A Comparison of Three Probabilistic Models of Binary Discrete Choice Under Risk

by

Nathaniel T. Wilcox*

Abstract

This paper compares the “out-of-context” predictive success of three probabilistic models of binary discrete choice under risk. One of the models is the conventional homoscedastic latent index or “strong utility” model that is widespread in applied econometrics: This model is “context-free” in the sense that its error part is homoscedastic with respect to decision sets. The other two models are also latent index models, but their error part is heteroscedastic with respect to decision sets, and in that sense are “context-dependent” models. Context-dependent models of choice under risk arise from several different theoretical perspectives. Here I consider my own “contextual utility” model (Wilcox 2009) and the “decision field theory” model (Busemeyer and Townsend 1993). A new experiment is performed on 80 subjects. Two-thirds of the data is used to estimate models at the individual level, and these estimates are used to predict the remaining third of choices. The data is divided up so that the decision sets in the estimating data and the prediction data have interestingly different contexts. The context-dependent error models consistently outperform the context-free error model in prediction.

JEL Classification Codes: C25, C91, D81

Keywords: risk, discrete choice, probabilistic choice, heteroscedasticity, prediction.

Working Draft, March 2011 (not yet complete).

*ESI (Chapman University) and CEAR (Georgia State University). Phone 714-628-7212, fax 714-628-2881, email nwilcox@chapman.edu. I thank Stacey Joldersma for her excellent research assistance.
Beginning with Mosteller and Nogee (1951), dozens of experiments on discrete choice under risk established that discrete choice under risk appears to have a strong probabilistic, random or stochastic component. These experiments involve repeated trials of binary choice pairs, and reveal substantial choice switching by the same subject between trials. In some cases, the trials span days (e.g. Tversky 1969; Hey and Orme 1994; Hey 2001) and one might worry that decision-relevant conditions may have changed between trials. Yet similarly substantial switching occurs even between trials separated by bare minutes, with no intervening change in wealth, background risk, or any other obviously decision-relevant variable (Camerer 1989; Starmer and Sugden 1989; Ballinger and Wilcox 1997; Loomes and Sugden 1998).

Since Kahneman and Tversky (1979) introduced Prospect Theory, most research on choice under risk has concerned its “structure,” that is the functional form or “representation” that describes how lottery characteristics (outcomes, events and their likelihoods) are functionally combined to represent binary preference directions. Econometrically, that discussion concerns the functional form taken by the fixed part of the latent index in a traditional discrete choice model. However, there is resurgent interest in the stochastic part of decision under risk. This has been driven both by theoretical questions and empirical findings. Theoretically, some or all of what passes for “an anomaly” (say, an apparent violation of expected utility or EU theory) can be attributed to stochastic models rather than the structure in question (Wilcox 2008). This old point goes back at least to Becker, DeGroot and Marschak’s (1963a,1963b) observation that violations of the “betweenness” property of EU are precluded by some probabilistic versions of EU (random preferences) but allowed by others (strong utility). But this general concern has been resurrected by many writers; Loomes (2005), Gul and Pesendorfer (2006) and Blavatskyy (2007) are just three relatively recent (but very different) examples.
In this paper I compare three probabilistic models of choice under risk. One of the models is the conventional homoscedastic latent index or “strong utility” model that is widespread in applied econometrics: This model is “context-free” in the sense that its random part is homoscedastic with respect to decision sets. The other two models are also latent index models, but their error part is heteroscedastic with respect to decision sets, and in that sense these two models are “context-dependent.” Context-dependent models of choice under risk arise from several different theoretical perspectives. Here I consider my own “contextual utility” model (Wilcox 2009) and the “decision field theory” model of Busemeyer and Townsend (1993). A new experiment is performed on 80 subjects. Two-thirds of the data is used to estimate models at the individual level, and these estimates are then used to predict the remaining third of choices. The data is divided up so that the decision sets in the estimating data and the prediction data have interestingly different contexts. The context-dependent error models consistently outperform the context-free error model in prediction.

1. Preliminaries

In my experiment, each choice pair is a set of two options \( \{ \text{risky, safe} \} \equiv \{(h, q, l), m\} \).

The option safe pays \( m \) dollars with certainty, while the option risky pays \( h \) dollars with probability \( q \) and \( l \) dollars with probability \( 1 - q \), where \( h > m > l \). Subjects choose between risky and safe in each pair presented to them. I call the vector of outcomes \( (l, m, h) \) the context of each pair. Figure 1 shows an example pair where \( \{ \text{risky, safe} \} \equiv \{(90,1/6,40), 50\} \) and the context of the pair is \( (40,50,90) \). One familiar interpretation of each specific context is that it names a specific Machina/Marschak triangle (see e.g. Machina 1987) representing all possible
pairs of lotteries composed solely from the three possible outcomes 40, 50 and 90. Figure 2 shows this representation of the example pair in a Machina/Marschak triangle.

I consider a class of probabilistic choice models of the form

$$P \equiv \Pr(\text{risky}) = F \left[ \lambda \frac{V(\text{risky}) - V(\text{safe})}{D(\text{risky}, \text{safe})} \right],$$

where $V(\text{risky}) - V(\text{safe})$ is a decision-theoretic representation of the difference between the values of the options \textit{risky} and \textit{safe}, $\lambda$ is a scale (or inverse standard deviation) parameter, $D(\text{risky}, \text{safe})$ is an adjustment to the scale parameter in heteroscedastic models, and $F$ is a cumulative distribution function or c.d.f. where $F(0) = 0.5$ and $F(x) = 1 - F(-x)$.

While my focus is on assumptions about the function $D(\text{risky}, \text{safe})$, I first discuss the “value difference” $\Delta V \equiv V(\text{risky}) - V(\text{safe})$. The function $V$ needs to be a decision-theoretic representation of lottery value with theoretical breadth and empirical strength. Rank-dependent utility or RDU, developed by Quiggin (1982), Chew (1983) and many others, fits this bill. Under RDU, the values of two-outcome options like \textit{risky}, and single outcome options like \textit{safe}, are

$$V_{\text{risky}} = w(q)u(h) + [1 - w(q)]u(l) \quad \text{and} \quad V_{\text{safe}} = u(m),$$

where $u(z)$ is the \textit{utility} of outcome $z$; and $w(q)$ is the \textit{weight} associated with the probability $q$ of receiving the highest outcome $h$ in the pair \{\textit{risky},\textit{safe}\}.

The \textit{RDU value difference} between \textit{risky} and \textit{safe} in a pair is thus

$$\Delta RDU = w(q)u(h) + [1 - w(q)]u(l) - u(m).$$
RDU nests the expected utility or EU representation: EU is just that special case of RDU where \( w(q) \equiv q \). Therefore we develop all choice models below in terms of \( \Delta RDU \). To convert those into EU-based models, just replace \( w(q) \) by \( q \) in 3 to get

\[
\Delta EU = qu(h) + (1-q)u(l) - u(m),
\]

the EU value difference between risky and safe.

Special experimental design choices also make the RDU representation indistinguishable from both Tversky and Kahneman’s (1992) cumulative prospect theory (or CPT) and Savage’s (1954) subjective expected utility (or SEU) representation. Cumulative prospect theory differs from RDU only in its treatment of negative outcomes or, more correctly, outcomes below some reference point (put differently, CPT posits loss aversion), and my experiment pairs contain only large positive outcomes of $40 to $120. In general, RDU is not a subjective expected utility model since the weight associated with an outcome will in general change when the rank order of an outcome differs in two different lotteries, regardless of whether the event(s) generating that outcome are held constant across those two lotteries. But if the mapping between events and outcome ranks is held constant across all lotteries, then SEU is indistinguishable from RDU. My experimental pairs intentionally satisfy this requirement as well.¹

This implies that the RDU representation in eq. 2 will be very broad, equivalent to (or nesting) all of RDU, CPT, SEU, EU and EV (expected value). If we wished to distinguish between these representations, this deliberate confounding would be a bug, but here it is a feature.

¹ More concretely: In the experiment, lotteries risky all have probabilities \( q \) of receiving their high outcome that are in sixths, generated by the roll of a six-sided die. All lotteries are constructed so that \( q = k/6 \) is always the roll “1 or 2 or...k”. So \( w(k/6) \), the weight on the high outcome \( h \) in \( \text{risky} \), can always be thought of as the subjective probability of the event “the die roll is 1 or 2 or...k”, while \( 1-w(k/6) \), the weight on the low outcome \( l \) in \( \text{risky} \), can always be thought of as the subjective probability of the event “the die roll is \( k+1 \) or \( k+2 \) or...6”.
since my interest lies with the scale adjustment $D(risky, safe)$.

By experimental design, the RDU representation of $\Delta V \equiv V(risky) - V(safe)$ will encompass this wide set of decision-theoretic representations, so all inferences concerning $D(risky, safe)$ will be robust for this set of decision-theoretic representations.

Decision theory knows the first probabilistic model as the “strong utility” or SU model (Debreu 1958; Block and Marschak 1960; Luce and Suppes 1965), and econometrics knows this as the homoscedastic latent index model. It imposes the restriction $D(risky, safe) \equiv 1$ on eq. 1, and with RDU it is

\begin{equation}
(5) \quad P^{\text{rdu}} \equiv \Pr(risky) = F(\lambda \Delta \text{RDU}).
\end{equation}

As is well-known (Luce 1959), if we choose the logistic c.d.f. $\Lambda(x)$ as $F(x)$, this is equivalent to a binary logit with RDU lottery values:

\begin{equation}
(6) \quad P^{\text{rdu}} \equiv \Pr(risky) = \frac{\exp(\lambda V_{\text{risky}})}{\exp(\lambda V_{\text{risky}}) + \exp(\lambda V_{\text{safe}})}, \text{ with } V_{\text{risky}} \text{ and } V_{\text{safe}} \text{ as given in eq. 2}.
\end{equation}

McFadden and others developed this model in economics, and it appears widely in experimental/behavioral applied theory (e.g. McKelvey and Palfrey 1995; Camerer and Ho 1999). I use the logistic c.d.f. as $F$ in all my estimations for this reason, so that my results speak clearly to these applications.

The contextual utility or CU model (Wilcox 2009) sets $D(risky, safe) \equiv u(h) - u(l)$, and with RDU it is
Contextual utility makes comparative risk aversion properties of the RDU representation and its stochastic implications consistent within and across contexts. For representations such as RDU and EU, utility functions \( u(z) \) are only unique up to a ratio of differences: Intuitively, contextual utility exploits this uniqueness to create a correspondence between structural and stochastic definitions of comparative risk aversion. To see this, consider any pair on a context. Under RDU and contextual utility, the choice probability in eq. 7 can be rewritten as

\[
P_{\text{rdu}} = F\left( \lambda \frac{\Delta \text{RDU}}{u(h) - u(l)} \right).
\]

This probability is decreasing in the ratio of differences \( \nu(l, m, h) \). Consider two subjects Anne and Bob: Assume that they have identical weighting functions (which includes the case where both have EU preferences) and identical scale parameters \( \lambda \). Also assume that Bob is globally more risk averse than Anne in Pratt’s sense—that Bob’s local absolute risk aversion \( -u''(z)/u'(z) \) exceeds that of Anne for all \( z \). The latter assumption, and simple algebra based on Pratt’s (1964) main theorem, then implies that \( \nu^{\text{Bob}}(l, m, h) > \nu^{\text{Anne}}(l, m, h) \), on all contexts, and as a result (8) implies that Bob will have a lower probability than Anne of choosing risky on all contexts. Wilcox shows that it is mathematically impossible for strong utility to share this property, and this is the primary motivation behind the contextual utility model.

The final model is decision field theory or DFT (Busemeyer and Townsend 1993). It sets \( D(\text{risky}, \text{safé}) = [u(h) - u(l)]\sqrt{w(q)[1 - w(q)]} \), and with RDU it is
Note that eq. 9 is DFT only for pairs like those found in this experiment, where every pair consists of a two-outcome risk versus a sure outcome. In general, the function $D(risky, safe)$ varies in a complex but theoretically well-specified manner with decision sets. Notice too that in this special case DFT shares CU’s main property: Holding constant scale parameters and weighting functions, globally greater risk aversion (in the sense of Pratt) will imply a lower probability of choosing risky in all pairs on all contexts. DFT has another attractive property: As $q$ approaches zero (or one)—that is, as safe (or risky) gets closer to stochastically dominating risky (or safe)—the probability of choosing the (nearly) stochastically dominating alternative approaches certainty.

Busemeyer and Townsend (1993) derive decision field theory from a sophisticated computational logic, but a simple intuition can be given for the model. Suppose that a decision maker’s computational resources can effortlessly and quickly provide utilities of outcomes, and also suppose the decision maker wishes to choose according to relative RDU value; but suppose she does not have an algorithm for effortlessly and quickly multiplying utilities and weights together. The decision maker could proceed by sampling the possible utilities in options in proportion to their decision weights, keeping running sums of these sampled utilities for each option, and stop (and choose) when the difference between the sums exceeds some threshold determined by the cost of sampling. In essence, the choice probability in eq. 9 results from this kind of sequential sampling decision procedure, which can be traced back to Wald (1947).
The superscripts in eqs. 5, 7 and 9 \((rdsu, rdcu \text{ and } rddft)\) index specific combinations of a decision-theoretic representation and a stochastic model: The prefix \(rd\) denotes a latent index based on the RDU representation, while the suffixes \(su\), \(cu\) and \(dft\) denote the three probabilistic models. Corresponding EU-based denotations are \(eusu\), \(eucu\) and \(eudft\). Call these specifications, and let \(spec\) stand for any one of them. The purpose of this study is to compare the “out-of-context” predictive power of these specifications. The next section describes the experiment that collects the data for this purpose. After that, we will return to comparing the specifications.

2. Experimental Design and Protocol

The subjects in this summer 2008 experiment were 80 University of Houston undergraduate students, recruited widely from registered students by means of a single undergraduate listserv email announcement. Each subject was individually scheduled for three separate sessions on three separate days of their own choosing, almost always finishing all three sessions within one week. Only one subject had to be replaced due to noncompletion of the three-day protocol. On each day, each subject made choices from the 100 choice pairs shown in Table 1, so that each made 300 choices in all by the end of their third day. On each day, the 100 choice pairs were split into two halves of 50 pairs each, separated by about ten to fifteen minutes of other tasks (demographic surveys, item response surveys, short tests of arithmetic and problem-solving ability, and so forth). Only rarely did any day’s session last more than an hour, and most sessions were substantially shorter than this. At the conclusion of each subject’s third day, one of their 300 choice pairs was selected at random (by means of the subject drawing a ticket from a bag) and the subject was paid according to their choice in that pair (this is called random task selection). If the subject’s choice in the selected pair was risky, the subject selected a six-sided
die from a box of six-sided dice (rolling them until satisfied if they wished), and their selected
die was then rolled by the attendant to determine the payment.

Here is the reasoning behind the protocol’s features. I want to estimate utilities and
weights without aggregation assumptions. Decision theories are about individuals, not
aggregates, and aggregation mutilates and destroys many observable properties of decision
theories (Wilcox 2008). A large amount of choice data from each subject is needed to estimate
utilities and weights with any precision at the individual level. A subject will become bored, and
will become careless, if she makes hundreds of decisions at one sitting. So the decisions are
divided up across three days, and on each day into two parts separated by unrelated tasks
providing a break from decisions. The separation across three days, in particular, introduces a
risk that some substantial event altering a subjects’ wealth or background risks will occur
between days, which could arguably undermine the assumption that utilities of outcomes and
hence choice probabilities are stationary throughout the protocol. This is a risk I am willing to
run to mitigate subject boredom with hundreds of choice tasks, and I can check whether
distributions of risky choice proportions across subjects appear to be stationary across subjects’
three days of decisions. Figure 3 shows these distributions. Although the first day distribution
appears to have slightly less dispersion, no parametric or nonparametric test finds any significant
difference between these three daily distributions. The within-subject difference between risky
choice proportions on the first and third day has a zero mean by all one-sample tests. Finally, the
first principal component of risky choice proportions on the three days accounts for 95% of their
collective variance, with no remotely significant second principal component associated with any
particular day. There is no sign of nonstationarity of choice probabilities across the three days, at
least in these comparisons of mean risky choice proportions.
Random task selection is meant to result in truthful, motivated and unbiased revelation of preferences in each pair: That is, subjects should make each of their 300 choices as if it was the only choice being made, for real, and there should be no portfolio or wealth effects making choices interdependent across the tasks. Both the independence axiom of EUT and the “isolation effect” of prospect theory would imply this. To see this for EUT, notice that the independence axiom in its “unreduced compounds” form implies

\[(\text{risky with Prob } = 1/300; \text{ } Z \text{ with Prob } = 299/300)\]

\[\text{risky } \geq \text{ safe } \quad \text{if and only if } \quad \geq \]

\[(\text{safe with Prob } = 1/300; \text{ } Z \text{ with Prob } = 299/300)\]

…where \(Z\) is any other outcome or risk, including the “grand lottery” created by the subject’s other 299 choices over the course of this experiment. Therefore, if subjects’ preferences satisfy independence in this unreduced compounds form, random task selection should be incentive compatible. Direct evidence does suggest that preferences generally satisfy the independence axiom in its unreduced compounds form (Kahneman and Tversky 1979; Conlisk 1989). Moreover, empirical examinations of random task selection in binary lottery choice experiments find no systematic choices differences between tasks selected with relatively low or high probabilities (Wilcox 1993) nor between tasks presented singly or under random task selection (Starmer and Sugden 1991), at least for relatively simple tasks like the pairs used here.
Two competing issues surround the resolution of risky lottery outcomes. On the one hand, experimenters want random devices to be concrete, observable and credible. This is why we use dice, cards, bingo cages and so forth. We also want subjects to have good reason to believe these devices are not rigged against them: This is why subjects select a die from an offered box of dice (and, if they wish, after rolling several to “test” them). However, the experimenter rolls the selected die because subjects may believe they exercise control over the die (whether they truly can or not; see e.g. Langer 1982). Here, the protocol compromises between the desire for credibility of randomizing devices and the possibility of subject beliefs in control over the die.

The choice pairs in Table 1 are organized into groups of four tasks (the rows of the table) by their shared outcome context. All risky lotteries are chances $q$ and $1-q$ (in sixths, generated by a six-sided die) of receiving the high and low outcomes $h$ and $l$ on the context, respectively: Four values of $q$ shown in each row in Table 1 ($q_a$, $q_b$, $q_c$ and $q_d$) create four risky lotteries on each context, and each of these is paired with safe (the middle outcome $m$ of the context with certainty) to create four pairs on the context. There are twenty-five distinct contexts, all constructed from nine positive money outcomes ($40$ to $120$ in $10$ increments).

Multiple outcome contexts serve several purposes. Nonparametric identification of all utilities and weights is impossible unless the same events (the die rolls) are matched with a multiplicity of outcomes on different contexts. My nonparametric ambitions here are high: In principle, I want to be able to estimate the utilities and weights without functional form assumptions. Multiple contexts improves the separate identification of utilities and weights. Additionally, the major difference between the context-free strong utility or SU model and the context-dependent DFT and CU models is how the latter models depend on $u(h) - u(l)$ through the $D(risky, safe)$ function. Therefore, the design deliberately creates a wide variety of contexts.
so that \( u(h) - u(l) \) is expected to vary widely across contexts: For instance, monotonicity of utilities in outcomes \( z \) implies that \( u(h) - u(l) \) must be greater on the context (40,60,110) than on the context (50,70,100) (these are contexts 19 and 20 in Table 1, respectively). This kind of variation is key to distinguishing between the probabilistic models.

Note also that Table 1 divides the 25 contexts (and hence the 100 pairs) into two disjoint sets of contexts—the “in” set of 16 contexts (64 pairs) on the left, and the “out” set of 9 contexts (36 pairs) on the right. For prediction comparisons, I estimate specifications using each subject’s choices from the in choice pairs (this is 64 pairs, for 192 observations over three days), and use these estimates to predict each subject’s choices from the “out” pairs (this is 36 pairs, for 108 choices over three days). The prediction is, therefore, an “out-of-context” prediction since the contexts of the in pairs and the out pairs are wholly distinct.

Finally, the choice of “sixths” as the “probability unit” for constructing risks serves several purposes. First, the six-sided die is perhaps the most culturally familiar randomizing device: This reduces some of the artificiality of laboratory risks. Second, sixths are well-suited to revealing a widely-believed shape of weighting functions. Figure 5 shows Prelec’s (1998) single-parameter weighting function 

\[
\gamma(q) = \exp(-[-\ln(q)])^{\gamma} \quad \forall q \in (0,1), \quad w(0)=0 \text{ and } w(1)=1, \quad \text{at various values of } \gamma \text{ from 0.5 to 1, covering widely-held priors about the shape of the function.}
\]

The linear function (heavy black line) is EU with \( \gamma = 1 \). Figure 5 shows that the maximum downward deflection of the nonlinear versions (from linearity) occurs very close to \( q = 5/6 \); and at \( q = 1/6 \) the upward deflection of nonlinear versions is about 75% of its maximum (which generally occurs at a somewhat smaller \( q \)). Finally, Monte Carlo simulations suggested that relatively coarse probability grids (fourths or sixths) over many different contexts permits relatively more precise estimation of utilities and weights.
3. Estimation

To discuss the estimation, it is helpful to define indices for pairs, trials (days) and subjects, as well as some important sets of indices:

\( i = 1, 2, \ldots, I \), indexing \( I \) distinct pairs. Here \( I = 100 \).

Pairs \( i \) are then \( \{ (h_i, q_i, l_i), m_i \} \), or \( \{ \text{risky}_i, \text{safe}_i \} \); and also note that

\( i = 1 \) to 64 are the \text{in} pairs, and \( i = 65 \) to 100 are the \text{out} pairs, in Table 1.

\( t = 1, 2, \ldots, T \), indexing \( T \) distinct trials (days) of each pair. Here \( T = 3 \) (the three days).

\( s = 1, 2, \ldots, S \), indexing the \( S \) distinct subjects. Here \( S = 80 \).

\( it \): A double subscript indicating the \( t \)th trial of choice pair \( i \).

\( r_{it}^s = 1 \) if subject \( s \) chose \text{risky}_i in her \( t \)th trial of pair \( i \), and zero otherwise.

\( r_{set}^s = (r_{it}^s \mid it \in \text{set}) \), the observed choice vector of subject \( s \) over those pairs and trials in \( \text{set} \). The \( \text{set} \) will be \text{in}, \text{out} or \text{all}, where \text{in} = \{ \text{it} \mid i \leq 64 \} (the estimation pairs/trials), \text{out} = \{ \text{it} \mid i \geq 65 \} (the prediction pairs/trials) and \text{all} = \{ \text{all it} \} is all 300 pairs/trials.

Let \( u^s(z) \) and \( w^s(q) \) denote utilities of outcomes \( z \) and weights associated with probabilities \( q \), respectively, of subject \( s \). The experiment involves nine distinct outcomes \( z \in \{ \$40, \$50, \ldots, \$110, \$120 \} \) across its 100 choice pairs, so there are nine utilities \([u'(40), u'(50), \ldots, u'(110), u'(120)]\) for each subject \( s \).

Because of the affine transformation invariance property of RDU and EU utilities, we can arbitrarily choose \( u^s(40) = 0 \) and \( u^s(50) = 1 \) for all subjects \( s \). So the unique utility vector \( u^s \) of each subject \( s \) is the utilities of the seven remaining outcomes,
\[ u^s = [u^s(60), u^s(70), \ldots, u^s(110), u^s(120)]. \]

The nonparametric treatment of utility makes each of those seven utilities a separate parameter to be estimated. I also examine a 2-parameter parametric alternative, the expo-power function (Saha 1993, Holt and Laury 2002) which blends the CARA and CRRA utilities in a flexible way,

\[ u^s(z) = -\exp(-\delta^s z^{1-\rho^s}) \] normalized so that \( u^s(40) = 0 \) and \( u^s(50) = 1. \)

The experiment also involves five distinct probabilities \( q \in [1/6, 2/6, 3/6, 4/6, 5/6] \), so there is a vector \( w^s \) of five weights to be estimated for each subject,

\[ w^s = [w^s(1/6), w^s(2/6), w^s(3/6), w^s(4/6), w^s(5/6)]. \]

The nonparametric approach again makes each of those five weights a separate parameter to be estimated. There are several parametric weighting functions in the literature such as the Prelec (1998) one-parameter function discussed above, but none are very flexible. Instead, I use the Beta distribution’s c.d.f. (with two parameters \( \alpha \) and \( \beta \)) as my parametric alternative, that is

\[ w(q|\alpha, \beta) = \frac{\int_0^q x^{\alpha-1}(1-x)^{\beta-1} dx}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx}. \]

It is the right kind of function (taking the unit interval onto itself, and monotone increasing) from the viewpoint of the general RDU representation theorem. It is much more flexible and (importantly) can take all major shapes suggested by the theorists who developed RDU and/or Cumulative Prospect Theory. It is also easily called in most nonlinear optimization software.

To summarize, the nonparametric latent index of the RDU representation, for subject \( s \) and pair \( i \), is

\[ \Delta RDU_{s,i}(u^s, w^s) = w^s(q_i)u^s(h_i) + [1 - w^s(q_i)]u^s(l_i) - u^s(m_i) \], where

\[ w^s = [w^s(1/6), w^s(2/6), w^s(3/6), w^s(4/6), w^s(5/6)], \]

\[ u^s = [u^s(60), u^s(70), \ldots, u^s(110), u^s(120)], \] and \( u^s(40) = 0 \) and \( u^s(50) = 1 \) for all \( s \).
Combining 10 with 5, 7 and 9, and choosing the logistic c.d.f. as \( F(x) \), we have the following choice probability specifications:

\[
\begin{align*}
(11) \quad P_{i}^{\text{spec}}(u^s, w^s, \lambda^s) &= \Lambda \left( \lambda^s \Delta RDU_i(u^s, w^s) \right), \\
(12) \quad P_{i}^{\text{du}}(u^s, w^s, \lambda^s) &= \Lambda \left( \lambda^s \frac{\Delta RDU_i(u^s, w^s)}{D_{i}^{\text{du}}(u^s)} \right), \quad \text{where} \quad D_{i}^{\text{du}}(u^s) = u^s(h_i) - u^s(l_i).
\end{align*}
\]

(13) \quad \lambda^s \frac{\Delta RDU_i(u^s, w^s)}{D_{i}^{\text{dii}}(u^s, w^s)}

Corresponding EU-based choice probabilities with either the strong utility or contextual utility model simply omit the vector of weights \( w^s \) from the function arguments and, in the case of DFT, set \( w^s(q_i) = q_i \) in the denominator expression.

Equations 10-13 define the probability of the event \( r^s_{it} = 1 \) (subject \( s \) chose risky in the \( t \)th trial of pair \( i \)). Letting \( P_{i}^{\text{spec}}(u^s, w^s, \lambda^s) \) generically denote any of those probabilities, the generic log likelihood of \( r^s_{it} \), given \( (u^s, w^s, \lambda^s) \), is

\[
\ell^s(\lambda^s \mid u^s, w^s, \lambda^s) = r^s_{it} \ln[P_{i}^{\text{spec}}(u^s, w^s, \lambda^s)] + (1 - r^s_{it}) \ln[1 - P_{i}^{\text{spec}}(u^s, w^s, \lambda^s)].
\]

Therefore, the total log likelihood over any particular set \( \text{in, out or all} \), for subject \( s \), is

\[
\ell^s(\lambda^s \mid u^s, w^s, \lambda^s) = \sum_{t \in \text{set}} \left[ r^s_{it} \ln[P_{i}^{\text{spec}}(u^s, w^s, \lambda^s)] + (1 - r^s_{it}) \ln[1 - P_{i}^{\text{spec}}(u^s, w^s, \lambda^s)] \right],
\]

and estimation of \( \theta^s = (u^s, w^s, \lambda^s) \) by maximum likelihood, for each subject, is straightforward.
Beginning with Manski (1975), econometricians made substantial progress eliminating the need for choosing a specific c.d.f. for \( F(x) \) in latent index models (e.g. Cosslett 1983; Klein and Spady 1993; Lewbel 2000). To my knowledge, none of these innovative methods can work with the fully nonparametric latent index of RDU (as in eq. 10). In essence, all “regressors” in eq. 10 are dummy variables (or interactions of dummy variables) indicating the presence or absence of particular outcomes and events in any pair \( i \). As is well-known in this literature, dispensing with knowledge of \( F(x) \) requires a so-called “special regressor” with several necessary properties: In particular, the special regressor must have an absolutely continuous distribution. I do not believe dummy regressors and their interactions, as in eq. 10, have that property, so I don’t believe I can use these innovative methods here. Moreover, in DFT, the logistic distribution arises as the limiting distribution of the sequential sampling process as the time interval between samples gets small, and this is a maintained assumption of DFT (Busemeyer and Townsend 1993). In other words, the choice of the logistic c.d.f as \( F(x) \) has a good theoretical motivation, at least from the perspective of one of the probabilistic models.

There is some interest in estimating \( \theta^* = (u^*, w^*, \lambda^*) \) using all of the data. This is particularly true for the fully nonparametric latent index models, since these provide a first look at function-free utilities and weights estimated simultaneously from discrete choice data. (I do not think this has been done before.) But for the purpose of comparing the probabilistic models, I prefer to compare their “out-of-context” prediction quality. To do that, I estimate \( \theta^* \) using just the \( in \) data, and use this estimate to predict choices on the \( out \) data. Let \( r_{in}^s \) and \( r_{out}^s \) denote the \( in \) and \( out \) choice vectors for each of the 80 subjects \( s \). Estimate each specification using just the data \( r_{in}^s \), producing pairs of estimated parameter vectors \( \hat{\theta}_{in}^{spec1,s} \) and \( \hat{\theta}_{in}^{spec2,s} \) for any two
specifications \textit{spec}1 and \textit{spec}2, for each \( s \). Then using these estimates, the \textit{out} set of data \( \mathbf{r}^s_{\text{out}} \), and eq. 15, we can calculate log likelihoods of the predictions based on the \textit{in} set estimates,

\[
\ell_{\text{out}}^{\text{spec}1}(s) = \ell_{\text{out}}^{\text{spec}1}(\mathbf{r}^s_{\text{out}} | \hat{\boldsymbol{\theta}}_{\text{in}}^{\text{spec}1,s}), \quad \text{and} \quad \ell_{\text{out}}^{\text{spec}2}(s) = \ell_{\text{out}}^{\text{spec}2}(\mathbf{r}^s_{\text{out}} | \hat{\boldsymbol{\theta}}_{\text{in}}^{\text{spec}2,s}).
\]

I use Vuong’s (1989) test to compare these prediction log likelihoods for pairs of specifications \textit{spec}1 and \textit{spec}2. It allows non-nested specifications and neither needs to be “correct” (the test is against the null “The specifications are equally close to the true DGP”). Intuitively, the test treats differences between the log likelihoods of two specifications as normal variates and does a \( z \) test on them. Define

\[
X^s_{\text{out}}(\text{spec}1,\text{spec}2) = \ell_{\text{out}}^{\text{spec}1}(s) - \ell_{\text{out}}^{\text{spec}2}(s),
\]

\[
\overline{X}_{\text{out}}(\text{spec}1,\text{spec}2) = \sum_{s=1}^{80} X^s_{\text{out}}(\text{spec}1,\text{spec}2)/80, \quad \text{and}
\]

\[
SD_{\text{out}}(\text{spec}1,\text{spec}2) = \sqrt{\sum_{s=1}^{80} (X^s_{\text{out}}(\text{spec}1,\text{spec}2) - \overline{X}_{\text{out}}(\text{spec}1,\text{spec}2))^2}/80.\]

Vuong’s test statistic is then

\[
z_{\text{out}}(\text{spec}1,\text{spec}2) = \frac{\overline{X}_{\text{out}}(\text{spec}1,\text{spec}2)}{(SD_{\text{out}}(\text{spec}1,\text{spec}2)/\sqrt{80})}, \quad \text{asymptotically } N(0,1) \text{ under the null.}
\]

\section*{4. Results}

Table 2 shows the seven representations (and treatments of utilities and/or weights) that I combine with the three probabilistic models to generate twenty-one specifications. All seven
representations of the latent index are nested within eq. 10, the fully nonparametric RDU latent index. By restricting eq. 10 to have either linear weights, linear utilities or both, we get (respectively) expected utility or EU, Yaari’s (1987) Dual Theory or “Yaari,” and expected value. RDU, EU and Yaari may be further restricted by replacing the nonparametric treatment of their utilities and/or weights by the 2-parameter functions discussed earlier (the expo-power function for utilities and/or the Beta c.d.f. for weights).

Figures 6 and 7 graph the 80 weight and utility functions (respectively) estimated using all of the data, the fully nonparametric RDU latent index in eq. 10, and the eq. 12 (CU) probability model. Figure 7 shows that the vast majority of utility functions are uniformly concave with perhaps a handful of exceptions. However, Figure 6 makes it clear that there is great heterogeneity of weight function shapes, so both Figures 6 and 7 employ a common color-coding for subjects with salient weight function shapes. Comparing Figures 5 and 6, it is curious but true that very few of the estimated weight functions appear to have the characteristic inverse-s shape posited by many theorists. In Figure 6, this expected shape is coded green: Although 14 of 80 weight functions have this general shape, most of those 14 cross the diagonal at a relatively high \( q \) above one-half, which is not true (for instance) of the one-parameter Prelec (1998) function shown in Figure 5 (this function always crosses the diagonal at \( e^{-1} \approx 0.37 \)).

The plurality of estimated weight functions (30 of the 80, coded blue in Figure 6) are uniformly concave and above the identity weights of EU, which Quiggin (1993) calls optimism. A very small number (4 of 80, coded red in Figure 6) are uniformly convex and below identity weights, which Quiggin calls pessimism. Finally, the second-most-common estimated weight function (26 of 80, coded orange in Figure 6) is s-shaped. One interpretation is that these people
tend to “round” low probabilities to zero and high probabilities to unity, so one might call these people “approximators.”

Comparison of Figures 6 and 7 suggests an odd relationship: The blue-coded optimists (concave weight functions) also tend to have the greatest concavity of their utility functions. This sounds counterintuitive since more concave utility means greater risk aversion while more concave weights (optimism) means less risk aversion. Part of the answer to this puzzle is undoubtedly poor identification for relatively risk-averse subjects. As a subject becomes increasingly risk-averse overall (for whatever reason), there will be less variation in her choices and separate identification of weights and utilities will be relatively poor. For such subjects, a positive finite sample correlation between estimated utility and weights concavity will be expected. Monte Carlo simulations of information matrices of relatively risk-averse RDU/CU data-generating processes confirm this econometric intuition. Therefore I suggest that this apparent finding be taken with a grain of salt.

Figure 8 displays the prediction comparisons between specifications in a graphical manner. The figure displays a linearly transformed version of average prediction log likelihoods,

(18) \[ \text{“Percent prediction metric”} = \frac{\ell_{\text{out}}^{\text{frontier}} - \ell_{\text{out}}^{\text{spec}}}{\ell_{\text{out}}^{\text{frontier}}} \text{, where } \ell_{\text{out}}^{\text{spec}} = \sum_s \ell_{\text{out}}^{\text{spec}}(s) / 80. \]

The specification \textit{evsu} is an expected value latent index with the SU probability model: This model has just one estimated parameter for each subject (the scale parameter \( \lambda \)) and I treat this as an upper prediction benchmark. The “specification” \textit{frontier} is not a model at all: This is simply the best possible out-of-context prediction likelihood with stationary choice probabilities. Since the experimental design involves repeated trials, and since there is some inconsistency of subject
choices across repeated trials, this best possible likelihood is nonzero for each subject, and so is its average across subjects. No model with stationary probabilities can predict better than this. If this metric takes the value 100%, the specification predicts no better than a strong utility expected value model. If this metric takes the value 0%, the specification predicts as well as any possible specification with stationary probabilities can predict.

Figure 8 shows that specifications with 2-parameter parametric treatments of utilities and/or weights, near the center of the figure, perform best in out-of-context prediction. This is not surprising since these specifications involve many fewer parameters than the fully nonparametric specifications. Yet even the fully nonparametric specifications at the left almost always predict noticeably better than the maximally lean EV (expected value) specifications at the far right (the sole exception here being the nonparametric Yaari representation with DFT). There are two other important findings. First, notice that holding representations constant, the context-dependent probability models CU and DFT are almost uniformly better at prediction than the context-free strong utility or SU model. Second, examine Figure 8’s comparative results for EU and RDU representations with parametric utilities and weights (near the center of Figure 8, labeled “EU 3 parms para” and “RDU 5 parms para”). Consider the 3-parameter EU specification with strong utility as a baseline. Notice that the improvement in prediction quality associated with changing from strong utility to CU or DFT (about 12% in both cases) exceeds the improvement in prediction quality associated with adding two parameters and changing to the 5-parameter RDU specification but staying with strong utility (about 9%). This confirms Wilcox’s (2008,2009) results using the well-known Hey and Orme (1994) data set: When we ask “what matters for prediction?” it seems that the greatest marginal gains from a strong utility and
EU starting point accrue from changing the probabilistic model rather than adopting a more expansive representation such as RDU.

Tables 3 (A, B, C and D) show the results of the Vuong (1989) tests that compare the probabilistic models. Each of these tables holds the representation constant and just compares changes in probabilistic models. Tables 3-A, 3-B, 3-C and 3-D show results for expected value, expected utility, Yaari’s dual model and rank-dependent utility, respectively. The tables all have left and right panels with in-context and out-of-context comparisons of log likelihoods. Additionally, except for table 3-A (expected value representation), the tables all have top and bottom panels showing results for the parametric and nonparametric versions of each representation. With the exception of Table 3-C (the Yaari dual model representations), the Vuong tests overwhelmingly and uniformly reject strong utility in favor of both of the context-dependent probability models. Even in Table 3-C (the Yaari representation), where strong utility is occasionally directionally better than the context-dependent models, it is only significantly better in a single comparison against DFT. The contextual utility probability model is always significantly better than strong utility, regardless of the representation or parametric expansiveness of the specifications.

The Vuong tests do not distinguish DFT and contextual utility in any consistent and persuasive way. Among the seven in-context comparisons between them, each is significantly best in three comparisons (with one insignificant comparison), and among the seven out-of-context comparisons, contextual utility significantly beats DFT in three comparisons and DFT significantly beats contextual utility in one comparison (with three insignificant comparisons).
5. Conclusions

For the purpose of predicting risky choices in new contexts, context-dependent models like contextual utility (Wilcox 2009) and decision field theory (Busemeyer and Townsend 1993) are overwhelmingly better than the traditional strong utility approach that dominates applied theoretical work in experimental and behavioral economics and much applied econometric work as well. Moreover, the greatest marginal gains in predictive success do not come from more expansive decision-theoretic representations that add parameters to be estimated: Instead they come from switching from context-free to context-dependent models, with no additional parameters to be estimated.

[Conclusions to be fleshed out more in next draft]
References


Loomes, G., 2005, Modeling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. Experimental Economics 8, 301-323.


Luce, R. D. and P. Suppes, 1965, Preference, utility and subjective probability, in R. D. Luce, R.
R. Bush and E. Galanter, (Eds.), Handbook of mathematical psychology Vol. III. Wiley, New
York, pp. 249-410.

Economic Perspectives 1, 121-154.

Manski, C., 1975, Maximum score estimation of the stochastic utility model of choice, Journal of
Econometrics 3, 205-228.

McKelvey, R. and T. Palfrey, 1995, Quantal response equilibria for normal form games. Games
and Economic Behavior 10, 6-38.

Economy 59, 371-404.


3, 323-343.

MS: Kluwer.

American Journal of Agricultural Economics 75, 905-913.


investigation of the common ratio effect. Journal of Risk and Uncertainty 2, 159-78.


Table 1: The 100 Choice Pairs—the “in set” for estimation, and the “out set” for prediction.

The “in” contexts of 64 pairs—4 pairs on each of 16 contexts (for estimation)

<table>
<thead>
<tr>
<th>#</th>
<th>(l,m,h)</th>
<th>qa</th>
<th>qb</th>
<th>qc</th>
<th>qd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(40,50,80)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>2</td>
<td>(40,50,90)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>3</td>
<td>(40,60,100)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>4</td>
<td>(40,60,120)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>5</td>
<td>(50,60,90)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>6</td>
<td>(50,70,110)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>7</td>
<td>(50,70,120)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>8</td>
<td>(60,70,90)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>9</td>
<td>(60,80,120)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>10</td>
<td>(70,80,100)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>11</td>
<td>(70,80,110)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>12</td>
<td>(70,80,120)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>13</td>
<td>(80,90,100)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>14</td>
<td>(80,90,110)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>15</td>
<td>(90,100,110)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>16</td>
<td>(100,110,120)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

The “out” contexts of 36 pairs—4 pairs on each of 9 contexts (use estimates to predict here)

<table>
<thead>
<tr>
<th>#</th>
<th>(l,m,h)</th>
<th>qa</th>
<th>qb</th>
<th>qc</th>
<th>qd</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>(40,50,60)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>18</td>
<td>(40,50,70)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>19</td>
<td>(40,60,110)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>20</td>
<td>(50,70,100)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>21</td>
<td>(60,80,110)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>22</td>
<td>(70,90,110)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>23</td>
<td>(80,90,120)</td>
<td>5/6</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
</tr>
<tr>
<td>24</td>
<td>(80,100,120)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
<tr>
<td>25</td>
<td>(90,100,120)</td>
<td>4/6</td>
<td>3/6</td>
<td>2/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>
Table 2. The seven representations estimated, including the treatment of utility and/or weigh functions and the resulting number of parameters estimated. Each of these is combined with one of the three probability models to produce twenty-one specifications in all, and each specification involves the estimation of one extra parameter, the scale parameter $\lambda$.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Utility function treatment</th>
<th>Weight function treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Value (EV)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Expected Utility (EU)</td>
<td>nonparametric (7 parms.)</td>
<td>–</td>
</tr>
<tr>
<td>Expected Utility (EU)</td>
<td>expo-power (2 parms.)</td>
<td>–</td>
</tr>
<tr>
<td>Yaari Dual Theory (Yaari)</td>
<td>–</td>
<td>nonparametric (5 parms.)</td>
</tr>
<tr>
<td>Yaari Dual Theory (Yaari)</td>
<td>–</td>
<td>beta c.d.f. (2 parms.)</td>
</tr>
<tr>
<td>Rank-Dependent Utility (RDU)</td>
<td>nonparametric (7 parms.)</td>
<td>nonparametric (5 parms.)</td>
</tr>
<tr>
<td>Rank-Dependent Utility (RDU)</td>
<td>expo-power (2 parms.)</td>
<td>beta c.d.f. (2 parms.)</td>
</tr>
</tbody>
</table>
Table 3-A. Vuong comparisons between the specifications: Expected value representation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Likelihood on the <em>in</em> data set, using the <em>in</em> parameter estimates</td>
<td>Log Likelihood on the <em>out</em> data set, using the <em>in</em> parameter estimates</td>
<td></td>
</tr>
<tr>
<td><strong>Contextual Utility</strong></td>
<td>Decision Field Theory</td>
<td>Strong Utility</td>
</tr>
<tr>
<td></td>
<td>$z = -4.18$</td>
<td>$z = 5.86$</td>
</tr>
<tr>
<td></td>
<td>$p = 4.2 \times 10^{-5}$</td>
<td>$p = 2.7 \times 10^{-9}$</td>
</tr>
<tr>
<td><strong>Decision Field Theory</strong></td>
<td>—</td>
<td>$z = 6.14$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p = 4.2 \times 10^{-10}$</td>
</tr>
</tbody>
</table>

Notes: All comparisons are based on parameters estimated from the 192 observations of the *in* data. The in-context fit comparisons compare the maximized log likelihoods of those 192 observations. The out-of-context fit comparisons use the parameter estimates to calculate log likelihoods in the *out* data, and compare those log likelihoods. In each cell, Vuong’s $z$ statistic is based on the difference between the row and column probability models’ log likelihood, so negative $z$ indicates that the column model fits best while positive $z$ indicates that the row model fits best.
Table 3-B. Vuong comparison between the specifications: Expected utility representation.

<table>
<thead>
<tr>
<th>Representation: Expected Utility (expo-power utilities, 3 parameters)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>“Explanatory performance” (in-context fit)</td>
<td>“Predictive performance” (out-of-context fit)</td>
<td></td>
</tr>
<tr>
<td>Log Likelihood on the in data set, using the in parameter estimates</td>
<td>Log Likelihood on the out data set, using the in parameter estimates</td>
<td></td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>Strong Utility</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>Contextual Utility</td>
<td>z = 2.00</td>
<td>z = 0.20</td>
</tr>
<tr>
<td>p = 0.023</td>
<td>p = 8.8x10^-10</td>
<td>p = 0.42</td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>z = 5.86</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>p = 2.3x10^-9</td>
<td>—</td>
<td>z = 5.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Representation: Expected Utility (nonparametric utilities, 8 parameters)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>“Explanatory performance” (in-context fit)</td>
<td>“Predictive performance” (out-of-context fit)</td>
<td></td>
</tr>
<tr>
<td>Log Likelihood on the in data set, using the in parameter estimates</td>
<td>Log Likelihood on the out data set, using the in parameter estimates</td>
<td></td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>Strong Utility</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>Contextual Utility</td>
<td>z = 1.99</td>
<td>z = 0.20</td>
</tr>
<tr>
<td>p = 0.02</td>
<td>p = 8.8x10^-10</td>
<td>p = 0.42</td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>z = 5.86</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>p = 2.3x10^-9</td>
<td>—</td>
<td>z = 5.73</td>
</tr>
</tbody>
</table>

Notes: All comparisons are based on parameters estimated from the 192 observations of the in data. The in-context fit comparisons compare the maximized log likelihoods of those 192 observations. The out-of-context fit comparisons use the parameter estimates to calculate log likelihoods in the out data, and compare those log likelihoods. In each cell, Vuong’s z statistic is based on the difference between the row and column probability models’ log likelihood, so negative z indicates that the column model fits best while positive z indicates that the row model fits best.
Table 3-C. Vuong comparison between the specifications: Yaari’s “dual theory” representation.

<table>
<thead>
<tr>
<th>Representation: Yaari (beta c.d.f weights, 3 parameters)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>“Explanatory performance” (in-context fit)</td>
<td>“Predictive performance” (out-of-context fit)</td>
<td></td>
</tr>
<tr>
<td>Log Likelihood on the <em>in</em> data set, using the <em>in</em> parameter estimates</td>
<td>Log Likelihood on the <em>out</em> data set, using the <em>in</em> parameter estimates</td>
<td></td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>Strong Utility</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>Contextual Utility</td>
<td>( z = 0.07 ) ( p = 0.47 )</td>
<td>( z = -0.65 ) ( p = 0.26 )</td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>( z = -0.66 ) ( p = 0.25 )</td>
<td>Strong Utility</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Representation: Yaari (nonparametric weights, 6 parameters)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>“Explanatory performance” (in-context fit)</td>
<td>“Predictive performance” (out-of-context fit)</td>
<td></td>
</tr>
<tr>
<td>Log Likelihood on the <em>in</em> data set, using the <em>in</em> parameter estimates</td>
<td>Log Likelihood on the <em>out</em> data set, using the <em>in</em> parameter estimates</td>
<td></td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>Strong Utility</td>
<td>Decision Field Theory</td>
</tr>
<tr>
<td>Contextual Utility</td>
<td>( z = -7.84 ) ( p = 2.3\times10^{-15} )</td>
<td>( z = -1.26 ) ( p = 0.10 )</td>
</tr>
<tr>
<td>Decision Field Theory</td>
<td>( z = 4.95 ) ( p = 3.7\times10^{-7} )</td>
<td>Strong Utility</td>
</tr>
</tbody>
</table>

Notes: All comparisons are based on parameters estimated from the 192 observations of the *in* data. The in-context fit comparisons compare the maximized log likelihoods of those 192 observations. The out-of-context fit comparisons use the parameter estimates to calculate log likelihoods in the *out* data, and compare those log likelihoods. In each cell, Vuong’s \( z \) statistic is based on the difference between the row and column probability models’ log likelihood, so negative \( z \) indicates that the column model fits best while positive \( z \) indicates that the row model fits best.
Table 3-D. Vuong comparison between specifications: Rank-dependent utility representation.

| Representation: Rank-Dependent Utility (expo-power utility, beta c.d.f. weights, 5 parameters) |  |  |
| “Explanatory performance” (in-context fit) | “Predictive performance” (out-of-context fit) |  |
| Log Likelihood on the \( in \) data set, using the \( in \) parameter estimates | Log Likelihood on the \( out \) data set, using the \( in \) parameter estimates |  |
| Decision Field Theory | Strong Utility | Decision Field Theory | Strong Utility |
| Contextual Utility | \( z = 4.63 \) \( p = 1.8 \times 10^{-6} \) | \( z = 6.63 \) \( p = 1.6 \times 10^{-11} \) | Contextual Utility | \( z = -0.13 \) \( p = 0.45 \) | \( z = 6.35 \) \( p = 1.0 \times 10^{-10} \) |
| Decision Field Theory | — | \( z = 4.17 \) \( p = 1.5 \times 10^{-5} \) | Decision Field Theory | — | \( z = 5.54 \) \( p = 1.5 \times 10^{-8} \) |

| Representation: Rank-Dependent Utility (nonparametric utilities and weights, 13 parameters) |  |  |
| “Explanatory performance” (in-context fit) | “Predictive performance” (out-of-context fit) |  |
| Log Likelihood on the \( in \) data set, using the \( in \) parameter estimates | Log Likelihood on the \( out \) data set, using the \( in \) parameter estimates |  |
| Decision Field Theory | Strong Utility | Decision Field Theory | Strong Utility |
| Contextual Utility | \( z = -2.85 \) \( p = 0.0022 \) | \( z = 5.82 \) \( p = 3.0 \times 10^{-9} \) | Contextual Utility | \( z = 1.78 \) \( p = 0.037 \) | \( z = 4.32 \) \( p = 7.9 \times 10^{-6} \) |
| Decision Field Theory | — | \( z = 6.37 \) \( p = 9.6 \times 10^{-11} \) | Decision Field Theory | — | \( z = 3.05 \) \( p = 0.0011 \) |

Notes: All comparisons are based on parameters estimated from the 192 observations of the \( in \) data. The in-context fit comparisons compare the maximized log likelihoods of those 192 observations. The out-of-context fit comparisons use the parameter estimates to calculate log likelihoods in the \( out \) data, and compare those log likelihoods. In each cell, Vuong’s \( z \) statistic is based on the difference between the row and column probability models’ log likelihood, so negative \( z \) indicates that the column model fits best while positive \( z \) indicates that the row model fits best.
Figure 1. An example pair, as displayed to subjects. The pair’s “context” in this example is (40,50,90) (in U.S. dollars).

Generally, **risky** is \((h,q,l)\), where \(h > l\), \(q = \Pr(h)\) and \(1-q = \Pr(l)\).

Here, \(h = $90\), \(q = \frac{1}{6}\) and \(l = $40\).

Generally, **safe** is \(m\) with Prob 1, where \(h > m > l\).

Here \(m = $50\).
Figure 2: Example pair in Machina-Marschak triangle representing all pairs on its context

\[ 1 - p_h - p_l = \text{probability of receiving} \]

\[ p_h = \text{probability of receiving high outcome 90} \]

\[ p_l = \text{probability of receiving low outcome 40} \]
Figure 3. Cumulative distributions of risky choice proportions across days

Locations of cumulative distributions of within-subject differences across pairs of days are not significantly different from zero by sign, signed-rank or t-tests; and these distributions pass all tests of normality.

The day 1 distribution appears to have slightly less variance across subjects than the days 2 and 3 distributions.
Figure 4. Cumulative distributions of switching rates between days

- Cumulative percent of subjects, days 1 to 2
- Cumulative percent of subjects, days 2 to 3
- Cumulative percent, days 1 to 2 minus days 2 to 3

Consistency improves as days pass.

Mean and median = 0.17 across days 1 and 2
Mean and median = 0.13 across days 2 and 3

Cumulative distribution of within-subject differences is significantly positive by sign, signed-rank and t-test at \( p < 0.0001 \). Consistency improves as days pass.
Figure 5: Prelec-type one-parameter weighting functions at 1/6 and 5/6 for widely-held priors ($\gamma = 1$ to 0.5)
Figure 6: 80 individually estimated probability weighting functions.

- **Blue** = 30 “Optimists” (above identity line).
- **Red** = 4 “Pessimists” (below identity line).
- **Green** = 14 “Prospect Theorists” (initially above, then below identity line).
- **Orange** = 26 “Approximators” (initially below, then above identity line).
- **Yellow** = 6 “Others” (crosses identity line more than once).
Figure 7: 80 individually estimated utility functions, color-coded to indicate weighting function type in Figure 6.

Blue = 30 “Optimists” (above identity line).

Red = 4 “Pessimists” (below identity line).

Green = 14 “Prospect Theorists” (initially above, then below identity line).

Orange = 26 “Approximators” (initially below, then above identity line).

Yellow = 6 “Others” (crosses identity line more than once).
Figure 8. Out-of-context percent prediction metric of various specifications: Average of individual estimation-prediction results

Structure (utility function treatment, weighting function treatment)