2015

Error and Generalization in Discrete Choice Under Risk

Nathaniel T. Wilcox
Chapman University, nwilcox@chapman.edu

Error and Generalization in Discrete Choice Under Risk

Comments

Working Paper 15-11

Error and Generalization in Discrete Choice Under Risk

by

Nathaniel T. Wilcox*

Abstract

I compare the generalization ability, or out-of-sample predictive success, of four probabilistic models of binary discrete choice under risk. One model is the conventional homoscedastic latent index model—the simple logit—that is common in applied econometrics: This model is "context-free" in the sense that its error part is homoscedastic with respect to decision sets. The other three models are also latent index models but their error part is heteroscedastic with respect to decision sets: In that sense they 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 2011), the "decision field theory" model of Busemeyer and Townsend (1993) and the "Blavatskyy-Fishburn" model (Fishburn 1978; Blavatskyy 2014). In a new experiment, all three context-dependent models outperform the context-free model in prediction, and significantly outperform a linear probability model (suggested by contemporary applied practice a la Angrist and Pischke 2009) when the latent preference structure is rank-dependent utility (Quiggin 1982). All of this holds true for function-free estimations of outcome utilities and probability weights as well as parametric estimations. Preoccupation with theories of the deterministic structure of choice under risk, to the exclusion of theories of error, is a mistake.

JEL Classification Codes: C25, C91, D81

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

May 2015.

*Economic Science Institute (Chapman University) and Center for the Economic Analysis of Risk (Georgia State University). Phone 714-628-7212, fax 714-628-2881, email nwilcox@chapman.edu. I have benefitted from conversations with Pavlo Blavatskyy, Jerome Busemeyer, John Hey, Chew Soo Hong, Jim Cox, Glenn Harrison, Stefan Hoderlein, Graham Loomes, Mark Machina, John Quiggin, Michel Regenwetter and Joerg Stoye. I thank Stacey Joldersma for her excellent research assistance.

Beginning with Mosteller and Nogee (1951), dozens of experiments on discrete choice under risk suggest that these choices have a strong probabilistic component. These experiments involve repeated trials of choice from pairs of risky options, and reveal high rates of choice switching by the same subject between trials of the same pair.[1] In some cases, the repeated trials span days (e.g. Tversky 1969; Hey and Orme 1994; Hey 2001) and one might worry that decision-relevant conditions 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—the functional or "representation" that describes how lottery characteristics (outcomes, events and their likelihoods) are combined to represent binary preference directions. Econometrically, that discussion concerns the functional form taken by the nonrandom part of the latent index in a conventional discrete choice model. However, there is renewed interest in the random part of decision under risk, driven both by theoretical questions and empirical findings. Sometimes, an anomaly (say, an apparent violation of expected utility or EU theory) can be attributed to probabilistic models rather than the structure in question (Wilcox 2008). The 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 not others (see also Blavatskyy 2006). Loomes (2005), Gul and Pesendorfer (2006) and Blavatskyy (2009) are just three relatively recent (but very different) examples of this renewed interest.

[1] For instance, Camerer (1989, p. 81) reported that "Overall, 31.6% of the subjects reversed preference [between a test and retest of the same lottery pair]. This number is distressingly close to…random, but comparable with numbers in other studies (e.g. Starmer and Sugden 1989)…"

I compare four probabilistic models of choice under risk. One of the models is the conventional homoscedastic latent index model—the simple logit—that was long a staple of applied econometrics: This model is "context-free" in the sense that its random error part is homoscedastic with respect to decision sets. The other three models are also latent index models, but their error part is heteroscedastic with respect to decision sets (though none require estimation of any new parameters), and in that sense these models are "context-dependent." Context dependence arises from several different theoretical perspectives. I consider my own "contextual utility" model (Wilcox 2011), the "decision field theory" model of Busemeyer and Townsend (1993) and the "Blavatskyy-Fishburn" model (Fishburn 1978; Blavatskyy 2014). A new experiment is performed on 80 subjects. Two-thirds of the data is used to estimate models for each individual, and these estimates predict the remaining third of choices. All the context-dependent models strongly outperform the context-free logit in prediction and, additionally, strongly outperform a simple linear probability model suggested by contemporary applied practice (a la Angrist and Pischke 2009) when the latent preference structure is rank-dependent utility (Quiggin 1982). My results strongly suggest that wholesale preoccupation with the deterministic structure of choice under risk, to the exclusion of theories of error, is a serious scientific mistake with widespread implications for applied theory and empirical applications.

In the literature on semiparametric estimation of discrete choice models, Monte Carlo evidence reveals the importance of heteroscedastic latent index errors (Manski and Thompson 1986; Horowitz 1992): Here, models that incorrectly impose a homoscedastic form can lead to highly biased estimation. Therefore we might expect that probabilistic models containing some of the truth of heteroscedastic error will predict discrete decisions much better than a homoscedastic misspecification. The heteroscedastic models I consider here all emerge from

some reasonable theoretical objection to the homoscedastic model, including concerns about violations of stochastic dominance, proper representation of comparative risk aversion and computational logic. Perhaps one or more of them catch some of the truth of decision error.

1. Preliminaries

In the experiment, each choice pair is a set of two options $\{risky, safe\} \equiv \{(h, q, l), m\}$. 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 $\langle l, m, h \rangle$ the context of each pair. Figure 1 shows an example pair where $\{risky, safe\}$ is $\{(90,1/6,40), 50\}$ and the context of the pair is $\langle 40,50,90 \rangle$.

I consider a class of probabilistic choice models of the form

$$(1) \qquad P \equiv Prob(risky\ chosen\ from\ \{risky, safe\}) = F\left(\lambda \frac{V(risky) - V(safe)}{D(risky, safe)}\right).$$

where $V(risky) - V(safe)$ is a decision-theoretic representation of the difference between the values of the options *risky* and *safe*, such as expected utility or rank-dependent utility, λ is a scale (or inverse standard deviation) parameter, $D(risky, safe)$ adjusts the scale parameter in heteroscedastic models, and $F: R \to [0,1]$ is an increasing function with $F(0) = 0.5$ and $F(x) = 1 - F(-x)$.

While my focus is on the function $D(risky, safe)$, first I consider the "value difference" $\Delta V = V(risky) - V(safe)$. The function *V* needs to be a decision-theoretic representation of lottery value with theoretical breadth and empirical strength. Rank-dependent utility or RDU, originally developed by Quiggin (1982), fits this bill. Under RDU, the values of two-outcome options like *risky*, and single outcome options like *safe*, are

(2) $V(risky) = w(q)u(h) + [1 - w(q)]u(l)$ and $V(safe) = u(m)$, where

$u(z)$ is the *utility* of outcome z; and

$w(q)$ is the *weight* associated with probability q of receiving outcome h in *risky*.

The RDU value difference between *risky* and *safe* in a pair is thus

(3) $\Delta RDU = w(q)u(h) + [1 - w(q)]u(l) - u(m)$.

RDU nests the expected utility or EU representation: EU is just that special case of RDU where $w(q) \equiv q$. Therefore I develop all choice models below in terms of ΔRDU. To convert those into EU-based models, just replace $w(q)$ by q in 3 to get

(4) $\Delta EU = qu(h) + (1 - q)u(l) - u(m)$, 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 outcomes below some reference point (put differently, CPT posits loss aversion), and my experiment pairs contain only large positive outcomes of $40 to $120.[2] 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. But if the mapping between events and outcome ranks is constant across all risks—as it is in this experiment—then SEU is indistinguishable from RDU.[3]

[2] It is possible that some subjects would have a reference point shaped by the payoff range of the experiment itself, in which case my claim here might be unjustified. However, my function-free estimations of utilities and weights will permit an *s*-shaped loss-averse array of outcome utilities around any reference point (including one interior to my outcome set)—as is the case with Prospect Theory—if the fitting of the choice data demands it.

[3] 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 *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 *risky*, can always be thought of as the subjective probability of the event "the die roll is k+1 or k+2 or…6". The states and outcome ranks are identically ordered across all option pairs (that is, the risky options are all comonotonic—see

This implies that the RDU representation in eq. 2 will be reasonably 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 since my interest lies with the scale adjustment $D(risky, safe)$. By experimental design, the RDU representation of $\Delta V = V(risky) - V(safe)$ will encompass this wide set of decision-theoretic representations, so inferences concerning $D(risky, safe)$ will hold for this set of decision-theoretic representations in this domain of option pairs.

2. The Probabilistic Models

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 it as the homoscedastic latent index model. It imposes the restriction $D(risky, safe) \equiv 1$ on eq. 1, and with RDU it is

(5) $$P^{rdsu} = Prob(risky) = F(\lambda \Delta RDU).$$

As is well-known (Luce 1959), if we let $F(x)$ be $\Lambda(x) = [1 + exp(-x)]^{-1}$, the logistic c.d.f. , this is equivalent to a binary logit:

(6) $$P^{rdsu} = \frac{exp[\lambda V(risky)]}{exp[\lambda V(risky)] + exp[\lambda V(safe)]},$$

with $V(risky)$ and $V(safe)$ as given in eq. 2. McFadden and others developed economic theory and application of this model and it appears widely in experimental and 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 in part for that reason, so that my results speak clearly to these applications.

Quiggin 1993), so rank-dependent weighting and subjective probability become indistinguishable. This feature is also necessary for applying Decision Field Theory to the RDU representation.

The contextual utility or CU model (Wilcox 2011) sets $D(risky, safe) \equiv u(h) - u(l)$, and with RDU it is

(7) $$P^{rdcu} = Prob(risky) = F\left(\frac{\lambda \Delta RDU}{u(h)-u(l)}\right).$$

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 functional and probabilistic definitions of comparative risk aversion. To see this, consider any of my pairs on a 3-outcome context. Under RDU and contextual utility, the choice probability in eq. 7 can be rewritten as

(8) $P^{rdcu} = F(\lambda[-v(l,m,h) + w(q)])$, where $v(l,m,h) = [u(m) - u(l)]/[u(h) - u(l)]$.

This probability is decreasing in the ratio of differences $v(l,m,h)$. Consider two subjects Anne and Bob with identical weighting functions (this includes the case where both have EU preferences) and identical scale parameters λ, and assume that Bob is globally more risk averse than Anne in Pratt's sense (Bob's local absolute risk aversion $-u''(z)/u'(z)$ exceeds that of Anne for all z). These assumptions and simple algebra based on Pratt's (1964) main theorem imply that $v^{Bob}(l,m,h) > v^{Anne}(l,m,h)$ on all contexts, and as a result (8) implies that Bob will have a lower probability than Anne of choosing *risky* on all contexts. Strong utility cannot share this property, and this was the primary motivation for the contextual utility model.

The third model is decision field theory or DFT (Busemeyer and Townsend 1992, 1993), one of the earliest "diffusion" models[4] of preferential choice (see Rangel 2009). It sets $D(risky, safe) \equiv [u(h) - u(l)]\sqrt{w(q)[1 - w(q)]}$, and with RDU it is

[4] The word "diffusion" appears in the text of Busemeyer and Townsend (1992) thirty times, and one of its keywords is "diffusion models." Decision field theory is most definitely a diffusion model.

(9) $$P^{rddft} = Prob(risky) = F\left(\frac{\lambda \Delta RDU}{[u(h)-u(l)]\sqrt{w(q)[1-w(q)]}}\right).$$

Note that eq. 9 is DFT only for pairs like those found in this experiment (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-motivated 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. The CU model does not share this property.

Busemeyer and Townsend (1992, 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). Busemeyer and Townsend also show that, as the sampling rate gets large, the function *F* will be the logistic c.d.f.—another reason I employ the logistic c.d.f. throughout this work.

The final model is called stronger utility by its author Blavatskyy (2014), but here I call it the BF model to avoid confusing it with strong utility. The BF model begins with a definition of two important benchmark options. Let $(risky \vee safe) = (h, q, m)$ and $(risky \wedge safe) = (m, q, l)$: These two options are the least upper bound and greatest lower bound, respectively, on both *risky* and *safe* in terms of stochastic dominance.[5] Then in the BF model, $D(risky, safe) = RDU(risky \vee safe) - RDU(risky \wedge safe)$, and

$$(10) \quad P^{rdbf} = Prob(risky) = H_\lambda\left(\frac{\Delta RDU}{RDU(risky \vee safe) - RDU(risky \wedge safe)}\right),$$

where $H_\lambda: [-1,1] \rightarrow [0,1]$ and otherwise has the same properties as F, and λ is again a scale parameter. The BF model is a general approach to constructing probabilistic models of risky choice that will respect stochastic dominance: That is, the model always attaches a zero probability to choice of stochastically dominated options. As mentioned above, the CU model does not do so.[6]

Although the H_λ function in the BF model differs from the F in the general class I defined earlier in eq. 1, a suitable choice of H_λ converts the BF model into the following form that uses the logistic c.d.f. (see Appendix I):

$$(11) \quad P^{rdbf} = Prob(risky) = \Lambda\left[\lambda \ln\left(\frac{w(q)[u(h) - u(m)]}{[1 - w(q)][u(m) - u(l)]}\right)\right].$$

Thus, all four models may be estimated using a common function F which, as mentioned above, will be the logistic c.d.f. throughout my estimations. When $w(q) \equiv q$ and we have an EU representation, the form taken by eq. 11 is an instance of Fishburn's (1978) incremental EU

[5] That is, $(risky \vee safe)$ stochastically dominates both *risky* and *safe*, but is itself stochastically dominated by every other option that stochastically dominates both *risky* and *safe*. Similarly, *risky* and *safe* both stochastically dominate $(risky \wedge safe)$, and every other option stochastically dominated by both *risky* and *safe* is itself stochastically dominated by $(risky \wedge safe)$.

[6] In the experiment reported here, there are no option pairs in which one option stochastically dominates the other. In Wilcox (2008) I provide a simple method for dealing with stochastic dominance in option pairs.

advantage model (see Appendix I). This is why I call this the BF ("Blavatskyy-Fishburn") model, as the form of eq. 11 is (with an EU representation) consistent with both Blavatskyy's and Fishburn's models.

The specifications denoted by the superscripts on P in eqs. 5, 7, 9 and 11 (*rdsu*, *rdcu*, *rddft* and *rdbf*) are specific combinations of a decision-theoretic representation (the prefix *rd* denotes the RDU representation) and a probabilistic model (denoted by the suffixes *su*, *cu*, *dft* and *bf*). Let *spec* stand for any specification. The purpose of the experiment described in the next section is to compare the generalization ability, or out-of-context prediction success, of these specifications as well as EU-based versions of them.

There are other ways to introduce probabilistic choice into models of decision under risk. One of these is random preferences (Loomes and Sugden 1995; Gul and Pesendorfer 2006): This approach treats vectors of outcome utilities and/or probability weights as random draws from a fixed distribution of these vectors. Random preference models also exhibit context dependence (Wilcox 2011, p. 101). Elsewhere I have shown that the generalization ability of the contextual utility model outperforms that of a random preference model (Wilcox 2008, 2011) as well as other models (including strong utility) using the data set of Hey and Orme (1994). There is, however, a difficult problem with considering a random preference RDU specification in this study, where the data contains 25 distinct outcome contexts: It is very difficult to generalize an RDU random preference specification across more than three outcome contexts without changing estimation techniques in fundamental ways (Wilcox 2008 pp. 252-256; Wilcox 2011 pp.101-102). I converted Blavatskyy's (2014) model into a form using the logistic c.d.f. not just to reveal its kinship with Fishburn's (1978) model, but also to control that parametric estimation

element, making it a common feature of the competing models. This simply cannot be done with a random preference RDU model across multiple outcome contexts.

3. Experimental Design and Protocol

The subjects in this experiment were 80 undergraduate students at a large urban university, recruited widely from registered students by means of a single email announcement to all undergraduates. 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, for each subject, the 100 choice pairs were randomly ordered 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 reliably estimate utilities and weights 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 2 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. There is no sign of nonstationarity of choice probabilities across the three days.

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 (i.e. "compound independence") implies

$$risky \succcurlyeq safe \quad \text{if and only if} \quad \begin{array}{c}(risky \text{ with Prob} = 1/300;\ Z \text{ with Prob} = 299/300) \\ \succcurlyeq \\ (safe \text{ with Prob} = 1/300;\ Z \text{ with Prob} = 299/300)\end{array}$$

…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. Some evidence suggests that preferences generally satisfy the independence axiom in its unreduced compounds form (Kahneman and Tversky 1979; Conlisk 1989), and older direct examinations of random task selection in binary lottery choice experiments found no systematic choice 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. There is renewed controversy on this point (Cox, Sadiraj and Schmidt 2014; Harrison and Swarthout 2014), but random task selection has been the standard experimental mechanism for a few decades.

Two competing issues surround the resolution of risky lottery outcomes. On the one hand experimenters want random devices to be concrete, observable and credible: I use a six-sided die for this reason. 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. I much prefer to carry off the comparison between probabilistic models without using functional form assumptions about the decision-theoretic entities (the utilities of outcomes and the probability weights). Therefore I want to be able to estimate the utilities and weights in the function-free manner Hey and Orme (1994) pioneered for utilities and as Blavatskyy (2013) did for utilities and weights.[7] Monte Carlo simulations showed that function-free identification utilities, weights and scale parameters is greatly improved when the same events (the die rolls) are matched with many different outcomes on different contexts. Additionally, the major feature of the context-dependent CU, DFT and BF models is how their choice probabilities vary with context. Therefore, the design contains a wide variety of contexts as shown in Table 1.

Finally, the choice of "sixths" as the "probability unit" for constructing risks serves several purposes. First, the six-sided die is perhaps the most familiar of all randomizing devices: This reduces some of the artificiality of laboratory risks. Second, sixths are well-suited to a widely-believed shape of weighting functions. Figure 5 shows Prelec's (1998) single-parameter weighting function $w(q|\gamma) = \exp(-[-\ln(q)]^{\gamma})\ \forall\ q \in (0,1)$, $w(0)=0$ and $w(1)=1$, at various values of γ 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

[7] Gonzalez and Wu (1999) also did this but not with binary choice data, as Blavatskyy did and as I do here. Gonzalez and Wu elicited and used option certainty equivalents as the dependent measure. There are long-standing doubts as to whether the elicitation of certainty equivalents produces the same weak order as binary choices, starting with the long literature on preference reversals (see Butler and Loomes 2007 for a relatively recent review). Additionally, Gonzalez and Wu's "choice list" methodology for eliciting certainty equivalents is not without critics. See for instance Cohen et al. (1987) and Loomes and Pogrebna (2014).

deflection (from linearity) of the nonlinear versions 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 contexts permits relatively more precise estimation of utilities and weights than a design with a finer probability grid and fewer contexts.

4. Estimation and Prediction

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

i = 1,2,…I, indexing I distinct pairs. Here I = 100.

Pairs i are then $\{(h_i, q_i, l_i), m_i\}$, or $\{risky_i, safe_i\}$; and also note that

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

s = 1,2,…S, indexing the S distinct subjects. Here S = 80.

it: A double subscript indicating the tth trial of pair i.

$r_{it}^s = 1$ if subject s chose $risky_i$ in her tth trial of pair i, and zero otherwise.

$\mathbf{r}_{set(k)}^s = (r_{it}^s | it \in set(k))$, the observed choice vector of subject s over the pairs and trials in some $set(k)$, which is either $in(k)$ or $out(k)$, where k = 1,2,…,10 indexes ten partitions of the 100 pairs into two sets—an $in(k)$ set for estimation, and an $out(k)$ set for prediction.

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 $z \in \{\$40, \$50, \ldots, \$120\}$ across its 100 choice pairs, but because of the affine transformation invariance property of RDU and EU utilities, we can choose $u^s(40) = 0$ and $u^s(120) = 1$ for

all subjects *s*. With this done, the unique estimable utility vector $\mathbf{u}^s$ for each subject *s* is the utilities of the seven remaining outcomes, $\mathbf{u}^s = \langle u^s(50), u^s(60), \ldots, u^s(110)\rangle$. The function-free estimations make each of those seven utilities a separate parameter to be estimated. I also examine a parametric alternative with one parameter ρ^s, the CRRA utility of money given by $u^s(z|\rho^s) = z^{1-\rho^s}/(1-\rho^s)$, normalized so $u^s(40) = 0$ and $u^s(120) = 1$.[8]

The experiment also involves five distinct probabilities $q \in \left\{\frac{1}{6}, \frac{2}{6}, \ldots, \frac{5}{6}\right\}$, so there is a vector $\mathbf{w}^s = \langle w^s\left(\frac{1}{6}\right), w^s\left(\frac{2}{6}\right), \ldots, w^s\left(\frac{5}{6}\right)\rangle$ of five weights to be estimated for each subject. The function-free estimations make each of those five weights a separate parameter to be estimated. For a parametric alternative I use Prelec's (1998) two-parameter function , given by $w^s(q|\beta^s, \gamma^s) = \exp\left(-\beta^s[-\ln(q)]^{\gamma^s}\right)$ $\forall$ $q \in (0,1)$, $w(0) = 0$ and $w(1) = 1$.

To summarize, the function-free latent index of the RDU representation, for subject *s* and pair *i*, is

(12) $\Delta RDU_i(\mathbf{u}^s, \mathbf{w}^s) = w^s(q_i)u^s(h_i) + [1 - w^s(q_i)]u^s(l_i) - u^s(m_i)$, where

$\mathbf{w}^s = \langle w^s\left(\frac{1}{6}\right), w^s\left(\frac{2}{6}\right), \ldots, w^s\left(\frac{5}{6}\right)\rangle$, and

$\mathbf{u}^s = \langle u^s(50), u^s(60), \ldots, u^s(110)\rangle$, with $u^s(40) = 0$ and $u^s(120) = 1$ $\forall$ *s*.

Combine eq. 12 with eqs. 5, 7, 9 and 11, let $\boldsymbol{\theta}^s \equiv (\mathbf{u}^s, \mathbf{w}^s, \lambda^s)$ and choose the logistic c.d.f. as *F*(*x*), and we have the following choice probability specifications:

(13) $P_i^{rdsu}(\boldsymbol{\theta}^s) = \Lambda[\lambda^s \Delta RDU_i(\mathbf{u}^s, \mathbf{w}^s)]$;

(14) $P_i^{rdcu}(\boldsymbol{\theta}^s) = \Lambda\left[\lambda^s \frac{\Delta RDU_i(\mathbf{u}^s, \mathbf{w}^s)}{u^s(h_i) - u^s(l_i)}\right]$;

(15) $P_i^{rddft}(\boldsymbol{\theta}^s) = \Lambda\left[\lambda^s \frac{\Delta RDU_i(\mathbf{u}^s, \mathbf{w}^s)}{[u^s(h_i) - u^s(l_i)]\sqrt{w^s(q_i)[1 - w^s(q_i)]}}\right]$; and

[8] This normalized version of CRRA utility is simply $u^s(z|\rho^s) = \left(z^{1-\rho^s} - 40^{1-\rho^s}\right)/\left(120^{1-\rho^s} - 40^{1-\rho^s}\right)$.

(16) $P_i^{rdbf}(\boldsymbol{\theta}^s) = \Lambda\left[\lambda^s \ln\left(\frac{w^s(q_i)[u^s(h_i)-u^s(m_i)]}{[1-w^s(q_i)][u^s(m_i)-u^s(l_i)]}\right)\right]$.

Corresponding EU-based choice probabilities simply omit the vector of weights $\mathbf{w}^s$ from function arguments and set $w^s(q_i) = q_i$ everywhere else.

Equations 13-16 define the probability of the event $r_{it}^s = 1$ (subject *s* chose *risky* in the *t*th trial of pair *i*). Letting $P_i^{spec}(\boldsymbol{\theta}^s)$ denote any of those probabilities, the log likelihood of r_{it}^s is

(17) $\ell^{spec}(r_{it}^s|\boldsymbol{\theta}^s) = r_{it}^s \ln\left[P_i^{spec}(\boldsymbol{\theta}^s)\right] + (1 - r_{it}^s)\ln\left[1 - P_i^{spec}(\boldsymbol{\theta}^s)\right]$;

the total log likelihood over any particular $set(k)$, for subject *s*, is

(18) $\mathcal{L}^{spec}\left(\mathbf{r}_{set(k)}^s \middle| \boldsymbol{\theta}^s\right) = \sum_{it \in set(k)} \ell^{spec}(r_{it}^s|\boldsymbol{\theta}^s)$;

and estimation of $\boldsymbol{\theta}^s$ by maximum likelihood, for each subject *s*, is performed using just the $in(k)$ choice vector $\mathbf{r}_{in(k)}^s$. Let $\widehat{\boldsymbol{\theta}}_{in(k)}^{spec,s}$ be the estimated parameter vector for any specification, for any *s*, using just the $in(k)$ data. Then using the estimate, the $out(k)$ choice vector $\mathbf{r}_{out(k)}^s$, and eq. 18, calculate average prediction log likelihoods (across the ten partitions *k* of the data), and their difference for any two specifications, as

(19) $\mathcal{L}_{pred}^{spec,s} = \frac{1}{10}\sum_k \sum_{it \in out(k)} \ell^{spec}\left(r_{it}^s \middle| \widehat{\boldsymbol{\theta}}_{in(k)}^{spec,s}\right)$,

and let $X_{pred}^s(spec1, spec2) \equiv \mathcal{L}_{pred}^{spec1,s} - \mathcal{L}_{pred}^{spec2,s}$.

$X_{pred}^s(spec1, spec2)$ is the average difference between the prediction log likelihoods of any two specifications, for subject *s*. If I was willing to assume that the 300 choices of subject *s* are independent trials, I might now talk about the statistical significance of X_{pred}^s for each subject *s*, using either the asymptotic approach of Vuong (1989) or Clarke's (2007) finite sample nonparametric variation on Voung's method. Instead, I will remain agnostic about statistical independence within each subject, and regard the probability models as marginal probabilities

across each subject's sequence of trials. This leads to conservatively treating each value of X^s_{pred} as a single observation coming from each subject *s*. I can inspect the distribution of X^s_{pred} across subjects and ask whether or not this distribution is significantly positive using a sign test—indicating whether or not *spec*1 is better than *spec*2 for the majority of my subjects.

Statistical significance is nice, but not everything: A sensible approach to judging the magnitudes of any improvements will round out the picture of results. For this purpose, I need two benchmark log likelihoods: A lower benchmark $\mathcal{L}^{low,s}_{pred}$ that any good specification ought to beat, and an upper benchmark $\mathcal{L}^{high,s}_{pred}$ that no specification could outperform. From these and the actual $\mathcal{L}^{spec,s}_{pred}$ of some specification, form a ratio of differences measure of prediction quality

$$(20) \qquad \bar{Y}^{spec}_{pred} = \frac{\sum_s \mathcal{L}^{spec,s}_{pred} - \sum_s \mathcal{L}^{low,s}_{pred}}{\sum_s \mathcal{L}^{high,s}_{pred} - \sum_s \mathcal{L}^{low,s}_{pred}}.$$

Because there are three trials of every option pair, there is a natural choice for the upper benchmark: The log likelihood of the observed choice proportions in the $out(k)$ choice vector. Letting $\bar{r}^s_i = \sum_t r^s_{it}/3$ be this observed choice proportion of subject *s* for pair *i*, this is

$$(21) \qquad \mathcal{L}^{high,s}_{pred} = \frac{1}{10}\sum_k \sum_{it \in out(k)} r^s_{it} \ln(\bar{r}^s_i) + (1 - r^s_{it}) \ln(1 - \bar{r}^s_i).$$

No model of fixed marginal choice probabilities can do better than this log likelihood.

The lower benchmark could be based on each subject's mean choice of $risky_i$ across all choice pairs, in which case the measure of prediction quality would resemble a pseudo R^2; but this seems like an uninformative straw man to me. In my view the lower benchmark ought to be a good atheoretical model of some sort. I take a few pages from Angrist and Pischke (2009) for this purpose, simply estimating a linear probability model or LPM of choices whose right-hand-side dummy regressors code all characteristics of option pairs in a plausible fashion. Regard each

option pair i as consisting of four dimensions: (i) m_i, the sure outcome of $safe_i$; (ii) $h_i - m_i$, the potential upside of $risky_i$; (iii) $m_i - l_i$, the potential downside of $risky_i$; and (iv) q_i, the probability of the upside of $risky_i$. Then the mostly harmless or MH specification is

$$(22)\ P_i^{mh}(\boldsymbol{\beta}^s) = \sum_{j=50,60,\ldots,110} a_j^s \cdot 1(m_i = j) + \sum_{j=10,20,\ldots,60} b_j^s \cdot 1(h_i - m_i = j) + \sum_{j=10,20} c_j^s \cdot 1(m_i - l_i = j) + \sum_{j=1/6,2/6,\ldots,5/6} d_j^s \cdot 1(q_i = j),$$

where $\boldsymbol{\beta}^s = (\mathbf{a}^s, \mathbf{b}^s, \mathbf{c}^s, \mathbf{d}^s)$ is a vector of parameters to be estimated. Though there appear to be twenty parameters, there are actually seventeen since, for estimability, the restrictions $\sum_j b_j^s = 0$, $c_{10}^s + c_{20}^s = 0$ and $\sum_j d_j^s = 0$ must be imposed. I estimate this by ordinary least squares for each subject, again staying close to Angrist and Pischke. With estimated values $\widehat{\boldsymbol{\beta}}_{in(k)}^s$ in hand for each partition *k*, it is straightforward to calculate

$$(23)\ \mathcal{L}_{pred}^{low,s} = \frac{1}{10}\sum_k \sum_{it \in out(k)} r_{it}^s \ln\left[P_i^{mh}\left(\widehat{\boldsymbol{\beta}}_{in(k)}^s\right)\right] + (1 - r_{it}^s)\ln\left[1 - P_i^{mh}\left(\widehat{\boldsymbol{\beta}}_{in(k)}^s\right)\right].$$

At first blush the MH specification may seem like a straw man. In fact the prediction log likelihoods of the MH specification improve on those of a mean-variance model—a result well-known from the history of psychological decision research (Payne 1973)—and my results suggest that it improves on most of the expected utility specifications (though significantly so only in the case of the strong utility model). So it is no straw man. Moreover, we seem to have entered a period where one regularly sees linear probability models used in applied microeconomics in preference to latent index models. Applied microeconomists may well feel

that a linear probability model will perform as well as (or better than) any of the context-dependent models examined here. We will see.

Finally, recall that $\mathcal{L}_{pred}^{spec,s}$ is an <u>average</u> prediction log likelihood across ten partitions k = 1,2,…,10 of the data into an $in(k)$ set for estimation and an $out(k)$ set for prediction. The partitions are not randomly drawn: They are highly constrained. Some of the specifications above, particularly the MH linear probability model, would not always be estimable if these partitions were constructed randomly, and identification of the other models would usually suffer too. The partitions are constructed subject to constraints that guarantee estimability and promote better identification. Moreover, all of the context-dependent models attach special importance to variation in context: To capture this, both the $in(k)$ and $out(k)$ sets are constructed to contain wide variation of contexts.

Each of the ten partitions of pairs are also a partition of contexts: The $in(k)$ set always contains sixteen of the twenty-five contexts in Table 1, while the $out(k)$ set always contains the remaining nine contexts. Thus the prediction task is always "out of context" since the estimation and prediction pairs are disjoint with respect to contexts. In all, there are always 64 pairs (and 192 choices across the three days) within the estimation data $in(k)$, and 36 pairs (and 108 choices across the three days) within the prediction data $out(k)$. Appendix II further discusses the construction of the ten partitions.

5. Results

The four panels of Figure 4 compare distributions of prediction log likelihoods. For this purpose, the prediction log likelihood of the MH specification is used as a baseline: The Figure 4 curves are empirical cumulative distributions (over subjects) of $X_{pred}^{s}(spec, mh)$, where *spec* is

any other specification. The four panels of Table 2 complement the Figure 4 panels, showing the number of the 80 subjects for whom $X^{s}_{pred}(spec1, spec2) > 0$ for all pairs of specifications, along with an associated 2-tailed p-value for a sign test against equal medians for each pair of specifications. Figure 5 in turn provides information about the magnitudes of any significant effects: It compares values of $\bar{Y}^{spec}_{pred}$ for the sixteen specifications formed by combining CU, DFT, BF or SU with any of the EU or RDU representations.

There are three main findings. First, differences between the three context-dependent heteroscedastic specifications CU, DFT and BF are, for the most part, small and inconsistent across the four decision-theoretic representations. The panels of Figure 4 show this visually, but the sign test results in Table 2 bring home the point: While CU significantly bests DFT and BF when using parametric EU or RDU specifications, BF significantly bests CU and DFT when using function-free EU, and DFT significantly bests CU when using function-free RDU. This mixed evidence does not order these context-dependent models in a convincing way.

The second finding is that the MH specification wins only in competitions with specifications using EU representations. This is easiest to see in Figure 5: There, almost all comparisons of MH against specifications using EU representations show that the latter actually predict a bit worse than the MH specification. Yet whenever a context-dependent model is combined with an RDU representation, it improves strongly on the MH specification (see panels a and b of Table 2).

The third finding concerns the truly dismal performance of the SU model. Figure 5 brings this finding home in convincing fashion. Recall that $\bar{Y}^{spec}_{pred} = 0$ is the performance of the MH specification. Figure 5 shows that $\bar{Y}^{eusu}_{pred}$ is strongly negative for both the parametric and function-free versions of EU: That is, a strong utility EU model has a prediction log likelihood

that is much worse than the atheoretic linear probability model. Even the RDU-based strong utility specifications do not significantly improve on the MH specification (see panels a and b of Table 2). Blavatskyy (2014) reports similar results using Hey's (2001) data for in-sample fit but did not explore prediction fit. My own earlier work (Wilcox 2008, 2011) finds the same failure of strong utility relative to contextual utility in prediction, using Hey and Orme's (1994) data.

Unlike in-sample log likelihoods, prediction log likelihoods do not need to be penalized for estimation degrees of freedom: Overfitting will automatically be punished in prediction. However, it is true that the linear probability model involves more independent parameters (seventeen) than the context-dependent models do (at most thirteen in the function-free RDU specifications). This might be the reason why the context-dependent RDU specifications outperform the MH specification. Pooled estimations (rather than the individual estimations considered so far) may make this concern less compelling since, in this case, data degrees of freedom far outnumber model degrees of freedom in all the estimation. Table 3 reports results of pooled estimations of the function-free RDU specifications and the MH specification. The three context-dependent specifications still strongly outperform the MH specification and the context-free SU specifications.

6. Conclusions

Binary choice under risk is an ubiquitous and frequently important situation in economic life, applied economic theory and empirical economics. Choices to pursue higher education or not, to have medical insurance or not, and to migrate or not are just three examples. In the case of binary choice under risk, the strong utility model—homoscedastic latent index models—should be regarded as mortally wounded. Theory tells us that strong utility models cannot respect

stochastic dominance (Falmagne 1985; Loomes and Sugden 1995), have no good computational interpretation (Busemeyer and Townsend 1993), and cannot coherently represent comparative risk aversion across agents in different choice contexts (Wilcox 2011). The laboratory evidence against strong utility for choice under risk is by now overwhelming (Loomes and Sugden 1998; Rieskamp 2008; Wilcox 2008, 2011; Butler, Isoni and Loomes 2012; Blavatskyy 2014).

There are constructive alternatives in place: The new heteroscedastic, context-dependent models of probabilistic choice can accommodate violations of the betweenness property (Blavatskyy 2006), the Allais phenomena (Blavatskyy 2007), and the preference reversal phenomenon (Blavatskyy 2009, 2014; Butler Isoni and Loomes 2012). They provide coherent models of probabilistic risk aversion across agents and decision situations (Wilcox 2011; Blavatskyy 2014), and they have solid grounding in cognitive science (Busemeyer and Townsend 1993). Here, I have added to this growing progressive and constructive research program by demonstrating the predictive superiority of the new context-dependent models of probabilistic discrete choice under risk.

Because it is convenient for IV estimation, the linear probability model will almost certainly be with us for some time (but see Lewbel, Dong and Yang 2012 for critique and alternatives). But if the application is a single equation analysis of binary choice under risk, my results show that an RDU-based representation combined with any of the three context-dependent probabilistic choice models strongly outperforms a linear probability model in prediction. Linear probability models have become very common in applied economics. I believe that in the case of choice under risk, there are now better ways to analyze the choice. Saying you tried a homoscedastic logit as well, and got similar results to the linear probability model, may soon fail to convince the audience.

References

Angrist, J. D. and J.-S. Pischke, 2009. Mostly Harmless Econometrics, Princeton, Princeton University Press.

Ballinger, T. P., and N. Wilcox, 1997, Decisions, error and heterogeneity. Economic Journal 107, 1090-1105.

Becker, G. M., M. H. DeGroot and J. Marschak, 1963a, Stochastic models of choice behavior. Behavioral Science 8, 41–55.

Becker, G. M., M. H. DeGroot and J. Marschak, 1963b, An experimental study of some stochastic models for wagers. Behavioral Science 8, 199-202.

Blavatskyy, P. R., 2006, Violations of betweenness or random errors? Economics Letters 91, 34-38.

Blavatskyy, P. R., 2007, Stochastic expected utility theory. Journal of Risk and Uncertainty 34, 259-286.

Blavatskyy, P. R., 2009, Preference reversals and probabilistic choice. Journal of Risk and Uncertainty 39, 237-250.

Blavatskyy, P. R., 2013, Which decision theory? Economics Letters 120, 40-44.

Blavatskyy, P. R., 2014, Stronger utility. Theory and Decision 76, 265-286.

Block, H. D. and J. Marschak, 1960, Random orderings and stochastic theories of responses, in I. Olkin et al., (Eds.), Contributions to probability and statistics: Essays in honor of Harold Hotelling. Stanford University Press, Stanford, pp. 97-132.

Busemeyer, J. and J. Townsend, 1992, Fundamental derivations from decision field theory. Mathematical Social Sciences 23, 255-282.

Busemeyer, J. and J. Townsend, 1993, Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review 100, 432-59.

Butler, D. and G. Loomes, 2007, Imprecision as an account of the preference reversal phenomenon. American Economic Review 97, 277-97.

Camerer, C., 1989, An experimental test of several generalized expected utility theories. Journal of Risk and Uncertainty 2, 61-104.

Camerer, C. and T-H. Ho, 1999, Experience weighted attraction learning in normal-form games. Econometrica 67:827-74.

Clarke, K. A., 2007, A simple distribution-free test for nonnested model selection. Political Analysis 15, 347-363.

Conlisk, J., 1989, Three variants on the Allais example. American Economic Review 79, 392-407.

Cox, J. C., V. Sadiraj and U. Schmidt, 2014, Paradoxes and mechanisms for choice under risk. Experimental Economics (forthcoming).

Debreu, G., 1958, Stochastic choice and cardinal utility. Econometrica 26, 440-444.

Falmagne, J.-C., 1985, Elements of Psychophysical Theory. Oxford: Oxford Univ. Press.

Fishburn, P., 1978, A probabilistic expected utility theory of risky binary choices. International Economic Review 19, 633-646.

Gonzalez, R. and G. Wu, 1999, On the shape of the probability weighting function. Cognitive Psychology 38, 129-166.

Gul, F., and W. Pesendorfer, 2006, Random expected utility. Econometrica 74, 121-146.

Harrison, G. W. and J. T. Swarthout, 2014, Experimental payment protocols and the bipolar behaviorist. Theory and Decision 77, 423-438.

Hey, J. D., 2001, Does repetition improve consistency? Experimental Economics 4, 5-54.

Hey, J. D. and C. Orme, 1994, Investigating parsimonious generalizations of expected utility theory using experimental data. Econometrica 62, 1291-1329.

Horowitz, J. L., 1992, A smoothed maximum score estimator for the binary response model. Econometrica 60, 505-531.

Kahneman, D. and A. Tversky, 1979, Prospect theory: An analysis of decision under risk. Econometrica 47, 263-291.

Langer, E. J., 1982, The illusion of control. In D. Kahneman, P. Slovic and A. Tversky, eds., Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press.

Lewbel, A., Y. Dong and T. T. Yang, 2012, Viewpoint: Comparing features of convenient estimators for binary choice models with endogenous regressors. Canadian Journal of Economics 45, 809-829.

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

Loomes, G. and G. Pogrebna, 2014, Measuring individual risk attitudes when preferences are imprecise. Economic Journal 124, 569-593.

Loomes, G. and R. Sugden, 1995, Incorporating a stochastic element into decision theories. European Economic Review 39, 641-648.

Loomes, G. and R. Sugden, 1998, Testing different stochastic specifications of risky choice. Economica 65, 581-598.

Luce, R. D., 1959. Individual Choice Behavior: A Theoretical Analysis. New York: Wiley.

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.

Manski, C. F. and T. S. Thompson, 1986, Operational characteristics of maximum score estimation. Journal of Econometrics 32, 65-108.

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

Mosteller, F. and P. Nogee, 1951, An experimental measurement of utility. Journal of Political Economy 59, 371-404.

Payne, J. W., 1973, Alternative approaches to decision making under risk: Moments versus risk dimensions. Psychological Bulletin 80, 439-453.

Pratt, J. W., 1964, Risk aversion in the small and in the large. Econometrica 32, 122-136.

Prelec, D., 1998, The probability weighting function. Econometrica 66, 497-527.

Quiggin, J, 1982, A theory of anticipated utility. Journal of Economic Behavior and Organization 3, 323-343.

Quiggin, J, 1993. Generalized Expected Utility Theory: The Rank-Dependent Model. Norwell, MS: Kluwer.

Rangel, A., 2009, The computation and comparison of value in goal-oriented choice. In P. Glimcher, C. Camerer, E. Fehr and R. Poldrack, eds., Neuroeconomics: Decision-Making and the Brain. London: Elsevier.

Rieskamp, J., 2008, The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory and Cognition 34, 1446-1465.

Savage, L. J., 1954. The Foundations of Statistics. New York: Wiley.

Starmer, C. and R. Sugden, 1989, Probability and juxtaposition effects: An experimental investigation of the common ratio effect. Journal of Risk and Uncertainty 2, 159-78.

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

Tversky, A., 1969, Intransitivity of preferences. Psychological Review 76, 31-48.

Tversky, A. and D. Kahneman, 1992, Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323.

Vuong, Q., 1989, Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.

Wald, A., 1947. Sequential Analysis. New York: Wiley.

Wilcox, N., 1993, Lottery choice: Incentives, complexity and decision time. Economic Journal 103, 1397-1417.

Wilcox, N., 2008, Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In J. C. Cox and G. W. Harrison, eds., Research in Experimental Economics Vol. 12: Risk Aversion in Experiments pp. 197-292. Bingley, UK: Emerald.

Wilcox, N., 2011, 'Stochastically more risk averse:' A contextual theory of stochastic discrete choice under risk. Journal of Econometrics 162, 89-104.

Appendix I: Equivalence of eqs. 10 and 11 given a suitable choice of the function H_λ.

Let $R = risky$ and $S = safe$. From eq. 10 and the definitions $U = (R \vee S) = (h, q, m)$ and $L = (R \wedge S) = (m, q, l)$ for the option pairs in this experiment, Blavatskyy's model is

(A1) $P^{rdbf} = Prob(R) = H_\lambda\left(\frac{V(R)-V(S)}{V(U)-V(L)}\right)$.

Choose $H_\lambda(x) = \Lambda\left[\lambda \ln\left(\frac{1+x}{1-x}\right)\right]$. For $x \in (-1,1)$, this has the needed properties $H_\lambda(0) = 0.5$ and $H_\lambda(x) = 1 - H_\lambda(-x)$. With $x = \frac{V(R)-V(S)}{V(U)-V(L)}$, we have

(A2) $\frac{1+x}{1-x} = \frac{1+\frac{V(R)-V(S)}{V(U)-V(L)}}{1-\frac{V(R)-V(S)}{V(U)-V(L)}} = \frac{V(U)-V(L)+V(R)-V(S)}{V(U)-V(L)+V(S)-V(R)} = \frac{[V(U)-V(S)]+[V(R)-V(L)]}{[V(U)-V(R)]+[V(S)-V(L)]}$.

Applying the RDU representation theorem to the four key options,

(A3) $V(R) = w(q)u(h) + [1 - w(q)]u(l)$, $V(S) = u(m)$,

$V(U) = w(q)u(h) + [1 - w(q)]u(m)$, and $V(L) = w(q)u(m) + [1 - w(q)]u(l)$.

Substitute these into the four bracketed terms at the right end of (A2) to get

(A4) $[V(U) - V(S)] = [w(q)u(h) + [1 - w(q)]u(m) - u(m)] = w(q)[u(h) - u(m)]$,

$[V(R) - V(L)] = [w(q)u(h) + [1 - w(q)]u(l) - w(q)u(m) - [1 - w(q)]u(l)] =$

$w(q)[u(h) - u(m)]$,

$[V(U) - V(R)] = [w(q)u(h) + [1 - w(q)]u(m) - w(q)u(h) - [1 - w(q)]u(l)] =$

$[1 - w(q)][u(m) - u(l)]$, and

$[V(S) - V(L)] = [u(m) - w(q)u(m) - [1 - w(q)]u(l)] =$

$[1 - w(q)][u(m) - u(l)]$.

Clearly $\frac{1+x}{1-x} = \frac{w(q)[u(h)-u(m)]}{[1-w(q)][u(h)-u(m)]}$, so the equivalence to eq. 11, given a suitable choice of H_λ, has been established. In turn, a bit of algebra on eq. 11 shows that it implies

(A5) $\frac{Prob(risky)}{1-Prob(risky)} = \left(\frac{w(q)[u(h)-u(m)]}{[1-w(q)][u(m)-u(l)]}\right)^{\lambda}$.

When we set $w(q) \equiv q$ so that the structural representation is expected utility, this expression is recognizable as an instance of the probabilistic form derived by Fishburn (1978, p. 635).

Appendix II: Estimation notes

All estimations were carried out in SAS 9.2 using the nonlinear programming procedure ("Proc NLP" in the SAS language) using the quasi-Newton algorithm. For function-free estimations all parameters bounded in the interval [0,1], that is utilities and weights, were constrained to lie in [0.0001,0.9999]; additionally, monotonicity was imposed on estimated utilities and weights. For parametric RDU estimations the parameters β^s and γ^s of the 2-parameter Prelec (1998) weighting function were constrained to the strictly positive reals—in practice the interval [0.0001,∞). No other constraints were imposed on any estimates.

Monte Carlo simulations showed that both finite sample biases of parameter estimates and prediction log likelihoods could be noticeably improved by penalizing estimation that produced fitted probabilities very close to zero or one. By a grid search across Monte Carlo simulations, the following piecewise quadratic penalty function $p_i(\boldsymbol{\theta}^s)$ was arrived at as a good kludge for penalizing such fitted probabilities:

$p_i(\boldsymbol{\theta}^s) = 0$ if $P_i^{spec}(\boldsymbol{\theta}^s) \in [0.001,0.999]$;

$p_i(\boldsymbol{\theta}^s) = -30 \cdot \left(1 - \left[P_i^{spec}(\boldsymbol{\theta}^s)/0.001\right]^2\right)$ if $P_i^{spec}(\boldsymbol{\theta}^s) < 0.001$; and

$p_i(\boldsymbol{\theta}^s) = -30 \cdot \left(1 - \left[1000 - 1000 \cdot P_i^{spec}(\boldsymbol{\theta}^s)\right]^2\right)$ if $P_i^{spec}(\boldsymbol{\theta}^s) > 0.999$.

This simply imposes a very steep but smoothly differentiable penalty on probabilities that wander within 0.001 of zero or one. The adjusted log likelihood function is

$$\mathcal{L}^{spec}\left(\mathbf{r}^{s}_{set(k)}\middle|\boldsymbol{\theta}^{s}\right) = \sum_{it\in set(k)} \ell^{spec}(r^{s}_{it}|\boldsymbol{\theta}^{s}) + \sum_{i} p_{i}(\boldsymbol{\theta}^{s})$$

This penalty $\sum_i p_i(\boldsymbol{\theta}^s)$ was imposed on all maximum likelihood estimations and (with a sign change) on the ordinary least squares estimation of the MH specification as well for purposes of comparability. Note that the penalty is imposed for all lottery pairs, including those in the $out(k)$ set (but the $out(k)$ data is not part of the penalty function), for estimation. However the penalty is not included in the prediction log likelihoods analyzed in the text.

For each subject, specification, and $in(k)$ data set, estimations were started from a moderate-sized grid of starting parameter vectors (six to sixty, depending on the dimensionality of the vector) to a "finalist" estimated vector from each starting vector, and the finalist with the best adjusted log likelihood was selected as the maximum likelihood estimate.

As mentioned in the text, the ten partitions of the data into $in(k)$ and $out(k)$ sets were constructed subject to several constraints. First, the MH specification must be estimable: For each of the MH specification's twenty indicator function regressors (see eq. 22) among the 64 pairs comprising any $in(k)$ set, the indicator must vary, taking some values of 1 and some values of 0. Among other requirements, this implies that every distinct value of m_i, the outcomes of $safe_i$ in pairs i, must occur at least once in any $in(k)$ set: This also aids the identification of utilities and the scale parameter λ^s in the other models.

Second, the context-dependent models differ from the context-free model mostly across pairs with different contexts: In particular, the utility difference $u^s(h_i) - u^s(l_i)$ is all or part of the D function of all three context-dependent models: This is transparent for the CU and DFT models in eqs. 7 and 9, and simple algebra shows that the D function of the BF model equals $0.5[u^s(h_i) - u^s(l_i)]$ at q_i such that $w^s(q_i) = 0.5$. Variation of $u^s(h_i) - u^s(l_i)$ across pairs also strengthens identification of the scale parameter λ^s in all of the models. Therefore, any partition

should include broad variation of $h_i - l_i$ across the pairs in both the $in(k)$ and $out(k)$ parts of the partition. In practice this meant requiring at least as many contexts with any particular value of $h_i - l_i$ in the $in(k)$ set as in the $out(k)$ set, while ensuring some instances in the $out(k)$ set as well. For example, Table 1 shows that there are six contexts where $h_i - l_i$ = \$40: The algorithm for constructing partitions required that of these six contexts, four would be found in any $in(k)$ set while two would be found in any $out(k)$ set. A similar constraint was imposed for each unique value of $h_i - l_i$ (\$20 to \$80) found among the contexts of Table 1.

Table 1: The 100 Choice Pairs

the contexts		four pairs			
#	$\langle l, m, h \rangle$	q_a	q_b	q_c	q_d
1	(40,50,60)	5/6	4/6	3/6	2/6
2	(40,50,70)	5/6	4/6	3/6	2/6
3	(40,50,80)	4/6	3/6	2/6	1/6
4	(40,50,90)	4/6	3/6	2/6	1/6
5	(40,60,100)	4/6	3/6	2/6	1/6
6	(40,60,110)	4/6	3/6	2/6	1/6
7	(40,60,120)	4/6	3/6	2/6	1/6
8	(50,60,90)	4/6	3/6	2/6	1/6
9	(50,70,100)	5/6	4/6	3/6	2/6
10	(50,70,110)	4/6	3/6	2/6	1/6
11	(50,70,120)	4/6	3/6	2/6	1/6
12	(60,70,90)	5/6	4/6	3/6	2/6
13	(60,80,110)	5/6	4/6	3/6	2/6
14	(60,80,120)	4/6	3/6	2/6	1/6

the contexts		four pairs			
#	$\langle l, m, h \rangle$	q_a	q_b	q_c	q_d
15	(70,80,100)	5/6	4/6	3/6	2/6
16	(70,80,110)	4/6	3/6	2/6	1/6
17	(70,80,120)	4/6	3/6	2/6	1/6
18	(70,90,110)	5/6	4/6	3/6	2/6
19	(80,90,100)	5/6	4/6	3/6	2/6
20	(80,90,110)	5/6	4/6	3/6	2/6
21	(80,90,120)	4/6	3/6	2/6	1/6
22	(80,100,120)	5/6	4/6	3/6	2/6
23	(90,100,110)	5/6	4/6	3/6	2/6
24	(90,100,120)	5/6	4/6	3/6	2/6
25	(100,110,120)	5/6	4/6	3/6	2/6

Table 2. Numbers of subjects for which the "row" specification has a higher prediction log likelihood than the "column" specification, with 2-tailed sign test p-values below numbers.

Table 2-a. Function-free RDU

	MH	SU	CU	BF
SU	45 0.31			
CU	66 <0.0001	61 <0.0001		
BF	71 <0.0001	62 <0.0001	46 0.22	
DFT	67 <0.0001	65 <0.0001	52 0.0097	41 0.91

Table 2-c. Function-free EU.

	SU	MH	DFT	CU
MH	52 0.0097			
DFT	61 <0.0001	46 0.22		
CU	62 <0.0001	45 0.31	48 0.093	
BF	68 <0.0001	50 0.033	55 0.0011	52 0.0097

Table 2-b. Parametric RDU.

	MH	SU	DFT	BF
SU	44 0.43			
DFT	52 0.0097	55 0.0011		
BF	62 <0.0001	64 <0.0001	54 0.0023	
CU	69 <0.0001	71 <0.0001	57 0.0002	51 0.0183

Table 2-d. Parametric EU.

	SU	DFT	BF	MH
DFT	70 <0.0001			
BF	71 <0.0001	45 0.31		
MH	59 <0.0001	45 0.31	46 0.22	
CU	70 <0.0001	51 0.018	58 <0.0001	45 0.31

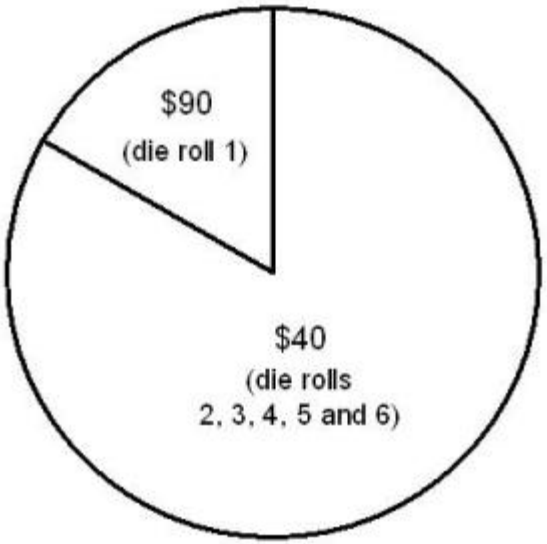
Table 3: Pooled function-free RDU specifications: Numbers of subjects for which the "row" specification has a higher prediction log likelihood than the "column" specification, with 2-tailed sign test p-values below numbers.

	SU	MH	DFT	BV
MH	62 <0.0001			
DFT	64 <0.0001	60 <0.0001		
BV	63 <0.0001	59 <0.0001	50 0.033	
CU	63 <0.0001	58 <0.0001	47 0.15	46 0.22

Notes (Tables 2 and 3). MH = mostly harmless model (linear probability model with dummy coding of option pair dimensions); SU = strong utility model (homoscedastic logit); CU = contextual utility; BF = Blavatskyy-Fishburn model; and DFT = decision field theory.

Figure 1. An example pair, as displayed to subjects.
The pair's "context" in this example is $\langle 40,50,90 \rangle$ (in U.S. dollars).

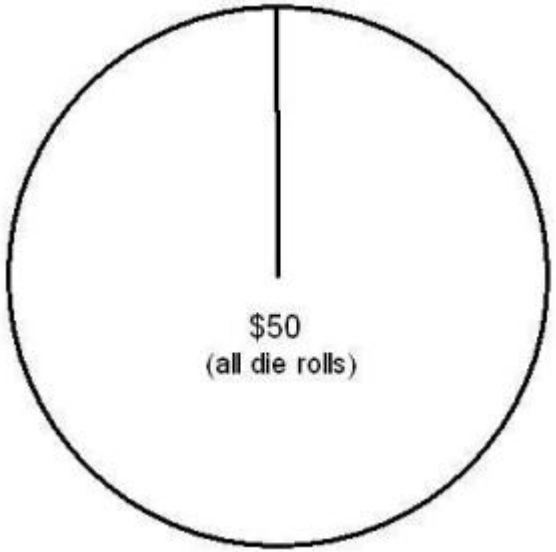
Left option ["*risky*"]



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

Here, h = \$90, q = 1/6 and l = \$40.

Right option ["*safe*"]



Generally, *safe* is m with Prob 1,
where $h > m > l$.

Here m = \$50.

Figure 2. Cumulative distributions of risky choice proportions across days

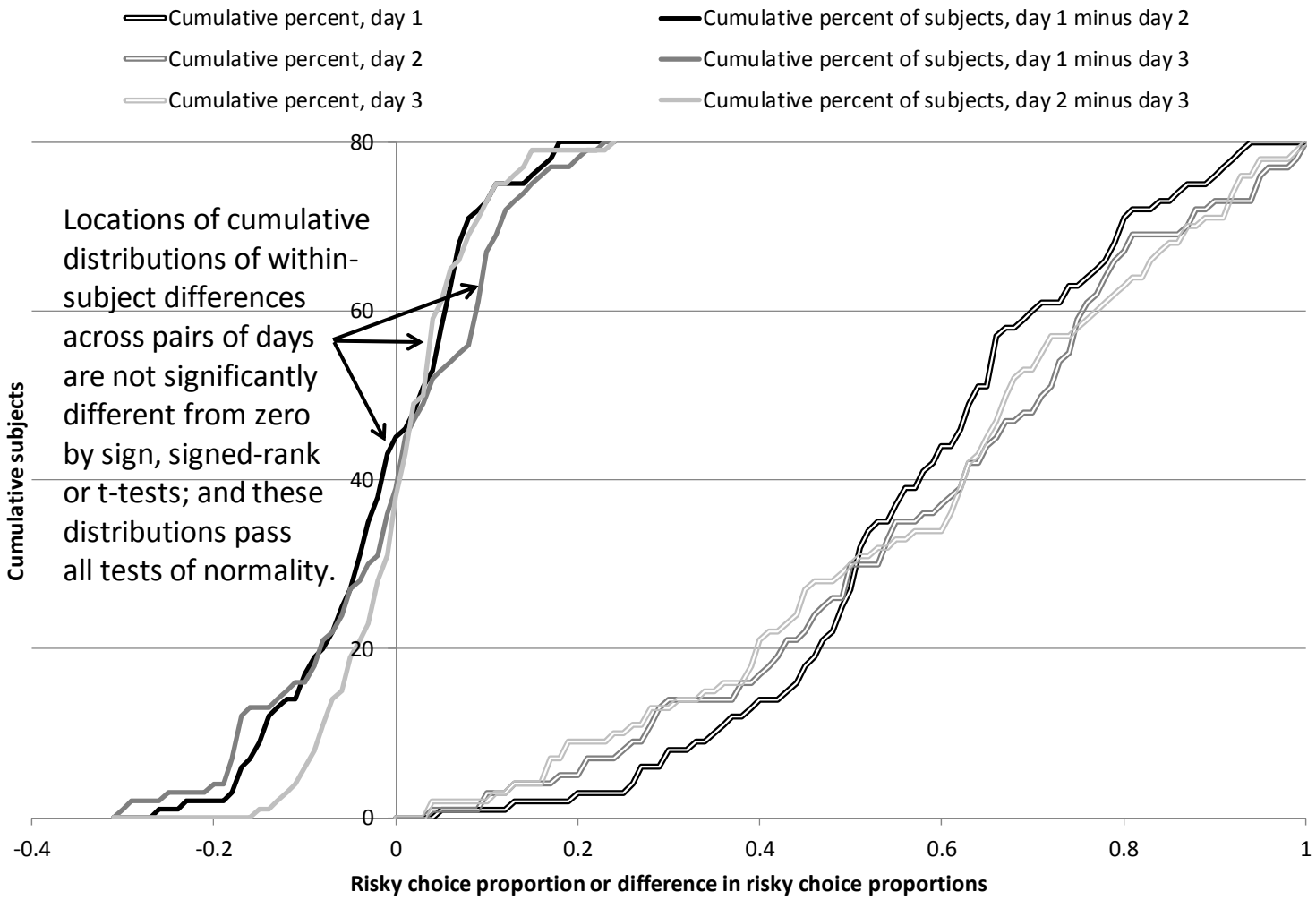
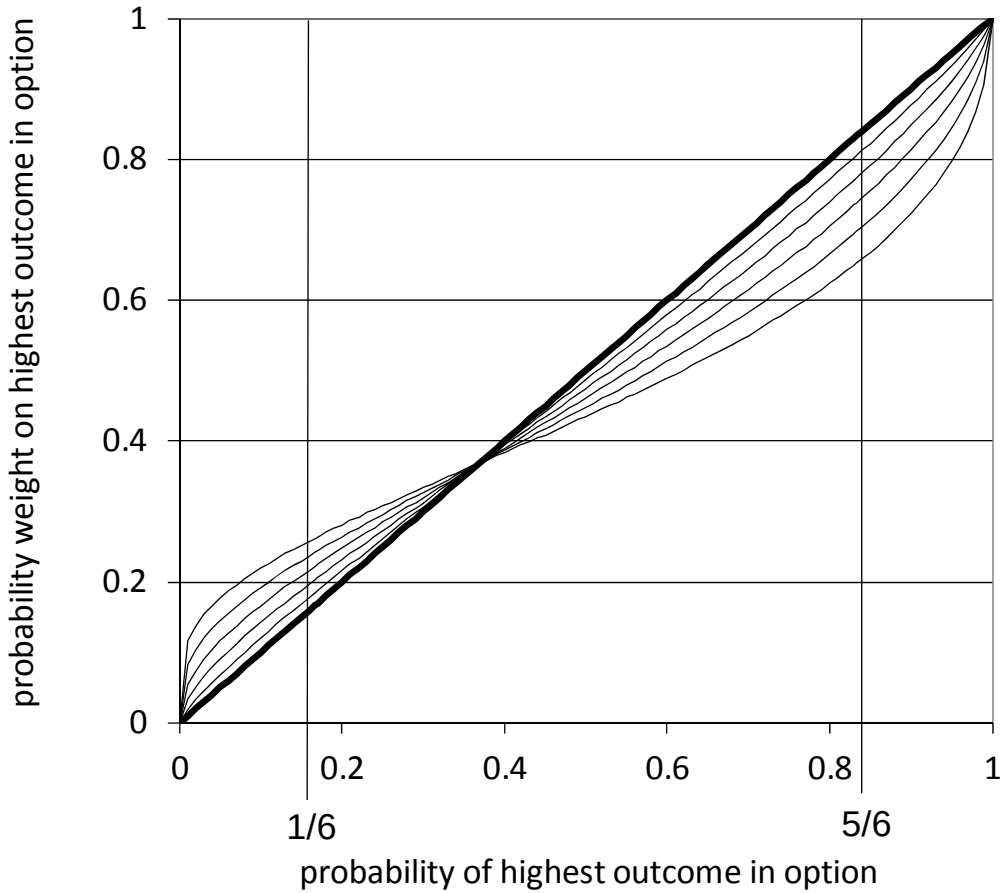
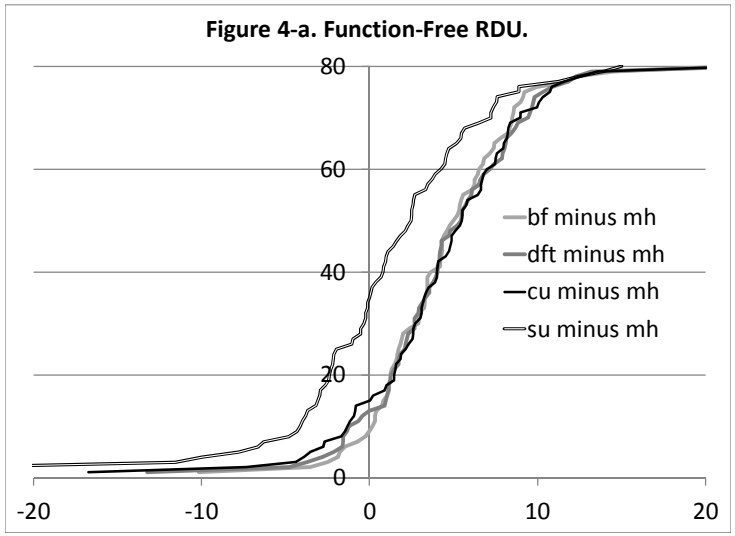
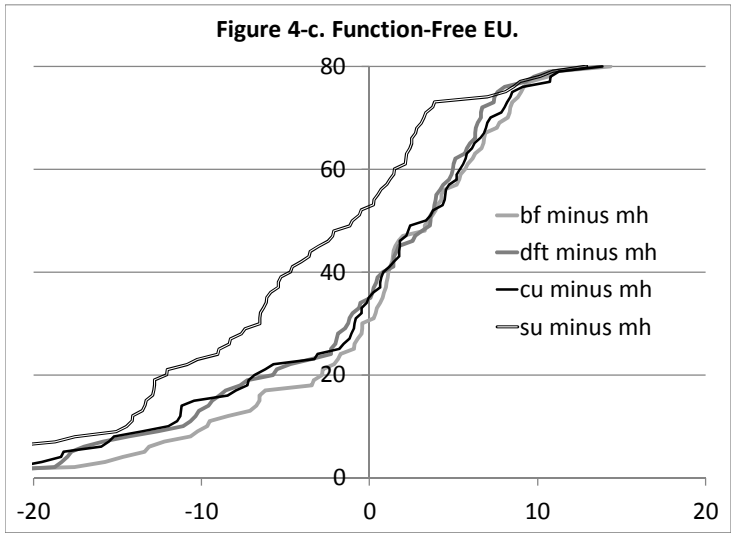


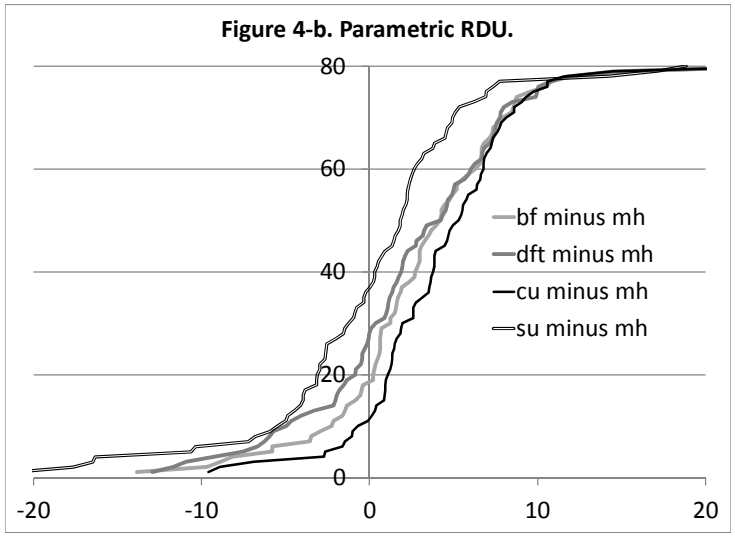
Figure 3: Prelec-type one-parameter weighting functions at 1/6 and 5/6 for widely-held priors (γ = 1 to 0.5)



Figures 4: Cumulative distributions of $X^{s}_{pred}(spec, mh)$ of the four probabilistic models (*spec*) and the mostly harmless model (*mh*). Vertical axes are cumulative subjects; horizontal axes are differences in prediction log likelihoods.







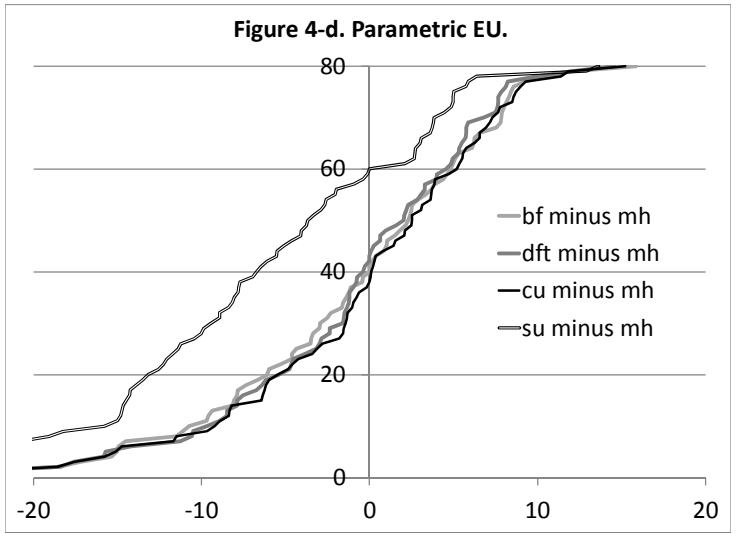
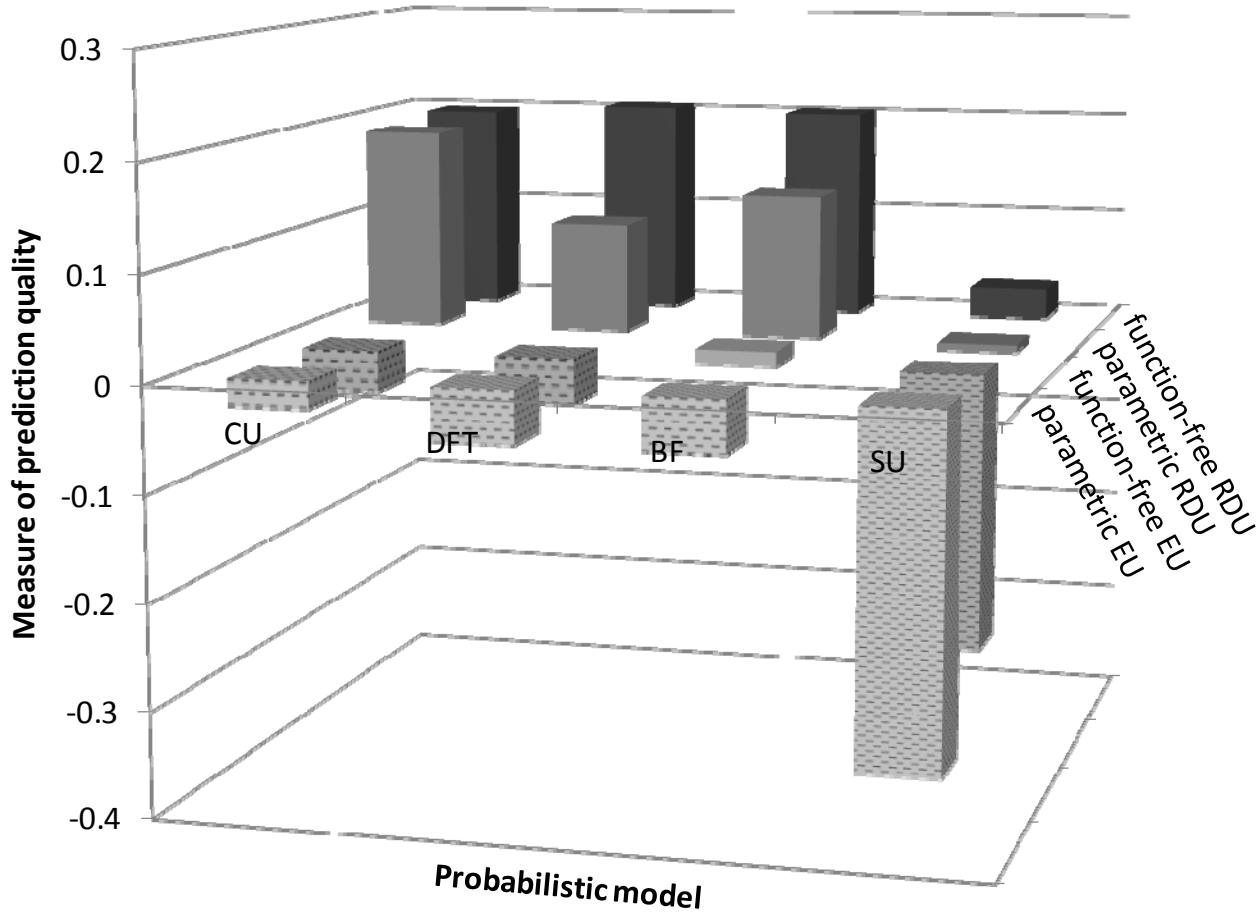


Figure 5. Measure of prediction quality for various specifications.



Notes. The measure of prediction quality is $\bar{Y}_{pred}^{spec}$, defined in eq. 20. Its zero value is the average prediction log likelihood of the mostly harmless model, that is $\mathcal{L}_{pred}^{low,s}$ (see eqs. 22 and 23). Its maximum (unity) is the average log likelihood of the observed choice proportions in the $out(k)$ data sets, that is $\mathcal{L}_{pred}^{high,s}$ (see eq. 21).

(Instructions to subjects) **Instructions**

You will participate in 3 different sessions—one session on each of 3 different days.

On **each** of the three days, you will make **100 choices** from each of 100 pairs of monetary options. Some of the options will involve chance, in the form of a die roll. Option pairs will be presented to you as pie charts, on a computer screen: In each option pair you see, you will choose the option you would prefer to play.

At the end of your third day with us, you will have made 300 choices over your three sessions. ONE of your 300 option choices will then be randomly selected using a bag of 300 tickets with the numbers 1, 2, 3,…, 299, 300 written on them. The numbers 1 to 100 correspond to the 100 choices you will make today, in the order you make them today. Likewise, the numbers 101 to 200 (and 201 to 300) correspond to the 100 choices you will make on your second day (and then on your third day) with us, in the order you make them on those days.

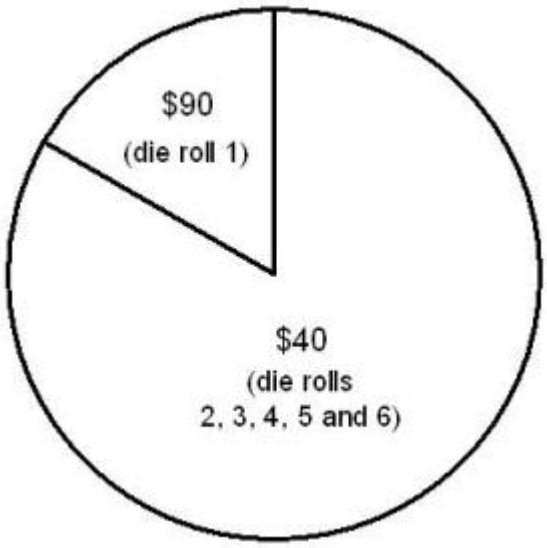
At the end of your third day with us, you will reach into the bag of tickets (without looking inside), pull one out and show us the number. We will then enter that number into the computer, and it will recall that option pair and show the option you chose. That option will determine your payment for participation in this project. If the option you chose requires a die roll, we will then roll a six-sided die to determine your payment.

Notice that since **every** option pair choice you make has a 1 in 300 chance of determining your payment for participation, you have a real reason to consider each option pair with equal care. Also, notice that **only one** of your 300 option pair choices **will** determine your payment.

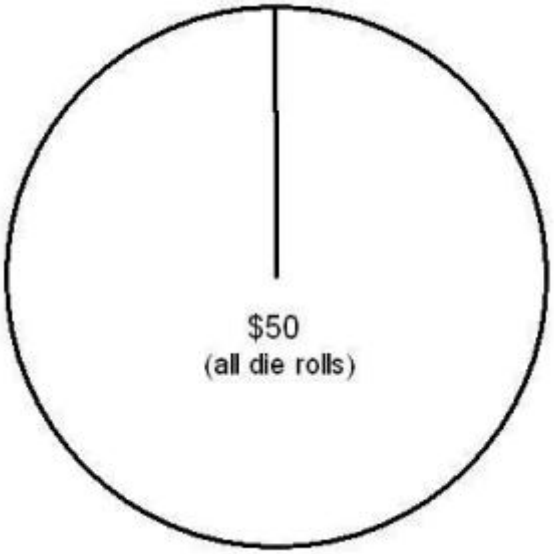
Please note that you won't be able to use a calculator, or pencil and paper, to make your choices. That would take too long for 100 choices…our lab schedule will not accommodate this.

An example of an option pair is shown below. The left option is a 1 in 6 chance of $90 and a 5 in 6 chance of $40: If you chose this option and it was selected to determine your payment, a die roll would be needed to determine the payment. The right option is a sure $50: If you chose this option and it was selected to determine your payment, no die roll would be needed.

Left Option



Right Option



(Instructions to subjects—continued)

The option pair you just saw is only one example. The money outcomes in the option pairs you see will range from $40 to $120, in ten dollar increments. Also, the connection between die rolls and money outcomes varies a lot over those options that involve a die roll, so remember to notice those die rolls when new option pairs appear on the screen for your consideration. Finally, note that the computer will present option pairs to you in a randomized order, and will also randomly select the left/right placement of the options in each pair. So you do not want to assume that option pairs appear in any kind of patterned sequence: They do not. The computer will remember the exact sequence, as well as what you chose, so that you can be paid properly on your last day with us.

Some questions for a break

It is difficult to maintain good attention over 100 choices. Even though the amounts of money in option pairs are not small, almost anyone will get a bit bored with making these kinds of choices after awhile.

Partly for that reason, the 100 option pair choices will be broken into two halves (50 pairs in each half) on each day. Between the halves, on each day, you will answer some survey questions and respond to some questionnaire items. This will go pretty quickly on all three days (a little longer on the second day), and will give you a break each day from the option pair choices.

You'll be able to do everything at your own pace. We believe that each session will last about one hour for most people on most days, but remember that we expect you to have 90 minutes available on each day, so that you are not rushed.

If there is anything you do not understand, please ask us. We will be happiest if you understand exactly how your decisions affect you: We want you to be able to do well for yourself, whatever your believe “doing well” is. We encourage you to do what you want.

Finally, the money for this study comes from grants. This money is earmarked for payment to student participants. We have no alternative use for this money: It must be paid out to participants like you. We must of course make payments only in accordance with the procedure we have described above. But do not worry about taking that money from us: It is specifically earmarked for this and we cannot use it for anything else. We say this, only because some students worry about taking such money from professors. You should not worry about it. The money is grant money, not Dr. Wilcox’s money, and it is earmarked specifically for paying out to student participants like yourself.