Exp Econ (2009) 12: 133–158
DOI 10.1007/s10683-008-9203-7

Expected utility theory and prospect theory: one wedding and a decent funeral

Glenn W. Harrison · E. Elisabet Rutström

Received: 20 July 2007 / Revised: 20 February 2008 / Accepted: 10 April 2008 /
Published online: 13 June 2008

Abstract Choice behavior is typically evaluated by assuming that the data is generated by one latent decision-making process or another. What if there are two (or more) latent decision-making processes generating the observed choices? Some choices might then be better characterized as being generated by one process, and other choices by the other process. A finite mixture model can be used to estimate the parameters of each decision process while simultaneously estimating the probability that each process applies to the sample. We consider the canonical case of lottery choices in a laboratory experiment and assume that the data is generated by expected utility theory and prospect theory decision rules. We jointly estimate the parameters of each theory as well as the fraction of choices characterized by each. The methodology provides the wedding invitation, and the data consummates the ceremony followed by a decent funeral for the representative agent model that assumes only one type of decision process. The evidence suggests support for each theory, and goes further to identify under what demographic domains one can expect to see one theory perform better than the other. We therefore propose a reconciliation of the debate over two of the dominant theories of choice under risk, at least for the tasks and samples we consider. The methodology is broadly applicable to a range of debates over competing theories generated by experimental and non-experimental data.

We thank the U.S. National Science Foundation for research support under grants NSF/IIS 9817518, NSF/HSD 0527675 and NSF/SES 0616746; Ryan Brossette, Harut Hovsepyan, David Millsom and Bob Potter for research assistance; and Steffen Andersen, Vince Crawford, Curt Eaton, John Hey, Peter Kennedy, Jan Kmenta, Peter Wakker, two referees, and numerous seminar participants for helpful comments. Supporting data and instructions are stored at the ExLab Digital Library at http://exlab.bus.ucf.edu.

G.W. Harrison (✉) · E.E. Rutström
Department of Economics, College of Business Administration, University of Central Florida, Orlando, FL, USA
e-mail: gharrison@research.bus.ucf.edu

E.E. Rutström
e-mail: erutstrom@bus.ucf.edu

Keywords Expected utility theory · Prospect theory · Mixture models

JEL Classification D81 · C91 · C51 · C12

One of the enduring contributions of behavioral economics is that we now have a rich set of competing models of behavior in many settings, with expected utility theory and prospect theory as the two front runners for choices under uncertainty. Debates over the validity of these models have often been framed as a horse race, with the winning theory being declared on the basis of some statistical test in which the theory is represented as a latent process explaining the data. In other words, we seem to pick the best theory by "majority rule." If one theory explains more of the data than another theory, we declare it the better theory and discard the other one. In effect, after the race is over we view the horse that "wins by a nose" as if it was the only horse in the race. The problem with this approach is that it does not recognize the possibility that several behavioral latent processes may coexist in a population.

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

Heterogeneity in responses is well recognized as causing statistical problems in experimental and non-experimental data. Nevertheless, allowing for heterogeneity in responses through standard methods, such as fixed or random effects, is not helpful when we want to identify which people behave according to what theory, and when. Heterogeneity can be partially recognized by collecting information on observable characteristics and controlling for them in the statistical analysis. For example, a given theory might allow some individuals to be more risk averse than others as a matter of personal preference. But this approach only recognizes *heterogeneity within a given theory*. This may be important for valid inferences about the ability of the theory to explain the data, but it does not allow for *heterogeneous theories* to co-exist in the same sample. One way to allow explicitly for alternative theories is to simply collect sufficient data at the individual level, and test the different theories for that individual, as demonstrated in Hey and Orme (1994). This is an approach that is not always feasible, however, due to the large amount of observations that are needed.[1]

The approach to heterogeneity and the possibility of co-existing theories adopted here is to propose a "wedding" of the theories. We specify and estimate a grand likelihood function that allows each theory to co-exist and have different weights, a so-called mixture model. The data can then identify what support each theory has. The

[1] In one experimental treatment Hey and Orme (1994) collected 200 binary choice responses from each of 80 subjects, and were able to test alternative estimated models for each individual. At the beginning of their discussion of results (p. 1300), they state simply, "We assume that all subjects are different. We therefore fit each of the 11 preference functionals discussed above to the subject's stated preferences for each of the 80 subjects *individually*."

wedding is consummated by the maximum likelihood estimates converging on probabilities that apportion non-trivial weights to each theory. We then offer the "decent funeral" requested by Kirman [1992; p. 119], who considers the dangers of using representative agent characterizations in macroeconomics and argues that "…it is clear that the 'representative' agent deserves a decent burial, as an approach to economics analysis that is not only primitive, but fundamentally erroneous."

A canonical setting for theories of choice under uncertainty has been the choice over pairs of monetary lotteries in laboratory experiments, reviewed by Camerer (1995). To keep things sharply focused, only two alternative models are posited here. One is a simple expected utility theory (EUT) specification, assuming a Constant Relative Risk Aversion (CRRA) utility function defined over the "final monetary prize" that the subject would receive if the lottery were played out. The other model is a popular specification of prospect theory (PT) due to Kahneman and Tversky (1979), in which the utility function is defined over gains and losses separately, and there is a probability weighting function that converts the underlying probabilities of the lottery into subjective probabilities.[2]

The evidence against EUT is extensive. Luce and Suppes (1965) and Schoemaker (1982) provide reviews of the earliest literature, and Camerer (1995) and Starmer (2000) of later developments. To the best of our knowledge, none of these tests allow for both processes to simultaneously explain the data. For example, there are many tests that propose extensions of EUT that nest EUT in an alternative model, and some might argue that these tests therefore allow either model to emerge if it is in fact generating the data. However, the tests are set up to identify a single explanation of EUT violations that can accommodate every type of violation, rather than allow for the possibility that one explanation accounts for certain violations and that other explanations account for different violations. For some violations it may be easy to write out specific parametric models of the latent EUT decision-making process that can account for the data. The problem is that the model that can easily account for one set of violations need not account for others. For example, the preference reversals of Grether and Plott (1979) can be explained by assuming risk neutral subjects with an arbitrarily small error process, since the paired lotteries are designed to have the same expected value. Hence each subject is indifferent, and the error process can account for the data.[3] But then such subjects should not violate EUT in other settings, such as common ratio tests. However, rarely does one encounter tests that confront subjects with a wide range of tasks and evaluates behavior simultaneously over that wider domain.[4]

[2]We use the language of EUT, but prospect theorists would instead refer to the utility function as a "value function," and to the transformed probabilities as "decision weights." Our implementation of PT uses the original version and assumes no editing processes. One could easily extend our approach to allow editing processes or Cumulative PT formulations.

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

[4]There are three striking counter-examples to this trend. Hey and Orme (1994) deliberately use lotteries that span a wide range of prizes and probabilities, avoiding "trip wire" pairs, and they conclude that EUT does an excellent job of explaining behavior compared to a wide range of alternatives. Similarly, Harless

Thus we view the state of the evidence as open, in the sense that nobody has explored the behavior of EUT and alternatives over a wide range of tasks *and* formally allowed the observed choices to be generated by both models simultaneously. Our primary methodological contribution is to illustrate how one can move from results that rely on assumptions that there is one single, true decision-making model to a richer characterization that allows for the *potential co-existence* of multiple decision-making models. Thus we are interested in investigating the co-existence of latent EUT *and* PT data-generating processes, and not just the heterogeneity of behavior given EUT *or* PT. Our data indicate that one should not think exclusively of EUT *or* PT as the correct model, but as models that are correct for distinct parts of the sample or the decisions. Thus the correct way to view these data is that they are best characterized by EUT *and* PT.

Section 1 discusses our approach in more detail. Section 2 reviews the experimental data, Sect. 3 the competing models, Sect. 4 presents the results of estimating a mixture model using both models, Sect. 5 compares our approach to other statistical procedures that might be used in this setting, and Sect. 6 draws conclusions.

1 Weddings, funerals and heterogeneity

We seek a statistical reconciliation of the debate over two dominant theories of choice under risk, EUT and PT. We collect experimental data in a setting in which subjects make risky decisions in a gain frame, a loss frame, or a mixed frame, explained in detail below. Such settings allow one to test the competing models on the richest domain. All data are collected using standard procedures in experimental economics in the laboratory: no deception, no field referents, fully salient choices, and with information on individual characteristics of the subjects. The experimental procedures are described in Sect. 2.

Our experiments are a replication and extension of the procedures of Hey and Orme (1994). The major extension is to consider lotteries in which some or all outcomes are framed as losses, as well as the usual case in which all outcomes are framed as gains. Each subject received an initial endowment that resulted in their final earnings opportunities being the same across all frames. A minor procedural extension is to collect individual demographic characteristics from each subject.

Our EUT specification is defined over the "final monetary prize" that the subject would receive if the lottery were played out. That is, the argument of the utility function is the prize plus the initial endowment, which are always jointly non-negative. Our PT specification defines the utility function over gains and losses separately, and there is a probability weighting function that converts the underlying probabilities of the lottery into subjective probabilities. The three critical features of the PT model are (i) that the arguments of the utility function be gains and losses relative to some

and Camerer (1994) consider a wide range of aggregate data across many studies, and find that EUT does a good job of explaining behavior if one places a value on parsimony. And Loomes and Sugden (1998) deliberately choose lotteries "…to provide good coverage of the space within each (implied Marschak-Machina probability) triangle, and also to span a range of gradients sufficiently wide to accommodate most subjects' risk attitudes." (p. 589).

reference point, taken here to be the endowment; (ii) that losses loom larger than gains in the utility function; and (iii) that there is a nonlinearity in the transformed probabilities that could account for apparently different risk attitudes for different lottery probabilities. There may be some debate over whether the endowment serves as a "reference point" to define losses, but that is one thing that is being tested with these different models.[5] We specify the exact forms of the models tested in Sect. 3.

It is apparent that there are many variants of these two models, and many others that deserve to be considered. Wonderful expositions of the major alternatives can be found in Camerer (1995) and Hey and Orme (1994), among others. But this is not the place to array every feasible alternative. Instead, the initial exploration of this approach considers two major alternatives and the manner in which they are characterized. Many of the variants involve nuances that would be hard to evaluate empirically with the data available here, rich as it is for the task at hand.[6] But the methodological point illustrated here is completely general.

Mixture models have a long pedigree in statistics, stretching back to Pearson (1894). Modern surveys of the development of mixture models are provided by Titterington et al. (1985), Everitt (1996) and McLachlan and Peel (2000). Mixture models are also virtually identical to "latent class models" used in many areas of statistics, marketing and econometrics, even though the applications often make them seem quite different (e.g., Goodman 1974a, 1974b; Vermunt and Magidson 2003). In experimental economics, El-Gamal and Grether (1995) estimate a finite mixture model of Bayesian updating behavior, and contrast it to a related approach in which individual subject behavior is classified completely as one type of the other. Stahl and Wilson (1995) develop a finite mixture model to explain behavior in a normal form game, differentiating between five types of boundedly rational players.[7]

One challenge with mixture models of this kind is the joint estimation of the probabilities and the parameters of the conditional likelihood functions. If these are conditional models that each have some chance of explaining the data, then mixture models will be characterized numerically by relatively flat likelihood functions. On the other hand, if there are indeed K distinct latent processes generating the overall sample, allowing for this richer structure is inferentially useful. Most of the applications of

[5] Some might argue that negative lottery prizes would count as losses only when the subject "earns" the initial stake. This is a fair point, but testable by simple modification of the design. There is evidence from related settings that such changes in how subjects receive their initial endowment can significantly change behaviour: see Cherry et al. (2002), Johnson et al. (2006) and George et al. (2007). Moreover, one can extend the PT specification to estimate the endogenous reference point that subjects employ: see Andersen et al. (2006a) for example.

[6] Quite apart from alternative models to EUT and PT, there are alternative parametric specifications within each of EUT or PT. For example, on the EUT side one may consider specifications that do not assume CRRA. But the reliable estimation of the parameters of such specifications likely requires a much wider range of prizes than considered here: hence the design of Holt and Laury (2002), where prizes were scaled by a factor of 20 for most of their sample in order to estimate such a flexible specification. For an example on the PT side, there are a plethora of specifications available for the probability weighting function (e.g., Wu and Gonzalez 1996 and Prelec 1998).

[7] Additional applications of mixture models to experimental data include Stahl (1996, 1998), Haruvy et al. (2001), Hurley and Shogren (2005), Andersen et al. (2006a, 2006b, 2008), Bardsley and Moffatt (2007), Bruhin et al. (2007), Conte et al. (2007), and Harrison et al. (2005).

mixture modeling in economics have been concerned with their use in better characterizing unobserved individual heterogeneity in the context of a given theory about behavior,[8] although there have been some applications to hypothetical survey data that consider individual heterogeneity over alternative behavioral theories.[9] There has been no earlier attempt to use mixture models to address the debate over theories of choice under uncertainty.

2 Experimental design

Subjects were presented with 60 lottery pairs, each represented as a "pie" showing the probability of each prize. The subject could choose the lottery on the left or the right, or explicitly express indifference (in which case the experimenter would flip a coin on the subject's behalf). After all 60 lottery pairs were evaluated, three were selected at random for payment.[10] The lotteries were presented to the subjects in color on a private computer screen,[11] and all choices recorded by the computer program. This program also recorded the time taken to make each choice. An appendix (available on request) presents the instructions given to our subjects, as well as an example of the "screen shots" they saw. In addition to the choice tasks, the subjects provided information on demographic and other personal characteristics.

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

[8] For example, Heckman and Singer (1984), Geweke and Keane (1999) and Araña and León (2005). The idea here is that the *disturbance term* of a given specification is treated as a mixture of processes. Such specifications have been used in many settings that may be familiar to economists under other names. For example, stochastic frontier models rely on (unweighted) mixture specifications of a symmetric "technical efficiency" disturbance term and an asymmetric "idiosyncratic" disturbance term; see Kumbhakar and Lovell (2000) for an extensive review.

[9] For example, Werner (1999) uses mixture models to characterize the "spike at zero" and "non-spike" responses common in contingent valuation surveys. Wang and Fischbeck (2004) provide an application to framing effects within prospect theory, using hypothetical field survey data on health insurance choices. Their approach is to view the frames as different data generation processes for responses.

[10] The typical application of the random lottery incentive mechanism in experiments such as these would have one choice selected at random. We used three to ensure comparability of rewards with other experiments in which subjects made choices over 40 or 20 lotteries, and where 2 lotteries or 1 lottery was respectively selected at random to be played out. Harrison and Rutström (2008) discuss these treatments, which have no effect on observed behavior.

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

be −$8, −$3, $3 and $8, generating final outcomes inclusive of the endowment of $0, $5, $11 and $16.[12]

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

The probabilities used in each lottery ranged roughly evenly over the unit interval. Values of 0, 0.13, 0.25, 0.37, 0.5, 0.62, 0.75 and 0.87 were used.[13] The presentation of a given lottery on the left or the right was determined at random, so that the "left" or "right" lotteries did not systematically reflect greater risk or greater prize range than the other.

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

Our data consists of responses from 158 subjects making 9311 choices that do not involve indifference. Only 1.7% of the choices involved explicit choice of indifference, and to simplify we drop those.[14] Of these 158 subjects, 63 participated in gain frame tasks, 37 participated in mixed frame tasks, and 58 participated in loss frame tasks.

3 Meet the bride and groom

3.1 Expected utility specification

We assume that utility of income is defined by $U(s, x) = (s + x)^r$ where s is the fixed endowment provided at the beginning of the experiment (excluding the show-up fee), x is the lottery prize, and r is a parameter to be estimated. With this specification

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

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

[14] For the specification of likelihoods of strictly binary responses, such observations add no information. However, one could augment the likelihood for the strict binary responses with a likelihood defined over "fractional responses" and assume that the fraction for these indifferent responses was exactly ½. Such specifications are provided by Papke and Wooldridge (1996), and used by Andersen et al. (2008), but add needless complexity for present purposes given the small number of responses involved.

we assume "perfect asset integration" between the endowment and lottery prize.[15] Probabilities for the kth prize, p_k, are those that are induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome in each lottery. Since there were up to 4 outcomes in each lottery i, $EU_i = \sum_{k=1,4}[p_k \times U_k]$.

A simple stochastic specification was used to specify likelihoods conditional on the model. The EU for each lottery pair was calculated for a candidate estimate of r, and the difference $\nabla EU = EU_R - EU_L$ calculated, where EU_L is the left lottery in the display and EU_R is the right lottery. The index ∇EU is then used to define the cumulative probability of the observed choice using the logistic function: $G(\nabla EU) = \exp(\nabla EU)/[1 + \exp(\nabla EU)]$.[16]

Thus the likelihood, conditional on the EUT model being true, depends on the estimates of r given the above specification and the observed choices. The conditional log-likelihood is

$$\ln L^{\text{EUT}}(r; y, X) = \sum_i \ln l_i^{\text{EUT}} = \sum_i [y_i \ln G(\nabla EU) + (1 - y_i) \ln(1 - G(\nabla EU))]$$

where $y_i = 1(0)$ denotes the choice of the right (left) lottery in task i, and X is a vector of individual characteristics that implicitly conditions ∇EU. Harrison and Rutström (2008) review procedures for structural estimation of models such as these using maximum likelihood.

3.2 Prospect theory specification

Tversky and Kahneman (1992) propose a popular parametric specification which is employed here. There are two components, the utility function and the probability weighting function.

A power utility function is defined separately over gains and losses: $U(x) = x^\alpha$ if $x \geq 0$, and $U(x) = -\lambda(-x)^\beta$ for $x < 0$. So α and β are the risk aversion parameters, and λ is the coefficient of loss aversion.

There are two variants of prospect theory, depending on the manner in which the probability weighting function is combined with utilities. The original version proposed by Kahneman and Tversky (1979) posits some weighting function which is separable in outcomes, and has been usefully termed Separable Prospect Theory (SPT) by Camerer and Ho [1994; p. 185]. The alternative version, proposed by Tversky and Kahneman (1992), posits a weighting function defined over the cumulative probability distributions. The form of the weighting function proposed by Tversky

[15] A valuable extension, inspired by Cox and Sadiraj (2006), would be to allow s and x to be combined in some linear manner, with weights to be estimated.

[16] The use of the logistic implies the usual random utility specification due to Marschak (1960). That is, that the expected utility of a given lottery equals the deterministic EU plus some extreme value error that is independent of the lottery (e.g., Train 2003; p. 55). This implies a well-known normalization of the error term to be $\pi^2/3$. Hey and Orme (1994; p. 1301) discuss in general terms how one could equivalently estimate this variance and impose constraints on the range of utility values under EUT. Such constraints are not applicable under standard specifications of PT with loss aversion (Köbberling and Wakker 2005; p. 121), so we avoid them.

and Kahneman (1992) has been widely used for both separable and cumulative versions of PT, and assumes weights $w(p) = p^{\gamma}/[p^{\gamma} + (1-p)^{\gamma}]^{1/\gamma}$.

Assuming that SPT is the true model, prospective utility PU is defined in much the same manner as when EUT is assumed to be the true model. The PT utility function is used instead of the EUT utility function, and $w(p)$ is used instead of p, but the steps are otherwise identical.[17] The same error process is assumed to apply when the subject forms a preference for one lottery over the other. Thus the difference in prospective utilities is defined similarly as $\nabla PU = PU_R - PU_L$.

Thus the likelihood, conditional on the SPT model being true, depends on the estimates of α, β, λ and γ given the above specification and observed choices.[18] The conditional log-likelihood is

$$\ln L^{\mathrm{PT}}(\alpha, \beta, \lambda, \gamma; y, X) = \sum_i \ln l_i^{\mathrm{PT}} = \sum_i [y_i \ln G(\nabla PU) + (1 - y_i) \ln(1 - G(\nabla PU))].$$

3.3 The nuptial

If we let π^{EUT} denote the probability that the EUT model is correct, and $\pi^{\mathrm{PT}} = (1 - \pi^{\mathrm{EUT}})$ denote the probability that the PT model is correct, the grand likelihood can be written as the probability weighted average of the conditional likelihoods. Thus the likelihood for the overall model estimated is defined by

$$\ln L(r, \alpha, \beta, \lambda, \gamma, \pi^{\mathrm{EUT}}; y, X) = \sum_i \ln[(\pi^{\mathrm{EUT}} \times l_i^{\mathrm{EUT}}) + (\pi^{\mathrm{PT}} \times l_i^{\mathrm{PT}})]. \quad (1)$$

This log-likelihood can be directly maximized[19] to find estimates of the parameters.[20]

[17] We ignore the editing processes discussed by Kahneman and Tversky (1979). They have not been used in the empirical implementations of SPT by Camerer and Ho (1994) or in the weighting functions used by Hey and Orme (1994), to take two prominent examples.

[18] The SPT specification has three more parameters than the EUT specification. There are numerous methods for accounting for differences in the number of parameters in different models, illustrated well by Hey and Orme (1994). Our view is that those corrections can be dangerous if applied mechanically, independently of the rationale for their inclusion in the original model. The "menu approach" proposed by Harless and Camerer (1994) strikes us as the correct way to view the effects of parsimony, as with all specification searches: the objective is to just document the trade-offs implied by more or less parameters in terms of some metric such as goodness of fit. The reader is then free to select from that menu in accord with their own preferences for parsimony. These issues also relate to our discussion in Sect. 5 of the relationship between mixture models and non-nested hypothesis test procedures.

[19] Estimation of mixture models requires some attention to the numerical properties of the log-likelihood, since there may easily be multiple modes. In most cases direct maximum likelihood will be well-behaved, but it is always useful to trace out the log-likelihood conditional on assumed values of the mixture probability to ensure that global maxima have been found. This is a simple matter when there are no covariates.

[20] This approach assumes that any one observation can be generated by both models, although it admits of extremes in which one or other model wholly generates the observation. One could alternatively define a grand likelihood in which observations or subjects are classified as following one model or the other on the basis of the latent probabilities π^{EUT} and π^{PT}. El-Gamal and Grether (1995) illustrate this approach in the context of identifying behavioral strategies in Bayesian updating experiments.

We allow each parameter to be a linear function of the observed individual characteristics of the subject. This is the X vector referred to above. Six characteristics are considered: binary variables to identify Females, subjects that self-reported their ethnicity as Black, those that reported being Hispanic, those that had a Business major, and those that reported a low cumulative GPA (below 3¼). We also included Age in years. The estimates of *each* parameter in the above likelihood function actually entails estimation of the coefficients of a linear function of these characteristics. For example, the estimate of r, $\hat{r}$, would actually be

$$\begin{aligned}\hat{r} = \hat{r}_0 &+ (\hat{r}_{\text{FEMALE}} \times \text{FEMALE}) + (\hat{r}_{\text{BLACK}} \times \text{BLACK}) \\ &+ (\hat{r}_{\text{HISPANIC}} \times \text{HISPANIC}) + (\hat{r}_{\text{BUSINESS}} \times \text{BUSINESS}) \\ &+ (\hat{r}_{\text{GPAlow}} \times \text{GPAlow}) + (\hat{r}_{\text{AGE}} \times \text{AGE})\end{aligned}$$

where $\hat{r}_0$ is the estimate of the constant. If we collapse this specification by dropping all individual characteristics, we would simply be estimating the constant terms for each of $r, \alpha, \beta, \lambda, \mu$. Obviously the X vector could include treatment effects as well as individual effects, or interaction effects.[21]

The estimates allow for the possibility of correlation between responses by the same subject, so the standard errors on estimates are corrected for the possibility that the 60 responses are clustered for the same subject. The use of clustering to allow for "panel effects" from unobserved individual effects is common in the statistical survey literature.[22]

3.4 Discussion

Flexible as this specification is, it rests on some assumptions, just as many marriages start with a pre-nuptial agreement.

First, we only consider two data generating processes, despite the fact that there are many alternatives to EUT and PT. Still, two processes is double the one process that is customarily assumed, and that is the main point of our analysis: to assume more than one data generating process. Furthermore, there is a natural "family similarity" between many of the alternatives that might make them hard to identify. This similarity is to be expected, given the common core of "empirical facts" about violations

[21] Haruvy et al. (2001) also explicitly consider within-type diversity in the context of a between-type mixture model, although applied to a different set of behavioral theories than what we use here. They find that a measure of the intensity of individual calculator use helps explain the allocation to their types.

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

of EUT that spawned them. But it implies low power for any statistical specification that seeks to identify their individual contribution to explaining behavior, unless one focuses on "trip wire" tasks that are designed just to identify these differences by setting up one or other models for a fall if choices follow certain patterns. Limiting our attention to the two most popular theories is therefore reasonable as a first approach.

Second, we adopt specific parameterizations of the EUT and PT models, recognizing that there exist many variants. Our parameterizations are not obviously restrictive compared to those found in the literature, but our results are conditional on them. We believe that allowing for greater flexibility in parametric form would distract attention from the principal methodological point, the formal consideration of multiple data generating models. Flexible parametric forms come at a price in terms of additional core parameters to be estimated, and since we are interested in allowing for heterogeneity with respect to observable characteristics as well as data generating models, it does not make sense to add that dimension at this stage.

Third, we assume that observable characteristics of the individual have a linear effect on the parameter, and that there are no interactions. Since these characteristics will provide some of the most entertaining moments of the marriage, as discussed below, this assumption would be worth examining in future work (with even larger data sets). Again, the objective is to show how one can allow for "traditional sources of heterogeneity" in the form of observable characteristics as well as non-traditional sources of heterogeneity in the form of competing models of the data generating process. Our goal is to illustrate how one does this jointly, so that claims about one source of heterogeneity are not made by assuming away the other source of heterogeneity. In this sense our approach encompasses both sources of heterogeneity in a natural manner.

Fourth, we deliberately avoid treating the model comparison inference as one of nesting one model in another. One could constrain the PT specification to be a generalization of a variant of our EUT specification, although that would be somewhat forced.[23] The problem with this approach is that it assumes that there is one true latent decision-making process for all choices, and that is exactly what we want to remain agnostic about and let the data decide.

Finally, we do not consider much variation in task domain, other than the obvious use of gain, loss or mixed frames. Our priors are that the relative explanatory power of EUT and PT will vary with demographics *and* task domain, and possibly some interaction of the two. To provide sharp results we therefore deliberately controlled the variability of the task domain, and focus on the possible effect of demographics.

These priors also imply that we prefer not to use mixture specifications in which *subjects* are categorized as completely EUT or PT. It is possible to rewrite the grand likelihood (1) such that $\pi_j^{\text{EUT}} = 1$ for subject j if $\sum_{i(j)} l_i^{\text{EUT}} > \sum_{i(j)} l_i^{\text{PT}}$ and $\pi_i^{\text{EUT}} = 0$ if $\sum_{i(j)} l_i^{\text{EUT}} < \sum_{i(j)} l_i^{\text{PT}}$, where the notation $i(j)$ denotes those observation i of subject j. The problem with this specification is that it assumes that there is

[23] Set $\gamma = \lambda = 1$, $\alpha = \beta = r$, and assume $U(s,x) = x^r$ for $x \geq 0$ and $U(s,x) = -\lambda(-x)^r$ for $x < 0$. It will, however, be useful later to quickly compare EUT and Rank-Dependent Utility specifications by testing the hypothesis that $\lambda = 1$.

no effect on the probability of EUT and PT from task domain. It is well known from experimental evidence that task domain can influence the strength of support for EUT. For example, Starmer (2000; p. 358) notes from his review that "...behavior on the interior of the probability triangle tends to conform more closely to the implications of EUT than behavior at the borders." Thus we want to allow the same subject to behave in accord with EUT for some choices, and in accord with PT for other choices, even for a relatively homogeneous task such as ours. Our approach allows that, by being agnostic about the interpretation of the mixing probability.

One could alternatively use a categorization mixture specification in which each binary *choice* was classified as wholly EUT or PT. We see no inferential advantage in this assumption for our purposes. Moreover, we want to allow the observed behavior to be interpreted in terms of "dual criteria decision making." This is the idea that the subject might use more than one criteria for evaluating whether to pick the left or right lottery. Although historically rare in economics, such models are easy to find in psychology.[24] There is also a wide class of "dual self" models becoming popular from behavioral economics, building on insights about the manner in which the brain resolves conflicts (e.g., see Cohen 2005; Benhabib and Bisin 2005; Fudenberg and Levine 2006).

For all of these reasons we want to avoid restricting the *formal* estimation and interpretation of the mixture probability as categorizing choices or subjects. We will, however, admit of informal *interpretations* of results in those terms, as long as the formal difference is noted.

4 Results

4.1 Heterogeneity of process

Table 1 presents maximum likelihood estimates of the conditional models as well as the mixture model when we assume no individual covariates. The estimates for the conditional models therefore represent the traditional representative agent assumption, that every subject has the same preferences *and* behaves in accord with one theory or the other. The estimates for the mixture model allow each theory to have positive probability, and therefore take one major step towards recognizing heterogeneity in decision making.

The first major result is that the *estimates for the probability of the EUT and PT specifications indicate that each is equally likely* for these data. Further, each estimated probability is significantly different from zero.[25] EUT wins by a (quantum) nose, with an estimated probability of 0.55, but the whole point of this approach

[24] See Starmer (2000) and Brandstätter et al. (2006) for reviews of many of these models. One important example is the SP/A model of Lopes (1995) and Lopes and Oden (1999), examined by Andersen et al. (2006b) from a mixture perspective. This model posits that subjects evaluate each choice using an SP criteria, which is akin to a rank-dependent utility evaluation, as well as an A criteria, which is a simple aspiration index akin to a utility threshold.

[25] The likelihood function actually estimates the log odds in favor of one model or the other. If we denote the log odds as κ, one can recover the probability for the EUT model as $\pi^{\text{EUT}} = 1/(1+\exp(\kappa))$. This non-

Table 1 Estimates of parameters of models with no individual covariates

Parameter or Test	Estimates from Conditional Models					Estimates from Mixture Model				
	Estimate	Standard Error	*p*-value	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Estimate	Standard Error	*p*-value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
r	0.867	0.029	0.000	0.809	0.924	0.846	0.044	0.000	0.759	0.933
α	0.710	0.046	0.000	0.620	0.801	0.614	0.057	0.000	0.501	0.727
β	0.723	0.065	0.000	0.695	0.851	0.312	0.132	0.019	0.052	0.572
λ	1.380	0.223	0.000	0.940	1.820	5.781	1.612	0.000	2.598	8.965
γ	0.911	0.061	0.000	0.790	1.033	0.681	0.047	0.000	0.587	0.774
π^{EUT}						0.550	0.072	0.000	0.407	0.692
π^{PT}						0.450	0.072	0.000	0.308	0.592
H_0: $\pi^{\text{EUT}} = \pi^{\text{PT}}$								0.490		
H_0: $\alpha = \beta$			0.861					0.046		
H_0: $\lambda = 1$			0.090					0.003		
H_0: $\gamma = 1$			0.151					0.000		

is to avoid such rash declarations of victory. The data suggest that the sample is roughly evenly split between those observations that are best characterized by EUT and those observations that are best characterized by PT. In fact, one cannot reject the formal hypothesis that the probabilities of each theory are identical and equal to ½ (p-value = 0.490).

The second major result is that the estimates for the PT specification are only weakly consistent with the *a priori* predictions of that theory when the specification is assumed to fit every subject, as in the conditional estimation, but *strikingly consistent with the predictions of the theory when it is only assumed to fit some of the subjects*, as in the mixture model. When the conditional PT model is assumed for all subjects and choices, the loss aversion parameter is greater than 1, but only by a small amount. The estimated coefficient of 1.380 is significantly different from 1 at the 5% significance level, since the p-value is 0.090 and a one-sided test is appropriate here given the priors from PT that $\lambda > 1$. But the size of the effect is *much* smaller than the 2.25 estimated by Tversky and Kahneman (1992) and since then almost universally employed.[26] However, when the same PT specification is estimated in the mixture model, where it is only assumed to account for the behavior of some of the subjects and choices, the estimated value for λ jumps to 5.781, clearly greater than 1 (although the estimated standard error also increases, from 0.223 to 1.612). Similarly, the γ parameter is estimated to be 0.911 with the conditional PT model, and is not statistically different from 1 (p-value = 0.151), which is the special EUT case of the probability weighting function where $w(p) = p$ for all p. But when the mixture model is estimated, the value of γ drops to 0.681 and is significantly smaller than 1, again consistent with many of the priors from PT. Finally, the risk aversion coefficients α and β are not significantly different from each other under the conditional model assumption, but are different from each other under the mixture model. This difference is not critical to PT, and is often assumed away in applied work with PT, but it is worth noting since it points to another sense in which gains and losses are evaluated differently by subjects, which is a primary tenet of PT.

4.2 Heterogeneity of process and parameters

Table 2 presents estimates from the mixture model with all individual characteristics included.[27] Each subject has a different implied coefficient for a core parameter, given by the set of estimates in Table 2 and their individual characteristics. For example, a

linear function of κ can be easily calculated from the estimates, and the "delta method" used to provide estimates of the standard errors and p-values (Oehlert 1992). Since $\pi^{EUT} = 1/2$ when $\kappa = 0$, the standard p-value on the estimate of κ provides the estimate for the null hypothesis $H_0 : \pi^{EUT} = \pi^{EUT}$ listed in Table 1.

[26] For example, Benartzi and Thaler (1995; p. 79) assume this value in their evaluation of the "equity premium puzzle," and note (p. 83) that it is the assumed value of the loss aversion parameter that drives their main result.

[27] One could further extend this analysis to allow for individual random effects on estimated coefficients (e.g., Train 2003; Chap. 6), but this would add considerable complexity to the estimation. Our goal is to demonstrate that one can combine an allowance for heterogeneity of parameters and process, and not to claim that this is the only way to do so.

Table 2 Estimates of parameters of mixture model with individual covariates

Parameter	Variable	Estimate	Standard Error	p-value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
r	Constant	−0.246	0.352	0.485	−0.940	0.449
	Female	−0.273	0.099	0.007	−0.469	−0.076
	Black	−0.043	0.195	0.824	−0.428	0.341
	Hispanic	−0.586	0.174	0.001	−0.929	−0.242
	Age	0.065	0.018	0.001	0.028	0.101
	Business	−0.104	0.099	0.291	−0.299	0.090
	GPAlow	0.042	0.083	0.616	−0.122	0.205
α	Constant	0.549	0.256	0.034	0.043	1.055
	Female	−0.201	0.215	0.350	−0.625	0.223
	Black	−0.101	0.216	0.640	−0.529	0.326
	Hispanic	−0.128	0.355	0.718	−0.829	0.572
	Business	0.053	0.292	0.857	−0.524	0.629
	GPAlow	0.059	0.196	0.765	−0.328	0.446
β	Constant	−0.202	1.199	0.866	−2.571	2.166
	Female	0.453	0.749	0.546	−1.026	1.933
	Black	0.300	0.291	0.303	−0.274	0.875
	Hispanic	0.172	0.504	0.733	−0.822	1.167
	Age	0.010	0.013	0.442	−0.016	0.036
	Business	0.130	0.451	0.774	−0.760	1.020
	GPAlow	0.095	0.175	0.587	−0.250	0.440
λ	Constant	1.592	7.164	0.824	−12.558	15.742
	Female	−4.007	10.037	0.690	−23.832	15.818
	Black	−4.494	2.029	0.028	−8.503	−0.486
	Hispanic	−5.083	2.053	0.014	−9.137	−1.028
	Age	0.523	0.566	0.357	−0.595	1.641
	Business	−2.981	2.226	0.183	−7.378	1.417
	GPAlow	−0.297	1.893	0.875	−4.036	3.441
γ	Constant	0.664	0.257	0.011	0.157	1.171
	Female	0.474	0.106	0.000	0.266	0.683
	Black	0.009	0.123	0.945	−0.234	0.252
	Hispanic	0.971	0.585	0.099	−0.185	2.127
	Age	−0.020	0.010	0.058	−0.041	0.001
	Business	0.333	0.180	0.065	−0.022	0.688
	GPAlow	−0.140	0.199	0.482	−0.533	0.253
κ[†]	Constant	0.558	1.268	0.660	−1.946	3.062
	Female	1.638	0.507	0.002	0.637	2.640
	Black	2.387	1.715	0.166	−1.001	5.775
	Hispanic	1.543	3.714	0.678	−5.793	8.880
	Age	−0.120	0.059	0.045	−0.237	−0.003
	Business	0.774	0.592	0.193	−0.395	1.943
	GPAlow	0.477	0.612	0.437	−0.732	1.686

[†] κ is the log odds of the probabilities of each model, where $\pi^{\text{EUT}} = 1/(1+\exp(\kappa))$

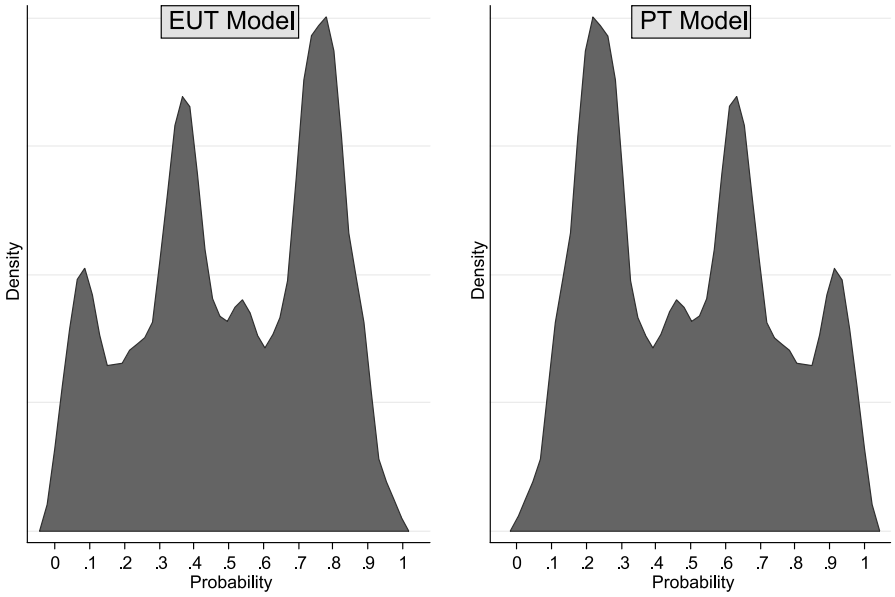

Fig. 1 Probability of competing models

black 21-year-old female that was not hispanic, did not have a business major, and had an average GPA, would have an estimate of r equal to $\hat{r}_0 + \hat{r}_{\text{FEMALE}} + \hat{r}_{\text{BLACK}} + \hat{r}_{\text{AGE}} \times 21$. She would have the same estimate of r as anyone else that had *exactly* the same set of characteristics. But the set of estimated characteristics is reasonably large, allowing considerable heterogeneity for a given subject.

The effect of allowing for this observable heterogeneity can be best seen by examining the distribution of coefficients across the sample, generated by predicting each parameter for each subject by evaluating the linear function with the characteristics of that subject. Such predictions are accompanied by estimated standard errors, just as the coefficient estimates in Table 1 are, so we also take the uncertainty of the predicted parameter value into account.

Figure 1 displays the main result, a kernel distribution of predicted probabilities for the competing models. The two panels are, by construction, mirror images of each other since we only have two competing models, but there is pedagogic value in displaying both. The support for the EUT specification is extremely high for some people, but once that support wanes the support for PT picks up. It is not the case that there are two sharp modes in this distribution, which is what one might have expected based on a prior that there are two distinct types of subjects. Instead, subjects are better characterized as either "clearly EUT" or "probably PT," although there is also a small proportion who are "clearly PT." It is as if EUT is fine for city folk, but PT rules the hinterland.[28] No doubt this is due to many subjects picking lotteries

[28] Andersen et al. (2006a) report the *reverse* qualitative pattern when considering *dynamic* lottery choices in which subjects could accumulate income. Conte et al. (2007) report roughly equal weight attached to

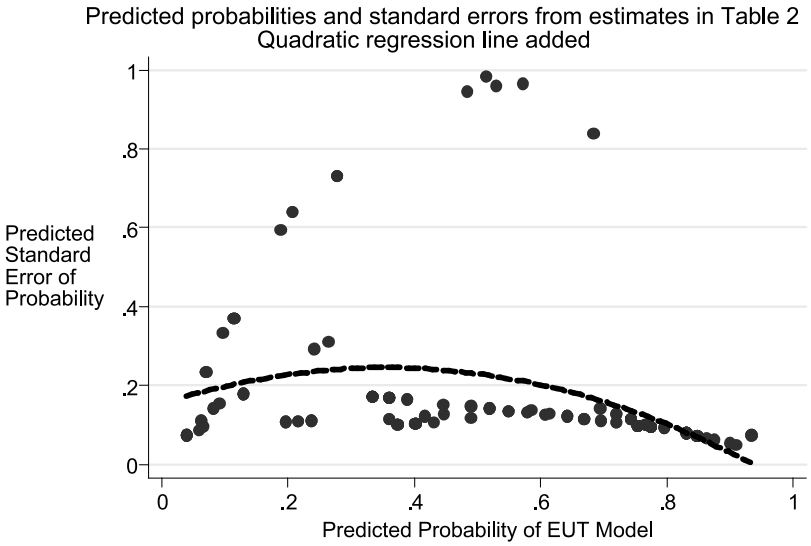

Fig. 2 Uncertainty in predicted probability of models

according to their estimate of the expected value of the lotteries, which would make them EUT-consistent and risk neutral ($r = 1$). The average probability for the EUT model is 0.51, as is the median, despite the apparent skewness in the distributions in Fig. 1.

Figure 2 indicates the uncertainty of these predicted probabilities. Consistent with the use of discrimination functions such as the logistic, uncertainty is smallest at the end-points and greatest at the mid-point. Since EUT has relatively more of its support closer to the upper end-point, this implies that one can express more confidence when declaring some subjects as better characterized by EUT than one can about the alternative model. There are some subjects that have significantly higher predicted standard errors in Fig. 2, reflecting their choices being either sharply consistent with EUT or sharply inconsistent.

Figures 3 and 4 display illustrative stratifications of the probability of the EUT model being correct in terms of individual characteristics. The results are striking. Figure 3 shows that men (52% of our sample) have a much stronger probability of behaving in accord with the EUT model than women. Figure 4 shows that the ethnicity group "Others" (including Whites, Asians and Mixed race and representing 75% of our sample) have a much higher probability of being EUT decision makers than "Blacks" (11% of the sample) or "Hispanics" (14% of the sample). Consider the implications of these two figures for the dramatic differences that List (2002, 2003, 2004) finds between subjects in field experiments on the floor of sports-card shows and conventional lab experiments with college students. The *vast* majority

an EUT and non-EUT specification, with very few subjects being in-between: subjects are either "clearly EUT" or "clearly not EUT" in their results. We therefore caution again that the relative explanatory power of EUT and PT likely interacts with task domain.

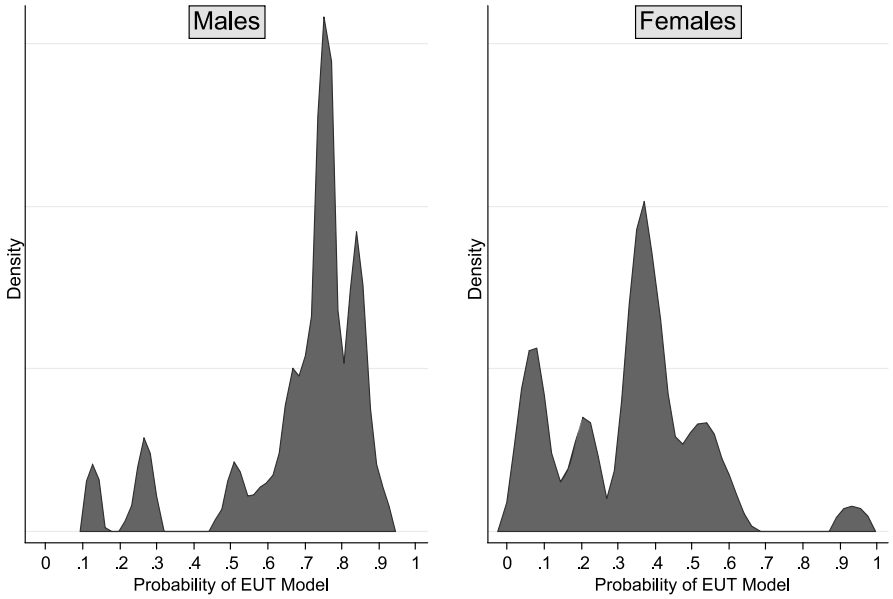

Fig. 3 Sex, EUT and prospect theory

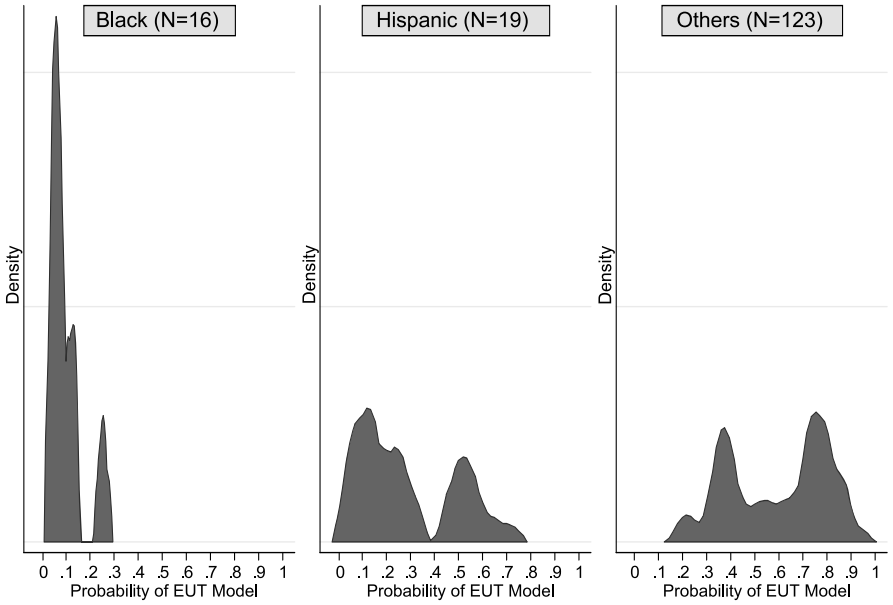

Fig. 4 Ethnicity, EUT and prospect theory

of participants on those shows are white males. No doubt there is much more going on in these field experiments than simply demographic mix, as stressed by Harrison

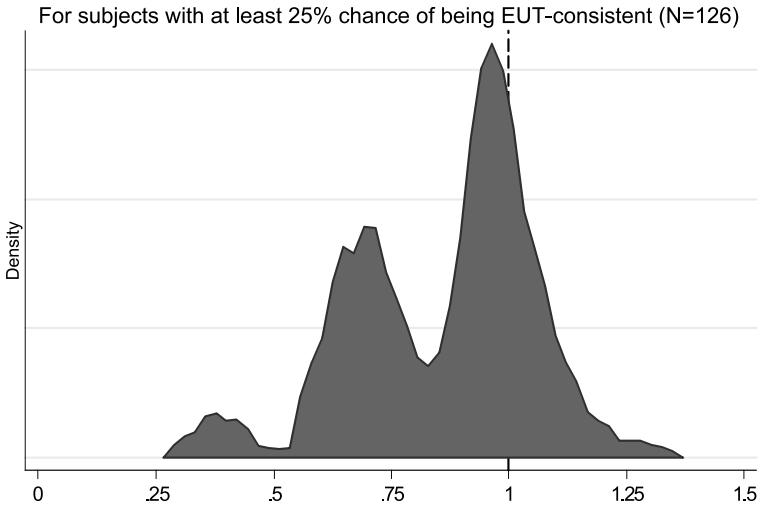

Fig. 5 CRRA parameter of the EUT model

and List (2004), but one might be able to go a long way to explaining the differences on these two demographics alone.[29]

Figures 5 and 6 display the distributions of the parameters of each model. The raw predictions for all 158 subjects include some wild estimates, but for completely sensible reasons. If some subject has a probability of behavior in accord with PT of less than 0.0001, for example, then one should not expect the estimates of α, β, λ or γ to be precise for that subject. In fact, most of the extreme estimates are associated with parameters that have very low support for that subject, or that have equally high standard errors.[30] Thus, we restrict the displays in Figs. 5 and 6 to those subjects whose probability of behaving in accord with EUT or PT, respectively, is at least ¼.

Figure 5 displays results on the risk aversion coefficient for the EUT subjects that are remarkably similar to those reported in Table 1. The average EUT subject has a CRRA coefficient r equal to 0.89 when we include observable individual characteristics and estimate a mixture specification, whereas the estimate from Table 1 was 0.87 when all EUT subjects were assumed to have the same risk attitude. In Fig. 5 we observe two modes, reflecting the ability to allow for heterogeneity of risk attitudes

[29] Haigh and List (2005) report contrary evidence, using tasks involving risky lotteries and comparing field traders from the Chicago Board of Trade and students. However, the behavior they observe can be easily explained under EUT using a specification of the utility function that allows non-constant relative risk aversion, such as the expo-power specification favored by Holt and Laury (2002) in their experiments. Thus one should not count this as contrary evidence to the claim in the text unless one constrains EUT to the special case of CRRA. Harrison and Rutström (2008) discuss this point in more detail.

[30] This accounts for some extremely wide 95% confidence interval estimates in Table 2. It would also be possible to constrain estimates to be within ranges predicted *a priori*, such as $\gamma \leq 1$ and $\lambda \geq 1$, but such restrictions are often the result of previous empirical analyses and intuition rather than derived from theory. As one example of a restriction with a theoretical rationale, Köbberling and Wakker (2005; p. 127ff.) argue that one should set $\alpha = \beta$ if one is using the CRRA functional form in a model used to identify loss aversion (λ).

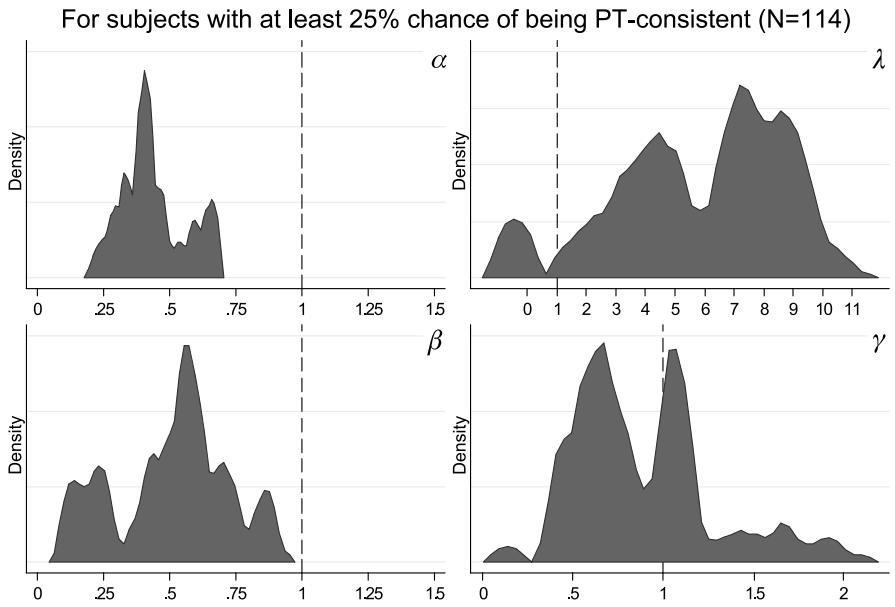

Fig. 6 Parameters of the PT moodel

in the mixture model by including individual characteristics. Some subjects exhibit considerable risk aversion ($r < 1$), and some even exhibit mild risk-loving behavior ($r > 1$). The most prominent mode is found right around risk-neutrality, $r = 1$. From Table 2 we see that the characteristics that drive this heterogeneity include sex, ethnicity, and age.

Figure 6 displays results for the PT parameters that are generally consistent with the estimates in Table 1, although there are some differences due to the use of observable characteristics in the model underlying Fig. 6. The parameters α and β are generally less than 1, consistent with mildly concave (convex) utility in the gain (loss) frame. The average estimate for α is 0.44, and the average estimate for β is 0.51. The loss aversion coefficient λ tends to be much greater than 1, averaging 5.81, consistent with the subjects that behave in accord with PT having significant loss aversion.[31] Of course, this applies only to the fraction of the sample deemed to have certain level of support for PT; the value of λ for EUT subjects is 1, by definition, so the weighted average for the whole sample would be much smaller. The probability weighting coefficient γ is generally less than 1, and averages 0.89 across the sample.

5 Comparing models

The idea of modeling mixtures of latent processes is a natural one, but how does it compare to other approaches used to compare models? Whenever one considers two non-nested models, readers expect to see some comparative measures of goodness of

[31] The restriction that $\lambda = 1$ is easily rejected using a Wald test, with a p-value of only 0.014.

fit. Common measures include R^2, pseudo-R^2, a "hit ratio," some other scalar appropriate for choice models (e.g., Hosmer and Lemeshow 2000; Chap. 5), and formal likelihood-ratio tests of one model against another (e.g., Cox 1961, 1962 or Vuong 1989). From the perspective adopted here, the *interpretation* of these tests suffers from the problem of implicitly assuming just one data-generating process. In effect, the mixture model provides a built-in comparative measure of goodness of fit—the mixture probability itself. If this probability is close to 0 or 1 by standard tests, one of the models is effectively rejected, in favor of the hypothesis that there is just one data-generating process.

In fact, if one traces back through the literature on non-nested hypothesis tests, these points are "well known." That literature is generally held to have been started formally by Cox (1961), who proposed a test statistic that generalized the usual likelihood ratio test (LRT). His test compares the difference between the actual LRT of the two models with the expected LRT, suitably normalized by the variance of that difference, under the hypothesis that one of the models is the true data-generating process. The statistic is applied symmetrically to both models, in the sense that each takes a turn at being the true model, and leads to one of four conclusions: one model is the true model, the other model is the true model, neither model is true, or both models are true.[32]

However, what is often missed is that Cox (1962; p. 407) briefly, but explicitly, proposed a multiplicative mixture model as an "alternative important method of tackling these problems." He noted that this "procedure has the major advantage of leading to an estimation procedure as well as to a significance test. Usually, however, the calculations will be very complicated." Given the computational limitations of the day, he efficiently did not pursue the mixture model approach further.

The next step in the statistical literature was the development by Atkinson (1970) of the suggestion of Cox. The main problem with this exposition, noted by virtually every commentator in the ensuing discussion, was the interpretation of the mixing parameter. Atkinson (1970; p. 324) focused on testing the hypothesis that this parameter equaled ½, "which implies that both models fit the data equally well, or equally badly." There is a colloquial sense in which this is a correct interpretation, but it can easily lead to confusion if one maintains the hypothesis that there is only one true data generating process, as the commentators do. In that case one is indeed confusing model specification tests with model selection tests. If instead the possibility that there are two data generating processes is allowed, then natural interpretations of tests of this kind arise.[33] Computational constraints again restricted Atkinson (1970) to deriving results for tractable special cases.[34]

[32] This possible ambiguity is viewed as an undesirable feature of the test by some, but not when the test is viewed as one of an armada of possible model *specification* tests rather than as a model *selection* tests. See Pollak and Wales (1991; p. 227ff.) and Davidson and MacKinnon (1993; p. 384) for clear discussions of these differences.

[33] Of course, as noted earlier, there are several possible interpretations in terms of mixtures occurring at the level of the observation (lottery choice) or the unit of observation (the subject or task). Quandt (1974) and Pesaran (1981) discuss problems with the multiplicative mixture specification from the perspective of the data being generated by a single process.

[34] These constraints were even binding on methodology as recently as Pollak and Wales (1991). They note (p. 228) that "If we could estimate the composite (the mixture specification proposed by Atkinson

This idea was more completely developed by Quandt (1974) in the additive mixture form we use. He did, however, add a seemingly strange comment that "The resulting pdf is formally identical with the pdf of a random variable produced by a mixture of two distributions. It is stressed that this is a formal similarity only." (p. 93/4) His point again derives from the tacit assumption that there is only one data generating process rather than two (or more). From the former perspective, he proposes viewing corner values of the mixture probability as evidence that one or other model is the true model, but to view interior values as evidence that some unknown model is actually used and that a mixture of the two proposed models just happens to provide a better approximation to that unknown, true model. But if we adopt the perspective that there are two possible data generating processes, the use and interpretation of the mixing probability estimate is direct.

Perhaps the most popular modern variant of the generalized LRT approach of Cox (1961, 1962) is due to Vuong (1989). He proposes the null hypothesis that both models are the true models, and then allows two one-sided alternative hypotheses.[35] The statistic he derives takes *observation-specific* ratios of the likelihoods under each model, so that in our case the ratio for observation i is the likelihood of observation i under EUT divided by the likelihood of observation i under PT. It then calculates the log of these ratios, and tests whether the expected value of these log-ratios over the sample is zero. Under reasonably general conditions a normalized version of this statistic is distributed according to the standard normal, allowing test criteria to be developed.[36] Thus the resulting statistic typically provides evidence in favor of one of the models which may or may not be statistically significant.

Applying the Vuong test to the EUT and PT models estimated independently in Table 1, we would conclude that there is overwhelming evidence in favor of the PT model.[37] Since we cannot reject equality between the α and the β parameters, and λ is only weakly significantly different from 1, there is really support for the more intermediate specification of the Rank-Dependent Utility (RDU) model nested within PT. However, when we use the Vuong test of the PT-only model against the mixture

1970 and Quandt 1974), then we could use the standard likelihood ratio test procedure to compare the two hypotheses with the composite and there would no reason to focus on choosing between the two hypotheses without the option of rejecting them both in favor of the composite. Thus, the model selection problem arises only when one cannot estimate the composite." They later discuss the estimation problems in their extended example, primarily deriving from the highly non-linear functional form (p. 232). As a result, they devise an ingenious method for ranking the alternative models under the *maintained assumption* that one cannot estimate the composite (p. 230).

[35] Some have criticized the Vuong test because the null hypothesis is often logically impossible, but it can also be interpreted as the hypothesis that one cannot say which model is correct.

[36] Clarke (2003) proposes a non-parametric sign test be applied to the sample of ratios. Clarke (2007) demonstrates that when the distribution of the log of the likelihood ratios is normally distributed then the Vuong test is better in terms of asymptotic efficiency. But if this distribution exhibits sharp peaks, in the sense that it is mesokurtic, then the non-parametric version is better. The likelihood ratios we are dealing with have the latter shape.

[37] The test statistic has a value of −10.33. There are often additional corrections for degrees of freedom, using one or other "information criteria" to penalize models with more parameters (in our case, the PT model). We do not accept the underlying premiss of these corrections, that smaller models are better, and do not make these corrections. The results reported below would be the same if we did.

model, the test statistic favors the mixture model; the test statistic is −0.56, with a p-value of 0.71 that the PT-only model is *not* the better model. Since the estimation of the mixture model also rejects the restrictions imposed on the PT parameters from assuming RDU, it is clear that the mixture model is showing something that is very different from a representative agent model specification that is simply intermediate between EUT and PT. The inferences that one draws from these test statistics therefore depend critically on the perspective adopted with respect to the data generating process. If we look for a single data generating process in our case, then PT dominates EUT. But if one allows the data to be generated by either model, the evidence is mixed—if one excuses the pun, and correctly interprets that as saying that both models receive roughly the same support. *Thus one would be led to the wrong qualitative conclusion if the non-nested hypothesis tests had been mechanically applied.*

6 Conclusion

Characterizing behavior in lottery choice tasks as generated by potentially heterogenous *processes* generates several important insights.

First, it provides a framework for systematically resolving debates over competing models.[38] This method does not rest on exploiting structural aspects of one type of task or another, which has been a staple in tests of EUT. Such extreme domains are still interesting of course, just as the experiments employed here deliberately generated loss frames and mixed frames. But the test of the competing models should not be restricted only to such "trip-wire" tests, which might not characterize behavior in a broader domain.

Second, it provides a more balanced metric for deciding which theory does better in a given domain, rather than extreme declarations of winners and losers. If some theory has little support, this approach will show that. A corollary is that one should avoid mechanical use of test statistics that are predicated on the idea that there can only be one latent data generating process. In fact, mixture models provide a natural alternative to the formal statistical tests that are popular for discriminating between non-nested hypotheses (and always have, at least in the older literature).

Third, the approach can provide some insight into *when* one theory does better than another, through the effects of individual characteristics and treatments on estimated probabilities of support. This insight might be fruitfully applied to other settings in which apparently conflicting findings have been reported. We fully expect that the relative explanatory power of EUT and PT will vary with task domain as well as demographics. This is one reason we reject *a priori* the restriction, explicit in categorical mixture specifications, that individuals employ one model or the other in all task domains. For the same reason we reject the assumption that one model or the other applies to every individual in a given task domain. The future, constructive challenge is to characterize those domains so that we know better when to use one

[38] The approach readily generalizes to other contexts, such as competing models of intertemporal discounting behavior (e.g., Andersen et al. 2008).

model or the other. To meet that challenge we must remain agnostic about the domain over which the choice process applies (viz., choices, individuals, or tasks).

Fourth, the approach readily generalizes to include more than two models, although likelihoods are bound to become flatter as one mixes more and more models that have a modicum of support from the data. The obvious antidote for those problems is to generate larger samples, to pool data across comparable experiments, and to marshal more exhaustive numerical methods.

References

Andersen, S., Harrison, G. W., & Rutström, E. E. (2006a). *Choice behavior, asset integration, and natural reference points* (Working Paper 06-07). Department of Economics, College of Business Administration, University of Central Florida.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2006b). *Dual criteria decisions* (Working Paper 06-11). Department of Economics, College of Business Administration, University of Central Florida.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

Araña, J. E., & León, C. J. (2005). Flexible mixture distribution modeling of dichotomous choice contingent valuation with heterogeneity. *Journal of Environmental Economics & Management*, *50*(1), 170–188.

Atkinson, A. C. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society, Series B*, *32*, 323–344.

Bardsley, N., & Moffatt, P. G. (2007). The experimetrics of public goods: inferring motivations from contributions. *Theory and Decision*, *62*(2), 161–193.

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

Benhabib, J., & Bisin, A. (2005). Modeling internal commitment mechanisms and self-control: a neuroeconomics approach to consumption-saving decisions. *Games and Economic Behavior*, *52*, 460–492.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*(2), 409–432.

Bruhin, A., Fehr-Duda, H., & Epper, T. F. (2007). *Risk and rationality: uncovering heterogeneity in probability distortion* (Working Paper 0705). Socioeconomic Institute, University of Zurich.

Camerer, C. F. (1995). Individual decision making. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics*. Princeton: Princeton University Press.

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

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218–1221.

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

Clarke, K. A. (2007). A simple distribution-free test for non-nested model selection. *Political Analysis*, *15*(3), 347–363.

Cohen, J. D. (2005). The vulcanization of the human brain: a neural perspective on interactions between cognition and emotion. *Journal of Economic Perspectives*, *19*(4), 3–24.

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

Cox, D. R. (1961). Tests of separate families of hypotheses. In E. G. Charatsis (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 105–123). Berkeley: University of California Press.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, *24*, 406–424.

Cox, J. C., & Sadiraj, V. (2006). Small- and large-stakes risk aversion: implications of concavity calibration for decision theory. *Games & Economic Behavior*, *56*(1), 45–60.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.
El-Gamal, M. A., & Grether, D. M. (1995). Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association*, *90*(432), 1137–1145.
Everitt, B. S. (1996). An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, *5*, 107–127.
Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *American Economic Review*, *96*(5), 1449–1476.
George, J. G., Johnson, L. T., & Rutström, E. E. (2007). Social preferences in the face of regulatory change. In T. Cherry, S. Kroll, & J. F. Shogren (Eds.), *Experimental methods, environmental economics*. Oxford: Routledge.
Geweke, J., & Keane, M. (1999). Mixture of normals probit models. In C. Hsio, K. Lahiri, L.-F. Lee, & M. H. Pesaran (Eds.), *Analysis of panel and limited dependent variables: a volume in honor of G.S. Maddala*. New York: Cambridge University Press.
Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I. A modified latent structure approach. *American Journal of Sociology*, *79*, 1179–1259.
Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.
Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, *69*(4), 623–648.
Haigh, M., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis. *Journal of Finance*, *60*(1), 523–534.
Harless, D. W., & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*, 1251–1289.
Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1013–1059.
Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. In J. C. Cox & G. W. Harrison (Eds.), *Research in experimental economics: Vol. 12. Risk aversion in experiments*. Bingley: Emerald.
Harrison, G. W., Humphrey, S. J., & Verschoor, A. (2005). *Choice under uncertainty: evidence from Ethiopia, India and Uganda* (Working Paper 05-29). Department of Economics, College of Business Administration, University of Central Florida.
Haruvy, E., Stahl, D. O., & Wilson, P. W. (2001). Modeling and testing for heterogeneity in observed strategic behavior. *Review of Economics and Statistics*, *83*(1), 146–157.
Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*(2), 271–320.
Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*(6), 1291–1326.
Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd edn.). New York: Wiley.
Hurley, T. M., & Shogren, J. F. (2005). An experimental comparison of induced and elicited beliefs. *Journal of Risk & Uncertainty*, *30*(2), 169–188.
Johnson, L. T., Rutström, E. E., & George, J. G. (2006). Income distribution preferences and regulatory change in social dilemmas. *Journal of Economic Behavior & Organization*, *61*(2), 181–198.
Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–291.
Kirman, A. P. (1992). Whom or what does the representative individual represent? *Journal of Economic Perspectives*, *6*(2), 117–136.
Köbberling, V., & Wakker, P. P. (2005). An index of loss aversion. *Journal of Economic Theory*, *122*, 119–131.
Kumbhakar, S. C., & Lovell, C. A. K. (2000). *Stochastic frontier analysis*. New York: Cambridge University Press.
Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.
List, J. A. (2002). Preference reversals of a different kind: the more is less phenomenon. *American Economic Review*, *92*, 1636–1643.

List, J. A. (2003). Does market experience eliminate market anomalies. *Quarterly Journal of Economics*, *118*, 41–71.
List, J. A. (2004). Neoclassical theory versus prospect theory: evidence from the marketplace. *Econometrica*, *72*(2), 615–625.
Loomes, G., & Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, *65*, 581–598.
Lopes, L. L. (1995). Algebra and process in the modeling of risky choice. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from a cognitive perspective*. San Diego: Academic Press.
Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: a comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, *43*, 286–313.
Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 249–410). New York: Wiley.
Marschak, J. (1960). Binary choice constraints on random utility indications. In K. Arow (Ed.), *Stanford symposium on mathematical models in the social sciences*. Stanford: Stanford University Press.
McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29.
Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics*, *11*, 619–632.
Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions A*, *185*, 71–110.
Pesaran, M. H. (1981). Pitfalls of testing non-nested hypotheses by the Lagrange multiplier method. *Journal of Econometrics*, *17*, 323–331.
Pollak, R. A., & Wales, T. J. (1991). The likelihood dominance criterion: a new approach to model selection. *Journal of Econometrics*, *47*, 227–242.
Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*(3), 497–527.
Quandt, R. E. (1974). A comparison of methods for testing nonnested hypotheses. *Review of Economics and Statistics*, *56*, 92–99.
Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13*, 19–23.
Schoemaker, P. (1982). The expected utility model: its variants, purposes, evidence and limitations. *Journal of Economic Literature*, *20*(2), 529–563.
Stahl, D. O. (1996). Boundedly rational rule learning in a guessing game. *Games and Economic Behavior*, *16*, 303–330.
Stahl, D. O. (1998). Is step-j thinking an arbitrary modelling restriction or a fact of human nature? *Journal of Economic Behavior & Organization*, *37*, 33–51.
Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: theory and experimental evidence. *Games and Economic Behavior*, *10*, 218–254.
Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*, 332–382.
Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.
Train, K. E. (2003). *Discrete choice methods with simulation*. New York: Cambridge University Press.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representations of uncertainty. *Journal of Risk & Uncertainty*, *5*, 297–323.
Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, *41*, 531–537.
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333.
Wang, M., & Fischbeck, P. S. (2004). Incorporating framing into prospect theory modeling: a mixture-model approach. *Journal of Risk & Uncertainty*, *29*(2), 181–197.
Werner, M. (1999). Allowing for zeros in dichotomous choice contingent valuation models. *Journal of Business and Economic Statistics*, *17*, 479–486.
Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56*, 645–646.
Wooldridge, J. (2003). Cluster-sample methods in applied econometrics. *American Economic Review (Papers & Proceedings)*, *93*, 133–138.
Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, *42*, 1676–1690.