the procedure of TCN is qualitatively the same as the first sheet, except that the probabilities of the high prize in each lottery are now 0.1 and 0.7. The specific prizes are different, but have the same structure as the first sheet. From the switching point in the second sheet one can derive another set of interval pairs for the parameters σ and α. The values for these intervals will be different than the intervals derived from the first sheet, because of differences in the value of the prizes and probabilities. By crossing the two sets of intervals one can reduce the implied intervals to the intersections from the two sheets. Since the prizes in these two sheets involve gains, the loss aversion parameter λ plays no role.

The third sheet in the procedure of TCN involves losses. There are seven options in which each lottery contains one positive prize and one negative prize, so these are "mixed lotteries." Probabilities of the high prize are fixed at 1/2 for all rows, and variations in three of the prizes occur from row to row. Conditional on a value of σ from responses to the first two sheets, the response in the third sheet implies an interval for λ. For example, if $\sigma = 0.2$ then somebody switching at, say, row 4 in the third sheet would have revealed a loss aversion parameter such that $1.88 < \lambda < 2.31$, but if $\sigma = 1$ then somebody switching at row 4 in the third sheet would have revealed a loss aversion parameter such that $1.71 < \lambda < 2.42$. The parameters for the third sheet were chosen, for a given observed response, so that the implied intervals for λ did not differ widely as σ varied over the expected range. Of course, the responses in the third sheet provides information on σ as well as λ. In other words, if one only observed responses from the third sheet there would be a number of interval pairs for σ and λ that could account for the data, just as there are a number of interval pairs of σ and α that could rationalize the observed response in the first or second sheet. So, the TCN procedure implicitly imposes a recursive estimation structure, so that σ is pinned down only from the responses in the first two sheets, and then the responses in the third sheet are used, conditional on some σ, to infer bounds for λ. This is a wily and parsimonious assumption, but might lead to different inferences than if one simply took all responses in these three sheets and simultaneously estimated σ, α, and λ, using ML methods discussed in Section 2.2.

The TCN procedure generates no information on standard errors of estimates, but such information would be provided automatically with the use of ML methods. Although the parameters they derive are conditional on the specific functional forms assumed, and in some cases (e.g., the third sheet) chosen to generate relatively robust inferences assuming those parameterizations, it should be possible to recover estimates for *some* minor variations in functional form (e.g., Constant Absolute Risk Aversion (CARA) instead of CRRA).

2. ESTIMATION PROCEDURES

Two broad methods of estimating risk attitudes have been used. One involves the calculation of bounds implied by the observed choices, typically using utility functions which only have a single-parameter to be inferred. The other involves the direct estimation by ML of some structural model of a latent choice process in which the core parameters defining risk attitudes can be estimated, in the manner pioneered by Camerer and Ho (1994; Section 6.1) and Hey and Orme (1994). The latter approach is particularly attractive for non-EUT specifications, where several core parameters combine to characterize risk attitudes. For example, one cannot characterize risk attitudes under Prospect Theory (PT) without making some statement about loss aversion and probability weighting, along with the curvature of the utility function. Thus, joint estimation of all parameters is a necessity for reliable statements about risk attitudes in such cases.

We first review examples of each approach (Sections 2.1 and 2.2), and then consider the role of stochastic errors (Section 2.3), the possibility of non-parametric estimation (Section 2.4), and a comparison of risk attitudes elicited from different procedures (Section 2.5), and treatments (Section 2.6). The exposition in this section focuses almost exclusively on EUT characterizations of risk attitudes. Alternative models are considered in Section 3.

2.1. Inferring Bounds

The HL data may be analyzed using a variety of statistical models. Each subject made 10 responses in each task, and typically made 30 responses over three different tasks. The responses in each task can be reduced to a scalar if one looks at the *lowest* row in panel A of Table 1 that the subject "switched" over to option B.[24] This reduces the response to a scalar for each subject and task, but a scalar that takes on integer values between 0 and 10. In fact, over 83% of their data takes on values of 4 through 7, and 94% takes on values between 3 and 8.

HL evaluate these data using ordinary least squares regression with the number of safe choices as the dependent variable, estimated on the sample generated by each task separately, and report univariate tests of demographic effects.[25] They also report semi-parametric tests of the number of safe choices with experimental condition as the sole control.

To study the effects of experimental conditions, while controlling for characteristics of the sample and the conduct of the experiment, one could employ an interval regression model, first proposed by Coller and Williams (1999) for an MPL experimental task (eliciting discount rates). The dependent variable in this analysis is the CRRA *interval* that each subject implicitly chose when they switched from option A to option B. For each row of panel A in Table 1, one can calculate the bounds on the CRRA coefficient that is implied, and these are in fact reported by Holt and Laury

(2002; Table 3). Thus, for example, a subject that made five safe choices and then switched to the risky alternatives would have revealed a CRRA interval between 0.15 and 0.41, and a subject that made seven safe choices would have revealed a CRRA interval between 0.68 and 0.97, and so on.[26] When we consider samples that pool responses over different tasks for the same individual, we would use a random effects panel interval regression model to allow for the correlation of responses from the same subject.

Using this panel interval regression model, we can control for all of the individual characteristics collected by HL, which includes sex, age, race (Black, Asian, or Hispanic), marital status, personal income, household income, household size, whether the individual is the primary household budget decision-maker, indicator of full-time employment, student status, faculty status, whether the person is a junior, senior, or graduate student, and whether the person has ever voted. In addition, dummy variables indicate specific sessions, and a separate indicator identifies those sessions conducted at Georgia State University. The treatment variables, of course, include the scale of payoffs (1, 20, 50, or 90), the order of the task (1, 2, 3, or 4), and the experimental income earned by the subject in task 3.

Table 2 presents ML estimates of this interval regression model. Since each subjects contributed several tasks, a random effects specification has been used to control for unobserved individual heterogeneity. One of the advantages of the use of inferred bounds for risk attitudes is that one can estimate detailed models such as in Table 2, since interval regression is a relatively stable statistical model, and a straightforward extension of ordinary least squares. It is also easy to correct for multiplicative heteroskedasticity using this estimation method, although that can introduce convergence problems as a practical matter. The main benefit of such an estimation is the ability to quickly ascertain treatment and demographic effects for the sample.

Consider first the question of order effects. Tasks 1 and 4 were identical in terms of the payoff scale, but differed because of their order and the fact that subjects had some experimental income from the immediately prior task 3. Controlling for that prior income, as well as other individual covariates, we find that there is an order effect: the CRRA coefficient increases by 0.16 in task 4 compared to task 1, and this is significant at the 2% level. Thus, order effects do seem to matter in these experiments, and in a direction that confound the inferences drawn about scale from the high-payoff treatments. There is also a significant scale effect, as seen for task 3 in Table 2, so the only way that one can ascertain the pure effect of order when there is a confounding change in scale, without such assumptions, would be

Table 2. Interval Regression Model of Responses in Holt and Laury Experiments[a].

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
scale5090	Payoffs scaled by 50 or 90	0.13	0.15	0.38	– 0.16	0.42
Task3	Third task	0.26	0.04	0.00	0.18	0.34
Task4	Fourth task	0.16	0.07	0.02	0.02	0.30
wealth	Wealth coming into the lottery choice	0.00	0.00	0.10	0.00	0.00
Sess2	Session B	– 0.18	0.20	0.37	– 0.58	0.21
Sess3	Session C	0.01	0.16	0.92	– 0.29	0.32
Sess4	Session D	– 0.16	0.20	0.43	– 0.54	0.23
Sess5	Session E	– 0.27	0.20	0.17	– 0.66	0.12
Sess6	Session F	– 0.14	0.15	0.34	– 0.44	0.15
Sess7	Session G	– 0.24	0.18	0.18	– 0.60	0.11
Sess8	Session H	– 0.45	0.20	0.02	– 0.84	– 0.06
Sess9	Session I	– 0.21	0.18	0.23	– 0.55	0.13
Sess10	Session J	– 0.31	0.18	0.08	– 0.67	0.04
Sess11	Session K	0.07	0.22	0.75	– 0.36	0.50
Sess13	Session M	0.10	0.21	0.62	– 0.31	0.52
female	Female	0.04	0.06	0.46	– 0.07	0.16
black	Black	0.05	0.16	0.75	– 0.26	0.36
asian	Asian	0.05	0.10	0.63	– 0.14	0.23
hispanic	Hispanic	– 0.39	0.12	0.00	– 0.62	– 0.16
age	Age	– 0.01	0.01	0.34	– 0.02	0.01
married	Ever married	0.12	0.09	0.18	– 0.06	0.30
Pinc2	Personal income between $5k and $15k	0.06	0.11	0.56	– 0.15	0.27
Pinc3	Personal income between $15k and $30k	– 0.14	0.11	0.24	– 0.36	0.09
Pinc4	Personal income above $30k	– 0.10	0.13	0.41	– 0.35	0.14
Hinc2	Household income between $5k and $15k	0.24	0.16	0.13	– 0.07	0.54
Hinc3	Household income between $15k and $30k	0.17	0.15	0.27	– 0.13	0.47
Hinc4	Household income between $30k and $45k	0.08	0.16	0.63	– 0.23	0.39
Hinc5	Household income between $45k and $100k	0.31	0.14	0.03	0.03	0.58
Hinc6	Household income over $100k	0.14	0.17	0.39	– 0.18	0.47
nhhd	Number in household	– 0.03	0.03	0.38	– 0.09	0.03

Table 2. (*Continued*)

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
decide	Primary household budget decision-maker	– 0.09	0.08	0.26	– 0.25	0.07
fulltime	Full time employment	0.15	0.10	0.16	– 0.06	0.35
student	Student	0.17	0.08	0.02	0.02	0.32
business	Business major	– 0.20	0.10	0.05	– 0.39	0.00
junior	Junior	– 0.16	0.13	0.23	– 0.41	0.10
senior	Senior	– 0.03	0.14	0.84	– 0.31	0.25
grad	Graduate student	0.18	0.15	0.22	– 0.11	0.46
faculty	Faculty	– 0.07	0.24	0.77	– 0.55	0.40
voter	Ever voted	– 0.01	0.07	0.86	– 0.15	0.12
gsu	Experiment at Georgia State University	– 0.40	0.22	0.07	– 0.83	0.03
Constant		0.63	0.27	0.02	0.10	1.15
σ_u	Standard deviation of random individual effect	0.29	0.03	0.00	0.24	0.34
σ_e	Standard deviation of residual	0.33	0.01	0.00	0.30	0.36

Notes: Log-likelihood value is – 838.24; Wald test for null hypothesis that all coefficients are zero has a χ^2 value of 118.44 with 40 degrees of freedom, implying a p-value less than 0.001; fraction of the total error variance due to random individual effects is estimated to be 0.433, with a standard error of 0.043.

[a]Random-effects interval regression. $N = 495$, based on 181 subjects from Holt and Laury (2002).

to modify the HL design and directly test for it. Harrison, Johnson, McInnes, and Rutström (2005b) provided such a test, and found that there were statistically significant order effects on risk attitudes; we consider their data below.

We observe no significant effect in Table 2 from sex: women are estimated to have a CRRA that is 0.04 higher than men, but the standard error of this estimate is 0.06. Hispanic subjects do have a statistically significant difference in risk attitudes: their CRRA is 0.39 lower on average, with a p-value of less than 0.001. Subjects with an annual household income that places them in the "upper middle class" (between $45,000 and $100,000) have a significantly higher CRRA that is 0.31 above the norm, with a p-value of 0.03. Students have a CRRA that is 0.17 higher on average (p-value = 0.02); the HL sample included faculty and staff in their experiments. Business majors were less risk averse on average, by about 0.20 (p-value = 0.05). There are some quantitatively large session effects, although only two sessions (H and J) have effects that are statistically significant in terms of the p-value. To preserve anonymity, the locations of these sessions apart from those at Georgia State University are confidential, so one can only detect individual session effects.

Fig. 6 shows the distribution of predicted CRRA coefficients from the interval regression model estimates of Table 2 from task 1 (top left panel) and task 3 (bottom left panel). The estimates for the high-payoff task 3 are only from those subjects that faced the payoffs that were scaled by a factor of 20. The average low-payoff CRRA is estimated to be 0.28, with a standard deviation of 0.20; the average high-payoff CRRA is estimated to be 0.54 with a standard deviation of 0.26. As Fig. 6 demonstrates, the distribution is normally shaped, with relatively few of the estimates exhibiting significant risk aversion above 0.9.

Harrison et al. (2005b) recruited 178 subjects from the University of South Carolina to participate in a series of non-computerized experiments using the MPL procedure of HL. Their design called for subjects to

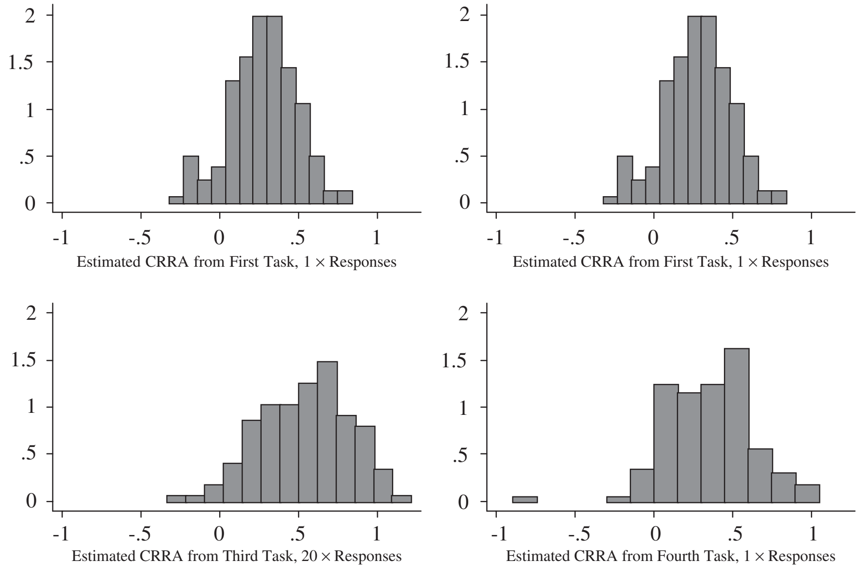


Fig. 6. Interval Regression Estimates of Risk Aversion From Holt and Laury (2002) Experiments (Fraction of the Sample, $N = 181$).

participate in either a 1× session, a 10× session, or a 1×10× session, where the "×" denotes the scalar applied to the basic payoffs used by HL in their 1× design (shown in panel A of Table 1). In the 1× session that is all that the subjects were asked to do; in the 10 × session they did one risk elicitation task but with payoffs scaled up by 10. In the 1×10× session subjects were asked to state their choices over 1× lotteries, and then given the opportunity to give up any earnings from that task and participate in a comparable 10× task. We examine the responses of the subjects in the 10× session and in the 10× part of the 1×10× session, with controls for whether their 10× responses were preceded by the 1× task or not. Table 3 reports the statistical analysis of these data, also using an interval regression model. Since each subject made only one 10× choice, no panel corrections are needed. The results show no significant effect from sex, and some effect from age, citizenship, and task order.

One limitation of this approach is that it assumes that all of the heterogeneity of the sample is captured by the individual characteristics measured by the experimenter. Although the socio-demographic questions typically used are relatively extensive, there is always some concern that there might be unobserved individual heterogeneity that could affect preferences towards risk. It is possible to undertake a statistical analysis of the responses of each individual, which implicitly controls for unobserved heterogeneity in the pooled analysis. However, the MPL design is not well suited to such an estimation task, even if it can be undertaken numerically, due to the small sample size for each individual. It is a simple matter to extend the HL design to have the subject consider several MPL tables for different lottery prizes, providing a richer data set with which to characterize individual risk attitudes (e.g., Harrison, Lau, & Rutström, 2007b). Apart from providing several interval responses per subject, such designs allow one to vary the prizes in the MPL design and pin down the latent CRRA more precisely by having overlapping intervals across tasks, as explained by Harrison et al. (2005d). Thus, if one task tells us that a given subject has a CRRA interval between 0.1 and 0.3, and another task tells us that the same subject has an interval between 0.2 and 0.4, we can infer a CRRA interval between 0.2 and 0.3 from the two tasks (with obvious assumptions about the absence of order effects, or some controls for them).

Another limitation of this approach, somewhat more fundamental, is that it restricts the analyst to utility functions that can characterize risk attitudes using one parameter. This is because one must infer the bounds that make the subject indifferent between the switch points, and such inferences become virtually incoherent statistically when there are two or more

Table 3. Interval Regression Model of Responses in Harrison, Johnson, McInnes, and Rutström Experiments[a].

Variable	Description	Estimate	Standard Error	*p*-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Female	Female	0.088	0.08	0.26	– 0.06	0.24
Black	Black	0.084	0.10	0.40	– 0.11	0.28
Age	Age in years	0.022	0.01	0.07	0.00	0.05
Business	Major is in business	– 0.043	0.07	0.56	– 0.19	0.10
Sophomore	Sophomore in college	– 0.068	0.11	0.54	– 0.29	0.15
Junior	Junior in college	– 0.035	0.12	0.77	– 0.27	0.20
Senior	Senior in college	– 0.023	0.13	0.85	– 0.27	0.22
GPAhi	High GPA (greater than 3.75)	0.004	0.09	0.97	– 0.18	0.19
GPAlow	Low GPA (below 3.24)	– 0.137	0.09	0.12	– 0.31	0.04
Graduate	Graduate student	0.034	0.16	0.83	– 0.27	0.34
EdExpect	Expect to complete a PhD or Professional Degree	– 0.119	0.09	0.18	– 0.29	0.05
EdFather	Father completed college	0.106	0.09	0.24	– 0.07	0.28
EdMother	Mother completed college	– 0.027	0.08	0.75	– 0.19	0.14
Citizen	U.S. citizen	0.234	0.12	0.05	0.00	0.47
Order	RA session 10× comes after 1×	0.166	0.08	0.03	0.01	0.32
	Constant	– 0.092	0.34	0.78	– 0.75	0.56

Notes: Log-likelihood value is – 290.2; Wald test for null hypothesis that all coefficients are zero has a χ^2 value of 18.36 with 15 degrees of freedom, implying a *p*-value of 0.244.

[a]All subjects facing 10× payoffs. $N = 178$ subjects from Harrison et al. (2005b).

parameters. Of course, for popular functions such as CRRA or CARA this is not an issue, but if one wants to move beyond those functions then there are problems. It is possible to devise one-parameter functional forms with more flexibility than CRRA or CARA in some dimension, as illustrated nicely by the one-parameter Expo-Power (EP) function developed by Abdellaoui, Barrios, & Wakker (2007a; Section 4). But in general we will need to move to structural modeling with ML to accommodate richer models, illustrated in Section 2.2.

We conclude that relatively consistent estimates of the CRRA coefficient of experimental subjects emerge from the HL experiments and the MPL design used in subsequent studies. There are, however, some apparent effects from task order, explored further in Harrison et al. (2005b) and Holt and Laury (2005). And there are significant limitations on the flexibility of the modeling of risk attitudes, pointing to the need for a complementary approach that allows structural estimation of latent models of choice under uncertainty.

2.2. Structural Estimation

Assume for the moment that utility of income is defined by

$$U(x) = \frac{x^{(1-r)}}{(1-r)} \tag{1}$$

where x is the lottery prize and $r \neq 1$ a parameter to be estimated. For $r = 1$, assume $U(x) = \ln(x)$ if needed. Thus, r is the coefficient of CRRA: $r = 0$ corresponds to RN, $r < 0$ to risk loving, and $r > 0$ to risk aversion. Let there be k possible outcomes in a lottery. Under EUT the probabilities for each outcome k, p_k, are those that are induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome in each lottery i:

$$\mathrm{EU}_i = \sum_{k=1,K} (p_k \times U_k) \tag{2}$$

The EU for each lottery pair is calculated for a candidate estimate of r, and the index

$$\nabla \mathrm{EU} = \mathrm{EU}_{\mathrm{R}} - \mathrm{EU}_{\mathrm{L}} \tag{3}$$

calculated, where EU_{L} is the "left" lottery and EU_{R} is the "right" lottery. This latent index, based on latent preferences, is then linked to the observed choices using a standard cumulative normal distribution function $\Phi(\nabla \mathrm{EU})$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1 using the function shown in Fig. 7. Thus, we have the probit link function,

$$\text{prob(choose lottery R)} = \Phi(\nabla \mathrm{EU}) \tag{4}$$

The logistic function is very similar, as illustrated in Fig. 7, and leads instead to the "logit" specification.

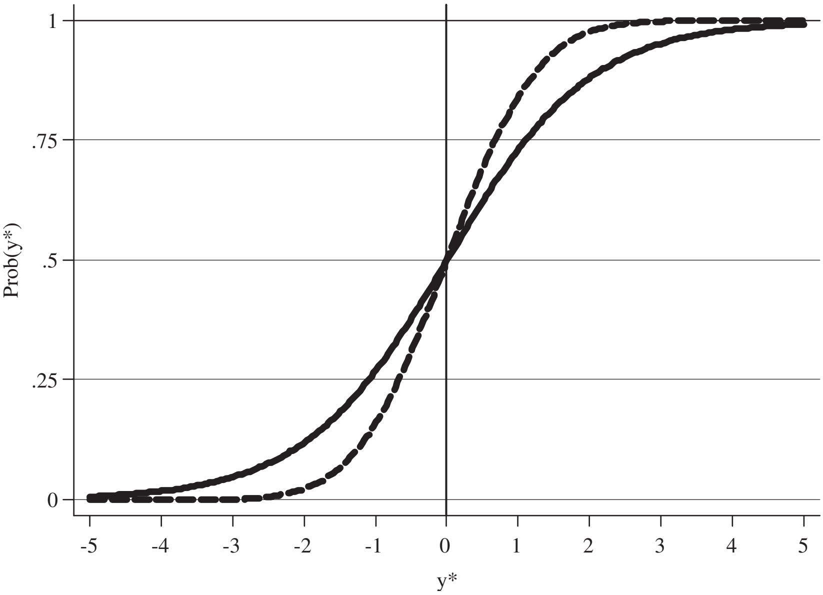


Fig. 7. Normal and Logistic Cumulative Density Functions (Dashed Line is Normal and Solid line is Logistic).

Even though Fig. 7 is common in econometrics texts, it is worth noting explicitly and understanding. It forms the critical statistical link between observed binary choices, the latent structure generating the index y^*, and the probability of that index y^* being observed. In our applications y^* refers to some function, such as Eq. (3), of the EU of two lotteries; or, later, the Prospective Utility (PU) of two lotteries. The index defined by Eq. (3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla\mathrm{EU}) > 1/2$, which is implied by Eq. (4).

Thus, the likelihood of the observed responses, conditional on the EUT and CRRA specifications being true, depends on the estimates of r given the above statistical specification and the observed choices. The ‘‘statistical specification’’ here includes assuming some functional form for the cumulative density function (CDF), such as one of the two shown in Fig. 7. If we ignore responses that reflect indifference for the moment the conditional log-likelihood would be

$$\ln \mathrm{L}(r; y, \mathbf{X}) = \sum_i ((\ln \Phi(\nabla\mathrm{EU}) | y_i = 1) + (\ln \Phi(1 - \nabla\mathrm{EU}) | y_i = -1)) \qquad (5)$$

where $y_i = 1(-1)$ denotes the choice of the Option R (L) lottery in risk aversion task i, and **X** is a vector of individual characteristics reflecting age, sex, race, and so on.

In most experiments the subjects are told at the outset that any expression of indifference would mean that if that choice was selected to be played out the experimenter would toss a fair coin to make the decision for them. Hence, one can modify the likelihood to take these responses into account by recognizing that such choices implied a 50:50 mixture of the likelihood of choosing either lottery:

$$\begin{aligned}\ln \mathrm{L}(r; y, \mathbf{X}) = \sum_i ((\ln \Phi(\nabla \mathrm{EU}) | y_i = 1) + (\ln \Phi(1 - \nabla \mathrm{EU}) | y_i = -1) \\ + (\ln(\tfrac{1}{2}\Phi(\nabla \mathrm{EU}) + \tfrac{1}{2}\Phi(1 - \nabla \mathrm{EU})) | y_i = 0))\end{aligned} \tag{5'}$$

where $y_i = 0$ denotes the choice of indifference. In our experience very few subjects choose the indifference option, but this formal statistical extension accommodates those responses.[27]

The latent index, Eq. (3), could have been written in a ratio form:

$$\nabla \mathrm{EU} = \frac{\mathrm{EU_R}}{(\mathrm{EU_R} + \mathrm{EU_L})} \tag{3'}$$

and then the latent index would already be in the form of a probability between 0 and 1, so we would not need to take the probit or logit transformation. We will see that this specification has also been used, with some modifications we discuss later, in HL.

Appendix F reviews procedures and syntax from the popular statistical package *Stata* that can be used to estimate structural models of this kind, as well as more complex models discussed later. The goal is to illustrate how experimental economists can write explicit ML routines that are specific to different structural choice models. It is a simple matter to correct for stratified survey responses, multiple responses from the same subject ("clustering"),[28] or heteroskedasticity, as needed, and those procedures are discussed in Appendix F.

Applying these methods to the data from the Hey and Orme (1994) experiments, one can obtain ML estimates of the core parameter r. Pooling all 200 of the responses from each subject over two sessions, and pooling over all subjects, we estimate $r = 0.66$ with a standard error of 0.04 assuming a normal CDF as in the dashed line in Fig. 7. These estimates correct for the clustering of responses by the same subject. If we instead assume a logistic CDF, as in the solid line in Fig. 7, we instead obtain an estimate $r = 0.80$ with a standard error of 0.04. This is not a significant economic difference, but it does point to the fact that parametric assumptions matter for estimation of risk attitudes using these methods. In particular, the choice of normal or logistic CDF is almost entirely arbitrary in this setting. One might apply some nested or non-nested hypothesis test to choose between specifications, but we will see that it is dangerous to rush into rejecting alternative specifications too quickly.

Extensions of the basic model are easy to implement, and this is the major attraction of the structural estimation approach. For example, one can easily extend the functional forms of utility to allow for varying degrees of RRA. Consider, as one important example, the EP utility function proposed by Saha (1993). Following Holt and Laury (2002), the EP function is defined as

$$U(x) = \frac{(1 - \exp(-\alpha x^{1-r}))}{\alpha} \tag{1'}$$

where α and r are parameters to be estimated. RRA is then $r + \alpha(1 - r)y^{1-r}$, so RRA varies with income if $\alpha \neq 0$. This function nests CRRA (as $\alpha \rightarrow 0$) and CARA (as $r \rightarrow 0$). We illustrate the use of this EP specification later.

It is also simple matter to generalize this ML analysis to allow the core parameter r to be a linear function of observable characteristics of the individual or task. In the HO experiments no demographic data were collected, but we can examine the effect of the subjects coming back for a second session by introducing a binary dummy variable (Task) for the second session. In this case, we extend the model to be $r = r_0 + r_1 \times \text{Task}$, where r_0 and r_1 are now the parameters to be estimated. In effect the prior model was to assume $r = r_0$ and just estimate r_0. This extension significantly enhances the attraction of structural ML estimation, particularly for responses pooled over different subjects, since one can condition estimates on observable characteristics of the task or subject. We illustrate the richness of this extension later. For now, we estimate $r_0 = 0.60$ and $r_1 = 0.10$, with standard errors of 0.04 and 0.02, respectively, using the probit specification. So there is some evidence of a session effect, with slightly greater risk aversion in the second session.

The effect of demographics and task can be examined using data generated by Harbaugh, Krause, and Vesterlund (2002). They examined lottery choices by a large number of individuals, varying in age between 5 and 64. Focusing on their lottery choices for dollars with individuals aged 19 and over, seven choices involved gambles in a gain frame, and seven over gambles in a loss frame. The loss frame experiments all involved subjects having some endowment up front, such that the loss was solely a framed loss, not a loss relative to the income they had coming into the session. In all cases the gamble was compared to a certain gain or loss, so these are relatively simple gambles to evaluate. The only demographic information included is age and sex, so we include those and interact them.[29] We also allow for quadratic effects of age.

Table 4 collects the estimates for models estimated separately on the choices made in the gain frame and choices made in the loss frame; later we

Table 4. Structural Maximum Likelihood Estimates of Risk Attitudes in Harbaugh, Krause, and Vesterlund Experiments[a].

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
A. Gain Domain						
Order2	Task order control	0.009	0.007	0.168	−0.004	0.023
Order3	Task order control	0.010	0.008	0.197	−0.005	0.026
Order4	Task order control	0.005	0.007	0.481	−0.009	0.019
Male	Male	0.016	0.029	0.594	−0.042	0.073
Age	Age in years	0.014	0.001	0.000	0.011	0.017
Age2	Age squared	0.000	0.000	0.000	0.000	0.000
Mage	Male × age	−0.001	0.002	0.776	−0.005	0.003
Mage2	Male × age2	0.000	0.000	0.852	0.000	0.000
Constant		0.476	0.021	0.000	0.434	0.517
B. Loss Domain						
Order2	Task order control	0.004	0.006	0.575	−0.009	0.016
Order3	Task order control	0.000	0.007	0.974	−0.013	0.014
Order4	Task order control	−0.005	0.007	0.494	−0.018	0.009
Male	Male	−0.030	0.024	0.205	−0.077	0.016
Age	Age in years	0.013	0.001	0.000	0.011	0.016
Age2	Age squared	0.000	0.000	0.000	0.000	0.000
Mage	Male × age	0.003	0.002	0.053	0.000	0.007
Mage2	Male × age2	0.000	0.000	0.026	0.000	0.000
Constant		0.483	0.016	0.000	0.452	0.514

Notes: Log-likelihood values are −8,070.56 in the gain domain, and −9,931.9 in the loss domain; Wald test for null hypothesis that all coefficients are zero has a χ^2 value of 339.2 with 8 degrees of freedom, implying a p-value less than 0.001 in the gain domain, and a value of 577.9 with 8 degrees of freedom in the loss domain.

[a] Maximum likelihood estimation of CRRA utility function using all pooled binary choices of adults. $N = 1092$, based on 156 adult subjects from Harbaugh et al. (2002).

consider the effect of assuming a model of loss aversion, rather than just viewing these as different frames.[30] There is virtually no effect from the loss frame, and in fact some evidence of a slight increase in risk aversion in that frame. The average of individual CRRA estimates is 0.476 in the gain frame, and is virtually identical in the loss frame. We find no evidence of a sex effect in the gain frame. The direct effect of sex is to change CRRA by 0.016, but this small effect has a p-value of 0.594 and a 95% confidence interval that easily spans zero. The joint effect of sex and age is also statistically insignificant: a test of the joint effect of sex and the sex–age interactions has a χ^2 value of 1.17, and with three degrees of freedom has a p-value of 0.761. Age has a significant effect on CRRA in the gain domain, at first increasing RRA and then eventually decreasing RRA as the individual gets older. The order dummies indicate no significant effect of task presentation order. There does appear to be an effect of sex on CRRA elicited in the loss frame. This effect is not direct, but is based on the interaction with age. Apart from the statistical significance of the individual interaction terms, a test that they are jointly zero has a χ^2 of 7.08 and a p-value of 0.069 with three degrees of freedom.

2.3. Stochastic Errors

An important extension of the core model is to allow for subjects to make some errors. The notion of error is one that has already been encountered in the form of the statistical assumption that the probability of choosing a lottery is not one when the EU of that lottery exceeds the EU of the other lottery. This assumption is clear in the use of a link function between the latent index ∇EU and the probability of picking one or other lottery; in the case of the normal CDF, this link function is $\Phi(\nabla \mathrm{EU})$ and is displayed in Fig. 7. If there were no errors from the perspective of EUT, this function would be a step function in Fig. 7: zero for all values of $y^* < 0$, anywhere between 0 and 1 for $y^* = 0$, and 1 for all values of $y^* > 0$. By varying the shape of the link function in Fig. 7, one can informally imagine subjects that are more sensitive to a given difference in the index ∇EU and subjects that are not so sensitive. Of course, such informal intuition is not strictly valid, since we can choose any scaling of utility for a given subject, but it is suggestive of the motivation for allowing for structural errors, and why we might want them to vary across subjects or task domains.

Consider the structural error specification used by HL, originally due to Luce. The EU for each lottery pair is calculated for candidate estimates of r, as explained above, and the ratio

$$\nabla \mathrm{EU} = \frac{\mathrm{EU}_{\mathrm{R}}^{1/\mu}}{(\mathrm{EU}_{\mathrm{L}}^{1/\mu} + EU_{\mathrm{R}}^{1/\mu})} \tag{3''}$$

calculated, where μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. The index ∇EU is in the form of a cumulative probability distribution function defined over differences in the EU of the two lotteries and the noise parameter μ. Thus, as $\mu \rightarrow 0$ this specification collapses to the deterministic choice EUT model, where the choice is strictly determined by the EU of the two lotteries; but as μ gets larger and larger the choice essentially becomes random. When $\mu = 1$, this specification collapses to Eq. (3′), where the probability of picking one lottery is given by the ratio of the EU of one lottery to the sum of the EU of both lotteries. Thus, μ can be viewed as a parameter that flattens out the link functions in Fig. 7 as it gets larger. This is just one of several different types of error story that could be used, and Wilcox (2008a, 2008b) provides masterful reviews of the implications of the alternatives.[31]

The use of this structural error parameter can be illustrated by a replication of the estimates provided by Holt and Laury (2002). Using the EP utility function in Eq. (1′), the Luce specification in Eq. (3″), and ignoring the fact that each subject made multiple binary choices, we estimate $r = 0.268$ and $\alpha = 0.028$ using the non-hypothetical data from HL. Panel A of Table 5 lists these estimates, which replicate the results reported by HL (p. 1653) almost exactly. Their estimates were obtained using optimization procedures in GAUSS, and did not calculate the likelihood at the level of the individual observation. Instead their data was aggregated according to the lottery choices in each row, and scaled up to reflect the correct sample size of observations. This approach works fine for a completely homogenous model in which one does not seek to estimate effects of individual characteristics or correct for unobserved heterogeneity at the level of the individual. But the approach adopted in our replication does operate at the level of the individual observation, so it is possible to make these extensions. In fact, allowing for unobserved individual heterogeneity does not affect these estimates greatly.

The role of the stochastic error assumption in Eq. (3″) can be evaluated by using Eq. (3′) instead, which is to assume that $\mu = 1$ in Eq. (3″). The effect, shown in panel B of Table 5, is to estimate more risk-loving behavior, with $r < 0$. Hence, at low levels of income subjects are now

Table 5. Structural Maximum Likelihood Estimates of Risk Attitudes in Holt and Laury Experiments[a].

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
A. Luce Error Specification and No Corrections for Clustering						
r	Utility function parameter	0.268	0.017	<0.001	0.234	0.302
α	Utility function parameter	0.028	0.002	<0.001	0.024	0.033
μ	Structural noise parameter	0.134	0.004	<0.001	0.125	0.143
B. No Luce Error Specification, No Corrections for Clustering						
r	Utility function parameter	–0.161	0.044	<0.001	–0.247	–0.074
α	Utility function parameter	0.015	0.003	<0.001	0.010	0.020
C. Probit Link Function, No Fechner Error Specification, Corrections for Clustering						
r	Utility function parameter	0.293	0.021	<0.001	0.251	0.334
α	Utility function parameter	0.038	0.003	<0.001	0.032	0.043
D. Probit Link Function, Fechner Error Specification, and Corrections for Clustering						
r	Utility function parameter	0.684	0.049	<0.001	0.589	0.780
α	Utility function parameter	0.045	0.059	0.452	–0.072	0.161
μ	Structural noise parameter	0.172	0.016	<0.001	0.140	0.203

[a] Maximum likelihood estimation of EP utility function using all pooled binary choices. $N = 3990$, based on 212 subjects from Holt and Laury (2002).

estimated to be risk loving. There is still evidence of Increasing Relative Risk Aversion (IRRA), with $\alpha > 0$. However, the log-likelihood of this specification is much worse than the original HL specification, and we can comfortably reject the null that $\mu = 1$. The point of this result is to demonstrate that the stochastic identifying restriction, to use the concept developed by Wilcox (2008a, 2008b), is not innocuous for inference about risk attitudes.

There is one other important error specification, due originally to Fechner and popularized by Hey and Orme (1994).[32] This error specification posits the latent index

$$\nabla \mathrm{EU} = \frac{(\mathrm{EU_R} - \mathrm{EU_L})}{\mu} \tag{3'''}$$

instead of Eq. (3), (3′), or (3″).

Wilcox (2008a) notes that as an analytical matter the evidence of IRRA in HL would be weaker, or perhaps even absent, if one had used a Fechner error specification instead of a Luce error specification. This important claim, that the evidence for IRRA may be an artifact of the (more or less arbitrary) stochastic identifying restriction assumed, can be tested with the HL data. The estimates in panels C and D of Table 5 confirm the claim of Wilcox (2008a). In panel C, we employ the probit link function Eq. (4) and the latent index function Eq. (3), and assume no Fechner error specification.[33] We confirm the original estimates of HL, with minor deviations: the path of estimated RRA in the left side of Fig. 9 mimics the original results from HL in Fig. 8. But when we add a Fechner error specification, in panel D of Table 5, we find striking evidence of CRRA over this prize domain. The path of RRA in this case is shown on the right side of Fig. 9, and provides a dramatic contrast to Fig. 8.

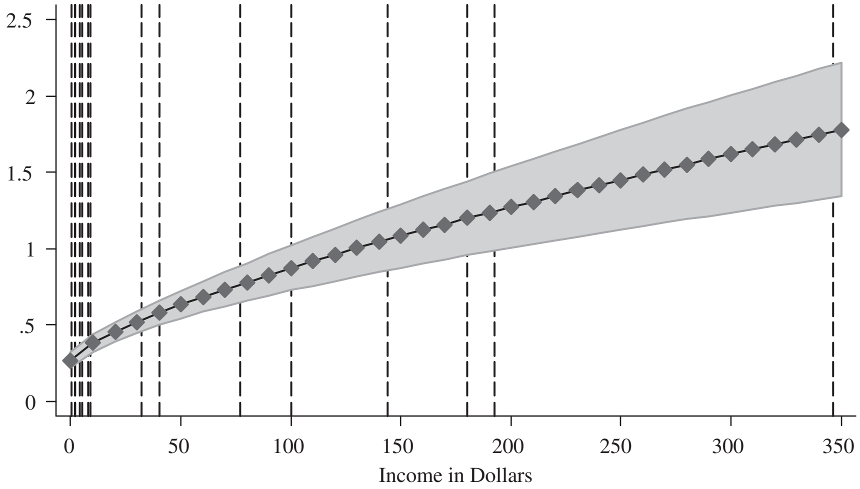


Fig. 8. Estimated Relative Risk Aversion Using the Holt–Laury Statistical Model. Estimated from Experimental Data of Holt & Laury (2002) Assuming Logit Likehood Function and Luce Noise.

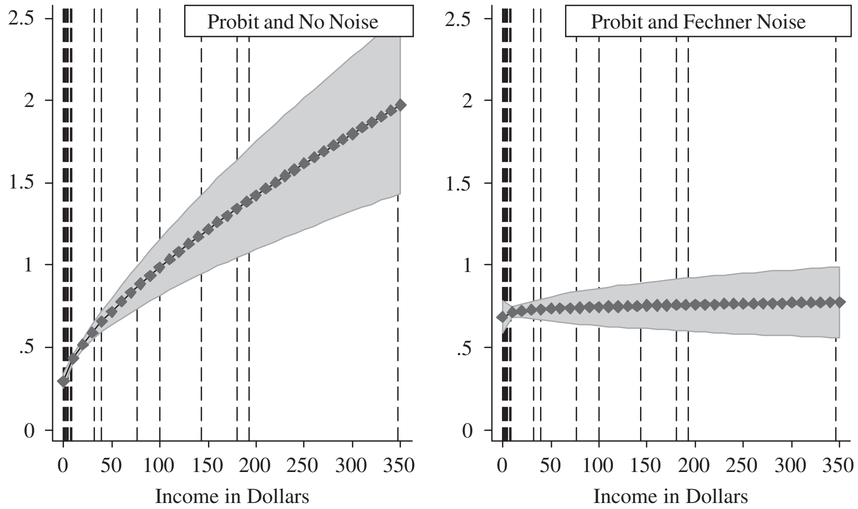


Fig. 9. Estimated Relative Risk Aversion with Expo-Power Utility and Fechner Noise. Estimated from Experimental Data of Holt and Laury (2002).

The log-likelihood of the Fechner specification is worse than the log-likelihood of the Luce specification. Since neither specification is nested in the other, a non-nested hypothesis test would seem to be called for. We reject the Fechner specification using either the Vuong (1989) test or the variant proposed by Clarke (2003). On the other hand, we prefer to avoid rejecting one specification out of hand just yet, since an alternative is to posit a latent data generating process in which two or more specifications have some validity. We return to consider this approach later.

2.4. Non-Parametric Estimation

It is possible to estimate the EUT model without assuming a functional form for utility, following Hey and Orme (1994). This approach works well for problem domains in which there are relatively few outcomes, since it involves estimation of one parameter for all but two of the outcomes. So if the task domain is constrained to just four outcomes, as in HO or HL, there are only two parameters to be estimated. But if the task domain spans many outcomes, these methods become inefficient and one must resort to a function defined by a few parameters, such as CRRA or EP utility functions.

To illustrate, we use the experimental data of HO, and then the replication of their experiments by Harrison and Rutström (2005). We also use the Fechner noise specification introduced above, to replicate the specification of HO. In HO there were only four monetary prizes of £0, £10, £20, and £30. We normalize to $u(£0) = 0$ and $u(£30) = 1$, and estimate $u(£10)$, $u(£20)$, and the noise parameter. As explained by HO, one could normalize the noise parameter to some fixed value and then estimate $u(£30)$ instead, but this choice of normalization seems the most natural. It is then possible to predict the values of the two estimated utilities: pooling over the two sessions and across subjects, we estimate $u(£10) = 0.66$ with a standard error of 0.02, and $u(£20) = 0.84$ with a standard error of 0.01, so $u(£0) < u(£10) < u(£20) < u(£30)$ as expected. The application of this estimation procedure in HO was at the level of the individual, which obviously allows variation in estimated utilities over individuals. This illustrative calculation does not.

The experiments of Harrison and Rutström (2005) were intended, in part, to replicate those of HO in the gain frame and additionally collect individual characteristics. In their case the prizes spanned \$0, \$5, \$10, and \$15. Employing the same non-parametric structure for this data as for the HO data above, the estimates are $u(\$5) = 0.60$ and $u(\$10) = 0.80$. In these data a set of demographic characteristics for each subject are known and we can therefore allow the estimated utilities to vary linearly with these characteristics. It is then possible to simply predict the estimated utilities, using the characteristics of each subject and the estimated coefficients on those characteristics, and plot them. Fig. 10 shows the distribution of estimated values. No subject had estimates that implied $u(\$10) < u(\$5)$.

2.5. Comparing Procedures

Do the various risk elicitation procedures imply essentially the same risk attitudes? In part this question requires that one agree on a standard way of representing lotteries, and that we understand the effect of those representations on elicited risk attitudes. It also requires that we agree on how to characterize risk attitudes statistically, and there are again many alternatives available in that direction that should be expected to affect inferred risk attitudes (Wilcox, 2008a). The older literature on utility elicitation was careful to undertake controlled comparisons of different procedures, as reviewed in Hershey, Kunreuther, and Schoemaker (1982) and illustrated by Hershey and Schoemaker (1985). But none of that

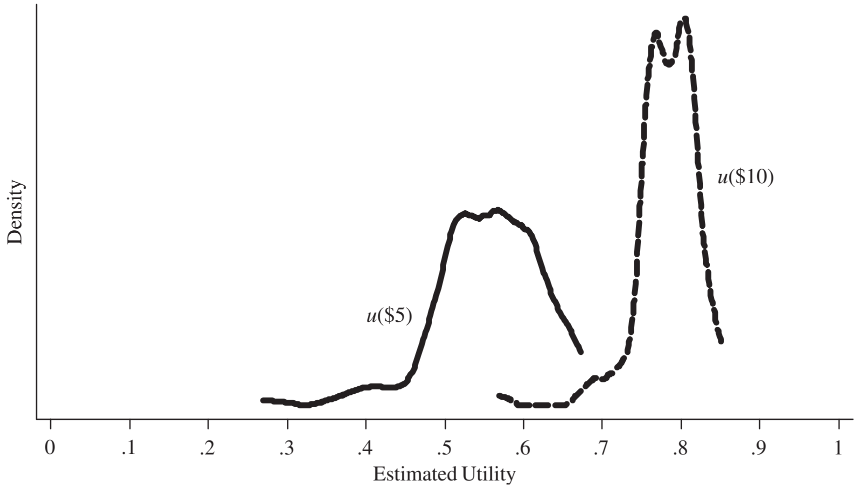


Fig. 10. Non-Parametric Estimates of Utility. (Assuming EUT and Normalized so that $u(\$0) = 0$ and $u(\$15) = 1$; Kernel Density of Predicted Utility Estimates for $N = 120$ Subjects; Data from Hey–Orme Replication of Harrison and Rutström (2005).)

literature seemed to be concerned with incentive compatibility and the effect of real rewards.

The striking counter-example, of course, is the preference reversal literature started for economists by Grether and Plott (1979), since they used methods for eliciting responses which were incentive compatible and they used real consequences to choices. And the phenomenon of preference reversals itself may be viewed as the claim that risk attitudes elicited from two procedures are not consistent, since the reversal is an "as if" change in risk attitudes when the elicitation mode changes. Unfortunately, the preference reversals in question involved a comparison of risk attitudes elicited with the RLP and BDM procedures, which both rely on strong assumptions to reliably elicit preferences.

It may therefore be useful to compare the three procedures that we do find attractive on *a priori* grounds: the MPL of Holt and Laury (2002), the RLP of Hey and Orme (1994), and the OLS of Binswanger (1980, 1981). Each procedure is applied to the same sample drawn from the same population: students at the University of Central Florida. In one session the MPL method was first and the OLS method last, in another session these orders were reversed, and the RLP method was always presented to subjects in between. The subjects learned what their payoffs were from each procedure at the end of the sequence of tasks for that procedure, so there is some potential in this design for income effects. There were 26 subjects in one session and 27 subjects in the second session, for a pooled sample of 53.

The parameters for the MPL procedure were scaled up by a factor of 10 to those used in the baseline experiments of Holt and Laury (2002), shown in panel A of Table 1. Thus, the prizes were \$1.00, \$16, \$20, and \$38.50. The parameters for the OLS procedure follow the broad pattern proposed by Binswanger (1980, 1981). The certain option offers \$10 whether a coin toss is heads or tails, and the next options offer \$19 or \$9, \$24 or \$8, \$25 or \$7, \$30 or \$6, \$32 or \$4, \$38 or \$2, and finally \$40 or \$0.[34] The RLP procedure used lotteries with probabilities and prizes that were each randomly drawn.[35] Each prize was randomly drawn from the uniform interval (\$0.01, \$15.00) in dollars and cents, and the number of prizes in each lottery pair was either 2, 3, or 4, also selected at random. For any lottery pair the cardinality of the outcomes was the same, so if one lottery had three prizes the other lottery would also have three prizes. The probabilities were also drawn at random, and represented to subjects to two decimal places. Each subject was given 60 pairs of lotteries to choose from, and three picked at random to be played out and paid. The expected value of each lottery was roughly \$7.50, with the expected value from the RLP procedure as a whole around \$22.70. Thus, the scale of prizes in the MPL and OLS procedures was virtually identical: up to \$38.50 and \$40, respectively. The scale of prizes in the RLP procedure was comparable: up to \$45 if all three selected lotteries generated an outcome of \$15 each.

In each case we estimate a CRRA model using Eq. (1). For the MPL and RLP procedures we use the probit link function, that is Eq. (4), defined over the difference in EU of the two lotteries for a candidate estimate of r and μ, and the Fechner error specification Eq. (3‴). For the OLS procedure we use the standard logit specification originally due to Luce (1959); McFadden (2001) reviews the starred history of this specification beautifully, and Train (2003) reviews modern developments. The EU for each lottery pair in this latter specification is calculated for a candidate estimate of r and μ, the exponential of the EU is taken as

$$eu_i = \exp(\mathrm{EU}_i^{1/\mu}) \tag{6}$$

and the index

$$\nabla \mathrm{EU}_i = \frac{eu_i}{(eu_1 + eu_2 + eu_3 + eu_4 + eu_5 + eu_6)} \tag{7}$$

calculated for each lottery *i*. This latent index, based on latent preferences, is in the form of a probability, and can therefore be directly linked to the observed choices; it is a multiple-lottery analogue of the Luce error specification Eq. (3″) for binary lottery choice.[36]

The results indicate consistency in the elicitation of risk attitudes, at least at the level of the inferred sample distribution. The point estimate (and 95% confidence intervals) for the MPL, RLP, and OLS procedures, respectively, are 0.75 (0.62, 0.88), 0.51 (0.42, 0.60), and 0.66 (0.44, 0.89). There is no significant order effect on the estimates from the OLS procedure: the estimates when it was first are 0.68 (0.43, 0.94), and when it was last they are 0.65 (0.25, 1.05). The 95% confidence intervals are wider in these estimates of the sub-samples, due to smaller samples. There is, however, a small but statistically significant order effect on the estimates from the MPL procedure: when it was first the CRRA estimate is 0.61 (0.46, 0.76) and when it was last the estimate is 0.86 (0.67, 1.05).

These results are suggestive that the procedures elicit roughly the same risk attitudes, apart from the sensitivity of the MPL procedure to order. Thus, one would tentatively conclude, based on the above analysis, that the procedures should be expected to generate roughly the same estimates of risk attitudes for a target population, and when used as the sole measuring instrument when used at the beginning of a session.[37]

A closely related issue is the temporal stability of risk preferences, even when one uses the same elicitation procedure. It is possible to define temporal stability of preferences in several different ways, reflecting alternative conceptual definitions and operational measures. Each definition has some validity for different inferential purposes.

Temporal stability of risk preferences can mean that subjects exhibit the same risk attitudes over time, or that their risk attitudes are a stable function of states of nature and opportunities that change over time. It is quite possible for risk preferences to be stable in both, either, or neither of these senses, depending on the view one adopts regarding the role preference stability takes in the theory. The temporal stability of risk preferences is one component of a broader set of issues that relate to the state-dependent approach to utility analysis.[38] This is a perfectly general approach, where the state of nature could be something as mundane as the weather or as fundamental as the individual's mortality risk. The states could also include the opportunities facing the individual, such as market prices and employment opportunities. Crucial to the approach, however, is the fact that all state realizations must be exogenous, or the model will not be identified and inferences about stability will be vacuous.

Problems arise, however, when one has to apply this approach empirically. Where does one draw the line in terms of the abstract "states of nature"? Many alleged violations of EUT amount to claims that a person behaved as if they had one risk preference for one lottery pair and another risk preference for a different lottery pair. Implicit in the claim that these are violations of EUT is the presumption that the differences in the two lottery pairs was not some state of nature over which preferences could be different.[39] Similarly, should we deem the preferences elicited with an open-ended auction procedure to be different from those elicited with a binary choice procedure, such as in the famous preference reversals of Grether and Plott (1979), because of some violation of EUT or just some change in the state of nature? Of course, it is a slippery inferential slope that allows "free parameters" to explain any empirical puzzle by shifting preferences. Such efforts have to be guided by direct evidence from external sources, lest they become open-ended specification searches.[40]

Several studies have begun to examine the temporal stability question. Limited exercises in laboratory settings are reported by Horowitz (1992) and Harrison, Johnson, McInnes, and Rutström (2005a), who demonstrate the temporal stability of risk attitudes in lab experiments over a period of up to 4 months. Horowitz (1992; p. 177) collects information on financial characteristics of the individual to control for changes in state of nature, but does not report if it changed the statistical inference about temporal stability. Harrison et al. (2005a) consider the temporal stability of risk attitudes in college students over a 4-week period, and do not control for changes in state of nature.

Andersen, Harrison, Lau, and Rutström (2008b) extend these simple designs in several ways. They use a much longer time span, control for changes in state of nature, use a stratified sample of a broader population, and report the results of a large-scale panel experiment undertaken in the field designed to examine this issue. Over a 17-month period they elicited risk preferences from subjects chosen to be representative of the adult Danish population. During this period many of the subjects were re-visited, and the same MPL risk aversion elicitation task repeated. In each visit information was also elicited on the individual characteristics of the subject, as well as their expectations about the state of their own economic situation and macroeconomic variables. The statistical analysis includes controls for changes in the subject's perceived states of nature, as well as the possible effects of endogenous sample selection into the re-test. There is evidence of some variation in risk attitudes over time, but there is no general tendency for risk attitudes to increase or decrease over a 17-month span.

Additionally, the small variation of risk attitudes over time is less prominent than variations across tasks and across individuals. The results also suggest that risk preferences are state contingent with respect to personal finances.

Of course, we could easily imagine target populations, such as the poor, that might be far less stable over time than the average adult Dane. There is some evidence from Dave, Eckel, Johnson, and Rojas (2007; Table 7) that the MPL instrument might exhibit some drift over time in such a population: estimated RRA increases by 0.12 compared to a baseline of 0.71, but the *p*-value of this change is 0.14, so it is not statistically significant. The real contribution of these studies is a systematic methodology for examining the issue of temporal stability with longitudinal experiments.

2.6. Comparing Treatments

The use of structural estimation of latent choice models also allows one to compare experimental treatments in terms of their effect on core parameters. Thus, we can answer questions such as "does treatment X affect risk attitudes" by directly estimating the effect on parameters determining risk attitudes, rather than relying on less direct measures of that effect. The value of inferences of this kind become more important when we allow for various parameters and processes to affect choice under uncertainty, such as when we consider rank-dependent preferences and/or sign-dependent preferences in Section 3.

To illustrate, consider the effect of providing information to subjects about the EV of lotteries they are to choose from. For simple, binary-outcome lotteries one often observes some subjects actually trying to do this arithmetic themselves on scrap paper, whether or not they then use that to decide which lottery to accept without adding or subtracting a risk premium. But when the cardinality of outcomes exceeds two, virtually all subjects tend to give up on those efforts to calculate EV. This raises the hypothesis that elicited risk attitudes might reflect underlying preferences or the interaction of those preferences and cognitive constraints on applying them to a particular lottery (if one assumes, for now, that subjects apply them the way economists theorize about them).

A direct measure of the effect of providing EV can be obtained by running these treatments and then estimating a model in which the treatment acts as a binary dummy on a core parameter of the latent structural model. For data we use the replication of the RLP procedures of Hey and Orme (1994) reported in Appendix B. These tasks were only over the gain frame;

63 subjects received no information over 60 binary choices, and 25 different subjects received information. For the structural model, we assume a CRRA power utility function, a Fechner error specification, and a probit linking function. If we introduce the binary dummy variable Info to capture those choices made under the treatment condition, we can estimate $r = r_0 + r_1 \times \text{Info}$ and directly assess the effect on risk attitudes by the sign and statistical significance of the coefficient r_1. It is also possible to allow for heteroskedasticity in the Fechner noise term, by estimating $\mu = \mu_0 + \mu_1 \times \text{Info}$ and examining the estimate of μ_1. Thus, we allow for the possibility that providing information on EV might not change risk attitudes, but might change the precision with which the subject makes choices given a latent preference for one lottery over the other.

The estimation results show that there is indeed a statistically significant effect on elicited risk attitudes from providing the EV of each lottery. The power function coefficient r increases by 0.15 from 0.47, which indicates a reduction in risk aversion towards RN. The p-value on the hypothesis test that this effect is zero is only 0.016, and the 95% confidence interval on the effect is between 0.03 and 0.28. So we conclude that there does appear to be a significant influence on elicited risk attitudes from providing information on EV. Whether this reflects better estimates of true preferences due to removing the confound of the cognitive burden of calculating EV, or reflects a simple anchoring response, cannot be determined. The point is that we can report the effect of the treatment in terms of its effect on the metric of interest, the core risk aversion parameter. In this specification there is no statistically significant effect on the Fechner noise parameter. Nor is there an effect on these conclusions from also controlling for the heterogeneity in preferences attributable to observed individual demographic effects.

3. EXTENSIONS AND FURTHER APPLICATIONS

We elicit risk attitudes to make inferences about different things. Obviously there is interest in the characterization of risk attitudes in general, and the previous section reviewed the estimation issues that arise under EUT. It is also important to consider the characterization of risk attitudes under alternatives to EUT. We consider the class of rank-dependent models due to Quiggin (1982) (Section 3.1), and then the class of sign-dependent models due to Kahneman and Tversky (1979) (Section 3.2). The implications for allowing several latent data generating processes to characterize risk attitudes are then considered (Section 3.3), concluding with a plea to avoid the assumption that there is one true model.

Risk attitudes also constitute a fundamental confound to inferences about behavior in stochastic settings, and it is here that we believe that the major payoff to better experimental controls for risk attitudes will be seen. We consider three major areas of investigation in which controls for risk should play a more significant role: identification of discount rates (Section 3.4), tests of EUT against competing models (Section 3.5), and tests of bidding behavior in auctions (Section 3.6). We also consider tests of a model of choice behavior that has radical implications for how one might think about risk aversion, Myopic Loss Aversion (Section 3.7). Finally, we consider the implications of the random lottery incentive procedure for risk elicitation (Section 3.8), and present some summary estimates using comparable modeling assumptions and designs that we believe to be the most reliable (Section 3.9).

3.1. Characterizing Risk Attitudes with Probability Weighting and Rank-Dependent Utility

One route of departure from EUT has been to allow preferences to depend on the rank of the final outcome through probability weighting. The idea that one could use non-linear transformations of the probabilities as a lottery when weighting outcomes, instead of non-linear transformations of the outcome into utility, was most sharply presented by Yaari (1987). To illustrate the point clearly, he assumed a linear utility function, in effect ruling out any risk aversion or risk seeking from the shape of the utility function *per se*. Instead, concave (convex) probability weighting functions would imply risk seeking (risk aversion).[41] It was possible for a given decision-maker to have a probability weighting function with both concave and convex components, and the conventional wisdom held that it was concave for smaller probabilities and convex for larger probabilities.

The idea of rank-dependent preferences had two important precursors.[42] In economics, Quiggin (1982, 1993) had formally presented the general case in which one allowed for subjective probability weighting in a rank-dependent manner and allowed non-linear utility functions. This branch of the family tree of choice models has become known as Rank-Dependent Utility (RDU). The Yaari (1987) model can be seen as a pedagogically important special case, and can be called Rank-Dependent Expected Value (RDEV). The other precursor, in psychology, is Lopes (1984). Her concern was motivated by clear preferences that experimental subjects exhibited for lotteries with the same expected value but alternative shapes of probabilities, as well as the verbal protocols those subjects provided as a possible indicator of their latent decision processes.

Formally, to calculate decision weights under RDU one replaces expected utility

$$\mathrm{EU}_i = \sum_{k=1,K} (p_k \times U_k) \tag{2}$$

with RDU

$$\mathrm{RDU}_i = \sum_{k=1,K} (w_k \times U_k) \tag{2$'$}$$

where

$$w_i = \omega(p_i + \ldots + p_n) - \omega(p_{i+1} + \ldots + p_n) \tag{8a}$$

for $i = 1, \ldots, n-1$, and

$$w_i = \omega(p_i) \tag{8b}$$

for $i = n$, the subscript indicates outcomes ranked from worst to best, and where $\omega(p)$ is some probability weighting function.

In the RDU model we have to define risk aversion in terms of the properties of the utility function and the probability weighting function, since both can affect risk attitudes. However, one can define conditional orderings, following Chew, Karni, and Safra (1987) and others, by considering the effects of more or less concave utility functions given a probability weighting function, and *vice versa*. Similarly, when we consider sign-dependent preferences in Section 3.2 the notion of risk aversion must include the effects of the sign of outcomes (e.g., possible loss aversion).

Picking the right probability weighting function is obviously important for RDU specifications. A weighting function proposed by Tversky and Kahneman (1992) has been widely used. It is assumed to have well-behaved endpoints such that $\omega(0) = 0$ and $\omega(1) = 1$ and to imply weights

$$\omega(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}} \tag{9}$$

for $0<p<1$. The normal assumption, backed by a substantial amount of evidence reviewed by Gonzalez and Wu (1999), is that $0<\gamma<1$. This gives the weighting function an “inverse S-shape,” characterized by a concave section signifying the overweighting of small probabilities up to a crossover-point where $\omega(\mathrm{p}) = \mathrm{p}$, beyond which there is then a convex section signifying underweighting. Under the RDU assumption about how these *probability* weights get converted into *decision* weights, $\gamma < 1$ implies overweighting of extreme outcomes. Thus, the probability associated with an outcome does not directly inform one about the decision weight of that outcome. If $\gamma > 1$ the function takes the less conventional "S-shape," with convexity for smaller probabilities and concavity for larger probabilities.[43] Under RDU $\gamma > 1$ implies *under*weighting of extreme outcomes.

We illustrate the effects of allowing for probability weighting using the experimental data from Holt and Laury (2005). We assume the EP functional form

$$U(x) = \frac{(1 - \exp(-\alpha x^{1-\rho}))}{\alpha} \tag{1''}$$

for utility. The remainder of the econometric specification is the same as for the EUT model with Luce error μ, generating

$$\nabla \mathrm{RDU} = \frac{\mathrm{RDU}_{\mathrm{R}}^{1/\mu}}{(\mathrm{RDU}_{\mathrm{L}}^{1/\mu} + \mathrm{RDU}_{\mathrm{R}}^{1/\mu})} \tag{3''''}$$

instead of Eq. (3‴). The conditional log-likelihood, ignoring indifference, becomes

$$\begin{aligned} \ln \mathrm{L}^{\mathrm{RDU}}(\rho, \gamma, \mu; y, X) = \sum_i l_i^{\mathrm{RDU}} = \sum_i ((\ln \Phi(\nabla \mathrm{RDU}) | y_i = 1) \\ + (\ln(1 - \Phi(\nabla \mathrm{RDU})) | y_i = 0)) \end{aligned} \tag{5''}$$

and requires the estimation of ρ, γ, and μ.

For RDEV one replaces Eq. (2′) with a specification that weights the prizes themselves, rather than the utility of the prizes:

$$\mathrm{RDEV}_i = \sum_{k=1,K} (\omega_k \times m_k) \tag{2''}$$

where m_k is the kth monetary prize. In effect, the RDEV specification is a special case of RDU.

The experimental data from Holt and Laury (2005) consists of 96 subjects facing their 1× condition or their 20× condition on a between-subjects basis.[44] The final monetary prizes ranged from a low of \$0.10 up to \$77. We only consider data in which subjects faced real rewards. Replicating their EUT statistical model, and allowing for clustering of responses, we estimate

$r = 0.40$ with a standard error of 0.07, and $\alpha = 0.076$ with a standard error of 0.02, closely tracking the estimates from Holt and Laury (2002). In particular, there is evidence of increasing RRA over this income domain.

When we estimate the RDU model using these data and specification, we find clear evidence of probability weighting. The estimate of γ is 0.37 with a standard error of 0.16, so we can easily reject the hypothesis that $\gamma = 1$ and that there is no probability weighting. Thus, we observe the conventional qualitative shape of the probability weighting function, an inverse S-shape. The effect of allowing for probability weighting is to lower the estimates of the curvature of the utility function – but we should be careful here not to associate curvature of the utility function with risk aversion. The risk aversion parameter ρ is estimated to be 0.26 and the α parameter to be 0.02, with standard errors of 0.05 and 0.012, respectively. Thus, there is some evidence for increasing curvature of the utility function as income increases ($\alpha > 0$), but it is not statistically significant (p-value of 0.16 that $\alpha = 0$). Fig. 11 displays the "relative risk aversion" associated with the curvature of the utility function, and the shape of the probability weighting function. Of course, RRA should actually be defined here in terms of both the curvature of the utility function and the effect of probability weighting, so the coefficients are not directly comparable to the EUT model. Nevertheless,

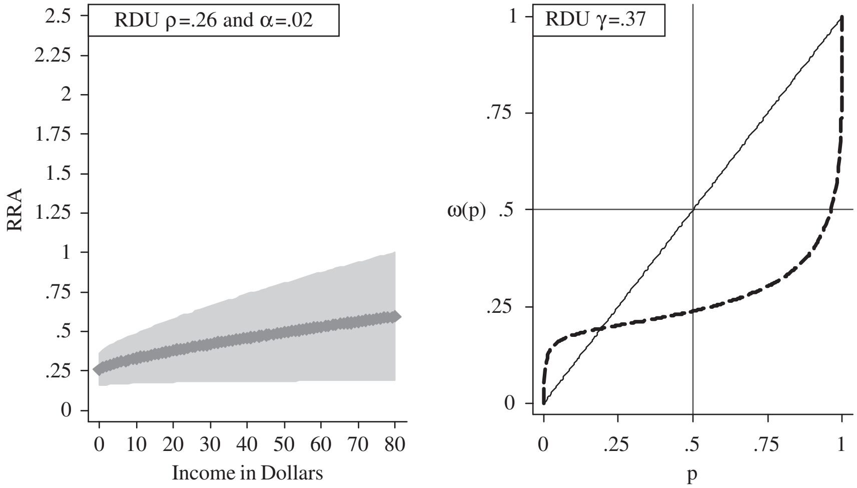


Fig. 11. Probability Weighting in Holt and Laury Risk Elicitation Task. RDU Parameters Estimated with $N = 96$ Subjects from Experimental Data of Holt and Laury (2005).

we can clearly say that inferences about increasing RRA depends on the assumptions one makes about probability weighting.

3.2. Characterizing Risk Attitudes with Loss Aversion and Sign-Dependent Utility

3.2.1. Original Prospect Theory

Kahneman and Tversky (1979) introduced the notion of sign-dependent preferences, stressing the role of the reference point when evaluating lotteries. In various forms, as we will see, PT has become the most popular alternative to EUT. Original Prospect Theory (OPT) departs from EUT in three major ways: (a) allowance for subjective probability weighting; (b) allowance for a reference point defined over outcomes, and the use of different utility functions for gains or losses; and (c) allowance for loss aversion, the notion that the disutility of losses weighs more heavily than the utility of comparable gains.

The first step is probability weighting, of the form $\omega(p)$ defined in Eq. (10), for example. One of the central assumptions of OPT, differentiating it from later variants of PT, is that $w(p) = \omega(p)$, so that the transformed probabilities given by $\omega(p)$ are directly used to evaluate PU:

$$\mathrm{PU}_i = \sum_{k=1,K} (\omega_k \times u_k) \tag{2'''}$$

The second step in OPT is to define a reference point so that one can identify outcomes as gains or losses. Let the reference point be given by χ for a given subject in a given choice. Consistent with the functional forms widely used in PT, we again use the CRRA functional form

$$u(m) = \frac{m^{1-\alpha}}{(1-\alpha)} \tag{1'''}$$

when $m \geq \chi$, and

$$u(m) = -\lambda \frac{(-m)^{1-\alpha}}{(1-\alpha)} \tag{1''''}$$

when $m < \chi$, and where λ is the loss aversion parameter. We use the same exponent α for the utility functions defined over gains and losses, even though the original statements of PT keep them theoretically distinct. Köbberling and Wakker (2005; Section 7) point out that this constraint is needed to identify the degree of loss aversion if one uses CRRA functional forms and does not want to make other strong assumptions (e.g., that utility is measurable only on a ratio scale).[45] Although λ is free in principle to be less than 1 or greater than 1, most PT analysts presume that $\lambda \geq 1$.

The specification of the reference point is critical to PT, and is discussed in Section 3.2.3. One issue is that it influences the nature of subjective probability weighting assumed, since different weights are allowed for gains and losses. Thus, we can again specify

$$\omega(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \tag{9}$$

for gains, but

$$\omega(p) = \frac{p^{\phi}}{(p^{\phi} + (1-p)^{\phi})^{1/\phi}} \tag{9$'$}$$

for losses. It is common in empirical applications to assume $\gamma = \phi$.

The remainder of the econometric specification would be the same as for EUT and RDU models. The latent index can be defined in the same manner, and the conditional log-likelihood defined comparably. Estimation of the core parameters α, λ, γ, ϕ, and μ is required.

The primary logical problem with OPT was that it implied violations of stochastic dominance. Whenever $\gamma \neq 1$ or $\phi \neq 1$, it is possible to find non-degenerate lotteries such that one lottery would stochastically dominate the other, but would be assigned a lower PU. Examples arise quickly when one recognizes that $\gamma(p_1 + p_2) \neq \gamma(p_1) + \gamma(p_2)$ for some p_1 and p_2. Kahneman and Tversky (1979) dealt with this problem by assuming that evaluation using OPT only occurred after dominated lotteries were eliminated. For specifications such as the one discussed here there is no modeling of an editing phase, but the stochastic error term μ could be interpreted as a reduced-form proxy for that editing process.[46] We do not provide any illustrative estimations of this model but move straight to the extensions provided by CPT.

3.2.2. Cumulative Prospect Theory

The notion of rank-dependent decision weights was incorporated into OPT by Starmer and Sugden (1989), Luce and Fishburn (1991), and Tversky and Kahneman (1992). Instead of implicitly assuming that $w(p) = \omega(p)$, it allowed $w(p)$ to be defined as in the RDU specification given by Eqs. (8a) and (8b). The sign-dependence of subjective probability weighting in OPT, leading to the estimation of different probability weighting functions, Eqs. (9) and (9′), for gains and losses, is maintained in CPT. Thus, there is a separate decumulative function used for gains and losses, but otherwise the logic is the same as for RDU.[47]

The estimation of a structural CPT model can be illustrated with data from the Harrison and Rutström (2005) replication and extension of the Hey and Orme (1994) RLP procedure. As explained in Appendix B, they had some subjects face lotteries defined over a gain frame, some face lotteries defined over a loss frame, and some face lotteries defined over a mixed gain–loss frame. In the mixed frame some prizes in a lottery were gains, and some were losses. In each case the subjects were endowed with cash to ensure that final outcomes were either exactly or approximately the same across frames.

Table 6 displays the ML estimates of the core parameters, and Fig. 12 displays the distributions over individuals of predicted values for each parameter. In each case the utility function is the CRRA power specification, a Fechner error story is included with a probit link function, and μ is a linear function of the same observable characteristics as every other parameter (Table 6 does not show the estimate for μ). The distribution of estimates of α are consistent with concave utility functions over gains and convex utility functions over losses, as expected. The estimates of γ are also consistent with expectations of an inverse S-shaped probability weighting function, implying greater decision weights on extreme prizes within each lottery. However, the estimates of λ are not at all consistent with loss aversion, and in fact suggest a clear tendency towards loss seeking. We reconsider the sensitivity of estimates of λ to the assumed reference point in more detail below.

Table 6 shows that there are some systematic effects of observable demographics on the EUT and CPT parameter estimates. Under EUT there is a slight effect from sex, with women being more risk averse, but it is not statistically significant. Similarly, ethnic characteristics show a large effect on risk attitudes, but they are not statistically significant. The only characteristic that has a statistically significant effect on risk attitudes under EUT is age, which is here shown in deviations from age 20. So every extra year leads to reduction in risk aversion. For completeness, we also estimate RDU on these data, not shown in Table 6, and find the curvature of the utility function similar to that of EUT, contrary to the estimates discussed above for the data of Holt and Laury (2005). For the RDU model the data here indicate a significant sex effect, with women being more risk averse

Table 6. Maximum Likelihood Estimates for EUT and CPT Models.

Parameter	Variable	Point Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
A. EUT Model (log-likelihood = − 7,665.0)						
r	Constant	0.952	0.149	0.00	0.66	1.24
	Female	− 0.133	0.094	0.16	− 0.32	0.05
	Black	− 0.138	0.133	0.30	− 0.40	0.12
	Hispanic	− 0.195	0.127	0.13	− 0.44	0.05
	Age (compared to 20)	0.039	0.009	0.00	0.02	0.06
	Major is in business	− 0.107	0.135	0.43	− 0.37	0.16
	Low GPA (below 3.24)	0.061	0.121	0.61	− 0.18	0.30
B. CPT Model (log-likelihood = − 7,425.5)						
α	Constant	0.761	0.079	0.00	0.61	0.91
	Female	− 0.160	0.109	0.14	− 0.37	0.05
	Black	− 0.132	0.277	0.63	− 0.67	0.41
	Hispanic	− 0.358	0.192	0.06	− 0.73	0.02
	Age (compared to 20)	0.017	0.009	0.07	0.00	0.04
	Major is in business	− 0.037	0.097	0.70	− 0.23	0.15
	Low GPA (below 3.24)	0.036	0.093	0.69	− 0.14	0.22
γ	Constant	1.017	0.061	0.00	0.89	1.14
	Female	− 0.050	0.074	0.49	− 0.20	0.09
	Black	− 0.300	0.133	0.02	− 0.56	− 0.04
	Hispanic	− 0.092	0.142	0.51	− 0.37	0.18
	Age (compared to 20)	− 0.001	0.004	0.75	− 0.01	0.01
	Major is in business	− 0.021	0.075	0.78	− 0.17	0.13
	Low GPA (below 3.24)	− 0.066	0.070	0.35	− 0.20	0.07
λ	Constant	0.447	0.207	0.03	0.04	0.85
	Female	0.432	0.416	0.30	− 0.38	1.25
	Black	0.233	1.062	0.83	− 1.85	2.31
	Hispanic	− 0.386	0.386	0.32	− 1.14	0.37
	Age (compared to 20)	0.033	0.018	0.08	0.00	0.07
	Major is in business	0.028	0.240	0.91	− 0.44	0.49
	Low GPA (below 3.24)	0.057	0.238	0.81	− 0.41	0.52

(− 0.09, p-value = 0.02), as well as Hispanics (− 0.17, p-value = 0.009). In addition, age has the same effect as under EUT. Although the extent of probability weighting is slight, and overall curvature of the utility function matches EUT, there are therefore some significant changes in the composition of the curvature of utility across the sample.

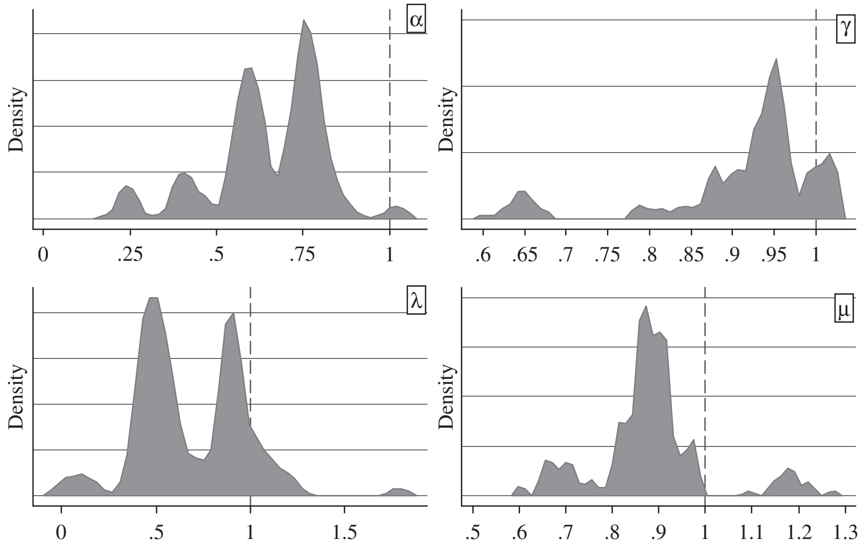


Fig. 12. Estimates of the Structural CPT Model. (Data from Hey–Orme Replication of Harrison and Rutström (2005); $N = 207$ Subjects: 63 Gain Frame, 57 Loss Frame, and 87 Mixed Frame.)

The CPT estimates in Table 6 also show some demographic effects on the composition of the curvature of utility across the sample. There is now a large and statistically significant effect from being Hispanic, in addition to a comparable age effect. The only characteristic that significantly affects the extent of probability weighting is whether the subject is Black, and it is a large effect. The effects on loss aversion appear to be poorly estimated, which of course may just be a reflection that this is not a stable parameter in terms of its effect, at least as currently modeled. Although these were static tasks, in the sense that there was no accumulation of earnings, subjects may have been adjusting their reference point during the 60 binary choices in some unspecified manner.

Finally, Fig. 13 collates estimates of the curvature of the utility function for these data using the three major alternative models of choice. In the top panel we include an EUT specification assuming the CRRA power utility function with parameter r. In the bottom left panel we estimate an RDU model with utility function parameter ρ, and that allows for rank-dependent probability weighting. The EUT and RDU models are estimated on the choices made in the loss frame, but with the actual net gain amount included

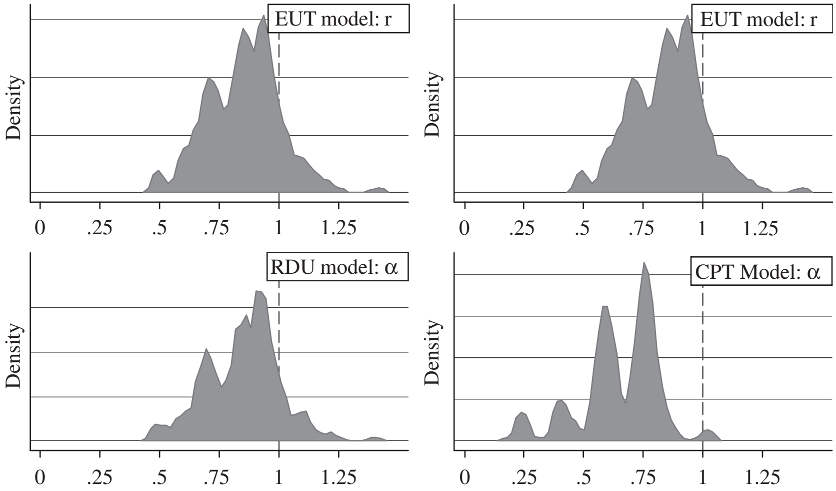


Fig. 13. Estimates of Curvature of Utility Function. (Data from Hey–Orme Replication of Harrison and Rutström (2005); $N = 207$ Subjects: 63 Gain Frame, 57 Loss Frame, and 87 Mixed Frame; Prizes for EUT and RDU Include Endowment.)

in the utility function.[48] In the bottom-right panel, we reproduce the estimate of α from Fig. 12, scaled to the EUT estimate above it for comparability. We see evidence that the RDU specification does not change the inferences we make about the curvature of the utility function significantly in comparison to EUT, so risk aversion here is not reflected in a transformation of probabilities. The CPT specification, which adds sign-dependence to utility, does result in a shift towards greater concavity of the utility function for gains, and more distinct modes reflecting a greater heterogeneity in preferences. Of course, curvature of the utility function under RDU and CPT is not the same as aversion to risk, but it is nonetheless useful to compare the implied shapes of the utility function.

3.2.3. The Reference Point and Loss Aversion

It is essential to take a structural perspective when estimating CPT models. Estimates of the loss aversion parameter depend intimately on the assumed reference point, as one would expect since the latter determines what are to be viewed as losses. So if we have assumed the wrong reference point, we will not reliably estimate the degree of loss aversion. However, if we do not get loss aversion leaping out at us when we make a natural assumption about the reference point, should we infer that there is no loss aversion or that there is loss aversion and we just used the wrong reference point? This question points to a key operational weakness of CPT: the need to specify what the reference point is. Loss aversion *may* be present for *some* reference point, but if it is not present for the one we used, and none others are "obviously" better, then should one keep searching for some reference point that generates loss aversion? Without a convincing argument about the correct reference point, and evidence for loss aversion conditional on that reference point, one simply cannot claim that loss aversion is always present. This specification ambiguity is arguably less severe in the lab, where one can frame tasks to try to induce a loss frame, but is a particularly serious issue in the field.

Similarly, estimates of the nature of probability weighting vary with changes in reference points, loss aversion parameters, and the concavity of the utility function, and *vice versa*. All of this is to be expected from the CPT model, but necessitates joint econometric estimation of these parameters if one is to be able to make consistent statements about behavior.

In many laboratory experiments it is simply assumed that the manner in which the task is framed to the subject defines the reference point that the subject uses. Thus, if one tells the subject that they have an endowment of $15 and that one lottery outcome is to have $8 taken from them, then the frame might be appropriately assumed to be $15 and this outcome coded as a loss of $8. But if the subject had been told, or expected, to earn only $5 from the experimental task, would this be coded instead as a gain of $2? The subjectivity and contextual nature of the reference point has been emphasized throughout by Kahneman and Tversky (1979), even though one often collapses it to the experimenter-induced frame in evaluating laboratory experiments. This imprecision in the reference point is not a criticism of PT, just a challenge to be careful assuming that it is always fixed and deterministic (see Schmidt, Starmer, & Sugden, 2005; Kőszegi & Rabin, 2006, 2007; Andersen, Harrison, & Rutström, 2006b).[49]

A corollary is that it might be a mistake to view loss aversion as a fixed parameter λ that does not vary with the context of the decision, *ceteris paribus* the reference point. See Novemsky and Kahneman (2005a) and Camerer (2005; pp. 132, 133) for discussion of this concern, which arises most clearly in dynamic decision-making settings with path-dependent earnings. This issue is particularly serious when one evaluates risk attitudes in some of the high-stakes game shows: see Andersen, Harrison, Lau, and Rutström (2008c) for a review of these studies and the modeling issues that arise.

To gauge the extent of the problem, we re-visit the estimation of a structural CPT model using our laboratory data (the replication of the Hey and Orme (1994) reported in Harrison and Rutström (2005)), but this time consider the effect of assuming different reference points than the one induced by the task frame. Assume that the reference point is χ, as in Eqs. (1‴) and (1⁗) above, but instead of setting $\chi = \$0$, allow it to vary between \$0 and \$10 in increments of \$0.10. The results are displayed in Fig. 14. The top left panel shows a trace of the log-likelihood value as the reference point is increased, and reaches a maximum at \$4.60. To properly interpret this value, note that these estimates are made at the level of the individual choice in this task, and the subject was to be paid for three of those choices. So the reference point for the overall task of 60 choices would be \$13.80 ($= 3 \times \$4.60$). This is roughly consistent with the range of estimates of expected session earnings elicited by Andersen et al. (2006b) for a sample drawn from the same population.[50]

The other interesting part of Fig. 14 is that the estimate of loss aversion increases steadily as one increases the assumed reference point. At the ML reference point of \$4.60, λ is estimated to be 2.51, with a standard error of 0.37 and a 95% confidence interval between 1.79 and 3.24. These estimates

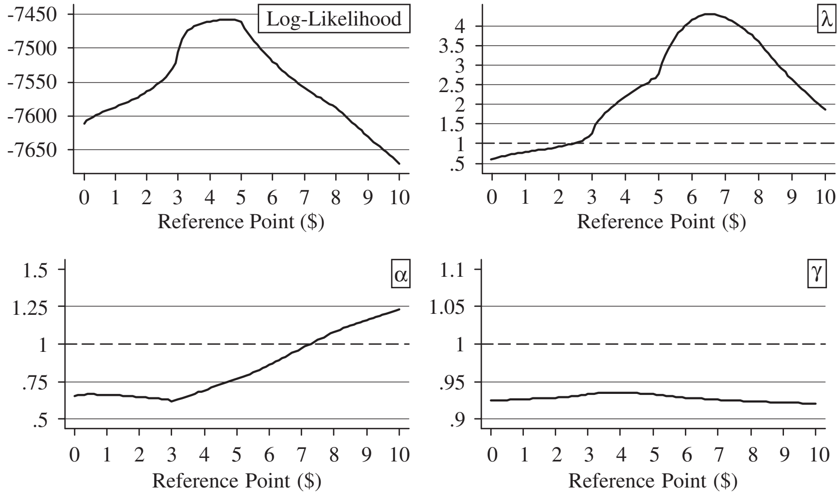


Fig. 14. Estimates of the Structural CPT Model with a Range of Assumed Reference Points. (Estimated with Subjects from the Harrison and Rutström (2005) design; $N = 207$ Subjects: 63 Gain Frame, 57 Loss Frame, and 87 Mixed Frame.)

raise an important methodological question: was it the data that led to the conclusion that loss aversion was significant, or the priors favoring significant loss aversion that led to the empirical specification of reference points? Our results may appear to be a confirmation of the argument made by some PT analysts that $\lambda \approx 2$, but it is important to recognize that the estimates presented here may not extend to other data sets or to other error specifications in the likelihood function. Further, in experimental subject pools with different reference points we would find something else entirely. At the very least, it is premature to proclaim "three cheers" for loss aversion (Camerer, 2005).

3.3. Characterizing Risk Attitudes with Several Latent Data Generating Processes

Since different models of choice behavior under uncertainty imply somewhat different characterizations of risk attitudes, it is important that we make some determination about which of these models is to be adopted. One of the enduring contributions of behavioral economics is that we now have a rich set of competing models of behavior in many settings, with EUT and PT as the two front-runners for choices under uncertainty. Debates over the validity of these models have often been framed as a horse race, with the winning theory being declared on the basis of some statistical test in which the theory is represented as a latent process explaining the data. In other words, we seem to pick the best theory by "majority rule." If one theory explains more of the data than another theory, we declare it the better theory and discard the other one. In effect, after the race is over we view the horse that "wins by a nose" as if it was the only horse in the race. The problem with this approach is that it does not recognize the possibility that several behavioral latent processes may co-exist in a population. Recognizing that possibility has direct implications for the characterization of risk attitudes in the population.

Ignoring this possibility can lead to erroneous conclusions about the domain of applicability of each theory, and is likely an important reason for why the horse races pick different winners in different domains. For purely statistical reasons, if we have a belief that there are two or more latent population processes generating the observed sample, one can make more appropriate inferences if the data are not forced to fit a specification that assumes one latent population process.

Heterogeneity in responses is well recognized as causing statistical problems in experimental and non-experimental data. Nevertheless, allowing for heterogeneity in responses through standard methods, such as fixed or random effects, is not helpful when we want to identify which people behave according to what theory, and when. Heterogeneity can be partially recognized by collecting information on observable characteristics and controlling for them in the statistical analysis. For example, a given theory might allow some individuals to be more risk averse than others as a reflection of personal preference. But this approach only recognizes *heterogeneity within a given theory*. This may be important for valid inferences about the ability of the theory to explain the data, but it does not allow for *heterogeneous theories* to co-exist in the same sample.

One approach to heterogeneity and the possibility of co-existing theories adopted by Harrison and Rutström (2005) is to propose a "wedding" of the theories. They specify and estimate a grand likelihood function that allows each theory to co-exist and have different weights, a so-called mixture model. The data can then identify what support each theory has. The wedding is consummated by the ML estimates converging on probabilities that apportion non-trivial weights to each theory.

Their results are striking: EUT and PT share the stage, in the sense that each accounts for roughly 50% of the observed choices. Thus, to the extent that EUT and PT imply different things about how one measures risk aversion, and the role of the utility function as against other constructs, assuming that the data is generated by one or the other model can lead to erroneous conclusions. The fact that the mixture probability is estimated with some precision, and that one can reject the null hypothesis that it is either 0 or 1, also indicates that one cannot claim that the equal weight to these models is due to chance.

The main methodological lesson from this exercise is that one should not rush to declare one or other model as a winner in all settings.[51] One would expect that the weight attached to EUT would vary across task domains, just as it can be shown to vary across observable socio-economics characteristics of individual decision makers.

Another approach to heterogeneity involves the use of "random parameters" in models, illustrated well by Wilcox (2008a, 2008b). Consider the simple EUT specification with no stochastic noise assumption, given by Eqs. (1)–(5). There is one parameter doing all the empirical work: the coefficient of RRA r. In the traditional statistical specification r is treated as the same across all individuals in the sample, or as a linear function of observable characteristics. An alternative approach is to view r as varying over the sample according to some distribution, commonly assumed to be Normal. In that specific case there are really two parameters to be estimated, the mean of r and the standard deviation of r.

If the heterogeneity of process takes a nested form, in the sense that one process is a restricted form of the other, then one can think of the correct statistical specification as either a finite mixture model or a random coefficients specification. In the latter case one would want to allow more flexible functional forms than Normal, to allow for multiple modes, but this is easy to generate as the sum of several uni-modal distributions. If the heterogeneity of process takes a non-nested form, such that the parameter sets are distinct for each process, then the mixture specification is more appropriate, or one should use a combination of mixture and random parameter specifications (Conte, Hey, & Moffatt, 2007).

3.4. Joint Elicitation of Risk Attitudes and Other Preferences

In many settings in experimental economics we want to elicit some preference from a set of choices that also depend on risk attitudes. Often these involve strategic games, where the uncertain ways in which behavior of others deviate from standard predictions engenders a lottery for each player. Such uncertain deviations could be due to, for example, unobservable social preferences such as fairness or reciprocity. One example is offers made in Ultimatum bargaining when the other player cannot be assumed to always accept a minuscule amount of money, and acceptable thresholds may be uncertain. Other examples include Public goods contribution games where one does not know the extent of free riding of other players, and Trust games in which one does not know the likelihood that the other player will return some of the pie transferred to him. Another source of uncertainty is the possibility that subjects make decisions with error, as predicted in Quantal Response Equilibria. Later we consider one example of this use of controls for risk attitudes in bidding in first-price auctions.

In some cases, however, we simply want to elicit a preference from choices that do not depend on the choices made by others in a strategic sense, but which still depend on risk attitudes. An example due to Andersen, Harrison, Lau, and Rutström (2008a) is the elicitation of individual discount rates. In this case, it is the concavity of the utility function that is important, and under EUT that is synonymous with risk attitudes. The implication is that we should combine a risk elicitation task with a time preference elicitation task, and use them jointly to infer discount rates over utility.

Assume EUT holds for choices over risky alternatives and that discounting is exponential. A subject is indifferent between two income options M_t and $M_{t+\tau}$ if and only if

$$U(\omega + M_t) + \left(\frac{1}{(1+\delta)^{\tau}}\right)U(\omega) = U(\omega) + \left(\frac{1}{(1+\delta)^{\tau}}\right)U(\omega + M_{t+\tau}) \quad (10)$$

where $U(\omega + M_t)$ is the utility of monetary outcome M_t for delivery at time t plus some measure of background consumption ω, δ the discount rate, τ the horizon for delivery of the later monetary outcome at time $t+\tau$, and the utility function U is separable and stationary over time. The left-hand side of Eq. (10) is the sum of the discounted utilities of receiving the monetary outcome M_t at time t (in addition to background consumption) and receiving nothing extra at time $t+\tau$, and the right-hand side is the sum of the discounted utilities of receiving nothing over background consumption at time t and the outcome $M_{t+\tau}$ (plus background consumption) at time $t+\tau$. Thus, Eq. (10) is an indifference condition and δ is the discount rate that equalizes the present value of the *utility* of the two monetary outcomes M_t and $M_{t+\tau}$, after integration with an appropriate level of background consumption ω.

Most analyses of discounting models implicitly assume that the individual is risk neutral,[52] so that Eq. (10) is instead written in the more familiar form

$$M_t = \left(\frac{1}{(1+\delta)^{\tau}}\right)M_{t+\tau} \quad (11)$$

where δ is the discount rate that makes the present value of the two monetary outcomes M_t and $M_{t+\tau}$ equal.

To state the obvious, Eqs. (10) and (11) are not the same. As one relaxes the assumption that the decision-maker is risk neutral, it is apparent from Jensen's Inequality that the implied discount rate decreases if $U(M)$ is concave in M. Thus, one cannot infer the level of the individual discount rate without knowing or assuming something about their risk attitudes. This identification problem implies that risk attitudes and discount rates cannot be estimated based on discount rate experiments alone, but separate tasks to identify the influence of risk preferences must also be implemented.

Andersen et al. (2008a) do this, and infer discount rates for the adult Danish population that are well below those estimated in the previous literature that assumed RN, such as Harrison, Lau, and Williams (2002), who estimated annualized rates of 28.1% for the same target population. Allowing for concave utility, they obtain a point estimate of the discount rate of 10.1%, which is significantly lower than the estimate of 25.2% for the same sample assuming linear utility. This does more than simply verify that discount rates and risk aversion coefficients are mathematical substitutes in the sense that either of them have the effect of lowering the influence from future payoffs on present utility. It tells us that, for risk aversion coefficients that are reasonable from the standpoint of explaining choices in the lottery choice task, the estimated discount rate takes on a value that is much more in line with what one would expect from market interest rates. To evaluate the statistical significance of adjusting for a concave utility function one can test the hypothesis that the estimated discount rate assuming risk aversion is the same as the discount rate estimated assuming RN. This null hypothesis is easily rejected. Thus, *allowing for risk aversion makes a significant difference to the elicited discount rates.*

3.5. Testing Expected Utility Theory

Much of the data collected with the direct intent of testing EUT involved choice pairs selected deliberately to provide a way of testing EUT *without having to know the risk attitudes of subjects*. Unfortunately they provide extremely weak tests, since one can only count a choice as a success or failure of the theory, and no transparent metric suggests itself to weight some violations rather than others as more serious.[53] This is why we generally use ML to estimate parameters in such binary choice settings, and not the "hit ratio," since some hits are closer than others and we want to take that into account by calculating the probability of the observed choice conditional on the parameters being evaluated.[54]

The problem is even more serious than devising a metric to test the seriousness of violations. In two respects, EUT is a hard theory to reject in these settings First, how does one know if the subjects are actually indifferent to the choice pairs on offer? Allowing subjects to express indifference does not suffice, since there is no way to know if they have randomized internally before picking out one lottery. Moreover, waiting for the data to exhibit 50–50 splits for indifference presumes that no artifactual presentation biases exist.[55] Second, how does one know if the subjects are not extremely risk averse? High levels of risk aversion mean that the CE of the utility values of the prizes are all close to "very small numbers." Hence, for sufficiently high levels of risk aversion, the CEs of the two lotteries are virtually identical and the subject should be rationally indifferent. Unfortunately, this free parameter gives EUT the formal leeway to escape from virtually any test one can think of. These problems lead one to question how operationally meaningful these tests are without some independent characterization of risk attitudes.

To provide one striking example of this issue, consider the Preference Reversal tests of EUT presented to economists by Grether and Plott (1979). In these experiments, the subject was asked to make a direct binary choice between lotteries A and B, and then to state a valuation on each of A and B. From the latter two valuations the experimenter can infer a binary preference. The reversal is said to occur when the inferred binary preference differs from the direct binary choice. One design feature of these tasks is that A and B had virtually identical expected value. Given this information, anthropomorphize and sympathize with a poor ML estimation routine trying to explain any sample of choices in which there are significant numbers of reversals. It could try assuming subjects were risk neutral, and then it could "explain" any observed choice since the subject would be indifferent between either option.

The best way to address these concerns is to characterize the risk attitudes of the subjects independently of the choice tasks, allowing the experimenter to identify those subjects that make for better tests of EUT. This identification can proceed independently of the choice data one is seeking to confront with EUT.

To illustrate, consider the Common Ratio tests of EUT from Cubitt, Starmer, and Sugden (1988a) (CSS). The CSS tests used 451 subjects, who were randomly given one of five problems.[56] The first and last problems in CSS were a choice between simple prospects. Problem 1 was a choice between option A, which was an 80% chance of £16, and option B, which was £10 for certain.

Problem 5 was a simple "common ratio" transformation which multiplied each option by 1/4, so that option A* was a 20% chance of £16 and option B* was a 25% chance of £10. Problems 2 through 4 were procedural variants on Problem 2, which are identical to Problem 5 from the perspective of EUT. We refer to these as problems AB and A*B* for present purposes, in new experiments discussed below. Thus, CSS Problems 2–5 correspond to problem A*B* in our design, and their Problem 1 corresponds to our problem AB.

Cubitt, Starmer, and Sugden (1988a; Table 2, p. 1375) report that 50% of their sample chose option A* in their Problems 2 through 5, which are qualitatively identical to problem A*B* in our design. Only 38% of their subjects chose option A in their Problem 1, which is qualitatively the same as problem AB in our design. Using the same χ^2 contingency table test employed by CSS, we can only reject the EUT hypothesis at a significance level of 11.2%; Fisher's exact test for the same two-sided comparison has a significance level of 15.3%. So there is weak evidence that EUT is violated, even if it does not strictly fail at conventional levels of significance.[57]

For a specific example of the Common Ratio test, in which we have independent information on risk attitudes, suppose Lottery A consists of prizes \$0 and \$30 with probabilities 0.2 and 0.8, and that Lottery B consists of prizes \$0 and \$20 with probabilities 0 and 1. Then one may construct two additional compound lotteries, A^* and B^*, by adding a front-end probability $q = 0.25$ of winning zero to lotteries A and B. That is, A^* offers a $(1-q)$ chance to play lottery A and a q chance of winning zero. Subjects choosing A over B and B^* over A^*, or choosing B over A and A^* over B^*, are said to violate EUT.

To show precisely how risk aversion does matter, assume that risk attitudes can be characterized by the popular CRRA function, Eq. (1). The CE of the lottery pairs AB and A^*B^* as a function of r are shown in the left and right upper panels, respectively, of Fig. 15. The CRRA coefficient ranges from -0.5 (moderately risk loving) up to 1.25 (very risk averse), with a risk-neutral subject at $r = 0$. The CE of lottery B, which offers \$20 for sure, is the horizontal line in the left panel of Fig. 15. The CE of A, A^*, and B^* all decline as risk aversion increases. The lower panels of Fig. 15 show the CE differences between the A and B (A^* and B^*) lotteries. Note that for

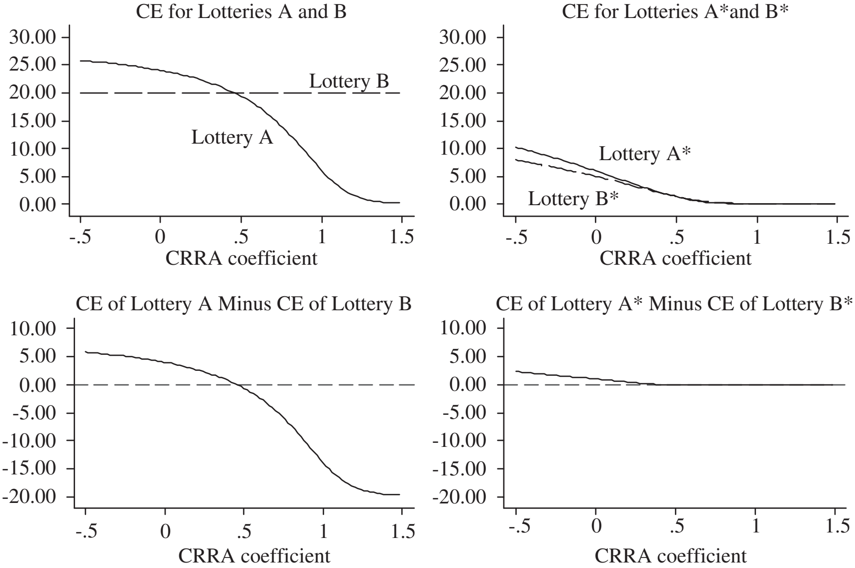


Fig. 15. Risk Attitudes and Common Ratio Tests of EUT.

the AB (A^*B^*) lotteries, the preferred outcome switches to lottery B (B^*) for a CRRA coefficient about 0.45.

Most evaluations of EUT acknowledge that one cannot expect any theory to predict perfectly, since any violation would lead one to reject the theory no matter how many correct predictions it makes. One way to evaluate mistakes is to calculate their costs under the theory being tested and to "forgive" those mistakes that are not very costly, while holding to account those that are. For each subject in our data and each lottery choice pair, we can calculate the CE *difference* given the individual's estimated CRRA coefficient, allowing us to identify those choice pairs that are most salient. A natural metric for defining "trivial EUT violations" can then be defined in terms of choices that involve a difference in CE below some given threshold.

Suppose for the moment that an expected utility maximizing individual will flip a coin to make a choice whenever the difference in CE falls below some cognitive threshold. If $r = 0.8$, the CE difference in favor of B is large in the first lottery pair and B will be chosen. In the second lottery pair, the difference between the payoffs for choosing A^* and B^* is trivial (less than a cent, in fact) and a coin is flipped to make a choice. Thus, with probability 0.5 the experimenter will observe the individual choosing B and A^*, a choice pattern inconsistent with EUT. In a sample with these risk attitudes, half the choices observed would then be expected to be inconsistent with EUT. With such a large difference between the choice frequencies, standard statistical tests would easily reject the hypothesis that they are the same. Thus, we would reject EUT in this case *even though* EUT is *essentially*[58] *true*.

Fig. 16 collates estimates of risk attitudes elicited by Harrison, Johnson, McInnes, and Rutström (2005b) from 152 subjects, described in Section 1.2 and Table 3. The idea is to simply align the CE differences for each of the CR lotteries (AB in the left panel, and A^*B^* in the right panel) with the distribution of risk attitudes expected from this sample (the bottom boxes). Clearly the subjects tend to have risk attitudes at *precisely the point at which these tests have least power to reject EUT*. This is particularly striking for the A^*B^* lottery choice, but even for the AB lottery choice it is only the few subjects "in the tails" of the risk distribution for which EUT has a strong prediction. Further, these risk attitude distributions refer to point estimates, and do not reflect the uncertainty of those estimates: it is quite possible that some subject that has a point estimate of his CRRA coefficient that makes the AB test powerful also has a large enough standard error on that point estimate that the AB test is not powerful. This issue of precision is addressed directly by Harrison, Johnson, McInnes, and Rutström (2007a).

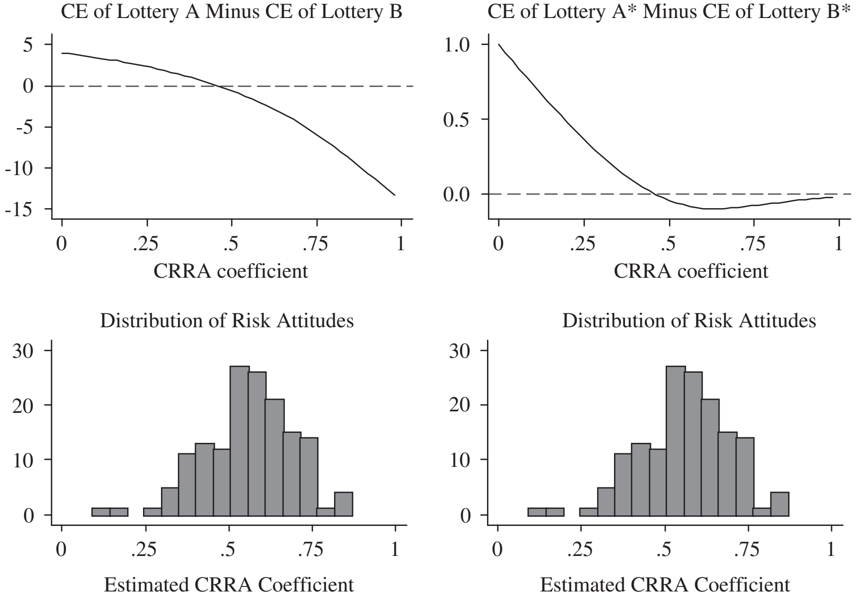


Fig. 16. Observed Risk Attitudes and Common-Ratio Tests of EUT.

For some violations it may be easy to write out specific parametric models of the latent EUT decision-making process that can account for the data. The problem is that the model that can easily account for one set of violations need not account for others. As already noted, the preference reversals of Grether and Plott (1979) can be explained by assuming risk-neutral subjects with an arbitrarily small error process, since the paired lotteries are designed to have the same expected value. Hence, each subject is indifferent, and the error process can account for the data.[59] But then such subjects should *not* violate EUT in other settings, such as common ratio tests.

However, rarely does one encounter tests that confront subjects with a wide range of tasks and evaluates behavior simultaneously over that wider domain. There are three striking counter-examples to this trend. First, Hey, and Orme (1994) deliberately use lotteries that span a wide range of prizes and probabilities, avoiding "trip wire" pairs, and they conclude that EUT does an excellent job of explaining behavior compared to a wide range of alternatives. Second, Harless and Camerer (1994) consider a wide range of aggregate data across many studies, and find that EUT does a good job of explaining behavior if one places sufficient value on parsimony. On the other hand, all of the data used by Harless and Camerer (1994) come from experimental designs that were intended to be tough on EUT compared to some alternative model; so their data is not as generic as Hey and Orme (1994). Third, Loomes and Sugden (1998) deliberately choose lotteries "… to provide good coverage of the space within each (implied Marschak–Machina probability) triangle, and also to span a range of gradients sufficiently wide to accommodate most subjects' risk attitudes." (p. 589). Their coverage is not as wide as Hey and Orme (1994) in terms of the range of CRRA values for which subjects would be indifferent under EUT, but the intent is clearly to provide some variability, and for the right reasons.

Maximal statistical power calls for what might be termed a "complementary slack experimental design": choose one set of tasks such that if subjects are risk averse (risk neutral) then the choice model is tested, recognizing that if they are risk neutral (risk averse) then the other set of tasks tests the choice model. Thus, the subjects that clearly provide little information about EUT in common ratio tests in Fig. 16 should provide significant information about EUT in preference reversal tests (Harrison et al., 2007a).[60] On the other hand, we know relatively little about what is the most "ecologically relevant" lottery pairs to use if we are trying to model task domains in a representative manner. Our only point is that this consideration deserves more attention by economists interested in making claims about the *general* validity of EUT or any other model, echoing similar calls from others (Smith, 2003).

3.6. Testing Auction Theory

To illustrate the potential importance of controlling for the risk attitude confound in a strategic setting, consider an important case in which there has been considerable debate over the ability of received theory to account for behavior: bidding in a first-price sealed-bid auction characterized by private and independent values.[61] Auction theory is very rich, and has been developed specifically for the parametric cases considered in experiments (e.g., Cox, Roberson, & Smith, 1982; Cox, Smith, & Walker, 1988). In a new series of laboratory experiments data are collected on observed valuations and bids, using standard procedures. However, information is also elicited that identifies the risk attitudes of the same subject, since that is a critical characteristic of the predicted bid under the standard model (e.g., Harrison, 1990). It is then straightforward to specify a joint likelihood function for the observed risk aversion responses and bids, estimate the risk aversion characteristic, and test if the implied NE bid systematically differs from the observed bid. The results are striking. In the simplest possible case, when there are only two bidders ($N = 2$), received theory does a wonderful job of characterizing behavior when one controls for the risk attitudes of the individual bidder.[62]

3.6.1. Theoretical Predictions

Cox et al. (1982) develop a model of bidding behavior in first-price sealed-bid auctions that assumes that each agent has a CRRA power utility function $U(y) = y^r$, where U is the utility of experimental income y and $(1 - r_i)$ is the Arrow–Pratt measure of risk aversion (RA). Each agent has their own r_i, so each agent is allowed to have distinct risk attitudes. However, r_i is restricted to lie on the closed interval (0,1), where $r_i = 1$ corresponds to RN. Hence, this model allows (weak) risk aversion, but does not admit risk-loving behavior.[63] Each agent in the model knows their own risk attitude, their own valuation v_i, that everyone's risk attitudes are drawn from the closed interval (0,1), and that everyone's valuation is drawn from a uniform distribution over the interval (v_0, v^1). It can then be shown that the symmetric Bayesian NE implies the following bid function:

$$b_i = v_0 + \left(\frac{(N-1)}{(N-1+r_i)}\right)(v_i - v_0) \tag{12}$$

where there are N active bidders. In the RN case in which $v_0 = 0$, $v^1 = 1$, and $r_i = 1$, this model is the one derived by Vickrey (1961), and calls for bidders to choose their optimal bid using a simple rule: take the valuation received and shade it down by $(N - 1)/N$. When $N = 2$, the RN NE bidding rule is therefore particularly simple: bid one-half of the valuation. Thus, one might expect the $N = 2$ case to provide a particularly compelling test of the general RA NE bidding rule, since the optimal RN NE bid is also an arithmetically simple heuristic.[64]

3.6.2. Experimental Design and Procedures

Each subject in our experiment participated in a single session consisting of two tasks. The first task involved a sequence of choices designed to reveal each subject's risk preferences. In the second task, subjects participated in a series of 10 first-price auctions against random opponents, followed by a small survey designed to collect individual characteristics. A total of 58 subjects from the student population of the University of Central Florida participated over three sessions. The smallest number of subjects in one session was 16, so there was little chance that the subjects would *rationally* believe that they could establish reputations over the 10 rounds of bidding against a random opponent.[65]

Each subject was told that they would be privately assigned induced values between \$0 and \$8, using a uniform distribution. Cox et al. (1982) show that for RN subjects the expected earning of each subject in a first-price auction is $(v^1 - v_0)/N(N+1)$, where v^1 and v_0 are the upper and lower bound for the support of the induced values. Thus, expected RN earnings were \$1.33 per subject in each period. Subjects in each session were also informed of the number of other bidders in the auction; that the other bidders' induced values were, like their own, drawn from a uniform support with bounds given above; and that their earnings in the auction would equal their induced value minus their bid if they have the highest bid, or zero otherwise.

We used the Holt and Laury (2002) design to elicit risk attitudes from the same subjects. In these experiments, we scaled these baseline prizes of their design, shown in panel A of Table 1, up by a factor of 2, so that the largest prize was \$7.70 and the smallest prize was \$0.20. The prizes in these lotteries effectively span the range of possible incomes in the auction, which range from \$8.00 to zero.

3.6.3. Results

Panel B of Fig. 17 displays observed bidding behavior. The induced value is displayed on the bottom axis, a 45° line is shown and corresponds to the subject just bidding their value, and then the RN bid prediction is shown under that 45° line. The standard behavior from a long series of such experiments is observed: subjects tend to bid higher than the RN prediction, to varying degrees.

The statistical model consists of a likelihood of observing the risk aversion responses *and* the observed bidding responses.

The likelihood of the risk aversion responses is modeled with a probit choice rule defined over the 10 binary choices that each subject made, exactly as illustrated in Section 1.2 but for the power utility function. To allow for subject heterogeneity with respect to risk attitudes, the parameter r is modeled as a linear function of observed individual characteristics of the subject. For example, assume that we only had information on the age and sex of the subject, denoted Age (in years) and Female (0 for males, and 1 for females). Then we would estimate the coefficients α, β, and γ in $r = \alpha + \beta \times \text{Age} + \gamma \times \text{Female}$. Therefore, each subject would have a different estimated r, $\hat{r}$, for a given set of estimates of α, β, and γ to the extent that the

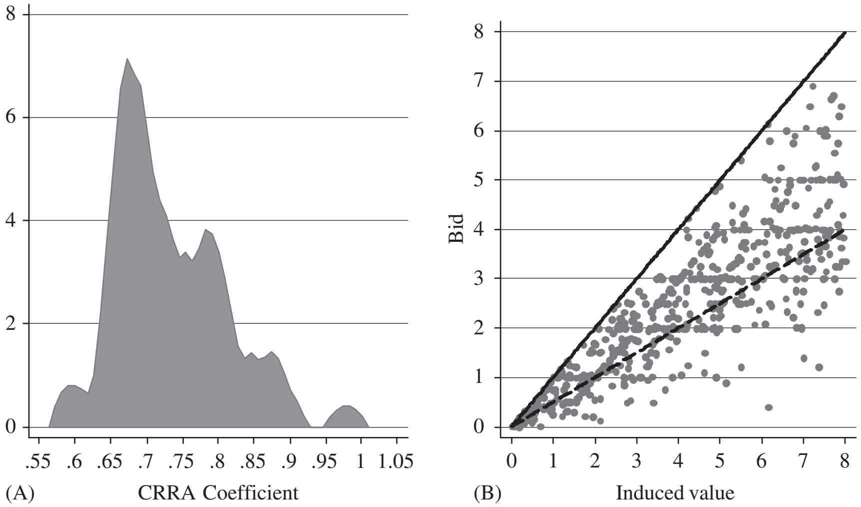


Fig. 17. Observed Risk Attitudes and Observed Bidding. (A) Risk Elicitation Task; Power Utility Function Assumed: $r<1$ is RA, $r=1$ is RN, and $r>1$ is RL. (B) First-Price Sealed Bid Auction; 2 bidders per auction over 10 rounds. $N=58$ Subjects with Random Opponents. Valuations between \$0 and \$8.

subject had distinct individual characteristics. So if there were two subjects with the same sex and age, to use the above example, they would literally have the same $\hat{r}$, but if they differed in sex and/or age they would generally have distinct $\hat{r}$. In fact, we use 12 individual characteristics in our model. Apart from age and sex, these include binary indicators for race (Non-White), a Business major, rich (parental or own income over \$80,000 in 2003), high GPA (above 3.75), low GPA (below 3.25), college education for the father of the subject, college education for the mother of the subject, whether the subject works, whether the subject is a Catholic, and whether the subject is some other Christian denomination. Panel A of Fig. 17 displays the predicted risk attitudes from this estimation exercise, using only the risk aversion task.

The likelihood of the bidding responses is then modeled as a multiplicative function of the predicted bid conditional on the estimated risk attitude for the subject. Thus, we estimate a coefficient b which scales up or down the predicted NE bid: if $b=1$ then the observed bid exactly tracks the predicted bid *for that subject*. The predicted NE bid for each subject i depends, of course, on the $\hat{r}_i$ for that subject, as well as the parameters N,

v_0, v^1, and v_i. Thus, if we observe two subjects with the same v_i but different bids, it is perfectly possible for this to be consistent with the predicted NE bid if they have distinct individual characteristics and hence distinct $\hat{r}_i$. The coefficient b is also modeled as a linear function of the same set of individual characteristics as the coefficient r.[66]

The full specification of the likelihood function for bidding allows for heteroskedasticity with respect to individual characteristics. Thus, the specification is $(b \times b^{\mathrm{NE}}) + \varepsilon$, where the variance of ε is again a linear function of the individual characteristics. Thus we obtain information from the coefficients of b on which types of subjects deviate systematically from the NE prediction, and we obtain information from the coefficients on ε on which types of subjects exhibit more noise in their bidding.

The overall likelihood consists of the likelihood of the risk aversion responses plus the likelihood of the bidding responses, conditional on estimates of r, b, and the variance of ε. In turn, these three parameters are linear functions of a constant and the individual characteristics of the subject. Since each subject provides 10 binary choices in the risk aversion task, and 10 bids in the auction task, we use clustering to allow for the responses of the same subject to be correlated due to unobserved individual effects.

Table 7 displays the ML estimates. The intercept for r is estimated to be 0.612, consistent with evidence from comparable experiments of risk aversion discussed earlier. The intercept for b is 1.02, consistent with bids being centered on the RA NE bid conditional on the estimated risk aversion for each subject. The top panel of Fig. 18 shows the distribution of predicted values of b for each of the 58 subjects. Some subjects have estimates of b as low as 0.8, or as high as 1.35, but the clear majority seem to be tracked well by the RA NE bidding prediction. The bottom panel of Fig. 18 displays a distribution of comparable estimates when we use the RN NE bidding prediction instead of the RA NE bidding prediction, and re-estimate the model. Observed bids are about 25% higher than predicted if one assumes, counter-factually, that subjects are all RN.

3.7. Testing Myopic Loss Aversion

Prospect Theory has forced economists to worry about the task domain over which decisions are evaluated, where a sequence of many tasks over time may be treated very differently from a single choice task. PT obviously focuses on the implications for loss aversion from this differential treatment.

Table 7. Maximum Likelihood Estimates for Model of Bidding Behavior.

Parameter	Variable	Point Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
r	Constant	0.612	0.320	0.06	– 0.02	1.24
	Age	0.003	0.015	0.82	– 0.03	0.03
	Female	– 0.052	0.079	0.51	– 0.21	0.10
	Non-white	0.011	0.081	0.89	– 0.15	0.17
	Major is in business	0.058	0.086	0.50	– 0.11	0.23
	Father completed college	– 0.004	0.083	0.96	– 0.17	0.16
	Mother completed college	0.003	0.093	0.97	– 0.18	0.19
	Income over $80k in 2003	0.036	0.073	0.62	– 0.11	0.18
	Low GPA (below 3.24)	– 0.024	0.098	0.81	– 0.22	0.17
	High GPA (greater than 3.75)	0.190	0.113	0.09	– 0.03	0.41
	Work full-time or part-time	– 0.022	0.074	0.77	– 0.17	0.12
	Catholic religious beliefs	– 0.046	0.130	0.72	– 0.30	0.21
	Other Christian religious beliefs	0.040	0.080	0.62	– 0.12	0.20
b	Constant	1.021	0.721	0.16	– 0.39	2.43
	Age	– 0.007	0.030	0.81	– 0.07	0.05
	Female	0.019	0.084	0.82	– 0.14	0.18
	Non-white	– 0.059	0.079	0.45	– 0.21	0.09
	Major is in business	0.023	0.083	0.78	– 0.14	0.19
	Father completed college	0.054	0.068	0.43	– 0.08	0.19
	Mother completed college	– 0.023	0.085	0.79	– 0.19	0.14
	Income over $80k in 2003	0.078	0.074	0.29	– 0.07	0.22
	Low GPA (below 3.24)	0.001	0.079	0.99	– 0.15	0.16
	High GPA (greater than 3.75)	0.210	0.124	0.09	– 0.03	0.45
	Work full-time or part-time	– 0.019	0.068	0.78	– 0.15	0.11
	Catholic religious beliefs	0.035	0.095	0.72	– 0.15	0.22
	Other Christian religious beliefs	0.157	0.080	0.05	0.00	0.31
ε	Constant	0.096	1.093	0.93	– 2.05	2.24
	Age	– 0.011	0.046	0.81	– 0.10	0.08
	Female	0.008	0.156	0.96	– 0.30	0.32
	Non-white	– 0.077	0.162	0.63	– 0.39	0.24
	Major is in business	0.123	0.133	0.36	– 0.14	0.38
	Father completed college	– 0.330	0.139	0.02	– 0.60	– 0.06
	Mother completed college	0.078	0.199	0.69	– 0.31	0.47

Table 7. (*Continued*)

Parameter	Variable	Point Estimate	Standard Error	*p*-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
	Income over $80k in 2003	0.044	0.119	0.71	− 0.19	0.28
	Low GPA (below 3.24)	0.116	0.111	0.29	− 0.10	0.33
	High GPA (greater than 3.75)	0.044	0.191	0.82	− 0.33	0.42
	Work full-time or part-time	− 0.144	0.148	0.33	− 0.43	0.15
	Catholic religious beliefs	− 0.341	0.157	0.03	− 0.65	− 0.03
	Other Christian religious beliefs	− 0.077	0.149	0.61	− 0.37	0.21

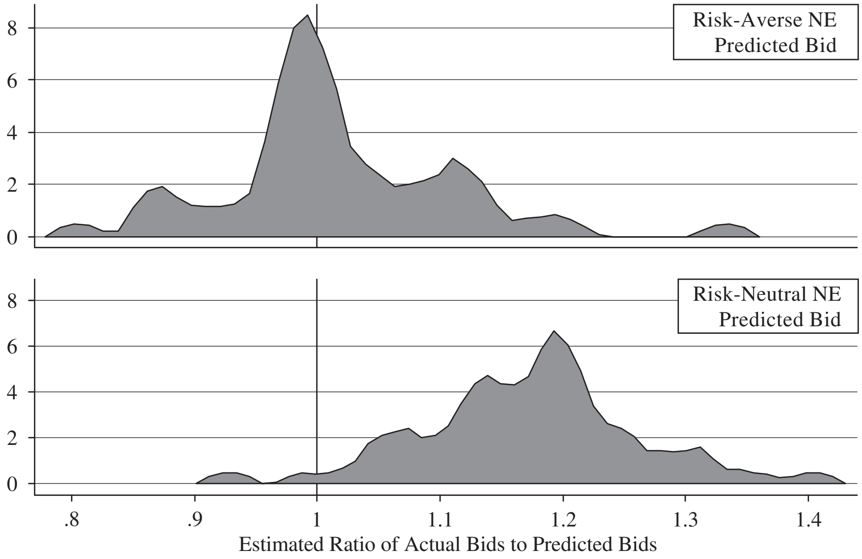


Fig. 18. Relative Support for Alternative Nash Equilibrium Bidding Models.

Unfortunately, the insight from PT that the evaluation period might differ from setting to setting, or from subject to subject, has not been integrated into EUT. In fact, this insight is often presented as one of the essential points of departure from EUT, and as one of the differentiating characteristics of PT. We argue that behavioral issues of the evaluation period is a more general and fundamental concern than concerns about loss aversion in PT. By considering recent experimental tests of aversion of this insight known as Myopic Loss Aversion (MLA), it is possible to see that the insight is just as relevant for EUT, and that a full characterization of risk attitudes must account for the evaluation period.

Camerer (2005; p. 130) explains why one naturally thinks of loss aversion and the evaluation period together:

> A crucial ingredient in empirical applications of loss aversion is decision isolation, or focusing illusion, in which single decisions loom large even though they are included in a stream of similar decisions. If many small decisions are integrated into a portfolio of choices, or a broad temporal view – the way a gambler might view next year's likely total wins and losses – the loss on any one gamble is likely to be offset by others, so aversion to losses is muted. Therefore, for loss aversion to be a powerful empirical force requires not only aversion to loss but also a narrow focus such that local losses are not blended with global gains. This theme emerges in the ten field studies that Camerer (2000) discusses, which show the power of loss aversion (and other prospect theory features) to explain substantial behaviors outside the lab.

However, there is very little direct experimental evidence, with real stakes, to support MLA. Furthermore, we argue that what evidence there is also happens to be consistent with EUT. By carefully considering those experimental tests from the perspective of EUT and the implications for the characterization of risk attitudes, it is easy to see that the behavioral issue of the evaluation period is a more general and fundamental concern.

Several recent studies propose experimental tests that purport to directly test EUT against the alternative hypothesis of MLA. Gneezy and Potters (1997) and Haigh and List (2005) use simple experiments in which many potential confounds are removed.[67] Unfortunately, those experiments only test a very special case of EUT against the alternative hypothesis. This special case is CRRA, and it fails rather dramatically. But it is easy to come up with other utility functions that are consistent with EUT and that can explain the observed data without relying on MLA. For example, any utility function with decreasing RRA and that exhibits risk aversion for low levels of income will suffice at a qualitative level. The empirical outcomes observed at the individual level can then be explained by simply fitting specific parameters to this utility function. Appendix E demonstrates this intuitively, as well as more formally.

Our new analysis of the GP data presented in Appendix E also identifies some unsettling implications of these experiments for MLA: that the key "loss aversion" parameters of the standard MLA model vary dramatically according to the exogenously imposed evaluation period and that the risk attitudes are the opposite of those generally assumed in PT, viz., risk loving in gains and risk averse in losses. Thus, the behaviorist explanation is hoisted on the same petard it alleged applied to the EUT explanation, the presence of anomalous behavior.

However, although it is useful and trivial to come up with a standard EUT story that accounts for the data, and even fun to find an anomaly for the behaviorists to ponder, these experiments force one to examine a much deeper question than "can EUT explain the data?" That question is whether utility is best defined over *each individual decision* that the subject faces or over the *full sequence of decisions* that the subject is asked to make in an experimental session, or perhaps even including extra-lab decisions. Depending on how the *subjects interpret the experimental task*, these frames could differ in this experimental task. This perspective suggests the hypothesis that behavior might be better characterized as a mixture of two latent data generating processes, as suggested by Harrison and Rutström (2005) and Section 3.3, with some subjects using one frame and other subjects using another frame.

A related issue underlying the assessment of behavior from these experiments is asset integration within the laboratory session. What incomes are arguments of the utility functions of the subjects? The common assumption in experimental economics is that it is simply the prizes over which they were making *choices* whenever they got to make a choice.[68] But what about asset integration of income earned during the sequence of rounds? Gneezy and Potters (1997; p. 636) note that this could affect risk attitudes in a more general specification, but assert that the effect is likely to be small given the small stakes. This may be true, but is just an assertion and deserves more complete study using the general framework proposed by Cox and Sadiraj (2006). The Gneezy and Potters (1997) data provide an opportunity to study this question, since subjects received information on their intra-session income flows at different rates. Hence one could, in principle, test what function of accumulated wealth was relevant for their choices.

We believe that the fundamental insight of Benartzi and Thaler (1995) of the importance of the evaluation horizon of decision makers is worthy of more attention, even though we find that the present tests of MLA have been somewhat misleading. The real contribution of the MLA literature and the experimental design of Gneezy and Potters (1997) is to force mainstream economists to pay attention to an issue they have neglected *within* their own framework.

3.8. The Random Lottery Incentive Procedure

The random lottery incentive procedure originated from the desire to avoid "wealth or portfolio effects" of subjects making multiple choices at once to determine their final experimental income.[69] It also has the advantage of saving scarce experimental subject payments, but that arose originally as a happy by-product. The procedure bothers theorists and non-experimenters, particularly when one is using the experimental responses to estimate risk attitudes. The reason is that there is some ambiguity as to whether the subject is evaluating the utility of the lottery in each choice, or the compound lottery that includes the random selection of one lottery for payment.

The procedure also imposes a motivational constraint on the level of incentives one can have in certain elicitation tasks. To generate better econometric estimates we would like to gather more choices from each subject: witness the glee that Wilcox (2008a) expresses over the sample size of the design in Hey (2001). Each subject in that design generated 500 binary choices, over five sessions at separate times, and was paid for one selected at random. But 1-in-500 is a small number, even if the prizes were as high as £125 and EV maximization would yield a payoff of just over £79. So there is a tension here, in which we want to gather more choices per subject, but run the risk that the probability of any one choice being realized drops as we do so. The experiments of Hey (2001) are remarkable because they appear to have motivated subjects well – aggregate error rates from repeated tasks are very low compared to those found in comparable designs with fewer tasks (Nathaniel Wilcox; personal communication). What we would like to do is run an experiment with as many choices as we believe that subjects can perform without getting bored, but ensure that they do not see each choice as having a vanishing chance of being salient. In our experience, 60 binary choices are about the maximum we can expect our subjects to undertake without visible signs of boredom setting in. But even 1-in-60 sounds small, and may be viewed that way by subjects, effectively generating hypothetical responses and the biases that typically come with them (see Section 4.1). Of course, this is a behavioral issue: do subjects focus on the task as if it were definitely the one to be paid, or do they mostly focus on the likelihood of the task determining their earnings?

Several direct tests of this procedure lead some critics of EUT to the conclusion that the procedure appears, as an *empirical* matter, to induce no cross-task contamination effects when choices are over *simple* lottery prospects; see Cubitt, Starmer, & Sugden (1988b; p. 129), for example. Related tests include Starmer and Sugden (1991) and Beattie and Loomes

(1997). So the empirical evidence suggests that it does not matter *behaviorally*.

On the other hand, doubts remain. Certain theories of decision-making under risk differ in terms of the predicted effect these procedures have on behavior. To take an important example, consider the use of the random lottery incentive procedure in the context of an MPL task. The theoretical validity of this procedure presumes EUT, and if EUT is invalid then it is possible that this procedure might be generating invalid inferences. Under EUT it does not matter if the subjects evaluate their choices in each task separately, make one big decision over the whole set of tasks, or anything in between, since the random incentive is just a "common ratio probability" applied to each task. However, under RDU or PT this common ratio probability *could* lead to very different choices, depending on the extent of probability weighting.

Hey and Lee (2005a, 2005b) provide evidence that subjects do not appear to consider all possible tasks, but their evidence is provided in the context of RLP designs discussed in Section 1.2. In that case the subject does not know the exact lotteries to be presented in the future, after the choice before him is made, so one can readily imagine the cognitive burden involved in anticipating what the future lotteries will be.[70] But for the MPL instrument the subject does know the exact lotteries to be presented in the whole task, and the set of responses can be plausibly reduced in number to just picking one switch point, rather than picking from the $2^{10} = 1024$ possible binary choices in 10 rows. Thus, the MPL instrument may be more susceptible to concerns with the validity of the random lottery incentive procedure than other instruments.[71]

On the other hand, it is not obvious *theoretically* that one wants to avoid "portfolio effects" when eliciting risk attitudes. These effects arise as soon as subjects are paid for more than one out of *K* choices. Again, consider the same type of binary choice experiments considered above. The standard implementation of the random lottery incentive mechanism in experiments such as these would have one choice selected at random. For the case of investigating "preference reversals" the reason for only using one choice is well explained by Cox and Epstein (1989; p. 409):

> Economic theories of decision making under risk explain how variations in wealth can affect choices. Thus an agent with wealth w may prefer lottery A to lottery B but that same agent with wealth $\hat{w} \neq w$ may prefer lottery B to A. Therefore, the results of preference reversal experiments that allow a subject's wealth to change between choices cannot provide a convincing challenge to economic theory unless it can be shown that wealth effects cannot account for the results.

> Economic theories of decision making under risk provide explanations of optimal portfolio choice. Such theories explain why an agent might prefer lottery A to lottery B but prefer the portfolio (A, B) to the portfolio (A, A). If the portfolio is accumulated by sequential choice of A over B and then B over A, an apparent preference reversal could consist of choices that construct an agent's optimal portfolio.

When the interest is in the inferred risk coefficient, however, the possibility of subjects choosing portfolios to match their preferences have different implications. To avoid risk-pooling incentives, the outcomes of the lotteries must be uncorrelated, which is normally the case in such experiments. Nevertheless, even then it is possible for a subject to prefer the portfolio (A, B) to (A, A) even if he would prefer A to B when being paid only for one of his choices. To see this, recall that the lottery options presented to subjects are always discrete. In the MPL, for example, a switch from lottery A to lottery B on row 6 would lead us to infer a risk aversion coefficient that is in a numeric interval, (0.14, 0.41) in the Holt and Laury (2002) experiments. An individual with a risk aversion coefficient close to the boundaries of this interval would always pick (B, B) or (A, A), but an individual with a risk aversion coefficient in the middle of the interval would have a preference for a mixed portfolio of (A, B). *Paying for more than one lottery therefore elicits more information and allows a more precise expression of the risk preference of each subject*. The point is that we then have to evaluate risk attitudes assuming that subjects compare portfolios, rather than comparing one individual lottery with another individual lottery. If we do that, then there is no theoretical reason for avoiding portfolio effects for this inferential purpose. There may be a practical and behavioral reason for avoiding that assumption in the design considered by Hey and Lee (2005a, 2005b), given the cognitive burden (to subject and analyst) of constructing all possible expected portfolios.

The behavioral significance of the portfolio effect can be directly tested by varying the number of lottery choices to be paid. In our replication of Hey and Orme (1994) we defaulted to having 60 binary lottery choices. Over 60 binary choices we used three choices for payment, to ensure comparability of rewards with other experiments in which subjects made choices over 40 or 20 lotteries, and where 2 lotteries or 1 lottery was respectively selected at random to be played out. Thus, the 1-in-20 treatment corresponds exactly to the random lottery incentive procedure that avoids portfolio effects, and the other two treatments raise the possibility of these effects. All of these tasks were in the gain frame, and all involved subjects being provided information on the EV of each lottery. The samples consisted of 11, 21, and 25 subjects in the 20, 40, and 60 lottery treatments, respectively, for a pooled sample of

57 subjects. All the lottery outcomes were uncorrelated by executing independent draws.

We find no evidence of portfolio effects, measured by the effect on the mean elicited risk attitudes. Assume an EUT model initially, and use the CRRA function given by Eq. (1), with a Fechner error specification. Pooling data over tasks in which the subject faced 20, 40, or 60 lotteries, on a between-subjects basis, and including a binary dummy for those sessions with 20 or 40 lotteries, there is no statistically significant effect on elicited risk attitudes. Quite apart from statistical insignificance, the estimated effect is small: around ±0.04 or less in terms of the risk aversion coefficient. The same conclusion holds with a comparable RDU model, whether one looks at the concavity of the utility function, Eq. (1), the curvature of the probability weighting function, Eq. (9), or both.

This valuable result is worth replicating with larger samples and in different elicitation procedures. We want to have more binary choices from the same subject to get more precise estimates of latent structural models, but on the other hand we worry that paying 1-in-*K* choices for *K* "large" might seriously dilute incentives for thoughtful behavior over consequential outcomes. If one can modestly increase the salience of each choice, as implemented here, and not worry about portfolio effects, then it is possible to use values of *K* that allow much more precise estimates of risk attitudes. Of course, the absence of the portfolio effect must be checked behaviorally, as illustrated here.

3.9. Summary Estimates

We finally collate some "preferred" estimates of simple specifications of risk attitudes from the various designs and statistical specifications in the literature. We do not mechanically list every estimate from every design and specification, in the spirit of some meta-analyses, ignoring the weaknesses we have discussed in each. Instead, we use *a priori* judgements to focus on two of the designs that we believe to be most attractive, the statistical specifications we believe to be the best available, and the studies we have the most reliable data. One design is the classic data set of Hey and Orme (1994), and the other is the classic design of Holt and Laury (2002, 2005). We favor the Holt and Laury (2005) study over Holt and Laury (2002), because of the contaminant of order effects in the earlier design, identified by Harrison, Johnson, McInnes, and Rutström (2005b). Similarly, we favor the Fechner error specification of Hey and Orme (1994) over the Luce specification of Holt and Laury (2002), for reasons detailed by Wilcox

(2008a).[72] We also augment the British data from Hey and Orme (1994) with results from our replications with U.S. college students.

We consider CRRA and EP variants for EUT, and also consider some simple RDU specifications. The CRRA utility function is specification Eq. (1) from Section 2.2, the EP utility function is specification Eq. (1′) from Section 2.2, and the probability weighting function is the popular specification Eq. (9) from Section 3.1. We do not consider CPT, due to ambiguity over the interpretation of reference points in the laboratory. So this is a selective summary, guided by our views on these issues.

We assume a homogenous preferences specification, with no allowances for heterogeneity across subjects. In part, this is to anticipate their use by theorists interested in using point estimates for ''calibration finger arithmetic'' (Cox & Sadiraj, 2008). We stress that these point estimates have standard errors, and structural noise parameters, and that out-of-sample predictions of utility will have ever-expanding confidence intervals for well-known statistical reasons. We encourage theorists not to forget this simple statistical point when taking estimates such as these to make predictions over domains that they were not estimated over. Of course, the corollary is that we should always qualify estimates such as these by referencing the domain over which the responses were made. Indeed, Cox and Sadiraj (2008) show that these point estimates can produce implausible thought experiments far enough out of sample. Cox and Harrison (2008) provide further discussion of this point, using estimates from Table 8, and we return to it below.

Table 8 collects the estimates.[73] In each case the preferred model is the RDU specification with the EP utility function. Some interesting patterns emerge. First, there appears to be very little substantive probability weighting in the Hey and Orme (1994) data, even if the coefficient γ is statistically significantly different from 1: indeed, the log-likelihood of the EUT and RDU specifications with EP are close. Second, the estimates from the two implementations of the Hey and Orme (1994) design generate estimates that are remarkably similar. Third, the extent and nature of probability weighting varies significantly in the Holt and Laury (2005) data depending on the assumed utility function. Fourth, there is evidence of decreasing RRA in the Hey and Orme (1994) data and our replication, with $\alpha < 0$, but evidence of very slightly increasing RRA in the Holt and Laury (2005) data. Finally, the estimates of the concavity of the utility function do not seem to depend so much on the EUT or RDU specification, as on the choice of utility function.

To return to the point about how estimates such as these should be ''read'' by theorists, and qualified by those presenting them, consider the

Table 8. Summary Estimates.

Specification	Parameter	Estimate	Standard Error	*p*-Value[a]	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Log Likelihood
Hey and Orme (1994): $N = 80$ subjects, pooled over both tasks; 15,567 responses, excluding indifference							
EUT with	r	0.61	0.03		0.56	0.66	– 8865.01
CRRA	μ	0.78	0.06		0.67	0.90	
EUT with	r	0.82	0.02		0.80	0.84	– 8848.03
Expo-	α	– 1.06	0.04		– 1.13	– 0.99	
Power	μ	0.47	0.04		0.39	0.55	
RDU with	ρ	0.61	0.03		0.56	0.66	– 8861.18
CRRA	γ	0.99	<0.01		0.98	1.00	
	μ	0.78	0.05		0.67	0.89	
RDU with	ρ	0.82	0.01		0.80	0.84	– 8844.11
Expo-	α	– 1.06	0.04		– 1.13	– 0.99	
Power	γ	0.99	<0.01		0.98	1.00	
	μ	0.46	0.04		0.38	0.54	
Our replication of Hey and Orme (1994): $N = 63$ subjects in gain domain; 3,736 responses, excluding indifference							
EUT with	r	0.53	0.05		0.44	0.62	– 2418.62
CRRA	μ	0.79	0.06		0.67	0.91	
EUT with	r	0.78	0.02		0.74	0.82	– 2412.26
Expo-	α	– 1.10	0.05		– 1.19	– 1.00	
Power	μ	0.58	0.05		0.48	0.69	
RDU with	ρ	0.53	0.04		0.45	0.62	– 2414.46
CRRA	γ	0.97	0.01		0.95	0.99	
	μ	0.78	0.05		0.66	0.90	
RDU with	ρ	0.78	0.02		0.74	0.82	– 2408.25
Expo-	α	– 1.10	0.05		– 1.19	– 1.01	
Power	γ	0.97	0.01		0.95	0.99	
	μ	0.57	0.05		0.47	0.67	
Holt and Laury (2005): $N = 96$ subjects, pooled over 1× and 20× tasks, with no order effects; 960 non-hypothetical responses							
EUT with	r	0.76	0.04		0.68	0.84	– 330.93
CRRA	μ	0.94	0.15		0.64	1.24	
EUT with	r	0.40	0.07		0.25	0.54	– 303.94
Expo-	α	0.07	0.02		0.04	0.11	
Power	μ	0.12	0.02		0.07	0.16	
RDU with	ρ	0.85	0.08		0.69	1.00	– 325.50
CRRA	γ	1.46	0.35	0.19[b]	0.77	2.15	
	μ	0.89	0.14		0.61	1.17	
RDU with	ρ	0.26	0.05		0.16	0.36	– 288.09
Expo-	α	0.02	0.01	0.16	– 0.01	0.04	
Power	γ	0.37	0.15		0.07	0.67	
	μ	0.06	0.02		0.02	0.11	

[a] Empty cells are *p*-values that are less than 0.005.

[b] The null hypothesis here is that $\gamma = 1$.

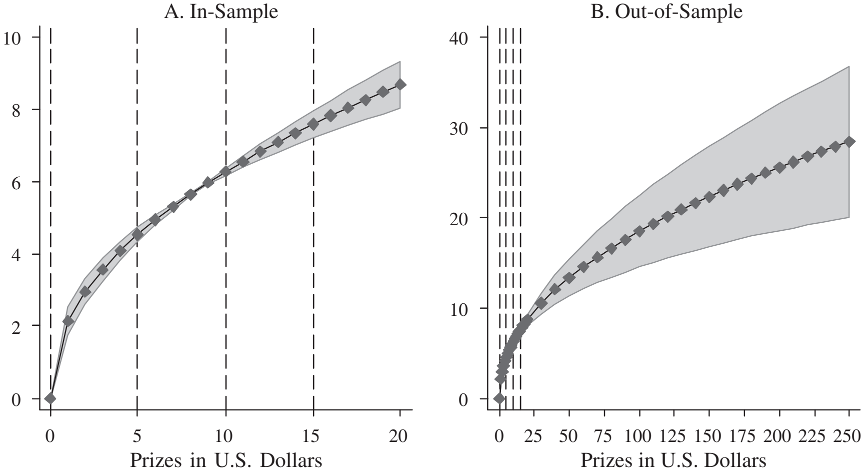


Fig. 19. Estimated In-Sample and Out-of-Sample Utility. (Estimated from Responses of 63 Subjects over 60 Binary Choices. Assuming EUT CRRA Specification with Fechner Error. Data from Our Replication of Hey and Orme (1994): Choices Over Prizes of $0, $5, $10, and $15. Point Prediction of Utility and 95% Confidence Intervals.) (A) In-sample. (B) Out-of Sample.

predicted utility values in Fig. 19. These predictions are from our replications of the Hey and Orme (1994) design, and the estimates for the EUT CRRA specification in Table 8. Fig. 19 displays predicted in-sample utility values and their 95% confidence interval using these estimates. Obviously the cardinal values on the vertical axis are arbitrary, but the main point is to see how relatively tight the confidence intervals are in relation to the changes in the utility numbers over the lottery prizes. Note the slight "flare" in the confidence interval in panel A of Fig. 19, as we start to modestly predict utility values beyond the top $15 prize used in estimation. Panel B extrapolates to provide predictions of out-of-sample utility values, up to $250, and their 95% confidence intervals. The widening confidence intervals are exactly what one expects from elementary econometrics. And these intervals would be even wider if we accounted for our uncertainty that this is the correct functional form, and our uncertainty that we had used the correct stochastic identifying assumptions. Moreover, the (Fechner) error specification used here allows for an extra element of imprecision when predicting what a subject would actually choose after evaluating the expected utility of the out-of-sample lotteries, and this does not show up in Fig. 19 since we only use the point estimate of μ.

The lesson here is that we have to be cautious when we make theoretical and empirical claims about risk attitudes. If the estimates displayed in panel A of Fig. 19 are to be used in the out-of-sample domain of panel B of Fig. 19, the extra uncertainty of prediction in that domain should be acknowledged. Cox and Sadiraj (2008) shows why we want to make such predictions, for both EUT and non-EUT specifications; we review the methods that can be used to generate these data, and econometric methods to estimate utility functions; and Wilcox (2008a) shows how alternative stochastic assumptions can have strikingly different substantive implications for the estimation of out-of-sample risk attitudes.

4. OPEN AND CLOSED QUESTIONS

We briefly review some issues which are, in our view, wide open for research or long closed.

4.1. Hypothetical Bias

Top of the ‘‘closed’’ list for us is the issue of hypothetical bias. This was a prime focus of Holt and Laury (2002, 2005), and again in Laury and Holt (2008), and has been reviewed in detail by Harrison (2007).

For some reason, however, many proponents of behavioral economics insist on using task responses that involve hypothetical choices. One simple explanation is that many of the earliest examples in behavioral economics came from psychologists, who did not use salient rewards to motivate subjects, and this tradition just persisted. Another explanation is that an influential survey by Camerer and Hogarth (1999) is widely mis-quoted as concluding that there is no evidence of hypothetical bias in such lottery choices.

What Camerer and Hogarth (1999) actually conclude, quite clearly, is that the use of hypothetical rewards makes a difference to the choices observed, but that it does not generally change the inference that they draw about the validity of EUT.[74] Since the latter typically involve paired comparisons of response rates in *two lottery pairs* (e.g., in common ratio tests), it is logically possible for there to be (i) differences in choice probabilities in *a given lottery* depending on whether one use hypothetical or real responses, and (ii) no difference between the effect of the EUT treatment on lottery *pair* responses rates depending on whether one uses hypothetical or real responses.

Furthermore, Camerer and Hogarth (1999) explicitly exclude from their analysis the mountain of data from experiments on valuation that show hypothetical bias.[75] Their rationale for this exclusion was that economic theory did not provide any guidance as to which set of responses was valid. This is an odd rationale, since there is a well-articulated methodology in experimental economics that is quite precise about the motivational role of salient financial incentives (Smith, 1982). And the experimental literature has generally been careful to consider elicitation mechanisms that provide dominant strategy incentives for honest revelation of valuations, and indeed in most instances explain this to subjects since it is not being tested. Thus, economic theory clearly points to the real responses as having a stronger claim to represent true valuations. In any event, the mere fact that hypothetical and real valuations differ so much tells us that at least *one* of them is wrong! Thus, one does not actually need to identify one as reflecting true preferences, even if that is an easy task *a priori*, in order to recognize that there are systematic and large differences in behavior between hypothetical and real responses.

4.2. Sample Selection

This is a wide-open issue that experimental economists will have to confront systematically before other researchers from labor economics do so for them. It is likely to be a significant factor in many experiments, since randomization to treatment is fundamental to statistical control in the design of experiments. But randomization implies some uncertainty about treatment condition, and individuals differ in their preferences towards taking on risk. Since human subjects volunteer for experiments, it is possible that the sample observed in an experiment might be biased because of the risk inherent in randomization. In the extreme case, subjects in experiments might be those that are *least* averse to being exposed to risk. For many experiments of biological response this might not be expected to have any influence on measurement of treatment efficacy, but other laboratory, field and social experiments measure treatment efficacy in ways that could be directly affected by randomization bias.[76]

On the other hand, the practice in experimental economics is to offer subjects a fixed participation fee to encourage attendance. These non-stochastic participation fees could offset the effects of randomization, by encouraging *more* risk-averse subjects to participate than might otherwise be the case. Thus, the term "randomization bias," in the context of economics experiments, should be taken to mean the net effects from these two latent sample selection effects.[77]

There is indirect evidence for these sample selection effects *within* the laboratory. One can recruit subjects to an experiment, conduct a test of risk attitudes, and then allow subjects to sort themselves into a given task rewarded by fixed or performance-variable payments. Cadsby, Song, and Tapon (2007) and Dohmen and Falk (2006) did just this, and show that more risk-averse subjects select into tasks with fixed rewards rather than rewards that vary with uncertain performance, and suffer in terms of expected pay. Of course, they were happy to forego some expected income in return for reduced variance. But these results strongly suggest that there would be an effect from risk attitudes if one moved the sample selection process one step earlier to include the choice to participate in the experimental session itself.[78]

Harrison, Lau, and Rutström (2005c) undertake a field experiment and a laboratory experiment to directly test the hypothesis that risk attitudes play a role in sample selection.[79] In both cases they followed standard procedures in the social sciences to recruit subjects. In their experiments the primary source of randomness had to do with the stochastic determination of final earnings, as explained below. They also employed random assignment to treatment in some experiments, but the general point applies whether the randomness is due to assignment to treatment or random determination of earnings, since the effect is the same on potential subjects. Nevertheless, it is reasonable to suspect that members of most populations from which experimenters recruit participants hold beliefs that the benefits from participating are uncertain. All that is required for sample selection to introduce a bias in the risk attitude of the participants is an expectation of uncertainty, not an actual presence of uncertainty in the experimental task.

In the field experiment it was possible to exploit the fact that the experimenter already knew certain characteristics of the population sampled, adults in Denmark in 2003, allowing a correction for sample selection bias using well-known methods from econometrics. The classic problem of sample selection refers to possible recruitment biases, such that the observed sample is generated by a process that depends on the nature of the experiment.[80] In principle, there are two offsetting forces at work in this sample selection process, as mentioned above. The use of randomization could attract subjects to experiments that are *less* risk averse than the population, if the subjects rationally anticipate the use of randomization.[81] Conversely, the use of guaranteed financial remuneration, common in experiments in economics for participation, could encourage those that are *more* risk averse to participate.

These field experiments therefore allowed an evaluation of the *net* effect of these opposing forces, which are intrinsic to any experiment in which subjects are voluntarily recruited with financial rewards. The results indicate that measured risk aversion is smaller after corrections for sample selection bias, consistent with the hypothesis that the *use of a substantial, guaranteed show-up fees more than offset any bias against attending an experiment that involved randomization*. This effect is statistically significant. The results also suggest that there is no evidence that any sample selection that occurred influenced inferences about the effects of observed individual demographic characteristics on risk aversion.

Harrison et al. (2005c) then conducted a laboratory experiment to complement the insights from their field experiment, and explore the conclusion that a larger *gross* sample selection effect might have been experienced due to randomization, but that the muted net sample selection effect actually observed was due to "lucky" choices of participation fees. The field design used the same fixed recruitment fee for all subjects, to ensure comparability of subjects in terms of the behavioral task. In the laboratory experiments this fixed recruitment fee was exogenously varied. If the level of the fixed fee affects the risk attitudes of the sample that choose to participate in the experiment, at least over the amounts they consider, one should then be able to directly see different risk attitudes in the sample. As expected *a priori*, they observed samples that were *more risk averse when a higher fixed participation fee* was used. In another treatment in the laboratory experiments they vary only the *range of the prizes* possible in the task, keeping the fixed participation fee constant, but announcing these ranges at the time of recruitment. In this case, they observed samples that were *more risk averse when the range of prizes was widened*, compared to the control. Hence, the level of the fixed recruitment fee and information on the range of prizes in the experiment had a direct influence on the composition of the sample in terms of individual risk attitudes.

The implication is that experimental economists should pay much more attention to the process that leads subjects to participate in the experiment if they are to draw reliable inferences in any setting in which risk attitudes play a role. This is true whether one conducts experiments in the laboratory or the field.[82]

A closely related issue is what role risk attitudes may play in affecting subjects' participation choices over different institutions or cohorts when such choices are allowed.[83] It is common in the experimental literature to study behavior in two or more institutions imposed exogenously on subjects, or to put subjects together exogenously. But in the naturally occurring world that our experiments are modeling, people choose institutions to some degree, and choose who to interact with to some degree. The effect of treatments may be completely different when people have some ability to select into them, or some ability to choose the cohorts to participate with, compared to the standard experimental paradigm. In effect, the experiment just has to be widened to include these processes of selection, if appropriate for the behavior under study. The broader experimental literature now identifies many possible mechanisms for this process, such as migration from one region to another in which local public policies exhibit differences (Ehrhart and Keser (1999), Page, Putterman and Unel (2005), Gürerk, Irlenbusch, and Rockenbach (2006)), voting in an explicit social choice setting (Botelho, Harrison, Pinto, & Rutström, 2005a; Ertan, Page, & Putterman, 2005; Sutter, Haigner, & Kocher, 2006), lobbying for policies (Bullock & Rutström, 2007), and even the evolution of social norms of conduct (Falk, Fehr, & Fischbacher, 2005). Each of these processes will interact with the risk attitudes of subjects.

4.3. Extending Lab Procedures to the Field

One of the main attractions of experimental methods is the control that it provides over factors that could influence behavior. The ability to control the environment allows the researcher to study the effects of treatments in isolation, and hence makes it easier to draw inferences as to what is influencing behavior. In most cases we are interested in making inferences about field behavior. We hypothesize that there is a danger that the imposition of an exogenous laboratory control might make it harder, in some settings, to make reliable inferences about field behavior. The reason is that the experimenter might not understand something about the factor being controlled, and might impose it in a way that is inconsistent with the way it arises naturally in the field, and that affects behavior.

Harrison et al. (2007c) take as a case study the elicitation of measures of risk aversion in the field. In the traditional paradigm, risk aversion is viewed in terms of diminishing marginal utility of the final prize in some abstract lottery. The concept of a lottery here is just a metaphor for a real lottery, although in practice the metaphor has been used as the primary vehicle for laboratory elicitation of risk attitudes. In general there is some commodity x and various levels i of x, x_i, that depend on some state of nature which occurs with a probability p_i that is known to the individual whose preferences are being elicited. Thus, the lottery is defined by $\{x_i; p_i\}$. Traditional measures of risk aversion under EUT are then defined in terms of the curvature of the utility function with respect to x.

Now consider the evaluation of risk attitudes in the field. This generally entails more than just "leaving the classroom" and recruiting outside of a university setting, as emphasized by Harrison and List (2004). In terms of sample composition, it means finding subjects who deal with that type of uncertainty to varying degrees, and trying to measure the extent of their field experience with uncertainty. Moreover, it means developing stimuli that more closely match those that the subjects have previously experienced, so that they can use whatever heuristics they have developed for that commodity when making their choices. Finally, it means developing ways of communicating probabilities that correspond with language that the subjects are familiar with. Thus, field experimentation in this case, and in general, involves several simultaneous changes from the lab setting with respect to subject recruitment and the development of stimuli that match the field setting.

Apart from sample and task selection issues a second theme that is important to the relevance of lab findings to field inferences is the influence of "background risk" on the attitudes towards a specific "foreground risk" that is the focus of the elicitation task. In many field settings it is not possible to artificially identify attitudes towards one risk source without worrying about how the subjects view that risk as being correlated with other risks. For example, mortality risks from alternative occupations tend to be highly correlated with morbidity risks. It is implausible to ask subjects their attitude toward one risk without some coherent explanation as to why a higher or lower level of that risk would not be associated with a higher or lower risk of the other.

Apart from situations where risks may be correlated, "background risk" can have an influence on elicited risk attitudes also when it is independent of the "foreground risk." The theoretical literature has also yielded a set of preferences that guarantee that the addition of an unfair background risk to wealth reduces the CE of any other independent risk. That is, the addition of background risk of this type makes risk-averse individuals behave in a more risk averse way with respect to any other independent risk. Gollier and Pratt (1996) refer to this type of behavior as "risk vulnerability," and show that all weakly Decreasing Absolute Risk Averse utility functions are risk vulnerable. This class includes many popular characterizations of risk attitudes, such as CARA and CRRA. Eeckhoudt, Gollier, and Schlesinger (1996) extend these results by providing the necessary and sufficient conditions on the characterization of risk aversion to ensure that any increase in background risk induces more risk aversion.

The field experiment in Harrison et al. (2007c) is designed to analyze such situations of independent multiple risk. The compare using monetary prizes to using prizes whose values involve some risk and conclude that the risk attitudes elicited are not the same in the two circumstances. These prizes are collector coins and the subjects are numismatists. They find that the subjects are generally more risk averse over the prizes when the latter involve additional, and independent, risk.[84] These results are consistent with the available theory from conventional EUT for the effects of background risk on attitudes to risk. Thus, applying risk preferences that have been elicited in the lab to field settings with background risks may underestimate the extent to which decisions will reflect risk aversion. In addition, eliciting risk attitudes in a natural field setting with natural tasks and non-monetary prizes requires one to consider the nature and degree of background risk, since it is inappropriate to ignore.[85]

A further virtue of extending lab procedures to the field, therefore, is to encourage richer lab designs by forcing the analyst to account for realistic features of the natural environment that have been placed aside. In virtually any market with asymmetric information, whether it is a coins market, an open-air market, or a stock exchange, a central issue is the quality of the object being traded. This issue, and attendant uncertainty, arises naturally. In many markets, the grade of the object, or professional certification of the seller, is one of the critical variables determining price. Thus, one could scarcely design a test of foreground risk in these markets without attending to the background risk. Harrison et al. (2007c) exploit the fact that such risks can be exogenously controlled in these settings, and in a manner consistent with the predictions of theory.[86]

In a complementary manner, Fiore, Harrison, Hughes, and Rutström (2007) consider how one can use simulation tools to represent "naturally occurring probabilities." As one moves away from the artifactual controls of the laboratory, distributions of outcomes are not always discrete, and probabilities are not given from outside. They are instead estimated as the result of some process that the subject perceives. One approach to modeling such naturally occurring probabilities in experiments is to write out a numerical simulation model that represents the physical process that stochastically generates the outcome as a function of certain inputs, render that process to subjects in a natural manner using tools of Virtual Reality, and study how behavior changes as one changes the inputs. For example, the probability that a wildfire will burn down a property "owned" by the subject might depend on the location of the property, the vegetation surrounding it, the location of the start of the wildfire, weather conditions, and interventions that the subject can choose to pay for to reduce the spread of a wildfire (e.g., prescribed burning). This probability can be simulated using a model, such as FARSITE developed by the U.S. Forest Service (Finney, 1998) to predict precisely these events. Thus, the subject sees a realistic rendering of the process generating a distribution over the binary outcome, "my property burns down or not." By studying how subjects react to this process, one can better approximate the manner in which risk attitudes affect decisions in naturally occurring environments.

5. CONCLUSION

At a substantive level, the most important conclusion is that the average subject is moderately risk averse, but there is evidence of considerable individual heterogeneity in risk attitudes in the laboratory. This heterogeneity is in evidence within given elicitation formats, so it cannot be ascribed to differences in elicitation formats. The range of risk attitudes is modest, however, and there is relatively little evidence of risk-loving behavior. The temptation to talk about a "central tendency" of "slight risk aversion" does not fit well with the bi-modal nature of the responses observed in several studies: a large fraction of subjects is well characterized as being close to risk neutral, or very slightly risk averse, and another large fraction as being quite risk averse.

At a methodological level, the evidence suggests some caution in expecting different elicitation formats to generate comparable data on risk attitudes. Both the framing of the questions and the implied incentives differ across instruments and may affect responses. One would expect the MPL and RLP procedures to generate comparable results, since they are so similar from a behavioral perspective, and they do. The OLS instrument is very portable in the field, has transparent incentives for truthful responses, and is easy to administer in all environments, so more work comparing its performance to the MPL and RLP instruments would be valuable. It suffers from not being able to provide a rich characterization of behavior when allowances are made for probability weighting, but that may be mitigated with extensions to consider probabilities other than 1/2.

In the Epilogue to a book-length review of the economics of risk and time, Gollier (2001; p.424ff.) writes that

> It is quite surprising and disappointing to me that almost 40 years after the establishment of the concept of risk aversion by Pratt and Arrow, our profession has not yet been able to attain a consensus about the measurement of risk aversion. Without such a consensus, there is no hope to quantify optimal portfolios, efficient public risk prevention policies, optimal insurance deductibles, and so on. It is vital that we put more effort on research aimed at refining our knowledge about risk aversion. For unclear reasons, this line of research is not in fashion these days, and it is a shame.

The most important conclusion we draw from our survey is that reliable laboratory methods exist to determine the individual aversion to risk of a subject, or to characterize the distribution of risk attitudes of a specific sample. These methods can now be systematically employed to ensure greater control over tests and applications of theory that depend on risk attitudes.

NOTES

1. For example, in virtually all experimental studies of non-cooperative bargaining behavior. A particularly striking example is provided by Ochs and Roth (1989), since Roth and Malouf (1979) pioneered the use of experimental procedures to induce risk neutral behavior in cooperative bargaining settings.

2. For example, in virtually all experimental studies of bidding behavior in first-price auctions, whether in private values settings (Cox et al., 1982) or common values settings (Kagel & Levin, 2002).

3. For example, the experimental literature on bidding behavior in first-price sealed bid auctions relies on predictions that are conditioned on the subjects following some Nash Equilibrium strategy as well as being characterized by risk in some way. Overbidding in comparison to the risk-neutral prediction could be due to failure of either the assumption of Nash bidding or the failure of the assumption of risk neutrality (Section 3.6). Harrison (1990) and Cox, Smith, and Walker (1985) attempt to tease these two possibilities apart using different designs.

4. We do not consider experimental designs that attempt to control for risk, or induce specific risk attitudes. Our general focus is on direct estimation of risk attitudes where rewards are real and there is some presumption that the procedure is incentive compatible. There is a huge, older literature on the elicitation of utility, but virtually none of it is concerned with incentive compatibility of elicitation, which we take as central. Great reviews include Fishburn (1967) and Farquhar (1984). Many components of the procedures we consider can be viewed as building on methods developed in this older literature. Biases in utility elicitation procedures are reviewed by Hershey et al. (1982), although again there is no discussion at all of incentive compatibility or hypothetical rewards bias.

5. Birnbaum (2004) illustrates the type of systematic comparison of representations that ought to be built in for broader research programs. He considers various representations of probability in terms of text, pie charts, natural frequencies, and alignments of equally likely consequences, as well as minor variants within each type of representation. One reason for this focus is his concern with violations of stochastic dominance, which is an elemental behavioral property of decisions, and presumed to be directly affected by task representation. In brief, he finds little effect on the extent of stochastic dominance of these alternative representations. That conclusion is limited to the hypotheses he considers, of course; there could still be an effect on structural estimates of underlying models, and other hypotheses derived from those estimates. Unfortunately, the procedures he uses, still common in the psychology and decision-making literature, employ hypothetical or near-hypothetical rewards for subjects to make salient decisions. Wakker, Erev, and Weber (1994) considered four types of representations, shown in Appendix A, in salient choices, but provide no evaluation of the effects of the alternatives.

6. There is an interesting question as to whether they should be provided. Arguably some subjects are trying to calculate them anyway, so providing them avoids a test of the joint hypothesis that "the subjects can calculate EV in their heads and will not accept a fair actuarial bet." On the other hand, providing them may cue the subjects to adopt risk-neutral choices. The effect of providing EV information deserves empirical study.

7. The last row does have the advantage of helping subjects see that they should obviously switch to option B by the last row, and hence seeing the ordered nature of the overall instrument. Arguably it would be useful to add a row 0 in which the lower prize for options A and B were obtained with certainty, to help the subject see that they should always choose A at the top and B at the bottom, and the only issue is where they should switch.

8. Schubert et al. (1999) present their method as the elicitation of a certainty-equivalent, but do not say clearly how they elicited the certainty-equivalent. In fact (Renate Schubert; personal communication) their procedures represent an early application of the MPL idea. Each subject was asked to choose between two lotteries, where one lottery was the risky one and the other degenerate lottery was a non-stochastic one. They asked subjects 98 binary choice questions, spanning 8 risky lotteries. These were arrayed in an ordered fashion on 98 separate sheets. The responses could then be ordered in increasing values for the non-stochastic lottery, and a "switch point" determined to identify the certainty-equivalent.

9. If the subject always chooses A, or indicates indifference for any of the decision rows, there are no additional decisions required and the task is completed.

10. Let the first stage of the iMPL be called Level 1, the second stage Level 2, and so on. After making all responses, the subject has one row from the first table of responses in Level 1 selected at random by the experimenter. In the MPL and sMPL procedures, that is all there is since there is only a Level 1 table. In the iMPL, that is all there is if the row selected at random by the experimenter is *not* the one at which the subject switched in Level 1. If it *is* the row at which the subject switched, another random draw is made to pick a row in the Level 2 table. For some tasks this procedure is repeated to Level 3.

11. In our experience subjects are suspicious of randomization generated by computers. Given the propensity of many experimenters in other disciplines to engage in deception, we avoid computer randomization whenever feasible.

12. Dave et al. (2007) draw similar conclusions, and include an explicit comparison in the field with the Holt and Laury (2002) MPL instrument. They also collect information on the cognitive abilities of subjects, to better identify the sources of any differences in behavior.

13. Millner, Pratt, and Reilly (1988) offered some important, critical observations on the design and analysis proposed by Harrison (1986). There is possible contamination from intra-session experimental earnings if the subject is paid for each selling price elicited, but this issue is common to all of the methods. One could either assume these wealth effects away (Harrison, 1986; Kachelmeier & Shehata, 1992), test for them (McKee, 1989), or pay subjects for just one of the stages. The last of these options is now the standard method when applying BDM, but raises the same issues with the validity of the random lottery incentive mechanism that have been discussed for other procedures (see Section 3.8).

14. One must also ensure that the buyout range exceeds the highest price that the subject would reasonably state, but this is not a major problem.

15. The same "payoff dominance problem" applies to first-price auctions, as noted by Harrison (1989). Hence, both of the institutions used by Isaac and James (2000) to infer risk attitudes are blighted with this problem. The same problem applies to two of the three institutions studied by Berg, Dickhaut, and McCabe (2005). Their third institution, the English auction, is known to have more reliable behavioral incentives for truthful responses (Harstad, 2000; Rutström, 1998).

16. Assume a risk neutral subject facing a MPL with prizes $20 and $16 for the safe lottery and $38.50 and $1 for the risky one. Such a subject should choose the risky lottery for rows 1 through 4 and then switch to the risky one. Not doing so would result in an expected earnings loss. For example, if he erroneously responds as if he is slightly risk loving by choosing the risky lottery already on row 4 he is forgoing $1.60, and if he erroneously responds as if he is slightly risk averse by still choosing the safe lottery on row 5, he is foregoing $1.75. Since the chances are 1 in 10 that the row with his erroneous choice is picked, his expected foregone earnings are about 16 to 17.5 cents. If he instead were asked to state his minimum WTA for each of the lotteries in a BDM, his true WTA when the probabilities correspond to those given in row 5 of the MPL (i.e., 50/50) would be $18 for the safe and $19.75 for the risky lottery. We can then calculate the expected loss from different misrepresentations of his preferences in ways that are comparable to those calculated for the MPL. To find the expected loss from representing his preferences as if they were defined over the safe MPL lottery given on row 4 we simply calculate the maximum WTA for the safe lottery on row 4 as $17.60. If this is his stated WTA he will experience a loss if the BDM selling price is between this report and his true WTA ($18).The likelihood for this is obviously quite small. On the other hand, the expected loss of a similarly erroneous report for the risky lottery would involve a report of $16 for a true maximum WTA of $19.75, a much stronger incentive. Again, the likelihood of this loss is the likelihood of the BDM selling price falling in between the stated and the true WTA. This likelihood is a function of the range of the buying prices used in

the particular implementation of the BDM. The narrower the range the higher is this likelihood. It is clear from this numeric example that the incentive properties of the BDM are much worse than those for the MPL for the safe lottery, but quite a bit better for the risky lottery. One problem with the BDM is that for a risk-loving subject who would state a high WTA for the lottery the probability of the BDM drawing a number higher than or equal to his WTA is low. Thus, the incentives for precision are low for such a subject.

17. Millner et al. (1988; p. 318) suggest that one should develop methods for identifying inconsistent responses, although they would agree that the original checks built in by BDM have some flaws, since the lotteries offered to subjects at later stages depend on earlier elicited selling prices. This sounds attractive in the abstract, but we caution against the use of mechanical rules for classifying subjects as inconsistent. For example, erratic responses could just be a reflection that the subject rationally perceives the absence of a strong incentive to respond truthfully. Classifying such subjects as inconsistent is inappropriate.

18. The former asks the subject to state a certain amount that makes them indifferent to the lottery, similar to what is done in the BDM, and the latter asks the subject to state some probability in the lottery that makes them indifferent to some fixed and certain amount, similar to what is done in the OLS. The latter method presumes that there are only two outcomes, and hence one probability.

19. Abdellaoui (2000) did introduce the use of a bisection method for establishing indifference in each stage that might mitigate some strategic concerns. The idea is to only allow subjects to pick one of two given lotteries, and not to state the indifference lottery directly. By starting this process at some *a priori* extreme pair, one can iterate down to the point of indifference using a conventional bisection search algorithm. In this instance the chaining strategy is limited to always picking the lottery with the highest possible prize. This method was also used by Kuilen, Wakker, and Zou (2007), and has the advantage of limiting the financial exposure of the experimenter to known bounds. Of course, subjects might not adopt the chaining strategy in the logically extreme form, perhaps to avoid being dismissed from the experiment or not being invited back again, but still be generating strategically biased responses.

20. The TO method has also been extended by Attema, Bleichrodt, Rohde, and Wakker (2006) to elicit discount rates. The same incentive compatibility problems apply, only hypothetical experiments are conducted, and there is no discussion of the problems of incentivizing responses.

21. We use the term "risk attitudes" here in the broader sense of including possible effects from non-linear utility functions, probability weighting *and* loss aversion.

22. Andersen et al. (2008a) and Section 3.4 discuss the elicitation of risk preference and time preferences, and the need for *joint* estimation of all parameters. The basic idea is that the discount rate involves the present value of utility streams, and not money streams, so one needs to know the concavity of the utility function to infer discount rates. In effect, the TCN procedure assumes risk neutrality when inferring discount rates, which will lead to overestimates of discount rates between utility flows.

23. We consider the use of such interval bounds for estimation in Section 2.1. Having some bounds that span a finite number and ∞ does not pose problems for

the "interval regression" methods widely available, although it does correctly lead to larger standard errors than collapsing this interval to just the lower bound.

24. Some subjects switched several times, but the minimum switch point is always well defined. It turns out not to make much difference how one handles these "multiple switch" subjects, but our analysis and the analysis of HL considers the effect of allowing for them in different ways explained below.

25. HL find that there is a significant sex effect in the low-payoff conditions, with women being more risk averse, and no effect in the high payoff conditions. We replicate this conclusion using their procedures and data. Unfortunately, the low-payoff sex effect does not hold if one controls for the other characteristics of the subject and uses a negative binomial regression model to handle the discrete nature of the dependant variable. HL also report that there is a significant Hispanic effect, with Hispanic subjects making fewer risk-averse choices in high payoff conditions. We confirm this conclusion, using their procedures as well as when one uses all covariates in a negative binomial regression.

26. A subject that switched from option A to option B after five safe choices, then switched back for one more option A before choosing all B's in the remaining rows, would therefore have revealed a CRRA interval between 0.15 and 0.97. Such subjects simply provide less precise information than subjects that switch once.

27. Our treatment of indifferent responses uses the specification developed by Papke and Wooldridge (1996; Eq. 5, p. 621) for fractional dependant variables. Alternatively, one could follow Hey and Orme (1994; p. 1302) and introduce a new parameter τ to capture the idea that certain subjects state indifference when the latent index showing how much they prefer one lottery over another falls below some threshold τ in absolute value. This is a natural assumption to make, particularly for the experiments they ran in which the subjects were told that expressions of indifference would be resolved by the experimenter, but not told how the experimenter would do that (p. 1295, footnote 4). It adds one more parameter to estimate, but for good cause.

28. Clustering commonly arises in national field surveys from the fact that physically proximate households are often sampled to save time and money, but it can also arise from more homely sampling procedures. For example, Williams (2000; p. 645) notes that it could arise from dental studies that "collect data on each tooth surface for each of several teeth from a set of patients" or "repeated measurements or recurrent events observed on the same person." The procedures for allowing for clustering allow heteroskedasticity between and within clusters, as well as autocorrelation within clusters. They are closely related to the "generalized estimating equations" approach to panel estimation in epidemiology (see Liang & Zeger, 1986), and generalize the "robust standard errors" approach popular in econometrics (see Rogers, 1993). Wooldridge (2003) reviews some issues in the use of clustering for panel effects, noting that significant inferential problems may arise with small numbers of panels.

29. Age was imputed as 20 for all subjects in the undergraduate class experiments conducted at the University of New Mexico, based on personal knowledge of the experimenters of the age distribution in those classes (Kate Krause, personal communication). Given the variation in age for non-student adults, this imputation is less likely to be a major factor compared to studies that only use student subjects.

30. That is, we treat the prizes here as gains measured as the net gain after deducting losses from the endowment. This analysis still allows for a framing effect, of course.

31. See Harless and Camerer (1994), Hey and Orme (1994) and Loomes and Sugden (1995) for the first wave of empirical studies including some formal stochastic specification in the version of EUT tested. There are several species of "errors" in use, reviewed by Hey (1995, 2002), Loomes and Sugden (1995), Ballinger and Wilcox (1997), Loomes, Moffatt, and Sugden (2002) and Wilcox (2008a). Some place the error at the final choice between one lottery or the other after the subject has decided deterministically which one has the higher expected utility; some place the error earlier, on the comparison of preferences leading to the choice; and some place the error even earlier, on the determination of the expected utility of each lottery.

32. This ends up being simple to formalize, but involves some extra steps in the economics. Let $\mathrm{EU_R}$ and $\mathrm{EU_L}$ denote the expected utility of lotteries R and L, respectively. If we ignore indifference, and the subject does not make mistakes, then R is chosen if $\mathrm{EU_R} - \mathrm{EU_L} > 0$, and otherwise L is chosen. If the subject makes measurement errors, denoted by ε, then the decision is made on the basis of the value of $\mathrm{EU_R} - \mathrm{EU_L} + \varepsilon$. That is, R is chosen if $\mathrm{EU_R} - \mathrm{EU_L} + \varepsilon > 0$, and otherwise L is chosen. If ε is random then the probability that R is chosen $= P(\mathrm{EU_R} - \mathrm{EU_L} + \varepsilon > 0) = P(\varepsilon > -(\mathrm{EU_R} - \mathrm{EU_L}))$. Now suppose that ε is normally distributed with mean 0 and standard deviation σ, then it follows that $Z = \varepsilon/\sigma$ is normally distributed with mean 0 and standard deviation 1: in other words, Z has a unit normal distribution. Hence, the probability that R is chosen is $P(\varepsilon > -(\mathrm{EU_R} - \mathrm{EU_L})) = P(\varepsilon/\sigma > -(\mathrm{EU_R} - \mathrm{EU_L})/\sigma)$. If $\Phi(\cdot)$ denotes the cumulative normal standard distribution, it follows that the probability that R is chosen is $1 - \Phi(-(\mathrm{EU_R} - \mathrm{EU_L})/\sigma) = \Phi((\mathrm{EU_R} - \mathrm{EU_L})/\sigma)$, since the distribution is symmetrical about 0. Hence, the probability that B is chosen is given by: $\Phi(-(\mathrm{EU_R} - \mathrm{EU_L})/\sigma) = 1 - \Phi((\mathrm{EU_R} - \mathrm{EU_L})/\sigma)$. If we denote by y the decision of the subject with $y = 1$ indicating that R was chosen and $y = -1$ indicating that L was chosen, then the likelihood is $\Phi((\mathrm{EU_R} - \mathrm{EU_L})/\sigma)$ if $y = 1$ and $1 - \Phi((\mathrm{EU_R} - \mathrm{EU_L})/\sigma)$ if $y = -1$.

33. We also correct for clustering, since it is the right thing to do statistically, but this again makes no essential difference to the estimates.

34. The instructions were brief: "Your decision sheet shows 8 options listed on the left. You should choose one of these options, which will then be played out for you. If the coin toss is a Heads you will receive the amount listed in the second column. If the coin toss is a Tail you will receive the amount listed in the third column." The transparency of the OLS procedure is apparent, and derives from only using probabilities of 1/2.

35. The secondary purpose of this design is to allow statistical examination of the hypothesis that subjects use "similarity relations" and "editing processes" to evaluate lotteries when prizes and probabilities are not pre-rounded, as in Hey and Orme (1994).

36. The use of the noise parameter μ in Eq. (8) is also familiar from the numerical literature on the smoothing of accept–reject simulators in discrete choice statistical modeling: see Train (2003; p. 125ff.), for example. This connection also reminds us that the use of specific linking functions such as logit or probit have a certain

arbitrariness to them, but embody implicit behavioral assumptions about responsiveness to latent indices.

37. A more complete statistical analysis would consider two factors: the effect of information about earnings in the prior procedure, and a more elaborate likelihood function that recognized that these are in-sample responses. Our estimates ignore both factors. It would also be useful to examine the experimental data from Engle-Warnick et al. (2006) using inferential methods such as ours, since their design used exactly the same lotteries in the RLP and OLS instruments. Dave et al. (2007) provide careful tests of the MPL and OLS instruments, concluding that the OLS instrument provides a more reliable measuring rod for risk attitudes in samples drawn from populations expected to have limited cognitive abilities.

38. Hirshleifer and Riley (1992) and Chambers and Quiggin (2000) demonstrate the elegant and powerful representations of decision-making under uncertainty that derive from adopting a state-contingent approach instead of popular alternatives.

39. Many of these claims involve evidence from between-sample designs, and rely on the assumption that sample sizes are large enough for randomization to ensure that between-sample differences in preferences (even if they are not state-contingent) are irrelevant. For two careful examples, see Conlisk (1989) and Cubitt et al. (1998a). There is also a rich literature on the contextual role of extreme lotteries, such that one often observes different behavior for "interior lotteries" that assign positive probability to all prizes as compared to "corner-solution lotteries" that assign zero weight to some prizes.

40. Stigler and Becker (1977; p. 76) note the nature of the impasse: "an explanation of economic phenomena that reaches a difference in tastes between people or times is the terminus of the argument: the problem is abandoned at this point to whoever studies and explains tastes (psychologists? anthropologists? phrenologists? socio-biologists?)."

41. Camerer (2005; p. 130) provides a useful reminder that "Any economics teacher who uses the St. Petersburg paradox as a "proof" that utility is concave (and gives students a low grade for not agreeing) is confusing the sufficiency of an explanation for its necessity."

42. Of course, many others recognized the basic point that the distribution of outcomes mattered for choice in some holistic sense. Allais (1979; p. 54) was quite clear about this, in a translation of his original 1952 article in French. Similarly, in psychology it is easy to find citations to kindred work in the 1960s and 1970s by Lichtenstein, Coombs and Payne, *inter alia*.

43. There are some well-known limitations of the probability weighting function Eq. (9). It does not allow independent specification of location and curvature; it has a crossover-point at $p = 1/e = 0.37$ for $\gamma < 1$ and at $p = 1 - 0.37 = 0.63$ for $\gamma > 1$; and it is not increasing in p for small values of γ. Prelec (1998) and Rieger and Wang (2006) offer two-parameter probability weighting functions that exhibits more flexibility than Eq. (9), but for our expository purposes the standard probability weighting function is adequate.

44. In this case, because each lottery only consists of two outcomes, the "rank dependence" of the RDU model does not play a distinctive role, but it will in later applications.

45. The estimates of the coefficient obtained by Tversky and Kahneman (1992) fortuitously happened to be the same for losses and gains, and many applications of PT assume that for convenience. The empirical methods of Tversky and Kahneman (1992) are difficult to defend, however: they report median values of the *estimates* obtained after fitting their model for each subject. The estimation for each subject is attractive if data permits, as magnificently demonstrated by Hey and Orme (1994), but the *median estimate* has nothing to commend it statistically.

46. In other words, evaluating the PU of two lotteries, without having edited out dominated lotteries, might lead to a dominated lottery having a higher PU. But if subjects always reject dominated lotteries, the choice would appear to be an error to the likelihood function. Apart from searching for better parameters to explain this error, as the ML algorithm does as it tries to find parameter estimates that reduce any other prediction error, our specification allows μ to increase. We stress that this argument is not intended to rationalize the use of separable probability weights in OPT, just to explain how a structural model with stochastic errors might account for the effects of stochastic dominance. Wakker (1989) contains a careful account of the notion of transforming probabilities in a "natural way" but without violating stochastic dominance.

47. One of the little secrets of CPT is that one must always have a probability weight for the residual outcome associated with the reference point, and that the reference outcome receive a utility of 0 for both gains and losses. This ensures that decision weights always add up to 1.

48. An alternative specification would be to take the negative of the utility function defined over the gross losses, in effect assuming $\lambda = 1$ from the CPT specification.

49. A corollary is that it might be a mistake to view loss aversion as a fixed parameter λ that does not vary with the context of the decision, *ceteris paribus* the reference point.

50. The mean estimate from their sample was $31, but there were clear nodes at $15 and $30. Our experimental sessions typically consist of several tasks, so expected earnings from the lottery task would have been some fraction of these expectations over session earnings. No subject stated an expected earning below $7.

51. A concrete implication, considered at length in Harrison and Rutström (2005; Section 5), is that the rush to use non-nested hypothesis tests is misplaced. If one reads the earlier literature on those tests it is immediately clear that they were viewed as poor, second-best alternatives to writing out a finite mixture model and estimating the weights that the data place on each latent process. The computational constraints that made them second-best decades ago no longer apply.

52. See Keller and Strazzera (2002; p. 148) and Frederick, Loewenstein, and O'Donoghue (2002; p. 381ff.) for an explicit statement of this assumption, which is often implicit in applied work. We refer to risk aversion and concavity of the utility function interchangeably, but it is concavity that is central (the two can differ for non-EUT specifications).

53. Harless and Camerer (1994) do consider different ways that one can compare different theories that have different numbers of "free parameters." They also

consider simple metrics for violations, but even these are still defined in terms of the number of failures of the theory in a given triple (e.g., one failure out of three predictions is considered better from the perspective of the theory than two failures out of three).

54. Some semi-parametric estimators, such as the Maximum Score estimator of Manski, do rely on "hit rates" as a metric.

55. Some experiments attempt to design checks for some of the more obvious biases, such as which lottery is presented on the left or right, or whether the lotteries are ordered best to worst or vice versa (e.g., see Harless, 1992; Hey & Orme, 1994).

56. Problem 2 in CSS involves losing three subjects at random for every one subject that was actually asked to make a choice, whereas the other problems involved all recruited subjects making a choice. Hence 200 subjects were recruited to Problem 2, and the eventual sample of choices was roughly 50 subjects for each problem, by design.

57. Comparing only Problems 1 and 5 in CSS, which involve choices only over simple lotteries, the evidence against EUT is even weaker.

58. The word "essentially" reminds us that this is EUT with some explicit stochastic error story. There are many alternative error stories, of course. Wilcox (2008a, 2008b) explores the deeper modeling issues of writing out a theory without specifying any stochastic process connecting it to data.

59. Some might object that even if the behavior can be formally explained by some small error, there are systematic behavioral tendencies that are not consistent with a white-noise error process. Of course, one can allow asymmetric errors or heteroskedastic errors.

60. Wakker et al. (1994) in effect adopted such a design. Their primary tasks deliberately had comparable expected values in the paired lotteries subjects were to choose over, but their "filler" tasks were then deliberately set up to have different expected values.

61. See Kagel (1995) and Harrison (1989, 1990) for a flavor of the debates.

62. Harrison, List, and Tra (2005e) show, however, that when auctions consist of more and more bidders, received theory does increasingly poorly in terms of characterizing "one shot" behavior. Their evidence suggests that received theory is relevant for "small auctions" but not for "large auctions." Thus, if one were testing received theory it would matter on what domain the data were generated. Cox et al. (1982) reported different results, with the smallest of their auctions ($N = 3$) generating the data that seemed to most obviously contradict the risk-averse Nash Equilibrium bidding model. However, this could have been due to collusion. In *all* of their experiments the same N bidders participated in multiple rounds, facilitating coordination of collusive under-bidding strategies that wreak havoc with the one-shot predictions of the theory.

63. Cox et al. (1988) offer a generalization that admits of some degrees of risk-loving behavior. Since we do not observe much risk loving in the population used in these experiments, college students in the United States, this extension is not needed for present purposes.

64. That is, 1/2 is arguably more focal than 2/3 or 3/4, and so on for $N>2$. It is certainly easier to implement arithmetically, absent calculating aids.

65. Unfortunately, there is evidence that subjects may not see it this way. In a generic public goods voluntary contribution game Botelho, Harrison, Pinto, and Rutström (2005b) show that Random Strangers designs do not generate the same behavior as Perfect Strangers designs in which the subject is guaranteed not to meet the same opponent twice.

66. One might be concerned that the full model fits the RA NE bidding model simply because it has a "free parameter r_i" to fit the bidding data to. In some sense this is true, since the joint likelihood of the data includes the effect of different $\hat{r}_i$'s on bids, and the estimates seek $\hat{r}_i$ values that explain the bidding data best. But it is not true entirely, since the joint likelihood must also explain the risk attitude choice data as well. One can formally compare the distribution of predicted risk attitudes if one only uses the risk aversion tasks and the distribution that is generated if one uses all data simultaneously. The two distributions are virtually identical. Kendall's τ statistic can be used to test for rank correlation; it has a value of 0.82, and leads one to reject the null hypothesis that the two sets of estimates of risk attitudes are independent at p-values below 0.0001.

67. Additional experimental tests include Thaler, Tversky, Kahneman, and Schwartz (1997) and Gneezy, Kapteyn, and Potters (2003). These provide results that are qualitatively identical, but harder to evaluate. Thaler et al. (1997) did not provide subjects with precise knowledge of the probabilities involved in the lotteries, but allowed them to infer that over time; hence behavior could have been driven by mistakes in the subjective inference of probabilities rather than MLA. Gneezy et al. (2003) embed the task in an asset market, which may have influenced individual behavior in other ways than predicted by EUT or MLA. These influences are of interest, since markets are the institution in which most stocks and bonds are traded, but from the perspective of wanting the cleanest possible test of competing theories those extra influences are just a confound. Camerer (2005) and Novemsky and Kahneman (2005a, 2005b) provide an overview of the history and current status of the loss aversion hypothesis.

68. It is also possible to augment the estimation procedure to include a parameter that can be interpreted as "baseline consumption," to which prizes are added before being evaluated using the utility function. This approach has been employed by Harrison et al. (2007c) and Heinemann (2008). Andersen et al. (2008a) consider the theoretical and empirical implications of this approach in detail.

69. The term "portfolio effects" is unfortunate, since it suggests a concern with correlated risks and risk pooling, which is not the issue here. Unfortunately, we cannot come up with a better expression, and this one has some currency in the literature.

70. For K binary choices it is 2^K, assuming that indifference is not an option. For $K = 10$ this is only 1,024, but for $K = 15$ it is 32,768, and one can guess the rest for larger K. The use of random lottery incentives in the context of the Random Lottery Pair elicitation procedure raises some deep modeling issues of sequential choice, since it introduces the interaction of risk aversion and ambiguity aversion with respect to future lotteries (Klibanoff, Marinacci, & Mukerji, 2005; Nau, 2006), as well as concerns with possible preferences over the temporal resolution of uncertainty (Kreps & Porteus, 1978). In effect, this is a setting in which the "small world" assumption of Savage (1972; Section 5.5), under which one focuses on

isolated decisions and ignores the broader context, may be particularly appropriate to apply. It may not be appropriate to apply for other tasks, as we discuss below.

71. Harrison et al. (2007; fn.16) report a direct test of the random lottery procedure with the MPL instrument, and note that it did not change elicited risk attitudes *assuming* EUT to infer risk attitudes.

72. In fact, Wilcox (2008a) recommends a third alternative specification, the Contextual Utility model developed in Wilcox (2008b), over both the Luce and Fechner specifications. If the choice is between Luce and Fechner, however, his discussion clearly favors Fechner. The estimates from Holt and Laury (2005) presented in Section 3.1 used the Luce specification, and hence differ from those presented here.

73. These do not exactly replicate all estimates presented earlier since there are slight differences in specifications.

74. With one exception, we do not believe that this inference is supported by the existing data and experimental designs. That exception is Beattie and Loomes (1997), an excellent example of the type of controlled study of incentives that is needed to address these issues.

75. The term ''valuation'' subsumes open-ended elicitation procedures as well as dichotomous choice, binary referenda, and stated choice tasks. See Harrison (2006a, 2006b) and Harrison and Rutström (2008) for reviews.

76. Heckman and Smith (1995; pp. 99–101) provide many examples, and coin the expression ''randomization bias'' for this possible effect. Harrison and List (2004) review the differences between laboratory, field, social, and natural experiments in economics, and all could be potentially affected by randomization bias. Palfrey and Pevnitskaya (2008) use thought experiments and laboratory experiments to illustrate how risk attitudes can theoretically affect the mix of bidders in sealed-bid auctions with endogenous entry, and thereby change behavior in the sample of bidders observed in the auction.

77. We hesitate to endorse practices in other fields, in which recruitment fees are not paid to subjects, since they open themselves up to abuse. We have considerable experience of faculty recruiting subjects for ''extra credit,'' but where the task and behavior bears no relationship at all to the learning objectives of the class, and no pedagogic feedback is provided to students even if it does bear some tangential relationship to the topic of the class. We have serious ethical problems with such practices, quite apart from the problems of motivation that they raise.

78. There is also evidence of differences in the demographics and behavior of volunteers and ''pseudo-volunteers,'' which are subjects formally recruited in a classroom to participate in an experiment during class time (Rutström, 1998; Eckel & Grossman, 2000). The disadvantage with pseudo-volunteers is that the subjects might simply not be interested in participating in the experiment, even with the use of salient rewards. The advantage, of course, is that the selection process that leads them to be in the classroom is unrelated to the characteristics of the experimental task, although even here one might just be replacing one ill-studied sample selection process with another. After all, even if we model the factors that cause subjects from a university population to select into an experiment, we have not modeled the factors that cause individuals to choose to become university students (Casari, Ham, & Kagel, 2007).

79. Endogenous subject attrition from the experiment can also be informative about subject preferences, since the subject's exit from the experiment indicates that the subject had made a negative evaluation of it. See Diggle and Kenward (1994) and Philipson and Hedges (1998) for discussion of this statistical issue.

80. More precisely, the statistical problem is that there may be some unobserved individual effects that cause subjects to be in the observed sample or not, and these effects could be correlated with responses once in the observed sample. For example, Camerer and Lovallo (1999) find that excess entry into competitive games occurs more often when subjects volunteered to participate knowing that payoffs would depend on skill in a sports or current events trivia. This treatment could encourage less risk-averse subjects to participate in the experiment and may explain the observed reference bias effect, or part of it.

81. It is well known in the field of clinical drug trials that persuading patients to participate in randomized studies is much harder than persuading them to participate in non-randomized studies (e.g., Kramer and Shapiro (1984; p. 2742ff.)). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralized bureaucracies to administer the random treatment (e.g., Hotz, 1992). For example, Heckman and Robb (1985) note that the refusal rate in one randomized job training program was over 90%.

82. Here we consider the role of preferences over risk, but the same concerns apply to the elicitation of other types of preferences, such as social preferences or time preferences (Eckel & Grossman, 2000; Lazear, Malmendier, & Weber, 2006; Dohmen & Falk, 2006). These concerns arise when subjects have some reason to believe that the task will lead them to evaluate those preferences, such as in longitudinal designs allowing attrition, or social experiments requiring disclosure of the nature of the task prior to participation. They might also arise if the sample is selected by some endogenous process in which selection might be correlated with those preferences, such as group membership or location choices.

83. In addition, we often just assign subjects to some role in an experiment, whether or not they would have selected for this role in any naturally occurring environment. This issue lies at the heart of the interest in field experiments initiated by Bohm (2002).

84. Lusk and Coble (2005) also report evidence consistent with this conclusion, comparing risk preferences elicited for an artificial monetary instrument and comparable preferences for an instrument defined over genetically-modified food. Lusk and Coble (2008) find that adding abstract background risk to an elicitation procedure using artificial monetary outcomes also generates more risk aversion, although they do not find the effect to be large quantitatively.

85. To make this point more succinctly, consider the elicitation of the value that a person places on safety, a critical input in the cost-benefit assessment of environmental policy such as the Clean Air Act (United States Environmental Protection Agency, 1997). Conventional procedures to measure such preferences focus on monetary values to avoid mortality risk, by asking subjects to value scenarios in which they face different risks of death. The traditional interpretation of responses to such questions ignores the fact that it is hard to imagine a physical risk that could kill you with some probability but that would leave you alive and have no effect whatsoever on your health. Of course, such risks exist, but most of the environmental

risks of concern for policy do not fall into such a category. In general, then, responses to the foreground risk question should allow for the fact that the subject likely perceived some background risk. This example represents an important policy issue and highlights the import of the theoretical literature on background risk.

86. However, since we do not know the subject probability distribution of background risk in the field, we cannot know if background risk is statistically independent with the foreground risk. We can think of no reason why the two might be correlated, but this illustrates again the type of trade-off one experiences with field experiments. It also points to the complementary nature of field and lab experiments: Lusk and Coble (2008) show that independent background risk in a lab setting is associated with an increase in foreground risk aversion.

87. The typical application of the random lottery incentive mechanism in experiments such as these would have one choice selected at random. We used three to ensure comparability of rewards with other experiments in which subjects made choices over 40 or 20 lotteries, and where 2 lotteries or 1 lottery was, respectively, selected at random to be played out.

88. The computer laboratory used for these experiments has 28 subject stations. Each screen is "sunken" into the desk, and subjects were typically separated by several empty stations due to staggered recruitment procedures. No subject could see what the other subjects were doing, let alone mimic what they were doing since each subject was started individually at different times.

89. These final outcomes differ by $1 from the two highest outcomes for the gain frame and mixed frame, because we did not want to offer prizes in fractions of dollars.

90. To ensure that probabilities summed to one, we also used probabilities of 0.26 instead of 0.25, 0.38 instead of 0.37, 0.49 instead of 0.50, or 0.74 instead of 0.75.

91. The control data in these three panels, for the 1× problem, are pooled across all task #1 responses. That is, the task #1 responses in the bottom left panel of Fig. 27 are not just the task #1 responses of the individuals facing the 90× problem. Nothing essential hinges on this at this stage of exposition. The statistical analysis in Section 2.1 does take this into account, using appropriate panel estimation procedures.

92. The experience was not with the same prize level, as noted earlier.

93. See Ortona (1994) and Kachelmeier and Shehata (1994).

94. These conclusions come from a panel regression model that controls for all of the factors discussed, and that allows for individual-level heteroskedasticity and individual-level first-order autocorrelation. All conclusions refer to effects that are statistically significant at the 1% level.

95. References to increases in risk aversion should also be understood, in this context, to refer to decreases in risk loving.

96. Although purely anecdotal, our own experience is that many subjects faced with the BDM task believe that the buying price depends in some way on their selling price. To mitigate such possible perceptions we have tended to use physical randomizing devices that are less prone to being questioned.

97. The stakes in the experiments of Gneezy and Potters (1997) were actually 2 Dutch guilders, which converted at the time of the experiment to roughly $1.20. Haigh and List (2005) used a stake of $1.00 for their students, to be comparable to the earlier stake. They quadrupled the stakes to $4.00 for the traders, on the grounds

that it would be more salient for them. Of course, this change in monetary stake size adds a potential confound to the comparability of results across students and traders, but one that has no obvious resolution without an elaborate investigation into the purchasing power of a dollar to students and traders.

98. Gneezy and Potters generously provided their individual data, and we used the same statistical model as Haigh and List (2005; Table II, specification 2, p. 530) on their data. Haigh and List also generously provided their individual data, and we replicated their statistical conclusions.

99. In fact, subjects tended to pick in round percentages. In the Low frequency treatment 76% of the choices were for 0, 25, 50, or 100% bets, and in the High frequency treatment 81% of the choices were for the 25, 50, or 100% bets.

100. For example, Kahneman and Lovallo (1993, p. 20), Benartzi and Thaler (1995, p. 79), Gneezy and Potters (1997, p. 632), Thaler et al. (1997. p. 650), Gneezy et al. (2003, p. 822), and Haigh and List (2005, p. 525).

101. In other words, there are settings in which a CRRA or even RN utility function might be appropriate for some theoretical, econometric, or policy exercise. But this experiment is not obviously one of those settings.

102. Yet another approach would be to modify the experimental design and allow subjects to leverage their bets beyond 100% of their stake. There are some logistical problems running such experiments in a laboratory setting, although of course stock exchanges and futures markets allow such trades.

103. The *a* parameter may be viewed as a counterpart in this specification of the noise parameter used by Holt and Laury (2002).

104. Benartzi and Thaler (1995, p. 80) are clear that this *evaluation* horizon is not the same thing as a planning horizon: "A young investor, for example, might be saving for retirement 30 years off in the future, but nevertheless experience the utility associated with the gains and losses of his investment every quarter when he opens a letter from his mutual fund. In this case, his (planning) horizon is 30 years but his evaluation period (evaluation horizon) is 3 months."

105. They prefer the expression "prospective utility," but there is no confusion as long we are clear about which utility functions and probabilities are being used to calculate expected utility.

106. Mankiw and Zeldes (1991) make the important observation that only 12% of Americans hold stocks worth more than $10,000, using a 1984 survey, so one really has to explain *their* indifference between holding bonds and stocks. Presumably, the remaining "corner-solution" individuals face some transactions costs to undertaking such investments. It would be an easy and important extension of the approach of BT to allow for such heterogeneity in the composition of stockholders and others.

107. The constant term in this linear function is suppressed, since it would be perfectly correlated with the sum of these two binary variables. To be explicit, denote these dummy variables for the treatments as L and H, respectively. Then we actually estimate α_L, α_L, β_L, β_H, λ_L, and λ_H, where $\alpha = \alpha_L \times L + \alpha_H \times H$, $\beta = \beta_L \times L + \beta_H \times H$, and $\lambda = \lambda_L \times L + \lambda_H \times H$. Thus, the logic of the likelihood function is as follows: candidate values of these six parameters are proposed, the linear function evaluated so that we know candidate value of α, β, and λ for each of the Low and High frequency treatments, the expected utility of the actual choice is evaluated using the Tversky and

Kahneman (1992) specification, and then the log-likelihood function specified above is evaluated.

108. The Arrow–Pratt coefficient of RRA is $1-\alpha$, so $\alpha = 1$ implies risk neutrality, $\alpha < 1$ implies risk aversion, and $\alpha > 1$ implies risk-loving behavior. These benchmarks are worth noting, to avoid confusion, given the popularity of specifications from Holt and Laury (2002) that estimate $1-\alpha$ directly (the risk-neutral value is 0 in that case, positive estimates indicate risk aversion, and negative estimates indicate risk loving).

109. The exposition is deliberately transparent to economists. Most of the exposition in Section F1 would be redundant for those familiar with Gould, Pitblado, and Sribney (2006) or even Rabe-Hesketh and Everitt (2004; ch.13). It is easy to find expositions of ML in *Stata* that are more general and elegant for their purpose, but for those trying to learn the basic tools for the first time that elegance can just appear to be needlessly cryptic coding, and actually act as an impediment to comprehension. There are good reasons that one wants to build more flexible and computationally efficient models, but ease of comprehension is rarely one of them. StataCorp (2007) documents the latest version 10 of *Stata*, but the exposition of the ML syntax is minimal in that otherwise extensive documentation.

110. Paarsch and Hong (2006; Appendix A.8) provide a comparable introduction to the use of MATLAB for estimation of structural models of auctions. Unfortunately their documentation contains no “real data” to evaluate the programs on.

111. Note that this is `euL’ and not ‘euL’: beginning *Stata* users make this mistake a lot.

112. Since the ML_eut0 program is called many, many times to evaluate Jacobians and the like, these warning messages can clutter the screen display needlessly. During debugging, however, one normally likes to have things displayed, so the command “quietly” would be changed to “noisily” for debugging. Actually, we use the “ml check” option for debugging, as explained later, and never change this to “noisily.” Or we can display one line by using the “noisily” option, to debug specific calculations.

ACKNOWLEDGMENT

We thank the U.S. National Science Foundation for research support under grants NSF/IIS 9817518, NSF/HSD 0527675, and NSF/SES 0616746, and the Danish Social Science Research Council for research support under project #24-02-0124. We are grateful to William Harbaugh, John Hey, Steven Kachelmeier, and Robert Sugden for making detailed experimental results available. Valuable comments were received from Steffen Andersen, James Cox, Morten Lau, Vjollca Sadiraj, Peter Wakker, and Nathaniel Wilcox. Harrison is also affiliated with the Durham Business School, Durham University, UK.

REFERENCES

Abdellaoui, M. (2000). Parameter-free elicitation of utilities and probability weighting functions. *Management Science*, *46*, 1497–1512.

Abdellaoui, M., Barrios, C., & Wakker, P. P. (2007a). Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory. *Journal of Econometrics*, *138*, 356–378.

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007b). Measuring loss aversion under prospect theory: A parameter-free approach. *Management Science*, *53*(10), 1659–1674.

Allais, M. (1979). The foundations of positive theory of choice involving risk and a criticism of the postulates and Axioms of the American school. In: M. Allais & O. Hagen (Eds), *Expected utility hypotheses and the Allais paradox*. Dordrecht, The Netherlands: Reidel.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2006a). Elicitation using multiple price lists. *Experimental Economics*, *9*(4), 383–405.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008a). Eliciting risk and time preferences. *Econometrica*, *76*, forthcoming.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008b). Lost in state space: Are preferences stable? *International Economic Review*, *49*, forthcoming.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008c). Risk aversion in game shows. In: J. C. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Andersen, S., Harrison, G. W., & Rutström, E. E. (2006b). *Choice behavior, asset integration and natural reference points*. Working Paper 06-04. Department of Economics, College of Business Administration, University of Central Florida.

Attema, A. E., Bleichrodt, H., Rohde, K. I. M., & Wakker, P. P. (2006). *Time-tradeoff sequences for quantifying and visualizing the degree of time inconsistency, using only pencil and paper*. Working Paper. Erasmus University, Rotterdam.

Ballinger, T. P., & Wilcox, N. T. (1997). Decisions, error and heterogeneity. *Economic Journal*, *107*, 1090–1105.

Barr, A. (2003). *Risk pooling, commitment, and information: An experimental test of two fundamental assumptions*. Working Paper 2003–05. Centre for the Study of African Economies, Department of Economics, University of Oxford.

Barr, A., & Packard, T. (2002). *Revealed preference and self insurance: Can we learn from the self employed in Chile*? Policy Research Working Paper #2754. World Bank, Washington DC.

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

Beattie, J., & Loomes, G. (1997). The impact of incentives upon risky choice experiments. *Journal of Risk and Uncertainty*, *14*, 149–162.

Beck, J. H. (1994). An experimental test of preferences for the distribution of income and individual risk aversion. *Eastern Economic Journal*, *20*(2), 131–145.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*, 226–232.

Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *Quarterly Journal of Economics*, *111*(1), 75–92.

Berg, J., Dickhaut, J., & McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences*, *102*, 4209–4214.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*, 395–407.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *Economic Journal*, *91*, 867–890.

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

Bleichrodt, H., & Pinto, J. L. (2000). A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management Science*, *46*, 1485–1496.

Bohm, P. (2002). Pitfalls in experimental economics. In: F. Andersson & H. Holm (Eds), *Experimental economics: Financial markets, auctions, and decision making*. Dordrecht: Kluwer.

Botelho, A., Harrison, G. W., Pinto, L. M. C., & Rutström, E. E. (2005a). *Social norms and social choice*. Working Paper 05-23. Department of Economics, College of Business Administration, University of Central Florida.

Botelho, A., Harrison, G. W., Pinto, L. M. C., & Rutström, E. E. (2005b). *Testing static game theory with dynamic experiments: A case study of public goods*. Working Paper 05–25. Department of Economics, College of Business Administration, University of Central Florida.

Bullock, D. S., & Rutström, E. E. (2007). Policy making and rent-dissipation: An experimental test. *Experimental Economics*, *10*(1), 21–36.

Cadsby, C. B., Song, F., & Tapon, F. (2007). Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal*, *50*(2), 387–405.

Calman, K. C., & Royston, G. (1997). Risk language and dialects. *British Medical Journal*, *315*, 939–942.

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

Camerer, C. F. (2000). Prospect theory in the wild: Evidence from the field. In: D. Kahneman & A. Tversky (Eds), *Choices, values and frames*. New York: Cambridge University Press.

Camerer, C. F. (2005). Three cheers – psychological, theoretical, empirical – for loss aversion. *Journal of Marketing Research*, *XLII*, 129–133.

Camerer, C., & Ho, T. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, *8*, 167–196.

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

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, *89*(1), 306–318.

Casari, M., Ham, J. C., & Kagel, J. H. (2007). Selection bias, demographic effects and ability effects in common value experiments. *American Economic Review*, *97*(4), 1278–1304.

Chambers, R. G., & Quiggin, J. (2000). *Uncertainty, production, choice, and agency: The state-contingent approach*. New York, NY: Cambridge University Press.

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.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, *47*(1), 72–93.

Cleveland, W. S., Harris, C. S., & McGill, R. (1982). Judgements of circle sizes on statistical maps. *Journal of the American Statistical Association*, *77*(379), 541–547.

Cleveland, W. S., Harris, C. S., & McGill, R. (1983). Experiments on quantitative judgements of graphs and maps. *Bell System Technical Journal*, *62*(6), 1659–1674.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*, 107–127.

Conlisk, J. (1989). Three variants on the Allais example. *American Economic Review*, *79*(3), 392–407.

Conte, A., Hey, J. D., & Moffatt, P. G. (2007). *Mixture models of choice under risk*. Discussion Paper No. 2007/06. Department of Economics and Related Studies, University of York.

Cox, J. C., & Epstein, S. (1989). Preference reversals without the independence axiom. *American Economic Review*, *79*(3), 408–426.

Cox, J. C., & Harrison, G. W. (2008). Risk aversion in experiments: An introduction. In: J. C. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Cox, J. C., Roberson, B., & Smith, V. L. (1982). Theory and behavior of single object auctions. In: V. L. Smith (Ed.), *Research in experimental economics* (Vol. 2). Greenwich: JAI Press.

Cox, J. C., & Sadiraj, V. (2006). Small- and large-stakes risk aversion: Implications of concavity calibration for decision theory. *Games & Economic Behavior*, *56*, 45–60.

Cox, J. C., & Sadiraj, V. (2008). Risky decisions in the large and in the small: Theory and experiment. In: J. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Cox, J. C., Smith, V. L., & Walker, J. M. (1985). Experimental development of sealed-bid auction theory: Calibrating controls for risk aversion. *American Economic Review (Papers & Proceedings)*, *75*, 160–165.

Cox, J. C., Smith, V. L., & Walker, J. M. (1988). Theory and individual behavior of first-price auctions. *Journal of Risk and Uncertainty*, *1*, 61–99.

Cubitt, R. P., Starmer, C., & Sugden, R. (1988a). Dynamic choice and the common ratio effect: An experimental investigation. *Economic Journal*, *108*, 1362–1380.

Cubitt, R. P., Starmer, C., & Sugden, R. (1988b). On the validity of the random lottery incentive system. *Experimental Economics*, *1*(2), 115–131.

Dave, C., Eckel, C., Johnson, C., & Rojas, C. (2007). *On the heterogeneity, stability and validity of risk preference measures*. Unpublished manuscript. Department of Economics, University of Texas at Dallas.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, *43*(1), 49–93.

Dohmen, T., & Falk, A. (2006). *Performance pay and multi-dimensional sorting: Productivity, preferences and gender*. Discussion Paper #2001. Institute for the Study of Labor (IZA), Bonn, Germany.

Eckel, C. C., & Grossman, P. J. (2000). Volunteers and pseudo-volunteers: The effect of recruitment method in dictator experiments. *Experimental Economics*, *3*, 07–120.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study of actual and forecast risk attitudes of women and men. *Journal of Economic Behavior & Organization*, forthcoming.

Eeckhoudt, L., Gollier, C., & Schlesinger, H. (1996). Changes in background risk and risk-taking behavior. *Econometrica*, *64*(3), 683–689.

Ehrhart, K.-M., & Keser, C. (1999). *Mobility and cooperation: On the run*. Working Paper 99s-24. CIRANO, University of Montreal.

Engle-Warnick, J., Escobal, J., & Laszlo, S. (2006). *The effect of an additional alternative on measured risk preferences in a laboratory experiment in Peru*. Working Paper 2006s-06. CIRANO, Montreal.

Ertan, A., Page, T., & Putterman, L. (2005). *Can endogenously chosen institutions mitigate the free-rider problem and reduce perverse punishment*? Working Paper 2005-13. Department of Economics, Brown University.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030.

Farquhar, P. H. (1984). Utility assessment methods. *Management Science*, *30*(11), 1283–1300.

Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, *97*(1), 298–317.

Fennema, H., & van Assen, M. (1999). Measuring the utility of losses by means of the trade off method. *Journal of Risk and Uncertainty*, *17*(3), 277–295.

Finney, M. A. (1998). *FARSITE: Fire Area Simulator – Model Development and Evaluation*, Research Paper RMRS-RP-4. Rocky Mountain Research Station, Forest Service, United States Department of Agriculture.

Fiore, S. M., Harrison, G. W., Hughes, C. E., & Rutström, E. E. (2007). *Virtual experiments and environmental policy*. Working Paper 07-01. Department of Economics, College of Business Administration, University of Central Florida.

Fishburn, P. C. (1967). Methods of estimating additive utilities. *Management Science*, *13*(7), 435–453.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *XL*, 351–401.

Gegax, D., Gerking, S., & Schulze, W. (1991). Perceived risk and the marginal value of safety. *Review of Economics and Statistics*, *73*, 589–596.

Gerking, S., de Haan, M., & Schulze, W. (1988). The marginal value of job safety: A contingent value study. *Journal of Risk and Uncertainty*, *1*, 185–199.

Gneezy, U., Kapteyn, A., & Potters, J. (2003). Evaluation periods and asset prices in a market experiment. *Journal of Finance*, *58*, 821–838.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, *112*, 631–645.

Gollier, C. (2001). *The economics of risk and time*. Cambridge, MA: MIT Press.

Gollier, C., & Pratt, J. W. (1996). Risk vulnerability and the tempering effect of background risk. *Econometrica*, *64*(5), 1109–1123.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*, 129–166.

Gould, W., Pitblado, J., & Sribney, W. (2006). *Maximum likelihood estimation with Stata* (3rd ed.). College Station, TX: Stata Press.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, *69*, 623–648.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*, 108–111.

Haigh, M. S., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis. *Journal of Finance*, *60*(1), 523–534.

Harbaugh, W. T., Krause, K., & Vesterlund, L. (2002). Risk attitudes of children and adults: Choices over small and large probability gains and losses. *Experimental Economics*, *5*, 53–84.

Harless, D. W. (1992). Predictions about indifference curves inside the unit triangle: A test of variants of expected utility theory. *Journal of Economic Behavior and Organization*, *18*, 391–414.

Harless, D. W., & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*(6), 1251–1289.

Harrison, G. W. (1986). An experimental test for risk aversion. *Economics Letters*, *21*(1), 7–11.

Harrison, G. W. (1989). Theory and misbehavior of first-price auctions. *American Economic Review*, *79*, 749–762.

Harrison, G. W. (1990). Risk attitudes in first-price auction experiments: A Bayesian analysis. *Review of Economics and Statistics*, *72*, 541–546.

Harrison, G. W. (1992). Theory and misbehavior of first-price auctions: Reply. *American Economic Review*, *82*, 1426–1443.

Harrison, G. W. (2006a). Experimental evidence on alternative environmental valuation methods. *Environmental and Resource Economics*, *34*, 125–162.

Harrison, G. W. (2006b). Making choice studies incentive compatible. In: B. Kanninen (Ed.), *Valuing environmental amenities using stated choice studies: A common sense guide to theory and practice* (pp. 65–108). Boston: Kluwer.

Harrison, G. W. (2006c). *Maximum likelihood estimation of utility functions using Stata*. Working Paper 06-12. Department of Economics, College of Business Administration, University of Central Florida.

Harrison, G. W. (2007). Hypothetical bias over uncertain outcomes. In: J. A. List (Ed.), *Using experimental methods in environmental and resource economics*. Northampton, MA: Elgar.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2005a). Temporal stability of estimates of risk aversion. *Applied Financial Economics Letters*, *1*, 31–35.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2005b). Risk aversion and incentive effects: Comment. *American Economic Review*, *95*(3), 897–901.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2007a). Measurement with experimental controls. In: M. Boumans (Ed.), *Measurement in economics: A handbook*. San Diego, CA: Elsevier.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2005c). *Risk attitudes, randomization to treatment, and self-selection into experiments*. Working Paper 05-01. Department of Economics, College of Business Administration, University of Central Florida; Journal of Economic Behaviour & Organization, forthcoming.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2007b). Estimating risk attitudes in Denmark: A field experiment. *Scandinavian Journal of Economics*, *109*(2), 341–368.

Harrison, G. W., Lau, M. I., Rutström, E. E., & Sullivan, M. B. (2005d). Eliciting risk and time preferences using field experiments: Some methodological issues. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field experiments in economics* (Vol. 10). Greenwich, CT: JAI Press, Research in Experimental Economics.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1013–1059.

Harrison, G. W., List, J. A., & Towe, C. (2007c). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. *Econometrica*, *75*(2), 433–458.

Harrison, G. W., List, J. A., & Tra, C. (2005e). *Statistical characterization of heterogeneity in experiments*. Working Paper 05-10. Department of Economics, College of Business Administration, University of Central Florida.

Harrison, G. W., & Rutström, E. E. (2005). *Expected utility theory and prospect theory: One wedding and a decent funeral*. Working Paper 05-18. Department of Economics, College of Business Administration, University of Central Florida; Experimental Economics, forthcoming.

Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In: C. R. Plott, V. L. Smith (Eds), *Handbook of experimental economics results*. North-Holland: Amsterdam, forthcoming.

Harstad, R. M. (2000). Dominant strategy adoption and bidders' experience with pricing rules. *Experimental Economics*, *3*(3), 261–280.

Heckman, J. J., Robb, R., Heckman, J., & Singer, B. (Eds). (1985). *Longitudinal analysis of labor market data*. New York: Cambridge University Press.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*(2), 85–110.

Heinemann, F. (2008). Measuring risk aversion and the wealth effect. In: J. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Greenwich, CT: JAI Press, Research in Experimental Economics.

Hershey, J. C., Kunreuther, H. C., & Schoemaker, P. J. H. (1982). Sources of bias in assessment procedures for utility functions. *Management Science*, *28*(8), 936–954.

Hershey, J. C., & Schoemaker, P. J. H. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, *31*(10), 1213–1231.

Hey, J. D. (1995). Experimental investigations of errors in decision making under risk. *European Economic Review*, *39*, 633–640.

Hey, J. D. (2001). Does repetition improve consistency?. *Experimental Economics*, *4*, 5–54.

Hey, J. D. (2002). Experimental economics and the theory of decision making under uncertainty. *Geneva Papers on Risk and Insurance Theory*, *27*(1), 5–21.

Hey, J. D., & Lee, J. (2005a). Do subjects remember the past?. *Applied Economics*, *37*, 9–18.

Hey, J. D., & Lee, J. (2005b). Do subjects separate (or are they sophisticated)?. *Experimental Economics*, *8*, 233–265.

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*(6), 1291–1326.

Hirshleifer, J., & Riley, J. G. (1992). *The analytics of uncertainty and information*. New York, NY: Cambridge University Press.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*(3), 902–912.

Horowitz, J. K. (1992). A test of intertemporal consistency. *Journal of Economic Behavior and Organization*, *17*, 171–182.

Hotz, V. J. (1992). Designing an evaluation of JTPA. In: C. Manski & I. Garfinkel (Eds), *Evaluating welfare and training programs*. Cambridge: Harvard University Press.

Isaac, R. M., & James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty*, *20*(2), 177–187.

James, D. (2007). Stability of risk preference parameter estimates within the Becker–DeGroot–Marschak procedure. *Experimental Economics*, *10*, 123–141.

Kachelmeier, S. J., & Shehata, M. (1992). Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *American Economic Review*, *82*(5), 1120–1141.

Kachelmeier, S. J., & Shehata, M. (1994). Examining risk preferences under high monetary incentives: Reply. *American Economic Review*, *84*(4), 1104.

Kagel, J. H. (1995). Auctions: A survey of experimental research. In: J. H. Kagel & A. E. Roth (Eds), *The handbook of experimental economics*. Princeton: Princeton University Press.

Kagel, J. H., & Levin, D. (2002). *Common value auctions and the winner's curse*. Princeton: Princeton University Press.

Kagel, J. H., MacDonald, D. N., & Battalio, R. C. (1990). Tests of 'Fanning Out' of indifference curves: Results from animal and human experiments. *American Economic Review*, *80*(4), 912–921.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17–31.

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

Keller, L. R., & Strazzera, E. (2002). Examining predictive accuracy among discounting models. *Journal of Risk and Uncertainty*, *24*(2), 143–160.

Kent, S. (1964). Words of estimated probability. *Studies in Intelligence*, *8*, 49–65.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, *73*(6), 1849–1892.

Köbberling, V., & Wakker, P. P. (2005). An index of loss aversion. *Journal of Economic Theory*, *122*, 119–131.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, *121*(4), 1133–1165.

Kőszegi, B., & Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, *97*(4), 1047–1073.

Kramer, M., & Shapiro, S. (1984). Scientific challenges in the application of randomized trials. *Journal of the American Medical Association*, *252*, 2739–2745.

Kreps, D. M., & Porteus, E. L. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, *46*(1), 185–200.

Krupnick, A., Alberini, A., Cropper, M., Simon, N., O'Brien, B., Goeree, R., & Heintzelman, M. (2002). Age, health and the willingness to pay for mortality risk reductions: A contingent valuation survey of Ontario residents. *Journal of Risk and Uncertainty*, *24*(2), 161–186.

Kuilen, G., Wakker, P. P., & Zou, L. (2007). *A midpoint technique for easily measuring prospect theory's probability weighting*. Working Paper. Econometric Institute, Erasmus University, Rotterdam, The Netherlands.

Laury, S. K., & Holt, C. A. (2008). Further reflections on prospect theory. In: J. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Lazear, E. P., Malmendier, U., & Weber, R. A. (2006). *Sorting in experiments with application to social preferences*. Working Paper #12041. National Bureau of Economic Research.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

List, J. A. (2003). Does market experience eliminate market anomalies?. *Quarterly Journal of Economics*, *118*, 41–71.

Loomes, G. (1988). Different experimental procedures for obtaining valuations of risky actions: Implications for utility theory. *Theory and Decision*, *25*, 1–23.

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, *24*(2), 103–130.

Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica*, *59*(2), 425–439.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648.

Loomes, G., & Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, *65*, 581–598.

Lopes, L. L. (1984). Risk and distributional inequality. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(4), 465–484.

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

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

Lusk, J. L., & Coble, K. H. (2005). Risk perceptions, risk preference, and acceptance of risky food. *American Journal of Agricultural Economics*, *87*(2), 393–404.

Lusk, J. L., & Coble, K. H. (2008). Risk aversion in the presence of background risk: Evidence from the lab. In: J. C. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Mankiw, N. G., & Zeldes, S. P. (1991). The consumption of stockholders and non-stockholders. *Journal of Financial Economics*, *29*(1), 97–112.

McFadden, D. (2001). Economic choices. *American Economic Review*, *91*(3), 351–378.

McKee, M. (1989). Intra-experimental income effects and risk aversion. *Economics Letters*, *30*, 109–115.

Miller, L., Meyer, D. E., & Lanzetta, J. T. (1969). Choice among equal expected value alternatives: Sequential effects of winning probability level on risk preferences. *Journal of Experimental Psychology*, *79*(3), 419–423.

Millner, E. L., Pratt, M. D., & Reilly, R. J. (1988). A reexamination of Harrison's experimental test for risk aversion. *Economics Letters*, *27*, 317–319.

Murnighan, J. K., Roth, A. E., & Shoumaker, F. (1987). Risk aversion and bargaining: Some preliminary results. *European Economic Review*, *31*, 265–271.

Murnighan, J. K., Roth, A. E., & Shoumaker, F. (1988). Risk aversion in bargaining: An experimental study. *Journal of Risk and Uncertainty*, *1*(1), 101–124.

Nau, R. F. (2006). Uncertainty aversion with second-order utilities and probabilities. *Management Science*, *52*(1), 136–145.

Novemsky, N., & Kahneman, D. (2005a). The boundaries of loss aversion. *Journal of Marketing Research*, *XLII*, 119–128.

Novemsky, N., & Kahneman, D. (2005b). How do intentions affect loss aversion? *Journal of Marketing Research*, *XLII*, 139–140.

Ochs, J., & Roth, A. E. (1989). An experimental study of sequential bargaining. *American Economic Review*, *79*(3), 355–384.

Ortona, G. (1994). Examining risk preferences under high monetary incentives: Comment. *American Economic Review*, *84*(4), 1104.

Paarsch, H. J., & Hong, H. (2006). *An introduction to the structural econometrics of auction data.* Cambridge, MA: MIT Press.

Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: Reciprocity, mimicry, and efficiency. *Economic Journal*, *115*, 1037–1058.

Palfrey, T. R., & Pevnitskaya, S. (2008). Endogenous entry and self-selection in private value auctions: An experimental study. *Journal of Economic Behavior & Organization*, 66, forthcoming.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics*, *11*, 619–632.

Philipson, T., & Hedges, L. V. (1998). Subject evaluation in social experiments. *Econometrica*, *66*(2), 381–408.

Plott, C. R., & Zeiler, K. (2005). The willingness to pay-willingness to accept gap, the 'Endowment Effect,' subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review*, *95*(3), 530–545.

Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*, 497–527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

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

Rabe-Hesketh, S., & Everitt, B. (2004). *A handbook of statistical analyses using Stata* (3rd ed.). New York: Chapman & Hall/CRC.

Rabin, M. (2000). Risk aversion and expected utility theory: A calibration theorem. *Econometrica*, *68*, 1281–1292.

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

Rieger, M. O., & Wang, M. (2006). Cumulative prospect theory and the St. Petersburg paradox. *Economic Theory*, *28*, 665–679.

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13*, 19–23.

Roth, A. E., & Malouf, M. W. K. (1979). Game-theoretic models and the role of information in bargaining. *Psychological Review*, *86*, 574–594.

Rutström, E. E. (1998). Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory*, *27*(3), 427–441.

Saha, A. (1993). Expo-power utility: A flexible form for absolute and relative risk aversion. *American Journal of Agricultural Economics*, *75*(4), 905–913.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Schmidt, U., Starmer, C., & Sugden, R. (2005). *Explaining preference reversal with third-generation prospect theory*. Working Paper. School of Economic and Social Science, University of East Anglia.

Schubert, R., Brown, M., Gysler, M., & Brachinger, H. W. (1999). Financial decision-making: Are women really more risk-averse? *American Economic Review (Papers & Proceedings)*, *89*(2), 381–385.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, *72*(5), 923–955.

Smith, V. L. (2003). Constructivist and ecological rationality in economics. *American Economic Review*, *93*(3), 465–508.

Starmer, C., & Sugden, R. (1989). Violations of the independence axiom in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operational Research*, *19*, 79–102.

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

StataCorp. (2007). *Stata statistical software: Release 10*. College Station, TX: Stata Corporation.

Stigler, G. J., & Becker, G. S. (1977). De gustibus non est disputandum. *American Economic Review*, *67*(2), 76–90.

Sutter, M., Haigner, S., & Kocher, M. (2006). *Choosing the stick or the carrot? Endogenous institutional choice in social dilemma situations*. Discussion Paper No. 5497. Centre for Economic Policy Research, London.

Tanaka, T., Camerer, C. F., & Nguyen, Q. (2007). *Risk and time preferences: Experimental and household survey data from Vietnam*. Working Paper. California Institute of Technology.

Thaler, R. H., Tversky, A., Kahneman, D., & Schwartz, A. (1997). The effect of myopia and loss aversion on risk taking: An experimental test. *Quarterly Journal of Economics*, *112*, 647–661.

Train, K. E. (2003). *Discrete choice methods with simulation*. New York: Cambridge University Press.

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

United States Environmental Protection Agency. (1997). *The benefits and costs of the clean air act: 1970 to 1990*. Washington, DC: Office of Air and Radiation, US EPA.

Vickrey, W. S. (1961). Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, *16*, 8–37.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333.

Wakker, P. P. (1989). Transforming probabilities without violating stochastic dominance. In: E. Roskam (Ed.), *Mathematical Psychology in Progress*. Berlin: Springer.

Wakker, P. P., & Deneffe, D. (1996). Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*, 1131–1150.

Wakker, P. P., Erev, I., & Weber, E. U. (1994). Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty*, *9*, 195–230.

Wilcox, N. T. (2008a). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In: J. Cox & G. W. Harrison (Eds), *Risk aversion in experiments* (Vol. 12). Bingley, UK: Emerald, Research in Experimental Economics.

Wilcox, N. T. (2008b). 'Stochastically more risk averse:' A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *142*, forthcoming.

Williams, R. (2000). A note on robust variance estimation for cluster-correlated. *Biometrics*, *56*, 645–646.

Wooldridge, J. (2003). Cluster-sample methods in applied econometrics. *American Economic Review (Papers & Proceedings)*, *93*, 133–138.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, *55*(1), 95–115.

APPENDIX A. REPRESENTATION AND PERCEPTION OF PROBABILITIES

There are two representational issues with probabilities. The first is that subjects may base their decisions on concepts of *subjective* probabilities such that we should expect them to deviate in some ways from objective probabilities. The second is that *perceptions* of probabilities may not correspond to the actual probabilities. Only with a theory that explains both the perception of probabilities and the relationship between subjective and objective probabilities we would be able to identify both of these deviations. Nevertheless, careful experimental design can be helpful in generating some robustness in subjective and perceived probabilities, and a convergence in both of these on the underlying objective ones when that is normatively desirable. The review in this appendix complements the discussion in Section 1 of the paper by showing some alternative ways to represent the lotteries to subjects.

Camerer (1989) used a stacked box display to represent his lotteries to subjects. The length of the box provided information on the probabilities of each prize, and the width of the box provided information on the relative size of the prizes. The example in Fig. 20 was used in his written instructions to subjects, to explain how to read the lottery. Those instructions were as follows:

> The outcomes of the lotteries will be determined by a random number between 01 and 100. Each number between (and including) 01 and 100 is equally likely to occur. In the example above, the left lottery, labeled "A", pays nothing (0) if the random number is between 01 and 40. Lottery A pays five dollars ($5) if the random number is between 41 and 100. Notice that the picture is drawn so that the height of the line between 01 and 40 is 40% of the distance from 01 to 100. The rectangle around "$5" is 60% of the distance from 01 to 100.
>
> In the example above the lottery on the right, labeled "B", pays nothing (0) if the random number is between 01 and 50, five dollars ($5) if the random number is between 51 and 90, and ten dollars ($10) if the random number is between 91 and 100. As with lottery A, the heights of the lines in lottery B represent the fraction of the possible numbers which yield each payoff. For example, the height of the $10 rectangle is 10% of the way from 01 to 100.
>
> The widths of the rectangles are proportional to the size of their payoffs. In lottery B, for example, the $10 rectangle is twice as wide as the $5 rectangle.

This display is ingenious in the sense that it compactly displays the "numbers" as well as visual referents for the probabilities and relative

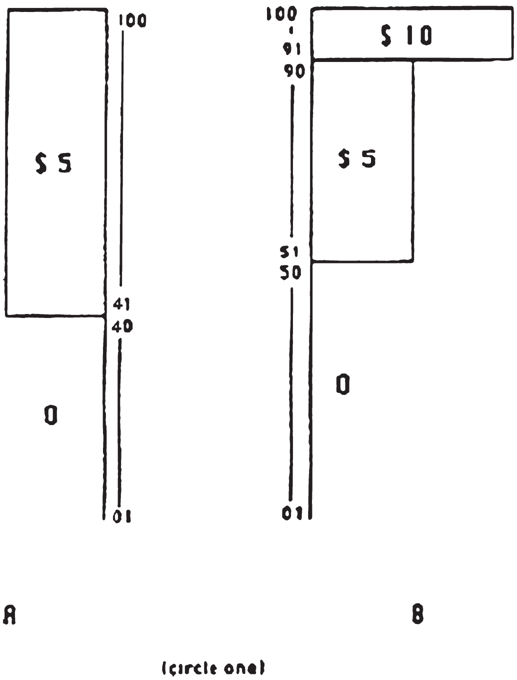


Fig. 20. Lottery Display Used by Camerer (1989).

prizes. The subject has to judge the probabilities for each prize from the visual referent, and is not directly provided that information numerically. There is a valuable literature on the ability of subjects to accurately assess quantitative magnitudes from visual referents of this kind, and it points to the need for individual-specific calibration in experts and non-experts (Cleveland, Harris, & McGill, 1982, 1983; Cleveland & McGill, 1984).

Battalio et al. (1990) and Kagel et al. (1990) employed purely numerical displays of their lotteries. For example, one such lottery was presented to subjects as follows:

A:	Winning $11 if 1–20	(20%)
	Winning $5 if 21–200	(80%)
B:	Winning $25 if 1–6	(6%)
	Winning $5 if 7–100	(94%)

Answer: (1) I prefer A. (2) I prefer B. (3) Indifferent.

This display presents all values numerically, with no visual referents. The numerical display shows the probability for each prize, rather than require the subject to infer that from the cumulative probabilities.

Beattie and Loomes (1997) used displays that were similar to those employed by Camerer (1989), although the probabilities were individually comparable since they were vertically aligned with a common base. Fig. 21 illustrates how they presented the lotteries to subjects. In addition, they provided text explaining how to read the display.

Wakker, Erev, and Weber (1994) considered four types of representations, shown in Fig. 22. One, on the far right, was a copy of the display employed by Camerer (1989), and the three on the left varied the extent to which information on outcomes was collapsed (top two panels on the left) and whether numerical information was provided in addition to the verbal information about probabilities (bottom panel on the left). The alternative representations were applied on a between-subjects basis, but no information is provided about the effect on behavior.

An example of the representation of probability using a verbal analogical scale is provided by Calman and Royston (1997; Table 4), using a distance analogue. For risks of 1 in 1, 1 in 10, 1 in 100, 1 in 1000, for example, the distance containing one "risk stick" 1 foot in length is 1 foot, 10 feet, 100 feet, and 1,000 feet, respectively. An older tradition seeks to "calibrate" words that are found in the natural English language with precise probability ranges. This idea stems from a concern that Kent (1964) had with the ambiguity in the use of colloquial expressions of uncertainty by intelligence operatives. He proposed that certain words be assigned specific numerical probability ranges. A study reported by von Winterfeldt and Edwards (1986, p. 98ff.) used these expressions and asked a number of NATO officials to state the probabilities that they would attach to the use of those words in sentences. The dots in Fig. 23 show the elicited probability judgements, and the shaded bars show the ranges suggested by Kent (1964). The fact that there

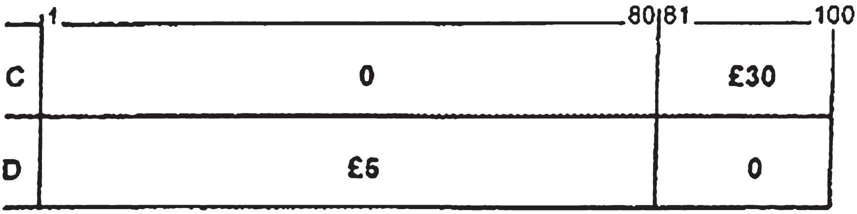
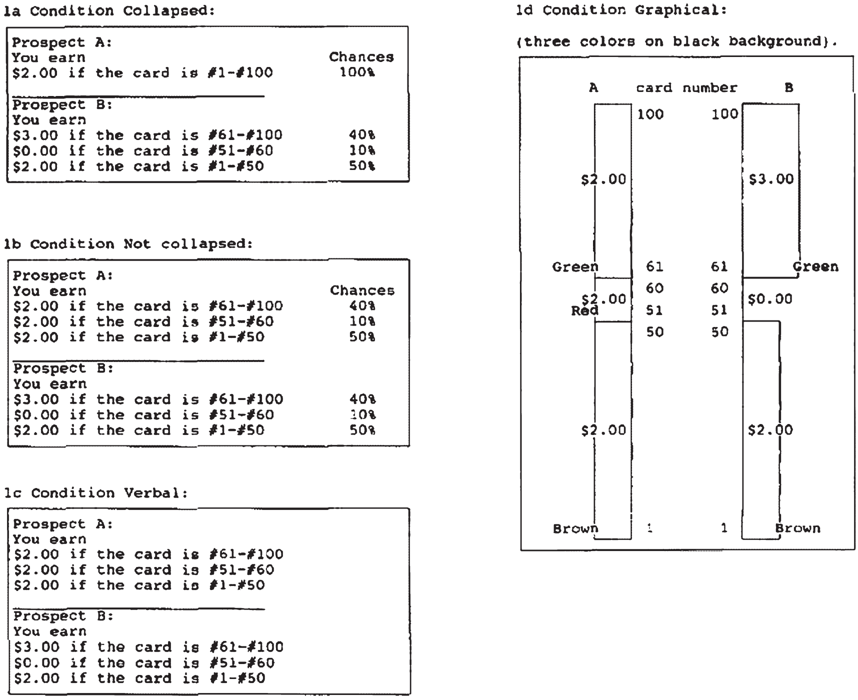


Fig. 21. Lottery Display Used by Beattie and Loomes (1997).



1a Condition Collapsed:

```
Prospect A:
You earn                                  Chances
$2.00 if the card is #1-#100               100%

Prospect B:
You earn
$3.00 if the card is #61-#100               40%
$0.00 if the card is #51-#60                10%
$2.00 if the card is #1-#50                 50%
```

1b Condition Not collapsed:

```
Prospect A:
You earn                                  Chances
$2.00 if the card is #61-#100               40%
$2.00 if the card is #51-#60                10%
$2.00 if the card is #1-#50                 50%

Prospect B:
You earn
$3.00 if the card is #61-#100               40%
$0.00 if the card is #51-#60                10%
$2.00 if the card is #1-#50                 50%
```

1c Condition Verbal:

```
Prospect A:
You earn
$2.00 if the card is #61-#100
$2.00 if the card is #51-#60
$2.00 if the card is #1-#50

Prospect B:
You earn
$3.00 if the card is #61-#100
$0.00 if the card is #51-#60
$2.00 if the card is #1-#50
```

1d Condition Graphical:

(three colors on black background).

Fig. 22. Lottery Displays Used by Wakker et al. (1994).

is a poor correspondence with untrained elicitors does not mean, however, that one could not undertake such a "semantic coordination game" using salient rewards, and try to encourage common usage of critical words.

The visual dots method is employed by Krupnick et al. (2002, p. 167), and provides a graphic image to complement the direct fractional, numerical representation of probability. An example of their visualization method is shown in Fig. 24.

Visual ladders have been used in previous research on mortality risk by Gerking, de Haan, and Schulze (1988) and Gegax, Gerking, and Schulze (1991). One such ladder, from their survey instrument, is shown in Fig. 25. An alternative ladder visualization is offered by Calman and Royston (1997; Fig. 1), and is shown in Fig. 26.

One hypothesis to emerge from this review of the representation of lotteries in laboratory and survey settings is that there is no single task representation for lotteries that is perfect for all subjects. It follows that

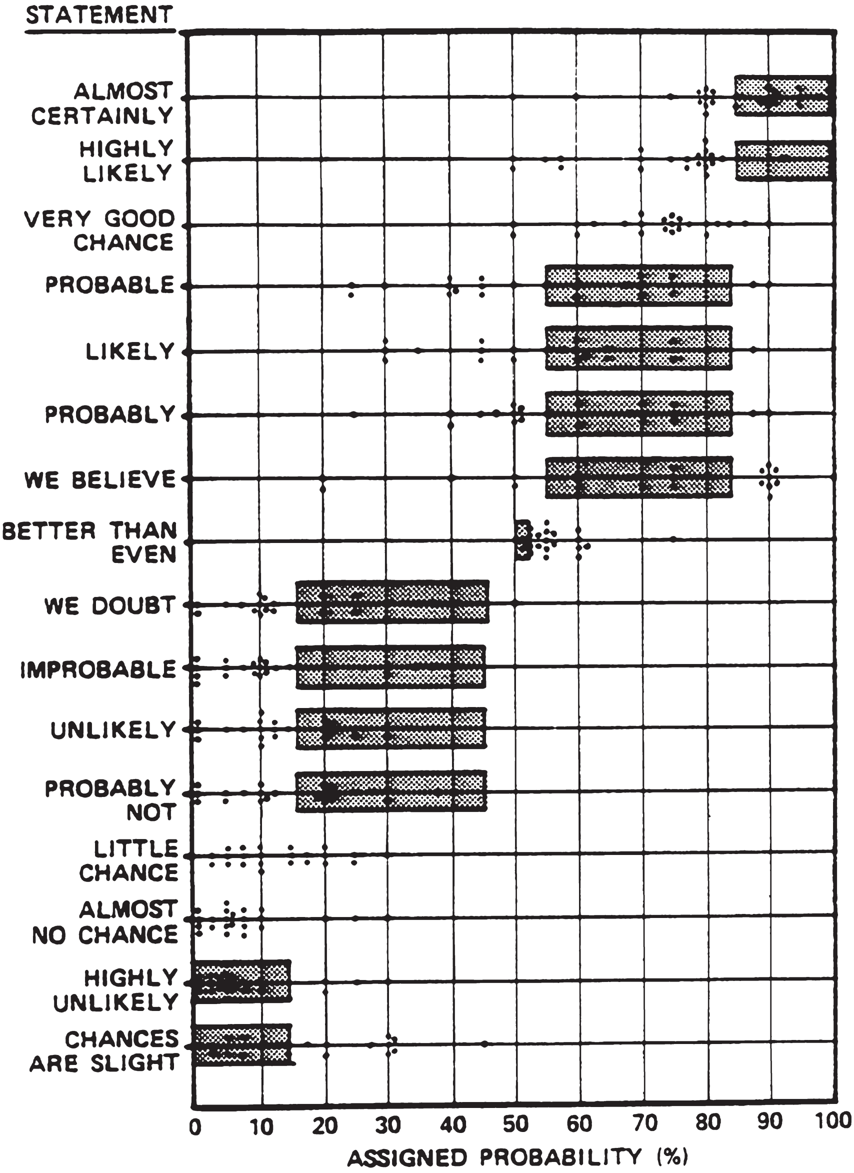


Fig. 23. Representing Risk on a Verbal Analogical Scale.

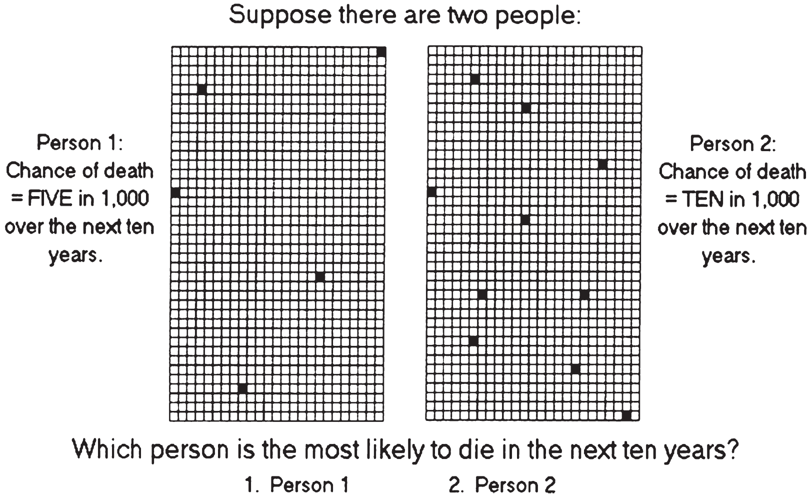


Fig. 24. Representing Risk with Dots.

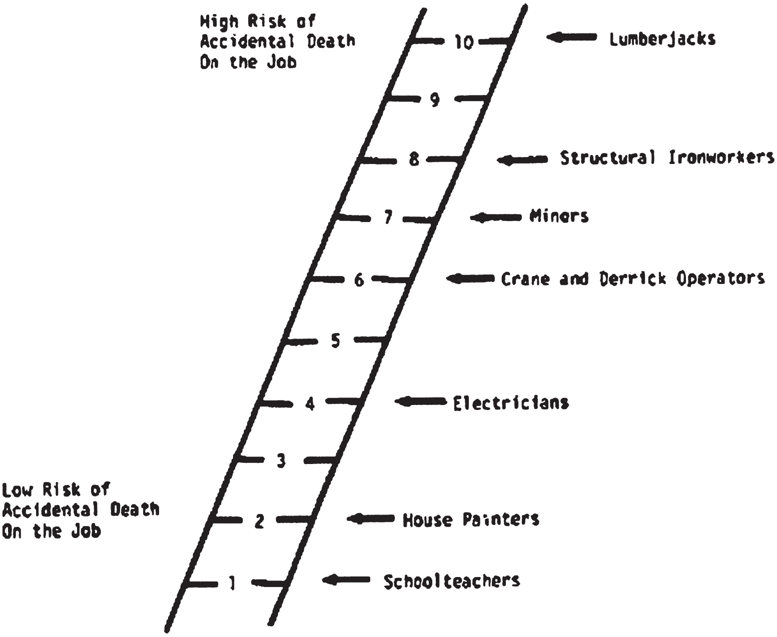


Fig. 25. Representing Risk with a 2D Ladder.

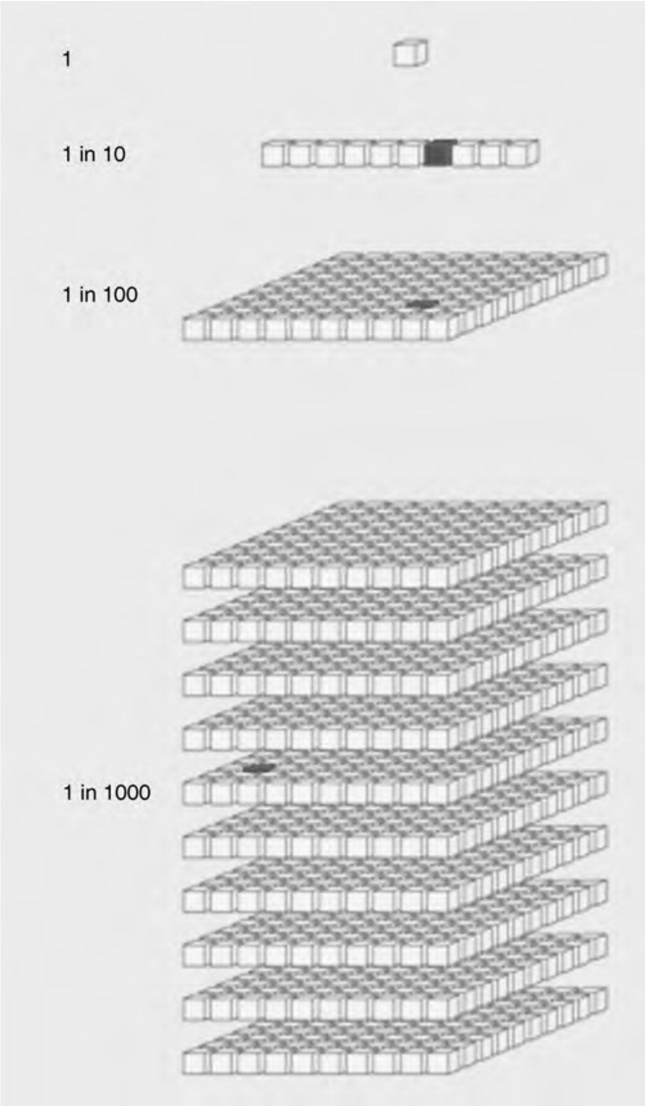


Fig. 26. Representing Risk with a 3D Ladder.

some of the evidence for framing effects in the representation of risk may be due to the implicit assumption that one form of representation works best for everyone: the "magic bullet" assumption. Rather, we should perhaps expect different people to perform better with different representations. To date no systematic comparison of these different methods have been performed and there is no consensus as to what constitutes a state of the art representation.

APPENDIX B. THE EXPERIMENTS OF HEY AND ORME (1994)

B1. The Original Experiments

The experiments of Hey and Orme (1994) are important in many respects. First, they use lottery tasks that are not designed as "trip wire" tests of one theory or another, but instead as representative lottery tasks. This design objective has strengths and weaknesses. The strength is that one can evaluate many different theories without the task domain being biased in favor of any one theory. Thus, tests of a theory will be based on tasks that are not just built to trick it into error. The weakness is that it might be inefficient as a domain for choosing between different theories. The second reason that these experiments are important, of course, is that they were evaluated using formal ML methods at the level of the individual, including explicit discussion of structural error models due to Fechner.

The basic experiments of HO are reviewed in Section 1.2, and the display subjects saw is presented in Fig. 3. Subjects were recruited from the University of York and participated in two sessions, each consisting of 100 binary lottery choices. The sample consisted of 80 students, who were allowed to proceed at their own pace. The lottery tasks took roughly 35 min to complete, and subjects earned an average of £17.50 per hour for this task and one other task.

B2. Replication

There are two limitations of the original HO experimental data, which make it useful to undertake a replication and extension. One is that there is no data on individual characteristics, so that it is impossible to pool data across subjects and condition estimation on those characteristics. Of course, this was not the objective of HO, who estimated choice functionals for each individual separately. But it does limit the use of these data for other purposes. Second, all of the lotteries were in the gain domain, and many theories require lotteries that are framed as losses or as mixtures of gains and losses. Hence, we review here the replications and extensions of Harrison and Rutström (2005), which address these two limitations.

Subjects were presented with 60 lottery pairs, each represented as a ''pie'' showing the probability of each prize. Fig. 4 illustrates one such representation. The subject could choose the lottery on the left or the right, or explicitly express indifference (in which case the experimenter would flip a coin on the subject's behalf). After all 60 lottery pairs were evaluated, and three were selected at random for payment.[87] The lotteries were presented to the subjects in color on a private computer screen,[88] and all choices recorded by the computer program. This program also recorded the time taken to make each choice. In addition to the choice tasks, the subjects provided information on demographic and other personal characteristics.

In the gain frame experiments the prizes in each lottery were \$0, \$5, \$10, and \$15, and the probabilities of each prize varied from choice to choice, and from lottery to lottery. In the loss frame experiments subjects were given an initial endowment of \$15, and the corresponding prizes from the gain frame lotteries were transformed to be $-$15, $-$10, $-$5, and \$0. Hence, the final outcomes, inclusive of the endowment, were the same in the gain frame and loss frame. In the mixed frame experiments subjects were given an initial endowment of \$8, and the prizes were transformed to be $-$8, $-$3, \$3, and \$8, generating final outcomes inclusive of the endowment of \$0, \$5, \$11, and \$16.[89]

In addition to the fixed endowment, each subject received a random endowment between \$1 and \$10. This endowment was generated using a uniform distribution defined over whole dollar amounts, operationalized by a 10-sided die. The purpose of this random endowment is to test for endowment effects on the choices.

The probabilities used in each lottery ranged roughly evenly over the unit interval. Values of 0, 0.13, 0.25, 0.37, 0.5, 0.62, 0.75, and 0.87 were used.[90] The presentation of a given lottery on the left or the right was determined at random, so that the ''left'' or ''right'' lotteries did not systematically reflect greater risk or greater prize range than the other.

Subjects were recruited at the University of Central Florida, primarily from the College of Business Administration, using the online recruiting application at ExLab (http://exlab.bus.ucf.edu). Each subject received a \$5 fee for showing up to the experiments, and completed an informed consent form. Subjects were deliberately recruited for "staggered" starting times, so that the subject would not pace their responses by any other subject. Each subject was presented with the instructions individually, and taken through the practice sessions at an individual pace. Since the rolls of die were important to the implementation of the objects of choice, the experimenters took some time to give each subject "hands-on" experience with the (10-sided, 20-sided, and 100-sided) die being used. Subjects were free to make their choices as quickly or as slowly as they wanted.

Our data consists of responses from 158 subjects making 9,311 choices that do not involve indifference. Only 1.7% of the choices involved explicit choice of indifference, and to simplify we drop those in estimation unless otherwise noted. Of these 158 subjects, 63 participated in gain frame tasks, 37 participated in mixed frame tasks, and 58 participated in loss frame tasks.

APPENDIX C. THE EXPERIMENTS OF HOLT AND LAURY (2002)

C1. Explaining the Data

Holt and Laury (2002) examine two main treatments with 212 subjects. The first is the effect of incentives. They vary the scale of the payoffs in the matrix shown in panel A of Table 1, which we take to be the scale of 1×. Every subject was presented with the first matrix of choices shown in panel A of Table 1, and with the exact same matrix at the end of the experiment. These two choices were always given to all subjects, and we will refer to them as task #1 and task #4. All subjects additionally had one *or* two intermediate choices, referred to here as task #2 and task #3. The question in task #2, if asked, was a *higher-scale, hypothetical* version of the initial matrix of payoffs. The question in task #3, if asked, was the *same* higher-scale version of payoffs but with *real* payoffs. Some subjects were asked one of these intermediate task questions; most subjects were asked both of them (hence for *some* subjects task #4 was actually their third and last task). Thus, we obtain the tabulation of individual responses shown in Table 9.

We see from Table 9 how each subject experienced different scales of payoffs in task #2 and/or task #3. This provides in-sample tests of the hypothesis that risk aversion does not vary with wealth, an important issue for those that assume specific functional forms such as CRRA or CARA.

Table 9. Sample Size and Design of the Holt and Laury (2002) Experiments.

Scale of Payoffs	Task				Total
	1	2	3	4	
1	212			212	424
20		118	150		268
50		19	19		38
90		18	18		36
All	212	155	187	212	766

A rejection of the "constancy" assumption in CRRA or CARA is not a rejection of EUT in general, of course, but just these particular (popular) parameterizations. In Section 3.7 and Appendix E, we see that some studies unfortunately equate "rejection of EUT" with "rejection of CRRA."

The second treatment in the HL design is the effect of hypothetical payoffs, which is why the questions in task #2 are included. Economic theory has no prediction when the task is not salient, and we have no control over subject behavior as an experimenter. The effect of using hypothetical responses is examined in depth in Harrison (2007) using these and other data, since the use of such data has been so prevalent in the empirical literature on the validity of EUT, but we do not consider them any further here. There is considerable evidence, bolstered by Holt and Laury (2005), that risk attitudes elicited with hypothetical responses are significantly different to risk attitudes elicited with real economic consequences, so this is a debate simply not worth pursuing.

Although having in-sample responses is valuable, it comes at a price in terms of control since there may be wealth effects from the subjects having earned some profit in the previous choice. To handle this HL use a nice trick: when the subjects proceed from task #1 to task #3, they are first asked if they are willing to give up their earnings in task #1 in order to play task #3. Since the stakes are so much higher in task #3, all subjects chose to do so. This means that the subjects face tasks #1 and #3 with no prior earnings from these experiments, although they do have experience with the type of task when facing task #3. No such trick can be applied for task #4, since the subjects would be unlikely to give up their earnings in task #3 in this instance. Thus, the responses to task #4 have no controls for wealth built in to the design. However, we do know the actual earnings of the subjects from the experimental data.

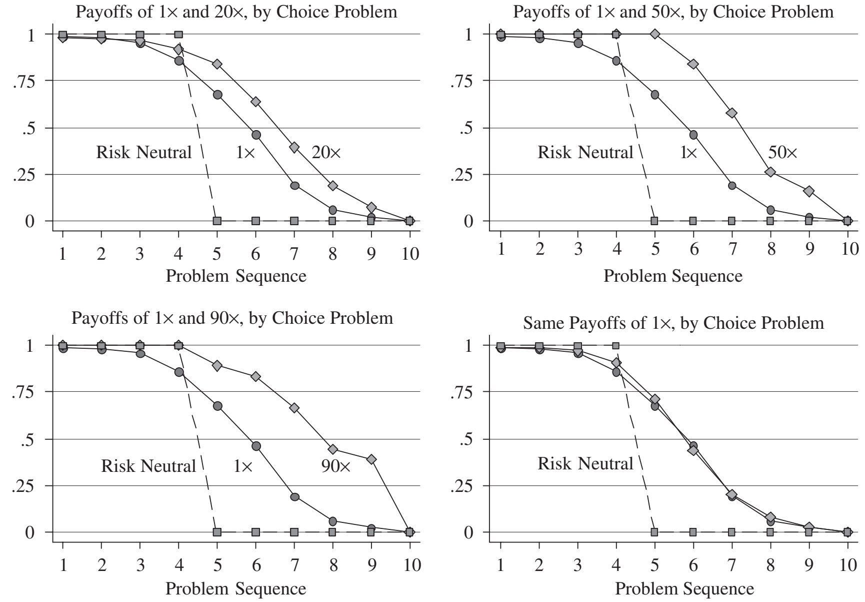


Fig. 27. Observed Choices in Holt and Laury (2002) Experiments.

HL also ask each subject to fill out a detailed question of individual demographic information, so their data include a rich set of controls for differences in risk preferences due to these characteristics.

Fig. 27 shows the main responses in the HL experiments. Consider the top left panel, which shows the average number of choices of the "safe" option A in each problem. In Problem 1, which is row 1 in panel A of Table 1, virtually everyone chooses option A (the safe choice). By the time the subjects get to Problem 10, which is the last row in panel A of Table 1, virtually everyone has switched over to problem B, the "risky" option. The dashed line marked RN shows the prediction if each and every subject were risk neutral: in this case everyone would choose option A up to Problem 4, then everyone would choose option B thereafter. The solid line marked with a circle shows the observed behavior in task #1, the low-payoff case. The solid line marked with a diamond shows the observed behavior in task #3, the high-payoff case. In the top left panel, the high payoff refers to payoff matrices that scale up the values in panel A of Table 1 by 20. The top right panel in Fig. 27 shows comparable data for the 50× problems, and the bottom left panel shows comparable data for the 90× problems.[91]

We examine the bottom-right panel later.

HL proceed with their analysis by looking at the first three pictures in Fig. 27 and drawing two conclusions. First, that one has to introduce some "noise" into any model of the data-generation process, since the observed choices are "smoother" than the risk-neutral prediction. A more general way of saying this is to allow subjects to have a specific degree of risk aversion, but to critically assume that they all have exactly the same degree of risk aversion. Thus, if subjects were a little risk averse the line marked RN would shift to the right and drop down a bit to the right, perhaps at Problem 6 or 7 instead of Problem 5. Of course, it would no longer represent risk-neutral responses, but it would still drop sharply, and that is the point being made by HL when arguing for a noise parameter. Second, and related to the previous explanation, the best-fitting line that assumes homogenous risk preferences would have to be a bit to the right of the risk-neutral line marked RN. So some degree of risk aversion, they argue, is needed to account for the *location* of the observed averages, quite apart from the need for a noise parameter to account for the *smoothness* of the observed averages.

Both conclusions depend critically on the assumption that every subject in the experiment has the same preferences over risk. The smoothness of the observed averages is easily explained if one allows heterogenous risk attitudes and no noise at all at the individual level: some people drop down at Problem 4, some more at Problem 5, some more at Problem 6, and so on. The smoothness that the eyeball sees in the aggregate data is just a counterpart of averaging this heterogenous process. The fact that *some* degree of risk aversion is needed for *some* subjects is undeniable, from the positive area above the RN line and below the circle or diamond lines from Problems 5 through 10. But it simply does not follow without further statistical analysis that all subjects, or even the typical subject, exhibit significant amounts of risk aversion. Nor does it follow that a noise parameter is needed to model these data.

These conclusions follow from inspection of each of the first three panels of Fig. 27, and just the RN and circle lines in each for that matter. Now turn to the comparison of the circle and diamond lines *within* each of the first three panels. The eyeball suggests that the diamond lines are to the right of the circle lines, which implies that risk aversion increases as the scale of payoffs increases. But this conclusion requires some measures of the uncertainty of these averages. Not surprisingly, the standard deviation in responses is the largest around Problems 5 through 7, suggesting that the confidence intervals around these diamond and circle lines could easily overlap. Again, this is a matter for an appropriate statistical analysis, not eyeball inspection of the averages.

Finally, compare the differences between the diamond and circle lines as one scans across the first three panels in Fig. 27. As the payoff scale gets larger, from 20× to 50× and then to 90×, it appears that the gap widens. That is, if one ignores the issue of standard errors around these averages, it appears that the degree of risk aversion increases. This leads HL to reject CRRA and CARA, and to consider generalized functional forms for utility functions that admit of increasing risk aversion. However, as Table 9 shows, the sample sizes for the 50× and 90× treatments were significantly smaller than those for the 20× treatment: 38 and 36 subjects, respectively, compared to 268 subjects for the 20× treatments. So one would expect that the standard errors around the 50× and 90× high-payoff lines would be much larger than those around the 20× high-payoff lines. This could make it difficult to statistically draw the eyeball conclusion that scale increases risk aversion.

Finally, one needs to account for the fact that all of the high-payoff data in the HL experiments was obtained in a task that followed the low-payoff task. Income effects were controlled for, in an elegant manner described above. But there could still be *simple order effects* due to experience with the qualitative task. HL recognize the possibility of order effects when discussing why they had the high hypothetical task before the high real task: "Doing the high hypothetical choice task before high real allows us to hold wealth constant and to evaluate the effect of using real incentives. For our purposes, it would not have made sense to do the high real treatment first, since the careful thinking would bias the high hypothetical decisions." The same (correct) logic applies to comparisons of the second real task with the first real task.

The bottom, right panel examines the data collected by HL in task #1 and task #4, which have the same scale but differ only in terms of the order effect and the accumulated wealth from task #3. These lines appear to be identical, suggesting no order effect, but a closer statistical analysis that conditions on the two differences shows that there is in fact an order effect at work.

C2. Modeling Behavior

One of the major contributions of HL is to present ML estimates of a relatively flexible utility function using their data. Recognizing the apparent changes in RRA with the scale treatments, they note that CRRA would not be appropriate, and use a parameterization of the EP function introduced by Saha (1993).

APPENDIX D. THE EXPERIMENTS OF KACHELMEIER AND SHEHATA (1992)

To illustrate the use of the BDM procedure, and to point to some potential problems, consider the ''high payoff'' experiments from China reported by Kachelmeier and Shehata (1992). These involved subjects facing lotteries with prizes equal to 0.5 yuan, 1 yuan, 5 yuan, or 10 yuan. Although 10 yuan only converted to about $2.50 at the time of the experiments, this represented a considerable amount of purchasing power in that region of China, as discussed by KS (p.1123). There were four treatments. One treatment used 25 lotteries with 5 yuan, one used 25 lotteries with 10 yuan, one used 25 lotteries with 0.5 yuan followed by 25 lotteries with 5 yuan, and one used 25 lotteries with 1 yuan followed by 25 lotteries with 10 yuan. In all cases, the first of the battery of 25 lotteries was a hypothetical trainer, and is ignored in the analysis shown below.

Figs. 28 and 29 show the data from the experiments of KS in China. The vertical axis shows the ratio of the elicited CE to the expected value of the lottery, and the horizontal axis shows the probability of winning each lottery. Each panel in Fig. 28 shows a scatter of data from each prize treatment. In Fig. 28 we only show data from the first series of lottery choices, for comparability in terms of experience. In Fig. 29 we show the results for the high-prize treatments, with the first series on top and the second series on the bottom, to show the effect of experience with the general task.[92] To orient the analysis, a simple cubic-spline is drawn through the median-bands; these lines are consistent with the formal statistical analysis reported below, but help explain certain features of the data.

Four properties of these responses are evident from the pictures. First, the general tendency towards risk-loving behavior at the lower three prize levels, as evidenced by the CEs being greater than the expected value in Fig. 28. Second, the dramatic reduction in the dispersion of selling prices as the probability of winning increases to 1, as evidenced by the pattern of the scatter within each panel. Indeed, these pictures discard data for probabilities less than 0.15, and for ratios greater than 2.5, to allow reasonable scaling. The discarded data exhibit even more dramatic dispersion than is already evident at probability levels of 0.25. Third, the

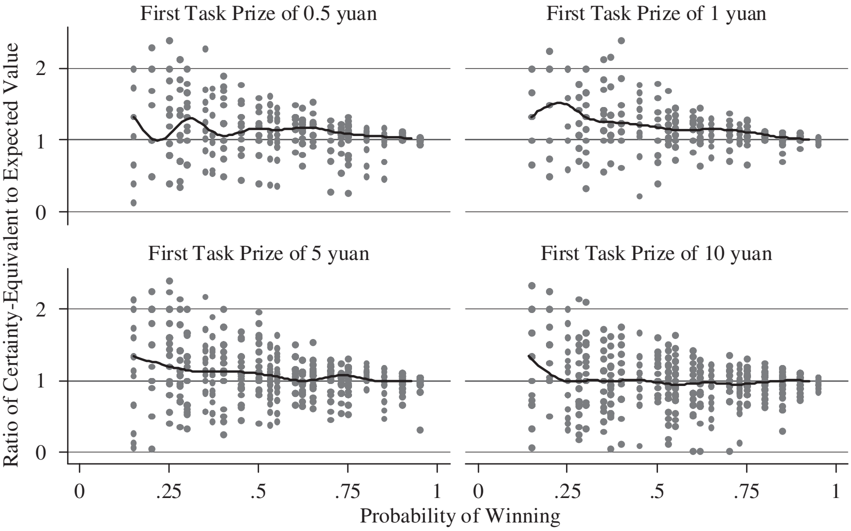


Fig. 28. Risk Premia and Probability of Winning in First Series of Kachelmeier–Shehata Experiments.

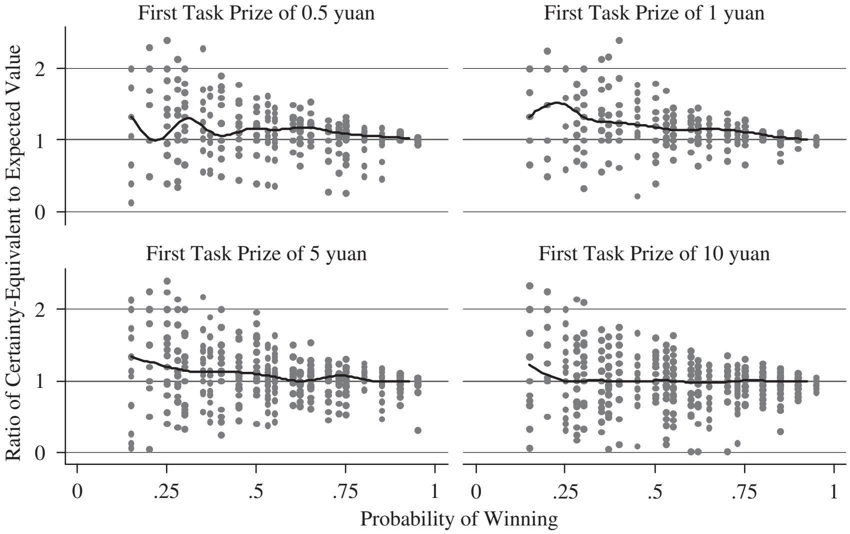


Fig. 29. Risk Premia and Probability of Winning in High Stakes Kachelmeier–Shehata Experiments.

responses for the highest prize treatment in Fig. 28 are much closer to being risk neutral or risk averse, at least for winning probabilities greater than around 0.2. Finally, the data in Fig. 29 suggests that subjects are less risk loving when they have experience with the task, and/or that there is an increase in risk aversion due to the accumulation of experimental income from the first series of tasks.

Since the BDM method generates a CE for each lottery, it is possible to estimate the CRRA coefficient directly for *each response* that a subject makes using the BDM method.[93] If p is the winning probability for prize y, and s is the CE elicited as a selling price from the subject, then the coefficient is equal to $1-\{\ln(p)/(\ln(s)-\ln(y))\}$. In this form, a value of zero indicates risk neutrality, and negative (positive) values risk-loving (risk averse) behavior.

The behavior of the CRRA coefficient elicited using the BDM method is extremely sensitive to experimental conditions, even if one restricts attention to the high-stakes lotteries and win probabilities within 15% of the boundaries.[94] First, the coefficients for low win probabilities imply extreme risk loving. This is perfectly plausible given the paltry stakes involved in such lotteries. Second, the coefficient depends on accumulated earnings, as hypothesized by McKee (1989). Increases in the *average accumulated* income earned in the task increase risk aversion, and increases in the three-round *moving average* of income decrease risk aversion.[95] Third, "bad joss," as measured by the fraction of random buying prices below the expected buying price of 50% of the prize, is associated with a large increase in risk-loving behavior.[96] Fourth, as Fig. 29 would suggest, experience with the general task increases risk aversion. Fifth, increasing the prize from 5 yuan to 10 yuan increases risk aversion significantly. Of course, this last result is consistent with non-constant RRA, and should not be necessarily viewed as a problem unless one insisted on applying the same CRRA coefficient over these two reward domains.

Fig. 30 summarizes the distribution of CRRA coefficients for the high-task decisions in KS. The dispersion of estimates is high, even though there is a marked tendency towards RN with the 10 yuan task and with experienced subjects. One of the key results here, as stressed by Kachelmeier and Shehata (1994), is that there is considerable variation in CRRA coefficients *within* each subject's sample of responses, as well as between subjects. The within-subjects' standard deviation in CRRA coefficients is 1.10, and the between-subjects' standard deviation is 1.13, around a mean of negative 1.36.

To deal with some of these problems we recommend paying subjects for just one stage to avoid intra-session income effects, the use of physical randomizing device to encourage subjects to see the random buyout price as independent of their selling price, the use of winning probabilities between 1/4 and 3/4 to avoid

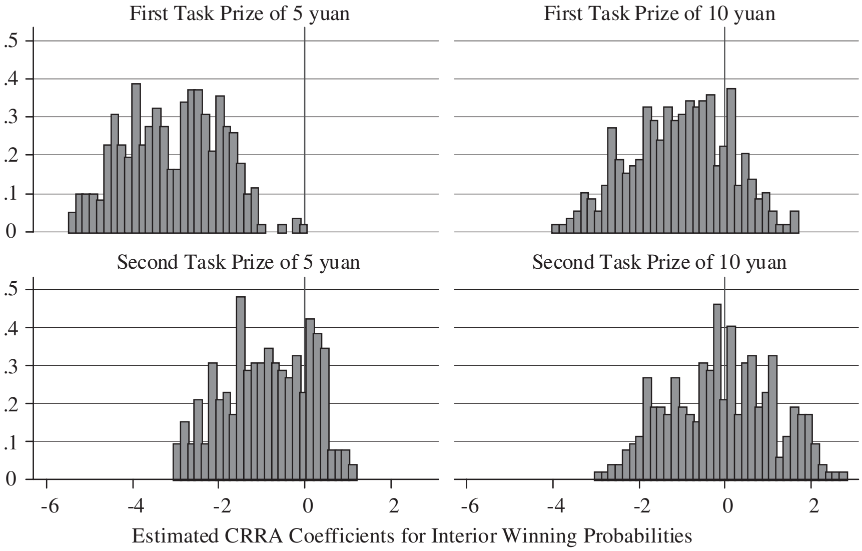


Fig. 30. Estimates of Risk Aversion from Kachelmeier–Shehata Experiments.

the more extreme effects of the end-point probabilities, and the provision of experience in the task in a completely prior session. We would also utilize extended instructions along the lines developed by Plott and Zeiler (2005).

APPENDIX E. THE EXPERIMENTS OF GNEEZY AND POTTERS (1997)

The experimental task of Gneezy and Potters (1997) was very simple, and was followed exactly by Haigh and List (2005). Each subject in the baseline treatment made nine decisions over a fixed stake. In GP this stake was 2 Dutch Guilders, which we will call \$2.00 for pedagogic ease. In each round they could choose a fraction of the stake to bet. If they chose to bet nothing then they received \$2.00 in that round for certain. If they bet \$$x$ then they faced a 2/3 chance of losing \$$x$ and a 1/3 chance of winning \$2.5x. These earnings were on top of the initial stake of \$2.00. Thus, the subject literally ended up with (\$2.00 – \$$x$) with probability 2/3 and (\$2.00 + \$2.5x) with probability 1/3. Since \$$x$ could not exceed \$2.00, by design, the subject actually faced no losses for the round as a whole. Of course, if one ignores the \$2.00 stake the subject did face a loss. In the baseline condition the subject chose a bet in each round, the random outcome was realized, their earnings in that round tabulated, and then the next round decision was made.

In the alternative treatment the subject made three decisions instead of nine. The first decision was a single amount to bet in each of rounds 1 through 3, the second decision was a single amount to bet in each of rounds 4 through 6, and the third decision was a single amount to bet in each of rounds 7 through 9. Thus, the subject made *one decision or choice* for each of the outcomes in rounds 1, 2, and 3. To state it equivalently, since this is critical to follow, one decision was simply applied three times: it is not the case that the subject made three separate decisions at round 1 that were applied in rounds 1, 2, and 3, respectively. The subject could not say "bet x, y, and z% in rounds 1, 2, and 3," but could only instead say "bet x%," meaning that x% would be bet for the subject in each of round 1, 2, and 3. In all other respects the experimental task was the same: the only thing that varied was the horizon over which the choices were made. This is referred to as the Low frequency treatment (L), and the baseline is referred to as the High frequency treatment (H).

The raw data in the two sets of experiments are presented in Figs. 31 and 32, which show the distribution of percentage bets. The general qualitative outcome is for subjects to bet more in the L treatment than in the H treatment. Gneezy and Potters (1997; Table I, p. 639) report that 50.5% was bet in their treatment H and 67.4% in their treatment L over all 9 rounds. They conducted their experiments with 83 Dutch students, split roughly evenly across the two treatments in a between-subjects design. Haigh and List (2005) (HLI) report virtually the same outcomes: for their sample of 64 American college students, the fractions were 50.9% and 62.5%, respectively, and for their sample of 54 current and former traders from the Chicago Board of Trade the fractions were 45% and 75%, respectively.[97] Using unconditional non-parametric tests or panel Tobit models, these differences are statistically significant at standard levels.[98] Thus, it appears that samples of subjects drawn from the same population behave as if more risk averse in treatment H compared to treatment L, and that the average subject is risk averse. The latter inference follows from the fact that a risk-neutral subject, according to EUT, would bet 100% of the stake.

Figs. 31 and 32 also alert us to one stochastic feature of these data that will play a role later: that there is a substantial spike at the 100% bet level. From an EUT perspective, this corresponds to subjects that are risk neutral or risk loving.

If we just consider "interior bets" then the same qualitative results obtain. In GP, the Low frequency treatment generates an average 42.1% bet

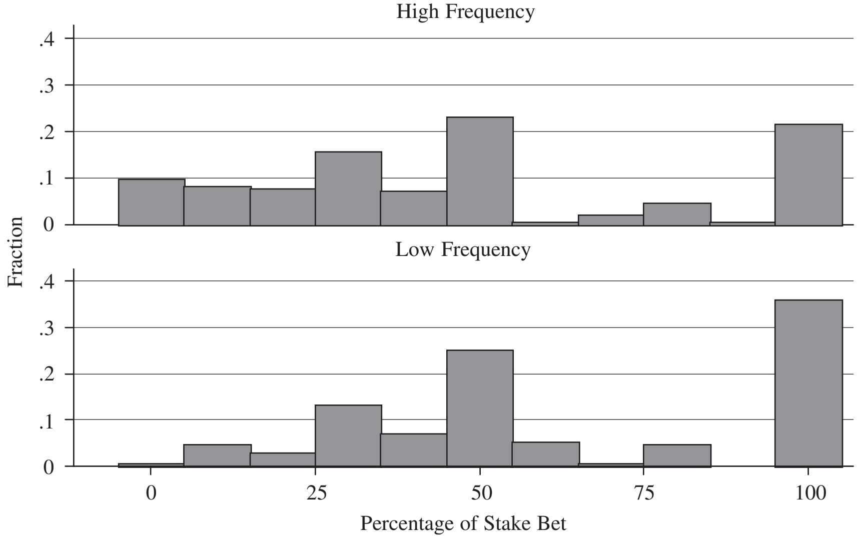


Fig. 31. Distribution of Percentage Bets in Gneezy and Potters (1997) Experiments.

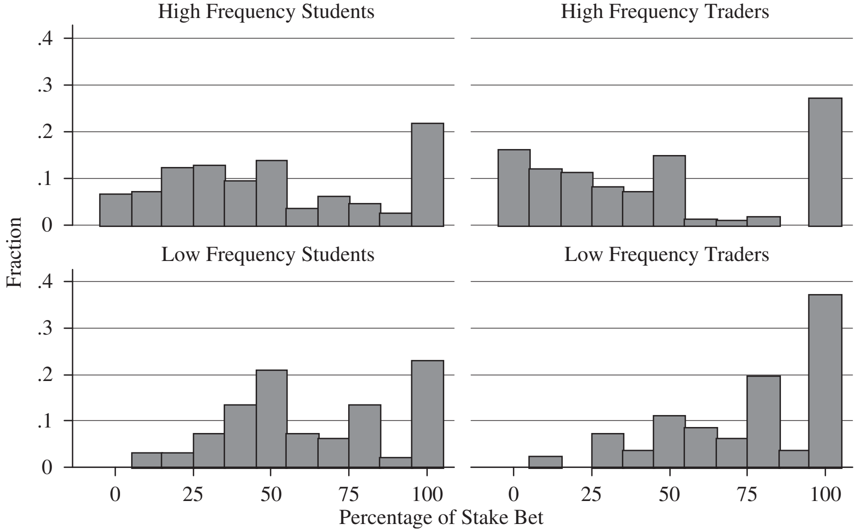


Fig. 32. Distribution of Percentage Bets in Haigh and List (2005) Experiments.

compared to an average 33.9% bet in the High frequency treatment. In HLI, the students (traders) bet an average of 37.7% (25.3%) and 51.4% (59.3%) in each treatment.

E1. Explaining the Data

When interpreting the experiments of GP and HLI it is important to view subjects as having a utility function that is defined over prize income that reflects the stakes that *choices* are being made over. The high frequency subjects can be viewed as making a series of nine choices over stakes defined, for each choice, by a vector *y* which takes on a range of integer values between \$0 and \$7. The subject could get \$0 if they bet 100% of the stake and lost it; or they could get as much as \$7 if they bet 100% of the stake and won 2.5 × \$2.00.

The low frequency subjects, on the other hand, made three choices over stakes defined by the possible combinations of gains and losses over three random draws. Thus, they could end up with three losses, 2 losses and 1 gain, 1 loss and 2 gains, or 3 gains. The probabilities for each outcome, irrespective of order, are 0.30, 0.44, 0.22, and 0.04, respectively. The monetary outcome in each case depends on the fraction of the stake that the subject chose to bet.

Table 10 spells out the arithmetic for different bets. For simplicity we evaluate the possible choices in increments of 10 cents, but of course the choices could be in pennies.[99] The second column shows the bet as a percent of the stake of \$2.00. Columns 3 though 7 show the components of the lottery facing the subject in the High frequency treatment for each possible bet, and columns 8 through 16 show the same components for the subject in Low frequency treatment. Consider, for example, a bet of 10 cents, which is 5% of the stake. If the subject is in the High treatment and loses, they earn 190 ($= 200 - 10$) cents in that period; this occurs with probability 2/3. If the subject is in the High treatment and wins, they earn 225($= 200 + 10 \times 2.5 = 200 + 25$) cents; this occurs with probability 1/3. In the corresponding entry for the subject in the Low treatment, the value of prizes is calculated similarly, but for three random draws. Thus, in the LLL outcome, the subject earns 570($= 200 - 10 + 200 - 10 + 200 - 10$) cents.

From Table 10 we see instantly that a risk-neutral subject that obeyed EUT would bet 100% of the pie in both treatments and thereby maximize expected value. It can also be inferred that a moderately risk-averse subject would bet some fraction of the pie in each treatment, less than 100%, and that a risk-loving subject would always bet 100% of the pie.

Table 10. Illustrative Calculations Assuming Risk Neutrality.

Possible Choices		High Frequency Treatment					Low Frequency Treatment								
Bet in cents	Bet as %	L	p(L)	G	p(G)	EV	LLL	p(LLL)	LLG	p(LLG)	LGG	p(LGG)	GGG	p(GGG)	EV
0	0	200	0.67	200	0.33	200.0	600	0.30	600	0.44	600	0.22	600	0.04	600
10	5	190	0.67	225	0.33	201.7	570	0.30	605	0.44	640	0.22	675	0.04	605
20	10	180	0.67	250	0.33	203.3	540	0.30	610	0.44	680	0.22	750	0.04	610
30	15	170	0.67	275	0.33	205.0	510	0.30	615	0.44	720	0.22	825	0.04	615
40	20	160	0.67	300	0.33	206.7	480	0.30	620	0.44	760	0.22	900	0.04	620
50	25	150	0.67	325	0.33	208.3	450	0.30	625	0.44	800	0.22	975	0.04	625
60	30	140	0.67	350	0.33	210.0	420	0.30	630	0.44	840	0.22	1050	0.04	630
70	35	130	0.67	375	0.33	211.7	390	0.30	635	0.44	880	0.22	1125	0.04	635
80	40	120	0.67	400	0.33	213.3	360	0.30	640	0.44	920	0.22	1200	0.04	640
90	45	110	0.67	425	0.33	215.0	330	0.30	645	0.44	960	0.22	1275	0.04	645
100	50	100	0.67	450	0.33	216.7	300	0.30	650	0.44	1000	0.22	1350	0.04	650
110	55	90	0.67	475	0.33	218.3	270	0.30	655	0.44	1040	0.22	1425	0.04	655
120	60	80	0.67	500	0.33	220.0	240	0.30	660	0.44	1080	0.22	1500	0.04	660
130	65	70	0.67	525	0.33	221.7	210	0.30	665	0.44	1120	0.22	1575	0.04	665
140	70	60	0.67	550	0.33	223.3	180	0.30	670	0.44	1160	0.22	1650	0.04	670
150	75	50	0.67	575	0.33	225.0	150	0.30	675	0.44	1200	0.22	1725	0.04	675
160	80	40	0.67	600	0.33	226.7	120	0.30	680	0.44	1240	0.22	1800	0.04	680
170	85	30	0.67	625	0.33	228.3	90	0.30	685	0.44	1280	0.22	1875	0.04	685
180	90	20	0.67	650	0.33	230.0	60	0.30	690	0.44	1320	0.22	1950	0.04	690
190	95	10	0.67	675	0.33	231.7	30	0.30	695	0.44	1360	0.22	2025	0.04	695
200	**100**	**0**	**0.67**	**700**	**0.33**	**233.3**	**0**	**0.30**	**700**	**0.44**	**1400**	**0.22**	**2100**	**0.04**	**700**

Note: Bold row show EUT-consistent choices.

p(LLL) = 2/3 × 2/3 × 2/3; p(LLG) = 2/3 × 2/3 × 1/3, and can occur in three equivalent ways (LLG, LGL, and GLL), so the probability shown is 2/3 × 2/3 × 1/3 × 3; p(LGG) = 2/3 × 1/3 × 1/3, and can also occur in three equivalent ways; and p(GGG) = 1/3 × 1/3 × 1/3.

The outcomes of the lotteries being evaluated by subjects in the High and Low treatments differ significantly. Consider the 50% bet, in the middle of Table 10. For subjects in the High treatment the two final outcomes from each choice are 100 and 450, occurring with the probabilities shown there. For subjects in the Low treatment there are four final outcomes from each choice: 300, 650, 1,000, and 1,350. Thus, *the monetary rewards from the same percentage choice differ significantly*. So, to explain why subjects in the High treatment are more risk averse than subjects in the Low treatment, it suffices at a qualitative level to find some utility function that has moderate amounts of risk aversion for "low" income levels and smaller amounts of risk aversion for "higher" income levels.

Although less obvious than the RN prediction, any subject exhibiting CRRA would choose *the same bet fraction in each row*. The more risk averse they were, the smaller would be the bet, but it would be the same bet in each of the High and Low treatments. This result is important since *every* statement of "the EUT null hypothesis" in the MLA literature that we can find uses RN or CRRA specifications for the utility function.[100] Thus, it is easy to see why evidence of a difference between the bet fractions in the High and Low treatments is viewed as a rejection of EUT.

Of course, this does not test EUT at all. It only tests a very special case of EUT, where the specific functional form seems to have been chosen to perform poorly.[101] It is easy to propose more flexible utility functions than CRRA. There are many such functions, but one of the most popular in recent work that is fully consistent with EUT has been the EP utility function proposed by Saha (1993). Following Holt and Laury (2002), the EP function is defined as

$$\frac{U(x) = (1 - \exp(-\alpha x^{1-r}))}{\alpha}$$

where α and r are parameters to be assumed or estimated. RRA is then $r + \alpha(1-r)y^{1-r}$, so RRA varies with income if $\alpha \neq 0$. This function nests CRRA (as $\alpha \to 0$) and CARA (as $r \to 0$). At a qualitative level, if $r > 0$ and $\alpha < 0$ one can immediately rationalize the qualitative data in these experiments: $\text{RRA} = r + \alpha(1-r)y^{1-r} \to r$ as $y \to 0$, and then one has declining RRA with higher prize incomes since $\alpha < 0$.

E2. Modeling Behavior

The qualitative insight that one can explain these data with a simple EUT specification can be formalized by estimating the parameters of a model that account for the data. Such an exercise also helps explain some differences between the traders and students in Haigh and List (2005).

As noted earlier, Figs. 31 and 32 alert us to the fact that the behavioral process generating data at the 100% bet level may be different than the process generating data at the "interior" solutions. From a statistical perspective, this is just a recognition that a model that tries to explain the interior modes of these data, and why they vary between the High and Low treatments, might have a difficult time also accounting for the spike at 100%. One approach is just to ignore that spike, and see what estimates obtain. Another approach is to construct a model and likelihood function that accounts for these two processes.[102] We apply both approaches, although favoring the latter *a priori*.

The dependent variable is naturally characterized as the fraction of the stake bet, denoted π. Therefore, the likelihood function is constructed using the specification developed by Papke and Wooldridge (1996) for fractional dependent variables. Specifically, the log-likelihood of observation i is defined as $l_i(\xi) = \pi_i \times \log(\mathrm{G}(x_i, \xi)) + (1-\pi_i) \times \log(1-\mathrm{G}(x_i, \xi))$ for parameter vector ξ, a vector of explanatory variables x, and some convenient cumulative distribution function $\mathrm{G}(\cdot)$. We use the cumulative Gamma distribution function $\mathrm{G}(z) = \Gamma(a,z)$, where a is a parameter that can be estimated.[103] The index z_i is the expected utility of the bet chosen, conditional on some parameter estimates of ξ and some characteristics x_i for observation i.

The index z is constructed using information on the lottery for the actual bet, reflecting a more detailed version of the arithmetic underlying Table 10. Thus, for a particular fractional bet, the parameters of the task imply that the subject was facing a particular lottery. So, one element of the x vector is whether or not the subject was in the High or Low treatment. Another element is the stake. Another element is the set of parameters of the experimental task defining the lottery outcomes (e.g., the probabilities of a loss or a gain, and the numbers defining how the bet is scaled to define the loss or the gain). Using this information, and candidate estimates of r and α for the EP utility function, the likelihood constructs the expected utility of the observed choice, and the ML estimates find the parameters of the EP utility function that best explain the observed choices.

This approach can be applied directly to the data in Figs. 31 and 32, recognizing that one model must explain the multiple modes of these distributions. Alternatively, one can posit a natural two-step decision process, where the subject first decides if they are going to bet everything or not, and then if they decide not to, decides how much to bet (including 0%). This might correspond to one way that a risk-averse or risk-loving subject might process such tasks: first figure out what a RN decision-maker would do, since that is computationally easier, and then shade one's choice in the direction dictated by risk preferences. Since the matrix in Table 10 was not presented to subjects in such an explicit form, this would be *one* sensible heuristic to use.

Irrespective of the interpretation, this proposed decision process implies a statistical "hurdle" model. First the subject makes a binary choice to contribute 100% or less. Then the subject decides what fraction to contribute, conditional on contributing less than 100%. The first stage can be modeled using a standard probit specification, although it is the second stage that is really of greatest interest.

A key feature of these estimates is that they pool the data from High and Low treatments. The objective is to ascertain if one EUT-consistent model can explain the shift in the distributions between these treatments in Figs. 31 and 32. Since each subject provided multiple observations there are clustering corrections for the possible correlation of errors associated with a given subject.

Table 11 reports the results of ML estimation of these models. Panel A provides estimates for the individual responses from Gneezy and Potters (1997). These estimates show some initial risk aversion at zero income levels ($r = 0.21$) and then some slight evidence of declining RRA as income rises ($\alpha = -10.019$). However, the evidence of declining RRA is not statistically significant, although the 95% confidence interval is skewed towards negative values. Much more telling evidence comes from comparable estimates for the interior bets, in panel B. Here we find striking evidence of the qualitative explanation presented earlier: initial risk aversion at zero income levels ($r = 1.12$) and sharply declining RRA as income rises ($\alpha = -0.57$). The point estimate of r exceeds 1 in this case, which violates the assumption of non-satiation. But the standard error on this estimate is 0.25, with a 95% confidence interval between 0.61 and 1.63. So we cannot reject the hypothesis that $r \leq 1$; in fact, the p-value that the coefficient equals 1 is 0.68, so we cannot reject that specific hypothesis.

Panels C through E report estimates for the treatments of Haigh and List (2005), estimated separately for traders and students since that was their main treatment. With the exception of the estimates in panel E, for *all* bets by University of Maryland (UMD) students, these results again confirm the qualitative explanation proposed above. Therefore, one must simply reject the conclusion of Haigh and List (2005, p. 531) that their "findings suggest that expected utility theory may not model professional traders' behavior well, and this finding lends credence to behavioral economics and finance

Table 11. Maximum Likelihood Estimates of Expo-Power Utility Function.

Coefficient	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
A. Gneezy and Potters (1997) – Estimates for All Bets by Dutch Students					
r	0.21	0.08	0.009	0.06	0.37
α	−0.02	0.03	0.463	−0.07	0.03
a	2.32	0.22	0.000	1.87	2.76
B. Gneezy and Potters (1997) – Estimates for Interior Bets by Dutch Students					
r[a]	1.12	0.25	0.000	0.61	1.63
α	−0.57	0.09	0.000	−0.74	−0.40
a	1.88	0.29	0.000	1.30	2.46
C. Haigh and List (2005) – Estimates for All Bets by CBOT Traders					
r	0.36	0.05	0.000	0.26	0.46
α	−0.13	0.02	0.000	−0.16	−0.10
a	3.67	0.42	0.000	2.82	4.53
D. Haigh and List (2005) – Estimates for Interior Bets by CBOT Traders					
r	0.67	0.04	0.000	0.60	0.74
α	−0.44	0.01	0.000	−0.46	−0.42
a	3.69	0.34	0.000	3.01	4.37
E. Haigh and List (2005) – Estimates for All Bets by UMD Students[b]					
r	−0.99	0.27	0.001	−1.54	−0.44
α	0.22	0.05	0.000	0.13	0.32
a	1.71	0.21	0.000	1.28	2.13

[a] See text for discussion of the point estimate for r exceeding 1, since that violates the non-satiation assumption for this specification.

[b] There are no estimates for the sub-sample of interior bets, since the estimate of r exceeds 1, and is statistically significantly greater than 1.

models, which are beginning to relax inherent assumptions used in standard financial economics.'' Whether MLA models the behavior of traders better than EUT is a separate matter, but EUT easily explains the data. In fact, these data are more consistent with the priors that motivated the Haigh and List (2005) study, illustrated by List (2003), that students would be *more* likely to exhibit anomalies than field traders.

E3. Coals To Newcastle: An Anomaly for the Behaviorists

The reason that MLA is interesting is that Benartzi and Thaler (1995) use it to provide an intuitive explanation for the equity premium puzzle. Their empirical approach is to assume a particular numerical specification of MLA, and then solve for the "evaluation horizon"[104] of returns to stocks and equities that makes their expected utility[105] equivalent. They find that this horizon is roughly 12 months, which strikes one as *a priori* plausible if one had to pick a single representative evaluation horizon for all investors.[106] Thus, they assume a particular empirical version of MLA and further assume that these coefficients do not change as they counter-factually calculate the effects of alternative evaluation horizons:

> According to our theory, the equity premium is produced by a combination of loss aversion and frequent evaluation. Loss aversion plays the role of risk aversion in standard models, and can be considered a fact of life (or, perhaps, a fact of preferences). In contrast, the frequency of evaluations is a policy choice that presumably could be altered, at least in principle. Furthermore, as the charts (…) show, stocks become more attractive as the evaluation period increases.

So the parameters of the MLA specification are assumed invariant to evaluation horizon, as an essential premiss of the empirical methodology.

Thus the motivation for the experiments of GP and HL. As GP note, Benartzi and Thaler (1995) "… do not present direct (experimental) evidence for the presence of MLA. The evidence presented in (BT) is only circumstantial. (…) We have experimental subjects making a sequence of risky choices. To analyze the presence of MLA, we do not try to estimate the period over which subjects evaluate financial outcomes, but rather we try to manipulate this evaluation period." Hence the data from GP can be used to recover the MLA preferences that are consistent with the observed behavior, and the empirical premiss of Benartzi and Thaler (1995) evaluated.

Since behavioral economists are so enamored of anomalies, it may be useful to point out one or two in the MLA literature being considered here. The first anomaly is that the data from the experiments of GP demonstrate that *the MLA parameters themselves depend on the evaluation horizon*, which of course was varied by experimental design in their data. Hence, one cannot assume that those parameters stay fixed as one calibrates the equity premium by varying the evaluation horizon. The second anomaly is that these data also *imply risk attitudes defined over the utility function that are qualitatively the opposite of those customarily assumed.*

The MLA parameterization adopted by Benartzi and Thaler (1995, p. 79) is taken directly from Tversky and Kahneman (1992), both in terms of the functional forms and parameter values. They assume a power utility function defined separately over gains and losses: $U(x) = x^{\alpha}$ if $x \geq 0$, and $U(x) = -\lambda(-x)^{\beta}$ for $x < 0$. So α and β are the risk aversion parameters, and

λ is the coefficient of loss aversion. Tversky and Kahneman (1992, p. 59) provide estimates that have been universally employed in applied work by behaviorists: $\alpha = \beta = 0.88$ and $\lambda = 2.25$.

Using the data from GP we estimate the parameters of this MLA model. For simplicity we assume no probability weighting, although that could be included. Benartzi and Thaler (1995, p. 83) and GP stress that it is the loss aversion parameter λ that drives the main prediction of MLA, rather than probability weighting or even risk aversion in the utility function. The likelihood function is again constructed using the specification developed by Papke and Wooldridge (1996) for fractional dependent variables. Since there are no data on personal characteristics in the GP data, the x vector refers solely to whether or not the decision was made in the Low frequency setting or the High frequency setting. Thus, $\xi = (\alpha, \beta, \lambda)$, and each of those fundamental parameters is estimated as a linear function of binary dummies for the Low and High frequencies.[107]

Table 12 reports the ML estimates obtained. The "good news" for MLA is that they provide strong evidence that the loss aversion parameter is greater than 1. The "bad news" for MLA is that they provide equally striking evidence that all of the parameters of the MLA specification vary with the evaluation horizon. The "awkward news" for MLA is that they provide inconsistent evidence about risk attitudes in relation to the received empirical wisdom.

The estimates for α indicate *risk-loving behavior over gains*.[108] There does not appear to be much difference in risk attitudes over gains, and indeed one cannot reject the null hypothesis that they are equal with a Wald test (p-value = 0.391). The estimates for β indicate a severe case of *risk aversion over losses*. Moreover, subjects appear to be *more* risk averse in the Low frequency setting than in the High frequency setting: a Wald test of the null

Table 12. Maximum Likelihood Estimates of Myopic Loss Aversion Utility Function[a].

Coefficient	Variable	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
α	Low frequency	1.48	0.04	0.000	1.40	1.55
	High frequency	1.38	0.10	0.000	1.18	1.59
β	Low frequency	0.03	0.07	0.689	−0.11	0.17
	High frequency	0.55	0.28	0.052	0.00	1.10
λ	Low frequency	1.90	0.08	0.000	1.74	2.07
	High frequency	4.28	0.64	0.000	2.99	5.56

[a] Estimates from responses in Gneezy and Potters (1997) experiments.

hypothesis of equality has a p-value of 0.074. Finally, the estimates for λ are consistent with loss aversion, since they are both each significantly greater than 1 (p-values < 0.0001). However, these subjects appear to be significantly *more* loss averse in the High frequency setting than in the Low frequency setting (p-value $= 0.0005$).

This new analysis of the GP data therefore imply that the MLA parameters depend on the evaluation horizon and that subjects are risk loving in gains and risk averse in losses, thus pointing to anomalies compared to the standard view of PT.

APPENDIX F. ESTIMATION USING MAXIMUM LIKELIHOOD

Economists in a wide range of fields are now developing customized likelihood functions to correspond to specific models of decision-making processes. These demands derive partly from the need to consider a variety of parametric functional forms, but also because these models often specify non-standard decision rules that have to be "written out by hand." Thus, it is becoming common to see user-written ML estimates, and less use of pre-packaged model specifications.

These pedagogic notes document the manner in which one can estimate ML models of utility functions within *Stata*.[109] However, we can quickly go beyond "utility functions" and consider a wide range of decision-making processes, to parallel the discussion in the text. We start with a standard CRRA utility function and binary choice data over two lotteries, assuming EUT. This step illustrates the basic economic and statistical logic, and introduces the core *Stata* syntax. We then quickly consider an extension to consider loss aversion and probability weighting from PT, the inclusion of "stochastic errors," and the estimation of utility numbers themselves to avoid any parametric assumption about the utility function. We then illustrate a replication of the ML estimates of HL. Once the basic syntax is defined from the first example, it is possible to quickly jump to other likelihood functions using different data and specifications. Of course, this is just a reflection of the "extensible power" of a package such as *Stata*, once one understands the basic syntax.[110]

F1. Estimating a CRRA Utility Function

Consider the simple CRRA specification in Section 2.2. This is an EUT model, with a CRRA utility function, and no stochastic error specification.

The following *Stata* program defines the model, in this case using the lottery choices of Harrison and Rutström (2005), which are a replication of the experimental tasks of Hey and Orme (1994):

```
* define Original Recipe EUT with CRRA and no errors
program define ML_eut0

    args lnf r
    tempvar prob0l prob1l prob2l prob3l prob0r prob1r prob2r prob3r y0 y1 y2 y3
    tempvar euL euR euDiff euRatio tmp lnf_eut lnf_pt p1 p2 f1 f2

    quietly {

        * construct likelihood for EUT
        generate double `prob0l' = $ML_y2
        generate double `prob1l' = $ML_y3
        generate double `prob2l' = $ML_y4
        generate double `prob3l' = $ML_y5

        generate double `prob0r' = $ML_y6
        generate double `prob1r' = $ML_y7
        generate double `prob2r' = $ML_y8
        generate double `prob3r' = $ML_y9

        generate double `y0' = ($ML_y14+$ML_y10)^`r'
        generate double `y1' = ($ML_y14+$ML_y11)^`r'
        generate double `y2' = ($ML_y14+$ML_y12)^`r'
        generate double `y3' = ($ML_y14+$ML_y13)^`r'

        gen double `euL' = (`prob0l'*`y0')+(`prob1l'*`y1')+(`prob2l'*`y2')+(`prob3l'*`y3')
        gen double `euR' = (`prob0r'*`y0')+(`prob1r'*`y1')+(`prob2r'*`y2')+(`prob3r'*`y3')

        generate double `euDiff' = `euR' - `euL'

        replace `lnf' = ln(normal( `euDiff')) if $ML_y1==1
        replace `lnf' = ln(normal(-`euDiff')) if $ML_y1==0

    }
end
```

This program makes more sense when one sees the command line invoking it, and supplying it with values for all variables. The simplest case is where there are no explanatory variables for the CRRA coefficient (we cover those below):

```
ml model lf ML_eut0 (r: Choices P0left P1left P2left P3left P0right P1right P2right

       P3right prize0 prize1 prize2 prize3 stake = ) if Choices~=., cluster(id)

       technique(nr) maximize
```

The "ml model" part invokes the *Stata* ML model specification routine, which essentially reads in the ML_eut0 program defined above and makes sure that it does not violate any syntax rules. The "lf" part of "lf ML_eut0" tells this routine that this is a particular type of likelihood specification (specifically, that the routine ML_eut0 does not calculate analytical derivatives, so those must be calculated numerically). The part in brackets defines the equation for the CRRA coefficient *r*. The "*r*:" part just labels this equation, for output display purposes and to help reference initial values if they are specified for recalcitrant models. There is no need for the "*r*:" here to match the "*r*" inside the ML_eut0 program; we could have referred to "*r*EUT:" in the "ml model" command. We use the same "*r*" to help see the connection, but it is not essential.

The "`Choices P0left P1left P2left P3left P0right P1right P2right P3right prize0 prize1 prize2 prize3 stake`" part tells the program what observed values and data to use. This allows one to pass parameter values as well as data to the likelihood evaluator defined in ML_eut0. Each item in this list translates into a $ML_y* variable referenced in the ML_eut0 program, where * denotes the order in which it appears in this list. Thus, the data in variable Choices, which consists of 0's and 1's for choices (and a dot, to signify "missing"), is passed to the ML_eut0 program as variable $ML_y1. Variable p0left, which holds the probabilities of the first prize of the lottery presented to subjects on the left of their screen, is passed as $ML_y2, and so on. Finally, variable stake, holding the values of the initial endowments provided to subjects, gets passed as variable $ML_y14. It is good programming practice to then define these in some less cryptic manner, as we do just after the "quietly" line in ML_eut0. This does not significantly slow down execution, and helps avoid cryptic code. There is no error if some variable that is passed to ML_eut0 is not referenced in ML_eut0.

Once the data is passed to ML_eut0 the likelihood function can be evaluated. By default, it assumes a constant term, so when we have " =)" in the above command line, this is saying that there are no other explanatory variables. We add some below, but for now this model is just assuming that one CRRA coefficient characterizes all choices by all subjects. That is, it assumes that everyone has the same risk preference.

We restrict the data that is passed to only include strict preferences, hence the "if Choices ~ = ." part at the end of the command line. The response of indifference was allowed in this experiment, and we code it as a "missing" value. Thus, the estimation only applies to the sub-sample of strict preferences. One could modify the likelihood function to handle indifference.

Returning to the ML_eut0 program, the "args" line defines some arguments for this program. When it is called, by the default Newton–Raphson optimization routine within *Stata*, it accepts arguments in the "*r*" array and returns a value for the log-likelihood in the "lnf" scalar. In this case, "*r*" is the vector of coefficient values being evaluated.

The "tempvar" lines create temporary variables for use in the program. These are temporary in the sense that they are only local to this program, and hence can be the same as variables in the main calling program. Once defined they are referred to with the ML_eut0 program by adding the funny left single-quote mark ` and the regular right single-quote mark ’. Thus temporary variable euL, to hold the expected utility of the left lottery, is referred to as `euL’ in the program.[111]

The “quietly” line defines a block of code that is to be processed without the display of messages. This avoids needless display of warning messages, such as when some evaluation returns a missing value. Errors are not skipped, just display messages.[112]

The remaining lines should make sense to any economist from the comment statements. The program simply builds up the expected utility of each lottery, using the CRRA specification for the utility of the prizes. Then it uses the probit index function to define the likelihood values. The actual responses, stored in variable Choices (which is internal variable $ML_y1), are used at the very end to define which side of the probit index function this choice happens to be. The logit index specification is just as easy to code up: you replace “normal” with “invlogit” and you are done! The most important feature of this specification is that one can “build up” the latent index with as many programming lines as needed. Thus, as illustrated below, it is an easy matter to write out more detailed models, such as required for estimation of PT specifications or mixture models.

The “cluster(id)” command at the end tells *Stata* to treat the residuals from the same person as potentially correlated. It then corrects for this fact when calculating standard errors of estimates. Invoking the above command line, with the “maximize” option at the end to tell *Stata* to actually proceed with the optimization, generates this output:

```
initial:       log pseudolikelihood = -8155.5697
alternative:   log pseudolikelihood = -7980.4161
rescale:       log pseudolikelihood = -7980.4161
Iteration 0:   log pseudolikelihood = -7980.4161  (not concave)
Iteration 1:   log pseudolikelihood = -7692.4056
Iteration 2:   log pseudolikelihood = -7689.4848
Iteration 3:   log pseudolikelihood = -7689.4544
Iteration 4:   log pseudolikelihood = -7689.4544

. ml display
                                                  Number of obs   =      11766
                                                  Wald chi2(0)    =          .
Log pseudolikelihood = -7689.4544                 Prob > chi2     =          .

                                   (Std. Err. adjusted for 215 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .7531553   .0204812    36.77   0.000     .7130128    .7932977
------------------------------------------------------------------------------
```

So we see that the optimization routine converged nicely, with no error messages or warnings about numerical irregularities at the end. The interim warning message is nothing to worry about: only worry if there is an error message of any kind at the end of the iterations. (Of course, lots of error message, particularly about derivatives being hard to calculate, usually flag convergence problems.) The "ml display" command allows us to view the standard output, and is given after the "ml model" command. For our purposes the critical thing is the "_cons" line, which displays the ML estimate and its standard error. Thus, we have estimated that $\hat{r} = 0.753$. This is the ML CRRA coefficient in this case. This indicates that these subjects are risk averse.

Before your program runs nicely it may have some syntax errors. The easiest way to check these is to issue the command

```
ml model lf ML_eut0 (r: Choices P0left P1left P2left P3left P0right P1right P2right
       P3right prize0 prize1 prize2 prize3 stake = )
```

which is the same as before except that it drops off the material after the comma, which tells *Stata* to maximize the likelihood and how to handle the errors. This command simply tells *Stata* to read in the model and be ready to process it, but not to begin processing it. You would then issue the command

```
ml check
```

and *Stata* will provide some diagnostics. These are extremely informative if you use them, particularly for syntax errors.

The power of this approach becomes evident when we allow the CRRA coefficient to be determined by individual or treatment characteristics. To illustrate, consider the effect of allowing the CRRA coefficient to differ depending on the individual demographic characteristics of the subject, as explained in the text. Here is a list and sample statistics:

```
    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      Female |     215    .4790698    .5007276          0          1
       Black |     215    .1069767     .309805          0          1
    Hispanic |     215    .1348837    .3423965          0          1
         Age |     215    19.95814    3.495406         17         47
    Business |     215    .4511628    .4987705          0          1
      GPAlow |     215    .4604651    .4995978          0          1
```

The earlier command line is changed slightly at the "$=$)" part to read "$=$ Female Black Hispanic Age Business GPAlow)", and no changes are made to ML_eut0. The results are as follows:

```
ml model lf ML_eut0 (r: Choices P0left P1left P2left P3left P0right P1right P2right
      P3right prize0 prize1 prize2 prize3 stake = Female Black Hispanic Age Business
      GPAlow), cluster(id) maximize

. ml display
                                                  Number of obs    =      11766
                                                  Wald chi2(6)     =      27.48
Log pseudolikelihood = -7557.2809                 Prob > chi2      =     0.0001

                                    (Std. Err. adjusted for 215 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     (95% Conf. Interval)
-------------+----------------------------------------------------------------
      Female |  -.0904283   .0425979    -2.12   0.034    -.1739187   -.0069379
       Black |  -.1283174   .0765071    -1.68   0.094    -.2782686    .0216339
    Hispanic |  -.2549614   .1149935    -2.22   0.027    -.4803446   -.0295783
         Age |   .0218001   .0052261     4.17   0.000     .0115571    .0320432
    Business |  -.0071756   .0401536    -0.18   0.858    -.0858753     .071524
      GPAlow |   .0131213   .0394622     0.33   0.740    -.0642233    .0904659
       _cons |    .393472   .1114147     3.53   0.000     .1751032    .6118408
------------------------------------------------------------------------------
```

So we see that the CRRA coefficient changes from $r = 0.753$ to $r = 0.393–0.090 \times$ Female $0.128 \times$ Black … and so on. We can quickly find out what the average value of r is when we evaluate this model using the actual characteristics of each subject and the estimated coefficients:

```
. predictnl r=xb(r)
. summ r if task==1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           r |       215    .7399284    .1275521   .4333093   1.320475
```

So the average value is 0.739, extremely close to the earlier estimate of 0.753. Thus, all we have done is provided a richer characterization of risk attitudes around roughly the same mean.

F2. Loss Aversion and Probability Weighting

It is a simple matter to specify different economic models. Two of the major structural features of PT are probability weighting and loss aversion. The code below implements each of these specifications, using the parametric forms of Tversky and Kahneman (1992). For simplicity we assume that the decision weights are the probability weights, and do not implement the rank-dependent transformation of probability weights into decision weights. Thus, the model is strictly an implementation of OPT from Kahneman and Tversky (1979). The extension to rank-dependent decision weights is messy from a programming perspective, and nothing is gained pedagogically here by showing it; Harrison (2006c) shows the mess in full. Note how much of this code is similar to ML_eut0, and the differences:

```
* define OPT specification with no errors
program define MLkt0

    args lnf alpha beta lambda gamma

    tempvar prob0l prob1l prob2l prob3l prob0r prob1r prob2r prob3r y0 y1 y2 y3
    tempvar euL euR euDiff euRatio tmp

    quietly {

        gen double `tmp' = (($ML_y2^`gamma')+($ML_y3^`gamma')+($ML_y4^`gamma')+($ML_y5^`gamma'))
        replace `tmp’ = `tmp’^(1/`gamma')
        generate double `prob0l' = ($ML_y2^`gamma')/`tmp'
        generate double `prob1l' = ($ML_y3^`gamma')/`tmp'
        generate double `prob2l' = ($ML_y4^`gamma')/`tmp'
        generate double `prob3l' = ($ML_y5^`gamma')/`tmp'

        replace `tmp' = (($ML_y6^`gamma')+($ML_y7^`gamma')+($ML_y8^`gamma')+($ML_y9^`gamma'))
        replace `tmp' = `tmp’^(1/`gamma')
        generate double `prob0r' = ($ML_y6^`gamma')/`tmp'
        generate double `prob1r' = ($ML_y7^`gamma')/`tmp'
        generate double `prob2r' = ($ML_y8^`gamma')/`tmp'
        generate double `prob3r' = ($ML_y9^`gamma')/`tmp'

        generate double `y0' = .
        replace `y0' =           ( $ML_y10)^(`alpha') if $ML_y10>=0
        replace `y0' = -`lambda'*(-$ML_y10)^(`beta')  if $ML_y10<0

        generate double `y1' = .
        replace `y1' =           ( $ML_y11)^(`alpha') if $ML_y11>=0
        replace `y1' = -`lambda'*(-$ML_y11)^(`beta')  if $ML_y11<0

        generate double `y2' = .
        replace `y2' =           ( $ML_y12)^(`alpha') if $ML_y12>=0
        replace `y2' = -`lambda'*(-$ML_y12)^(`beta')  if $ML_y12<0

        generate double `y3' = .
        replace `y3' =           ( $ML_y13)^(`alpha') if $ML_y13>=0
        replace `y3' = -`lambda'*(-$ML_y13)^(`beta')  if $ML_y13<0

        gen double `euL'=(`prob0l'*`y0')+(`prob1l'*`y1')+(`prob2l'*`y2')+(`prob3l'*`y3')
        gen double `euR'=(`prob0r'*`y0')+(`prob1r'*`y1')+(`prob2r'*`y2')+(`prob3r'*`y3')

        generate double `euDiff' = `euR' - `euL'

        replace `lnf' = ln(normal( `euDiff')) if $ML_y1==1
        replace `lnf' = ln(normal(-`euDiff')) if $ML_y1==0

    }
end
```

The first thing to notice is that the initial line “`args lnf alpha beta lambda gamma`” has more parameters than with ML_eut0. The “lnf” parameter is the same, since it is the one used to return the value of the likelihood function for trial values of the other parameters. But we now have four parameters instead of just one.

When we estimate this model we get this output:

```
. ml model lf MLkt0 (alpha: Choices P0left P1left P2left P3left P0right P1right
P2right P3right prize0 prize1 prize2 prize3 = ) (beta: ) (lambda: ) (gamma: ),
cluster(id ) maximize

ml display
                                                Number of obs   =      11766
                                                Wald chi2(0)    =          .
Log pseudolikelihood = -7455.1001               Prob > chi2     =          .

                                   (Std. Err. adjusted for 215 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     (95% Conf. Interval)
-------------+----------------------------------------------------------------
alpha        |
       _cons |   .6551177   .0275903    23.74   0.000     .6010417    .7091938
-------------+----------------------------------------------------------------
beta         |
       _cons |   .8276235   .0541717    15.28   0.000      .721449     .933798
-------------+----------------------------------------------------------------
lambda       |
       _cons |   .7322427   .1163792     6.29   0.000     .5041436    .9603417
-------------+----------------------------------------------------------------
gamma        |
       _cons |    .938848   .0339912    27.62   0.000     .8722265     1.00547
------------------------------------------------------------------------------
```

So we get estimates for all four parameters. *Stata* used the variable "_cons" for the constant, and since there are no characteristics here, that is the only variable to be estimated. We could also add demographic or other characteristics to any or all of these four parameters. We see that the utility curvature coefficients α and β are similar, and indicate concavity in the gain domain and convexity in the loss domain. The loss aversion parameter λ is less than 1, which is a blow for PT since "loss aversion" calls for $\lambda > 1$. And γ is very close to 1, which is the value that implies that $w(p) = p$ for all p, the EUT case. We can readily test some of these hypotheses:

```
. test [alpha]_cons=[beta]_cons

 ( 1)  [alpha]_cons - [beta]_cons = 0

           chi2(  1) =    8.59
         Prob > chi2 =    0.0034

. test [lambda]_cons=1

 ( 1)  [lambda]_cons = 1

           chi2(  1) =    5.29
         Prob > chi2 =    0.0214

. test [gamma]_cons=1

 ( 1)  [gamma]_cons = 1

           chi2(  1) =    3.24
         Prob > chi2 =    0.0720
```

So we see that PT is not doing so well here in relation to the *a priori* beliefs it comes packaged with, and that the deviation in λ is indeed statistically significant. But γ is less than 1, so things are not so bad in that respect.

F3. Adding Stochastic Errors

In the text the Luce and Fechner "stochastic error stories" were explained. To add the Luce specification, popularized by HL, we return to base camp, the ML_eut0 program, and simply make two changes. We augment the arguments by one parameter, μ, to be estimated:

```
args lnf r mu
```

and then we revise the line defining the EU difference from

```
generate double `euDiff' = `euR' - `euL'
```

to

```
generate double `euDiff' = (`euR'^(1/`mu'))/((`euR'^(1/`mu'))
                          +(`euL'^(1/`mu')))
```

So this changes the latent preference index from being the difference to the ratio. But it also adds the $1/\mu$ exponent to each expected utility. Apart from this change in the program, there is nothing extra that is needed. You just add one more parameter in the "ml model" stage, as we did for the PT extensions. In fact, HL cleverly exploit the fact that the latent preference index defined above is already in the form of a cumulative probability density function, since it ranges from 0 to 1, and is equal to 1/2 when the subject is indifferent between the two lotteries. Thus, instead of defining the likelihood contribution by

```
replace `lnf' = ln(normal( `euDiff')) if $ML_y1==1
replace `lnf' = ln(normal(-`euDiff')) if $ML_y1==0
```

we can use

```
replace `lnf' = ln(`euDiff') if $ML_y1==1
replace `lnf' = ln(1-`euDiff') if $ML_y1==0
```

instead.

The Fechner specification popularized by Hey and Orme (1994) implies a simple change to ML_eut0. Again we add an error term "noise" to the arguments of the program, as above, and now we have the latent index

```
generate double `euDiff' = (`euR' - `euL')/`noise'
```

instead of the original

```
generate double `euDiff' = `euR' - `euL'
```

Here are the results:

```
. ml model lf ML_eut (r: Choices P0left P1left P2left P3left P0right P1right P2right
P3right prize0 prize1 prize2 prize3 stake = ) (noise: ), cluster(id) maximize

. ml display
                                                  Number of obs   =      11766
                                                  Wald chi2(0)    =          .
Log pseudolikelihood = -7679.9527                 Prob > chi2     =          .

                                   (Std. Err. adjusted for 215 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r            |
       _cons |   .7119379   .0303941    23.42   0.000     .6523666    .7715092
-------------+----------------------------------------------------------------
noise        |
       _cons |   .7628203    .080064     9.53   0.000     .6058977    .9197429
------------------------------------------------------------------------------
```

So the CRRA coefficient declines very slight, and the noise term is estimated as a normal probability with standard deviation of 0.763.

F4. Non-Parametric Estimation of the EUT Model

It is possible to estimate the EUT model without assuming a functional form for utility, following Hey and Orme (1994). The likelihood function is evaluated as follows:

```
* define Original Recipe EUT with Fechner errors: non-parametric
program define ML_eut0_np

    args lnf u5 u10 noise

    tempvar prob0l prob1l prob2l prob3l prob0r prob1r prob2r prob3r y0 y1 y2 y3
    tempvar euL euR euDiff euRatio tmp lnf_eut lnf_pt p1 p2 f1 f2 u0 u15

    quietly {

        * construct likelihood for EUT
        generate double `prob0l' = $ML_y2
        generate double `prob1l' = $ML_y3
        generate double `prob2l' = $ML_y4
        generate double `prob3l' = $ML_y5

        generate double `prob0r' = $ML_y6
        generate double `prob1r' = $ML_y7
        generate double `prob2r' = $ML_y8
        generate double `prob3r' = $ML_y9

        generate double `u0'  = 0
        generate double `u15' = 1

        generate double `y0' = `u0'
        generate double `y1' = `u5'
        generate double `y2' = `u10'
        generate double `y3' = `u15'

        gen double `euL'=(`prob0l'*`y0')+(`prob1l'*`y1')+(`prob2l'*`y2')+(`prob3l'*`y3')
        gen double `euR'=(`prob0r'*`y0')+(`prob1r'*`y1')+(`prob2r'*`y2')+(`prob3r'*`y3')

        generate double `euDiff' = (`euR' - `euL')/`noise'

        replace `lnf' = ln(normal( `euDiff')) if $ML_y1==1
        replace `lnf' = ln(normal(-`euDiff')) if $ML_y1==0

    }
end
```

and estimates can be obtained in the usual manner. We include demographics for each parameter, and introduce the notion of a "global" macro function in *Stata*. Instead of typing out the list of demographic variables, one gives the command

```
global demog "Female Black Hispanic Age Business GPAlow"
```

and then simply refer to $global. Every time *Stata* sees "$demog" it simply substitutes the string "Female Black Hispanic Age Business GPAlow" without the quotes. Hence, we have the following results:

```
. ml model lf ML_eut0_np (u5: Choices P0left P1left P2left P3left P0right P1right
P2right P3right prize0 prize1 prize2 prize3 stake = $demog ) (u10: $demog ) (noise: )
if expid=="ucf0", cluster(id) technique(dfp) maximize difficult

. ml display
                                                  Number of obs   =       3736
                                                  Wald chi2(6)    =      18.19
Log pseudolikelihood = -2321.8966                 Prob > chi2     =     0.0058

                                    (Std. Err. adjusted for 63 clusters in id)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     (95% Conf. Interval)
-------------+----------------------------------------------------------------
u5           |
      Female |    .096698   .0453102     2.13   0.033     .0078916    .1855044
       Black |   .0209427   .0808325     0.26   0.796    -.1374861    .1793715
    Hispanic |   .0655292   .0784451     0.84   0.404    -.0882203    .2192787
         Age |  -.0270362   .0093295    -2.90   0.004    -.0453217   -.0087508
    Business |   .0234831   .0493705     0.48   0.634    -.0732813    .1202475
      GPAlow |  -.0101648   .0480595    -0.21   0.832    -.1043597    .0840301
       _cons |   1.065798   .1853812     5.75   0.000     .7024573    1.429138
-------------+----------------------------------------------------------------
u10          |
      Female |   .0336875   .0287811     1.17   0.242    -.0227224    .0900973
       Black |   .0204992   .0557963     0.37   0.713    -.0888596    .1298579
    Hispanic |   .0627681   .0413216     1.52   0.129    -.0182209     .143757
         Age |  -.0185383   .0072704    -2.55   0.011     -.032788   -.0042886
    Business |   .0172999   .0308531     0.56   0.575    -.0431711    .0777708
      GPAlow |  -.0110738   .0304819    -0.36   0.716    -.0708171    .0486696
       _cons |   1.131618   .1400619     8.08   0.000     .8571015    1.406134
-------------+----------------------------------------------------------------
noise        |
       _cons |   .0952326   .0079348    12.00   0.000     .0796807    .1107844
------------------------------------------------------------------------------
```

It is then possible to predict the values of the two estimated utilities, which will vary with the characteristics of each subject, and plot them. Fig. 10 in the text shows the distributions of estimated utility values.

F5. Replication of Holt and Laury (2002)

Finally, it may be useful to show an implementation in *Stata* of the ML problem solved by HL:

```
program define HLep1

        args lnf r alpha mu

        tempvar theta lnfj prob1 prob2 scale euSAFE euRISKY euRatio
               mA1 mA2 mB1 mB2 yA1 yA2 yB1 yB2 wp1 wp2

        quietly {


         /* initializations */
         generate double `prob1' = $ML_y2/10
         generate double `prob2' = 1 - `prob1'
         generate double `scale' = $ML_y7

         /* add the endowments to the prizes */
         generate double `mA1' = $ML_y8 + $ML_y3
         generate double `mA2' = $ML_y8 + $ML_y4
         generate double `mB1' = $ML_y8 + $ML_y5
         generate double `mB2' = $ML_y8 + $ML_y6

        /* utility of prize m */
         generate double `yA1' = (1-exp(-`alpha'*((`scale'*`mA1')^(1-`r'))))/`alpha'
         generate double `yA2' = (1-exp(-`alpha'*((`scale'*`mA2')^(1-`r'))))/`alpha'
         generate double `yB1' = (1-exp(-`alpha'*((`scale'*`mB1')^(1-`r'))))/`alpha'
         generate double `yB2' = (1-exp(-`alpha'*((`scale'*`mB2')^(1-`r'))))/`alpha'

        /* classic EUT probability weighting function */
         generate double `wp1' = `prob1'
           generate double `wp2' = `prob2'

        /* expected utility */
         generate double `euSAFE' = (`wp1'*`yA1')+(`wp2'*`yB1')
           generate double `euRISKY' = (`wp1'*`yA2')+(`wp2'*`yB2')

        /* EU ratio */
         generate double `euRatio' = (`euSAFE'^(1/`mu'))/
                                    ((`euSAFE'^(1/`mu'))+(`euRISKY'^(1/`mu')))

        /* contribution to likelihood */
         replace `lnf' = ln(`euRatio')   if $ML_y1==0
         replace `lnf' = ln(1-`euRatio') if $ML_y1==1

    }
end
```

The general structure of this routine should be easy to see. The routine is called with this command `ml model lf HLep1 (r: Choices problem m1a m2a m1b m2b scale wealth = ) (alpha: ) (mu: )`where variable “Choices” is a binary variable defining the subject’s choices of the safe or risky lottery; variable “problem” is a counter from 1 to 10 in the usual implementation of the design; the next four variables define the fixed prizes; variable “scale” indicates the multiples of the basic payoffs used (e.g., 1, 10, 20, 50, or 90), and variable “wealth” measures initial endowments prior to the risk aversion task (typically $0). Three parameters are estimated, as defined in the EP specification discussed in the text. The only new steps are the definition of the utility of the prize, using the EP specification instead of the CRRA specification, and the definition of the index of the likelihood.

Use of this procedure with the original HL data replicates the estimates in Holt and Laury (2002, p. 1653) exactly. The advantage of this formulation is that one can readily extend it to include covariates for any of the parameters. One can also correct for clustering of observations by the same subject. And extensions to consider probability weighting are trivial to add.

F6. Extensions

There are many possible extensions of the basic programming elements considered here. Harrison (2006c) illustrates the following:

- modeling rank-dependent decision weights for the RDU and RDEV structural model;
- modeling rank-dependent decision weights and sign-dependent utility for the CPT structural model;
- the imposition of constraints on parameters to ensure non-negativity (e.g., $\lambda > 1$ or $\mu > 0$), or finite bounds (e.g., $0 < r < 1$);
- the specification of finite mixture models;
- the coding of non-nested hypothesis tests; and
- maximum simulated likelihood, in which one or more parameters are treated as random coefficients to reflect unobserved individual heterogeneity (e.g., Train (2003)).

In each case template code is provided along with data and illustrative estimates.