4

The Methodologies of Behavioral Econometrics

Glenn W. Harrison

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 preexisting 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 4.1 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 looks 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 misapplication of econometric methods, akin to the story that Leamer (1978: 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 4.1, and the illustrations of two methodological innovations in Sections 4.2 and 4.3. They are directly illustrated with some real case studies in Sections 4.4, 4.5, and 4.6, with emphasis on the measurement and analysis of risk preferences. Section 4.4 considers the empirical evidence for Cumulative Prospect Theory (CPT) and asks if anyone is even reading the evidence with any methodological care. Section 4.5 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

4.6 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 4.7 draws some general conclusions, and a call to arms for methodologists.

4.1 Best-Practice Econometric Methods

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

4.1.1 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 in 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 nonstructural 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 arise 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 it 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 nonlinear 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 the mean of the remaining covariates.

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 nonzero 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 Ordinary Least Squares (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 nonzero 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 nonzero characteristics of the data, and then a constrained OLS for the nonzero level of the data conditional on it being nonzero, 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 nonzero 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 nonobservable 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]

4.1.2 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 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 toward 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 (EU) of lotteries of outcomes, and the difference in EU of alternative lotteries in a binary choice setting. In turn, these latent constructs can be characterized with parametric, semi-parametric, or nonparametric functional forms. Within the parametric family, there can be flexible functional forms 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 (CR) 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 an 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 CR violations in that case 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 et al. (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 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 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 (CS) of observed choices, given the 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 CS 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 CS, positive or negative, depending on the choice, than others. By themselves, choice patterns can only tell us the sign of the CS, not the magnitude. Section 4.2 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 et al. (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 probabilities, as shown by Andersen, Fountain et al. (2014), and when estimating the intertemporal risk preferences, as shown by Andersen et al. (2018). Section 4.3 reviews applications of joint estimation, and the methodological issues that arise.

4.2 Behavioral Econometrics and Behavioral Welfare Economics

Consider the evaluation of CS 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 CS from doing so. But how large a surplus? The agent will also purchase the product at premia with positive loadings, but CS 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 CS 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 CS 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.

4.2.1 Risk Preferences

There are now many published statements of the structural models of risk preferences underlying EUT and RDU models, starting with Harrison and Rutström (2008, §2). Appendix A (online) reviews the formal econometric specification. The latest generation of these models is now commonly estimated at the level of the individual, as demonstrated by Harrison and Ng (2016) and Harrison and Ross (2018). Assume that a subject has been classified as an EUT or RDU decision-maker, using these methods, and that we have estimates (point estimates and covariance matrices) of their risk preferences condition on the type of risk preferences.

4.2.2 Welfare Evaluation

If the subject is assumed to be an EUT type, the 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 percent 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. It is easy to demonstrate, as in Harrison and Ng (2016), that it is critical to not only identify the right type of risk preferences (EUT or RDU) for each individual but also to estimate specific parameters of those risk preferences, if one is to correctly identify the sign *and* size of welfare gain or loss from insurance choices.

4.2.3 The Welfare Metric

To evaluate RDU preferences one can estimate an RDU model for each individual. For the purpose of classifying subjects as EUT or RDU it does not matter which probability weighting functions characterize behavior: the only issue here is at what statistical confidence level we can reject the EUT hypothesis that there is no probability weighting. This hypothesis takes the form of testing $\omega(p) = p$, where $\omega(p)$ is some probability weighting function defined over objective probabilities 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 will be virtually no subjects classified as EUT since RDU nests EUT. But if we use metrics of a 10 percent, 5 percent, or 1 percent significance level on the test of the EUT hypothesis that $\omega(p) = p$, then Harrison and Ng (2016) classify 39 percent, 49 percent, or 68 percent, respectively, of 102 subjects with valid estimates as being EUT-consistent.

4.2.4 Welfare Evaluation

Expected welfare gain is foregone if the subject chooses to purchase insurance when that purchase decision has a negative CS, and similarly when the subject chooses not to purchase insurance when the purchase decision has a positive CS. For example, if we compare the expected welfare gain from each decision to the actual decisions made by subject 8 of Harrison and Ng (2016), 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 twenty-four 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 percent. 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, we find that 49 percent of decisions made resulted in negative predicted CS. 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 is 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 percent. The modal efficiency is slightly less than 50 percent, 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. Estimated distributions of CS changes, or efficiency, stand as explicit, rigorous "target practice" for anyone proposing nudges or clubs to improve welfare from insurance decisions.

4.2.5 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 (constant relative risk aversion) 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 (2017, 2018) and Monroe (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 provides "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.

4.3 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 three applications here, and one open area for future research, limiting attention to nonstrategic settings.[12]

4.3.1 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 et al. (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 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. Appendix B (online) presents the formal theoretical specification.

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 nonlinearity 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.[13]

Andersen et al. (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 percent for the same target population. Allowing for concave utility, they obtain a point estimate of the discount rate of 10 percent, which is significantly lower than the estimate of 25 percent 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 makes a significant difference to the elicited discount rates.*

4.3.2 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 has to either 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 quadratic scoring rules applied to elicit subjective probabilities of binary events, theory tells us that EUT subjects that are risk averse will report a probability closer to 0.5 than their true, latent probability. This is due to an aversion to variability of payoffs under the two states of nature: in the extreme, reporting 0.5 ensures the same payoffs under each state of nature. If we know how risk averse the individual is, we can infer what subjective probability rationally led them to make any observed report. Andersen, Fountain et al. (2014) show how to operationalize this logic econometrically and jointly estimate risk preferences and subjective probabilities if the subject is EUT. As expected, each subjective probability estimate comes with a standard error, and imprecision in estimating risk attitudes propagates, as it should as a matter of theory, to imprecise inferences about subjective probabilities.

The same logic extends to RDU models of risk preferences, although here one must account for the "first-order" effect of probability weighting, by effectively taking the inverse of the probability weighting function. This adds some complexity, particularly for reports close to 0.5, but it is also econometrically tractable, as demonstrated by Andersen, Fountain et al. (2014).

The same ideas extend to application of proper scoring rules to elicit beliefs over nonbinary events, or discrete representations of continuous events. In this case risk-averse EUT subjects will "flatten" their optimal reports over events they assign any subjective probability to: again, just reducing the variability of payoffs across events that have nonzero chance of occurring (see Harrison et al. 2017). RDU subjects will again have a more dramatic distortion of their reports than EUT subjects, although one can also recover their true, latent subject belief distributions (see Harrison and Ulm 2015).

4.3.3 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.

The concept of intertemporal risk aversion arises from theoretical deviations from an 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 $t + \tau$; 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.

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 thus requires that one controls for atemporal risk aversion and the individual discount rate. All three of these parameters are intrinsically, conceptually connected as a matter of theory, unless one makes strong assumptions otherwise. The experimental design and econometric logic of Andersen et al. (2018) follow from this theoretical point. The experimental procedures needed are a direct extension of those employed by Andersen et al. (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, an SS amount and an LL amount. In some cases, the sooner amount was paid out today, and in some cases it will be 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 percent 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 percent 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 percent chance of receiving L_t *and* $L_{t+\tau}$ and a 90 percent 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, as explained by Andersen et al. (2018) and reviewed in Appendix C (online).

The nature of the implied 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 depend on atemporal risk preferences; the discounting tasks over non-stochastic outcomes paid out today or sometime in the future depend on atemporal risk preferences (via U″) and time preferences; and the discounting tasks over stochastic outcomes paid out today or sometime in the future depend on atemporal risk preferences, time preferences, and intertemporal risk aversion. Putting behavioral error terms aside, if we were to try to estimate atemporal risk preferences and time preferences using either the lottery choices over stochastic lotteries paid out today or the discounting tasks over non-stochastic outcomes, we would be unable to identify both parameters. Similarly, if we were to try to estimate atemporal risk preferences, time preferences and intertemporal risk preferences 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 the utility function is, because of poor estimates of risk preferences, then one simply cannot make precise inferences about time preferences or intertemporal risk preferences. Similarly, poor estimates of time preferences, even if U″ is estimated relatively precisely, imply that one cannot make precise inferences about intertemporal risk preferences.

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 et al. (2014b).[14]

4.3.4 Social Preferences

It is a commonplace that individuals care about others. The concept of social preferences is a reflection of the attitudes that one individual has for the well-being of others, and the extent to which that trades off with the well-being of the individual.

Just as preferences over different commodities are a latent theoretical construct to explain observed choice behavior by an individual over those commodities, social preferences are a latent theoretical construct to explain observed choice behavior over allocations by an individual to others and the individual. But if we find it useful to think of the utility that commodities bring, it follows that social preferences defined over allocations of commodities might also usefully be defined in terms of the utility of those allocations. That is, someone might choose to allocate commodities to another person because they behave as if they care about the *actual utility* of the other person, and not because they care about the commodities received by the other person per se. But then I cannot make inferences about the social preferences of one individual without jointly making inferences about the utility function of that individual and the utility function of the other person.

Another implication of adopting this approach is that the social preference of an individual might take into account their *subjective perception* of the utility that allocations to others brings to the other person. Even if the individual knows what allocation is being made to the other person, they may not know the well-being that this allocation brings. To take an example, imagine that the allocation to the other player is a lottery: my perception of the income-equivalent of that allocation depends on what I believe to be the risk attitude of the other person. In this case, to make conceptually valid inferences about social preferences requires that one jointly estimate subjective beliefs about the risk preferences of others as well as my social preference toward that perceived EU for the other person.

Yet another reason for adopting this approach is that the social preference of an individual might utilize a *normative* utility function for allocations to others. I may know that my child is a risk-lover, but treat her as if she is risk-neutral or risk-averse when deciding on my allocations to her. Again, the challenge for joint estimation is to make inferences about my normative judgments of utility functions for others at the same time as making inferences about my social preferences.

In effect, we are proposing that one characterize social preferences the same way that we characterize social welfare functions, where the arguments are almost always the utilities of the affected individuals. In some sense the main insight from this change in characterization is the possibility of developing a structural model of different social preferences that accord with the way we characterize social preferences over income distribution for society as a whole. After all, the social preferences of an individual for one other individual, or a member of their household, is just a "little social welfare function" defined over those individuals. If we are attracted to assuming "welfarism" when characterizing social welfare functions, the assumption that social welfare is defined over individual welfare values, then the same should follow for social preferences. The three ways of thinking about the utility of the other person,[15] then, would be viewed as distinct social preference functionals, but would instead simply be viewed as different *arguments* of a single social welfare function.

The methodological point is that we cannot begin to discuss social preferences in any general form without worrying about the identification and estimation issues of jointly estimating those social preferences *and* the arguments of any social preference function.

4.3.5 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.[16]

4.4 Just Read the Literature: A Case Study of CPT

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: 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 $m < 0$, 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).

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 values for α, β, and λ.[17] One can usefully refer to this source of loss aversion as *probabilistic loss aversion*, following Schmidt and Zank (2008: 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, $\gamma+$, and $\gamma-$, although extensions are easy to consider (e.g., to the Prelec (1998) probability weighting function, which significantly generalizes the Inverse-S function).

It is remarkable to see how light the existing evidence for CPT is when one weighs 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 literally evidence for RDU. Harrison and Swarthout (2016) provide a detailed review of the literature, focusing only on controlled experiments, which has been the original basis of empirical claims for CPT. Here we focus on several of the more prominent studies.

Tversky and Kahneman (1992) gave their twenty-five subjects a total of sixty-four 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 seven certain amounts were presented in a logarithmic scale, with values spanning the extreme payouts of the risky lottery. The subject made seven 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 seven CEs for the same risky lottery, zeroing in on the interval selected in the first stage.[18] Furthermore, "switching" was ruled out, with the computer program enforcing a single switch between the risky lottery and the certain values.[19] All risky prospects used two prizes, and there were fifty-six 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 set of eight 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 fifty-six 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 eight mixed-frame choices was ½.

Tversky and Kahneman (1992) estimate a structural model of CPT using nonlinear least squares, and at the level of the individual. Remarkably, they then report the *median* point estimate, for each structural parameter, over the twenty-five estimated values. So, over all twenty-five 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.[20]

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,[21] 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.[22] 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 study of Tversky and Kahneman (1992) 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 estimates as statistically representative of anything must not care about rigor in empirical work.

Camerer and Ho (1994) was 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 (188). Hence it should be viewed as the first structural estimation of the RDU model, but not of a CPT model.

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 … " (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.

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 (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," 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 thirty subjects, each making 180 binary choices. They

find no evidence for utility loss aversion. 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.

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. They note clearly (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 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-percent 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.

Murphy and ten Brincke (2018) 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, 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. Estimated distributions, "given by medians of estimates" (fn. 9) for the pooled sample, show that there appears to be no statistically significant loss aversion, with $\lambda \approx 1$, and virtually no probability weighting on average, with $\eta \approx \varphi \approx 1$.

4.5 There Is a Reason We Compute Likelihoods: A Case Study of the PH

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.[23] 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 an 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 nonnegative outcomes, the heuristic is,

1. 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.
2. Otherwise, if one lottery has a probability of the minimum gain that is at least $\acute{\omega}$ percent better than the other, pick it.
3. Otherwise, pick the lottery with the maximum gain.

The parameters ω and $\acute{\omega}$ are each set to 10, based on arguments (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 of prizes and their probabilities.[24] This is a bold departure from the traditions embodied in EUT, RDU, CPT, and even the SP/A (security-potential/aspiration) 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, the Reflection Effect, the Certainty Effect, the Fourfold Pattern, the 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 (expected value) maximization.

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

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 et al. (1996), and Fan (2002).[25] Subjects appear to crank out the EV when given real tasks to perform, and the vast majority behave consistently with EUT as a result.[26] This is not to claim that all anomalies or stylized facts are untrue, but there is a 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 Manhattan 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.[27] 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: 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: 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 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 US bond and stock investors. Such exercises are fine as "finger mathematics" exemplars, but they are no substitute for estimation on the comparable samples. 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 an 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. The minimum gain is \$1.60 for A and \$0.10 for B, and 10 percent 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 ten 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 et al. (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: 903, Table 1) 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 four choices of lottery A, and a modest extension of EUT to allow small levels of risk aversion would explain five or six safe choices quite well. In fact, using the usual CRRA utility function, any RRA between 0.15 and 0.41 would predict five choices, and any RRA between 0.41 and 0.68 would predict six choices (Holt and Laury 2002: 1649, Table 3).

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.[28] In the ten 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 five times and then switches to B for the other five choices, hence predicting almost perfectly in each of the ten choices. But the PH gets almost four out of ten wrong *every time*, and yet is viewed as a 100 percent successful theory by this metric.

This example shows exactly why it is a mistake to casually use the "hit rate" as a metric of evaluation in such settings.[29] 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.

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 Section 2.

4.6 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 the tightest test of this proposition. Unfortunately, their conclusions are an artefact of estimation procedures that do not worry about sampling errors.[30] 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 centerpiece of their analysis. The known urn, K, has some objective distribution of balls with five 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.[31] 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-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 $w(\pi)\ u_U(X) + [1\text{-}w(\pi)]\ u_U(x)$. 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 thirty-two bets based on the K urn, and thirty-two bets based on the U urn, generating sixty-four observations per subject. They propose a power utility function defined over prizes z normalized to lie between 0 and 1, $u(z) = z^{\rho}$, 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 $\rho = \rho_K + \rho_U\ S$ and then there is an obvious hypothesis test that $\rho_U = 0$ 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: $w(p) = \exp\{-\eta(-\ln p^{\varphi}\}$, 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 sixty-six subjects was presented with thirty-two 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 fifty rows in a multiple price list elicitation. The end-result for each subjective lottery is a certain amount of money evaluated as being just less valuable than the lottery, and a certain amount of money evaluated as being just more valuable than the lottery. The switch point is enforced for the subject and involves an increment of 0.5€. Thus we have sixty-four binary lottery comparisons for each subject over thirty-two tasks. Each subject was told that one of the thirty-two tasks would be selected for payment, thereby incentivizing them to respond truthfully. Appendix D (online) reviews these estimates, which show no support for the hypothesis of source dependence.

Although 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.[32] But we should not mistake our intuition for the evidence, as comforting as that might be.

4.7 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.[33]

The subtle methodological issues involve the selection of metrics for normative evaluation, now that behavioral economics has given us a rich array of alternative *descriptive* models to the traditional models.[34] 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.

Notes

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.

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 are 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.

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), whether the individual currently smokes daily or occasionally (binary), whether the individual is a former smoker (binary), the number of cigarettes smoked per day (discrete, multivalued), and the number of years that current daily smokers have smoked (discrete, multivalued).

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 percent 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.

6 EUT does not, in these circumstances, 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 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.

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, Bayesians 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 et al. (2014a).

9 A classic example is the binary choice procedure, which is self-evidently incentive-compatible, compared to the Becker, DeGroot, and Marschak (BDM) (1964) 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: 537). Moreover, even if subjects understand the incentives, the mechanism is known to generate *extremely* weak incentives for accurate reports: see Harrison (1992; 1994).

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."

12 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.

13 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; 78–9). A similar debate rages with respect to structural assumptions about statistical 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.

14 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 nonadditive, then risk preferences over time streams of money need to be sharply distinguished from risk preferences over atemporal 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 out the latter hypothesis, but there is a simpler explanation well within received theory. Evidence for intertemporal risk aversion in experiments is provided

by Andersen et al. (2018), 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 atemporal risk aversion encourages mean-preserving reductions in the variability of atemporal 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 nonadditive intertemporal utility function.

15 The actual utility of the other subject, the subjective perception I have about the utility of the other subject, or the normative utility I choose to apply to the other subject.

16 For example, I have seen so little evidence for CPT that I no longer automatically build in (longer) risk batteries with mixed frames or loss frames. Others might demur.

17 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.

18 This variant is called an *iterative* multiple price list by Andersen et al. (2006).

19 This variant is called a *sequential* multiple price list by Andersen et al. (2006).

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

21 This issue is the focus of the use of "hierarchical" methods by Nilsson, Rieskamp, and Wagenmakers (2011) and Murphy and ten Brincke (2018), which are in principle well suited to handling this particular problem, which is not unique to CPT.

22 Tversky and Kahneman (1992: 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 fifty-six choices, and the median values of $x for the final eight 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. Glöckner and Pachur (2012) used the same unfortunate style of reporting results.

23 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.

24 Of course, there are many such heuristics from psychology and the judgment and decision-making literature, noted explicitly by Brandstätter et al. (2006: 417, Table 3).

25 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 (105).

26 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 et al. (2007) pay attention to this issue in different ways in their designs.

27 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 dataset after dataset.
28 To see this, follow carefully the explanation in Brandstätter et al. (2006: 418) of how the vertical axis on their figure 1 is created. There are fourteen choice tasks being evaluated here. The PH predicted the *majority* choice in each of the fourteen tasks, so it is given a predictive score of 100 percent. The "equiprobable" heuristic predicted ten out of fourteen 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.
29 There are some noncasual, 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: 483ff., §14.7.2).
30 These estimation procedures are defended by Wakker (2010, Appendix A), so this is not just an inadvertent slip.
31 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.
32 For example, by the application of Bayes Rule or the reduction of compound lotteries.
33 The point here is the role of methodologists in addressing these issues. It is descriptively easy to see the effects of negative externalities generated by the popularity of the "mostly harmless" school of econometrics (Angrist and Pischke 2009).
34 See Harrison and Ross (2017, 2018) for a statement of the philosophical issues raised.

References

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

Andersen, Steffen, John Fountain, Glenn W. Harrison, and E. Elisabet Rutström. 2014. "Estimating Subjective Probabilities." *Journal of Risk & Uncertainty* 48: 207–29.

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

Andersen, Steffen, Glenn W. Harrison, Morten Igel Lau, and E. Elisabet Rutström. 2008. "Eliciting Risk and Time Preferences." *Econometrica* 76 (3): 583–618.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2014a. "Dual Criteria Decisions." *Journal of Economic Psychology* 41 (April): 101–13.

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

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2018. "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion." *International Economic Review* 59 (2): 537–55.

Andreoni, James, and Charles Sprenger. 2012. "Risk Preferences Are Not Time Preferences." *American Economic Review* 102 (7): 3357–76.
Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.
Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9 (July): 226–32.
Benartzi, Shlomo, and Richard H. Thaler. 1995. "Myopic Loss Aversion and the Equity Premium Puzzle." *Quarterly Journal of Economics* 111 (1): 75–92.
Birnbaum, Michael H. 2004. "Causes of Allais Common Consequence Paradoxes: An Experimental Dissection." *Journal of Mathematical Psychology* 48: 87–106.
Birnbaum, Michael H. 2008. "Evaluation of the Priority Heuristic as a Descriptive Model of Risky Decision Making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006)." *Psychological Review* 115 (1): 253–60.
Birnbaum, Michael H., and Juan B. Navarrete. 1998. "Testing Descriptive Utility Theories: Violations of Stochastic Dominance and Cumulative Independence." *Journal of Risk and Uncertainty* 17: 17–49.
Brandstätter, Eduard, GerdGigerenzer, and Ralph Hertwig. 2006. "The Priority Heuristic: Making Choices without Trade-Offs." *Psychological Review* 113 (2): 409–32.
Bruhin, Adrian, Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica* 78 (4): 1375–412.
Burke, Michael S., John R. Carter, Robert D. Gominiak, and Daniel F. Ohl. 1996. "An Experimental Note on the Allais Paradox and Monetary Incentives." *Empirical Economics* 21: 617–32.
Camerer, Colin F., and Teck-Hua Ho. 1994. "Violations of the betweenness Axiom and Nonlinearity in Probability." *Journal of Risk and Uncertainty* 8: 167–96.
Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.
Conlisk, John. 1989. "Three Variants on the Allais Example." *American Economic Review* 79 (3): 392–407.
Fan, Chinn-Ping. 2002. "Allais Paradox in the Small." *Journal of Economic Behavior & Organization* 49: 411–21.
Glöckner, Andreas, and Thorsten Pachur. 2012. "Cognitive Models of Risky Choice: Parameter Stability and Predictive Accuracy of Prospect Theory." *Cognition* 123 (1): 21–32.
Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2003. "Risk Averse Behavior in Generalized Matching Pennies Games." *Games and Economic Behavior* 45: 97–113.
Grether, David M., and Charles R. Plott. 1979. "Economic Theory of Choice and the Preference Reversal Phenomenon." *American Economic Review* 69 (September): 623–48.
Gul, Faruk, and Wolfgang Pesendorfer. 2007. "The Case for Mindless Economics." In *Handbook of Economic Methodologies*, ed. A. Caplin and A. Schotter. New York: Oxford University Press.
Harless, David W., and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62 (6): 1251–89.
Harrison, Glenn W. 1992. "Theory and Misbehavior of First-Price Auctions: Reply." *American Economic Review* 82 (December): 1426–43.

Harrison, Glenn W. 1994. "Expected Utility Theory and The Experimentalists." *Empirical Economics* 19 (2): 223–53.
Harrison, Glenn W. 2008. "Neuroeconomics: A Critical Reconsideration." *Economics and Philosophy* 24: 203–44.
Harrison, Glenn W., Eric Johnson, Melayne M. McInnes, and E. Elisabet Rutström. 2005. "Risk Aversion and Incentive Effects: Comment." *American Economic Review* 95 (3): 897–901.
Harrison, Glenn W., Eric Johnson, Melayne M. McInnes, and E. Elisabet Rutström. 2007. "Measurement with Experimental Controls." In *Measurement in Economics: A Handbook*, ed. M. Boumans. San Diego, CA: Elsevier.
Harrison, Glenn W., Morten I. Lau, and Melonie B. Williams. 2002. "Estimating Individual Discount Rates for Denmark: A Field Experiment." *American Economic Review* 92 (5): 1606–17.
Harrison, Glenn W., Jimmy Martínez-Correa, J. Todd Swarthout, and Eric Ulm. 2017. "Scoring Rules for Subjective Probability Distributions." *Journal of Economic Behavior & Organization* 134: 430–48.
Harrison, Glenn W., and Jia Min Ng. 2016. "Evaluating the Expected Welfare Gain from Insurance." *Journal of Risk and Insurance* 83 (1): 91–120.
Harrison, Glenn W., and Don Ross. 2017. "The Empirical Adequacy of Cumulative Prospect Theory and Its Implications for Normative Assessment." *Journal of Economic Methodology* 24 (2): 150–65.
Harrison, Glenn W., and Don Ross. 2018. "Varieties of Paternalism and the Heterogeneity of Utility Structures." *Journal of Economic Methodology* 25 (1): 42–67.
Harrison, Glenn W., and E. Elisabet Rutström. 2008. "Risk Aversion in the Laboratory." In *Risk Aversion in Experiments*, Vol. 12, ed. J. C. Cox and G. W. Harrison. Bingley: Emerald, Research in Experimental Economics.
Harrison, Glenn W., and E. Elisabet Rutström. 2009. "Expected Utility and Prospect Theory: One Wedding and a Decent Funeral." *Experimental Economics* 12 (2): 133–58.
Harrison, Glenn W., and J. Todd Swarthout. 2016. "Cumulative Prospect Theory in the Laboratory: A Reconsideration." *CEAR Working Paper 2016-05*. Atlanta, GA: Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.
Harrison, Glenn W., and Eric R. Ulm. 2015. "Recovering Subjective Probability Distributions." *CEAR Working Paper 2015-01*. Atlanta, GA: Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.
Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55.
Holt, Charles A., and Susan K. Laury. 2005. "Risk Aversion and Incentive Effects: New Data without Order Effects." *American Economic Review* 95 (3): 902–4.
Kadane, Joseph B. 1992. "Healthy Skepticism as an Expected-Utility Explanation of the Phenomena of Allais and Ellsberg." *Theory and Decision* 32 (1): 57–64.
Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47: 263–91.
Keane, Michael P. 2010. "Structural *vs.* Atheoretic Approaches to Econometrics." *Journal of Econometrics* 156: 3–20.
Köbberling, Veronika, and Peter P. Wakker. 2005. "An Index of Loss Aversion." *Journal of Economic Theory* 122: 119–31.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
Leamer, Edward E. 2011. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24 (2): 31–46.
Loomes, Graham, and Robert Sugden. 1998. "Testing Different Stochastic Specifications of Risky Choice." *Economica* 65: 581–98.
Lopes, Lola L. 1984. "Risk and Distributional Inequality." *Journal of Experimental Psychology: Human Perception and Performance* 10 (4): 465–84.
Manski, Charles F. 1975. "The Maximum Score Estimator of the Stochastic Utility Model of Choice." *Journal of Econometrics* 3: 205–28.
Monroe, Brian A. August 2017. *Stochastic Models in Experimental Economics*. PhD Thesis. South Africa: School of Economics, University of Cape Town.
Murphy, Ryan O., and Robert H. W. ten Brincke. 2018. "Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates." *Management Science* 64 (1): 308–26.
Nilsson, Håkan, Jörg Rieskamp, and Eric-Jan Wagenmakers. 2011. "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory." *Journal of Mathematical Psychology* 55: 84–93.
Plott, Charles R., and Vernon L. Smith. 1978. "An Experimental Examination of Two Exchange Institution." *Review of Economic Studies* 45 (1): 133–53.
Plott, Charles R., and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review* 95: 530–45.
Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica* 66: 497–527.
Rieskamp, Jörg, 2008. "The Probabilistic Nature of Preferential Choice." *Journal of Experimental Psychology: Learning, Memory and Cognition* 34 (6): 1446–65.
Samuelson, Paul A. 1947. *Foundations of Economic Analysis*. Boston, MA: Harvard University Press.
Schmidt, Ulrich, and Horst Zank. 2008. "Risk Aversion in Cumulative Prospect Theory." *Management Science* 54: 208–16.
Schneeweiss, Hans. 1973. "The Ellsberg Paradox from the Point of View of Game Theory." *Inference and Decision* 1: 65–78.
Smith, Vernon L. 1969. "Measuring Nonmonetary Utilities in Uncertain Choices: The Ellsberg Urn." *Quarterly Journal of Economics* 83 (2): 324–9.
Starmer, Chris. 2000. "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk." *Journal of Economic Literature* 38 (June): 332–82.
Stott, Henry P. 2006. "Cumulative Prospect Theory's Functional Menagerie." *Journal of Risk and Uncertainty* 32: 101–30.
Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representations of Uncertainty." *Journal of Risk & Uncertainty* 5: 297–323.
von Gaudecker, Hans-Martin, Arthur van Soest, and Erik Wengström. 2011. "Heterogeneity in Risky Choice Behavior in a Broad Population." *American Economic Review* 101 (April): 664–94.
Wakker, Peter P. 2010. *Prospect Theory for Risk and Ambiguity*. New York: Cambridge University Press.

Wilcox, Nathaniel T. 2008. "Predicting Individual Risky Choices Out-of-Context: A Critical Stochastic Modeling Primer and Monte Carlo Study." In *Risk Aversion in Experiments*, Vol. 12, ed. J. Cox and G. W. Harrison. Bingley: Emerald, Research in Experimental Economics.

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