The Methodologies of Behavioral Econometrics

by

Glenn W. Harrison †

March 2017

Forthcoming in Michiru Nagatsu and Attilia Ruzzene (eds.),
Philosophy and Interdisciplinary Social Science: A Dialogue
(London: Bloomsbury, 2017 forthcoming)

ABSTRACT.

Behavioral econometrics is one part of a methodological trinity that includes theory, data collection and econometrics. Sometimes, on a good methodological day, there are few demands on the econometrician and the data can just be described and summarized, or elementary statistical tests applied. More often, latent structures from theory have to be estimated, and data collection has to be designed to allow identification and powerful estimation to be undertaken. In these cases, more common than many believe, appropriate econometric methods have to be used. The various methodologies of behavioral econometrics are reviewed, with illustrative case studies that showcase appropriate and inappropriate states of the art. Significant methodological challenges are identified.

† Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. E-mail: gharrison@gsu.edu.

Table of Contents

There is an essential methodological connection between theory, the collection of data, and econometrics. Theory can consist of simple or complex hypotheses, the comparative static predictions of structural theory, or the latent components of that structural theory itself. The collection of data might be as simple as using pre-existing data, the development of survey instruments, or the design of controlled experiments. In many cases of interest to behavioral econometrics, the data consist of controlled lab or field experiments.

Most of the behavioral data encountered from controlled experiments is relatively easy to evaluate with known econometric methods. Section 1 reviews a range of methods for different types of experiments. In some cases simple, "agnostic" statistical modeling is appropriate, since the experiment "does the work of theory" for the analyst, by controlling for treatments and potential confounds. In other cases more nuanced structural modeling is appropriate, and we now have a rich array of econometric tools that have been applied and adapted to the specific needs of behavioral economists.

On the other hand, there is a methodological tension in the air, with widely differing statistical and econometric methods being applied to what look to be the same type of inference. There are two major problems with the methodologies applied in behavioral econometrics. One is a separation of skills, with statistical and econometric methods just being appended as an afterthought.[1] The other is the simple mis-application of econometric methods, akin to the story that Leamer [1978; p. vi] told of his teachers preaching econometric cleanliness in the classrooms on the top floor of a building, and then descending into the basement to build large-scale macro-econometric models that violated almost every tenet from the classroom.

These problems are anticipated in the review in section 1, and the illustrations of two methodological innovations in sections 2 and 3. They are directly illustrated with some real case studies in sections 4, 5 and 6, with emphasis on the measurement and analysis of risk preferences. Section 4

[1] Adam Smith preached the virtues of a division of labor, but only under the assumption that trade occurred to allow the efficiency gains to be realized.

considers the empirical evidence for Cumulative Prospect Theory (CPT), and asks if anyone is even reading the evidence with any methodological care. Section 5 considers the empirical evidence for the Priority Heuristic (PH) from psychology, and offers a sharp reminder of why we worry about the likelihood of observations from the perspective of theory. Section 6 considers empirical evidence for the notion of "source dependence," the hypothesis that risk preferences depend on the source of risk, and shows why we must not confuse point estimates with data. Section 7 draws some general conclusions, and a call to arms for methodologists.

1. Best-Practice Econometric Methods

There is a useful divide between non-structural econometric methods and structural methods. The two should be seen as complementary, depending on the inferential question at hand.

A. Non-Structural Methods

It is appropriate to dispense with a structural specification when the experimental design has controlled for the factors of importance for inference. Obviously, a randomized treatment is attractive and widely viewed as facilitating identification of the treatment effect, and this has been long recognized in laboratory experiments as well as field experiments. There is also widespread recognition that sometimes it is not as easy to fully randomize as one would want, and in this case one might resort to evaluating the "intent to treat" instead of the treatment itself. Or one might engage in some sort of modeling of the sample selection process, by which subjects present for the control or the treatments. These econometric methods are well known and understood.

Although it is popular to use Ordinary Least Squares (OLS) methods for non-structural econometrics, there is a growing awareness that alternative specifications are just as easy to estimate

and interpret, and can avoid some major pitfalls of OLS.[2] These issues arises when dependent variables are not real-valued between $\pm\infty$. The first item of business is to just plot the data, normally with a histogram or kernel density. The advantage of a histogram is that it might show a "spike" better, whether the spike is at some natural boundary or at some prominent value. These plots are not intended to see if the unconditional distribution is bell-shaped, since it is the distribution of the *residual* that we want to be Gaussian for the proper application of OLS. Unless the only covariate is a constant term, these are not the same thing.

Once the plot shows us if the data are bounded, dichotomous, ordered, or nominal (e.g., integer-valued), we all know what to do. In the old days is was not a trivial matter to compute marginal effects using proper econometric methods that kept track of standard errors, and allowed hypothesis testing, but those days have long passed. Marginal effects can be calculated using the "delta method," allowing non-linear functional relationships of estimated parameters to be calculated along with the (approximately) correct standard error. An important extension is to evaluate marginal effects for all values of the remaining covariates, and average those estimates: these are commonly called "average marginal effects," and convey a better sense of the marginal effect than when that effect is evaluated at

[2] Occasionally one encounters defenders of OLS, even when we know that the conditions for OLS are violated. None of these arguments hold much water when confronted. One argument is that it is "easier to interpret OLS estimates directly as marginal effects." Yes, but that is only because one has to assume away anything that might cause OLS to generate unreliable marginal effects. That is just circular reasoning. What might be easier, might just as well be wrong: ease of calculation and cognitive effort are not the same thing as validity of estimates. And modern software completely removes the ease argument. Another argument for OLS is that "you get the same results anyway." Really? In the old days one might have seen a wide table of OLS estimates, with gaps here and there to reflect specification searches, and one column in which estimates from the appropriate model is included. But not the myriad of specification searches using the appropriate model, the validity of *ad hoc* specification searches aside. So we do not know if the "robustness" shown with OLS is indeed a robustness that carries over to the appropriate model. Another argument for OLS, common in some finance journals, is that "I don't believe the results unless I see them in OLS." This is just bad epistemology, and should be called out as such. And if this is the theological ritual needed to get published, why not put the knowingly-incorrect estimates in the online appendix? Another argument for OLS is that, "I checked and the average is in the interior of the natural boundary." Perhaps some share, bounded between 0 and 1, has an average of 0.24. But that is the average, which is swept out by the OLS estimate (on a good day with respect to other assumptions). It says nothing about the residuals, which are the things we would like to be Gaussian, and lie unconstrained between $\pm\infty$. Are we just to ignore the residual that is below 0 or above 1? Finally, one sometimes hears, "well, everyone else does it," and surely that statement does not even need a rebuttal in scientific discourse.

the mean of the remaining covaiates.

One important insight from modern methods is to recognize the important distinction between a "hurdle specification" and a censored specification (e.g., a Tobit). Each of these arise in the common situation in which there is a spike in the data at some prominent value, typically zero. The classic example in economics is an expenditure on some good, and in health economics the utilization or expenditure on medical services. In this case the hurdle model recognizes that the data-generating process that causes the zero observations may be very different than the data-generating process that causes the non-zero observations. For instance, women may be less likely to go to hospital than men, but once there they may use more costly resources. Hence an OLS estimate of the effect of gender on health expenditure might see no net effect, but that is because the two data-generating processes are generating large gross effects that offset each other. Hurdle models can only ever improve inferences in settings like this, by allowing two latent processes to generate the data independently. Specifications that handle censoring, such as Tobit models, assume that there is one latent data-generating process, but that it is transformed into a zero or non-zero observation in some manner that is independent of covariates.

Hurdle models are extremely easy to estimate. Limited-information methods, where one estimates, say, a probit model for the zero or non-zero characteristics of the data, and then a constrained OLS for the non-zero level of the data conditional on it being non-zero, generate consistent estimates. Efficient estimates require maximum likelihood (ML) methods for the joint estimation of both the probit and constrained OLS, but these are trivial now. One can easily extend the specification to consider two-sided hurdles, two-step hurdles, and non-zero data-generating-processes that are integer-valued or bounded. Again, marginal effects can be readily calculated to correctly take into account both stages of the generation of an observation, or just one stage alone if that is of interest.

Randomization to treatment is one way to try to ensure that the effects of heterogeneity are

controlled for. If sample sizes are large enough, and assignment to treatment random enough, then many observable and non-observable characteristics will be "balanced" and hence play no significant role as a confound for inference. There also exist techniques to "re-balance" the samples that are used in treatments with the samples that are in the control, so as to make inferences about treatment effect even more reliable. These techniques are most widely used in observational settings where no assignment to treatment has occurred, or cannot occur for ethical reasons. However they may also be used to improve inferences when sample sizes do not allow one to rely solely on the effects of randomization.[3]

B Structural Methods

Behavioral economics now provides a rich array of competing structural models of decision-making in many areas of interest. In terms of risk preferences, major alternatives to Expected Utility Theory (EUT) include Rank-Dependent Utility (RDU) and Cumulative Prospect Theory (CPT). In terms of time preferences, major alternatives to Exponential Discounting include Hyperbolic Discounting and Quasi-Hyperbolic Discounting. We now also have rich, structural characterizations of attitudes towards uncertainty and ambiguity, as well as social preferences. All of these models consist of latent structures: they posit latent constructs that individuals behave as if they evaluate when making decisions. For example, in EUT the latent constructs consist of the utility of outcomes, the expected utility of lotteries of outcomes, and the difference in expected utility of alternative lotteries in a binary choice setting. In turn, these latent constructs can be characterized with parametric, semi-parametric or non-parametric functional forms. Within the parametric family, there can be flexible functional forms

[3] One limitation is that the "treatment" has to be binary, continuous *or* multilevel, but cannot be a mix of these. Unfortunately many treatments of interest are best characterized by a rich mixture of all of these. Consider the evaluation of the effect of smoking on health expenditures. Smoking history might depend on whether the individual had ever smoked 100 cigarettes (binary), wether the individual currently smokes daily or occasionally (binary), whether the individual is a former smoker (binary), the number of cigarettes smoked per day (discrete, multi-valued), and the number of years that current daily smokers have smoked (discrete, multi-valued).

that generalize many others, or there can be relatively restrictive functional forms. For simplicity, most of our remarks focus on risk preferences.

Sometimes one can avoid estimating the full structure by just studying comparative static predictions of different theories. Indeed, the vast bulk of the behavioral literature testing EUT consists of the careful design of pairs of lotteries that provide tests of EUT by just examining the patterns of choice: see Starmer [2000] for a masterful review. In the renowned Allais Paradox, for instance, observed choices between one lottery pair A and B lead to precise predictions over another lottery pair A* and B* that are transformations of A and B: if the subject picks A (B) then under EUT the subject must also pick A* (B*). If the purpose is to test EUT against an alternative, then one might just study patterns such as these for consistency.[4]

One immediate problem is that choice patterns might have extremely low power when it comes to testing EUT. The reason is that many of the popular tests, such as the Allais Paradox and Common Ratio tests, use lottery pairs where the individual might reasonably be close to indifferent between the two. To avoid this problem, Loomes and Sugden [1998] design an ingenious battery of lottery choices which vary the "gradient" of the EUT-consistent indifference curves within a Marschak-Machina (MM) triangle.[5] The reason for this design feature is to generate some choice patterns that are more powerful tests of EUT for any given risk attitude. Under EUT the slope of the indifference curve within a MM triangle is a measure of risk aversion. So there always exists some risk attitude such that the subject is indifferent, as stressed by Harrison [1994], and evidence of Common Ratio (CR) violations in that case

[4] One issue here is that we cannot compare the choices over A and B of one subject with the choices over A* and B* of another subject, without making the unwarranted assumption that they have the same preferences over risk. In practice the same subject can have both pairs presented in the context of a wider battery, and then direct comparisons can be made for each subject.

[5] In the MM triangle there are always one, two or three prizes in each lottery that have positive probability of occurring. The vertical axis in each panel shows the probability attached to the high prize of that triple, and the horizontal axis shows the probability attached to the low prize of that triple. So when the probability of the highest and lowest prize is zero, 100% weight falls on the middle prize. Any lotteries strictly in the interior of the MM triangle have positive weight on all three prizes, and any lottery on the boundary of the MM triangle has zero weight on one or two prizes.

has virtually zero power.[6]

The beauty of this design is that even if the risk attitude of the subject makes the tests of a CR violation from some sets of lottery pairs have low power, then the tests based on other sets of lottery pairs must have higher power for this subject. By presenting subjects with several such sets, varying the slope of the EUT-consistent indifference curve, one can be sure of having some tests for CR violations that have decent power for each subject, without having to know *a priori* what their risk attitude is. Harrison, Johnson, McInnes and Rutström [2007] refer to this as a "complementary slack experimental design," since low-power tests of EUT in one set mean that there must be higher-power tests of EUT in another set.[7]

This design illustrates how smart experimenters can mitigate "downstream" econometric problems, when they know the theory they are testing. But the need for structural estimation remains. We still need to know if the subject has sufficiently precise risk preferences to make any of these tests powerful. What if the subject does not have a temporally stable or deterministic utility function? If we can estimate an EUT model for each subject we can then weight the evidence across a sample, where the greatest weight is given to those with relatively precisely estimated risk preferences.

There are four deeper methodological reasons why the need for structural estimation remains.

The earliest tests of EUT were tests of the point-null hypothesis of EUT against the composite-alternative hypothesis of "anything but EUT." In this setting the subject either behaved consistently with EUT or not, and that translated into non-rejection of the null or not. But the most interesting tests now are horse races of one specification against another: for instance, does EUT or RDU best

---

[6] EUT does not, then, predict 50:50 choices, as some casually claim. It does say that the expected utility differences will not explain behavior, and that then allows all sorts of psychological factors to explain behavior. In effect, EUT has *no* prediction in this instance, and that is not the same as predicting an even split.

[7] The famous "preference reversal" experiments of Grether and Plott [1979], for instance, have virtually no power when the individual is risk neutral, since the lotteries in each pair were chosen to have roughly the same expected value. But a given subject cannot simultaneously have a a low-power test of EUT from preference reversal choices *and* a low-power test of EUT from CR choices, assuming we have some reasonably precise estimate of the risk attitudes of the subject.

characterize behavior? This happens to be an easy horse race to judge, since EUT is nested within RDU. So the goal becomes the estimation of a reasonably flexible RDU model, and then a test if the restriction to EUT is rejected or not at conventional statistical levels. Horse races of non-nested models involve more careful hypothesis tests or mixture models, discussed by Harrison and Rutström [2009], but the need for structural estimation remains.[8]

The second reason for structural estimation is to be able to compare the latent risk preferences generated by different elicitation methods. An unfortunate cottage industry designing new elicitation methods has grown up, and a natural question to ask is whether they generate the same latent risk preferences or not. There are any number of reasons why theoretically incentive-compatible elicitation methods might not elicit the same risk preferences: the most important behaviorally is that some tasks are easier to explain than others.[9] The point is not whether there is some pairwise correlation between observed choices or reports across elicitation methods, but rather whether they lead one to recover the same latent risk preferences. For this comparison one must specify a structural model for each method that connects observed responses to risk preferences, and then generate the likelihoods of each observation for that method. Then do the same for other methods, and then generate a grand model in which the likelihoods for both models are estimated simultaneously, allowing a direct test that one

[8] Mixture models change the language of horse races, in important ways, as well as allowing one to see how non-nested hypothesis tests have historically been "second best" alternatives to a fully-specified mixture. Rather than posing these as binary outcomes, where one model wins and the other is rejected, mixture models estimate the weight of the evidence consistent with one model over the other. And that weight can vary predictably with demographic characteristics or task characteristics. As usual, Bayesian handle all of this in a natural manner, with posterior odds being the basis for assessing the weight of one model over another, and Hierarchical Bayesian methods allow meta-parameters to affect these weights. Mixture models also provide an insight into the use of multiple criteria by an individual decision-maker in a given choice, in the spirt of the SP/A model of Lopes [1984] from psychology: see Andersen, Harrison, Lau and Rutström [2014a].

[9] A classic example is the binary choice procedure, which is self-evidently incentive-compatible, compared to the Becker, DeGroot and Marschak [1964] (BDM) elicitation method. Although formally incentive compatible, the BDM elicitation method is widely avoided by experimental economists since subjects often fail to understand it without a great deal of hands-on training: see Plott and Zeiler [2005; p.537]. Moreover, even if subjects understand the incentives, the mechanism is known to generate *extremely* weak incentives for accurate reports: see Harrison [1992][1994]. In the interests of full disclosure, it was Harrison [1986] that first proposed the use of the BDM method to elicit risk attitudes.

method generates different structural parameters.[10]

The third reason for structural estimation is to be able to characterize risk preferences for normative purposes. It is one thing to say that a subject is better characterized by EUT or RDU, and another thing to be able to evaluate the consumer surplus of observed choices given estimates of the risk preferences of the subject. In other words, when someone makes a risky choice, and we "know" their risk preferences from some other battery of risky choices and structural estimates, what is the *size* of the consumer surplus gained or foregone? Data on choice patterns is silent on this, even if we have intelligently designed a battery to tell us that some choices involve a larger consumer surplus, positive or negative, depending on the choice, than others. By themselves, choice patterns can only tell us the sign of the consumer surplus, not the magnitude. Section 2 provides a case study to illustrate the role of structural estimation in behavioral welfare economics.

The fourth reason for structural estimation is to be able to correctly infer some latent construct that depends on some latent characteristic of another construct. This seemingly abstract point is of great practical significance. For example, to estimate time preferences, where the discount factor is defined as the scalar that equates the present discounted utility of a larger, later (LL) amount to the present discounted utility of a smaller, sooner (SS) amount, one needs to know the utility function for the amounts. Concavity of the utility function has a first-order impact on inferred discount rates, as shown by Andersen, Harrison, Lau and Rutström [2008], who introduced the idea of joint estimation and applied it to risk and time preferences. To correctly infer discount rates from observed choices over LL and SS outcomes, one must know or assume some value for U″, and this comes most easily from estimates of a parametric utility function.[11] Similar applications arise when estimating subjective

[10] It is not *a priori* obvious that this exercise is interesting if one has access to a transparent elicitation method that is attractive by making minimal demands on the understanding of subjects. Arguably this is true of binary choice methods, even if other methods would provide greater information *if behaviorally reliable* (e.g., knowing a certainty equivalent takes one immediately to the risk premium).

[11] Since risk attitudes only equate to U″ under EUT, it is a mistake to equate joint estimation in this application with "risk attitudes and time preferences being correlated."

probabilities, as shown by Andersen, Fountain, Harrison and Rutström [2014], and when estimating the intertemporal risk preferences, as shown by Andersen, Harrison, Lau and Rutström [2017]. Section 3 reviews applications of joint estimation, and the methodological issues that arise.

2. Behavioral Econometrics and Behavioral Welfare Economics

Consider the evaluation of consumer surplus from a simple, full indemnity insurance contract, following Harrison and Ng [2016]. We know from theory that a risk averse EUT agent should always purchase this product at premia equal or below the actuarially fair premium, and would garner a positive consumer surplus from doing so. But how large a surplus? The agent will also purchase the product at premia with positive loadings, but consumer surplus drops as the loading increases, and at some point the product should not be purchased. But how quickly does the surplus diminish, and at what point should the agent decline to buy?

To answer these questions we need to know the risk preferences of the agent, and then use those to evaluate the consumer surplus of observed insurance choices. That surplus may be positive or negative, depending on whether the "correct" purchase decision is made, conditional on the risk preferences of the agent. The first step is to estimate risk preferences, the second step is to calculate consumer surplus conditional on risk preferences, the third step is to determine the best characterization of risk preferences for the agent, and the final step is to assess the impact on welfare.

A. Risk Preferences

Assume that utility of income is defined by

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

where x is the lottery prize and $r \neq 1$ is a parameter to be estimated. Thus r is the coefficient of CRRA: $r=0$ corresponds to risk neutrality, $r<0$ to risk loving, and $r>0$ to risk aversion. Let there be J possible outcomes in a lottery. Under EUT the probabilities for each outcome x_j, $p(x_j)$, are those that are

induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome in each lottery i:

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

The EU for each lottery pair is calculated for a candidate estimate of r, 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 as presented to subjects. This latent index, 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,

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

Even though this "link function" 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 being observed. The index defined by (3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla EU)>½$, which is implied by (4).

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

$$\ln L(r; y, \mathbf{X}) = \sum_i [\, (\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) \,] \tag{5}$$

where $\mathbf{I}(\cdot)$ is the indicator function, $y_i = 1 (-1)$ denotes the choice of the right (left) lottery in risk aversion task i, and $\mathbf{X}$ is a vector of individual characteristics reflecting age, sex, race, and so on.[12]

[12] Harrison and Rutström [2008; Appendix F] review procedures 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 researches can write explicit 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"), as needed. It is also a simple matter to generalize this ML analysis to allow the core parameter r to be a linear function of observable

An important extension of the core model is to allow for subjects to make some *behavioral* errors. The notion of error is one that has already been encountered in the form of the statistical assumption that the probability of choosing a lottery is not 1 when the EU of that lottery exceeds the EU of the other lottery. This assumption is clear in the use of a non-degenerate link function between the latent index ∇EU and the probability of picking one or other lottery; in the case of the normal CDF, this link function is $\Phi(\nabla EU)$. If there were no errors from the perspective of EUT, this function would be a step function: zero for all values of $\nabla EU<0$, anywhere between 0 and 1 for $\nabla EU=0$, and 1 for all values of $\nabla EU>0$.

We employ the error specification originally due to Fechner and popularized by Hey and Orme [1994]. This error specification posits the latent index

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

instead of (3), where μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. 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. As $\mu \to 0$ 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), where the probability of picking one lottery is given by the ratio of the EU of one lottery to the sum of the EU of both lotteries. Thus μ can be viewed as a parameter that flattens out the link functions as it gets larger.

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 posits the latent index

---

characteristics of the individual or task. We would then extend the model to be $r = r_0 + R \times \mathbf{X}$, where r_0 is a fixed parameter and R is a vector of effects associated with each characteristic in the variable vector **X**. In effect the unconditional model assumes $r = r_0$ and just estimates r_0. This extension significantly enhances the attraction of structural ML estimation, particularly for responses pooled over different subjects and treatments, since one can condition estimates on observable characteristics of the task or subject.

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

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 pair to lottery choice pair: hence it is said to be "contextual." For the Fechner specification, dividing by ν ensures that the *normalized* EU difference $[(EU_R - EU_L)/\nu]$ remains in the unit interval. The term ν does not need to be estimated in addition to the utility function parameters and the parameter for the behavioral error tern, since it is given by the data and the assumed values of those estimated parameters.

The specification employed here is the CRRA utility function from (1), the Fechner error specification using contextual utility from (3″), and the link function using the normal CDF from (4). The complete log-likelihood is then

$$\ln L(r, \mu; y, \mathbf{X}) = \sum_i [(\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1))] \qquad (5'')$$

and the parameters to be estimated are r and μ given observed data on the binary choices y and the lottery parameters in **X**.

The RDU model of Quiggin [1982] extends the EUT model by allowing for decision weights on lottery outcomes. The specification of the utility function is the same 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 consider three popular probability weighting functions. The first is the simple "power" probability weighting function proposed by Quiggin [1982], with curvature parameter γ:

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

So $\gamma \neq 1$ is consistent with a deviation from the conventional EUT representation. Convexity of the probability weighting function is said to reflect "pessimism" and generates, if one assumes for simplicity a linear utility function, a risk premium since $\omega(p) < p \ \forall p$ and hence the "RDU EV" weighted by $\omega(p)$ instead of p has to be less than the EV weighted by p. 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 second probability weighting function is the "inverse-S" function popularized by Tversky and Kahneman [1992]:

$$\omega(p) = p^{\gamma} / (p^{\gamma} + (1-p)^{\gamma})^{1/\gamma} \tag{8}$$

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 $\gamma>1$.

The third probability weighting function is a general functional form proposed by Prelec [1998] that exhibits considerable flexibility. This function is

$$\omega(p) = \exp\{-\eta(-\ln p)^{\varphi}\}, \tag{9}$$

and is defined for $0<p\leq 1$, $\eta>0$ and $\varphi>0$. When $\varphi=1$ this function collapses to the Power function $\omega(p) = p^{\eta}$. Of course, EUT assumes the identity function $\omega(p)=p$, which is the case when $\eta = \varphi = 1$. Many apply the Prelec [1998; Proposition 1, part (B)] function with constraint $0 < \varphi < 1$, which requires that the probability weighting function exhibit subproportionality (so-called "inverse-S" weighting). Contrary to received wisdom, many individuals exhibit estimated probability weighting functions that violate subproportionality, so we use the more general specification from Prelec [1998; Proposition 1, part (C)], only requiring $\varphi > 0$, and let the evidence determine if the estimated φ lies in the unit interval. This seemingly minor point often makes a major difference empirically.[13]

[13] One often finds applications of the one-parameter Prelec [1988] function, on the grounds that it is "flexible" and only uses one parameter. The additional flexibility over the inverse-S probability weighting function is real, but minimal compared to the full two-parameter function.

The construction of the log-likelihood for the RDU model with power or inverse-S probability weighting follows the same pattern as for EUT, with the parameters r, γ and μ to be estimated. Similarly, for the RDU model with Prelec probability weighting the parameters r, η, φ and μ are to be estimated.

B. Welfare Evaluation

If the subject is assumed to be an EUT type, the consumer surplus (CS) of the insurance decision is calculated as the difference between the certainty equivalent (CE) of the EU with insurance and the CE of the EU without insurance. CS is calculated the same way using the RDU instead of EU if the subject is classified as a RDU type.

Assume a simple indemnity insurance product, which provides full coverage in the event of a loss. We assume an initial endowment of $20, with a 10% chance of a $15 one-time loss occurring. If an individual purchased the insurance, she could avoid the loss with certainty by paying the insurance premium up front. There are four possible payoff outcomes. If no insurance is purchased, the individual keeps her $20 if no loss occurs, but is only left with $5 if there is a loss. If insurance is purchased, the individual keeps $20 less the premium if no loss occurs, and still keeps $20 less the premium if the loss does occur.

Using the decision-making models discussed above, the EU or RDU across the two possible states, loss or no loss, can be calculated for each choice, to purchase or not to purchase insurance. The CE from the EU or RDU of each choice can be derived, and the difference between the CE from choosing insurance and the CE from not choosing insurance is then the expected welfare gain of purchasing insurance for that individual.

Figure 1 shows how this CS from purchasing insurance would vary for an EUT individual following the above example, for premiums ranging from $0.01 to $4.50. Each bar shows the CS for a CRRA coefficient ranging from 0.3 to 0.7, typical values expected for a risk averse EUT individual in

an experiment. We see that the CS is larger if the individual is more risk averse, which follows from the fact that more risk averse individuals are willing to pay more for insurance. As premiums increase, CS becomes negative, showing that there is a threshold premium above which the subject would experience negative expected welfare from purchasing the insurance product.

Similar graphs are generated by Harrison and Ng [2016] for RDU using an inverse-S probability weighting function and a power weighting function, respectively. For both models γ ranges from 0.7 to 1.3, and the CRRA coefficient is held constant at 0.6. As γ increases, CS decreases when the inverse-S probability weighting function is used, but CS increases when the power function is used. Assigning the right decision making model, even from this basic set of EUT and RDU specifications, is important for measuring individual welfare evaluation. In general, estimating the right risk parameters for the individual, conditional on the decision making model, will also affect the identification of the correct decision as well as the opportunity cost of that decision.

C. The Welfare Metric

To evaluate RDU preferences one can estimate an RDU model for each individual. We consider the CRRA utility function (1) and one of three possible probability weighting functions defined earlier by (7), (8) and (9). For the purpose of classifying subjects as EUT or RDU it does not matter which of these probability weighting functions characterize behavior: the only issue here is at what statistical confidence level we can reject the EUT hypothesis that $\omega(p) = p$.

Of course, if the sole metric for deciding if a subject were better characterized by EUT and RDU was the log-likelihood of the estimated model, then there were be virtually no subjects classified as EUT since RDU nests EUT. But if we use metrics of a 10%, 5% or 1% significance level on the test of the EUT hypothesis that $\omega(p) = p$, then we classify 39%, 49% or 68%, respectively, of 102 subjects with valid estimates as being EUT-consistent.

D. Welfare Evaluation

Expected welfare gain is foregone if the subject chooses to purchase insurance when that decision has a negative CS, and similarly when the subject chooses not to purchase insurance when the decision has a positive CS. For example, if we consider the expected welfare gain from each decision to the actual decisions made by subject 8, based on her EUT classification, we find that the subject has foregone $10.37 out of a possible $31.36 of expected welfare gain from insurance. This subject's total expected welfare gain for all 24 decisions was $10.62; hence the efficiency for this subject, in the spirit of the traditional definition by Plott and Smith [1978], is 33.9%. In this experiment the efficiency is the expected CS given the subject's actual choices and estimated risk preferences as a percent of total possible expected CS given her predicted choices and estimated risk preferences. The efficiency metric is defined at the level of the individual subject, whereas the expected welfare gain is defined at the level of each choice by each subject. In addition, efficiency provides a natural normalization of expected welfare gain on loss by comparing to the maximal expected welfare gain for that choice and subject. Both metrics are of interest, and are complementary.

Expanding this analysis to look across all subjects, the left panel of Figure 2 shows the kernel density of the expected CS of each decision made. We find that 49% of decisions made resulted in negative predicted CS. The distribution of expected CS from these results is similar to the distribution if the insurance purchase decision was randomized. If insurance was randomly purchased, 50% of decision made would result in negative predicted CS, and average expected welfare gain would not be significantly different from $0. Although the average expected welfare gain of $0.27 from actual decisions made is statistically greater than zero at a p-value of less than 0.001, there are still a large proportion of decisions where take-up is not reflecting the welfare benefit of the insurance product to the individual.

The efficiency of all decisions made is only 14.0%. The right panel of Figure 2 shows the distribution of efficiency of decisions made by each individual. The modal efficiency is slightly less than

50%, and a significant proportion of individuals make decisions that result in negative efficiency. In other words, these subjects have made choices that resulted in a larger expected welfare loss than the choices that resulted in any expected welfare gain.

One objective of this exercise is to define conceptually and demonstrate empirically how one could undertake a field evaluation of the welfare of insurance products. We also view the laboratory as the appropriate place to "wind tunnel" the normative welfare evaluation of new products or decision scaffolds. Figure 2 stands as explicit, rigorous "target practice" for anyone proposing nudges or clubs to improve welfare from insurance decisions.

E. What Should the Normative Welfare Metric Be?

Our statement of welfare losses takes as given the type of risk preferences each individual employs, and uses that as the basis for evaluating welfare effects of insurance decisions: *periculum habitus non est disputandum*. One could go further and question if the RDU models themselves embody an efficiency loss for those subjects we classify as RDU. Many would argue that RDU violates some normatively attractive axioms, such as the independence axiom. Forget whether that axiom is descriptively accurate or not. If RDU is not normatively attractive then we should do a calculation of CS in which we only assume EUT parameters for subjects: we could estimate the EUT model and get the corresponding CRRA coefficient estimate (we would not just use the CRRA coefficient estimate from the RDU specification). Then we repeat the calculations. For subjects best modeled as EUT there is no change in the inferred CS, of course.

This issue raises many deeper issues with the way in which one should undertake behavioral welfare economics, discussed by Harrison and Ross [2016][2017]. For now, we take the agnostic view that the risk preferences we have modeled as best characterizing the individual are those that should be used, in the spirit of the "welfarism" axiom of welfare economics. Even though the alternatives to EUT were originally developed to relax one of the axioms of EUT that some consider attractive normatively,

it does not follow that one is unable to write down axioms that make those alternatives attractive normatively.

We view this methodological issue as urgent, open, and important. There is a large, general literature on behavioral welfare economics. Our general concern with this literature is that although it identifies the methodological problem well, none provide "clear guidance" so far to practical, rigorous welfare evaluation with respect to risk preferences as far as we can determine. We know of no way to undertake robust, general welfare evaluations of risky decisions without knowing structural risk preferences.

3. The Many Applications of Joint Estimation

The idea of joint estimation, again, is that one jointly estimates preferences from one structural model in order to correctly identify and estimate preferences of another structural model. The need for joint estimation comes from theory – it is not just an empirical matter of attending to behavioral correlations. We review applications here, limiting attention to non-strategic settings.[14]

A. Time Preferences

In many settings in experimental economics we want to elicit some preference from a set of choices that also depend on risk attitudes. An example due to Andersen, Harrison, Lau and Rutström [2008] is the elicitation of individual discount rates. In this case it is the concavity of the utility function, U'', that is important, and under EUT that is synonymous with risk attitudes. Thus the risk aversion task is just a (convenient) vehicle to infer utility over deterministic outcomes. One methodological

[14] The same concepts apply in strategic settings, but with the added complexity that the likelihood of behavior of all subjects in the game must be constrained by some equilibrium concept. Goeree, Holt and Palfrey [2003] illustrate the joint estimation of risk attitudes for a representative agent playing a generalized matching pennies game, with a "quantal response equilibrium" constraint. Harrison and Rutström [2008; §3.6] illustrate the joint estimation of risk attitudes and bidding behavior in a First-Price sealed bid auction, with a Bayesian Nash Equilibrium constraint.

implication is that we should combine a risk elicitation task with a time preference elicitation task, and use them jointly to infer discount rates over utility.

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

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

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

Most early analyses of discounting models implicitly assume that the individual has a linear utility function, so that (10) is instead written in the more familiar form

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

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

To state the obvious, (10) and (11) are not the same. As one relaxes the assumption that the decision maker has a linear utility function, it is apparent from Jensen's Inequality that the implied discount rate decreases if U(M) is concave in M. Thus one cannot infer the level of the discount rate without knowing or assuming something about the utility function. This identification problem implies that discount rates cannot be estimated based on discount rate experiments with choices defined solely over time-dated money flows, and that separate tasks to identify the extent of diminishing marginal

utility must also be implemented.

Thus there is a clear implication from theory to experimental design: you need to know the non-linearity of the utility function before you can *conceptually* define the discount rate. There is also a clear implication for econometric method: you need to jointly estimate the parameters of the utility function and the discount rate, to ensure that sampling errors in one propagate correctly to sampling errors of the other. In other words, if we know the parameters of the utility function less precisely, due to small samples or poor parametric specifications, we have to use methods that reflect the effect of that imprecision on our estimates of discount rates.[15]

Andersen, Harrison, Lau and Rutström [2008] do this, and infer discount rates for the adult Danish population that are well below those estimated in the previous literature that assumed linear utility functions, such as Harrison, Lau and Williams [2002], who estimated annualized rates of 28% for the same target population. Allowing for concave utility, they obtain a point estimate of the discount rate of 10%, which is significantly lower than the estimate of 25% for the same sample assuming linear utility. This does more than simply verify that discount rates and diminishing marginal utility are mathematical substitutes in the sense that either of them have the effect of lowering the influence from future payoffs on present utility. It tells us that, for utility function coefficients that are reasonable from the standpoint of explaining choices in the lottery choice task, the estimated discount rate takes on a value that is much more in line with what one would expect from market interest rates. To evaluate the statistical significance of adjusting for a concave utility function one can test the hypothesis that the estimated discount rate assuming risk aversion is the same as the discount rate estimated assuming linear utility functions. This null hypothesis is easily rejected. Thus, *allowing for diminishing marginal utility*

[15] It is true that one must rely on structural assumptions about the form of utility functions, probability weighting functions, and discounting functions, in order to draw inferences. These assumptions can be tested, and have been, against more flexible versions and even non-parametric versions (e.g., Harrison and Rutström [2008; p. 78-79]). A similar debate rages with respect to structural assumptions about error specifications, as illustrated by the charming title of the book by Angrist and Pischke [2009], *Mostly Harmless Econometrics*. But it is an illusion, popular in some quarters, that one can safely dispense with all structural assumptions and draw inferences: see Keane [2010] and Leamer [2010] for spirited assaults on that theology.

makes a significant difference to the elicited discount rates.

We can write out the likelihood function for the choices that our subjects made and jointly estimate the risk parameter r in (1) and the discount rate δ in (10). We use the same stochastic error specification as in §2, and the contribution to the overall likelihood from the risk aversion responses is given by (5″).

A similar specification is employed for the discount rate choices. Equation (3) is replaced by the discounted utility of each of the two options, conditional on some assumed discount rate, and equation (4) is defined in terms of those discounted utilities instead of the expected utilities. The discounted utility of Option A is given by

$$PV_A = (\omega+M_A)^{(1-r)}/(1-r) + (1/(1+\delta)^{\tau})\ \omega^{(1-r)}/(1-r) \tag{12}$$

and the discounted utility of Option B is

$$PV_B = \omega^{(1-r)}/(1-r) + (1/(1+\delta)^{\tau})\ (\omega+M_B)^{(1-r)}/(1-r) \tag{13}$$

where M_A and M_B are the SS and LL monetary amounts in the choice tasks presented to subjects, and the utility function is assumed to be stationary over time.

An index of the difference between these present values, conditional on r and δ, can then be defined as

$$\nabla PV = (PV_A + PV_B)/\xi \tag{14}$$

where ξ is a noise parameter for the discount rate choices, just as μ was a noise parameter for the risk aversion choices. It is not obvious that μ = ξ, since these are cognitively different tasks.

Thus the likelihood of the discount rate responses, conditional on the EUT, CRRA and Exponential discounting specifications being true, depend on the estimates of r, δ, μ and ξ, given the assumed value of ω and the observed choices. If we ignore responses that reflect indifference, the conditional log-likelihood is

$$\ln L^{DR} = \ln L(r, \delta, \mu, \xi; y, \omega, \mathbf{X}) = \sum_i [\ (\ln \Phi(\nabla PV) \times \mathbf{I}(y_i=1)) + (\ln (1-\Phi(\nabla PV)) \times \mathbf{I}(y_i=-1))\ ] \tag{15}$$

where $y_i = 1(-1)$ again denotes the choice of Option B (A) in discount rate task i, and **X** is a vector of

individual characteristics. The joint likelihood of the risk aversion and discount rate responses can then be written as

$$\ln L\,(r, \delta, \mu, \xi; y, \omega, \mathbf{X}) = \ln L^{RA} + \ln L^{DR} \tag{16}$$

where L^{RA} is defined by (5″) and L^{DR} is defined by (15). This expression can then be maximized using standard numerical methods.

B. Subjective Probabilities

Exactly the same joint estimation methodology can be used to infer subjective probabilities over some binary event. Subjective probabilities are operationally defined as those probabilities that lead an agent to choose some prospects over others when outcomes depend on events that are not yet actualized. These choices could be as natural as placing a bet on a horse race, or as experimentally structured as responding to the payoff prizes provided by some scoring rule. In order to infer subjective probabilities from observed choices of this kind, however, one either has to make some strong assumptions about risk attitudes or jointly estimate risk attitudes and subjective probabilities. Joint estimation of a structural model of choice across the two types of tasks, one to elicit risk attitudes and the other to (recursively) elicit beliefs conditional on risk attitudes, allows one to make inferences about subjective probabilities from observed behavior in relatively simple choice tasks.

For simplicity assume that the events in question only have two outcomes.[16] A scoring rule asks the subject to make some report θ, and then defines how an elicitor pays a subject depending on their report and the outcome of the event. This framework for eliciting subjective probabilities can be formally viewed from the perspective of a trading game between two agents: you give me a report, and I agree to pay you $X if one outcome occurs and $Y if the other outcome occurs. The scoring rule defines the terms of the exchange quantitatively, explaining how the elicitor converts the report from

[16] Extensions to eliciting subjective beliefs over continuous events are considered by Matheson and Winkler [1976] and Harrison, Martínez-Correa, Swarthout and Ulm [2017].

the subject into a lottery. We use the terminology "report" because we want to view this formally as a mechanism, and do not want to presume that the report is in fact the subjective probability π of the subject. In general, it is not.

The popular Quadratic Scoring Rule (QSR) is defined in terms of two positive parameters, α and β that determine a fixed reward the subject gets and a penalty for error. Assume that the possible outcomes are A or B, where B is the complement of A, that θ is the reported probability for A, and that Θ is the true binary-valued outcome for A. Hence $\Theta=1$ if A occurs, and $\Theta=0$ if it does not occur (and thus B occurs instead). The subject is paid $S(\theta \mid \text{A occurs}) = \alpha - \beta(\Theta-\theta)^2 = \alpha - \beta(1-\theta)^2$ if event A occurs and $S(\theta \mid \text{B occurs}) = \alpha - \beta(\Theta-\theta)^2 = \alpha - \beta(0-\theta)^2$ if B occurs. In effect, the score or payment penalizes the subject by the squared deviation of the report from the true binary-valued outcome, Θ, which is 1 and 0 respectively for A and B occurring. An omniscient seer would obviously set $\theta = \Theta$. The fixed reward is a convenience to ensure that subjects are willing to play this trading game, and the penalty function simply accentuates the penalty from not being an omniscient seer. In the experiments of Andersen, Fountain, Harrison and Rutström [2014] experiments $\alpha = \beta = \$100$, so subjects could earn up to \$100 or as little as \$0. If they reported 1 they earned \$100 if event A occurred or \$0 if event B occurred; if they reported ¾ they earned \$93.75 or \$43.75; and if they reported ½ they earned \$75 no matter what event occurred.

Assume for the moment that we have an EUT specification. The subject who selects report θ from a given scoring rule receives the following *subjective* EU

$$\begin{aligned} EU_\theta = \quad & \pi_A \times u(\text{payout if A occurs} \mid \text{report } \theta) + \\ & (1-\pi_A) \times u(\text{payout if B occurs} \mid \text{report } \theta) \end{aligned} \tag{17}$$

where π_A is the subjective probability that A will occur. The payouts that enter the utility function are defined by the scoring rule and of course the specific report θ, and span the interval [\$0, \$100]. For the QSR and a report of 75%, for example, we have

$$EU_{75\%} = \pi_A \times u(\$93.75) + (1-\pi_A) \times u(\$43.75) \tag{2''}$$

and so on for other possible reports. We observe the report made by the subject for QSR. This report can take 101 different integer values defined over percentage points. Then we can calculate the likelihood of that choice given values of r, π_A and μ, where the likelihood is the multinomial analogue of the binary logit specification used for lottery choices. We define

$$eu_{\Theta} = \exp[(EU_{\Theta}/\nu)/\mu] \tag{18}$$

for any report Θ,where μ is a Fechner error and ν is a contextual utility transformation, and then

$$\nabla EU = eu_{\theta}/(eu_{0\%} + eu_{1\%} + ... + eu_{100\%}) \tag{19}$$

for the specific report θ observed, analogously to the comparable expression for EUT or RDU.

We need r to evaluate the utility function in (17), we need π_A to calculate the EU_{θ} in (17) for each *possible* report Θ in {0%, 1%, 2%, ..., 100% } once we know the utility values, and we need μ to calculate the latent indices (18) and (19) that generate the subjective probability of observing the choice of specific report θ when we allow for some noise in that process. The *joint* ML problem is to find the values of these four parameters that best explain observed choices in the belief elicitation tasks as well as in the lottery tasks. These parameters are estimated simultaneously.

Exactly the same logic extends to the model in which we assume an RDU latent structure instead of an EUT latent structure. In effect, the lottery task allows us to identify r under EUT, and r *and* γ under RDU, thanks to the variations in both prizes and probabilities in this task.

C. Intertemporal Risk Preferences

Joint estimation scales "vertically upwards," as needed by theory. The concept of intertemporal risk aversion, also known as correlation aversion, is all about preferences over the *interaction* of risk preferences and time preferences. As such, one must jointly estimate atemporal risk preferences, time preferences, and the intertemporal utility function, building on the joint estimation approach to the first two developed by Andersen, Harrison, Lau and Rutström [2008].

The concept of intertemporal risk aversion arises from theoretical deviations from additively

separable intertemporal utility function. Define the lottery ψ as a 50:50 mixture of {x, Y} and {X, y}, and the lottery Ψ as a 50:50 mixture of {x, y} and {X, Y}, where X>x and Y>y. So ψ is a 50:50 mixture of both bad and good outcomes in time t and t+τ; and Ψ is a 50:50 mixture of only bad outcomes or only good outcomes in the two time periods. These lotteries ψ and Ψ are defined over all possible "good" and "bad" outcomes. If the individual is indifferent between ψ and Ψ we say that he is neutral to intertemporally correlated payoffs in the two time periods. If the individual prefers ψ to Ψ we say that he is averse to intertemporally correlated payoffs: it is better to have a given chance of being lucky in one of the two periods than to have the same chance of being very unlucky or very lucky in both periods. The correlation averse individual prefers to have non-extreme payoffs *across* periods, just as the risk averse individual prefers to have non-extreme payoffs *within* periods. One can also view the correlation averse individual as preferring to avoid correlation-increasing transformations of payoffs in different periods.

We consider first an intertemporal decision model with weakly separable preferences under EUT. The intertemporal utility function at time t=0 is written as:

$$U(X_1, X_2,..., X_n) = E[\ \Xi\ (\ \textstyle\sum_{t=1}^{n} (1/(1+\delta)^t)\ u(x_t)\ )\ ] \tag{20}$$

where $x_t \in X_t$, $u(x_t)$ is the atemporal utility of money at time t, and δ is the exponential discount rate. For now, let Ξ be an identity function; we return to it momentarily. The decision tasks in the experiments of Andersen, Harrison, Lau and Rutström [2017] provide lotteries with payments at two different points in time, where the time horizon between sooner and later payments varied between 2 weeks and 12 months. We can simplify the model in (20) and consider decisions that involve payments at two different points in time. Assuming u(0)=0, the intertemporal utility function is specified as:

$$U(X_t, X_{t+\tau},) = E[\ \Xi\ (\ D_t\ u(x_t) + D_{t+\tau}\ u(x_{t+\tau})\ )\ ] \tag{21}$$

where $D_t = 1/(1+\delta)^t$ is the discounting function with a constant discount rate δ.

Let the atemporal utility function be the CRRA specification (1) defined earlier. The key point in this setting is that it is defined solely over risky *atemporal* tradeoffs in t or t+τ.

Given the popularity of the CRRA function in the microeconomic and macroeconomic literature, it is natural to consider this alternative structural specification of the intertemporal utility function Ξ:

$$U(X_t, X_{t+\tau}) = E\left[\left(D_t\, u(x_t) + D_{t+\tau}\, u(x_{t+\tau}) \right)^{(1-\zeta)}/(1-\zeta) \right] = E\left[\Lambda(x_t, x_{t+\tau})^{(1-\zeta)}/(1-\zeta) \right] \quad (22)$$

where ζ is the intertemporal relative risk aversion parameter ($\zeta \neq 1$), and the expression for the weighted sum of atemporal utility flows, $\Lambda(x_t, x_{t+\tau})$, is useful below. The intertemporal utility function is separable but not additive when $\zeta \neq 0$, and collapses to $E[\Lambda(x_t, x_{t+\tau})]$ when there is intertemporal risk neutrality at $\zeta = 0$.

If the intertemporal utility function in (21) is additively separable, then the inverse of the intertemporal elasticity of substitution is equal to the coefficient of atemporal risk aversion. This assumption is one of convenience and is popular in models of intertemporal choice. The linear specification of intertemporal utility is then equal to a weighted sum of atemporal utility flows, where the weights are determined by the discount rate.

To elicit intertemporal risk aversion one has to present subjects with choices over lotteries that have different income profiles over time. Proper identification of intertemporal risk aversion (ζ) thus requires that one controls for atemporal risk aversion (r) and the individual discount rate (δ). All three parameters are intrinsically, conceptually connected as a matter of theory, unless one makes strong assumptions otherwise. The experimental design and econometric logic of Andersen, Harrison, Lau and Rutström [2017] follow from this theoretical point. The experimental procedures needed are a simple extension of those employed by Andersen, Harrison, Lau and Rutström [2008][2014b].

One task elicited atemporal risk attitudes for lotteries payable today, as a vehicle for inferring the concavity of the atemporal utility function. Another task elicited time preferences over non-stochastic amounts of money payable at different times: in general, a SS amount and a LL amount. In some cases the sooner amount was paid out today, and in some cases it was paid out in the future. A third task, new to this design, elicited intertemporal risk attitudes by asking subjects to make a series of

choices over risky profiles of outcomes that are paid out at different points in time. For example, lottery A might give the individual a 10% chance of receiving a larger amount L_t at time t *and* a smaller amount $S_{t+\tau}$ at time $t+\tau$, $(L_t, S_{t+\tau})$ and a 90% chance of receiving the smaller amount S_t at time t *and* the larger amount $L_{t+\tau}$ at time $t+\tau$, $(S_t, L_{t+\tau})$. Lottery B might give the individual a 10% chance of receiving L_t *and* $L_{t+\tau}$ and a 90% chance of receiving S_t *and* $S_{t+\tau}$. The subject picks A or B.

The econometric implications for joint estimation follow rigidly from the theory and experimental design presented above.

Consider non-additive separable specifications of the intertemporal utility function and estimate the coefficient of intertemporal risk aversion. Equation (22) implies that the expected utility of Option A in the intertemporal risk aversion task is given by

$$PEU_A = p(L_t , S_{t+\tau})\times[\ \Lambda(L_t, S_{t+\tau})^{(1-\zeta)}/(1-\zeta)\ ] + p(S_t, L_{t+\tau})\times[\ \Lambda(S_t, L_{t+\tau})^{(1-\zeta)}/(1-\zeta)\ ] \quad (23)$$

and the expected utility of Option B is given by

$$PEU_B = p(L_t, L_{t+\tau})\times[\ \Lambda(L_t, L_{t+\tau})^{(1-\zeta)}/(1-\zeta)\ ] + p(S_t, S_{t+\tau})\times[\ \Lambda(S_t, S_{t+\tau})^{(1-\zeta)}/(1-\zeta)\ ] \quad (24)$$

where $p(L_t, L_{t+\tau})$ is the probability of receiving L_t in period t and $L_{t+\tau}$ in period $t+\tau$. We can write out the likelihood function for the choices that the subjects made and jointly estimate the risk parameter r, the discount rate parameter δ, and the intertemporal risk parameter ζ. We again employ the contextual error specification proposed by Wilcox [2011], and the latent index is specified by

$$\nabla PEU = ((PEU_B - PEU_A)/\lambda)/\mu^{SDR} \quad (25)$$

where μ^{SDR} is a noise parameter for the ("stochastic discounting") intertemporal risk aversion choices.[17]

The likelihood of the intertemporal risk aversion responses, conditional on the specification of intertemporal utility being true, depends on the estimates of r, δ, ζ, μ^{SDR}, μ^{RA} and μ^{DR}, given the observed choices, where μ^{RA} is a noise parameter for the atemporal risk aversion choices and μ^{DR} is a

[17] The normalizing term λ is defined as the maximum intertemporal utility over all prize profiles in this lottery pair $(L_t, L_{t+\tau})$ minus the minimum utility over all prize profiles in this lottery pair $(S_t, S_{t+\tau})$. The maximum intertemporal utility over all prize profiles in the lottery pair is $[D_t\ u(L_t) + D_{t+\tau}\ u(L_{t+\tau})]^{(1-\zeta)}/(1-\zeta)$, and the minimum intertemporal utility is $[D_t\ u(S_t) + D_{t+\tau}\ u(S_{t+\tau})]^{(1-\zeta)}/(1-\zeta)$.

noise parameter for the discount rate choices. The conditional log-likelihood is

$$\ln L(r, \delta, \zeta, \mu^{RA}, \mu^{DR}, \mu^{SDR}; c, \mathbf{X}) = \sum_i [(\ln \Phi(\nabla PEU) \times \mathbf{I}(c_i=1)) + (\ln (1-\Phi(\nabla PEU)) \times \mathbf{I}(c_i=-1))] \quad (26)$$

where $c_i = 1(-1)$ denotes the choice of Option B (A) in intertemporal risk aversion task i, and all other notation is defined previously.

The joint likelihood of the atemporal risk aversion, discount rate and intertemporal risk aversion responses can then be written as

$$\ln L(r, \delta, \zeta, \mu^{RA}, \mu^{DR}, \mu^{SDR}; c, \mathbf{X}) = \ln L^{RA} + \ln L^{DR} + \ln L^{SDR} \quad (27)$$

where L^{RA} is the conditional log-likelihood of the atemporal risk aversion responses, L^{DR} is the conditional log-likelihood of the discount rate responses and L^{SDR} is defined by (26).

The nature of this joint likelihood function is matched by the recursive experimental design. Ignoring the objective parameters of the tasks, the lottery choices over stochastic lotteries paid out today (RA) depend on r, ζ and μ^{RA}; the discounting tasks over non-stochastic outcomes paid out today or some time in the future (DR) depend on r, μ^{RA}, δ and μ^{DR}; and the discounting tasks over stochastic outcomes paid out today or some time in the future (SDR) depend on r, μ^{RA}, δ, μ^{DR}, ζ and μ^{SDR}. Putting the behavioral error terms aside, if we were to try to estimate r and δ using either the RA or the DR choices, we would be unable to identify both parameters. Similarly, if we were to try to estimate r, δ and ζ using only two of three tasks, we would face an identification problem.

These identification problems are inherent to the *theoretical* definitions of the discount rate and intertemporal risk aversion, and demand a recursive experimental design that combines multiple types of choices and an econometric approach that recognizes the complete structural model. The general principle is joint estimation of all structural parameters so that uncertainty about the parameters defining the utility function propagates in a "full information" sense into the uncertainty about the parameters defining the discount function and the intertemporal utility function. Intuitively, if the experimenter only has a vague notion of what u(.) is, because of poor estimates of r, then one simply cannot make precise inferences about δ or ζ. Similarly, poor estimates of δ, even if r is estimated

relatively precisely, imply that one cannot make precise inferences about ζ.

This inferential procedure about intertemporal risk aversion does not rely on the use of EUT, or the CRRA functional form. Nor does it rely on the use of the exponential discounting function; the method generalizes immediately to alternative specifications that use alternative discounting functions, as illustrated in Andersen, Harrison, Lau and Rutström [2014b].

The implication for the claim by Andreoni and Sprenger [2012] that "risk preferences are not time preferences" is immediate. If the intertemporal utility function that subjects use is actually non-additive, then risk preferences over time streams of money need to be sharply distinguished from risk preferences over a-temporal payoffs. In effect, there are two possible types of risk aversion when one considers risky choices over time, not one. To be more precise, if one gives subjects choices over differently-time-dated payoffs, which is what Andreoni and Sprenger [2012] did, one sets up exactly the thought experiment that *defines* intertemporal risk aversion. They compare behavior when subjects make choices over time-dated payoffs that are not stochastic with choices over time-dated payoffs that are stochastic, and observe different behavior. In the former case virtually all choices in their portfolios were at extreme allocations, either all payoffs sooner or all payoffs later; in the latter case they observed more choices in which subjects picked an interior mix of sooner and later payoffs, diversifying intertemporally. Evidence that subjects behave differently, when there is an opportunity for intertemporal risk aversion to affect their choices compared to a setting in which it has no role, is evidence of intertemporal risk aversion. It is not necessarily evidence for the claim that there is a "different utility function" at work when considering stochastic and non-stochastic choices. We do not rule the latter hypothesis out, but there is a simpler explanation well within received theory.

Evidence for intertemporal risk aversion in experiments is provided by Andersen, Harrison, Lau and Rutström [2017], who also provide extensive cites to the older literature. Intertemporal risk aversion provides an immediate explanation for the observed behavior in Andreoni and Sprenger [2012]. Just as a-temporal risk aversion encourages mean-preserving reductions in the variability of

a-temporal payoffs (imagine lotteries defined solely over x and X or defined solely over y and Y), intertemporal risk aversion or intertemporal risk aversion encourages mean-preserving reductions in the variability of the time stream of payoffs (imagine lotteries ψ and Ψ defined above over x, X, y and Y).

Hence, when Andreoni and Sprenger [2012] claim that "risk preferences are not time preferences," one can restate this correctly as "a-temporal risk aversion is not the same as intertemporal risk aversion," and of course that is true whenever there is a non-additive intertemporal utility function.

D. A General Lesson

One general methodological lesson from these examples is that there is some considerable virtue in having experimental tasks that are "agnostic" about what latent structural model will be applied to them. We do not want an elicitation method for atemporal risk preferences that assumes EUT, RDU or CPT, or any of the myriad of alternative possible models one could consider (e.g., Disappointment Aversion or Regret Theory). Nor do we want an elicitation method for time preferences that assumes Exponential discounting. Inferences about intertemporal risk aversion should not be held methodological hostage to elicitation methods that lock in one theoretical specification or another, unless there are good *a priori* reasons for doing so.[18]

4. Just Read the Literature: A Case Study of Cumulative Prospect Theory

The key innovation of CPT, in comparison to RDU, is to allow sign-dependent preferences, where risk attitudes depend on whether the individual is evaluating a gain or a loss. Tversky and Kahneman [1992; p. 309] popularized the functional forms we often see for loss aversion, using a CRRA specification of utility: $U(m) = m^{1-\alpha}/(1-\alpha)$ when $m \geq 0$ and $U(m) = -\lambda[(-m)^{1-\beta}/(1-\beta)]$ when $m <$

[18] For example, we have seen so little evidence for CPT that we no longer automatically build in (longer) risk batteries with mixed frames or loss frames. We accept that others might demur.

0, where λ is the utility loss aversion parameter, and α and β are coefficients of utility curvature in the gain and loss frame, respectively. Here we have the assumption that the degree of utility loss aversion for small unit changes is the same as the degree of loss aversion for large unit changes: the same λ applies locally to gains and losses of the same monetary magnitude around 0 as it does globally to any size gain or loss of the same magnitude. This is not a criticism, just a restrictive parametric turn in the specification compared to Kahneman and Tversky [1979].

There is a clear statement of the critical "exchange rate assumptions" needed to define utility loss aversion in Abdellaoui, Bleichrodt and Paraschiv [2007; p.1662], as well as a tabulation of the range of definitions that have been proposed in the literature. For instance, Fishburn and Kochenberger [1979] and Pennings and Smidts [2003] defined loss aversion as $U'(-x)/U'(x)$, Tversky and Kahneman [1992] as $-U(-1)/U(1)$, Bleichrodt, Pinto and Wakker [2001] as $-U(-x)/U(x)$, and Schmidt and Traub [2002; p.235] as $U(x)-U(y) \leq U(-y)-U(-x)\ \forall\ x>y\geq 0$. One can make the exchange rate assumptions formally *de minimus* by defining an index of loss aversion solely in terms of the directional derivatives at the reference point, $U'_-(0)/U'_+(0)$, as proposed by Köbberling and Wakker [2005] and Booij and van de Kuilen [2009]. But this has the very unfortunate effect, as honestly emphasized by Wakker [2010; p. 247], that *global* properties of loss aversion are being driven by very, very *local* properties of estimated utility functionals,[19] and that puts a great strain on empirics and functional form assumptions.

Probability weighting for gains is identical to RDU, and the logic for losses is similar. Following Tversky and Kahneman [1992], one often sees the use of the inverse-S function, resulting in $\omega(p) = p^{\gamma+}/(\ p^{\gamma+} + (1-p)^{\gamma+}\ )^{1/\gamma+}$ for $m \geq 0$ and $\omega(p) = p^{\gamma-}/(\ p^{\gamma-} + (1-p)^{\gamma-}\ )^{1/\gamma-}$ for $m < 0$. The application of probability weighting for loss-frame and mixed-frame lotteries is not obvious, and is spelled out by Harrison and Swarthout [2016; Appendix B]. Probability weighting can easily lead to differences in the decision weights for gains and losses, and hence generate loss aversion or loss seeking, ceteris paribus

[19] In other words, the utility loss aversion for a loss of one penny is the same proportionally as the utility loss aversion of one million dollars.

values for α, β and λ.[20] One can usefully refer to this source of loss aversion as *probabilistic loss aversion*, following Schmidt and Zank [2008; p.213]. Thus loss aversion comes from *two* possible psychological pathways: utility loss aversion *and* probabilistic loss aversion. This is not a radical interpretation of CPT, but a direct consequence of the general form of CPT.

The upshot is that the conventional CPT model can be defined by parameters α, β, λ, γ+ and γ-, although extensions are easy to consider (e.g., to the Prelec probability weighting function (9), which significantly generalizes the Inverse-S function).

One final characteristic of CPT is that the argument of the utility functions is the difference between outcomes and a subjective reference point. Put aside how that reference point is determined, as proposed by Kahneman and Tversky [1979].[21] The issue is that EUT and RDU instead are viewed from the perspective of CPT as assuming that the argument of utility is net wealth. To be specific, assume a house endowment of $100 and a 50:50 lottery of +$10 or -$9 relative out of that endowment. If the endowment is viewed as the reference point, for pedagogic purposes, then CPT would assume that the arguments of the utility functions are +$10 and -$9, whereas EUT and RDU would be viewed as assuming that the arguments are $110 and $91. In this form, CPT does not nest EUT and RDU.

It is remarkable to see how light the previous evidence for CPT is when one weights the experimental and econometric procedures carefully. Moreover, a recent trend seems to be to declare any evidence for probability weighting, even if only in the gain domain, as evidence for CPT when it is

---

[20] Imagine that there is no probability weighting on the gain domain, so the decision weights are the objective probabilities, but that there is some probability weighting on the loss domain. Then one could easily have losses weighted more than gains, from the implied decision weights.

[21] There is a literature considering the endogenous determination of the reference point. For example, Kőszegi and Rabin [2007] consider the implications of loss aversion relative to a *stochastic reference point*, defined in terms of *subjective beliefs* about outcomes of the lottery. Recognizing that "... relatively little evidence on the determinants of reference points currently exists," (p. 1051), they make this notion operational by assuming that individuals use the EV of the lottery as their subjective belief about the lottery outcome. Gul [1991] developed the first endogenous reference point in his Disappointment Aversion, defined as the CE of the reference lottery. The notion of a stochastic reference point was developed by Sugden [2003] for subjective EUT and by Schmidt, Starmer and Sugden [2008] for CPT, although they still defined the reference in terms of some generic "status quo."

literally evidence for RDU. Table 1 summarizes a review of the literature, focusing only on controlled experiments, which has been the original basis of empirical claims for CPT.

Tversky and Kahneman [1992] gave their 25 subjects a total of 64 choices. Their subjects received $25 to participate in the experiment, but rewards were not salient, so their choices had no monetary consequences. The majority of data from their experiments used an elicitation procedure that we would now call a multiple price list, in the spirit of Holt and Laury [2002]. Subjects were told the expected value of the risky lottery, and 7 certain amounts were presented in a logarithmic scale, with values spanning the extreme payouts of the risky lottery. The subject made 7 binary choices between the given risky lottery and the series of certain amounts. To generate more refined choices, the subject was given a second series of 7 certainty equivalents for the same risky lottery, zeroing in on the interval selected in the first stage.[22] Furthermore, "switching" was ruled out, with the computer program enforcing a single switch between the risky lottery and the certain values.[23] All risky prospects used two prizes, and there were 56 prospects evaluated in this manner. One half of these prospects were in the gain frame, and one half were in the loss frame, with the latter being a "reflection" of the former in terms of the values employed.

A further 8 tasks involved mixed-frame gambles. In these choices the subject was asked to Fill-In-the-Blank (FIB) by entering a value $x that would make the risky lottery ($a, ½; $b, ½) equivalent to ($c, ½; $x, ½), for given values of a, b and c. The probabilities for the initial 56 choices over gain frame or loss frame choices were 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95 and 0.01, whereas the sole probability for the 8 mixed-frame choices was ½.[24]

---

[22] This variant is now called an *iterative* multiple price list by Andersen, Harrison, Lau and Rutström [2006].

[23] This variant is now called a *sequential* multiple price list by Andersen, Harrison, Lau and Rutström [2006].

[24] Wakker [2010; p. 175] sharply admonishes anyone that only uses one probability to elicit risk attitudes. Of course, Tversky and Kahneman [1992] used several probabilities in the gain frame and in the loss frame, so it is surprising that they did not do likewise in the mixed frame. No obvious "all-or-nothing" identification problems arise from their choice set design overall, but identification of probabilistic loss aversion is surely improved, in the broader sense, if one allows various probabilities in mixed frame lotteries.

Tversky and Kahneman [1992] estimate a structural model of CPT using non-linear least squares, and at the level of the individual. Remarkably, they then report the *median* point estimate, for each structural parameter, over the 25 estimated values. So over all 25 subjects, and using the earlier notation, the median value for α was 0.88, the median value of λ was 2.22, the median value of $\gamma+$ was 0.61, and the median value of $\gamma-$ was 0.69.[25]

These parameter estimates are remarkable in three respects, given the prominence they have received in the literature. First, whenever one sees point estimates estimated for individuals, one can be certain that there are many "wild" estimates from an *a priori* perspective,[26] so reporting the median value alone might be quite unrepresentative of the average value, and provides no information whatsoever on the variability across subjects. Second, there is no mention at all of standard errors, so we have no way of knowing, for example, if the oft-repeated value of λ is statistically significantly different from 1. Third, the median value of any given parameter is not linked in any manner to the median value of any other parameter: these are *not the values of some representative, median subject*, which is often how they are implicitly portrayed.[27] The subject that actually generated the median value of λ, for instance, might have had any value for α, β, $\gamma+$ and $\gamma-$.

These shortcomings of the Tversky and Kahneman [1992] study have not, to our knowledge, led anyone to replicate their experiments with salient rewards and report complete sets of parameter estimates with standard errors. The fault is not that of Tversky and Kahneman [1992], who otherwise employed quite modern methods, but the subsequent CPT literature. Anybody casually using these

---

[25] They also estimated β and apparently obtained *exactly* the same median value as α, which is quite remarkable from a numerical perspective.

[26] This issue is the focus of the use of "hierarchical" methods by Nilsson, Rieskamp and Wagenmakers [2011] and Murphy and ten Brincke [2017], which are in principle well-suited to handling this particular problem, which is not unique to CPT.

[27] Tversky and Kahneman [1992; p. 312] do note that the "parameters estimated from the median data were essentially the same." It is not clear how to interpret this sentence. It may mean that the median certainty-equivalents for the initial 56 choices, and the median values of $x for the final 8 choices, were combined to form a synthetic "median subject," and then estimates obtained from those data. The expression "median data" does not lead one to suspect that it was any one actual subject. Nor is there any reference to standard errors for these estimates.

estimates as statistically representative of anything must not care about rigor in empirical work.

Camerer and Ho [1994] is a remarkable study, with many insights. It was also one of the first to claim to estimate a structural model of CPT using ML (§6.1). The data employed were choice patterns from a wide range of studies, but the analysis was explicitly restricted to the gain frame (p. 188). Hence it could be said to be the first structural estimation of the RDU model, but not of a CPT model including losses.

Wu and Gonzalez [1996] focus entirely on the probability weighting function. They stress the point that they estimate the probability weighting function without having to make assumptions about utility functions, and view the need to make those assumptions as a methodological flaw. The reason it is said to be a flaw is that inferences about the probability weighting function could be confounded by mis-specifications of the true utility function (p.1678). This is true, but misses the critical point about structural estimation: estimates of one parameter should affect estimates of all other parameters, in general, since they are all "working" to explain the risk premium. So if one changes the estimate of α, *ceteris paribus* the estimates of β (and allowing $\beta \neq \alpha$), the estimate of λ must change according to which "exchange rate assumption" is employed. And that will flow on to changes in $\gamma+$ and $\gamma-$, since $\gamma+$, $\gamma-$ and λ jointly explain loss aversion. This is all *theory*: one cannot suddenly decouple estimates of one parameter from estimates of other parameters.[28]

Wu and Gonzalez [1996] propose a simple method for eliciting probability weights based on a series of choices with only two common outcomes, $200 or $240. Hence one could normalize utilities of these outcomes to 0 and 1, and avoid making any further assumptions about the utility function. Unfortunately this procedure was implemented in a non-salient, hypothetical choice task, and only for the gain frame (§4). When Wu and Gonzalez [1996] undertake ML estimation, via a non-linear least squares method, they assume a power utility function and also restrict themselves to gain frame choices

[28] The corollary is that one can only smile at attempts to identify utility loss aversion from one or two questions.

(§5). One could adapt the Wu and Gonzalez [1996] method for eliciting a probability weighting function for the gain frame to eliciting functions for the gain *and* loss frame, but they did not do so. Gonzalez and Wu [1999] estimate (non-parametric) probability weighting functions *and* utility functions for 10 subjects based on elicited certainty-equivalents for two-outcome lotteries solely in the gain frame. They at least employed salient rewards for their small number of subjects, but this is again just an RDU model, not CPT.

Harbaugh, Krause and Vesterlund [2002] paid for one of the 24 lotteries studied. Each lottery had two outcomes, with zero payment possible in every lottery. In half the lotteries the second payment was positive, and the other half of lotteries had a negative second payment; thus, there were no mixed frame lotteries. Each decision was between one of the lotteries and a certain amount, which was usually the expected value of the lottery. Decisions were presented to subjects on separate plastic cards, with each lottery presented as a pie chart with a "spinner" in the middle of the circle. Extra care was given to the method of task presentation, since subjects were as young as five years old. They do not undertake structural estimation of the CPT model, claiming (p.83) that, "Given our data it is not possible to simultaneously estimate both the probability weighting function and the value function." They do not consider utility loss aversion at all.

Mason, Shogren, Settle and List [2005] evaluate behavior over risky lotteries defined solely in a loss frame. They do not consider gain frame choices or mixed frame choices, but they do employ salient, real rewards.

Stott [2006] examines a wide range of parametric functional forms for CPT, but only considers data from hypothetical tasks defined over the gain frame.[29]

Fehr-Duda, Gennaro and Schubert [2006] paid subjects for one of 50 binary choices over lotteries with two outcomes. Half of the battery of losses were for gains, half were for losses, and there

[29] Stott [2006; p.113] notes that one choice was incentivized by scaling prizes down from nominal amounts up to £40,000 to an actual payment amounts up to £5. Average salient payments were just £2.13, which is effectively hypothetical.

were no mixed frame choices. For each lottery, an ordered MPL with 20 certain amounts was used to elicit a certainty equivalent. The certain amounts spanned the two outcomes of the lottery, so each subject faced 50 MPLs each with 20 rows. The utility loss aversion parameter λ was not estimated because of the absence of mixed frame lotteries (p. 295).

Fennema and van Assen [1998], Abdellaoui [2000], Etchart-Vincent [2004], Schunk and Betsch [2006], Abdellaoui, Bleichrodt and l'Haridon [2008] and Booij and van de Kuilen [2009] are widely cited for using the "tradeoff method" to estimate the utility function for losses. Fennema and van Assen [1998], Etchart-Vincent [2004] and Booij and van de Kuilen [2009] used hypothetical survey questions, with no real consequences. Abdellaoui [2000; p. 1502] and Schunk and Betsch [2006; p. 389] used real incentives for the gain frame, but hypothetical survey questions for the loss frame; neither asked any questions, hypothetical or real, in the mixed frame. Abdellaoui, Bleichrodt and l'Haridon [2008] asked real questions in the gain frame, but only hypothetical survey questions in the loss and mixed frames. Brooks and Zank [2005] used real losses, and focused on testing certain implications for choice patterns from utility loss aversion, not estimating the full CPT structure. In a similar vein, Brooks, Peters and Zank [2014] used real losses from a house endowment, and generated choice predictions based on assumed parametric values for a standard CPT specification. No CPT model was estimated from the 105 binary choices each subject made over gain, mixed and loss frames.

Rieskamp [2008] used "slightly real" rewards and all three frames. Subjects made binary choices over lotteries with outcomes between +€100 and -€100, one of 180 choices was selected for payment and realization, and then 5% of the outcome added or subtracted from an endowment of €15. So the rewards were salient, but not substantial. Nonetheless, this is a great advance from virtually all other studies. The structural estimates employed both α and β in power utility functions, with no discussion of the implications for identifying utility loss aversion. As it happened, the estimates of these two parameters were virtually identical, as in Tversky and Kahneman [1992]. The utility loss aversion parameter was constrained to be greater than 1, ruling out utility loss seeking. And the parameters for

the Inverse-S probability weighting functions were constrained to be less than 1 for both gains and losses. Pooled over all subjects, the estimates (p. 1455) were $\alpha = \beta = 0.91$, $\lambda = 1$, $\gamma+ = 0.69$ and $\gamma- = 0.71$. It is an open question what these estimates would be if λ had not "hit" the imposed lower boundary value.

Boij, van Praag and van de Kuilen [2010] estimate parametric models of CPT, but use hypothetical survey questions.

Bruhin, Fehr-Duda and Epper [2010] estimated parametric models of CPT that assumed that the utility loss aversion parameter λ was 1, noting wryly that "our specification of the value function seems to lack a prominent feature of prospect theory, loss aversion..." (p. 1382). They did this because their design only included lotteries in the gain frame and the loss frame, and none in the mixed frame. Estimation of utility loss aversion is logically impossible without mixed frame choices. They did provide real incentives for decisions, and employed an endowment of house money just as we did.

Pachur, Hanoch and Gummerum [2010] studied inmates in a UK prison, as well as UK non-prisoners. Choices were hypothetical, as the inmates received no compensation of any kind, and the non-prisoners received only a fixed £3 pound participation payment that was non-salient.

Nilsson, Rieskamp and Wagenmakers [2011] utilized the same "slightly real" data of Rieskamp [2008] and applied a Bayesian hierarchical model to estimate structural CPT parameters. They recognized the identification problem with power utility specifications when $\alpha \neq \beta$ indirectly. They initially simulated data using the popular point estimates from Tversky and Kahneman [1992], to test the ability of their model to recover them. They found that their model underestimated λ and that α was estimated to be much lower than β, rather than $\alpha \approx \beta$. They concluded (p.89) as follows:

> It is likely that these results are caused by a peculiarity of CPT, that is, its ability to account for loss aversion in multiple ways. The most obvious way for CPT to account for loss aversion is by parameter λ (after all, the purpose of λ is to measure loss aversion). A second way, however, is to decrease the marginal utility at a faster pace for gains than for losses. This occurs when α is smaller than β. Based on this reasoning, we hypothesized that the parameter estimation routines compensate for the underestimation of λ by assigning lower values to α than to β; in this way, CPT accounts

for the existing loss aversion indirectly in a manner that we had not anticipated.

Of course, this is just the *theoretical* identification issue that requires an "exchange rate assumption," as noted earlier and discussed in Köbberling and Wakker [2005; §7] and Wakker [2010; §9.6]. In any event, they optionally estimate all models with $\alpha = \beta$, and avoid this identification problem. Using the Inverse-S probability weighting function they reported Bayesian posterior modes (standard deviations) over the pooled sample of $\alpha = \beta = 0.91$ (0.16), $\lambda = 1.02$ (0.26), $\gamma+ = 0.68$ (0.11) and $\gamma- = 0.89$ (0.19). Unlike Rieskamp [2008], they did not constrain λ to be greater than 1.

These estimates are the Bayesian counterparts of random coefficients: hence each parameter is a distribution, which can be summarized in several ways. Reporting the mode is a more robust alternative to the mean, given the symmetric nature of their visual display of estimates, and the standard deviation provides information on the estimated variability across the 30 subjects, each making 180 binary choices. They find no evidence for utility loss aversion. Figure 3 shows the two probability weighting functions estimated, and implied decision weights. There is *very* slight evidence of probabilistic loss aversion for small probabilities, since there is slight risk loving over gains and extremely slight risk aversion for losses. For large probabilities this evidence suggests probabilistic loss seeking, albeit modest.[30]

Glöckner and Pachur [2012] undertook incentivized experiments, presented subjects with 138 binary choices over two-outcome lotteries spanning the gain, loss and mixed frame. A house endowment of €22 was used to cover potential losses of up to €9.90, from one lottery choice that was

[30] They also report (Table 2, p.91) ML estimates for each of the 30 subjects, and comment about the relative imprecision of these estimates compared to those obtained from the pooled Bayesian hierarchical methods. We agree with this likely outcome from individual-level estimates, as noted earlier, even when there are 180 binary choices per subject. Earlier they anticipated this finding, noting (p. 87) that they "... illustrate how single-subject ML, one of the most popular estimation methods for CPT (e.g. Harrison & Rutström 2009; Harless & Camerer, 1994; Stott, 2006), can produce extreme, implausible point estimates for parameters estimated with high uncertainty." The first two studies references here did not in fact estimate at the level of the individual, as claimed, and Stott [2006] used hypothetical choice data.

selected to play out.[31] Structural CPT estimates were generated, and one of their metrics for selecting parameters reflected likelihoods, rather than the unweighted hit rate. However, it appears that their estimation procedures do not generate standard errors, as illustrated by the tests of the hypothesis of stability of choices over two sessions.[32] Median estimates of parameters across individuals are reported (Table 4, p.27), following the unfortunate procedure of Tversky and Kahneman [1992], so one cannot say what any individual or representative agent's parameters were. EUT is compared (p. 29), but only with respect to the unweighted hit rate; there is no comparison to RDU, although a long list of *ad hoc* heuristics (Table 2, p. 26) are compared in terms of unweighted hit rates.

von Gaudecker, van Soest and Wengström [2011] estimated parametric models of CPT that assumed a complete absence of probability weighting, on both gain and loss frames.[33] They note clearly (p.675) that their specification entails

> ...departures from the original prospect theory specification. [...] it does not involve nonlinear probability weighting because our goal is to estimate individual-level parameters, and the dimension of the estimation problem is large already. Adding a parameter that is highly collinear with utility curvature in our experimental setup would result in an infeasibly large number of parameters, given the structure of our data. Furthermore, typical probability weighting functionals develop the highest impact at extreme probabilities, which are absent from our experiment.

Unfortunately these justifications are tenuous. The fact that the goal is individual-level estimation does not, by itself, have any theoretical implications for why one can pick and choose aspects of the CPT

---

[31] An unfortunate, but popular, use of a "lab currency" allowed them to state outcomes ranging between -€1000 and +€1200. These amounts were scaled down by 100 if chosen for payment. This procedure is unattractive, since it only affects behavior if subjects exhibit money illusion and are unable to infer the true payoff in the natural currency. If subjects exhibit money illusion then there is a loss of control over stimuli, by definition, since one does not know how the illusion manifests itself (e.g., non-linearly). In general it is better to deal with the budgetary consequences of presenting monetary amounts in the natural currency.

[32] They consider correlations of parameter estimates for each subject between the two sessions (p. 28), rather than a direct test of the hypothesis that the estimate *distributions* are the same.

[33] von Gaudecker, van Soest and Wengström [2011] employed a design in which all payments were to be sent to participants 3 months after their choices were made. This was to allow the design to vary the time of resolution of risk (now or in the future), without confounding that treatment with the timing of payment and discount rates. Their payoff configurations (Table 1, p. 669) include gain frame lotteries, mixed-frame lotteries, and no loss frame lotteries. Four of the seven payoff configurations have all risk resolved at the time of choice, although by means of a computer realization (raising issues of credibility).

model. Indeed, adding two parameters for probability weighting, does add minimally to the dimensionality of the estimation problem. But numerical convenience is hardly an acceptable rationale for mis-specification of the CPT model.

Colinearity with utility curvature is actually a theoretical point of some importance, and to be expected, and not an econometric nuisance. Indeed, it extends to colinearity with the utility loss aversion parameter, unless one assumes away *a priori* the possibility of probabilistic loss aversion by not estimating any probability weights. If one parameter plays a significant role in explaining the risk premium for an individual, then assuming it away surely biases conclusions about the strength and even sign of other psychological pathways. The final point, about not having sufficient variability in probabilities to estimate probability weighting functions, is even less clear. Their initial lottery choices varied the probability of the high prize from 0.25 to 0.5, 0.75 and 1; then their second stage choice interpolated the probability weights between one of these gaps (0 to 0.25, 0.25 to 0.5, 0.5 to 0.75, or 0.75 to 1) in grids of roughly 10 percentage points. Even from the first stage choices, if one assumes the popular Power or Inverse-S function then formally one only needs one interior probability to allow estimation. In fact, their design always has three interior probabilities of the first stage, and typically have refinements within one of those intervals. In sum, these arguments sound as though they were constructed "after the fact" of extensive numerical and econometric experimentation, and in the face of *a priori* unreliable numerical results.

Zeisberger, Vrecko and Langer [2012] estimate a structural CPT model from experimental data from 89 students, who earned €60 in an experiment a month prior, with payment only for the two sessions. One in ten students were paid, based on their choices for one random task out of 30. They elicited CE for lotteries in the gain, mixed and loss frames, using the Becker, De Groot and Marschak [1964] procedure. They estimated a "full" model for each subject in which all CPT parameters are jointly estimated using ML methods. For some reason standard errors needed to be generated by bootstrapping (e.g., Table 5, p. 375), and no hypothesis tests of parameters are presented. Median

estimates are presented (Table 4, p. 373), but at least interquartile ranges are also presented. No estimates for a representative agent are presented. Individual point estimates are presented (Table 5, p. 375ff.), and exhibit some "wild" estimates. This may be due to the small number of choices for each subject, although if the CE is reliably elicited it embeds more information than a binary choice. No comparison between CPT and other models is presented.

Abdelloui, l'Haridon and Paraschiv [2013] estimated parametric models of an RDU model defined over gains, but referred to this as a CPT model even if there were no losses at all in the stimuli. They did use real incentives, and told 65 couples that "they could be selected to play out one of their choices for real..."; it is not clear if one of the 65 would be selected for salient rewards, or this means that there was some probability that each couple could be selected. In any event, this is not a CPT model since losses played no part.

Murphy and ten Brincke [2017] estimate parametric structural models of CPT at the individual level, using mixed estimation methods to condition individual estimates based on pooled estimates. They assume that $\alpha = \beta$ in order to avoid making any "exchange rate assumption," but of course that is an assumption nonetheless. Although they used the flexible Prelec [1998] probability weighting function (9), they assumed the same probability weighting function for gains and losses, another restrictive assumption; their rationale (fn. 4) was "... parsimony and as a first pass, given the relatively low number of binary observations compared to the number of model parameters." They report (§6.1) values for λ of 1.11 and 1.18 in two sessions, one later than the other, but do not say if these were statistically significantly different from 1. Figure 4 displays estimated distributions, "given by medians of estimates" (fn. 9) for the pooled sample.[34] There appears to be no statistical significant loss aversion, with $\lambda \approx 1$, and virtually no probability weighting on average, with $\eta \approx \varphi \approx 1$. Figure 5 shows the implied utility functions and probability weighting functions, using the means of the distributions in Figure 4.

[34] Figure 4 is generated by considering the distribution for each parameter separately, since no information is provided about the full multivariate distribution spanning all parameters. It is unlikely that this would affect qualitative statements about the statistical (in)significance of utility loss aversion.

In summary, Table 1 shows that very few studies that use real, salient incentives for gain, loss and mixed frames. Those that meet these methodological criteria are shaded.

5. There Is a Reason We Compute Likelihoods: A Case Study of the Priority Heuristic

One of the valuable contributions of psychology is the focus on the *process* of decision-making. Economists have tended to focus on the characterization of properties of equilibria, and neglected the connection to explicit or implicit processes that might bring these about (Harrison [2008; §4]). Of course, this was not always so, as the correspondence principle of Samuelson [1947] dramatically illustrated. But it has become a common methodological difference in practice.[35] Brandstätter, Gigerenzer and Hertwig [2006] illustrate the extreme alternative, a process model that is amazingly simple and that apparently explains a lot of data. Their "priority heuristic" is therefore a useful case study in the statistical issues considered here, and the role of a ML estimation framework applied to a structural model.

The PH proposes that subjects evaluate binary choices using a sequence of rules, applied lexicographically. For the case of two non-negative outcomes, the heuristic is:

- If one lottery has a minimum gain that is larger than the minimum gain of the other lottery by ω percent or more of the maximum possible gain, pick it.
- Otherwise, if one lottery has a probability of the minimum gain that is at least ώ percent better than the other, pick it.
- Otherwise, pick the lottery with the maximum gain.

The parameters ω and ώ are each set to 10, based on arguments (p. 412ff.) about "cultural prominence." The heuristic has a simple extension to consider the probability of the maximum gain when there are more than two outcomes per lottery.

The key feature of this heuristic is that it completely eschews the notion of trading off the utility

[35] Some would seek to elevate this practice to define what economics is: see Gul and Pesendorfer [2007]. This is simply historically inaccurate and unproductive, quite apart from the debate over the usefulness of "neuroeconomics" that prompted it.

of prizes and their probabilities.[36] This is a bold departure from the traditions embodied in EUT, RDU, CPT, and even the SP/A theory of Lopes [1984]. What is striking, then, is that it appears to blow *every* other theory out of the water when applied to *every* conceivable decision problem. It explains the Allais Paradox, it explains the Reflection Effect, it explains the Certainty Effect, it explains the Fourfold Pattern, it explains Intransitivities, and it even predicts choices in "diverse sets of choice problems" better than a very long list of alternatives. It is notable that the list of opponents arrayed in the dramatic Figures 1 through 5 of Brandstätter, Gigerenzer and Hertwig [2006] do not include EUT with some simple CRRA specification and modest amounts of risk aversion, or even simple EV maximization.

However, there are three problems with the evidence for the PH.[37]

First, one must be extraordinarily careful of claims about "well known stylized facts" about choice, since the behavioral economics literature has become somewhat untethered from the facts in this regard. Consider behavioral Ground Zero, the Allais Paradox. It is now well documented that experimental subjects just do not fall prey to the Allais Paradox like decision-making lemmings when one presents the task for real payments and drops the word "millions" after the prize amount: see Conlisk [1989], Harrison [1994], Burke, Carter, Gominiak and Ohl [1996] and Fan [2002].[38] Subjects appear to crank out the EV when given real tasks to perform, and the vast majority behave consistently with EUT as a result.[39] This is not to claim that all anomalies or stylized facts are untrue, but there is a

---

[36] Of course, there are many such heuristics from psychology and the judgement and decision-making literature, noted explicitly by Brandstätter, Gigerenzer and Hertwig [2006; Table 3, p.417].

[37] Birnbaum [2008] also argues that the data used by Brandstätter, Gigerenzer and Hertwig [2006] was selective. Unfortunately, most of the data he would like to have included is hypothetical or effectively hypothetical. One might reasonably argue that much of the data originally used to test the PH was hypothetical, but one cannot mitigate such problems by just having more such data.

[38] This finding may be well documented, but it is apparently not well known. Birnbaum [2004] provides a comprehensive review of his own experimental studies of the Allais common consequence paradoxes, does not mention any of the studies referenced here, and then claims as a general matter that using real, credible payments does not affect behavior (p.105).

[39] Another concern with many of these stylized examples is that they are conducted on a between-subjects basis, and rely on comparable choices in two pairs of lotteries. Thus one must account for the presumed heterogeneity in risk attitudes when evaluating the statistical power of claims that EUT is rejected. Loomes and Sugden [1998] and Harrison, Johnson, McInnes and Rutström [2007] pay attention to this issue in different ways in their designs.

casual tendency in the behavioral economics literature to repeatedly assume stylized facts that are simply incorrect. Thus, to return to the Allais Paradox, if the PH predicts a violation, and in fact the data says otherwise for *motivated* subjects, doesn't this count directly as evidence *against* the PH?

The second problem with the evaluation of the performance of the PH against alternative models is that the *parameters* of those models, when the model relies on parameters, are taken from studies of different subjects and choice tasks. It is as if the CRRA of an EUT model from an Iowa potato farmer making fertilizer choices had been applied to the portfolio choices of a Manhatten investment banker. The naïve idea is that there is one, true set of parameters that define the model, and that is the model for all time and all domains.[40] This flies in the face of the default assumption by economists, and not a few psychologists (e.g., Birnbaum [2008]), that individuals might have different preferences over risk. It is notable that many applied researchers disregard that presumption and build tests of theories that assume homogenous preferences, but at least they are well aware that this is simply an auxiliary assumption made for tractability (e.g., Camerer and Ho [1994; p.186]). In any event, in those instances the researcher at least estimates parameters afresh in some ML sense for the sample of interest.

It is a different matter to estimate parameters for a model from responses from a random sample from a given population, and then see if those parameters predict data from another random sample from the *same population*. Although this tends not to be commonly done in economics, it is different than assuming that parameters are universal constants. For example, Birnbaum and Navarrete [1998; p.50] clearly seek to test model predictions "in the manner predicted in advance of the experiment" using parameters from comparable samples. One must take care that the stimuli and recruitment procedures match, of course, so that one is comparing apples to apples.

This issue is not peculiar to psychologists: behavioral economists have an embarrassing

[40] There is a folklore joke about how psychologists treat their models the way economists treat their toothbrush: everyone has their own. In this case it seems as though an old, discarded toothbrush is getting passed around to brush data set after data set.

tendency to just assume certain critical parameters casually, relying inordinately on the illustrative estimates of Tversky and Kahneman [1992] (very) critically reviewed in §4. For one celebrated example, consider Benartzi and Thaler [1995], who use laboratory-generated estimates from college students to calibrate a model of the behavior of U.S. bond and stock investors. Such exercises are fine as "finger mathematics" exemplars, but are no substitute for estimation on the comparable samples.[41] In general, economists tend to focus on in-sample comparisons of estimates from different models, although some have considered the formal estimation issues that arise when one seeks to undertake out-of-sample comparisons (Wilcox [2008][2011]). An example would be comparing behavior in one task context to behavior in another task context, albeit a context that is comparable.

The third problem with the PH is the fundamental one from the present perspective of thinking about models using a ML approach: it predicts with probability one or zero. So, surely, aren't there *some* interesting settings in which the heuristic must be completely wrong most or all the time? Indeed there are. Consider the comparison of lottery A in which the subject gets \$1.60 with probability p and \$2.00 with probability 1−p, and lottery B in which the subject gets \$0.10 with probability p and \$3.85 with probability 1−p. The PH picks A *every time*, no matter how low p is.[42] The minimum gain is \$1.60 for A and \$0.10 for B, and 10% of \$1.60 is \$0.16, greater than \$0.10.

At this point experimental economists are jumping up and down, waving their hands and pointing to the data from a massive range of experiments initiated by Holt and Laury [2002] with exactly these parameters. Their baseline experimental task presented subjects with an ordered list of 10

---

[41] This example also illustrates the danger of using estimates from one structural model and applying them casually to a different structural model. In this case the prospect theory parameters were held fixed and the best-fitting "evaluation horizon" determined from data. But when one estimates these parameters from responses in controlled experiments in which the evaluation horizon is varied as a treatment, they are not the same (Harrison and Rutström [2008; Appendix E3]).

[42] In fact, there is a threshold ϖ of the ratio of the expected values of the lotteries, *above* which the PH is assumed not to apply, and where probabilities and prizes are traded off in the usual EUT manner assuming risk neutrality (p. 425ff.). The parameter ϖ is set to 2, but is apparently not applied in the main tests of the predictive power of alternative theories in Brandstätter, Gigerenzer and Hertwig [2006; p.416ff]. With this modification, the PH predicts that A might also be selected because of EUT heuristics for $p \leq 0.1972$.

such choices, with p ranging from 0.1 to 1 in increments of 0.1. Refer to these prizes as their 1x prizes, where the number indicates a scale factor applied to all prizes. Identical tasks are reported by Holt and Laury [2002][2005] with 20x, 50x and 90x prizes, and by Harrison, Johnson, McInnes and Rutström [2005] with 10x prizes. Although we will want to do much, much better than just look at average choices, it is apparent from these data that the PH must be in trouble as a general model. Holt and Laury [2005; Table 1, p. 903] report that the average number of choices of lottery A is 5.2, 5.3, 6.1 and 5.7 over hundreds of subjects facing the 1x task, 6.0 over 178 subjects facing the 10x task, and 6.7 over 216 subjects facing the 20x task, in all cases for real payments and with no order effects. The predicted outcome for an EUT model assuming risk neutrality is for 4 choices of lottery A, and a modest extension of EUT to allow small levels of risk aversion would explain 5 or 6 safe choices quite well. In fact, using the CRRA utility function (1), any RRA between 0.15 and 0.41 would predict 5 choices, and any RRA between 0.41 and 0.68 would predict 6 choices (Holt and Laury [2002; Table 3, p.1649]).

But using the metric of evaluation of Brandstätter, Gigerenzer and Hertwig [2006] the PH would predict behavior here perfectly as well! This is because they count a success for a theory based on whether it predicts the *majority* choice correctly.[43] In the 10 choices of the Holt and Laury [2002] task, imagine that subjects picked A on average 5.000000001 times. An EUT model, in which the CRRA was set to around 0.25, would predict that the average subject picks lottery A 5 times and then switches to B for the other 5 choices, hence predicting almost perfectly in each of the 10 choices. But the PH gets almost 4 out of 10 wrong *every time*, and yet is viewed as a 100% successful theory by this metric.

This example shows exactly why it is a mistake to casually use the "hit rate" as a metric of

[43] To see this follow carefully the explanation in Brandstätter, Gigerenzer and Hertwig [2006; p.418] of how the vertical axis on their Figure 1 is created. There are 14 choice tasks being evaluated here. The PH predicted the *majority* choice in each of the 14 tasks, so it is given a predictive score of 100%. The "equiprobable" heuristic predicted 10 out of 14 of the *majority* choices, so it is given a predictive score of 71.4% = (10÷14)×100. The predictive accuracy measure is not calculated at the level of the individual choice, but instead using a summary statistic of those choices. Rieger and Wang [2008; Figure 3] make essentially the same point, but do not suggest the preferred evaluation in terms of likelihoods.

evaluation in such settings.[44] The likelihood approach instead asks the model to state the probability of observing the actual choice, conditional on some trial values of the parameters of the theory. ML then just finds those parameters that generate the highest probability of observing the data. For binary choice tasks, and independent observations, we know that the likelihood of the sample is just the product of the likelihood of each choice conditional on the model and the parameters assumed, and that the likelihood of each choice is just the probability of that choice. So if we have any observation that has zero probability, and the PH has many, the log-likelihood for that observation zooms off to minus infinity. Even if we set the likelihood to some minuscule amount, so we do not have to evaluate the logarithm of zero, the overall likelihood of the PH is *a priori* abysmal without even firing up any statistical package.[45]

Of course, this is true for any theory that predicts deterministically, including EUT. This is why one needs some formal statement about how the deterministic prediction of the theory translates into a probability of observing one choice or the other, and then perhaps also some formal statement about the role that structural errors might play, as explained in §2.

6. Point Estimates Are Not Data: A Case Study of Source Dependence

Abdellaoui, Baillon, Placido and Wakker [2011] (ABPW) conclude that different probability weighting functions are used when subjects face risky processes with known probabilities and uncertain processes with subjective processes. They call this "source dependence," where the notion of a source is relatively easy to identify in the context of an artefactual laboratory experiment, and hence provides

[44] There are some non-casual, semi-parametric estimation procedures for binary choice models that use the hit rate, such as the "maximum score" estimator of Manski [1975]. The literature on this estimator is reviewed by Cameron and Trivedi [2005; §14.7.2, p.483ff.].

[45] How would one modify the PH to make it worth testing against any real data at an individual level? Perhaps one could count how many of the criteria are pointing towards one lottery, and use that as an indicator of strength of preference. But this path seems *ad hoc*, would need weights on the criterion to avoid discontinuities in any likelihood maximization process using gradient methods, and is contrary to the *raison d'être* of the model.

the tightest test of this proposition. Unfortunately, their conclusions are an artefact of estimation procedures that do not worry about sampling errors.[46] These procedures are now often used in behavioral economics, and need to be examined carefully. In this case, they make a huge difference to the inferences one draws.

Consider the simple two-urn Ellsberg design, the centrepiece of their analysis. The known urn, K, has some objective distribution of balls with 5 colors. Design an experiment to elicit CE for a number of these urns, where the probabilities are generated objectively and vary from urn to urn. Assume the subject believes that.[47] The unknown urn, U, has some mix of balls of the same colors. Define some lotteries from the U urn, such as "you get $100 if blue comes out, otherwise $0 if any other color comes out" or "you get $100 if blue or red comes out, otherwise $0 if any other color comes out." Then elicit CE for these bets.

Now write out some models to describe behavior. For the K urn, which we call risk, and restricting to two prizes, X and x, for X>x, we have $w_K(p)\, u_K(X) + [\, 1 - w_K(p)\,]\, u_K(x)$ for some objective probability p of the bet being true and the subject earning X. We assume some specific functional forms for the probability weighting functions and utility functions, and estimate those parameters. For the U urn, which we call uncertainty, we propose $w_U(\pi)u_U(X) + [1-w_U(\pi)\,]u_U(x)$ for some subjective probability π of the bet being true and the subject earning X. So in the general models shown here the probability weighting function *and* the utility function are source-dependent. This is the model that ABPW propose: source dependence in both utility and probability weighting functions, which seems reasonable to test.

On the basis of *a priori* reasoning, some have suggested instead that we only have source-

---

[46] These estimation procedures are defended by Wakker [2010; Appendix A], so this is not just an inadvertent slip.

[47] If there is even the slightest concern by the subject that the experimenter might be manipulating the unknown urn strategically to reduce payouts, the Ellsberg paradox is explained: see Kadane [1992] and Schneeweis [1973]. This is why one should not rely on computer-generated realizations of random processes in behavioral research if at all possible. The experiment in ABPW was conducted entirely on a computer.

dependence in the probability weighting function, so we would have $w_K(p)\, u(X) + [\, 1 - w_K(p)\,]\, u(x)$ and $w_U(\pi)\, u(X) + [\, 1 - w_U(\pi)\,]\, u(x)$. Of course this is a testable restriction of the general model to $u_K(z) = u_U(z)$ for $z \in \{X, x\}$. There is an obvious, symmetric special case with source-dependence only in the utility function: $w(p)\, u_K(X) + [\, 1 - w(p)\,]\, u_K(x)$ and $w(\pi) u_U(X) + [1-w(\pi)]\, u_U(x)$. Again this is a testable restriction of the general model to $w_K(p) = w_U(\pi)$ for $p=\pi$. Indeed, it is the alternative hypothesis offered by (Vernon) Smith [1969] in a comment on Ellsberg.

These models can be estimated using data generated from the "Ellsberg experiment" of ABPW. In this experiment each subject was asked to state CE for 32 bets based on the K urn, and 32 bets based on the U urn, generating 64 observations per subject. They propose a power utility function defined over prizes z normalized to lie between 0 and 1, $u(z) = z^{\varrho}$, where the parameter ϱ is allowed to take on different values depending on the source K or U. So if S is defined to be a binary variable such that S=1 when the U process was used and S=0 when the K process was used, one estimates ϱ_K and ϱ_U in $\varrho = \varrho_K + \varrho_U\, S$ and then there is an obvious hypothesis test that $\varrho_U = 0$ in order to test for source independence with respect to the utility function.

The probability weighting function is due to Prelec [1998], which exhibits considerable flexibility and was defined earlier in (9): $w(p) = \exp\{-\eta(-\ln p^{\varphi}\}$, where w(p) is for choices from the K process. The same function $w(\pi)$ can be defined for the choices from the U process. It is similarly possible to estimate linear functions of the structural parameters φ and η to test for source-independence: $\varphi = \varphi_K + \varphi_U\, S$ and $\eta = \eta_K + \eta_U\, S$. The obvious hypothesis test for source independence in probability weighting is that $\varphi_U=0$ *and* $\eta_U=0$.

The experimental data of ABPW can be used to estimate these structural parameters and undertake the hypothesis tests for source independence. Each of 66 subjects was presented with 32 tasks in which they were asked to indicate "switch points" between a bet on some outcome from drawing a ball from the urn and a certain amount of money. Half of the bets were based on draws from the K urn, and half from bets based on the U urn. The CE were ordered increments between 0€ and

25€, using 50 rows in a multiple price list elicitation. The end-result for each subjective lottery is a certain amount of money which is evaluated as being just less valuable than the lottery, and a certain amount of money which is evaluating as being just more valuable than the lottery. The switch point is enforced for the subject, and involves an increment of 0.5€.[48] Thus we have 64 binary lottery comparisons for each subject over 32 tasks.[49] Each subject was told that one of the 32 tasks would be selected for payment, thereby incentivizing them to respond truthfully.

These binary comparisons can be used to generate ML estimates of the structural parameters. Each comparison involves the "left" lottery $RDU_L = w_K(p)\, u_K(X) + [\, 1 - w_K(p)\,]\, u_K(x)$ or $RDU_L = w_U(\pi)\, u_U(X) + [\, 1 - w_U(\pi)\,]\, u_U(x)$, and the "right" lottery $RDU_R = u_K(Z)$ for the certain amount Z.[50] The latent index $\nabla RDU = RDU_R - RDU_L$ can then be calculated. This latent index, based on latent preferences, is then linked to observed choices using a standard cumulative normal distribution function as explained in §2: prob(choose lottery R) = $\Phi(\nabla RDU)$. In addition, we assume a behavioral error specification, also discussed in §2, leading to the latent index $\nabla RDU = ((RDU_R - RDU_L)/\nu)/\mu$ instead, where ν is a familiar normalizing term for each lottery pair L and R.

Thus the likelihood of the observed responses, conditional on these specifications being true, depends on the estimates of ϱ, φ, η and μ given the above statistical specification and the observed choices. The conditional log-likelihood is then

$$\ln L(\varrho, \varphi, \eta, \mu; y, S) = \sum_i [\, (\ln \Phi(\nabla RDU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla RDU)) \times \mathbf{I}(y_i = -1))\,]$$

where $\mathbf{I}(\cdot)$ is the indicator function, $y_i = 1 (-1)$ denotes the choice of the Option R (L) lottery in choice

---

[48] The use of an enforced switch points of this kind is studied in detail in Andersen, Harrison, Lau and Rutström [2006], and is referred to there as the sequential multiple price list procedure.

[49] Each subject actually made 3,200 choices, since each of the 50 rows involved a binary choice. However, the choices either side of the switch point are correlated by design in the sequential multiple price list procedure, and contain no extra information. The results are qualitatively the same if one includes all choices, including the implied ones.

[50] An alternative specification would use $RDU_R = u_K(Z)$ for the certain amount when comparing to the risky lottery based on the K urn, and $RDU_R = u_U(Z)$ for the certain amount when comparing to the risky lottery based on the U urn. In effect, this specification assumes that the source-dependence is "contextual" and defined by the choice context. Using this specification makes no difference to the qualitative conclusions.

task i, and S is the binary variable defined earlier to denote the U source.

Table 2 reports hypothesis tests and selected estimates from ML estimation of this model. Column 1 shows the ID number of the subject, columns 2 through 6 report *p*-values of hypothesis tests of source dependence, and columns 8 and 9 report the point estimate and standard error for the φ_U parameter of the probability weighting function. As often happens with estimation at the level of the individual, numerical instability arises for some subjects: in the present case 13 of the 66 subjects were dropped due to the inability to estimate the model.

The *p*-values indicate striking evidence for source *independence*, and are sorted using the values on column 6. Those *p*-values less than 0.1, implying rejection of the null hypothesis of source independence, are shaded: there are very few shaded cells. Column 2 shows the *p*-values on the hypothesis test for the utility function, and the lowest three *p*-values are 0.14, 0.17 and 0.18 for subjects 18, 8 and 29, respectively. Columns 3 and 4 report *p*-values for each of the parameters of the probability weighting function, and column 5 and 6 report joint hypothesis tests. Only 4 subjects violate source independence with respect to the φ and β parameters.

Columns 8 and 9 report the point estimates of the φ_U parameter to illustrate a concern with the manner in which ABPW draw inferences from these data. For each subject they calculate the values of the parameters in two steps. First, they use non-linear least squares for ϱ, $w_K(0.5)$ and $w_U(0.5)$ using choice tasks where one can *a priori* assume the value of $p = \pi = 0.5$. Thus they do not estimate the parameters φ and η in this step, but directly estimate the decision weights. Second, conditional on the point estimate for ϱ, they calculate the values of φ and η that minimize a quadratic distance metric using choice tasks for which one can *a priori* assume the value of p or π to be ⅛, ¼, ⅜, ⅝, ¾ or ⅞.

One immediate concern with this approach is that sample errors in the estimation of ϱ in the first step are assumed away in the second step, likely resulting in an understatement of sample errors in the estimation of φ and η. The fact that sample errors are not reported in the first step does not mean that they are zero. Indeed, it is common for all statistical packages to have non-linear least squares

procedures with several ways of calculating standard errors. Another concern with this approach is that the estimates of $w_K(0.5)$ and $w_U(0.5)$ in the first step appear to play no role in constraining the estimates of φ and η in the second step: for any values of φ and η there is an implied value of $w_K(0.5)$ and $w_U(0.5)$, and these procedures do not respect that connection, which is a matter of theoretical consistency.

These problems are compounded when inferences are drawn solely on the vector of point estimates of some parameter, with no regard for possible sample errors. For example, ABPW conclude that utility is linear because the *median* values of the *point estimates* of ϱ for the K and U processes are not statistically different from 1 using a sign test. As it happens, this conclusion is generally correct, but for a very different reason: the point estimates of ϱ have very large sample errors. So the statistical result arises because of poor estimates, and does not arise because subjects fail to exhibit diminishing marginal utility.[51] The median point estimate from our ML specification is 1.003, but the median standard error is 0.67 (and significantly different from zero, if one is can use this metric descriptively). This finding, of course, makes one particularly concerned about the assumption that the standard error of $\varrho=0$ when making inferences about the parameters of the probability weighting, φ and η.[52]

The key finding from ABPW is that there is source dependence in the probability weighting function. They first examine the median of the *point estimates* of φ and η, noting that in their estimations $(\varphi_K+\varphi_U) < \varphi_K < 1$ and $(\eta_K+\eta_U) \approx \eta_U$ for these median values and our notation.[53] Of course, without any sense of the precision of these estimates of median values, they have no inferential value whatsoever. In fact, columns 3, 4 and 5 confirm that these inferences from the median point estimates are generally invalid.

---

[51] Of course, the latter claim may be true: with large sample errors one simply cannot say. The claim that utility functions are linear in RDU models flies in the face of a wide range of data collected from laboratory experiments on the matter, surveyed in Harrison and Rutström [2008].

[52] ABPW also use inferences based on the experiment with objective probabilities to constrain their experimental design with subjective probabilities, specifically the inference that there was no source dependence in utility. Although we agree with that inference, for rather different statistical reasons, it is perilous to build experimental designs in the domain of subjective probability that rely on such maintained behavioral assumptions that are inferred from the domain of objective probability.

[53] In which the subscript U denotes the deviation from the estimate with the subscript K.

ABPW next examine the values of two indices that are intended to convey the "insensitivity" and "pessimism" of the probability weighting function. These are derived from ordinary least squares approximations of the probability weighting function evaluated at the point estimates of φ and η. They conclude that these two indices are significantly different for the K and U processes, but of course this is an approximation based solely on point estimates. The same hypothesis test can be undertaken directly, and more powerfully, by examining the φ and η parameters themselves for each process and individual. Table 2 does precisely that, recognizing the standard errors in these estimates, and the conclusion is apparent.

The need to pay attention to the precision of estimates is so important, in terms of the way in which empirical inference in behavioral economics seems to be progressing, as to warrant an alternative demonstration. Consider the φ parameter. Table 2 lists the point estimate and standard error of the φ_U coefficient for each subject, where $\varphi_U = 0$ is consistent with source independence with respect to this parameter. If one tests whether the point estimates of φ_U in column 7 of Table 2 are significantly different from 0, one would conclude that they are. A two-sided sign test has a *p*-value of 0.027, and a two-sided *t*-test has a *p*-value of 0.0001. But these tests completely ignore the imprecision of these estimates, shown in column 8. If one just looks at the standard errors, it is apparent that these point estimates are not precisely estimated. The *p*-values in column 3 do formally and properly what these sign tests and *t*-tests do incorrectly, and the difference in conclusions is dramatic.

Of course, any failure to reject a null hypothesis could be an artefact of sample sizes being too small. True, but that provides no rationale for ignoring sample errors. Indeed, this point is so obvious that it leads one to question any statistical methodology that could claim to draw inferences from samples that were arbitrarily small. What if the sample of ABPW had been 6 instead of 66? They could still have drawn their conclusions, based on looking at the median values of individual estimates, despite the patent imprecision that N=6 should alert anyone to. The use of a sample of N=66 just confers this method with an illusion of statistical, large-sample validity.

ABPW (p. 704) note that

> We also analyzed our data using probabilistic choice-error theories and econometric maximum likelihood estimations. The results [...] all agree with the results reported here. All estimations of utilities and weighting functions were done at the individual level.

In fact, the *maximum likelihood* estimation referred to here was only undertaken with the pooled sample, and not at the level of the individual. The individual "estimations" were only undertaken with the least squares and quadratic distance procedures.

In conclusion, the evidence for source dependence is missing. This does not mean that the behavioral phenomenon is missing. Indeed, it is intuitively plausible once one moves to the domain of subjective probabilities, or where objective probabilities are presumed to arise from some inferential process.[54] But we should not mistake our intuition for the evidence, as comforting as that might be.

7. Conclusion: Where Are the Methodologists?

The overall methodological lesson is that one cannot do behavioral econometrics effectively without knowing structural theory, and one cannot design experiments efficiently without knowing structural theory, and having an eye to what identification issues will arise. Of course, "identification" is a matter for theory, as much as econometric method: it basically means the same thing as proposing an operationally meaningful theory. So there is a methodological trinity here.

There are some low-hanging methodological issues reviewed here, and some subtle issues. To take the low-hanging cases first, how have philosophers of science and methodologists allowed CPT to survive on the basis of the flimsy empirical evidence transparently before us? If it is not their job to maintain intellectual standards across erstwhile intellectual silos, then whose is it? One reasonable response is that this is what experimental economists should do, since they are the methodological bridge between theory and evidence. In effect, they have to operate at both coalfaces.

The subtle methodological issues involve the selection of metrics for normative evaluation, now

[54] For example, by the application of Bayes Rule or the reduction of compound lotteries.

that behavioral economics has given us a rich array of alternative *descriptive* models to the traditional models.[55] It is not automatically true that the traditional models are the normatively attractive models, even if they are often mis-characterized as such. To motivate richer discussion of these issues we need more examples where "getting the positive economics right" matters for the welfare evaluation of policies of substance. Armed with normative tradeoffs of substance, rather than abstract constructions *per se*, we will then have to address the normative methodological issues.

[55] See Harrison and Ross [2016][2017] for a statement of the philosophical issues raised.

Figure 1. Consumer Surplus Across EUT CRRA Coefficients

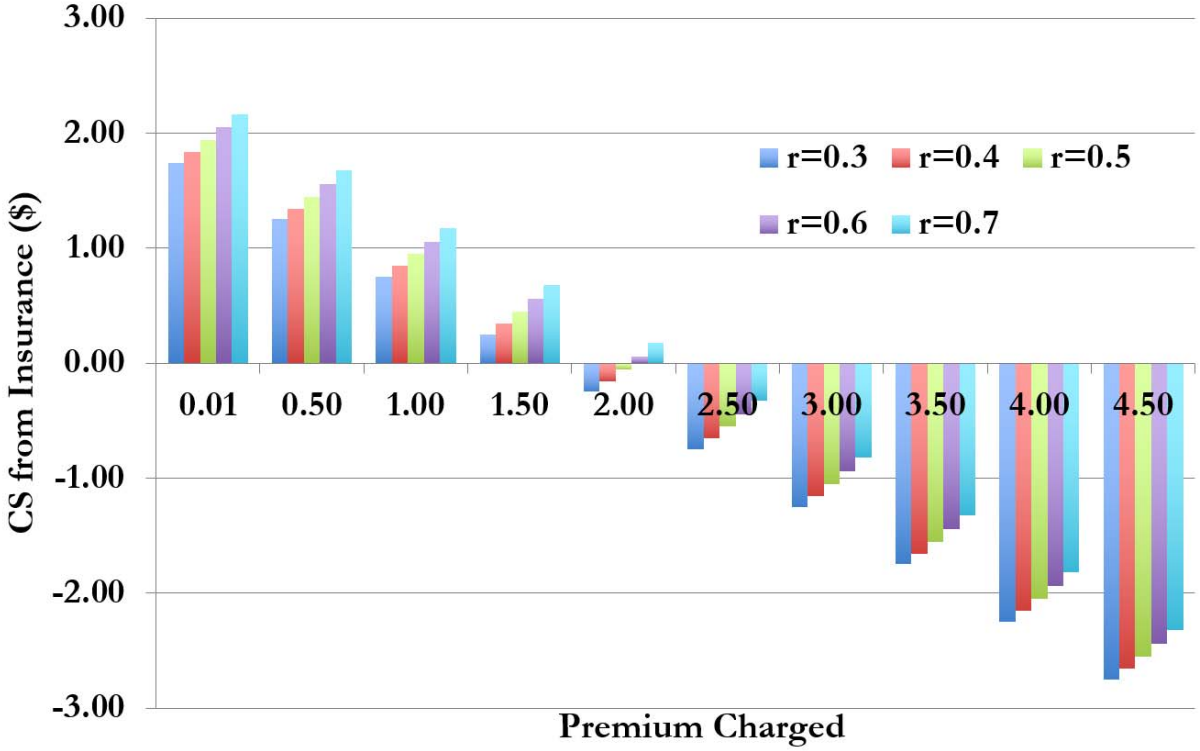

Table 1: The Existing Literature Claiming to Estimate CPT

Study	Rewards	Frames	Comments
Tversky and Kahneman [1992]	Non-salient	Gain, Loss	"Median" estimates reported.
Camerer and Ho [1994]	Real	Gain	
Wu and Gonzalez [1996]	Hypothetical	Gain	
Gonzalez and Wu [1999]	Real	Gain	
Fennema and van Assen [1998]	Hypothetical	Gain, Loss, Mixed	
Abdellaoui [2000]	Real Hypothetical	Gain Loss	
Schmidt and Traub [2002]	Hypothetical	Gain, Loss	
Harbaugh, Krause and Vesterlund [2002]	Real	Gain, Loss	Assumes no utility loss aversion. Claim to be unable to jointly estimate probability weighting and the value function.
Pennings and Smidts [2003]	Hypothetical	Gain, Loss†	
Etchart-Vincent [2004]	Hypothetical	Loss	
Mason, Shogren, Settle and List [2005]	Real	Loss	
Schunk and Betsch [2006]	Real Hypothetical	Gain Loss	
Stott [2006]	"Slightly Real"¶	Gain	Does not mention loss aversion.
Fehr-Duda, Gennaro, and Schubert [2006]	Real	Gain, Loss	Assumes no utility loss aversion.
Abdellaoui, Bleichrodt and l'Haridon [2008]	Real Hypothetical	Gain Loss, Mixed	
Rieskamp [2008]	"Slightly Real"‡	Gain, Loss, Mixed	Constrained to show loss aversion.
Booij and van de Kuilen [2009]	Hypothetical	Gain, Loss	

Booij, van Praag and van de Kuilen [2010]	Hypothetical	Gain, Loss, Mixed	
Bruhin, Fehr-Duda and Epper [2010]	Real	Gain, Loss	Assumes no utility loss aversion.
Pachur, Hanoch and Gummerum [2010]	Hypothetical	Gain, Loss Mixed	
von Gaudecker, van Soest and Wengström [2011]	Real	Gain, Mixed	Assumes no probability weighting.
Nilsson, Rieskamp and Wagenmakers [2011]	"Slightly Real"‡	Gain, Loss, Mixed	
Glöckner and Pachur [2012]	Real	Gain, Loss, Mixed	"Median" estimates reported, apparently with no standard errors
Zeisberger, Vrecko and Langer [2012]	Real §	Gain, Loss, Mixed	Becker, DeGroot and Marschak [1964] method used to elicit certainty-equivalents.
Abdelloui, l'Haridon and Paraschiv [2013]	Real	Gain	
Scholten and Read [2014]	Non-Salient	Gain, Loss	Assumes no utility loss aversion.
Balcombe and Fraser [2016]		Gain	Does not mention loss aversion.
Bouchouicha and Vieder [2016]	Hypothetical	Gain, Loss	Assumes no utility loss aversion.
Murphy and ten Brincke [2017]	Real	Gain, Loss, Mixed	

Notes: † Subject elicitations were all in the gain frame, but the authors' assumed (p. 1254) some positive reference point in their analysis and treated gains below that as "losses" for the purposes of analysis.

‡ Subjects made binary choices over lotteries with outcomes between +€100 and -€100, one of 180 choices was selected for payment and realization, and then 5% of the outcome added or subtracted from an endowment of €15.

¶ Lottery prizes up to £40,000 were included in binary lottery choices. Each subject was given a fixed £3, and one of the 90 choices selected, re-scaled so that the maximum prize would be £5, and then played out.

§ One subject in ten was selected for payment, but the losses in that case were substantial (up to €60) out of an endowment that had been earned in a previous task in that session.

Figure 3: Probability Weighting and Decision Weights from Mode of Bayesian Posterior Distributions Estimated by Nilsson, Rieskamp and Wagenmakers [2011]

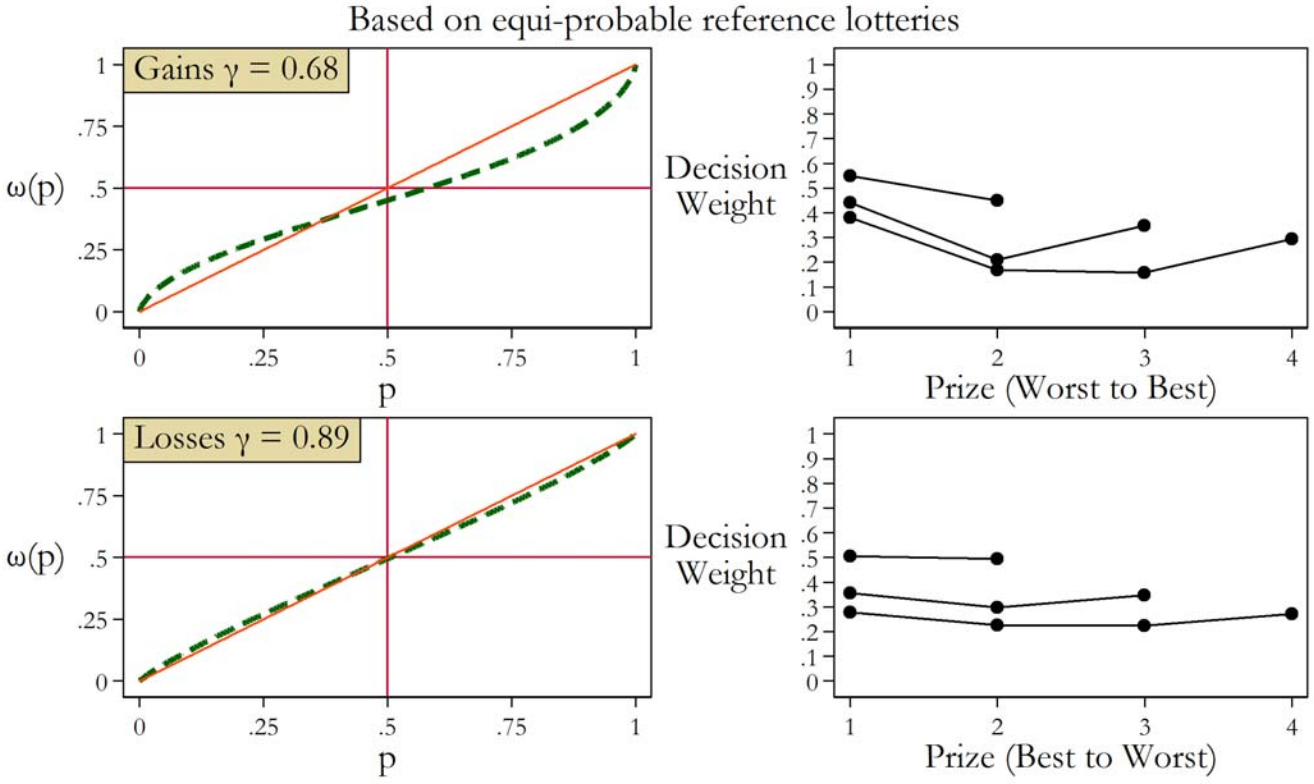

Figure 4: CPT Model Estimates from Murphy and ten Brincke [2017]

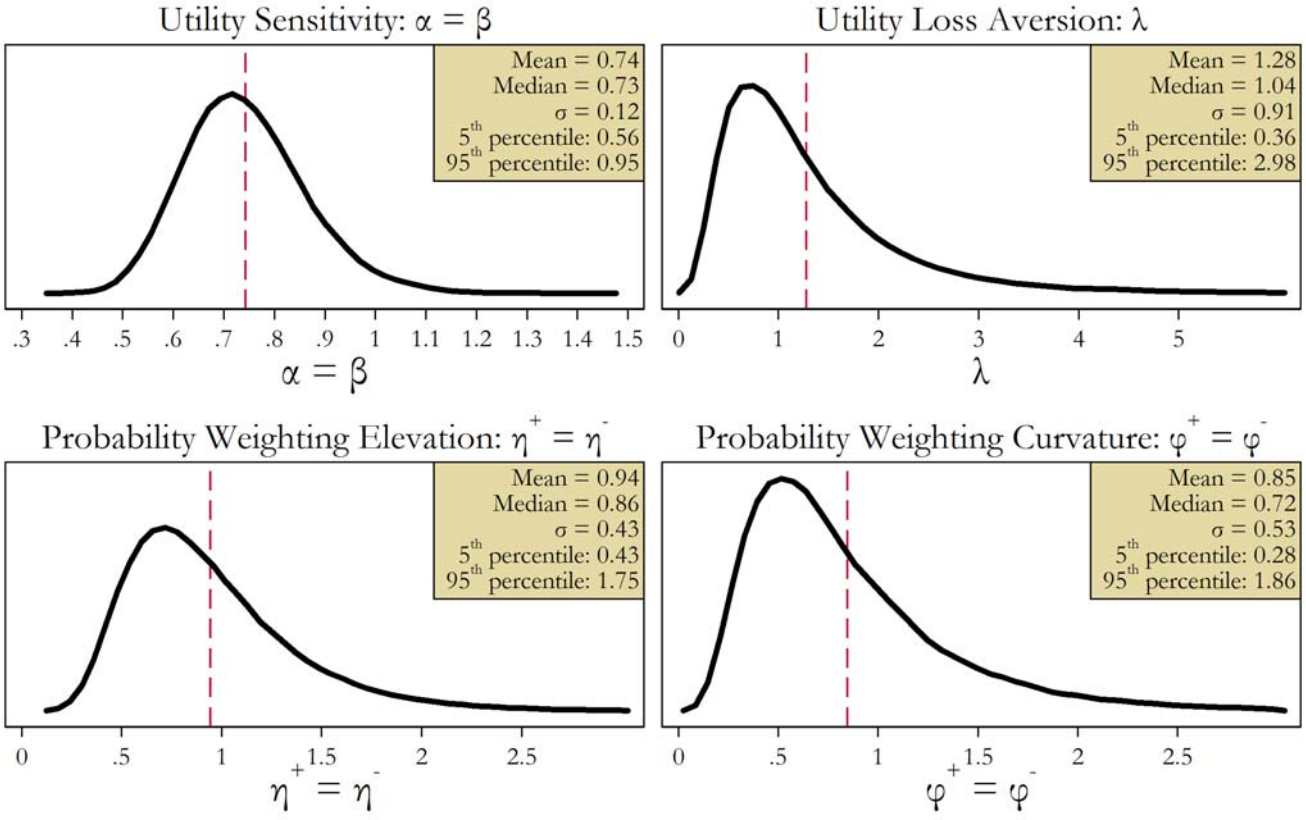

Figure 5: CPT Model Implied by Estimated Means from Murphy and ten Brincke [2017]

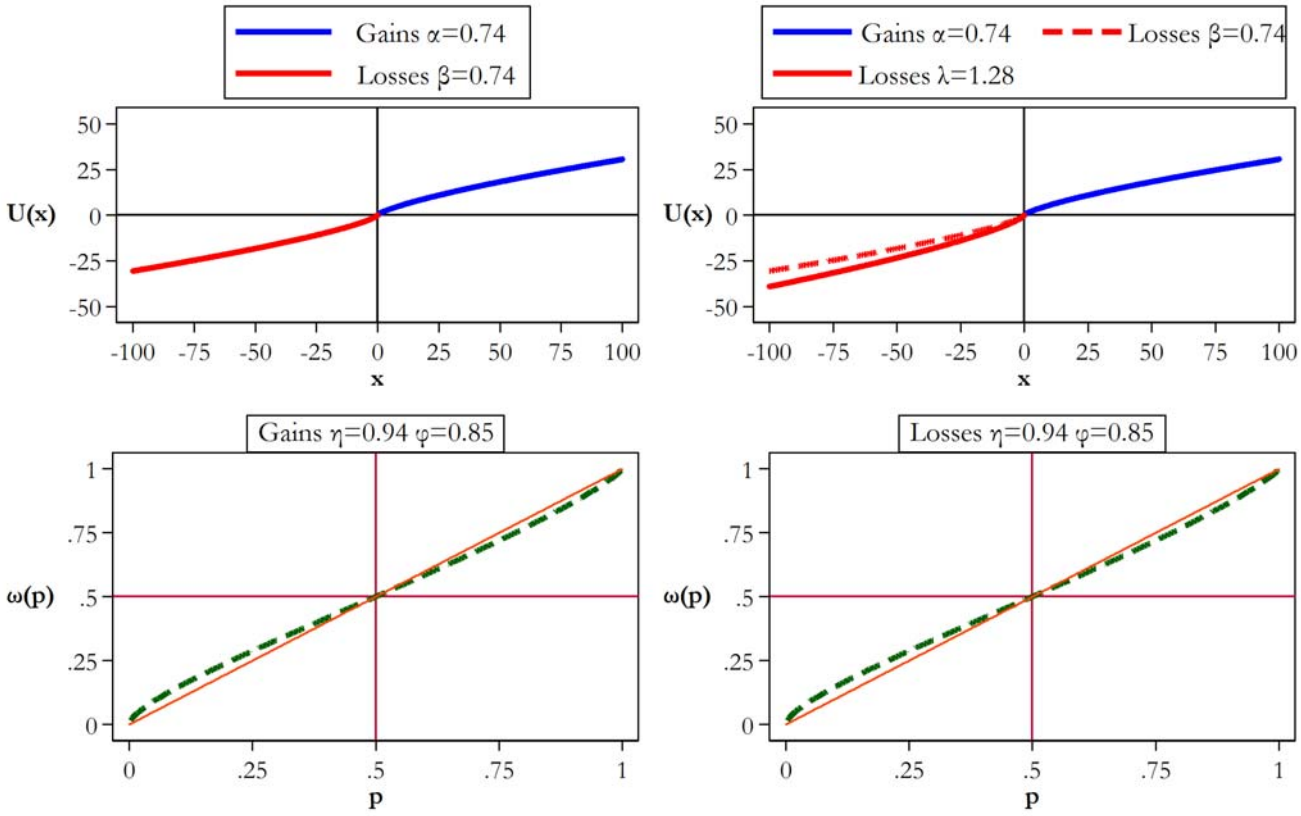

Table 2: Maximum Likelihood Estimates of the Source-Dependence Model

Subject	*p*-value on tests of source-dependence					Point Estimate of φ_U	Standard Error on Estimate of φ_U
	ϱ_U	φ_U	η_U	φ_U and η_U	ϱ_U, φ_U and η_U		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
53	0.59	0.01	0	0	0	-0.494	0.19
57	0.99	0.07	0.01	0	0	-0.301	0.17
29	0.18	0.03	0	0	0.01	-0.182	0.08
6	0.46	0.39	0	0.02	0.04	-0.232	0.27
71	0.92	0.48	0.39	0.38	0.22	-0.428	0.61
8	0.17	0.4	0.1	0.22	0.22	-0.413	0.49
50	0.37	0.23	0.61	0.23	0.23	-0.68	0.57
61	0.69	0.88	0.16	0.33	0.28	-0.069	0.44
18	0.14	0.44	0.19	0.27	0.36	0.223	0.29
7	0.64	0.27	0.15	0.3	0.5	0.43	0.39
45	0.53	0.54	0.37	0.61	0.51	-0.509	0.84
46	0.66	0.23	0.25	0.38	0.52	-0.54	0.45
55	0.43	0.9	0.18	0.41	0.56	0.067	0.51
34	0.83	0.61	0.8	0.38	0.58	0.387	0.75
5	0.95	0.86	0.3	0.58	0.63	0.059	0.33
10	0.61	0.62	0.5	0.75	0.69	0.222	0.45
3	0.55	0.3	0.54	0.49	0.7	-0.459	0.44
59	0.89	0.9	0.4	0.68	0.73	0.095	0.75
20	0.51	0.39	0.46	0.65	0.73	-0.615	0.71
40	0.62	0.87	0.28	0.54	0.74	0.073	0.46
48	0.56	0.53	0.52	0.53	0.74	-0.232	0.37
66	0.83	0.98	0.32	0.55	0.75	-0.006	0.26
38	0.41	0.65	0.33	0.62	0.75	-0.309	0.68
26	0.49	0.68	0.36	0.64	0.75	-0.279	0.68
60	0.69	0.55	0.34	0.6	0.77	-0.361	0.6
27	0.65	0.93	0.37	0.6	0.78	0.065	0.71
16	0.6	0.93	0.36	0.62	0.81	0.033	0.4
11	0.88	0.92	0.51	0.78	0.81	-0.053	0.51
24	0.65	0.9	0.62	0.86	0.87	0.126	1
39	0.42	0.95	0.86	0.98	0.88	0.048	0.72
17	0.54	0.91	0.88	0.99	0.9	-0.153	1.36
23	0.86	0.58	0.93	0.85	0.9	-0.561	1
54	0.66	0.73	0.91	0.79	0.92	-0.333	0.96
31	1	0.96	0.64	0.9	0.93	-0.095	1.9
28	0.89	0.91	0.53	0.82	0.93	0.034	0.31
2	0.86	0.52	0.76	0.81	0.93	-0.292	0.45

64	0.59	0.85	0.8	0.95	0.93	-0.271	1.41
33	0.8	0.78	0.77	0.91	0.94	-0.329	1.18
15	0.8	0.96	0.55	0.83	0.94	0.033	0.75
44	0.62	0.97	0.88	0.98	0.94	-0.03	0.88
80	0.77	0.69	0.82	0.87	0.96	-0.442	1.11
42	0.82	0.85	0.85	0.9	0.98	-0.41	2.12
62	0.8	0.91	0.84	0.98	0.98	-0.331	2.9
30	0.87	0.96	0.7	0.93	0.99	0.155	2.8
43	0.98	0.99	0.74	0.94	0.99	0.021	1.65
9	0.89	0.82	0.87	0.94	0.99	-0.403	1.79
47	0.97	0.93	0.77	0.96	0.99	0.151	1.63
13	0.89	0.91	0.82	0.96	0.99	-0.083	0.7
106	1	0.91	0.88	0.98	1	-0.393	3.49
35	0.9	0.94	0.92	0.99	1	-0.097	1.22
49	1	0.98	0.87	0.99	1	-0.093	3.93
19	0.96	0.91	0.91	0.98	1	0.38	3.19
41	0.88	0.95	0.96	1	1	-0.142	2.24

References

Abdellaoui, Mohammed, "Parameter-Free Elicitation of Utilities and Probability Weighting Functions," *Management Science*, 46, 2000, 1497-1512.

Abdellaoui, Mohammed; Baillon, Aurélien; Placido, Lætitia and Wakker, Peter P., "The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation," *American Economic Review*, 101, April 2011, 695-723.

Abdellaoui, Mohammed; Bleichrodt, Han, and l'Haridon, Olivier, "A Tractable Method to Measure Utility and Loss Aversion under Prospect Theory," *Journal of Risk and Uncertainty*, 36, 2008, 245-266.

Abdellaoui, Mohammed; Bleichrodt, Han, and Paraschiv, Corina, "Loss Aversion under Prospect Theory: A Parameter-Free Approach," *Management Science*, 53(10), October 2007, 1659-1674.

Abdellaoui, Mohammed; Bleichrodt, Han, and Paraschiv, Corina, "Measuring Loss Aversion under Prospect Theory: A Parameter-Free Approach," *Management Science*, 53(10), October 2007, 1659-1674.

Abdellaoui, Mohammed; l'Haridon, Olivier, and Paraschiv, Corina, "Individual vs. Couple Behavior: An Experimental Investigation of Risk Preferences," *Theory and Decision*, 75(2), 2013, 175-191.

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Journal of Risk & Uncertainty*, 48, 2014, 207-229.

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 Igel, and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), May 2008, 583-618.

Andersen, Steffen; Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet, "Dual Criteria Decisions," *Journal of Economic Psychology*, 41, April 2014a, 101-113.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Discounting Behavior: A Reconsideration," *European Economic Review*, 71, November 2014b, 15-33.

Andersen, Steffen; Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet, "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion," *International Economic Review*, 2017 forthcoming.

Andreoni, James, and Sprenger, Charles, "Risk Preferences Are Not Time Preferences,"*American Economic Review*, 102(7), December 2012, 3357-3376.

Angrist, Joshua D. and Pischke, Jörn-Steffen, *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press, 2009.

Becker, Gordon M.; DeGroot, Morris H., and Marschak, Jacob., "Measuring Utility By A Single-Response Sequential Method," *Behavioral Science*, 9, July 1964, 226-232.

Benartzi, Shlomo, and Thaler, Richard H., "Myopic Loss Aversion and the Equity Premium Puzzle," *Quarterly Journal of Economics*, 111(1), February 1995, 75-92.

Birnbaum, Michael H., "Causes of Allais Common Consequence Paradoxes: An Experimental Dissection," *Journal of Mathematical Psychology*, 48, 2004, 87-106.

Birnbaum, Michael H., "Evaluation of the Priority Heuristic as a Descriptive Model of Risky Decision Making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006)," *Psychological Review*, 115(1), January 2008, 253-260.

Birnbaum, Michael H., and Navarrete, Juan B., "Testing Descriptive Utility Theories: Violations of Stochastic Dominance and Cumulative Independence," *Journal of Risk and Uncertainty*, 17, 1998, 17-49.

Bleichrodt, Han; Pinto, J.L, and Wakker, Peter P., "Using Descriptive Findings of Prospect Theory to Improve the Prescriptive Use of Expected Utility," *Management Science*, 47, 2001, 1498-1514.

Booij, Adam S., and van de Kuilen, Gijs, "A Parameter-Free Analysis of the Utility of Money for the General Population Under Prospect Theory," *Journal of Economic Psychology*, 30, 2009, 651-666.

Booij, Adam S.; van Praag, Bernard M.S., and van de Kuilen, Gijs, "A Parametric Analysis of Prospect Theory's Functionals for the General Population," *Theory and Decision*, 68, 2010, 115-148.

Brandstätter, Eduard; Gigerenzer, Gerd, and Hertwig, Ralph, "The Priority Heuristic: Making Choices Without Trade-Offs," *Psychological Review*, 113(2), 2006, 409-432.

Brooks, Peter; Peters, Simon, and Zank, Horst, "Risk Behavior for Gain, Loss, and Mixed Prospects," *Theory and Decision*, 77, 2014, 153-182.

Brooks, Peter, and Zank, Horst, "Loss Averse Behavior," *Journal of Risk & Uncertainty*, 31(3), 2005, 301-325.

Bruhin, Adrian; Fehr-Duda, and Epper, Thomas, "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica*, 78(4), July 2010, 1375-1412.

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, 617-232.

Camerer, Colin F., and Ho, Teck-Hua, "Violations of the Betweeness Axiom and Nonlinearity in Probability," *Journal of Risk and Uncertainty*, 8, 1994, 167-196.

Cameron, A. Colin, and Trivedi, Pravin K., *Microeconometrics: Methods and Applications* (New York: Cambridge University Press, 2005).

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

Etchart-Vincent, Nathalie, "Is Probability Weighting Sensitive to the Magnitude of Consequences? An Experimental Investigation on Losses," *Journal of Risk & Uncertainty*, 28, 2004, 217-235.

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

Fehr-Duda, Helga; Gennaro, Manuelle, and Schubert, Renate, "Gender, Financial Risk, and Probability Weights," *Theory and Decision*, 60, 2006, 283-313.

Fennema, Hein, and van Assen, Marcel, "Measuring the Utility of Losses by Means of the Trade-off Method," *Journal of Risk & Uncertainty*, 17, 1998, 277-295.

Fishburn, Peter C., and Kochenberger, Gary A,. "Two-Piece von Neumann-Morgenstern Utility Functions," *Decision Sciences*, 10, 1979, 503-518.

Glöckner, Andreas, and Betsch, Tilmann, "Do People Male Decisions Under Risk Based on Ignorance? An Empirical Test of the Priority Heuristic Against Cumulative Prospect Theory," *Organizational Behavior and Human Decision Processes*, 107, 2008, 75-95.

Glöckner, Andreas, and Pachur, Thorsten, "Cognitive Models of Risky Choice: Parameter Stability and Predictive Accuracy of Prospect Theory," *Cognition*, 123(1), 2012, 21-32.

Goeree, Jacob K.; Holt, Charles A., and Palfrey, Thomas R., "Risk Averse Behavior in Generalized Matching Pennies Games," *Games and Economic Behavior*, 45, 2003, 97-113.

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, September 1979, 623-648.

Gul, Faruk, "A Theory of Disappointment Aversion," *Econometrica*, 59, 1991, 667-686.

Gul, Faruk, and Pesendorfer, Wolfgang, "The Case for Mindless Economics," in A. Caplin and A. Schotter (eds.), *Handbook of Economic Methodologies* (New York: Oxford University Press, 2007).

Harbaugh, William T.; Krause, Kate, and Vesterlund, Lise, "Risk Attitudes of Children and Adults: Choices over Small and Large Probability Gains and Losses," *Experimental Economics*, 5, 2002, 53-84.

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., "An Experimental Test for Risk Aversion," *Economics Letters*, 21(1), 1986, 7-11.

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions," *American Economic Review*, 79, September 1989, 749-762

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions: Reply," *American Economic Review*, 82, December 1992, 1426-1443.

Harrison, Glenn W., "Expected Utility Theory and The Experimentalists," *Empirical Economics*, 19(2), 1994, 223-253.

Harrison, Glenn W., "Neuroeconomics: A Critical Reconsideration," *Economics and Philosophy*, 24, 2008, 203-244.

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.; Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric "Scoring Rules for Subjective Probability Distributions," *Journal of Economic Behavior & Organization*, 2017 forthcoming.

Harrison, Glenn W.; Lau, Morten I., and Williams, Melonie B., "Estimating Individual Discount Rates for Denmark: A Field Experiment," *American Economic Review*, 92(5), December 2002, 1606-1617.

Harrison, Glenn W., and Ng, Jia Min, "Evaluating the Expected Welfare Gain from Insurance," *Journal of Risk and Insurance*, 83(1), 2016, 91-120.

Harrison, Glenn W., and Ross, Don, "Varieties of Paternalism and the Heterogeneity of Utility Structures," *CEAR Working Paper 2016-06*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2016.

Harrison, Glenn W., and Ross, Don, "The Empirical Adequacy of Cumulative Prospect Theory and its Implications for Normative Assessment," *CEAR Working Paper 2017-01*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2017.

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), 2009, 133-158.

Harrison, Glenn W., and Swarthout, J. Todd, "Cumulative Prospect Theory in the Laboratory: A Reconsideration," *CEAR Working Paper 2016-05*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2016.

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., 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, 902-904.

Kadane, Joseph B., "Healthy Skepticism as an Expected-Utility Explanation of the Phenomena of Allais and Ellsberg," *Theory and Decision*, 32(1), January 1992, 57-64.

Kahneman, Daniel, and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 1979, 263-291.

Keane, Michael P., "Structural *vs.* Atheoretic Approaches to Econometrics," *Journal of Econometrics*, 156, 2010, 3-20.

Köbberling, Veronika, and Wakker, Peter P., "An Index of Loss Aversion," *Journal of Economic Theory*, 122, 2005, 119-131.

Kőszegi, Botond, and Rabin, Matthew, "Reference-Dependent Risk Attitudes," *American Economic Review*, 97(4), September 2007, 1047-1073.

Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York: Wiley, 1978).

Leamer, Edward E., "Tantalus on the Road to Asymptopia," *Journal of Economic Perspectives*, 24(2), Spring 2011, 31-46.

Loomes, Graham, "Modelling the Stochastic Component of Behavior in Experiments: Some Issues for the Interpretation of Data," *Experimental Economics*, 8, 2005, 301-323.

Loomes, Graham; Moffatt, Peter G., and Sugden, Robert, "A Microeconometric Test of Alternative Stochastic Theories of Risky Choice," *Journal of Risk and Uncertainty*, 24(2), 2002, 103-130.

Loomes, Graham, and Sugden, Robert, "Incorporating a Stochastic Element Into Decision Theories," *European Economic Review*, 39, 1995, 641-648.

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

Lopes, Lola L., "Risk and Distributional Inequality," *Journal of Experimental Psychology: Human Perception and Performance*, 10(4), August 1984, 465-484.

Manski, Charles F., "The Maximum Score Estimator of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 1975, 205-228.

Mason, Charles F.; Shogren, Jason F.; Settle, Chad, and List, John A., "Investigating Risky Choices Over Losses Using Experimental Data," *Journal of Risk and Uncertainty*, 31(2), 187-215, 2005.

Matheson, James E., and Winkler, Robert L, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22(10), June 1976, 1087-1096.

Murphy, Ryan O., and ten Brincke, Robert H.W., "Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates," *Management Science*, 2017 forthcoming.

Nilsson, Håkan; Rieskamp, Jörg, and Wagenmakers, Eric-Jan, "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory," *Journal of Mathematical Psychology*, 55, 2011, 84-93.

Novemsky, Nathan, and Kahneman, Daniel, "The Boundaries of Loss Aversion," *Journal of Marketing Research*, XLII, May 2005, 119-128.

Pachur, Thorsten; Hanoch, Yaniv, and Gummerum, Michaela, "Prospects Behind Bars: Analyzing Decisions Under Risk in a Prison Population," *Psychonomic Bulletin and Review*, 17, 2010, 630-636.

Pennings, Joost M.E, and Smidts, Ale, "The Shape of Utility Functions and Organizational Behavior," *Management Science*, 24, 2003, 1251-1263.

Plott, Charles R., and Smith, Vernon L., "An Experimental Examination of Two Exchange Institution," *Review of Economic Studies*, 45(1), February 1978, 133-153.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66, 1998, 497-527.

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

Rieger, Marc Oliver and Wang, Mei, "What is Behind the Priority Heuristic? A Mathematical Analysis," *Psychological Review*, 115(1), January 2008, 274-280.

Rieskamp, Jörg, "The Probabilistic Nature of Preferential Choice," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 34(6), 2008, 1446-1465.

Samuelson, Paul A., *Foundations of Economic Analysis* (Boston: Harvard University Press, 1947).

Schmidt, Ulrich; Starmer, Chris, and Sugden, Robert, "Third-Generation Prospect Theory," *Journal of Risk and Uncertainty,* 36(3), June 2008, 203-223.

Schmidt, Ulrich, and Traub, Stefan, "An Experimental Test of Loss Aversion," *Journal of Risk & Uncertainty*, 25, 2002, 233-249.

Schmidt, Ulrich, and Zank, Horst, "Risk Aversion in Cumulative Prospect Theory," *Management Science*, 54, 2008, 208–216.

Schneeweiss, Hans, "The Ellsberg Paradox from the Point of View of Game Theory," *Inference and Decision*, 1, 1973, 65-78

Schunk, Daniel, and Betsch, Cornelia, "Explaining Heterogeneity in Utility Functions by Individual Differences in Decision Modes," *Journal of Economic Psychology*, 27, 2006, 386-401.

Smith, Vernon L., "Measuring Nonmonetary Utilities in Uncertain Choices: the Ellsberg Urn," *Quarterly Journal of Economics*, 83(2), May 1969, 324-329.

Starmer, Chris, "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk," *Journal of Economic Literature*, 38, June 2000, 332-382.

Stott, Henry P., "Cumulative Prospect Theory's Functional Menagerie," *Journal of Risk and Uncertainty*, 32, 2006, 101-130.

Sugden, Robert, "Reference-Dependent Subjective Expected Utility," *Journal of Economic Theory*, 111, 2003, 172-191.

Tversky, Amos, and Kahneman, Daniel, "Advances in Prospect Theory: Cumulative Representations of Uncertainty," *Journal of Risk & Uncertainty*, 5, 1992, 297-323.

von Gaudecker, Hans-Martin; van Soest, Arthur, and Wengström, Erik, "Heterogeneity in Risky Choice Behavior in a Broad Population," *American Economic Review*, 101, April 2011, 664-694.

Wakker, Peter P., *Prospect Theory for Risk and Ambiguity* (New York: Cambridge University Press, 2010).

Wilcox, Nathaniel T., "Predicting Individual Risky Choices Out-of-Context: A Critical Stochastic Modeling Primer and Monte Carlo Study," 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., "'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice Under Risk," *Journal of Econometrics*, 162(1), May 2011, 89-104.

Zeisberger, Stefan; Vrecko, Dennis, and Langer, Thomas, "Measuring the Time Stability of Prospect Theory Preferences," *Theory and Decision*, 72, 2012, 359-386.