# The Independence Axiom and the Bipolar Behaviorist

by

Glenn W. Harrison and J. Todd Swarthout [†]

January 2012

ABSTRACT.

Developments in the theory of risk require yet another evaluation of the behavioral validity of the independence axiom. This axiom plays a central role in most formal statements of expected utility theory, as well as popular alternative models of decision-making under risk, such as rank-dependent utility theory. It also plays a central role in experiments used to characterize the way in which risk preferences deviate from expected utility theory. If someone claims that individuals behave as if they "probability weight" outcomes, and hence *violate* the independence axiom, it is invariably on the basis of experiments that must *assume* the independence axiom. We refer to this as the Bipolar Behavioral Hypothesis: behavioral economists are pessimistic about the axiom when it comes to characterizing how individuals directly evaluate two lotteries in a binary choice task, but are optimistic about the axiom when it comes to characterizing how individuals evaluate multiple lotteries that make up the incentive structure for a multiple-task experiment. Building on designs that have a long tradition in experimental economics, we offer direct tests of the axiom and the evidence for probability weighting. We reject the Bipolar Behavioral Hypothesis: we find that non-parametric preferences estimated for the rank-dependent utility model are significantly affected when one elicits choices with *procedures* that require the independence assumption, as compared to choices with *procedures* that do not require that assumption. We also demonstrate this result with familiar parametric preference specifications, and draw general implications for the empirical evaluation of theories about risk.

[†] Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison); and Department of Economics, Andrew Young School of Policy Studies, Georgia State University, USA (Swarthout). E-mail contacts: gharrison@gsu.edu and swarthout@gsu.edu. We are grateful to Jim Cox, Jimmy Martínez, John Quiggin, Vjollca Sadiraj, Ulrich Schmidt, Uzi Segal and Nathaniel Wilcox for helpful discussions.

## Table of Contents

Developments in the theory of risk require yet another an evaluation of the behavioral validity of the independence axiom. This axiom plays a central role in most formal statements of expected utility theory (EUT), as well as popular alternative models of decision-making under risk, such as rank-dependent utility (RDU) theory. It also plays a central role in most experiments used to characterize the way in which risk preferences deviate from EUT. For example, if someone claims that individuals behave as if they "probability weight" outcomes, and hence *violate* the independence axiom (IA), it is usually on the basis of experiments that must *assume* the IA if the incentives are to be taken seriously. But there is an obvious inconsistency with saying that individuals behave as if they violate the IA on the basis of evidence collected under the maintained assumption that the axiom is magically valid.

This inconsistency has long been noted in the literature, with some ingenious experimental designs intended to trap the IA under some circumstances. But these indirect tests of the IA have been inconclusive. This is frustrating: either the axiom applies or it does not. The uneasy state of the literature has evolved to assuming the axiom for the purposes of making the payment protocol of an experiment valid, but not assuming it when characterizing the risk preferences exhibited in the same experiment. Those characterizations seem to show evidence of rank-dependent probability weighting, when that very evidence calls into question the maintained assumption of the payment protocol used to generate the evidence. We refer to this as the Bipolar Behavioral Hypothesis: behavioral economists are pessimistic about the IA when it comes to characterizing how individuals directly evaluate two lotteries in a binary choice task, but are optimistic about the IA when characterizing how individuals evaluate multiple lotteries that make up the incentive structure for a multiple-task experiment.

The standard payment protocol involves a subject making K>1 choices, and then selecting one choice at random for payment. We call this protocol 1-in-K. Following Conlisk [1989], Starmer and Sugden [1991], Beattie and Loomes [1997], Cubitt, Starmer and Sugden [1998] and Cox, Sadiraj and Schmidt [2011], an alternative payment protocol, which we call 1-in-1, involves a subject making only one choice, and then being paid with certainty for the single choice.[1] The IA can have no role to play in the validity of the 1-in-1 protocol *per se* if we restrict choice to simple lotteries, but plays a defining role in the 1-in-K protocol. And the role that the IA plays in the theoretical and behavioral validity of the experimental payment protocol is quite distinct from the role that it might play in evaluating the actual binary choice or choices. Even with the 1-in-1 protocol being used, it is possible to ask if behavior is better characterized by violations of IA or not. Indeed, the whole point of our design is to highlight the dual role of the IA in 1-in-K protocols that seek to test violations of IA.

We offer direct tests of the effect of IA on preferences for risk in general, and the evidence for probability weighting in particular, by using both of these payment protocols. We reject the Bipolar Behavioral Hypothesis. We find evidence of RDU probability weighting with the 1-in-1 protocol that does *not* rely on the validity of the IA. So this result establishes that there is theoretical and behavioral "cause for concern" when one assumes the validity of the IA for the 1-in-K protocol. We then find that this theoretical concern is empirically relevant. Estimated RDU risk preferences *are different* depending on whether one infers them from data collected with the 1-in-1 payment protocol or the 1-in-K payment protocol.

Many studies invoke something referred to as "the isolation effect," which is often a *behavioral* assertion that a subject views each choice in an experiment as independent of other choices in the experiment. When used formally, this hypothesis is usually the same as the IA, and is indeed

[1] Conlisk [1989; p.406] has a very clear statement of the problem, and the need for the 1-in-1 protocol. He uses the 1-in-1 protocol in his test of the Allais Paradox, incidentally finding no evidence whatsoever for the alleged anomaly, but does not test it behaviorally against the 1-in-K protocol. Starmer and Sugden [1991] were the first to undertake that behavioral comparison.

exactly the same as the IA in our choice context. We do recognize that it is often invoked informally as "an empirical matter," much as a magic talisman is used to ward off evil spirits.

In section 1 we describe the theoretical constructs needed for our design, in particular the various axioms that are at issue. In section 2 we present our experimental design, which allows comparison of risk preferences obtained from tasks that do not require the IA with risk preferences obtained from tasks that do require the assumption. We also explain why we focus on differences in estimated preferences across treatments rather than just examine raw choice patterns. In section 3 we develop the econometric model used to estimate preferences. We pay particular attention to the manner in which between-subject heterogeneity is modeled. The reason for this attention is that the simplest way of avoiding reliance on the IA is to give some individuals one task, *necessitating* the pooling of choices across individuals. In the absence of an assumption of homogeneity of risk preferences, or samples of sufficient power to allow randomization to mitigate the need for that assumption, we must address the econometric modeling of heterogeneity. In section 4 we examine the data from our experiments and econometric analysis. In section 5 we draw some general implications of our results, and in section 6 draw some general conclusions. Appendices A and B document the parameters and instructions used in our experiments, and appendix C reviews the previous experimental literature.

## 1. Theory

### *A. Basic Axioms*

Following Segal [1988][1990][1992], we distinguish between three axioms. In words, the **Reduction of Compound Lotteries** (ROCL) axiom states that a decision-maker is indifferent between a compound lottery and the actuarially-equivalent simple lottery in which the probabilities of the two stages of the compound lottery have been multiplied out. To use the language of

Samuelson [1952; p.671], the former generates a *compound income-probability-situation*, and the latter defines an *associated income-probability-situation*, and that "...only algebra, not human behavior, is involved in this definition."

To state this more explicitly, with notation to be used to state all axioms, let X, Y and Z denote simple lotteries, A and B denote compound lotteries, $\succ$ express strict preference, and ~ express indifference. Then the ROCL axiom says that A ~ X if the probabilities and prizes in X are the actuarially-equivalent probabilities and prizes from A. Thus if A is the compound lottery that pays "double or nothing" from the outcome of the lottery that pays \$10 if a coin flip is a head and \$2 if the coin flip is a tail, then X would be the lottery that pays \$20 with probability ½×½ = ¼, \$4 with probability ½×½ = ¼, and nothing with probability ½. From an observational perspective, one would have to see choices between compound lotteries and the actuarially-equivalent simple lottery to test ROCL.

The **Compound Independence Axiom** (CIA) states that a compound lottery formed from two simple lotteries by adding a positive common lottery with the same probability to each of the simple lotteries will exhibit the same preference ordering as the simple lotteries. So this is a statement that the preference ordering of the two constructed compound lotteries will be the same as the preference ordering of the different simple lotteries that distinguish the compound lotteries, provided that the common prize in the compound lotteries is the same and has the same (compound lottery) probability. It says nothing about how the compound lotteries are to be evaluated, and in particular *it does not assume ROCL*. It only restricts the preference *ordering* of the two constructed compound lotteries to match the preference *ordering* of the original simple lotteries.

The CIA says that if A is the compound lottery giving the simple lottery X with probability α and the simple lottery Z with probability (1-α), and B is the compound lottery giving the simple lottery Y with probability α and the simple lottery Z with probability (1-α), then A $\succ$ B iff X $\succ$ Y ∀ α

$\in$ (0,1). So the construction of the two compound lotteries A and B has the "independence axiom" cadence of the common prize Z with a common probability (1-$\alpha$), but the implication is only that the *ordering* of the compound and constituent simple lotteries are the same.[2]

Finally, the **Mixture Independence Axiom** (MIA) says that the preference ordering of two simple lotteries must be the same as the actuarially-equivalent simple lottery formed by adding a common outcome in a compound lottery of each of the simple lotteries, where the common outcome has the same value and the same (compound lottery) probability. So stated, it is clear that the MIA strengthens the CIA by making a definite statement that the constructed compound lotteries are to be evaluated in a way that is ROCL-consistent. Construction of the compound lottery in the MIA is actually implicit: the axiom only makes observable statements about two pairs of simple lotteries. To restate Samuelson's point about the definition of ROCL, the experimenter testing the MIA could have constructed the associated income-probability-situation without knowing the risk preferences of the individual (although the experimenter would need to know how to multiply).

The MIA says that X $\succ$ Y iff the actuarially-equivalent simple lottery of $\alpha$X + (1-$\alpha$)Z is strictly preferred to the actuarially-equivalent simple lottery of $\alpha$Y + (1-$\alpha$)Z, $\forall$ $\alpha \in$ (0,1). The verbose language used to state the axiom makes it clear that MIA embeds ROCL into the usual independence axiom construction with a common prize Z and a common probability (1-$\alpha$) for that prize.

The reason these three axioms are important is that the failure of MIA does not imply the failure of CIA and ROCL. It does imply the failure of one or the other, but it is far from obvious

[2] For example, Segal [1992; p.170] defines the CIA by assuming that the second-stage lotteries are replaced by their certainty-equivalent, "throwing away" information about the second-stage probabilities before one examines the first-stage probabilities at all. Hence one cannot then define the actuarially-equivalent simple lottery, by construction, since the informational bridge to that calculation has been burnt.

which one. Indeed, one could imagine some individuals or task domains where only CIA might fail, only ROCL might fail, or both might fail. Moreover, specific types of failures of ROCL lie at the heart of many important models of decision-making under uncertainty and ambiguity. We use the acronym IA when we mean "CIA or MIA" and the acronyms CIA or MIA directly when the difference matters.

### *B. Experimental Payment Protocols*

Turning now to experimental procedures, as a matter of theory the most popular payment protocol assumes the validity of the CIA. This payment protocol is called the **Random Lottery Incentive Mechanism** (RLIM). It entails the subject undertaking K tasks and then one of the K choices being selected at random to be played out. Typically, and without loss of generality, assume that the selection of the k$^{th}$ task to be played out uses a uniform distribution over the K tasks. Since the other K-1 tasks will generate a payoff of zero, the payment protocol can be seen as a compound lottery that assigns probability $\alpha = 1/k$ to the selected task and $(1-\alpha) = (1-(1/k))$ to the other K-1 tasks as a whole. If the task consists of binary choices between simple lotteries X and Y, then the RLIM can be immediately seen to entail an application of the CIA, where Z = U($0) and $(1-\alpha) = (1-(1/k))$, for the utility function U(·). Hence, under the CIA, the preference ordering of X and Y is independent of all of the choices in the other tasks (Holt [1986]).

If the K objects of choice include any compound lotteries, directly or indirectly, then one might naturally think of the RLIM as requiring the stronger MIA instead of just the CIA. Indeed, this was the setting for the classic discussions of the interaction of the IA with the RLIM, the commentaries of Holt [1986], Karni and Safra [1987] and Segal [1988] on the "preference reversal" findings of Grether and Plott [1979]. In those experiments the elicitation procedure for the certainty-equivalents of simple lotteries was, itself, a compound lottery. Hence the validity of the incentives for this design required both CIA and ROCL, hence MIA. Holt [1986] and Karni and Safra [1987] showed that if CIA was violated, but ROCL and transitivity was assumed, one might still observe choices that suggest "preference reversals." Segal [1988] showed that if ROCL was violated, but CIA and transitivity was assumed, that one might also still observe choices that suggest "preference reversals."[3] Again, the only reason that ROCL was implicated in these discussions is because the experimental task implicitly included choices over compound lotteries. In our case we only consider choices over simple lotteries, so the validity of RLIM rests solely on the validity of the CIA.

The CIA can be avoided by setting K=1, and asking each subject to answer one binary choice task for payment. Unfortunately, this comes at the cost of another assumption if one wants to compare choice patterns over two simple lottery pairs, as in most of the popular tests of EUT such as the Allais Paradox and Common Ratio test: the assumption that risk preferences across subjects are the same. This is a strong assumption, obviously, and one that leads to inferential tradeoffs in terms of the "power" of tests of EUT relying on randomization that will vary with sample size. Sadly, plausible estimates of the degree of heterogeneity in the typical population imply massive sample sizes for reasonable power, well beyond those of most experiments.

The assumption of homogeneous preferences can be diluted, however, by changing it to a conditional form: that risk preferences are homogeneous conditional on a finite set of observable characteristics.[4] Although this sounds like an econometric assumption, and it certainly has statistical implications, it is as much a matter of theory as formal statements of the CIA, ROCL and MIA.

[3] Guala [2005; p.97ff] contains an excellent discussion of these issues surrounding the "preference reversal" debates.

[4] Another way of diluting the assumption is to posit some (flexible) parametric form for the distribution of risk attitudes in the population, and use econometric methods that allow one to estimate the extent of that unobserved heterogeneity across individuals. Tools for this "random coefficients" approach to estimating non-linear preference functionals are developed in Andersen, Harrison, Hole, Lau and Rutström [2010].

## 2. Experiment

### *A. Basic Design Issues*

Our basic experimental design focuses directly on the risk preferences that one can infer from binary choices over pairs of simple lotteries. This task is canonical, in terms of testing EUT against alternatives such as RDU, as well as for estimating risk preferences. Our design builds on a comparison of the risk preferences implied by 1-in-1 and 1-in-K choice tasks. We let K equal 30, to match the typical risky choice experiment in which there are many choices (e.g., Hey and Orme [1994]). Figure 1 shows the interface given to our subjects in this case of sequential presentation of choice tasks. A standard, fixed show-up fee, in our case $7.50, was paid to every subject independently of their lottery choices.

An important dimension of choice tasks, for K>1, is whether the individual gets to see the lotteries prior to making any choices. Again, the typical case in the experimental literature is when the choices are presented sequentially. Although there is often some similarity in prizes and probabilities from choice task to choice task, the subject does not know the exact lotteries to come, and that can make the task of forming portfolios very demanding (Hey and Lee [2005a][2005b]). On the other hand, presenting subjects with a "multiple price list" of ordered lottery choices is a justifiably popular task for eliciting risk attitudes (Holt and Laury [2002][2005], Harrison, Johnson, McInnes and Rutström [2005] and Andersen, Harrison, Lau and Rutström [2006]). In this case the subject sees all binary choices arrayed on one page, and is virtually encouraged to form some portfolio.[5]

It is common in many experimental settings for the individual to face one or more paid tasks after the K tasks of focus in the elicitation of risk preferences. A good example is the joint

[5] It is easy to show that first order stochastic dominance implies that K rows or binary choice tasks imply only K+1 efficient portfolios, in which there are only 0 or 1 switch points.

estimation design of Andersen, Harrison, Lau and Rutström [2008], in which subjects completed risky lottery choices designed to infer the concavity of their utility function so that inferences about discount rates defined over utility could be made from a later task involving choices over time-dated monetary amounts. Hence we also examine the effect of there being an extra task after the binary lottery choice of primary interest.

### *B. Specific Design*

Table 1 summarizes our experimental design. In **treatment A** subjects undertake 1-in-1 binary choices, where the one pair they face is drawn at random from a set of 69 lottery pairs shown in Table A1 of Appendix A. These lottery pairs span five monetary prize amounts, \$5, \$10, \$20, \$35 and \$70, and five probabilities, 0, ¼, ½, ¾ and 1. The prizes are combined in ten "contexts," defined as a particular triple of prizes.[6] They are based on a battery of lottery pairs developed by Wilcox [2010] for the purpose of robust estimation of EUT and RDU models.[7] Figure 2 shows the coverage of these lottery pairs in terms of the Marschak-Machina triangle. Each prize context defines a different triangle, but the patterns of choice overlap considerably. Figure 2 shows that there are many lottery pair chords that assume parallel indifference curves, as expected under EUT, but that the slope of the indifference curve can vary, so that the tests of EUT have reasonable power for a wide range of risk attitudes under the EUT null hypothesis (Loomes and Sugden [1998] and Harrison, Johnson, McInnes and Rutström [2007]). These lotteries also contain a number of pairs in

[6] For example, the first context consists of lotteries defined over the prizes \$5, \$10 and \$20, and the tenth context consists of lotteries defined over the prizes \$20, \$35 and \$70. The significance of the prize context is explained by Wilcox [2010][2011].

[7] The original battery includes repetition of some choices, to help identify the "error rate" and hence the behavioral error parameter, defined later. In addition, the original battery was designed to be administered in its entirety to every subject. We decided *a priori* that 30 choice tasks was the maximum that our subject pool could focus on in any one session, given the need in some sessions for there to be later tasks.

which the "EUT-safe" lottery has a *higher* EV than the "EUT-risky" lottery: this is designed deliberately to evaluate the extent of risk premia deriving from probability pessimism rather than diminishing marginal utility.

In treatment A we do *not* have to assume the CIA in order for observed choices to reflect risk preferences under EUT or RDU. In effect, it represents the behavioral Gold Standard benchmark, against which the other payment protocols are to be evaluated.

In **treatment B** we move to the 1-in-30 case, which is typical of the usual risk elicitation setting. In all cases, unless otherwise noted, we explicitly told subjects that there were no further salient tasks affecting their earnings after the risky lottery task, to avoid them even tacitly thinking of forming a portfolio over the risky lottery tasks and any future tasks.

**Treatment C** extends the 1-in-30 case to the most common in the experimental literature, where the risky lottery choice task is followed by some other paid task. Payments for the lottery choices are not affected by payments for the other task, but the prospect of another paid task might encourage subjects to form some sort of "anticipated portfolio." The instructions in treatment C raised the possibility of a future task for payment, but the instructions in treatments A, B, and D clearly stated that there would be no further paid task.[8] Common practice and expectation in our lab might have led subjects to expect multiple tasks, and that could obviously vary with the experiences of each subject.

Every random event determining payouts was generated by the rolling of one or more dice.

[8] To be precise, at the end of the instructions for treatment C subjects were told that "All payoffs are in cash, and are in addition to the $7.50 show-up fee that you receive just for being here, as well as any other earnings in other tasks." In the other treatments subjects were told that "All payoffs are in cash, and are in addition to the $7.50 show-up fee that you receive just for being here. The only other task today is for you to answer some demographic questions. Your answers to those questions will not affect your payoffs."

These dice were illustrated visually during the reading of the instructions,[9] and each subject rolled their own dice.

### *C. Why Not Just Look At Raw Choice Patterns?*

We focus on the risk preferences implied by the observed choice data, and do not examine the choice patterns themselves. The reason is that there are limits on what can be inferred by just looking at choice patterns. Since much of the literature on the evaluation of the axioms of EUT has done precisely that, we explain why we believe this to be less informative than trying to make inferences about the underlying latent preferences. This may be particularly important because many might wonder how they *could* differ: after all, if preferences are just rationalizing observed choices, and if observed choices appear to violate the predictions of EUT or IA, how can it be that the implied preferences might not?

#### Behavioral Errors

In an important sense, our task would be easier if humans never made mistakes. This would allow us to test deterministic theories of choice, and *any* deviation from the predictions of the theory would provide *prima facie* evidence of a failure of the theory. However, humans do make errors in behavior, and so our task is more complex. The canonical evidence for behavioral errors is the fraction of "switching behavior" observed when subjects are given literally the same lottery pair at different points in a session (e.g., Wilcox [1993]). Any analysis of individual choice ought to account for such behavioral errors. Indeed, some previous analyses of choice patterns have attempted to

[9] The lab contains a video projector from the front table to the displays throughout the room. Apart from a large front-screen display, there are 3 wide-screen TV displays throughout the lab so that every cubicle has a clear view.

account for "mistakes" by implementing "trembles" (e.g., Conlisk [1989; Appendix I] and Harless and Camerer [1994]). Such trembles are agnostic about the way any behavioral error might affect the latent components of the choice. A more satisfactory approach would incorporate behavioral errors into the choice process in a more coherent manner, as discussed in detail by Wilcox [2008].

It is worth emphasizing that behavioral errors are quite distinct conceptually from sampling errors. The former refer to some latent component of the theoretical structure generating a predicted choice. The latter refer to the properties of an estimate of the parameters of that theoretical structure. To see the difference, and assuming a consistent estimator, if the sample size gets larger and larger the sampling errors must get smaller and smaller, but the (point estimate of the) behavioral error need not.[10] In the first instance behavioral errors are the business of theorists, not econometricians.[11]

<u>Do Choice Patterns Use All Available Information?</u>

Once we recognize that there can be some imprecision in the manner in which preferences translate into observed choices, we obtain another informational advantage from making inferences about preferences estimated from a structural model: a theory about how the intensity of a preference for one lottery over another matters. For any given utility function and set of parameter values, and assuming EUT for exposition, a larger difference in the EU of two lotteries matters more for the likelihood of the presumed preferences than a difference in the EU that is close to

[10] An additional complication arises if one posits random coefficients. In this case, the estimates for any structural parameter, such as the behavioral error parameter, will have a distribution that characterizes the population. If that population distribution is assumed to be Gaussian, as is often the case, there will be a point estimate and standard error estimate of the population mean, and a point estimate and standard *error* estimate of the population standard *deviation*. With a consistent estimator, increased sample sizes imply that both standard *error* estimates will decrease, but the point estimate of the population standard *deviation* need not.

[11] Of course they interact, as stressed by Wilcox [2008][2011].

zero. To see this, assume some parameter values characterizing preferences, and two lottery pairs. One lottery pair, evaluated at those parameter values, implies an EU for the left lottery that is ε greater than the EU for the right lottery. Another lottery pair, similarly evaluated at those parameter values, implies an EU for the left lottery that is much greater than ε more than the EU for the right lottery. An observed choice that is inconsistent with the predicted choice for the second lottery pair matters more for the validity of the assumed parameter values than an inconsistent observed choice for the first lottery pair. This is not the case when one simply looks at the number of consistent and inconsistent choice pairs, as all inconsistent choices are treated as informationally equivalent.

Of course, one has to define the term "intensity" for a given utility representation, and there are theoretical and econometric subtleties involved in normalizing EU differences over different choice contexts, discussed later and in Wilcox [2008][2011]. And structural estimation does entail some parametric assumptions, also discussed later, that are not involved with the usual analysis of choice patterns. But there is simply more information used when one evaluates estimated preferences with a structural model. The difference is akin to limited-information inference versus full-information inference in statistics: *ceteris paribus*, it is always better to use more information than less. Now we admit immediately that things are not all equal, and that *some* parametric assumptions will be needed to undertake what we call the full-information approach here.[12] But we do argue that the preference estimation approach is complementary to studying choice patterns, and not an inferior and less direct method of conducting the same analysis.

### Are the Stimuli Representative?

Comparison of choice patterns from a paradox test with two pairs of lotteries may support

[12] We will see that the parametric assumptions can be a lot fewer than one usually makes.

or refute the theory under consideration, but how confident are we that the result is representative of choices over all lottery pairs? What if multiple tests using distinct choice *patterns* are conducted and only a single test *pattern* suggests a failure of the theory? Perhaps some theorists are content with a single case of falsification, but others may be concerned that the single failure is a rare exception. For example, it is well-known that violations of EUT tend to occur less frequently when lotteries are in the "interior" of the Marschak-Machina triangle (e.g., Starmer [2000; p.358]). Hence one might draw one negative set of qualitative conclusions about EUT from one battery of stimuli and a different, positive set of qualitative conclusions about EUT from a different battery of stimuli.[13] As a general model for all sets of stimuli, EUT is still in trouble in this case, to be sure, but inferences about the validity of EUT then need to be nuanced and conditional.

Model estimation can address this "representativeness" issue by presenting subjects with a wide range of lottery pairs, a point first stressed in the experimental economics literature by Hey and Orme [1994]. Of course, there is a tradeoff in doing this: with the 1-in-1 protocol we cannot conduct choice pattern comparisons due to low sample sizes for any given lottery pair.

### The Homogeneity Assumption

Another theoretical reason one might want to estimate a structural model of preferences, rather than examine choice data alone, is to better account for heterogeneity of preferences in the 1-in-1 treatment. The analysis of choice patterns must assume preference homogeneity, or perhaps minimally condition on a factor, such as assuming homogeneity within samples of men or women. Some might appeal to large-sample randomization in an attempt to avoid the assumption of homogeneity, but rarely does anyone conduct appropriate power analyses to justify that appeal. By

[13] For example, contrast Camerer [1989] and Camerer [1992] for an illustration of this precise issue.

using structural model estimation, observed preference heterogeneity can be ameliorated through the use of demographics controls (e.g., Harrison and Rutström [2008]), and unobserved preference heterogeneity can be ameliorated through the use of random coefficient models (e.g., Andersen, Harrison, Hole, Lau and Rutström [2010]).

*D. Data*

A total of 348 subjects were recruited to participate in experiments at Georgia State University between February 2011 and April 2011. The general recruitment message did not mention the show-up fee or any specific range of possible earnings, and subjects were undergraduate students recruited from across the campus. Table 1 shows the allocations of subjects across our main treatments. Instructions for all treatments are presented in Appendix B. Every subject received a copy of the instructions, printed in color, and had time to read them after being seated in the lab. The instructions were then projected on-screen and read out word-for-word by the same experimenter. Every subject also completed a demographic survey covering standard characteristics. All subjects were paid in cash at the end of each session.

## 3. Econometrics

Our interest is in making inferences about the latent risk preferences underlying observed choice behavior. The estimation approach is typically to write out a structural model of decision-making, assuming some functional forms if necessary. We focus initially on EUT as the appropriate null, but also consider RDU and Dual Theory models of decision-making under risk. The lottery parameters in our design also allow us to estimate the structural model assuming non-parametric specifications of the utility and probability weighting functions, and these non-parametric estimations will be the main focus of inferences whenever possible.

*A. The Basic Model*

Assume that the utility of income is defined by a completely non-parametric utility function. We exploit the fact that, by design, the lottery pairs in our experiment span only 5 monetary prize amounts, \$5, \$10, \$20, \$35 and \$70. Set the utility for the smallest prize to 0 and the utility of the largest prize to 1, and directly estimate the utility of the intermediate prizes:

$$U(\$0) = 0,\ U(\$10) = \varkappa_{10}\ ,\ U(\$20) = \varkappa_{20}\ ,\ U(\$35) = \varkappa_{35}\ ,\ U(\$70) = 1 \tag{1}$$

with the constraint that $\varkappa_{10}$ , $\varkappa_{20}$ and $\varkappa_{35}$ lie in the unit interval. This is precisely the approach employed by Hey and Orme [1994] and Wilcox [2010].

Let there be J possible outcomes in a lottery. The probability $p(M_j)$ of each outcome $M_j$ is induced by the experimenter, so expected utility of lottery i is simply the probability weighted utility of each outcome j:

$$EU_i = \sum_{j=1,J} [\ p(M_j) \times U(M_j)\ ]. \tag{2}$$

The EU for each lottery pair is calculated for candidate estimates of $\varkappa_{10}$ , $\varkappa_{20}$ and $\varkappa_{35}$, and the index

$$\nabla EU = EU_R - EU_L \tag{3}$$

calculated, where $EU_L$ is the "left" lottery and $EU_R$ is the "right" lottery of a given lottery pair as presented to subjects. The latent index $\nabla EU$, based on latent preferences, is then linked to observed choices using a standard cumulative normal distribution function $\Phi(\nabla EU)$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1. Thus we have the probit link function,

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

The logistic function is very similar and leads instead to the "logit" specification.[14]

Thus the likelihood of the observed responses, conditional on the EUT specifications being true, depends on the estimates of $\varkappa_{10}$ , $\varkappa_{20}$ and $\varkappa_{35}$ given the above statistical specification and the observed choices. The "statistical specification" here includes assuming some functional form for the cumulative density function (CDF). The conditional log-likelihood is then

$$\ln \mathrm{L}(\varkappa_{10} , \varkappa_{20}, \varkappa_{35} ; \mathrm{y}, \mathbf{X}) = \textstyle\sum_i [ (\ln \Phi(\nabla \mathrm{EU})\times\mathbf{I}(\mathrm{y}_i = 1)) + (\ln (1\text{-}\Phi(\nabla \mathrm{EU}))\times\mathbf{I}(\mathrm{y}_i = -1)) ] \qquad (5)$$

where $\mathbf{I}(\cdot)$ is the indicator function, $\mathrm{y}_i = 1(-1)$ 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.

It is a simple matter to generalize this analysis to allow the core parameters $\varkappa_{10}$ , $\varkappa_{20}$ and $\varkappa_{35}$ to each be a linear function of observable characteristics of the individual or task. We would then extend the model to allow $\varkappa_{10}$ , for example, to be $\varkappa_{10} + \mathrm{K}\times\mathbf{X}$, where $\varkappa_{10}$ is a fixed parameter and K is a vector of effects associated with each characteristic in the variable vector **X**. In effect the unconditional model just estimates $\varkappa_{10}$ and assumes implicitly that K is a vector of zeroes. This extension significantly enhances the attraction of structural ML estimation, particularly for responses pooled over different subjects, which is a central issue here because of treatment A, since one can condition estimates on observable characteristics of the task or subject.

Harrison and Rutström [2008; Appendix F] review 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 non-EUT models. The goal is to illustrate how experimental economists can write explicit

[14] Even though (4) 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 ∇EU, and the probability of that index ∇EU being observed. In our applications ∇EU refers to some function, such as (3), of the EU of two lotteries; or, if one is estimating an RDU model, the rank-dependent utility of two lotteries. The index defined by (3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla \mathrm{EU})>\frac{1}{2}$, which is implied by (4) and the functional form of the cumulative normal distribution function $\Phi(\cdot)$.

maximum likelihood (ML) routines that are specific to different structural choice models. It is a simple matter to correct for multiple responses from the same subject ("clustering"),[15] or heteroskedasticity, as needed.

*B. Behavioral Errors*

An important extension of the core structural model is to allow for subjects to make some behavioral errors. We employ a Fechner error specification, popularized by Hey and Orme [1994], that posits the latent index

$$\nabla \text{EU} = (\text{EU}_R - \text{EU}_L)/\mu \qquad (3')$$

instead of (3). In this specification μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model.[16] 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 (3). Thus μ can be viewed as a parameter that flattens out the link function in (4) as μ gets larger.

[15] 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 and 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.

[16] This is just one of several different types of error story that could be used, and Wilcox [2008] provides a masterful review of the implications of the alternatives. Some specifications place the error at the final choice between one lottery or after the subject has decided 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.

An important contribution to the characterization of behavioral errors is the "contextual error" specification proposed by Wilcox [2011]. It is designed to allow robust inferences about the primitive "more stochastically risk averse than," and consistent inferences when one estimates over prize contexts in order to get better estimates (Figure 2). It posits the latent index

$$\nabla EU = ((EU_R - EU_L)/\nu)/\mu \qquad (3'')$$

instead of (3′), where ν is a normalizing term for each lottery pair L and R. The normalizing term ν is defined as the maximum utility over all prizes in this lottery pair minus the minimum utility over all prizes in this lottery pair. The value of ν varies, in principle, from lottery choice to lottery choice: hence it is said to be "contextual." For the Fechner error specification, dividing by ν ensures that the *normalized* EU difference $[(EU_R - EU_L)/\nu]$ remains in the unit interval. Our utility normalization (1) automatically ensures that the EU difference remains in the unit interval, but later specifications relax that, and normalization is needed then.

### *C. Rank-Dependent Models*

The RDU model extends the EUT model by allowing for decision weights on lottery outcomes. The specification of the utility function is the same non-parametric specification (1) considered for EUT. To calculate decision weights under RDU one replaces expected utility defined by (2) with RDU

$$RDU_i = \sum_{j=1,J} [\, w(p(M_j)) \times U(M_j) \,] = \sum_{j=1,J} [\, w_j \times U(M_j) \,] \qquad (2')$$

where

$$w_j = \omega(p_j + ... + p_J) - \omega(p_{j+1} + ... + p_J) \qquad (6a)$$

for j=1,... , J-1, and

$$w_j = \omega(p_j) \qquad (6b)$$

for j=J, with the subscript j ranking outcomes from worst to best, and ω(·) is some probability weighting function.

We could adopt the simple "power" probability weighting function proposed by Quiggin [1982], with curvature parameter γ:

$$\omega(p) = p^{\gamma} \tag{7}$$

So γ≠1 is consistent with a deviation from the conventional EUT representation. Convexity of the probability weighting function is said to reflect "pessimism." If one assumes for simplicity a linear utility function, this implies a risk premium.[17]

We use instead a non-parametric specification of the probability weighting function which exploits the fact that our main lottery parameters only use the 5 probabilities, 0, ¼, ½, ¾ and 1. If we constrain the extremes to have weight 0 and 1, we then have

$$\omega(0) = 0,\ \omega(\tfrac{1}{4}) = \varphi_{1/4}\ ,\ \omega(\tfrac{1}{2}) = \varphi_{1/2}\ ,\ \omega(\tfrac{3}{4}) = \varphi_{3/4}\ \text{ and }\ \omega(1) = 1 \tag{8}$$

and directly estimate $\varphi_{1/4}$ , $\varphi_{1/2}$ and $\varphi_{3/4}$ with the constraint that each lie in the unit interval. This is the approach employed by Gonzalez and Wu [1996] and Wilcox [2010]. The rest of the ML specification for the RDU model is identical to the specification for the EUT model, but with different parameters to estimate.

The Dual Theory (DT) specification of Yaari [1987] is the special case of the RDU model in which the utility function is assumed to be linear. Hence diminishing marginal utility can have no influence on the risk premium, and the only thing that can explain the risk premium is "probability pessimism."

[17] Since $\omega(p) < p \ \ \forall p$, the "RDU EV" in which monetary prizes are weighted by ω(p) instead of p has to be less than the EV weighted by p. Hence the CE under RDU has to be less than the true EV.

## 4. Results

We initially focus on behavior observed under treatments A, B and C, and evaluate the Bipolar Behavioral Hypothesis that risk preferences are the same across the three treatments.[18] We present the initial estimates assuming preference homogeneity across subjects, to be able to focus on the interpretation of non-parametric estimates of the utility and probability weighting functions. We then allow for preference heterogeneity. Although everyone says that they prefer to see non-parametric functions for utility and probability weighting, the corollary is that the resulting estimates can become detailed, since one eschews "boiling" down to just one or two parameters. So we recap at the end with some homely and intelligible parametric estimates, confirming our qualitative findings with non-parametric forms.

### *A. Non-Parametric Estimates Assuming Preference Homogeneity*

Baseline Estimates

Start with non-parametric estimates of the EUT, DT and RDU models in the payoff environment that does not assume IA: the 1-in-1 treatment A. Of course, EUT assumes IA, so EUT estimates under payoff environments that require IA, such as the 1-in-30 treatment B, will also be theoretically consistent with EUT estimates from treatment A. But the estimates for DT and RDU will not generally be theoretically consistent unless we use the 1-in-1 payoff environment.[19] So the estimates in Table 2 provide the first estimates, to the best of our knowledge, of DT and RDU when those estimates are not contaminated by having to assume the IA in the form of the Bipolar Behavioral Hypothesis. The estimates also provide the basis for testing our main hypothesis: that risk preferences estimated under EUT or RDU change when one moves away from payoff

[18] We implicitly view treatments B and C as the same here, and check for differences in due course.

[19] Or somehow model the full portfolio of 30 sequential choices as if it were one choice.

environments that assume the IA to be valid. Of course, as stressed earlier, the "bad news" theoretically is that one must make an assumption of homogeneous preferences across individuals to interpret these estimates as reflecting risk preferences. Popular as that assumption is, we can and will relax it.

Panel A in Table 2 shows the EUT estimates for each interior prize. The point estimates are increasing in the prize value, consistent with non-satiation, $\partial U(x)/\partial x > 0$. The 95% confidence intervals are generally tight, in the sense of allowing one to rule out the hypothesis that these estimates are statistically indistinguishable from 0 or 1.[20] They also suggest that the estimates satisfy non-satiation even when one allows for sampling error. For example, the 95% confidence interval for the U($10) estimate is between 0.05 and 0.27, and the 95% confidence interval for the U($20) estimate is between 0.34 and 0.56, so there is no overlap. There is some slight overlap between the 95% confidence interval for U($20) and the interval for U($35), which is between 0.51 and 0.79. The statistical significance of this overlap is tested directly in the next two lines with $\Delta U_{20 \leftrightarrow 35}$, which is the difference in the utilities: if this is positive, and statistically significantly different from zero, as it is, then we can be confident that these estimates satisfy non-satiation. The same is true, as expected, of the increment from U($10) to U($20), shown by $\Delta U_{10 \leftrightarrow 20}$.

We also directly test for diminishing marginal utility, $\partial^2 U(x)/\partial x^2 < 0$, by evaluating the marginal utility of each increment in utility, and then seeing if the difference between the first and second marginal utility is positive. The estimates show that each of the marginal utilities is positive, as one would expect from the non-satiation result, and that there is evidence of statistically

[20] In a numerical sense this might not be surprising, since we estimate these parameters by using a non-linear transform that ensures that they lie in the unit interval, as theory suggests. But it is still possible for the sampling errors to be large enough that the 95% confidence intervals get very close to 0 or 1, and as a practical matter for finite samples this can occur. The "delta method" is used to infer point estimates and standard errors from non-linear transformations of this kind (Oehlert [1992]), and it includes some approximation error which can be particularly noticeable when point estimates are close to the boundary.

significant diminishing marginal utility.

Turning to the DT estimates in Panel B of Table 2, the aggregate log-likelihood is better than the aggregate log-likelihood for EUT. We later consider the evidence for and against different models more carefully, since DT and EUT are non-nested, but this is an intriguing finding for the most interesting, parsimonious alternative to EUT, at least under the assumption of homogeneous preferences.[21] Since the EUT estimates show diminishing marginal utility, we infer that the risk premium is positive, so it is no surprise to see that the point estimates for the DT model show probability "pessimism." The estimated probability weights for the ¼, ½ and ¾ probabilities are only 0.21, 0.27 and 0.56, respectively. From the 95% confidence intervals on these point estimates, and the *p*-values on the increments in probability weight ($\Delta p_{1/4 \leftrightarrow 1/2}$ and $\Delta p_{1/2 \leftrightarrow 3/4}$), we see that these estimates indicate a non-decreasing probability weighting function from ¼ to ½, and an increasing probability weighting function from ½ to ¾. Finally, we confirm that the probability weights for ½ and ¾ are indeed statistically significantly below the true probability, by evaluating the estimated differences between the probability weights and the true probability: $\varphi_{1/4}$ - ¼, $\varphi_{1/2}$ - ½ and $\varphi_{3/4}$ - ¾. This is true for two of the three individual probability weight differences, and for all of the differences considered jointly, so there is clearly some violation of the IA that could interact with the payoff environment once we consider the 1-in-30 treatment.

Panel C presents estimates for the RDU model, combining the two "risk premium stories" from EUT and DT. Not surprisingly, it has an aggregate log-likelihood that is better than either of those two nested alternatives. The most interesting feature of these estimates is the striking role of diminishing marginal utility and the minor role of probability weighting. The estimated probability weights for the ¼, ½ and ¾ probabilities are 0.30, 0.38 and 0.69, respectively, and in each case the

[21] Expected value is the most parsimonious alternative, but not interesting.

95% confidence interval includes the true probability. The 95% confidence interval for $\varphi_{1/2}$ is between 0.18 and 0.58, and overlaps with the 95% confidence interval for $\varphi_{1/4}$. In fact, the increase of 8.8 percentage points from $\varphi_{1/4}$ to $\varphi_{1/2}$ has a *p*-value of 0.115; although a one-sided hypothesis test would be appropriate here, given our prior of an increasing probability weighting function, this still implies a *p*-value of 0.057. A $\chi^2$ test of the hypothesis that all three of these estimated probability weights are equal to the corresponding probability has a *p*-value of 0.03, implying that *there is evidence of statistically significant probability weighting*. The estimated utility function under RDU exhibits the familiar properties of non-satiation and diminishing marginal utility. Again, these conclusions are all under the maintained assumption of preference homogeneity across subjects.

### The Effect of Being Bipolar

These estimates provide the baseline for evaluating the effect of the 1-in-30 payoff treatment on risk preferences. Table 3 shows more estimates, again assuming that risk preferences are homogeneous across individuals. In this case we employ all of the data from Table 1, and include binary dummy variables for the variations in treatments B and C compared to treatment A. The first three lines in Panel A of Table 3 show estimates of $\varkappa_{10}$ , $\varkappa_{10}^{\text{pay1}}$ and $\varkappa_{10}^{\text{ra\_idr}}$ from U($10) = $\varkappa_{10}$ + $\varkappa_{10}^{\text{pay1}}$×**pay1** + $\varkappa_{10}^{\text{ra\_idr}}$×**ra_idr**, where **pay1** is a binary dummy variable equal to 1 for the 1-in-1 treatment and 0 otherwise, and **ra_idr** is a binary dummy variable equal to 1 for the 1-in-30 treatment in which there was an additional, salient, individual discount rate elicitation task after the lottery choices. In each case we show the marginal effect of the binary variable, so we see that U($10) = 0.21 - 0.072×**pay1** + 0.037×**ra_idr**.

We find no statistically significant effect of the treatments on the estimated utility values under EUT. In one respect this is just comforting, and not "news," since EUT assumes the IA and the IA is what makes treatments B and C formally the same as treatment A.

There is a different story with DT, which of course relies on probability weighting and relaxations of IA to explain the risk premium. Here we do see some statistically significant effects when comparing the 1-in-1 and 1-in-30 treatments. For the ¼ probability weight, we find that the 1-in-1 treatment increases the weighted probability from 0.07 by 0.18, and that this increase is statistically significant with a *p*-value of 0.013. Similarly, for the ½ probability weight there is an effect from having a paid task follow the lottery choice task; it makes the probability weight even more pessimistic, by 6.7 percentage points, and has a *p*-value of 0.077. Overall, a $\chi^2$ test confirms that the pay1 and ra_idr treatments are jointly significant across all three probability weight coefficients, with a *p*-value of 0.004. The aggregate log-likelihood for the DT model in this case is worse than the aggregate log-likelihood for the EUT model. Hence *the inferred DT preferences are sensitive to the use of a payment protocol that assumes the IA.*

In many respects the RDU results are the most interesting, since Table 2 suggested that there was evidence for probability weighting overall, and that the IA axiom was therefore significantly violated. If the IA is significantly violated, then we might expect to see different risk preferences under RDU when we merge in the 1-in-30 choices, just as we did with the DT specification that assumes that all of the risk premium derives from a IA violation. This is indeed what we see in Panel C of Table 3, although it is not obvious from examination of the individual significance levels. None of the treatment dummies are individually statistically significant, even though there is a hint of some effect on the probability weights for the ¼ and ½ probabilities of the 1-in-1 treatment; the *p*-values on these estimated effects are 0.19 and 0.12, respectively, but they are large in size.

Overall, a $\chi^2$ test indicates that the treatment dummies are *not* a significant factor across *all* estimated coefficients, with a *p*-value of 0.13. But the effect is significant for the probability weighting coefficients, with a *p*-value of 0.05 for those taken jointly (the *p*-value for the effect on the utility coefficients is 0.76). So we *do see some statistically significant effect of the payoff treatment on elicited preferences under RDU, deriving from effects on the estimated degree of probability weighting*. Again, however, we stress that this is still under the maintained assumption of preference homogeneity across subjects. It is time to relax that assumption and re-evaluate the inferences about the payoff treatments.

### *B. Non-Parametric Estimates Allowing Preference Heterogeneity*

We extend the estimation to include a set of observable characteristics of the individual. We employ a series of binary variables: **female** is 1 for women, and 0 otherwise; **freshman**, **sophomore**, and **senior** are 1 for whether that was the current stage of undergraduate education at GSU, and 0 otherwise; **asian** and **white** are 1 based on self-reported ethnic status, and 0 otherwise; and **gpaVHI** is 1 for those reporting a cumulative GPA between 3.5 and 4.0 (mostly A's), and 0 otherwise.[22] Table 4 shows the detailed effect of allowing for this observable heterogeneity in the EUT model, and Table 5 shows the effect on the estimates of the treatment variables in the DT and RDU models. So in Table 5 we suppress all of the estimates of demographics, and focus just on the estimates of interest for our inferences. The demographic characteristics as a whole are statistically significant for all three models.[23]

Table 4 shows that *allowing for subject heterogeneity does not change the inferences about risk preferences under EUT*. Again, this is expected, given that the 1-in-30 treatments should theoretically have no effect on elicited risk preferences if the IA holds, and EUT assumes the IA. A $\chi^2$ test of the joint

[22] We would normally include a measure of age as well, but the sample variation was too small for this to be useful, and highly correlated with the levels of undergraduate standing.

[23] For the EUT model a $\chi^2$ test on this hypothesis has a *p*-value of 0.016. For the DT model the *p*-value is 0.02, and for the RDU model the *p*-value is 0.04 for the utility parameters and 0.01 for the probability weighting parameters (and less than 0.0001 for all parameters).

significance of these treatment variables across all estimates has a *p*-value of 0.70, confirming that conclusion. Figure 4 illustrates the predicted values of utility across all subjects, using the estimated model in Table 4 to generate these predictions.[24]

Much more interesting results arise with the DT and RDU model estimates in Table 5. In the case of DT, we have a significant effect of the variable pay1 on the probability weight for ¼, and a close to significant effect of the variable ra_idr on the probability weight for ½. Overall, a $\chi^2$ test shows a significant effect on all estimates with a *p*-value of 0.033, confirming that relying *entirely* on a certain deviation from the IA to explain risk preferences does lead to different estimates of risk preferences when one has to assume the IA with respect to the payment procedures in order to make inferences. The aggregate log-likelihood of the DT model is worse than the aggregate log-likelihood of the comparable EUT model in Table 4. This reverses the, mildly surprising, relationship obtained when assuming homogeneous preferences.

For the RDU model we observe only one significant individual effect at conventional levels, from the pay1 variable on the probability weight for ½. However, we *do find a significant overall effect from the 1-in-1 treatment on probability weights*. A $\chi^2$ test on the hypothesis that this treatment has no effect on all three probability weights can be rejected with a *p*-value of 0.045. The 1-in-1 treatment has no significant effect on the utility parameters. Figure 5 illustrates the predicted probability weights generated from the full model, with heterogeneity, underlying the estimates reported in Panel B of Table 5.

In summary, and allowing for observable heterogeneity in preferences, we conclude that

- there is no evidence that estimated EUT preferences are affected by the two experimental payment protocols employed;

[24] These predictions reflect the point estimates in Table 4, and not the sampling errors. Formal hypothesis tests must take those sampling errors into account.

- there is evidence that estimated DT preferences are affected by the use of an experimental payment protocol that assumes the validity of the very axiom that DT relaxes in order to explain the risk premium; and
- there is evidence that estimated RDU preferences are also affected by the use of an experimental payment protocol that requires the validity of the IA.

These results imply that the Bipolar Behaviorist is in urgent need of medication. It is not possible to simultaneously maintain that (a) the IA is invalid in the latent specification of choices over pairs of lotteries, and that (b) the IA is magically valid when paying subjects for more than one choice. We often hear the "isolation effect" invoked to allow this discord to stand, as noted earlier, but we have not seen that effect stated in a formal manner that explains how it differs from the IA. It is used in scientific rhetoric more in the manner of a behavioral "get out of jail free card" in the parlor game *Monopoly*.

### *C. Parametric Estimates*

We employ familiar specifications for the parametric utility and probability weighting functions. Instead of (1) for the utility function, we use the Expo-Power (EP) utility function proposed by Saha [1993]. Following Holt and Laury [2002], the EP function can be defined as

$$U(x) = [1-\exp(-\alpha x^{1-r})]/\alpha, \tag{9}$$

where $\alpha$ and r are parameters to be estimated. RRA is then $r + \alpha(1-r)x^{1-r}$, so RRA varies with income x if $\alpha \neq 0$. This function nests CRRA (as $\alpha \rightarrow 0$) and CARA (as $r \rightarrow 0$), so can be unbounded or bounded depending on particular parameter values. Instead of (8) for the probability weighting function, we employ the power function $\omega(p) = p^\gamma$ defined earlier by (7) and the inverse-S function

$$\omega(p) = p^\gamma / ( p^\gamma + (1-p)^\gamma )^{1/\gamma} \tag{10}$$

This function exhibits inverse-S probability weighting (optimism for small p, and pessimism for large p) for $\gamma<1$, and S-shaped probability weighting (pessimism for small p, and optimism for large p) for $\gamma>1$. We are aware that there are more exotic functional forms, particularly for probability weighting, but we have already evaluated a completely non-parametric form in (8), so we use the simplest, popular, one-parameter functions (7) and (10).

Figures 6 and 7 show the effects of moving from the 1-in-1 payment protocol to the 1-in-30 payment protocol for DT and RDU models, assuming for now homogeneous preferences across all subjects. The differences are striking, quantitatively *and* qualitatively, no matter which probability weighting function is used. Since we know that the primary effect of the payment protocol is on the estimated probability weighting, it is to be expected that the effects would be more dramatic for DT than for RDU. For both DT and RDU the preferred probability weighting function is the inverse-S, which we use for the heterogenous preferences specifications.

Turning to specifications which control for observable characteristics of individual decision makers, we can formally test the statistical significance of the effect of the 1-in-30 payment protocol using the 1-in-1 payment protocol as the baseline. For the EUT model, the joint hypothesis that the 1-in-30 dummy on the structural coefficients r and $\alpha$ are both equal to zero cannot be rejected, with a *p*-value of 0.65 (and the *p*-values for r and $\alpha$ separately are 0.36 and 0.73, respectively). This confirms our earlier finding that under EUT there is no statistically significant difference in elicited risk preferences across the two payment protocols.

For DT the hypothesis that the 1-in-30 dummy on the structural coefficient $\gamma$ is equal to zero can be rejected with a *p*-value of 0.026. The qualitative effect on probability weighting, allowing for observed heterogeneity, is the same as shown in Figure 6.

For RDU the joint hypothesis that the dummy on the structural coefficients r, $\alpha$ and $\gamma$ are all equal to zero can be rejected with a *p*-value of 0.019. In this case it is noteworthy that, consistent with the non-parametric findings, that the culprit is the probability weighting parameter: the *p*-values for the r, α and γ coefficients alone are 0.57, 0.58 and 0.003, respectively. The qualitative effect on probability weighting, allowing for observed heterogeneity, is also the same as shown in Figure 7.

## 5. Implications

*A. Immediate Implications*

A first implication of our results is to encourage theorists to come up with payment protocols that allow one to elicit multiple choices but do not require that one violate an assumption required for the coherent specification of the particular decision model. This challenge has been directly addressed, and partially met, by Cox, Sadiraj and Schmidt [2011]. For the DT of Yaari [1987] and the Linear Cumulative Prospect Theory model of Schmidt and Zank [2009], they devise payment protocols that *should* generate estimates of the same preferences as the 1-in-1 protocol.[25] There are no known, or obvious, payment protocols that can be used for RDU and Cumulative Prospect Theory.

A second implication of our results is to question inferences made about *specific alternative hypotheses* to EUT when the 1-in-K protocol has been employed. That is, in literally every test of specific alternatives to EUT that we are aware of. This is not to say that EUT is valid, just that tests of the validity of specific alternatives rest on a maintained assumption that is false. Our results suggest an obvious research strategy to properly evaluate the validity of EUT in an efficient manner. Examine the catalog of anomalies that arise in choice tasks over simple lotteries using a 1-in-K payment protocol, for some large K, and then for those anomalies that survive, drill down with the

[25] The Linear Cumulative Prospect Theory model assumes linear utility, but allows the probability weighting of DT with the addition of loss aversion over utilities.

more expensive 1-in-1 protocol. This strategy does run the risk that there could be "offsetting violations" of EUT in the 1-in-K payment protocol, but that is a tradeoff that many scholars would, we believe, be willing to take. And the alternative to the tradeoff is simple enough: replicate every anomaly using the 1-in-1 payment protocol.

A third, costly implication of our results, then, is to place a premium on collecting choice data in smaller doses, using 1-in-1 payment protocols. Anyone proposing new anomalies should be encouraged to take their Bipolar Behaviorist medication, and demonstrate that the alleged misbehavior persists when one removes the obvious theoretical confound.

A fourth, modeling implication of the need for 1-in-1 choice data is to place greater urgency on the use of rigorous econometric methods to flexibly characterize heterogeneous preferences. Random coefficient methods can be used to better characterize unobserved individual heterogeneity for *non-linear structural* econometric models.[26] Or one can consider semi-parametric stochastic specifications, to complement the non-parametric specifications of utility and probability weighting functions employed here.

A fifth implication is to consider more rigorously the learning behavior that might change behavior towards lottery choices such as these. Binmore [2007; p. 6ff.] has long made the point that we ought to recognize that the artefactual nature of the usual laboratory tasks, and indeed some tasks in the field, means that we should allow subjects to learn how to behave in that environment before drawing unconditional conclusions. Although his immediate arguments are about the study

[26] We stress the words non-linear and structural here. The "mixed logit" theorem shows that the *linear* mixed logit specification can approximate arbitrarily well any random-utility model (McFadden and Train [2000]). One needs a non-linear structural specification because these results only go in one direction: for any specification of a latent structure, defined over "deep parameters" such as risk preferences, they show that there *exists* an equivalent linear mixed logit. But they do not allow the direct recovery of those deep parameters in the estimates from the linear mixed logit. The deep parameters, which are typically the very things of interest, are buried in the estimates from the mixed logit, but can only be identified with extremely restrictive assumptions about functional form of the structural model.

of strategic behavior in games, they are general. Thus the argument is that one would expect 1-in-1 behavior to differ from 1-in-30 behavior since the latter reflects some learning behavior. The problem with this line of argument is that it is silent as to what should be compared to what, and does not provide a metric for defining when learning is finished. One could mitigate the issue by providing subjects with lots of experience in one session, and then invite them back for further experiments, either 1-in-1 or 1-in-30, arguing on *a priori* grounds that any behavior differences then should reflect longer-run, steady-state behavior for this task. We are sympathetic to this view, and indeed it was implicit in the early days of experimental economics where "experience" meant that subjects has participated in some task and then had time to "sleep on it" before the next session. The hypothesis implied here is that the differences we find would diminish if subjects were given "enough" experience, which is of course testable if one can define what "enough" means.

### *B. A More Subtle Implication: Modeling Portfolios*

A final implication is to model the effects of treating behavior as if generated by portfolio formation for the experiment as a whole. Indeed, an important subtlety emerges when properly interpreting our results, which we believe to be significant for future research. We find from our 1-in-1 tasks that procedures for estimating non-EUT risk preferences are required, but that they do not generate *consistent* estimates of preferences when one uses the standard 1-in-K payment protocol. We stress the word consistent for a reason: the results tell us that there are differences in DT and RDU estimates when one assumes that the 1-in-K payment protocol generates the same risk preferences as the 1-in-1 payment protocol. However, the estimated risk preferences need not be the same under these two payment protocols, and indeed there are theoretical grounds for expecting them not to be if the IA is violated. Payment protocol 1-in-1 has the advantage that it does not rely on IA, and that provides a critical behavioral Gold Standard to use for our purposes. But these results only show that data generated under payment protocol 1-in-K cannot be used to estimate DT or RDU risk preferences that are the *same* as those estimated under payment protocol 1-in-1. The implication is that one has to account for the effects of the violation in the IA in protocol 1-in-K in order to correctly estimate DT or RDU risk preferences from data generated under protocol 1-in-K. It is possible that these *theoretically correct estimates of DT or RDU* in protocol 1-in-K are the same as those obtained from protocol 1-in-1.

Table 6 shows the possible interactions between assumptions used for estimating risk preferences and payment protocols. Since the IA does not influence the 1-in-1 payment protocol, the risk preferences estimated in cell III are identical, by construction, to those estimated in cell V. But the risk preferences in III and V need not be the same as those estimated in cell I, since the IA plays a role in the evaluation of the lotteries that are the object of the sole choice under the 1-in-1 protocol. Our first result is that the risk preferences in cell I are indeed different from those in cells III and V.

Since EUT assumes the IA, in theory the risk preferences estimated in cell II should be the same as those in cell I, and indeed they are behaviorally, as we have demonstrated. But one can estimate DT or RDU preferences in two ways. One way assumes the Bipolar Behavioral Hypothesis, in cell IV. The other way, in cell VI, assumes that the same violation of the IA that applies for the evaluation of the constituent lotteries of the choice in cell V and choices in cell VI also applies to the evaluation of the compound lottery implied by the payment protocol. Hence the subtle point we are making is that evidence of differences in risk preferences in cells II and IV does not imply that there would be differences in the risk preferences in cells V and VI. Cell VI is what we referred to above as the theoretically consistent estimates of DT or RDU.

Another way of stating this is that we do *not* label choices under other payment protocols "incentive compatible" if they happen to match the choices under the 1-in-1 payment protocol. An allocative mechanism or institution is said to be incentive compatible when its rules provide individuals with incentives to truthfully and fully reveal their preferences. The fact that preferences are *different* in a 1-in-K setting to the preferences in a 1-in-1 setting does not make the 1-in-K preferences untruthful in any useful sense of the word. Instead, they might just reflect true risk preferences when selecting a compound lottery, which is inapplicable by construction in the 1-in-1 setting.

The research implication is to design experiments in which it is tractable to model the portfolio explicitly. Using K=2 would be sufficient for this purpose, with each binary choice again defined over simple lotteries.[27] Then the task is to write out explicit structural models that relax the IA of EUT in one or other manner to evaluate the portfolio of 4 combinations that could be chosen. It is also feasible to consider K=3 or K=4 as well, generating portfolios of 8 or 16

[27] Choosing K=2, the smallest integer greater than 1, allows easy visualization of the complete set of lottery pairs using a display format akin to the one we use, and facilitates tractable evaluation of the hypothesis that subjects are evaluating the "grand" compound lottery by considering the experiment as one single decision problem. This is simply infeasible with K=30, whether or not the 30 pairs are presented sequentially. Hey and Lee [2005b; p. 234] document the extent of the problem, and the sad outcome for them: "The crucial point is that, if the subject does not have EU preferences, and if the subject considers the experiment as a whole, then the responses on individual questions may well not reflect the true preferences of that subject with respect to the individual questions. This objection was raised by a referee on an experiment carried out by one of the authors in which subjects were asked 30 pairwise choice questions. The referee asked: 'how do you know that the subjects were answering the questions individually and not answering to the experiment as a whole? How do you know that subjects were not choosing the best strategy for the experiment as a whole?' The response made to the referee was that if the subjects tried to do the latter, then they would have to choose between $2^{30}$ = 1,073,741,824 different strategies, and that this was computationally difficult and therefore unlikely. The referee was not satisfied by this response and countered with the usual 'as-if' arguments. These were enough to convince the editor." The problem is obviously exacerbated dramatically when the specific lotteries to come in future stages are not known, and have to be guessed at if the subject is to choose the best strategy for the experiment as a whole. This turns a problem of decision making under objective risk into a challenging problem of decision making under subjective ambiguity. Although one could envisage procedures to address this concern, it is easier to focus on the simplest case in which this information can be communicated in a way that does not dramatically change the cognitive burden of the series of tasks.

combinations. One would then estimate the risk preferences for those models and compare them to those obtained from the 1-in-1 choice tasks.

## 6. Conclusions

Bipolar disorders have several manifestations, apart from making it hard to lead a stable, productive life. One important manifestation is that sufferers are often mis-diagnosed as being depressives, since that is what typically leads them to present themselves for scrutiny by trained specialists. The serious consequence of this is that the treatment for depression often makes bipolar disorders much worse. So it is important that our powerful diagnostic test, the 1-in-1 payment protocol, confirms that what appears to be a bipolar disorder among behaviorists is indeed straightforward depression about the Independence Axiom. The treatment then shifts to untangling the way in which that axiom fails when one does not have inferences confounded by the payment protocol.

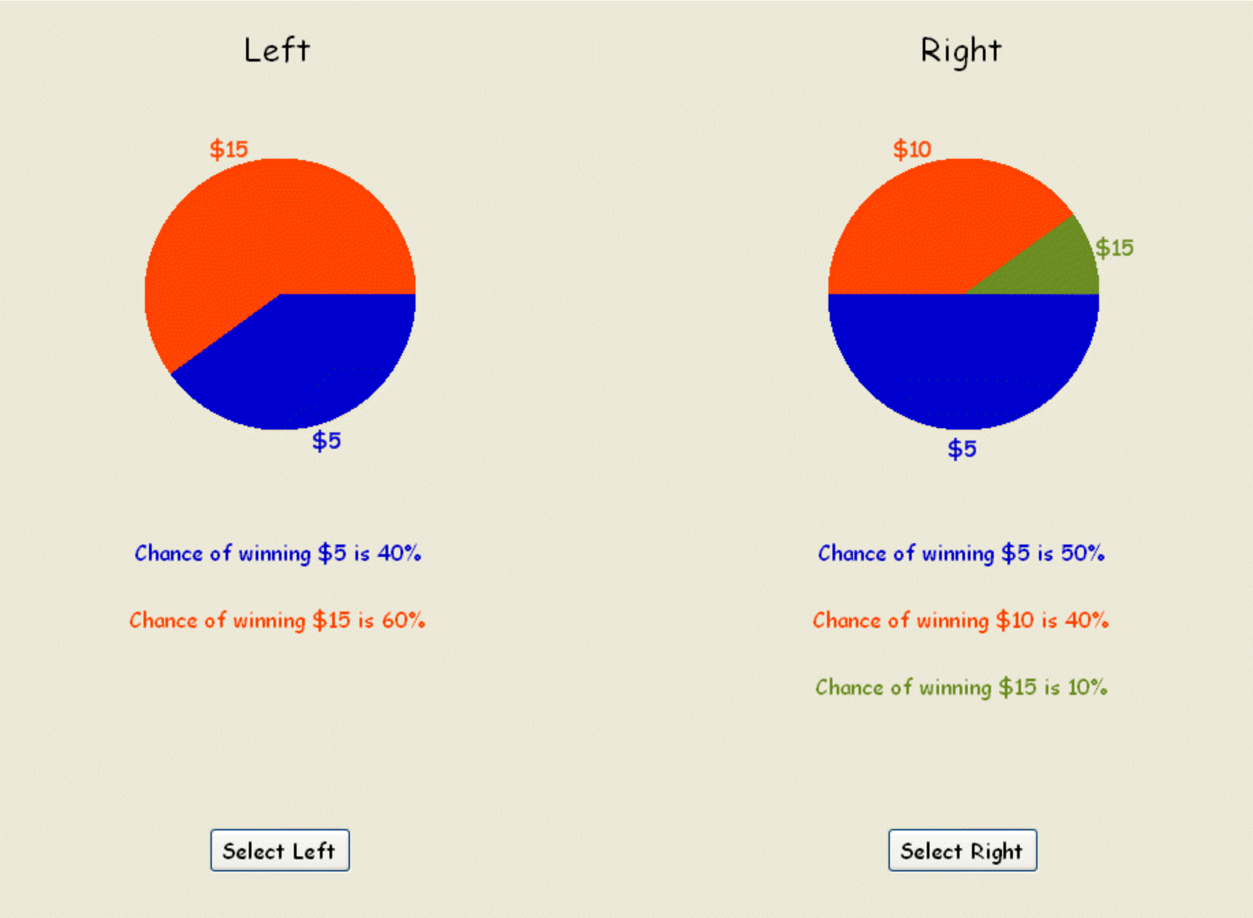
### Table 1: Experimental Design

All choices drawn from the same battery of 69 lottery pairs at random.
All subjects receive a $7.50 show-up fee.
Unless otherwise noted for treatment C, subjects were told that there would be no other salient task in the experiment.

| Treatment | Subjects | Choices |
|---|---|---|
| **A**. 1-in-1 | 75 | 75 |
| **B**. 1-in-30 Sequential | 37 | 1110 |
| **C**. 1-in-30 Sequential with an additional paid task † | 236 | 7080 |

Notes: † additional task was a time-discounting choice, after the risky lottery choices, and the subjects were told at the outset that there could be additional salient tasks.

**Figure 1: Default Binary Choice Interface**



Chance of winning $5 is 40%

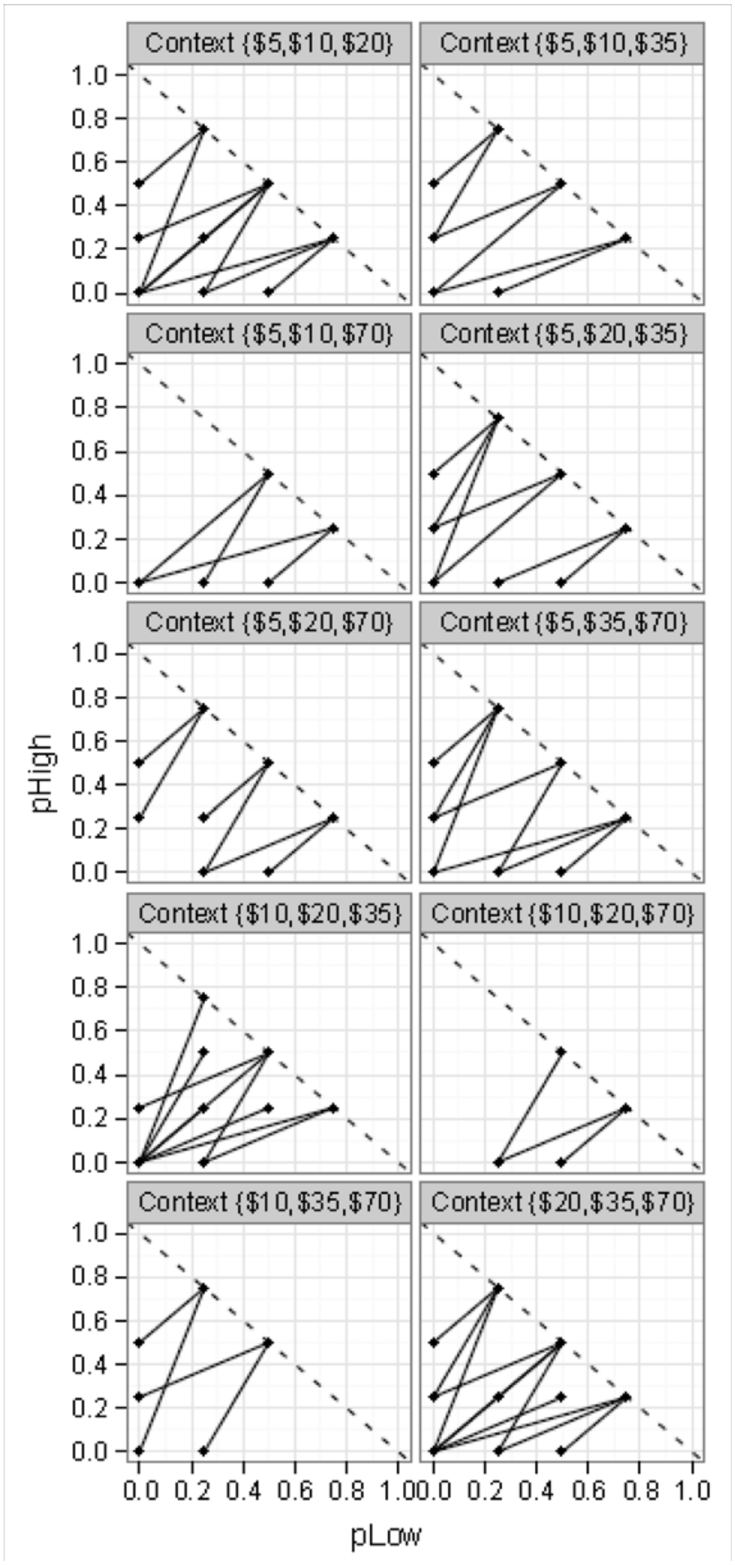
Chance of winning $15 is 60%

Chance of winning $5 is 50%

Chance of winning $10 is 40%

Chance of winning $15 is 10%

## Figure 2: Lotteries in the Marschak-Machina Triangle



## Table 2: Non-Parametric Estimates Assuming Homogeneity and 1-in-1 Choices

| Parameter | Point Estimate | Standard Error | *p*-value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | *A. Expected Utility Theory* (LL = -37.8) | | |
| $\varkappa_{10}$ | 0.163 | 0.056 | 0.004 | 0.053 | 0.273 |
| $\varkappa_{20}$ | 0.449 | 0.058 | <0.001 | 0.336 | 0.562 |
| $\varkappa_{35}$ | 0.653 | 0.071 | <0.001 | 0.513 | 0.793 |
| $\Delta U_{10 \leftrightarrow 20}$ | 0.286 | 0.052 | <0.001 | 0.184 | 0.388 |
| $\Delta U_{20 \leftrightarrow 35}$ | 0.203 | 0.049 | <0.001 | 0.108 | 0.299 |
| $\Delta U_{10 \leftrightarrow 20} \div 10$ | 0.029 | 0.005 | <0.001 | 0.018 | 0.039 |
| $\Delta U_{20 \leftrightarrow 35} \div 15$ | 0.014 | 0.003 | <0.001 | 0.007 | 0.020 |
| $\partial^2 U(x)/\partial x^2$ | 0.015 | 0.007 | 0.021 | 0.002 | 0.028 |
| | | | *B. Dual Theory* (LL = -35.7) | | |
| $\varphi_{1/4}$ | 0.207 | 0.042 | <0.001 | 0.125 | 0.289 |
| $\varphi_{1/2}$ | 0.271 | 0.041 | <0.001 | 0.191 | 0.352 |
| $\varphi_{3/4}$ | 0.561 | 0.052 | <0.001 | 0.458 | 0.663 |
| $\Delta p_{1/4 \leftrightarrow 1/2}$ | 0.064 | 0.048 | 0.178 | -0.029 | 0.158 |
| $\Delta p_{1/2 \leftrightarrow 3/4}$ | 0.289 | 0.047 | <0.001 | 0.197 | 0.382 |
| $\varphi_{1/4} - 1/4$ | -0.043 | 0.042 | 0.303 | -0.125 | 0.039 |
| $\varphi_{1/2} - 1/2$ | -0.229 | 0.041 | <0.001 | -0.309 | -0.148 |
| $\varphi_{3/4} - 3/4$ | -0.189 | 0.052 | <0.001 | -0.292 | -0.087 |
| | | | *C. Rank-Dependent Utility Theory* (LL = -34.7) | | |
| $\varkappa_{10}$ | 0.169 | 0.086 | 0.049 | 0.000 | 0.337 |
| $\varkappa_{20}$ | 0.395 | 0.124 | 0.001 | 0.152 | 0.639 |
| $\varkappa_{35}$ | 0.608 | 0.121 | <0.001 | 0.369 | 0.846 |
| $\Delta U_{10 \leftrightarrow 20}$ | 0.227 | 0.054 | <0.001 | 0.122 | 0.332 |
| $\Delta U_{20 \leftrightarrow 35}$ | 0.212 | 0.039 | <0.001 | 0.135 | 0.289 |
| $\Delta U_{10 \leftrightarrow 20} \div 10$ | 0.023 | 0.005 | <0.001 | 0.012 | 0.033 |
| $\Delta U_{20 \leftrightarrow 35} \div 15$ | 0.014 | 0.003 | <0.001 | 0.009 | 0.019 |
| $\partial^2 U(x)/\partial x^2$ | 0.009 | 0.007 | 0.188 | -0.004 | 0.021 |
| $\varphi_{1/4}$ | 0.297 | 0.086 | 0.001 | 0.128 | 0.465 |
| $\varphi_{1/2}$ | 0.385 | 0.102 | <0.001 | 0.185 | 0.584 |
| $\varphi_{3/4}$ | 0.686 | 0.095 | <0.001 | 0.500 | 0.871 |
| $\Delta p_{1/4 \leftrightarrow 1/2}$ | 0.088 | 0.056 | 0.118 | -0.022 | 0.199 |
| $\Delta p_{1/2 \leftrightarrow 3/4}$ | 0.301 | 0.046 | <0.001 | 0.210 | 0.392 |
| $\varphi_{1/4} - 1/4$ | 0.047 | 0.086 | 0.588 | -0.122 | 0.215 |
| $\varphi_{1/2} - 1/2$ | -0.115 | 0.102 | 0.258 | -0.315 | 0.084 |
| $\varphi_{3/4} - 3/4$ | -0.064 | 0.095 | 0.496 | -0.250 | 0.121 |

## Table 3: Non-Parametric Estimates Assuming Homogeneity

Data from treatment A, B and C

| Parameter | Point Estimate | Standard Error | *p*-value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| *A. Expected Utility Theory* (LL = -3761.8) | | | | | |
| $\varkappa_{10}$ constant | 0.207 | 0.034 | <0.001 | 0.141 | 0.273 |
| $\varkappa_{10}$ pay1 | -0.072 | 0.075 | 0.333 | -0.218 | 0.074 |
| $\varkappa_{10}$ ra_idr | 0.037 | 0.038 | 0.330 | -0.037 | 0.111 |
| $\varkappa_{20}$ constant | 0.451 | 0.045 | <0.001 | 0.361 | 0.540 |
| $\varkappa_{20}$ pay1 | -0.021 | 0.082 | 0.798 | -0.182 | 0.140 |
| $\varkappa_{20}$ ra_idr | 0.053 | 0.050 | 0.293 | -0.046 | 0.152 |
| $\varkappa_{35}$ constant | 0.662 | 0.039 | <0.001 | 0.585 | 0.739 |
| $\varkappa_{35}$ pay1 | -0.037 | 0.095 | 0.694 | -0.223 | 0.148 |
| $\varkappa_{35}$ ra_idr | 0.020 | 0.044 | 0.647 | -0.066 | 0.105 |
| *B. Dual Theory* (LL = -3806.1) | | | | | |
| $\varphi_{1/4}$ constant | 0.073 | 0.032 | 0.021 | 0.011 | 0.135 |
| $\varphi_{1/4}$ pay1 | 0.181 | 0.073 | 0.013 | 0.039 | 0.324 |
| $\varphi_{1/4}$ ra_idr | 0.006 | 0.035 | 0.868 | -0.063 | 0.075 |
| $\varphi_{1/2}$ constant | 0.395 | 0.034 | <0.001 | 0.328 | 0.463 |
| $\varphi_{1/2}$ pay1 | -0.081 | 0.068 | 0.233 | -0.213 | 0.052 |
| $\varphi_{1/2}$ ra_idr | -0.067 | 0.038 | 0.077 | -0.142 | 0.007 |
| $\varphi_{3/4}$ constant | 0.611 | 0.045 | <0.001 | 0.524 | 0.699 |
| $\varphi_{3/4}$ pay1 | 0.009 | 0.093 | 0.924 | -0.173 | 0.191 |
| $\varphi_{3/4}$ ra_idr | -0.050 | 0.048 | 0.298 | -0.145 | 0.044 |
| *C. Rank-Dependent Utility Theory* (LL = -3724.6) | | | | | |
| $\varkappa_{10}$ constant | 0.236 | 0.044 | <0.001 | 0.149 | 0.323 |
| $\varkappa_{10}$ pay1 | -0.097 | 0.087 | 0.266 | -0.268 | 0.074 |
| $\varkappa_{10}$ ra_idr | -0.002 | 0.052 | 0.967 | -0.104 | 0.100 |
| $\varkappa_{20}$ constant | 0.501 | 0.063 | <0.001 | 0.377 | 0.625 |
| $\varkappa_{20}$ pay1 | -0.100 | 0.113 | 0.374 | -0.322 | 0.121 |
| $\varkappa_{20}$ ra_idr | -0.007 | 0.072 | 0.918 | -0.148 | 0.134 |
| $\varkappa_{35}$ constant | 0.705 | 0.051 | <0.001 | 0.604 | 0.805 |
| $\varkappa_{35}$ pay1 | -0.120 | 0.122 | 0.325 | -0.358 | 0.119 |
| $\varkappa_{35}$ ra_idr | -0.035 | 0.059 | 0.554 | -0.152 | 0.081 |
| $\varphi_{1/4}$ constant | 0.208 | 0.048 | <0.001 | 0.114 | 0.302 |
| $\varphi_{1/4}$ pay1 | 0.130 | 0.099 | 0.189 | -0.064 | 0.324 |
| $\varphi_{1/4}$ ra_idr | -0.004 | 0.054 | 0.943 | -0.110 | 0.102 |
| $\varphi_{1/2}$ constant | 0.594 | 0.045 | <0.001 | 0.505 | 0.683 |
| $\varphi_{1/2}$ pay1 | -0.163 | 0.105 | 0.120 | -0.369 | 0.043 |
| $\varphi_{1/2}$ ra_idr | -0.079 | 0.053 | 0.137 | -0.182 | 0.025 |
| $\varphi_{3/4}$ constant | 0.802 | 0.041 | <0.001 | 0.723 | 0.881 |
| $\varphi_{3/4}$ pay1 | -0.054 | 0.102 | 0.596 | -0.253 | 0.145 |
| $\varphi_{3/4}$ ra_idr | -0.051 | 0.047 | 0.276 | -0.144 | 0.041 |

**Table 4: Non-Parametric Estimates of EUT Model Allowing Heterogeneity**

Data from treatments A, B and C (LL = -3718.6)

| Parameter | Point Estimate | Standard Error | *p*-value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| $\varkappa_{10}$ constant | 0.160 | 0.039 | <0.001 | 0.084 | 0.236 |
| $\varkappa_{10}$ pay1 | -0.063 | 0.060 | 0.298 | -0.180 | 0.055 |
| $\varkappa_{10}$ ra_idr | 0.026 | 0.031 | 0.401 | -0.035 | 0.087 |
| $\varkappa_{10}$ female | 0.069 | 0.030 | 0.021 | 0.010 | 0.128 |
| $\varkappa_{10}$ freshman | 0.055 | 0.048 | 0.256 | -0.040 | 0.149 |
| $\varkappa_{10}$ sophomore | 0.016 | 0.032 | 0.622 | -0.047 | 0.079 |
| $\varkappa_{10}$ senior | -0.010 | 0.031 | 0.739 | -0.070 | 0.050 |
| $\varkappa_{10}$ asian | -0.007 | 0.032 | 0.823 | -0.069 | 0.055 |
| $\varkappa_{10}$ white | 0.017 | 0.030 | 0.582 | -0.042 | 0.075 |
| $\varkappa_{10}$ gpaVHI | -0.021 | 0.030 | 0.479 | -0.081 | 0.038 |
| $\varkappa_{20}$ constant | 0.355 | 0.060 | <0.001 | 0.237 | 0.473 |
| $\varkappa_{20}$ pay1 | -0.052 | 0.083 | 0.533 | -0.215 | 0.112 |
| $\varkappa_{20}$ ra_idr | 0.039 | 0.045 | 0.390 | -0.050 | 0.127 |
| $\varkappa_{20}$ female | 0.096 | 0.039 | 0.014 | 0.019 | 0.173 |
| $\varkappa_{20}$ freshman | 0.172 | 0.064 | 0.007 | 0.047 | 0.298 |
| $\varkappa_{20}$ sophomore | 0.064 | 0.047 | 0.177 | -0.029 | 0.157 |
| $\varkappa_{20}$ senior | 0.005 | 0.046 | 0.918 | -0.086 | 0.095 |
| $\varkappa_{20}$ asian | -0.005 | 0.048 | 0.910 | -0.099 | 0.088 |
| $\varkappa_{20}$ white | 0.049 | 0.043 | 0.260 | -0.036 | 0.133 |
| $\varkappa_{20}$ gpaVHI | -0.050 | 0.045 | 0.262 | -0.138 | 0.038 |
| $\varkappa_{35}$ constant | 0.554 | 0.063 | <0.001 | 0.430 | 0.679 |
| $\varkappa_{35}$ pay1 | -0.077 | 0.114 | 0.497 | -0.301 | 0.146 |
| $\varkappa_{35}$ ra_idr | 0.012 | 0.046 | 0.786 | -0.077 | 0.102 |
| $\varkappa_{35}$ female | 0.101 | 0.041 | 0.013 | 0.022 | 0.181 |
| $\varkappa_{35}$ freshman | 0.177 | 0.058 | 0.002 | 0.064 | 0.290 |
| $\varkappa_{35}$ sophomore | 0.106 | 0.045 | 0.018 | 0.018 | 0.193 |
| $\varkappa_{35}$ senior | 0.029 | 0.047 | 0.541 | -0.063 | 0.121 |
| $\varkappa_{35}$ asian | -0.042 | 0.051 | 0.415 | -0.142 | 0.059 |
| $\varkappa_{35}$ white | 0.023 | 0.044 | 0.599 | -0.063 | 0.109 |
| $\varkappa_{35}$ gpaVHI | -0.041 | 0.049 | 0.403 | -0.137 | 0.055 |

## Table 5: Non-Parametric Estimates of DT and RDU Model Allowing Heterogeneity

Data from treatments A, B and C. Estimates of demographic variables and constant omitted

| Parameter | Point Estimate | Standard Error | *p*-value | 95% Confidence Interval | |
|---|---|---|---|---|---|
| *A. Dual Theory* (LL = -3749.6) | | | | | |
| $\varphi_{1/4}$ pay1 | 0.234 | 0.123 | 0.058 | -0.008 | 0.475 |
| $\varphi_{1/4}$ ra_idr | 0.009 | 0.045 | 0.846 | -0.080 | 0.097 |
| $\varphi_{1/2}$ pay1 | -0.093 | 0.078 | 0.232 | -0.245 | 0.059 |
| $\varphi_{1/2}$ ra_idr | -0.069 | 0.040 | 0.089 | -0.148 | 0.010 |
| $\varphi_{3/4}$ pay1 | 0.005 | 0.072 | 0.942 | -0.135 | 0.145 |
| $\varphi_{3/4}$ ra_idr | -0.032 | 0.042 | 0.444 | -0.115 | 0.050 |
| *B. Rank-Dependent Utility Theory* (LL = -3641.0) | | | | | |
| $\varkappa_{10}$ pay1 | -0.151 | 0.117 | 0.198 | -0.381 | 0.079 |
| $\varkappa_{10}$ ra_idr | 0.014 | 0.055 | 0.800 | -0.094 | 0.122 |
| $\varkappa_{20}$ pay1 | -0.179 | 0.143 | 0.211 | -0.459 | 0.101 |
| $\varkappa_{20}$ ra_idr | 0.007 | 0.073 | 0.920 | -0.135 | 0.150 |
| $\varkappa_{35}$ pay1 | -0.179 | 0.154 | 0.246 | -0.481 | 0.123 |
| $\varkappa_{35}$ ra_idr | -0.024 | 0.061 | 0.691 | -0.145 | 0.096 |
| $\varphi_{1/4}$ pay1 | 0.125 | 0.148 | 0.399 | -0.166 | 0.415 |
| $\varphi_{1/4}$ ra_idr | 0.023 | 0.060 | 0.709 | -0.096 | 0.141 |
| $\varphi_{1/2}$ pay1 | -0.192 | 0.133 | 0.148 | -0.453 | 0.068 |
| $\varphi_{1/2}$ ra_idr | -0.057 | 0.054 | 0.295 | -0.163 | 0.049 |
| $\varphi_{3/4}$ pay1 | -0.068 | 0.069 | 0.323 | -0.203 | 0.067 |
| $\varphi_{3/4}$ ra_idr | -0.019 | 0.033 | 0.575 | -0.083 | 0.046 |

**Table 6: Preference Estimates and Payment Protocols**

| Assumptions used to estimate risk preferences | Payment protocol A:<br>pay 1-in-1 | Payment protocol B:<br>pay 1-in-K |
|---|---|---|
| EUT | I | II |
| RDU *and* IA for payment protocol B | III | IV |
| RDU | V | VI |

Figure 4: Non-Parametric Estimates of Utility

Assuming EUT and normalized so that U($5)=0 and U($70)=1
Kernel density of predicted utility estimates for N=283 subjects
Data from treatments A, B and C

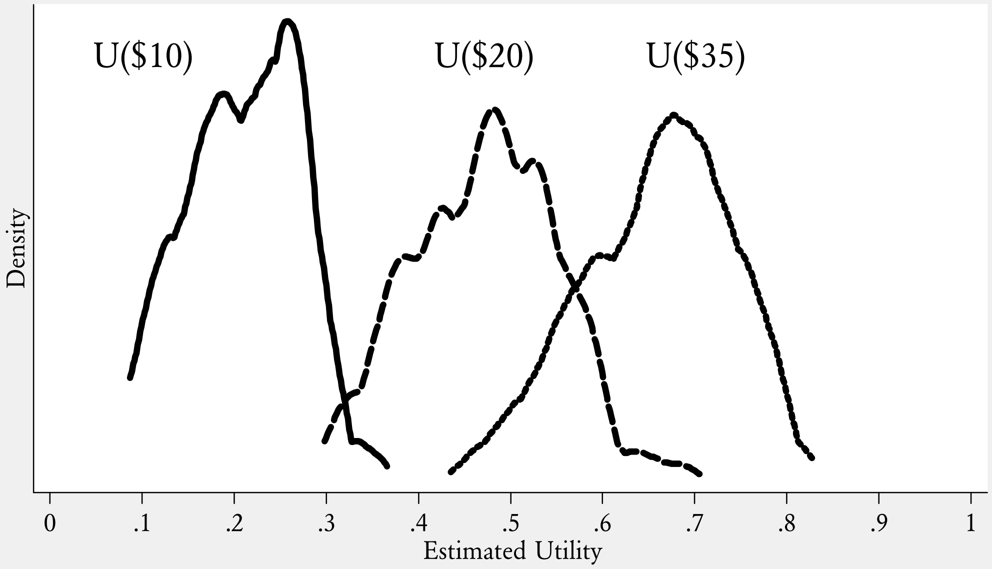


Figure 5: Non-Parametric Estimates of Probability Weights

Assuming RDU and normalized so that U($5)=0 and U($70)=1
Kernel density of predicted utility estimates for N=283 subjects
Data from treatments A, B and C

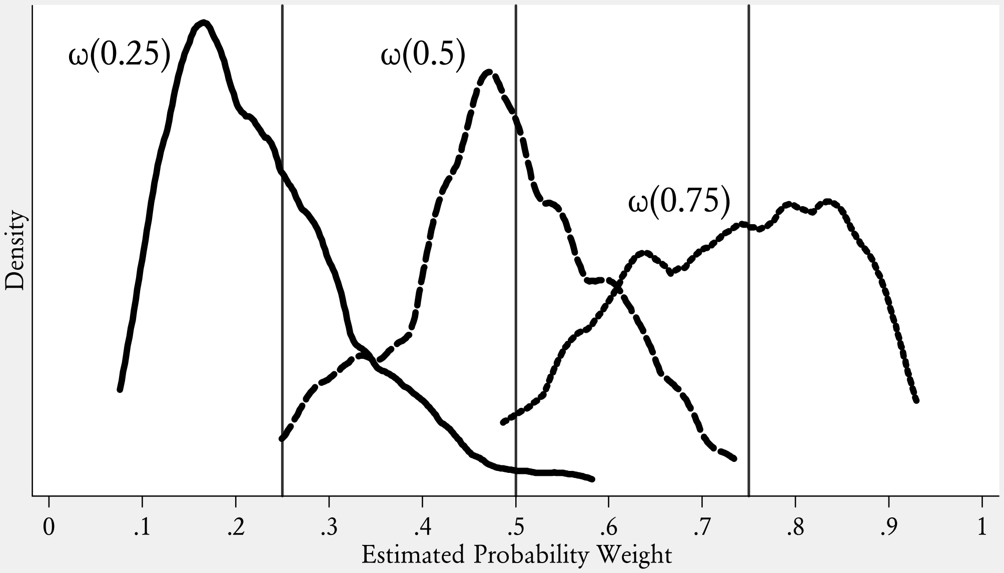


Figure 6: Bipolar Probability Weighting Functions for Dual Theory Models

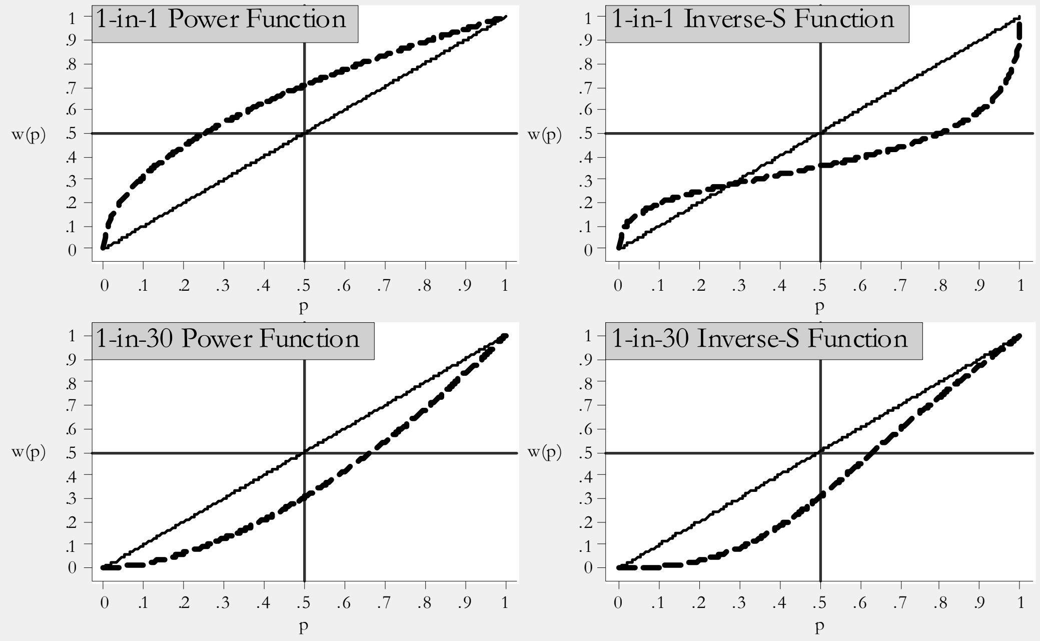
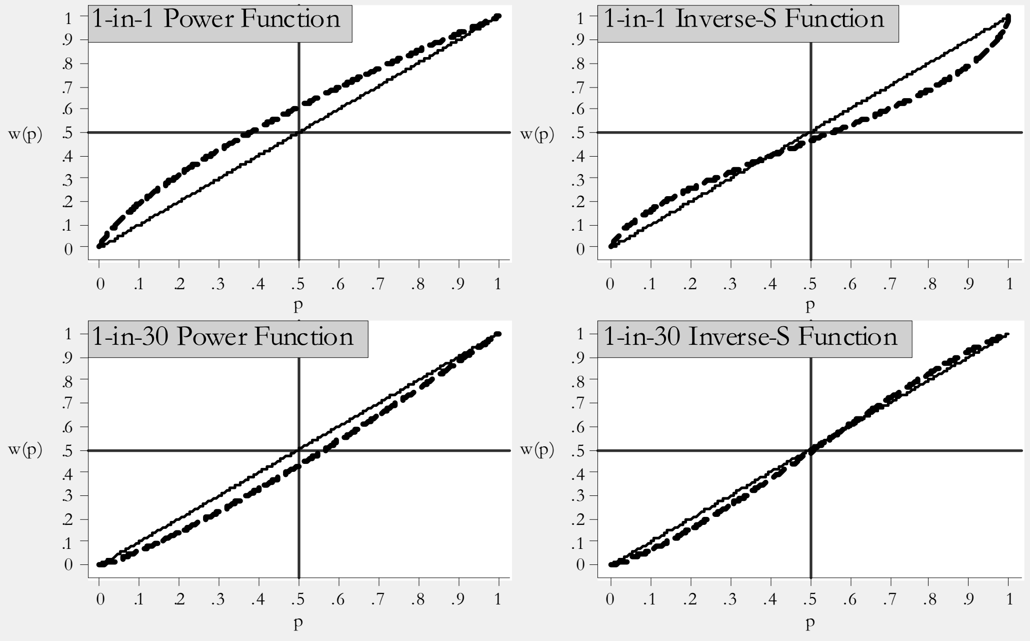


Figure 7: Bipolar Probability Weighting Functions for the Rank-Dependent Utility Model



## References

Andersen, Steffen; Harrison, Glenn W.; Hole, Arne Rise, Lau, Morten I., and Rutström, E. Elisabet, "Non-Linear Mixed Logit," *Working Paper 2010-07*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2010; forthcoming, *Theory and Decision*.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Elicitation Using Multiple Price Lists," *Experimental Economics*, 9(4), December 2006, 383-405.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 2008, 583-619.

Beattie, J., and Loomes, Graham, "The Impact of Incentives Upon Risky Choice Experiments," *Journal of Risk and Uncertainty*, 14, 1997, 149-162.

Binmore, Ken, *Does Game Theory Work? The Bargaining Challenge* (Cambridge, MA: MIT Press, 2007).

Camerer, Colin F., "An Experimental Test of Several Generalized Utility Theories," *Journal of Risk and Uncertainty*, 2, 1989, 61-104.

Camerer, Colin F., "Recent Tests of Generalizations of Expected Utility Theory," in W. Edwards (ed.), *Utility Theories: Measurements and Applications* (Boston: Kluwer, 1992).

Conlisk, John, "Three Variants on the Allais Example," *American Economic Review*, 79(3), June 1989, 392-407.

Cubitt, Robin P.; Starmer, Chris, and Sugden, Robert, "On the Validity of the Random Lottery Incentive System," *Experimental Economics*, 1(2), 1998, 115-131.

Cox, James C., and Epstein, Seth, "Preference Reversals Without the Independence Axiom," *American Economic Review*, 79(3), June 1989, 408-426.

Cox, James C.; Sadiraj, Vjollca, and Schmidt, Ulrich, "Paradoxes and Mechanisms for Choice under Risk," *Working Paper 2011-12*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2011.

Gonzalez, Richard, and Wu, George, "On the Shape of the Probability Weighting Function," *Cognitive Psychology*, 38, 1999, 129-166.

Grether, David M., and Plott, Charles R., "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, 69(4), September 1979, 623-638.

Guala, Fancesco, *The Methodology of Experimental Economics* (New York: Cambridge University Press, 2005).

Harless, David W., and Camerer, Colin F., "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62(6), November 1994, 1251-1289.

Harrison, Glenn W.; Johnson, Eric; McInnes, Melayne M., and Rutström, E. Elisabet, "Risk Aversion and Incentive Effects: Comment," *American Economic Review*, 95(3), June 2005, 897-901.

Harrison, Glenn W.; Johnson, Eric; McInnes, Melayne M., and Rutström, E. Elisabet, "Measurement With Experimental Controls," in M. Boumans (ed.), *Measurement in Economics: A Handbook* (San Diego, CA: Elsevier, 2007).

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Harrison, Glenn W., and Rutström, E. Elisabet, "Expected Utility *And* Prospect Theory: One Wedding and A Decent Funeral," *Experimental Economics*, 12(2), June 2009, 133-158.

Hey, John D., "Does Repetition Improve Consistency?" *Experimental Economics*, 4, 2001, 5-54.

Hey, John D., and Lee, Jinkwon, "Do Subjects Remember the Past?" *Applied Economics*, 37, 2005a, 9-8.

Hey, John D., and Lee, Jinkwon, "Do Subjects Separate (or Are They Sophisticated)?" *Experimental Economics*, 8, 2005b, 233-265.

Hey, John D., and Orme, Chris, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, 62(6), November 1994, 1291-1326.

Holt, Charles A., "Preference Reversals and the Independence Axiom," *American Economic Review*, 76, June 1986, 508-514.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), December 2002, 1644-1655.

Holt, Charles A., and Laury, Susan K., "Risk Aversion and Incentive Effects: New Data Without Order Effects," *American Economic Review*, 95(3), June 2005, 902-912.

Karni, Edi, and Safra, Zvi, "Preference Reversals and the Observability of Preferences by Experimental Methods," *Econometrica*, 55, 1987, 675-685.

Liang, K-Y., and Zeger, S.L., "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 1986, 13-22.

Loomes, Graham, and Sugden, Robert, "Testing Different Stochastic Specifications of Risky Choice," *Economica*, 65, 1998, 581-598.

McFadden, Daniel, and Train, Kenneth, "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 2000, 447-470.

Oehlert, Gary W., "A Note on the Delta Method," *The American Statistician*, 46(1), February 1992, 27-29.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3(4), 1982, 323-343.

Rogers, W. H., "Regression standard errors in clustered samples," *Stata Technical Bulletin*, 13, 1993, 19-23.

Saha, Atanu, "Expo-Power Utility: A Flexible Form for Absolute and Relative Risk Aversion," *American Journal of Agricultural Economics*, 75(4), November 1993, 905-913.

Samuelson, Paul A., "Probability, Utility, and the Independence Axiom," *Econometrica*, 20, 1952, 670-678.

Schmidt, Ulrich, and Zank, Horst, "A Simple Model of Cumulative Prospect Theory," *Journal of Mathematical Economics*, 45(3-4), March 2009, 308–319.

Segal, Uzi, "Does the Preference Reversal Phenomenon Necessarily Contradict the Independence Axiom?" *American Economic Review*, 78(1), March 1988, 233-236.

Segal, Uzi, "Two-Stage Lotteries Without the Reduction Axiom," *Econometrica*, 58(2), March 1990, 349-377.

Segal, Uzi, "The Independence Axiom Versus the Reduction Axiom: Must We Have Both?" in W. Edwards (ed.), *Utility Theories: Measurements and Applications* (Boston: Kluwer Academic Publishers, 1992).

Starmer, Chris, and Sugden, Robert, "Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation," *American Economic Review*, 81, 1991, 971-978.

Wilcox, Nathaniel T., "Lottery Choice: Incentives, Complexity, and Decision Time," *Economic Journal*, 103, 1993, 1397-1417.

Wilcox, Nathaniel T., "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison," in J. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Wilcox, Nathaniel T., "A Comparison of Three Probabilistic Models of Binary Discrete Choice Under Risk," *Working Paper*, Economic Science Institute, Chapman University, March 2010.

Wilcox, Nathaniel T., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice Under Risk," *Journal of Econometrics*, 162(1), May 2011, 89-104.

Williams, Rick L., "A Note on Robust Variance Estimation for Cluster-Correlated Data," *Biometrics*, 56, June 2000, 645-646.

Wooldridge, Jeffrey, "Cluster-Sample Methods in Applied Econometrics," *American Economic Review (Papers & Proceedings)*, 93, May 2003, 133-138.

Yaari, Menahem E., "The Dual Theory of Choice under Risk," *Econometrica*, 55(1), 1987, 95-115.

## Appendix A: Parameters of Experiments

### Table A1: Lotteries in Experiments

| | | Prizes | | | "Safe" Lottery Probabilities | | | "Risky" Lottery Probabilities | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pair | Context | Low | Middle | High | Low | Middle | High | Low | Middle | High | EV Safe | EV Risky |
| 1 | 1 | $5 | $10 | $20 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $10.00 | $16.25 |
| 2 | 1 | $5 | $10 | $20 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $8.75 | $12.50 |
| 3 | 1 | $5 | $10 | $20 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $10.00 | $12.50 |
| 4 | 1 | $5 | $10 | $20 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $7.50 | $8.75 |
| 5 | 1 | $5 | $10 | $20 | 0 | 1 | 0 | 0.25 | 0.5 | 0.25 | $10.00 | $11.25 |
| 6 | 1 | $5 | $10 | $20 | 0.25 | 0.5 | 0.25 | 0.5 | 0 | 0.5 | $11.25 | $12.50 |
| 7 | 1 | $5 | $10 | $20 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $15.00 | $16.25 |
| 8 | 1 | $5 | $10 | $20 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $12.50 | $12.50 |
| 9 | 1 | $5 | $10 | $20 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $8.75 | $8.75 |
| 10 | 1 | $5 | $10 | $20 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $10.00 | $8.75 |
| 11 | 2 | $5 | $10 | $35 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $10.00 | $20.00 |
| 12 | 2 | $5 | $10 | $35 | 0 | 0.75 | 0.25 | 0.25 | 0 | 0.75 | $16.25 | $27.50 |
| 13 | 2 | $5 | $10 | $35 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $8.75 | $12.50 |
| 14 | 2 | $5 | $10 | $35 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $22.50 | $27.50 |
| 15 | 2 | $5 | $10 | $35 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $16.25 | $20.00 |
| 16 | 2 | $5 | $10 | $35 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $10.00 | $12.50 |
| 17 | 3 | $5 | $10 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $8.75 | $37.50 |
| 18 | 3 | $5 | $10 | $70 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $10.00 | $37.50 |
| 19 | 3 | $5 | $10 | $70 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $7.50 | $21.25 |
| 20 | 3 | $5 | $10 | $70 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $10.00 | $21.25 |
| 21 | 4 | $5 | $20 | $35 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $20.00 | $27.50 |
| 22 | 4 | $5 | $20 | $35 | 0 | 0.75 | 0.25 | 0.25 | 0 | 0.75 | $23.75 | $27.50 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | 4 | $5 | $20 | $35 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $27.50 | $27.50 |
| 24 | 4 | $5 | $20 | $35 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $20.00 | $20.00 |
| 25 | 4 | $5 | $20 | $35 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $12.50 | $12.50 |
| 26 | 4 | $5 | $20 | $35 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $23.75 | $20.00 |
| 27 | 4 | $5 | $20 | $35 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $16.25 | $12.50 |
| 28 | 5 | $5 | $20 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $16.25 | $37.50 |
| 29 | 5 | $5 | $20 | $70 | 0 | 0.75 | 0.25 | 0.25 | 0 | 0.75 | $32.50 | $53.75 |
| 30 | 5 | $5 | $20 | $70 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $12.50 | $21.25 |
| 31 | 5 | $5 | $20 | $70 | 0.25 | 0.5 | 0.25 | 0.5 | 0 | 0.5 | $28.75 | $37.50 |
| 32 | 5 | $5 | $20 | $70 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $16.25 | $21.25 |
| 33 | 5 | $5 | $20 | $70 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $45.00 | $53.75 |
| 34 | 6 | $5 | $35 | $70 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $35.00 | $53.75 |
| 35 | 6 | $5 | $35 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $27.50 | $37.50 |
| 36 | 6 | $5 | $35 | $70 | 0 | 0.75 | 0.25 | 0.25 | 0 | 0.75 | $43.75 | $53.75 |
| 37 | 6 | $5 | $35 | $70 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $20.00 | $21.25 |
| 38 | 6 | $5 | $35 | $70 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $52.50 | $53.75 |
| 39 | 6 | $5 | $35 | $70 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $43.75 | $37.50 |
| 40 | 6 | $5 | $35 | $70 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $27.50 | $21.25 |
| 41 | 6 | $5 | $35 | $70 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $35.00 | $21.25 |
| 42 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $20.00 | $28.75 |
| 43 | 7 | $10 | $20 | $35 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $17.50 | $22.50 |
| 44 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.25 | 0.25 | 0.5 | $20.00 | $25.00 |
| 45 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $20.00 | $22.50 |
| 46 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.25 | 0.5 | 0.25 | $20.00 | $21.25 |
| 47 | 7 | $10 | $20 | $35 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $23.75 | $22.50 |
| 48 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.5 | 0.25 | 0.25 | $20.00 | $18.75 |
| 49 | 7 | $10 | $20 | $35 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $17.50 | $16.25 |
| 50 | 7 | $10 | $20 | $35 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $20.00 | $16.25 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 51 | 8 | $10 | $20 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $17.50 | $40.00 |
| 52 | 8 | $10 | $20 | $70 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $15.00 | $25.00 |
| 53 | 8 | $10 | $20 | $70 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $17.50 | $25.00 |
| 54 | 9 | $10 | $35 | $70 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $35.00 | $55.00 |
| 55 | 9 | $10 | $35 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $28.75 | $40.00 |
| 56 | 9 | $10 | $35 | $70 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $52.50 | $55.00 |
| 57 | 9 | $10 | $35 | $70 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $43.75 | $40.00 |
| 58 | 10 | $20 | $35 | $70 | 0 | 1 | 0 | 0.25 | 0 | 0.75 | $35.00 | $57.50 |
| 59 | 10 | $20 | $35 | $70 | 0.25 | 0.75 | 0 | 0.5 | 0 | 0.5 | $31.25 | $45.00 |
| 60 | 10 | $20 | $35 | $70 | 0 | 0.75 | 0.25 | 0.25 | 0 | 0.75 | $43.75 | $57.50 |
| 61 | 10 | $20 | $35 | $70 | 0 | 1 | 0 | 0.5 | 0 | 0.5 | $35.00 | $45.00 |
| 62 | 10 | $20 | $35 | $70 | 0.5 | 0.5 | 0 | 0.75 | 0 | 0.25 | $27.50 | $32.50 |
| 63 | 10 | $20 | $35 | $70 | 0 | 1 | 0 | 0.25 | 0.5 | 0.25 | $35.00 | $40.00 |
| 64 | 10 | $20 | $35 | $70 | 0.25 | 0.5 | 0.25 | 0.5 | 0 | 0.5 | $40.00 | $45.00 |
| 65 | 10 | $20 | $35 | $70 | 0 | 0.5 | 0.5 | 0.25 | 0 | 0.75 | $52.50 | $57.50 |
| 66 | 10 | $20 | $35 | $70 | 0 | 1 | 0 | 0.5 | 0.25 | 0.25 | $35.00 | $36.25 |
| 67 | 10 | $20 | $35 | $70 | 0.25 | 0.75 | 0 | 0.75 | 0 | 0.25 | $31.25 | $32.50 |
| 68 | 10 | $20 | $35 | $70 | 0 | 0.75 | 0.25 | 0.5 | 0 | 0.5 | $43.75 | $45.00 |
| 69 | 10 | $20 | $35 | $70 | 0 | 1 | 0 | 0.75 | 0 | 0.25 | $35.00 | $32.50 |

## Appendix B: Instructions

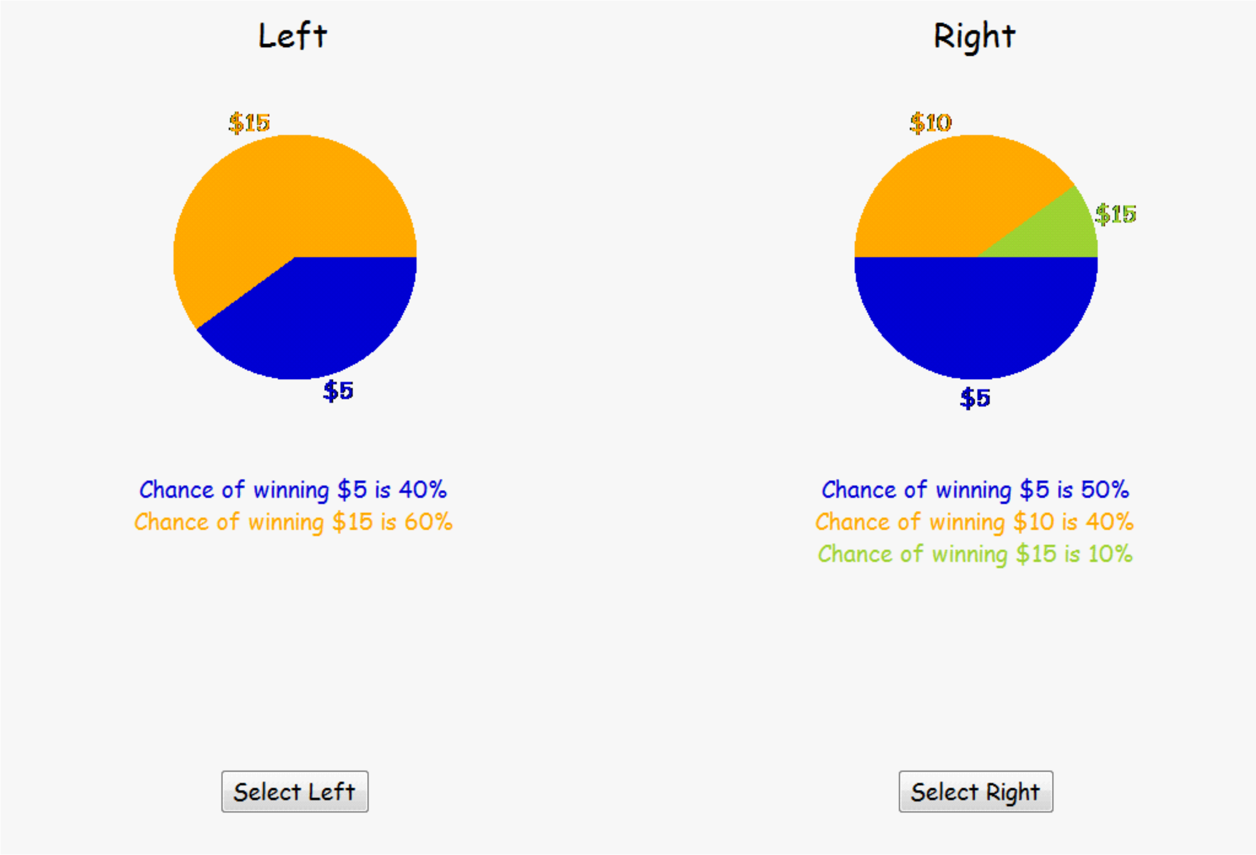
The original instructions used in all experiments are available on request.

*Treatment A: 1-in-1*

### Choices Over Risky Prospects

This is a task where you will choose between prospects with varying prizes and chances of winning. You will be presented with one pair of prospects where you will choose one of them. You should choose the prospect you prefer to play. You will actually get the chance to play the prospect you choose, and you will be paid according to the outcome of that prospect, so you should think carefully about which prospect you prefer.

Here is an example of what the computer display of a pair of prospects will look like.



Chance of winning $5 is 40%
Chance of winning $15 is 60%

Chance of winning $5 is 50%
Chance of winning $10 is 40%
Chance of winning $15 is 10%

The outcome of the prospects will be determined by the draw of a random number between 1 and 100. Each number between, and including, 1 and 100 is equally likely to occur. In fact, you will be able to draw the number yourself using two 10-sided dice.

In the above example the left prospect pays five dollars ($5) if the number drawn is between

1 and 40, and pays fifteen dollars ($15) if the number is between 41 and 100. The blue color in the pie chart corresponds to 40% of the area and illustrates the chances that the number drawn will be between 1 and 40 and your prize will be $5. The orange area in the pie chart corresponds to 60% of the area and illustrates the chances that the number drawn will be between 41 and 100 and your prize will be $15.

Now look at the pie in the chart on the right. It pays five dollars ($5) if the number drawn is between 1 and 50, ten dollars ($10) if the number is between 51 and 90, and fifteen dollars ($15) if the number is between 91 and 100. As with the prospect on the left, the pie slices represent the fraction of the possible numbers which yield each payoff. For example, the size of the $15 pie slice is 10% of the total pie.

The pair of prospects you choose from is shown on a screen on the computer. On that screen, you should indicate which prospect you prefer to play by clicking on one of the buttons beneath the prospects.

After you have made your choice, raise your hand and an experimenter will come over. It is certain that your one choice will be played out for real. You will roll the two ten-sided dice to determine the outcome of the prospect you chose.

For instance, suppose you picked the prospect on the left in the above example. If the random number was 37, you would win $5; if it was 93, you would get $15. If you picked the prospect on the right and drew the number 37, you would get $5; if it was 93, you would get $15.

Therefore, your payoff is determined by two things:

- by which prospect you selected, the left or the right; and
- by the outcome of that prospect when you roll the two 10-sided dice.

Which prospects you prefer is a matter of personal taste. The people next to you may be presented with a different prospect, and may have different preferences, so their responses should not matter to you. Please work silently, and make your choices by thinking carefully about the prospect you are presented with.

All payoffs are in cash, and are in addition to the $7.50 show-up fee that you receive just for being here. The only other task today is for you to answer some demographic questions. Your answers to those questions will not affect your payoffs.

*Treatment B: 1-in-30 Sequential*

## Choices Over Risky Prospects

This is a task where you will choose between prospects with varying prizes and chances of winning. You will be presented with a series of pairs of prospects where you will choose one of them. There are 30 pairs in the series. For each pair of prospects, you should choose the prospect you prefer to play. You will actually get the chance to play one of the prospects you choose, and you will be paid according to the outcome of that prospect, so you should think carefully about which prospect you prefer.

Here is an example of what the computer display of such a pair of prospects will look like.

SAME DISPLAY AS FOR TREATMENT A

The outcome of the prospects will be determined by the draw of a random number between 1 and 100. Each number between, and including, 1 and 100 is equally likely to occur. In fact, you will be able to draw the number yourself using two 10-sided dice.

In the above example the left prospect pays five dollars ($5) if the number drawn is between 1 and 40, and pays fifteen dollars ($15) if the number is between 41 and 100. The blue color in the pie chart corresponds to 40% of the area and illustrates the chances that the number drawn will be between 1 and 40 and your prize will be $5. The orange area in the pie chart corresponds to 60% of the area and illustrates the chances that the number drawn will be between 41 and 100 and your prize will be $15.

Now look at the pie in the chart on the right. It pays five dollars ($5) if the number drawn is between 1 and 50, ten dollars ($10) if the number is between 51 and 90, and fifteen dollars ($15) if the number is between 91 and 100. As with the prospect on the left, the pie slices represent the fraction of the possible numbers which yield each payoff. For example, the size of the $15 pie slice is 10% of the total pie.

Each pair of prospects is shown on a separate screen on the computer. On each screen, you should indicate which prospect you prefer to play by clicking on one of the buttons beneath the prospects.

After you have worked through all of the pairs of prospects, raise your hand and an experimenter will come over. You will then roll a 30-sided die to determine which pair of prospects will be played out. Since there is a chance that any of your 30 choices could be played out for real, you should approach each pair of prospects as if it is the one that you will play out. Finally, you will roll the two ten-sided dice to determine the outcome of the prospect you chose.

For instance, suppose you picked the prospect on the left in the above example. If the random number was 37, you would win $5; if it was 93, you would get $15. If you picked the prospect on the right and drew the number 37, you would get $5; if it was 93, you would get $15.

Therefore, your payoff is determined by three things:

- by which prospect you selected, the left or the right, for each of these 30 pairs;
- by which prospect pair is chosen to be played out in the series of 30 such pairs using the 30-sided die; and
- by the outcome of that prospect when you roll the two 10-sided dice.

Which prospects you prefer is a matter of personal taste. The people next to you may be presented with different prospects, and may have different preferences, so their responses should not matter to you. Please work silently, and make your choices by thinking carefully about each prospect.

All payoffs are in cash, and are in addition to the $7.50 show-up fee that you receive just for being here. The only other task today is for you to answer some demographic questions. Your answers to those questions will not affect your payoffs.

*Treatment C: 1-in-30 With an Additional Paid Task*

These instructions were identical to those for Treatment B, apart from the language changes in the final paragraph described in the text.

## Appendix C: Literature Review

Starmer and Sugden [1991], Beattie and Loomes [1997] and Cubitt, Starmer and Sugden [1998] have directly studied the Random Lottery Incentive Method, which of course relies on the validity of the IA. All of these studies consider direct and indirect violations of the IA.[28] Direct violations come from comparisons of choices 1-in-1 with 1-in-K payment procedures in the experiments, and indirect violations come from comparisons of choices that have a "trip-wire" prediction from EUT (and any decision-making model that assumes IA). These indirect violations are variants of the Allais [1953] phenomena known as "Common Ratio" effects and "Common Consequence" effects.[29]

Following **Cubitt, Starmer and Sugden** [1998; p.119], let a and b be monetary prizes, such that a>b>0. Consider the risky prospects

R1: {a, λ; 0, 1-λ} R2: {a, λp; b, 1-p; 0, (1-λ)p} R3: {a, λp; 0, 1-λp}

and the safe prospects

S1: {b, 1} S2: {b, 1} S3: (b,p; 0, 1-p},

[28] This is clearly recognized by Starmer and Sugden [1991; p.973]: "The success of the experiment depended on our finding systematic violations of expected-utility theory for real choices. Provided we found these, we would be able to test whether subjects behaved according to the reduction principle by investigating whether the same violations were found with the random-lottery design. The experiment also allows a second kind of test of the random-lottery design: if random-lottery experiments elicit true preferences, we should expect to find no significant difference between subjects' responses to the random lottery and real-choice designs." Their reduction principle applies the IA, and their "random lottery" design is what we refer to as a 1-in-K payment protocol.

[29] The first test of the common consequence form of the Allais Paradox using incentivized 1-in-1 choices, that do not assume IA, is due to Conlisk [1989]. He found striking evidence of no violations of IA. On the other hand, this result is not welcome in some circles. Cubitt, Starmer and Sugden [1989; p. 130] comment that these results are "... sometimes quoted as evidence that violations of EUT are less frequent in single choice than in random lottery designs. Conlisk investigated the Common Consequence effect using a single choice design. In each of the two relevant tasks, almost all subjects (26 out of 27 in one case, 24 out of 26 in the other) chose the riskier option. Clearly, this distribution of responses between riskier and safer choices is far too asymmetric for the experiment to be a satisfactory test for systematic deviations from EUT." The logic of the final sentence is hard to ascertain. Moreover, the evidence for the Common Consequence effect in *incentivized* 1-in-K choices is decidedly mixed: Burke, Carter, Gominiak and Ohl [1996] and Fan [2002] find no evidence of an EUT violation, whereas Starmer and Sugden [1991] do, as discussed below.

where $0<\lambda<1$ and $0<p\leq 1$. With these lotteries, the IA under EUT implies that preferences over R1 and S1 will be the same as preferences over R2 and S2 and also over R3 and S3. The Common Consequence effect is said to occur when one observes a greater fraction of risky choices over R3 and S3 than over R2 and S2, and a Common Ratio effect is said to occur when one observes a greater fraction of risky choices over R3 and S3 than over R1 and S1. These differences in the fractions of risky choices indirectly imply a statistically significant difference in risk preferences.

**Starmer and Sugden** [1991] present subjects with two pairs of lotteries, making up a Common Consequence test of EUT. In one 1-in-1 treatment 40 subjects were given each lottery pair to make a choice over, and over two 1-in-2 treatments 80 subjects were given the two lotteries to make a choice over. Their Common Consequence test between 1-in-1 choices shows evidence of a clear violation of the EUT prediction, and the IA. Using a *one-sided* Fisher Exact test, since there is an *a priori* prediction of direction, albeit from previously observed behavior from hypothetical tasks, we calculate a *p*-value of 0.021 on the prediction of the EUT hypothesis.[30] Their direct tests of the IA axiom, from comparisons of choices in the same lotteries across the 1-in-1 and 1-in-2 payment protocols, show mixed results. Since there is no prior hypothesis as to the direction of the effect of relying on the IA in the 1-in-2 treatment, it is appropriate in this case to use *two-sided* Fisher Exact tests of the hypothesis that the patterns of choice in each treatment are the same. For one lottery pair the *p*-value on this hypothesis is 0.23, and for the other lottery pair the *p*-value is 0.055. These data provide clear evidence for outright pessimism with respect to the IA: nothing bipolar here.

**Beattie and Loomes** [1997] examined 4 lottery choice tasks. The first 3 tasks involved a binary choice between two lotteries, and the fourth task involved the subject selecting one of four possible lotteries. For each of the 4 choice tasks they had a 1-in-1 treatment, and a 1-in-4 treatment,

[30] A comparable test, using the data from the 1-in-2 choices, also rejects the EUT hypothesis, in this case with a *p*-value of 0.018.

conducted on a between subjects basis. Sample sizes were 48, 47, 48 and 50 subjects in each of the 1-in-1 treatments for the 4 tasks, and 48 subjects for the 1-in-4 treatment. The *p*-values for the two-sided test of the 1-in-1 and 1-in-4 choices for the same lottery pairs are 0.42, 0.84, 0.77, and 0.058, respectively, for each choice task. Over all 4 tasks, the *p*-value is only 0.51. But there is a significant effect for one of the 4 tasks, and this is a task that is essentially the same as the popular method developed by Binswanger [1980]: subjects are offered an ordered set of choices that increase the average payoff while increasing variance. There is no direct evidence of an effect from the RLIM in the binary choice tasks.

On the other hand, there is clear, but indirect, evidence of a violation of the IA in binary choices by a comparison of two Common Ratio pairs in their set of choice tasks. For the 1-in-1 choices for these two pairs, a Fisher Exact test can reject the hypothesis of the same choices, as predicted under EUT, with a *p*-value of less than 0.001.[31] Taken with the direct evidence for these two pairs, when the IA is tested and *not* rejected via the RLIM payment procedure, these results provide striking support for the Bipolar Hypothesis advanced earlier.

**Cubitt, Starmer and Sugden** [1998] focus exclusively on Common Consequence and Common Ratio pairs of pairs, across three sets of experiments.

In the first set of experiments they compare 1-in-1 choices with 1-in-3 choices. Their comparison rests on subjects *not* having extreme risk-loving preferences over the other lotteries in the 1-in-3 treatment, but this is an *a priori* plausible assumption, and generally supported by their data. The two-sided *p*-values for these tests are 0.14 and 0.045, providing evidence against the application of the IA.

In the second set of experiments they compare 1-in-1 choices with 1-in-4 choices, with samples of 51 and 46 for the 1-in-1 treatments and 53 for the 1-in-4 treatment. Tests of the

[31] The same result occurs for the 1-in-4 pairs in this comparison.

hypothesis of the same choices in each leg of the Common Ratio pair of pairs have *p*-values of 0.84 and 0.16, implying that the direct test of the IA via the payment procedure had no significant effect. But in this case, in contrast to Beattie and Loomes [1997], there is no evidence for the Common Ratio effect in the 1-in-1 comparisons: the *p*-value on these choice patterns, spanning *both* legs of the Common Ratio pair of pairs, is 0.31. Of course, if EUT appears to be alive and well in terms of this familiar trip-wire test, then there is no *theoretical* expectation that the IA would be violated via the RLIM payment procedure.

In the third set of experiments with virtually the same Common Ratio pairs of pairs, and in fact the same lottery probabilities and prizes as the comparable Common Ratio pair of Beattie and Loomes [1997], they compare 1-in-1 choice patterns with 1-in-20 choice patterns. Sample sizes are 49 and 56 for the 1-in-1 treatments, and 97 for the 1-in-20 treatments. The direct tests of the IA via the RLIM procedure have *p*-values of 0.41 and 0.32, and the indirect Common Ratio test of the IA using the 1-in-1 treatment has a *p*-value of 0.10. At the risk of mixing psychiatric disorder metaphors, this is evidence for a Borderline Bipolar Hypothesis.

One common feature of virtually all of these studies is the use of a small value of K in the 1-in-K treatments. The rationale for this is explained by Cubitt, Starmer and Sugden [1998; p. 125]:

> First, as in the case of Experiments 1 and 2 [which used K=3 and K=4, respectively], we wanted to test the contamination hypothesis in a context in which we could expect the independence axiom to be violated. For Experiment 3 [which used K=20], however, we chose a somewhat different approach. In Experiments 1 and 2, the random lottery treatments involved only two tasks. In practical applications of the random lottery design, there are usually many tasks, and we wished to test the contamination hypothesis in such a setting. There are some reasons for expecting the extent of any bias in the random lottery design to depend on the number of tasks. On the one hand, it might be argued that, the more tasks there are in a random lottery experiment, the more likely subjects are to use the simplifying heuristic of treating each task in isolation. On the other hand, the more tasks there are, the more incentives are diluted; thus if bias is a product of dilution, its extent will increase with the number of tasks.

With the notable exception of the multiple price list design of Holt and Laury [2002], which uses

K=10, most applications of the RLIM do use large values of K. Hey and Orme [1994], for example, had K=100, Harrison and Rutström [2009] had K=60, Hey and Lee [2005a][2005b] have employed K=30 like us, and Wilcox [2010] and Hey [2001] bravely use K=300 and K=500, respectively, to obtain a rich data set for each individual subject.

Another common feature of all of these studies is that subjects were able to see all lotteries before having to make any choices. Starmer and Sugden [1998b] provided all lotteries in a booklet, and allowed subjects to make choices at any order they wanted. The specific lotteries of interest here were presented together on the same page of the booklet (their Figure 1, p. 975). Beattie and Loomes [1997; p.157] note that their 1-in-4 lotteries were "... presented together on a single sheet of paper." Cubitt, Starmer and Sugden [1998; p. 127] note that the software interface they used "... allowed subjects in all groups to backtrack at any point in the experiment, going back to previous tasks and changing their responses if they wished. After they had made all [...] responses they were reminded of this option. In this way [...] we gave subjects the opportunity to treat the whole experiment as a single decision problem if they so wished." They find that only one-third of subjects used this backtrack option, and they did not record if there were any changes in choices. Of course, the remaining two-thirds of subjects could still have viewed all tasks as one decision problem, making choices in later stages as a function of choices in earlier stages (e.g., "tend to pick safe options early, to ensure a certain payoff, then go for more risky options"). Cox, Sadiraj and Schmidt [2011] simply gave subjects all choices at the outset, each on one of K unbound sheets of paper, and then allowed them to enter choices on a computer interface that presented them sequentially.

Camerer [1989] studied the IA, literally in his design as an "afterthought." After subjects had made a number of choices using the RLIM, one was selected for payment, and the subject asked if he wanted to change the choice. Very few did, as Camerer [1989] notes:

> Only two of 80 subjects did change. Therefore, either the independence axiom holds or subjects exhibit an isolation effect. Since the data below suggest that

> independence is often violated, we must conclude that there is an isolation effect. This is puzzling for theorists, but comforting for experimenters because it implies that allowing subjects to play some randomly chosen gambles generates meaningful responses for all gambles.

Our design provides a more direct test along these lines, without the 1-in-1 choice being an afterthought where the subject might feel compelled to stick with the initial choice rather than appear confused or capricious to the experimenter

## Additional References

Allais, Maurice, "Le Comportement de L'homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de L'école Américaine," *Econometrica*, 21, 1953, 503-546.

Binswanger, Hans P., "Attitudes Toward Risk: Experimental Measurement in Rural India," *American Journal of Agricultural Economics*, 62, August 1980, 395-407.

Burke, Michael S.; Carter, John R.; Gominiak, Robert D., and Ohl, Daniel F., "An Experimental Note on the Allais Paradox and Monetary Incentives," *Empirical Economics*, 21, 1996, 617-632.

Fan, Chinn-Ping, "Allais Paradox in the Small," *Journal of Economic Behavior & Organization*, 49, 2002, 411-421.