Theory Dec. (2014) 77:423–438
DOI 10.1007/s11238-014-9447-y

Experimental payment protocols and the Bipolar Behaviorist

Glenn W. Harrison · J. Todd Swarthout

Published online: 12 June 2014

Abstract If someone claims that individuals behave as if they *violate* the independence axiom (IA) when making decisions over simple lotteries, it is invariably on the basis of experiments and theories that must *assume* the IA through the use of the random lottery incentive mechanism (RLIM). We refer to someone who holds this view as a Bipolar Behaviorist, exhibiting pessimism about the axiom when it comes to characterizing how individuals directly evaluate two lotteries in a binary choice task, but optimism about the axiom when it comes to characterizing how individuals evaluate multiple lotteries that make up the incentive structure for a multiple-task experiment. We reject the hypothesis about subject behavior underlying this stance: we find that preferences estimated with a model that assumes violations of the IA are significantly affected when one elicits choices with *procedures* that require the independence assumption, as compared to choices elicited with *procedures* that do not require the assumption. The upshot is that one cannot consistently estimate popular models that relax the IA using data from experiments that assume the validity of the RLIM.

Electronic supplementary material The online version of this article (doi:10.1007/s11238-014-9447-y) contains supplementary material, which is available to authorized users.

G. W. Harrison
Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta, GA, USA
e-mail: gharrison@gsu.edu

J. T. Swarthout (✉)
Department of Economics and Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA, USA
e-mail: swarthout@gsu.edu

Keywords Experiment · Independence axiom · Payment protocols · Random lottery incentive mechanism

1 Introduction

The independence axiom (IA) plays a central role in most formal statements of expected utility theory (EUT), as well as popular alternative models of decision-making under objective or subjective risk. One such alternative is rank-dependent utility (RDU) theory (Quiggin 1982), which assumes that the IA is invalid in a certain way. The axiom also plays a central role in virtually every experiment used to characterize the way in which risk preferences deviate from EUT, through the use of the random lottery incentive mechanism (RLIM). For example, if someone claims that individuals behave as if they "probability weight" outcomes, and hence *violate* the IA, it is almost always on the basis of experiments and theories that *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 bothered theorists confronted with experimental data, and there have been responses from theorists and experimentalists. The primary theoretical response has been to argue that there is a way to write out a model of decision-making under risk that allows one to relax the IA but to still allow risk preferences to deviate from EUT predictions and for RLIM to be valid. In effect, to argue an existence proof: even though the inconsistency is real for the most popular alternatives to EUT, there exists a formal alternative to EUT where there is no inconsistency. We discuss this theoretical response later in Sect. 6. The primary experimental response has been to develop some direct tests and some ingenious designs intended to trap the IA under some circumstances.[1] But these direct and indirect experimental 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 rejecting it when characterizing the risk preferences exhibited in the same experiment using the standard alternatives to EUT.[2] Those characterizations seem to show evidence of rank-dependent probability weighting, when that very evidence calls into question a maintained assumption of the payment protocol used to generate the evidence. We refer to someone who holds this view as a Bipolar[3] Behaviorist, exhibiting pessimism about

[1] Cubitt et al. (1998, p. 119) explain the logic of the *indirect* tests: "Our strategy is to take as the maintained hypothesis that the random lottery design is unbiased. We test this hypothesis in situations in which we have a priori expectations that individuals' preferences violate the IA in ways which, if the contamination hypothesis were true, would induce observable biases." All studies using indirect tests of this kind, which of course rest on premisses that might be false, also report direct tests.

[2] An illustrative sample of studies estimating or testing models of RDU, for instance, without questioning the inconsistent use of RLIM include Camerer (1989), Starmer (1992), Camerer and Ho (1994), Hey and Orme (1994), Wakker et al. (1994), and Harrison and Rutström (2008, 2009).

[3] Our use of the term "bipolar" is to convey diametrically opposed views, and not to imply mental illness. Of course, one could instead view the term as a colorful metaphor, with a little bite to it. Indeed, we openly admit to such bipolar attitudes at times in our own research.

the IA when it comes to characterizing how individuals directly evaluate two lotteries in a binary choice task, but optimism 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 in individual risky choice experiments involves a subject making $K > 1$ binary choices over objective lotteries, 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 et al. (1998), and Cox et al. (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.[4] 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.

Testing the manner in which the IA interacts with payment protocols used to collect data on observed choice behavior is complicated by the possibility that the reduction of compound lotteries (ROCL) axiom may be invalid behaviorally. This possibility lies at the heart of the theoretical response to the hypothesis about subject behavior underlying the stance of the Bipolar Behaviorist. If the objects of choice are themselves compound lotteries, as is the case in some famous experimental tasks such as the "preference reversal" experiments of Grether and Plott (1979), then one has to take a stand on the validity of ROCL anyway.[5] But if the objects of choice are simple lotteries, as here, then one can test the *implications of RLIM for the standard alternatives to EUT* without taking a position on the validity of ROCL.[6]

We offer direct tests of the effect of payment protocols on preferences for risk in general, and the evidence for probability weighting in particular. We do find statistically significant evidence for a difference in estimated risk preferences deriving from the use of different payment protocols and experimental tasks.

Using choices over simple lotteries, 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 estab-

[4] 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 with real monetary consequences, 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.

[5] 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 Compound IA and ROCL, hence Mixture IA. Holt (1986) and Karni and Safra (1987) showed that if Compound IA 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 Compound IA and transitivity was assumed, that one might also still observe choices that suggest "preference reversals."

[6] To anticipate the language explained in Sect. 2, we are then directly testing the Compound IA and not testing the Mixture IA (the Mixture IA implies the validity of the Compound IA *and* ROCL). Starmer and Sugden (1991), in fact, test the RLIM payment protocol by testing for the validity of ROCL.

lishes 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. It is not the existence of evidence for probability weighting that is the issue, it is the fact that the nature of probability weighting differs in the 1-in-1 *versus* 1-in-K protocol.

In order to justify the use of the 1-in-K payment protocol, many studies appeal to the "isolation effect." This effect is often presented as a *behavioral* assertion that a subject views each choice in an experiment as independent of other choices in the experiment. When stated formally, the isolation effect is often expressed the same as the IA, and is indeed exactly the same as the IA in our choice context. We recognize that the isolation effect is often invoked informally as "an empirical matter," with either an appeal to prior evidence[7] or simply a conjecture that the isolation effect is a reasonable description human behavior. Given limited empirical support from prior studies and the tautology of support via conjecture, we present an experiment which provides a new test of the isolation effect.

In Sect. 2 we describe the theoretical constructs needed for our design, in particular the various axioms that are at issue. In Sect. 3 we present our experimental design, which allows comparison of risk preferences obtained from choice tasks over simple lotteries 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 Sect. 4 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 only one choice to make, *necessitating* the pooling of choices across different 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 Sect. 5 we examine the data from our experiments, and present the econometric analy-

[7] Given the ubiquity of the RLIM in the laboratory, surely past studies have definitively verified the empirical validity of the isolation effect? Unfortunately, this is not the case. A few studies have focused on this issue, but conclusions differ and the verdict is still out on whether use of the RLIM biases behavior. All of these studies consider direct and indirect violations of the IA underlying the RLIM. Direct violations come from comparisons of choices 1-in-1 with 1-in-K payment procedures in the experiments, exactly as in our design, 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 phenomena known as "Common Ratio" effects and "Common Consequence" effects. Focusing just on the direct tests, comparable to our design, we find mixed results in the previous literature. Starmer and Sugden (1991) find the same pattern of choices in one 1-in-1 versus 1-in-2 comparison, and a different pattern in another comparison. Their design only had two pairs of choices, so $K = 2$; indeed, all of the previous studies had a small K. Beattie and Loomes (1997) used $K = 4$, and found no difference between the 1-in-1 and 1-in-4 choices over three pairs of binary lottery choices. They did find a difference between the 1-in-1 and 1-in-4 choices over the single "multiple lottery choice" task, patterned after Binswanger (1980). Online Appendix C contains a more complete summary of the previous literature.

sis of hypotheses. In Sect. 6 we draw some general implications of our results, and in Sect. 7 offer general conclusions.

2 Theory

2.1 Basic axioms

Following Segal (1987, 1988, 1990, 1992), we distinguish between three axioms defined over objective lotteries.[8] In words, the 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 $\sim$ express indifference. Then the ROCL axiom says that $A \sim X$ 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 $1/2 \times 1/2 = 1/4$, \$4 with probability $1/2 \times 1/2 = 1/4$, and nothing with probability 1/2. 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 $(1-\alpha)$, and B is the compound lottery giving the simple lottery Y with probability α and the simple lottery Z with probability $(1-\alpha)$, then $A \succ B$ iff $X \succ Y \, \forall \alpha \in (0, 1)$. So the construction of

[8] In general we focus throughout on lotteries defined over objective probabilities. Remarkably, Bade (2011) shows that the 1-in-K payment protocol does not immediately generate inferential problems for *some* models of choices over lotteries defined over ambiguous acts, such as the maxmin expected utility model. However, for the popular "smooth" models of ambiguity aversion, the 1-in-K protocol does generate problems if the smooth model is compatible with the notion of stochastic independence.

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. For example, Segal (1992, p. 170) defines the CIA by assuming that the second-stage lotteries are replaced by their certainty-equivalent according to some (possibly non-EUT) model, "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 and hence state the ROCL axiom, by construction, since the informational bridge to that calculation has been burnt.

Finally, the *mixture independence axiom* (MIA) says that the preference ordering of two simple lotteries must be the same as the two actuarially-equivalent simple lotteries derived from the two compound lotteries formed by combining a common outcome with one of the original simple lotteries, where the common outcome has the same (compound lottery) probability. That is, $X \succ Y$ iff the actuarially-equivalent simple lottery of $\alpha X + (1 - \alpha)Z$ is strictly preferred to the actuarially-equivalent simple lottery of $\alpha Y + (1 - \alpha)Z, \forall \alpha \in (0, 1]$. 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 reason these three axioms are important for the evaluation of alternatives to EUT 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 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.

2.2 Experimental payment protocols

Turning now to experimental procedures, the most popular payment protocol used in individual choice experiments assumes the validity of the CIA. This payment protocol is called the RLIM. It entails the subject making K choices 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 kth task to be played out uses a discrete 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 experiment 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 $Z = \text{U}(\$0)$ and $(1 - \alpha) = (1 - (1/k))$, for the utility function $U(\cdot)$. Hence,

under the CIA, the preference ordering of X and Y is independent of all of the choices in the other tasks (Holt 1986).

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. Further, experimenters in this area are wont to ignore the implications this assumption has on power calculations, and so the results from small-sample studies which assume homogeneity of preferences can be questionable due to low power.

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. Although this sounds like an econometric assumption, and it certainly has statistical implications, it is as much a matter of (operationally meaningful) theory as formal statements of the CIA, ROCL, and MIA.

3 Experiment

3.1 Basic design issues

Our 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. A standard, fixed show-up fee, in our case $7.50, was paid to every subject independently of their lottery choices.

3.2 Specific design

Table 1 summarizes our experimental design. In *treatment A*, each subject undertakes one 1-in-1 binary choice, where each subject's single lottery pair is drawn at random from a set of 69 lottery pairs shown in online Appendix A. These lottery pairs span five monetary prize amounts, $5, $10, $20, $35, and $70, and five probabilities, 0, 1/4, 1/2, 3/4, and 1. The prizes are combined in ten "contexts," defined as a particular triple of prizes.[9] They are based on a battery of lottery pairs developed by Wilcox (2010)

[9] 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).

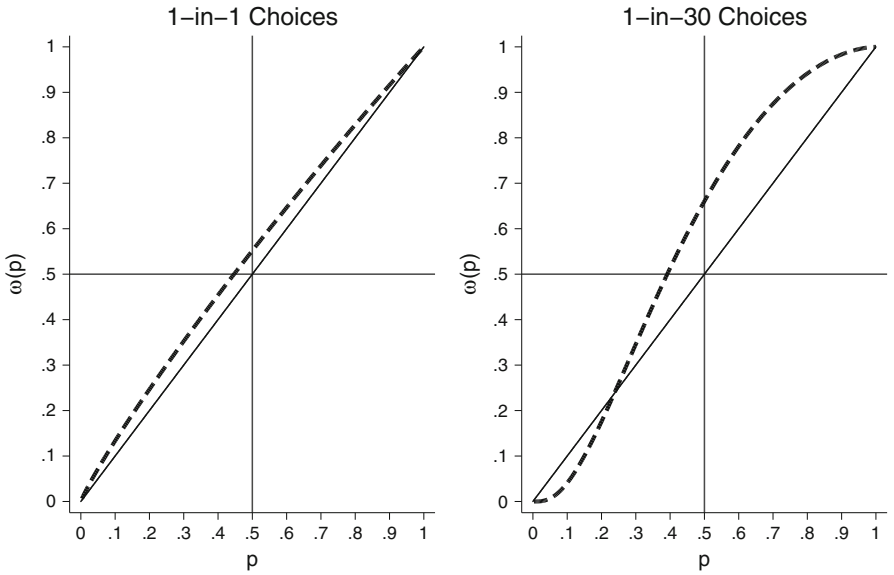

Fig. 1 Bipolar probability weighting functions for rank-dependent utility models

Table 1 Experimental design

Treatment	Subjects	Choices
A. 1-in-1	75	75
B. 1-in-30[a]	208	6240

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 B, subjects were told that there would be no other salient task in the experiment

[a] For 171 subjects, and 5130 choices, there was an additional task after the 30 lottery choices. This was a time-discounting choice, and these subjects were told at the outset that there could be additional salient tasks

for the purpose of robust estimation of EUT and RDU models.[10] These lotteries also contain a number of pairs in 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 in terms of internal validity, against which the other payment protocols are to be evaluated.

[10] 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.

In *treatment B* we move to the 1-in-30 case, which is typical of the usual risk elicitation setting. In some cases there was an additional task after the 30 lottery choices, and in some cases there was no other task. In the former case the instructions raised the possibility of a future task for payment.[11] In the latter case 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. We later test for an effect of there being an extra task and find none, so pool these minor variants into one treatment B.

Every random event determining payouts was generated by the rolling of one or more dice. These dice were illustrated visually during the reading of the instructions,[12] and each subject rolled their own dice.

3.3 Why not just look at raw choice patterns?

Prior tests looked directly at choice patterns. In contrast, 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 one 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?

3.3.1 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 account for "mistakes" by implementing "trembles" (e.g., Conlisk (1989, Appendix I) and Harless and Camerer (1994)). Such trembles are agnostic about the

[11] To be precise, when there *was* an extra task 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 all other cases 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."

[12] A video camera captured images at the front table and broadcasted the images to displays throughout the lab. In addition to a large projection screen in the front, there are three wide-screen TV displays spread throughout the lab so that every cubicle has a clear view of the images.

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 structural manner (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.[13] In the first instance behavioral errors are the business of theorists, not econometricians, as stressed by Wilcox (2008, 2011).

3.3.2 Do choice patterns use all available information?

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 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 greater, by ε, than the EU for the right lottery. Another lottery pair, similarly evaluated at those same parameter values, implies an EU for the left lottery that is greater, by much more than ε, 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). Structural estimation also typically entails 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, and the only (statistically) efficient estimators use all information. We

[13] An additional subtlety 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.

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. But we 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.

3.3.3 Are the stimuli representative?

Comparison of choice patterns from a paradox test with two pairs of lotteries may support or refute the theory under consideration, but how confident are we that the result is representative of choices over all possible 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 [e.g., Camerer (1989, 1992) on the importance of "boundary effects"]. 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.

3.3.4 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 an observable characteristic or two, such as assuming homogeneity within samples of men and 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 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 et al. 2012).

3.4 Data

A total of 283 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 over our treatments. Instructions for all treatments are presented in online 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 in every session. Every subject also completed a demographic survey covering standard characteristics. All subjects were paid in cash at the end of each session.

4 Econometrics

Our interest is in making inferences about the latent risk preferences underlying observed choice behavior. The estimation approach is 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 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.

Online Appendix D provides the specifications of our econometric model, and generally follows Harrison and Rutström (2008).

5 Results

We consider behavior observed under treatments A and B, and evaluate the hypothesis that risk preferences are the same across the treatments. The initial estimates assume 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 using observable demographic characteristics. Finally, we consider parametric estimates, which may be more intuitive and familiar.

Online Appendix D contains the detailed estimates of all structural models. If we use a non-parametric specification, we cannot find any statistically significant effect of the payment protocols when we assume EUT. This is true where we assume preference homogeneity (p value $= 0.78$) or allow for preference heterogeneity (p value $= 0.81$). But we do find significant differences when we assume RDU. When we assume preference homogeneity we observe a statistically significant effect at a p value of 0.078. With preference heterogeneity the p value across all parameters, those characterizing utility and probability weighting, is only 0.15; but if we evaluate the effect for just the probability weighting parameters the p value is 0.07.

Turning to parametric estimates, since they are familiar and easier to visualize, we again find no effect of payment protocols under EUT (p value $= 0.98$). In this case we use the Expo-Power utility function proposed by Saha (1993), and implemented by Holt and Laury (2002), since it allows for non-constant relative risk aversion. We find a p value of 0.06 with the RDU model, across all parameters, and note again that the culprit appears to be the probability weighting behavior (the p value on just

those parameters is 0.07). In this case we again use the Expo-Power utility function, as well as the flexible two-parameter Prelec (1998) probability weighting function These results assume preference heterogeneity across individuals Fig. 1 shows the difference between RDU probability weighting functions when moving from the 1-in-1 payment protocol to the 1-in-K payment protocol. The differences are striking, quantitatively *and* qualitatively.

6 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 et al. (2011). There are no known, or obvious, payment protocols that can be used for RDU and cumulative prospect theory (CPT).

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 of which we are aware. This is not to say that EUT is valid, just that tests of the validity of specific alternatives rest on a maintained assumption that appears to be false.[14]

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 or robust anomalies should be encouraged to do this, 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.

A fifth implication is to consider more rigorously the learning behavior that might change behavior towards lottery choices such as these.[15] 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

[14] Our results suggest a 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 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 in the interests of efficient use of an experimental budget. And the alternative to the tradeoff is simple enough: replicate every anomaly using the 1-in-1 payment protocol, as in the non-hypothetical experiments of Conlisk (1989).

[15] 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 of strategic behavior in games, they are general. These arguments also suggest that a 1-in-1 payment protocol might not be the Gold Standard if one is interested in external validity, whatever its role in terms of internal validity.

to what should be compared to what, and does not provide a metric for defining when learning is finished.[16]

A sixth implication is to formally model the effects of treating behavior as if generated by portfolio formation for the experiment as a whole. In effect, this is the implication of the theoretical "Recursive RDU" response to the problems posed by the RLIM payment procedure.[17] Even though nobody has ever estimated or empirically implemented such a model, the fact that one can formally write one out is regarded as enough to validate the use of the RLIM. The Recursive RDU model is not the standard RDU model, and it is the latter model that is the one that appears in all empirical work. It is apparent that the estimation of a Recursive RDU model for $K \ggg 0$, as in our experimental design, is infeasible.

A final implication is to just be honest when presenting experimental findings on RDU and CPT models about the assumed neutrality of the experimental payment protocol. In effect this is just saying that there might be two independence axioms at work: one for the evaluation of a given lottery in a binary choice setting, and another one for an evaluation of sets of choices in 1-in-K settings. If one estimates RDU and CPT models with a 1-in-K protocol one might claim to be allowing the first axiom to be relaxed while maintaining the second. It is logically possible for the latter axiom to be empirically false while the former axiom is empirically true. In the absence of better alternatives, we do this in our own ongoing research using 1-in-K protocols.

7 Conclusions

We have demonstrated that estimated non-EUT preferences are sensitive to whether behavior is elicited with the RLIM or instead with a "one-shot" design. We do so by considering a simple and popular alternative to EUT: a standard RDU model that relaxes the CIA. Of course, this RDU specification is identical to the "gain frame" part of CPT, so CPT inherits all of the issues we have raised with respect to inferring risk preferences with RDU. Experimental economists should take this result seriously, and recognize the apparent problem of inferring preferences through use of the RLIM and treating these results "as if" they are the same as those from a 1-in-1 scenario.

Acknowledgments We are grateful to Jim Cox, Jimmy Martínez, John Quiggin, Elisabet Rutström, Vjollca Sadiraj, Ulrich Schmidt, Uzi Segal, and Nathaniel Wilcox for helpful discussions.

[16] 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 differences in behavior 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 a subject 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 someone can ever define what "enough" means.

[17] In fact, one of the earliest statements of this Recursive RDU model by Segal (1988) was in the context of offering an explanation of preference reversals behavior being logically consistent with the validity of the Compound IA. It is therefore theoretically possible that the RLIM procedure is suspect but the Compound IA is valid, so that one does not have to take a bipolar stance about the Compound IA after all. On the other hand, evidence that payment protocols do affect behavior is then evidence against the Mixture IA, so the hypothesis to be tested to support the stance of the Bipolar Behaviorist is then with respect to ROCL.

References

Andersen, S., Harrison, G. W., Hole, A. R., Lau, M. I., & Rutström, E. E. (2012). Non-linear mixed logit. *Theory and Decision*, *73*, 77–96.

Bade, S. (2011). Independent randomization devices and the elicitation of ambiguity averse preferences. Working Paper, Max Planck Institute for Research on Collective Goods, Bonn.

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

Binmore, K. (2007). *Does game theory work? The bargaining challenge*. Cambridge, MA: MIT Press.

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

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

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Ed.), *Utility theories: Measurements and applications*. Boston, MA: Kluwer.

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

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

Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, *1*(2), 115–131.

Cox, J. C. Sadiraj, V. & Schmidt, U. (2011). 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 (revised 2014); forthcoming, Experimental Economics.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, *69*(4), 623–638.

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

Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. In J. C. Cox & G. W. Harrison (Eds.), *Risk aversion in experiments* (Vol. 12). Bingley: Emerald, Research in Experimental Economics.

Harrison, G. W., & Rutström, E. E. (2009). Expected utility and prospect theory: One wedding and a decent funeral. *Experimental Economics*, *12*(2), 133–158.

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

Holt, C. A. (1986). Preference reversals and the independence axiom. *American Economic Review*, *76*, 508–514.

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

Karni, E., & Safra, Z. (1987). Preference reversals and the observability of preferences by experimental methods. *Econometrica*, *55*, 675–685.

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

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

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

Samuelson, P. A. (1952). Probability, utility, and the independence axiom. *Econometrica*, *20*, 670–678.

Segal, U. (1987). The Ellsberg paradox and risk aversion: An anticipated utility approach. *International Economic Review*, *28*, 145–154.

Segal, U. (1988). Does the preference reversal phenomenon necessarily contradict the independence axiom? *American Economic Review*, *78*(1), 233–236.

Segal, U. (1990). Two-stage lotteries without the reduction axiom. *Econometrica*, *58*(2), 349–377.

Segal, U. (1992). The independence axiom versus the reduction axiom: Must we have both? In W. Edwards (Ed.), *Utility theories: measurements and applications*. Boston, MA: Kluwer.

Starmer, C. (1992). Testing new theories of choice under uncertainty using the common consequence effect. *Review of Economic Studies*, *59*, 813–830.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332–382.

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

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

Wilcox, N. T. (1993). Lottery choice: Incentives, complexity, and decision time. *Economic Journal*, *103*, 1397–1417.

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

Wilcox, N. T. (2010). A comparison of three probabilistic models of binary discrete choice under risk. Working Paper, Economic Science Institute, Chapman University.

Wilcox, N. T. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89–104.