Predicting Risky Choices Out-of-Context: A Monte Carlo Study

Nathaniel T. Wilcox
Department of Economics
University of Houston
Houston, TX 77204-5019
713-743-3840
nwilcox@mail.uh.edu

Abstract

Risk attitude is frequently taken to be a general and stable characteristic of an agent, explaining how that agent's decisions cohere across different economic choice contexts. In this view, between-agent differences in risk attitudes should also be a primary source of differences in the behavior of those agents across a given choice context. Accordingly, two primary reasons for measuring risk attitudes is to predict the behavior of agents, and explain variations in the behavior of different agents, in choice contexts that are new, or at least not the same context in which measurement occurs. I refer to this as "out-of-context" prediction and explanation. Here, I examine one elicitation/estimation paradigm in detail—the random lottery pair or RLP design, combined with individual maximum likelihood estimation—and ask how well we can expect to predict and explain out-of-context, given sample sizes (number of lottery pairs) that are typical of RLP designs, under various null hypotheses concerning underlying structural models of decision under risk and stochastic models governing choice probabilities. I conclude that these "out-of-context" prediction and explanation tasks are, statistically speaking, much more difficult than intuition might suggest. Simple model-free and estimation-free approaches perform nearly as well, or even better, than the more formal and asymptotically correct approach of individual maximum likelihood estimation, given the finite sample sizes we typically have. We should expect to be frequently disappointed by the explanatory and predictive power of individually estimated risk attitudes for purely statistical reasons, even if risk attitudes are a stable and general cause of an agent's economic behavior across economic contexts.

I thank Glenn Harrison, John Hey, Graham Loomes, Chris Orme, Lisa Rutström and Robert Sugden for making their data available to myself and others. I am especially grateful to Glenn Harrison for conversation, commentary and support during this work. I also thank the U.S. National Science Foundation for research support under grant SES 0350565.

They have humbled themselves; I will not destroy them, but I will grant them some deliverance…

—2 Chronicles 12:7 (Revised Standard Version)

1. Introduction

Suppose we have data on lottery choices made by seventy subjects. The subjects chose a lottery from a pair of lotteries, and each subject made one hundred such choices. We now want to predict similar choices made by the same seventy subjects from fifty new lottery pairs. The new pairs are different enough from the original hundred that simple projection of observed choice proportions in the original pairs to the new pairs seems wrong: We are predicting individual behavior in a new choice "context." So we pick a model of risky choice, and a stochastic specification of choice probabilities, and estimate the parameters of this model separately for each subject, using the one hundred pair choices and maximum likelihood. We use these estimated parameters and the model to create predicted choice probabilities: Let $\hat{P}_{nm}$ be the predicted probability that subject *n* chooses the "safe"[1] lottery from pair *m* of the fifty new lottery pairs, and let $y_{nm} = 1$ if that actually occurs (zero otherwise).

Curious about our success, we construct a "calibration curve" (Lichtenstein, Fischoff and Phillips 1977). Define subscript sets $nm(\hat{j}) = \{nm \mid \hat{j} - 0.025 < \hat{P}_{nm} \leq \hat{j} + 0.025\}$, denoting subject/pair combinations *nm* for which our predicted probabilities fall into 0.05-wide intervals around midpoints $\hat{j} \in \{0.025, 0.075, 0.125 \ldots, 0.975\}$. Also, let $\bar{\bar{y}}_{\hat{j}} = \sum_{nm \in nm(\hat{j})} y_{nm} / \#nm(\hat{j})$, the actual proportion of safe choices for $nm \in nm(\hat{j})$.[2] If our model is a good one, and our

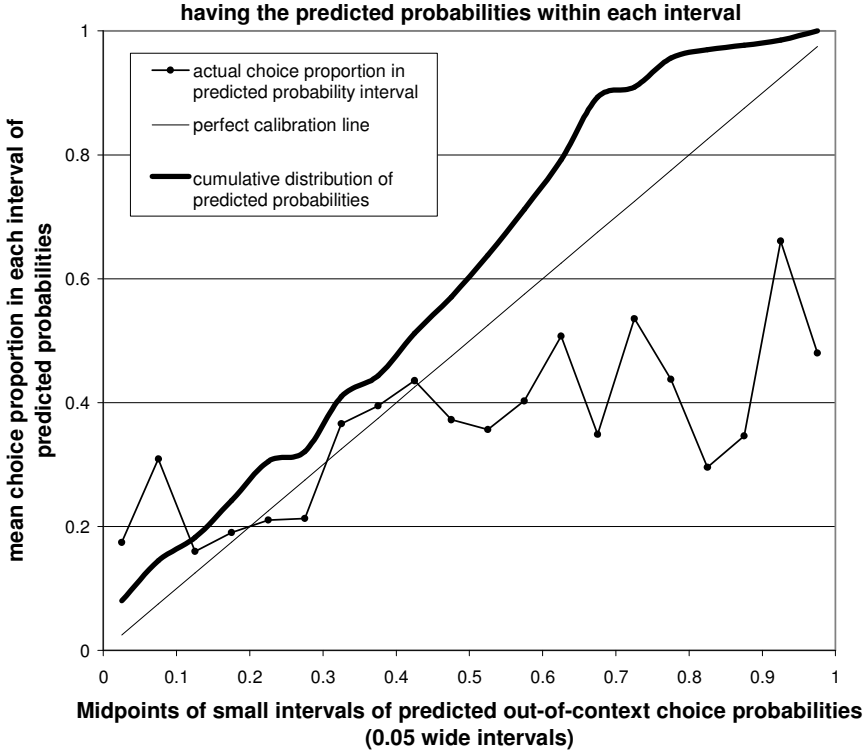
[1] The "safer" lottery in a pair has no obvious normative status; that is, it is not necessarily "correct" to choose it. "Safer than" is a technical but intuitive relation between certain lotteries, defined below at an appropriate time.

[2] The double bar denotes an average taken across pairs *m* and subjects *n*; a single bar denotes an average taken only across pairs for a given subject. The symbol # in front of a set denotes the number of elements in the set.

estimation proceeds with sufficient precision, we should roughly expect $\bar{\bar{y}}_{\hat{j}} \approx \hat{j}$: A graph of $\bar{\bar{y}}_{\hat{j}}$ against $\hat{j}$ is a calibration curve, and we hope it will look roughly like the identity function or a 45 line. Instead, our calibration curve appears similar to Figure 1. The horizontal axis plots the midpoints $\hat{j}$ of small intervals of our predicted probabilities $\hat{P}_{nm}$ of safe choices in the new sample of 50 choices, coming from some model we estimated using the 100 observations of the original sample; the height of each plotted black dot is the actual proportion of safe choices observed for subject/pair combinations in those intervals. The smooth, heavy black line is the cumulative distribution of our predicted probabilities across the predicted probability intervals $\hat{j}$; it shows that the particularly poor-looking right half of the figure involves upwards of forty percent of our predicted probabilities, making it clear that the poor-looking results on the right are not rare and hence are of practical concern.

Should we draw the conclusion that there is something seriously wrong with either the risky choice model, or the stochastic choice specification, we used to make our predictions? We will see that pictures like Figure 1 are expected at such sample sizes, even with a true model specification: The poor-looking result is as likely due to the small sample behavior of maximum likelihood estimation as any misspecification. At these sample sizes, we will almost certainly have poor calibration of actual choices to predicted probabilities regardless of the model of risky choice or the stochastic choice specification we use: One hundred observations per subject is simply not enough data to get very good results here with individual subject estimation. The problem only gets worse if we estimate less parsimonious models than expected utility, such as rank-dependent expected utility; and it becomes especially awful for parametrically tractable versions of the "random preference" model of stochastic choice.

Figure 1. Example of a calibration curve, showing relationship between small intervals of predicted "out-of-context" choice probabilities and actual choice proportions of observations having the predicted probabilities within each interval



In both laboratory and field applications, empirical economists try to build individual-level measures of risk attitudes by using data on either real or hypothetical choices from lottery pairs, both in laboratory settings (Hey and Dardanoni 1987; Archibald and Wilcox 2002) and in large field surveys such as the U.S. Health and Retirement Study (Barsky et al. 1997) or the U.S. Panel Study of Income Dynamics (Hryshko, Luengo-Prado and Sorensen 2006; Kimball, Sahm and Shapiro 2007).[3] There are very good reasons for these efforts. Some behavior of interest to researchers, such as bidding in auctions or investment in assets, theoretically depends on risk attitudes. At the very least, a good measure of risk attitudes might absorb cross-sectional heterogeneity and hence improve the efficiency of hypothesis tests and the precision of

[3] These surveys include a small sequence of hypothetical choices between a respondent's current job, or a job that is identical in all respects except that it will deliver either a proportionally higher or lower salary, with known probabilities. Thus, the sequence of choices are precisely choices of one lottery from a pair of lotteries.

estimates. In certain applications where lagged dependent variables appear in models, such as panel data studies of consumption and saving, or risky occupational and/or health choices, taking account of such heterogeneity can be econometrically crucial.

I show that we cannot really expect such efforts to provide us with very precise information on individual preferences, given the number of discrete lottery choices that are typically present in experiments and surveys that attempt this. For instance, we will see from a combination of random parameters estimation and Monte Carlo simulation that if we had the "true" choice probabilities P_{nm} in our hands (as we would asymptotically, with a "very large" sample for estimation), rather than noisy finite sample estimates $\hat{P}_{nm}$ of these, we might expect to explain as much as sixty or seventy percent of the between-individuals variance in observed proportions of safe lottery choices in the new sample of 50 choices; by contrast, our estimates $\hat{P}_{nm}$ based on the 100 choices of the original sample will typically permit us to explain barely ten percent of that variance—nearly an order of magnitude shrinkage in explained variance. We must have modest expectations of what we can achieve, in terms of out-of-sample prediction and explanation, with the small samples of discrete lottery choices we typically have in our hands: We must humble ourselves.

If we expect more from these predictive and explanatory exercises than is statistically possible, we may make three kinds of mistakes. First, we may unnecessarily dismiss perfectly adequate models of risk preference and stochastic choice for the wrong reasons. Second, many researchers report that risk preferences appear to be at best weakly related across different economic institutions in which risk preferences purportedly influence individual behavior (Berg, Dickhaut and McCabe 2005); indeed some claim to find significantly negative relationships (Isaac and James 2000). Such findings have counterparts in an older psychological literature

(Hershey and Schoemaker 1985). The Monte Carlo work reported here shows that even within a single "institution" or "response mode" (here simple binary lottery choice), we should expect at best weak predictive success from one choice context to another—even with samples of fifty to one hundred observations per subject, which is well beyond the typical depth of individual sampling in many of these experimental designs. Note carefully that this is an entirely statistical phenomenon here, having nothing to do with unstable risk preferences across choice contexts in any interesting sense: In the Monte Carlo simulations that I undertake here, all "simulated subjects" have decision making parameters that are fixed and constant across choice contexts.

Finally, in our unwarranted disappointment, we may go looking for alternative kinds of heterogeneity, such as cognitive heterogeneity,[4] to explain individual differences in behavior, when in fact our first hypothesis (heterogeneity of risk attitudes and stochastic choice parameters) was the major source of cross-sectional heterogeneity (though "poor results" made that appear unlikely). I do not wish to detract from explorations of cognitive heterogeneity (or other potential sources of heterogeneity), a phenomenon I believe is real and important (Ballinger et al. 2006). Indeed, I conjecture that measures of demographic, personality and cognitive heterogeneity may very soon permit us to estimate risk attitudes and stochastic choice parameters with greater precision than we now can from observed lottery choices alone (either because they are sources of nuisance variance, and/or because they are themselves causally related to heterogeneity of attitudes toward risk and randomness of behavior). Rather, it is simply to point out the limits of what we can infer from a lack of association between discrete choice predictions and outcomes when our predictions are based on a handful (say one hundred, a "big" handful) of discrete choices per subject. Those limits are severe.

[4] For instance, see Frederick (2005) or Benjamin, Brown and Shapiro (2006).

Some deliverance will be forthcoming too. In particular, we will see that there are simple estimation-free covariates based on observed choices in the original sample—in fact, simple proportions of safe choices—that explain between-individual variation in proportions of safe choices in the new sample much better than estimated models do, even when the model estimated is known to be the true model (something we can only know in a Monte Carlo situation). This is a particularly damning result, though: It shows how badly behaved maximum likelihood estimation can be in the relatively small samples of lottery choices we typically have.

As a part of this study, I provide a critical primer on stochastic choice models that have commonly been used by experimental economists to model stochastic choice under risk. In the course of this, I will provide extensions of some of these models, and introduce one class of models, called moderate utility models, that have not been considered much in this literature. It will emerge that when it comes to out-of-context prediction and explanation, certain stochastic models face peculiar and special econometric and theoretical challenges which, to my knowledge, have at best received only brief attention in the literature.

2. Three predictive and explanatory tasks

Let $Z = (0,1,2,...,z,...,(I-1))$ be the I non-negative integers $z = 0, 1, 2,\ldots,(I-1)$, representing I equally spaced non-negative money outcomes including zero.[5] The unit outcome varies between experiments and surveys (i.e. "1" could represent \$5 or £50 in Z). For my purposes in this chapter, a lottery S_m is a discrete probability distribution $(s_{m0}, s_{m1},..., s_{mz},..., s_{m(I-1)})$ on Z. A pair m is a set $\{S_m, R_m\}$ of two lotteries. Let Z_m be the "subvector" of Z formed by deleting all

[5] I restrict attention here to lotteries that do not involve losses to allow for expositional clarity and focus; negative outcomes in lotteries would require extensive notational embellishments due to the phenomenon of loss aversion. Ignoring loss aversion here has no consequence for my main econometric points.

elements for which $s_{mi} = r_{mi} = 0$; in words, Z_m is the vector of outcomes that are possible in at least one of the lotteries in pair m. I call Z_m the outcome context or, more simply, the context of pair m. In many experiments, the context of all pairs is a triple (that is, all pairs involve just three possible outcomes). For instance in Hey and Orme (1994), Hey (2001) and Harrison and Rutström (2005), we have $I = 4$ equally spaced money outcomes including zero, or $Z = (0,1,2,3)$; this quadruple yields four distinct contexts, namely (0,1,2), (0,1,3), (0,2,3) and (1,2,3), and all pairs in those experiments are on one of those four contexts. Let $mz = \{m \mid s_{mz} = r_{mz} = 0\}$; for instance, $m0$ is the set of pairs on context (1,2,3) or, equivalently, the pairs which do not have $z = 0$ as a possible outcome.

An experiment consists of presenting the pairs $m = 1, 2,\ldots, M$ to subjects $n = 1, 2,\ldots, N$. In some experiments, each pair is presented repeatedly; in such cases, let $t = 1, 2,\ldots,T$ denote these repeated trials. Let $P_{nmt} = \Pr(y_{nmt} = 1)$ be the probability that subject n chooses S_m in trial t of pair m, where $y_{nmt} = 1$ is this event in the experiment ($y_{nmt} = 0$ if n instead chooses R_m). The trial subscript t may be dropped for experiments with single trials (T=1) or if one assumes that choices are independent across trials and choice probabilities do not change across trials,[6] writing P_{nm} instead of P_{nmt}. I will generally suppress t whenever possible.

Suppose that we have observations y_{nmt} of risky choices from pairs m on contexts (0,1,3), (0,1,2) and (1,2,3) (that is, for pairs in $m2$, $m3$ and $m0$),[7] for subjects n, as from the experiments mentioned above. Let $Y_n^{set} = \{y_{nmt} \mid m \in set\}$: For instance, letting $m23 \equiv m2 \cup m3$ be the set of

[6] See Loomes, Moffatt and Sugden (2002) for ways to treat violations of this assumption. There is substantial evidence of a "drift" with repetition toward increasingly safe choices in several experiments (Hey and Orme 1994; Ballinger and Wilcox 1997; Loomes and Sugden 1998), probably attributable to a small decrease in stochastic parts of decision making with repetition. Harrison et al. (2005) find order effects with just two trials. I abstract from these phenomena here.

[7] I Ignore pairs in $m1$, that is all pairs on context (0,2,3). The reason I do this will be clear later, when I discuss the random preference model of stochastic choice. Briefly, the random preference model is not conveniently generalizable across all contexts for any simple parametric family of random preference distributions.

pairs on the contexts (0,1,2) and (0,1,3), Y_n^{m23} is subject *n*'s choice data for pairs in that set. How well can we predict or explain variation in $y_{nmt} \in Y_n^{m0}$, that is choices made from pairs on context (1,2,3), using only the choice data in Y_n^{m23}? And, how well should we be able to do these things? Consider these three tasks we might ask a model of discrete choice under risk to do well:

Point Prediction (Task I): We would like our model to produce an estimator $\hat{P}_{nmt}(Y_n^{m23})$ of the true choice probabilities P_{nmt} of choices $y_{nmt} \in Y_n^{m0}$, so that the forecasts made by that estimator are as good as possible according to some agreeable criterion, such as a sum of log likelihoods or squared errors.

Explaining Within-Subject, Between-Pair Variation (Task II): We would like our model to produce a covariate $\hat{W}_{nm}(Y_n^{m23})$, perhaps $\hat{P}_{nm}(Y_n^{m23})$ itself, that best explains variations in the $y_{nmt} \in Y_n^{m0}$, according to some agreeable criterion such as the Pearson correlation.

Explaining Between-Subject Variation (Task III): Let $\bar{y}_n^{m0} = \sum_{m \in m0} y_{nmt} / (\#m0)$, the sample proportion of "safe" choices of S_m made by subject *n* across all $m \in m0$. We would like our model to produce a covariate $\hat{A}_n(Y_n^{m23})$, perhaps $\hat{\bar{P}}_n^{m0}(Y_n^{m23}) \equiv \sum_{m \in m0} \hat{P}_{nm}(Y_n^{m23}) / (\#m0)$, that best explains variations in $\bar{y}_n^{m0}$ across subjects *n*, according to some agreeable criterion such as the Pearson correlation.

If we have a properly specified model of the true choice probabilities P_{nmt}, and our estimator of them $\hat{P}_{nmt}(Y_n^{m23})$ is consistent, then for large enough samples Y_n^{m23} we can do no better than this estimator as far as Task I (prediction) goes. It follows that, for sufficiently large samples of data Y_n^{m23}, the best possible covariates for Tasks II and III (explaining within- and between-variation) will in fact be $\hat{P}_{nmt}(Y_n^{m23})$ and $\hat{\bar{P}}_n^{m0}(Y_n^{m23})$, respectively. Are the sample sizes of existing experiments "sufficiently large" for these asymptotic facts to hold, even approximately? If not, how well can we really do at the three tasks relative to the asymptotic ideal where we would have "true" choice probabilities in hand? Are the asymptotically best covariates for tasks II and III actually the best choice, given the sample sizes we typically have available to us? Or are there better and simpler alternatives?

Analytical treatment of finite sample properties of estimators of highly nonlinear models is almost never possible: We are compelled to use time-consuming and highly specific Monte Carlo study to answer these questions. But we <u>can</u> get a good sense of what we are up against by considering a Monte Carlo study of an exceptionally well-suited experimental design with a sample that is large relative to existing experiments. Therefore, I will focus on the experimental design of Hey and Orme (1994), which has #m23 = 100 observations (that is, 100 observations of choices from pairs on contexts (0,1,2) and (0,1,3)) from which to estimate models for each subject n, and #m0 = 50 "out-of-context" observations (that is, 50 observations of choices from pairs on context (1,2,3)) to predict on the basis of those estimated models. Hey and Orme's experimental design also has a specific desirable feature: The probability vectors they use are identical across contexts. Thus we cannot attribute out-of-context prediction failures in the Hey and Orme data to different probability vectors used on different contexts to construct pairs (as we might using the still larger Hey 2001 data set).

3. Two structures of choice under risk

For transitive theories, the structure of choice under risk is a function V of lotteries and a vector of structural parameters β^n such that $V(S_m \mid \beta^n) \geq V(R_m \mid \beta^n) \Leftrightarrow P_{nm} \geq 0.5$. For intransitive theories, it is a function D of lottery pairs and a structural parameter vector β^n such that $D(m \mid \beta^n) \geq 0 \Leftrightarrow P_{nm} \geq 0.5$. Such statements equate the theoretical primitive relation "S_m is weakly preferred to R_m by n" or $S_m \succeq_n R_m$ with "n is not more likely to choose R_m from m," a probability statement, a common approach since Edwards (1954).[8] Structure maps pairs into a set of possible probabilities rather than a single unique probability, and hence underdetermines choice probabilities (unless P_{nm} is always either zero, 0.5 or unity, an empirically dubious hypothesis I henceforth ignore[9]). Stochastic modeling additions to the structure remedy this; these are discussed in the next section.

The expected utility (or EU) structure is $V(S_m \mid U^n) = \sum_{z=0}^{I-1} s_{mz} u_z^n$, where $U^n = (0,1,u_2^n,u_3^n,...u_{I-1}^n)$, such that $\sum_{z=0}^{I-1} s_{mz} u_z^n \geq \sum_{z=0}^{I-1} r_{mz} u_z^n \Leftrightarrow P_{nm} \geq 0.5$. The structural parameters of EU are $\beta^n = (u_2^n,u_3^n,...,u_{I-1}^n)$, subject n's utilities of the highest I–2 outcomes z.[10] Rank-dependent expected utility (or RDEU), with Prelec's (1998) single-parameter parametric form for the weighting function $w(q) = \exp(-[-\ln(q)]^{\gamma^n}) \ \forall \ q \in (0,1)$, $w(0)=0$ and $w(1)=1$, is the

[8] There are alternative stochastic choice models under which this would not be as obvious as it seems, such as Machina (1985); but existing evidence offers little support for this alternative (Hey and Carbone 1995).

[9] A large number of experiments that involve multiple trials of identical pairs document substantial choice switching between trials, even at very short time intervals of a few minutes (Camerer 1989, Starmer and Sugden 1989, Hey and Orme 1994, Ballinger and Wilcox 1997, Loomes and Sugden 1998, Hey 2001). A determined skeptic might still say that the hypothesis $(P_{nm1} = 1 \cup P_{nm2} = 0) \cap (P_{nm1} = 0 \cup P_{nm2} = 1)$ is consistent with this switching across two trials t = 1 and 2. I freely grant the point, but am aware of no theory motivating that hypothesis, so I choose to ignore it here.

[10] Because the utility of money is only unique up to an affine transformation, one can always arbitrarily pick two money outcomes to have the utilities 0 and 1; here, this is done here to the outcomes z = 0 and 1, respectively.

structure $V(S_m \mid U^n, \gamma^n) = \sum_{z=0}^{I-1} \left(w\left(\sum_{i \geq z} s_{mi}\right) - w\left(\sum_{i>z} s_{mi}\right) \right) u_z^n$. Here, $\beta^n = (u_2^n, u_3^n, ..., u_{I-1}^n, \gamma^n)$; EU is a special case of this where $\gamma^n = 1$ so that $w(q) = q$. I use these two structures as exemplars, but the general issues at stake are not specific to them or any specific parametric instantiation of them. That RDEU is a one-parameter generalization of EU is ideal for my purposes: The addition of parameters may have a high cost in our out-of-context prediction and explanation game at the sample size of Hey and Orme (1994). Although I do not use it for estimation here, I have some occasions to discuss EU or RDEU structures with a "constant relative risk aversion" or CRRA utility of money function $u_z^n = z^{1-\varphi^n}/(1-\varphi^n)$, where φ^n is called subject *n*'s "coefficient of relative risk aversion." With CRRA utility, EU is a one-parameter structure where $\beta^n \equiv \varphi^n$.

Finally, here are some terminological and notational conventions. The EU and RDEU structures have an affine transformation invariance property that I use frequently, saying they are affine structures. Specifically, if the utility of money u_z represents an EU or RDEU order, then so does $\dot{u}_z = a + \lambda u_z$, where a and λ are any real numbers and $\lambda > 0$. With this fact in mind, the expression $\lambda[V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]$, while only unique for a particular choice of λ, will frequently be called *V*-distance; this plays a central role in several stochastic models. Finally, consider any pair $m = \{S_m, R_m\} \equiv \{(s_{mj}, s_{mk}, s_{ml}), (r_{mj}, r_{mk}, r_{ml})\}$ on a three-outcome context (j,k,l), $l > k > j$. Whenever possible, choose lottery names S_m and R_m so that $s_{mk} + s_{ml} > r_{mk} + r_{ml}$ and $s_{ml} < r_{ml}$. Lottery S_m then has less probability of either the lowest or highest outcomes *j* and *l*, but a larger probability of the middle outcome *k*, than lottery R_m. In this case, we say that S_m is safer than R_m and call S_m the safe lottery in pair *m*. This labeling is henceforth adopted, and applies to all pairs in Hey and Orme's (1994) experiment. There are other experiments that include some pairs where one lottery first-order stochastically dominates the other, that is where lottery names

can be chosen so that $s_{mk} + s_{ml} \geq r_{mk} + r_{ml}$ and $s_{ml} \geq r_{ml}$, with at least one inequality strict. Such lotteries are not ordered by the "safer than" relation as I have just described it, but I will nevertheless let S_m denote the dominating lottery in such pairs, which I will call "FOSD pairs." The intended meaning of S_m should be clear below by context.

4. The stochastic choice models

The very first decision one makes in stochastic modeling of binary choice is what general class of stochastic models to employ. There are many possibilities here (Becker, DeGroot and Marschak 1963a; Luce and Suppes 1965; Fishburn 1999), but I will confine my discussion to three of these. Two are well-known to experimental economics: These are the random preference and strong utility models. The third are moderate utility models, which are virtually unused by experimental economists, though related stochastic modeling assumptions are found in Hey (1995) and Buschena and Zilberman (2000). I use these here to provide alternatives to the better-known strong utility and random preference models. I discuss each of these in turn, along with their special properties and some empirical evidence about them. The choice between these classes of stochastic models is every bit as much an identifying restriction as any parametric assumptions one makes within any of them. Therefore, a discussion of their empirical implications, successes and failures is important.

To begin, note that some randomness of observed choice has been thought to arise from attention lapses or simple responding mistakes that are independent of pairs *m*. Following Moffatt and Peters (2001), call such events trembles and assume they occur with probability ω^n independent of *m* and, in the event of a tremble, assume that choices of S_m or R_m are

equiprobable.[11] In concert with this, draw a distinction between overall choice probabilities P_{nm} (that in part reflect trembles) and considered choice probabilities P^c_{nm} which depend on characteristics of *m* and govern choice behavior when no tremble occurs. Under these assumptions and definitions, we have

(4.0) $\quad P_{nm} = (1-\omega^n)P^c_{nm} + \omega^n/2$.

Note that under equation (4.0), $P_{nm} \geq 0.5$ iff $P^c_{nm} \geq 0.5$, $\forall\ \omega^n \in [0,1]$. In words, we may give a stochastic definition of preference in terms of either P_{nm} or P^c_{nm} since trembles do not reverse preference directions relative to stochastic indifference (defined as $P_{nm} = P^c_{nm} = 0.5$).

4.1 Random preferences

From an econometric viewpoint, the random preference model views random choice as arising from randomness of structural parameters. We think of each individual subject *n* as having an urn filled with structural parameter vectors β^n. (For instance, a CRRA EU structure with random preferences could be thought of as a situation where subject *n* has an urn filled with various values of her coefficient of relative risk aversion φ^n.) At each new trial *t* of any pair *m*, the subject draws a new parameter vector from this urn (with replacement) and uses it to calculate both $V(S_m \mid \beta^n)$ and $V(R_m \mid \beta^n)$ without error, choosing S_m iff $V(S_m \mid \beta^n) \geq V(R_m \mid \beta^n)$. Let $B^n_m = \{\beta^n \mid V(S_m \mid \beta^n) - V(R_m \mid \beta^n) \geq 0\}$; then under random preferences, $P^c_{nm} = \Pr(\beta^n \in B^n_m)$.[12]

[11] One may also condition ω_n on a task order subscript τ if, for instance, one believes that trembles become less likely with experience, as in Loomes, Moffatt and Sugden (2002) and Moffatt (2005).

[12] For simplicity's sake I assume throughout this chapter that parameter vectors producing indifference in any pair *m* have zero measure for all *n*, so that the sets B_m and $B_{m*} = \{\beta \mid V(S_m \mid \beta) - V(R_m \mid \beta) > 0\}$ have equal measure.

To carry out parametric estimation for any subject *n*, one then needs to specify a joint distribution function $H_\beta(x \mid \alpha^n)$ for β^n, conditioned on some vector α^n of parameters determining the shape of the distribution H_β. I will call the vector $\eta^n = (\alpha^n, \omega^n)$ the stochastic parameters because they determine choice probabilities but are not themselves structural parameters. In random preference models, the stochastic parameters α^n determine a subject's distribution of structural parameters. We then have the overall choice probability

$$(4.11) \qquad P_{nm} = (1-\omega^n)P^c_{nm} + \omega^n/2 = (1-\omega^n)\left(\int_{\beta \in B^n_m} dH_\beta(x \mid \alpha^n)\right) + \omega^n/2 .$$

Substituting (α,ω) for the true parameter vector, one may then use (4.11) to construct a likelihood function for observations y_{nm} conditional on (α,ω), for some subject *n*. This likelihood function would then be maximized in (α,ω) to estimate (α^n,ω^n).

Since EU is a special case of RDEU where $\gamma^n = 1$, we can develop a random preferences specification for RDEU and it will apply equally to EU with this added restriction. The technique I outline here generalizes that of Loomes, Moffatt and Sugden (2002), which was pioneered by Carbone (1997) for random preference EU models. Like Loomes, Moffatt and Sugden, I will simplify the problem by assuming that γ^n is nonstochastic, that is, that the only structural parameters that vary in a subject's "random preference urn" are her outcome utilities.

As developed by them, Loomes, Moffatt and Sugden's (2002) technique is for pairs *m* on a single three-outcome context $Z_m = (j,k,l)$, $l > k > j$. Begin by using the affine structure property of RDEU to express the outcome utilities as $(0,1,1+v^n_m)$, so that $v^n_m \in \mathbb{R}^+$ is a single random

---

However, one may make indifference a positive probability event in various ways; for a strong utility approach based on a threshold of discrimination, see Hey and Orme (1994).

variate describing the distribution of subject *n*'s random utilities on context $Z_m = (j,k,l)$. Then the RDEU structural preference $S_m \succeq_n R_m$ in pair *m* is written

(4.12) $W_{mk} + W_{ml}(1+v_m^n) \geq 0$, where

$$W_{mk} = w(s_{mk}+s_{ml}) - w(s_{ml}) - [w(r_{mk}+r_{ml}) - w(r_{ml})] \text{ and } W_{ml} = w(s_{ml}) - w(r_{ml}).$$

Hence, $S_m \succeq_n R_m$ is equivalent to

$$(4.13)\ v_m^n \underset{<}{\overset{\geq}{}} -1 - W_{mk}/W_{ml} \equiv \frac{w(s_{mk}+s_{ml}) - w(r_{mk}+r_{ml})}{w(r_{ml}) - w(s_{ml})} \text{ as } W_{ml} \equiv w(s_{ml}) - w(r_{ml}) \underset{<}{\overset{\geq}{}} 0\ .$$

Adopt the one-parameter (γ) Prelec weighting function and write $w(q \mid \gamma^n)$ to take notice of this. Let $H_{vm}(x \mid \alpha_m^n)$ be the c.d.f. of v_m^n on context Z_m. Then considered choice probabilities for RDEU under random preferences are[13]

$$(4.14)\ P_{nm}^c = \begin{cases} 1 - H_{vm}\left(\dfrac{w(s_{mk}+s_{ml} \mid \gamma^n) - w(r_{mk}+r_{ml} \mid \gamma^n)}{w(r_{ml} \mid \gamma^n) - w(s_{ml} \mid \gamma^n)} \mid \alpha_m^n \right) \text{ when } w(s_{ml} \mid \gamma^n) - w(r_{ml} \mid \gamma^n) > 0\ , \\ \text{and } H_{vm}\left(\dfrac{w(s_{mk}+s_{ml} \mid \gamma^n) - w(r_{mk}+r_{ml} \mid \gamma^n)}{w(r_{ml} \mid \gamma^n) - w(s_{ml} \mid \gamma^n)} \mid \alpha_m^n \right) \text{ when } w(s_{ml} \mid \gamma^n) - w(r_{ml} \mid \gamma^n) < 0\ . \end{cases}$$

In Hey and Orme's (1994) experiment, our convention that S_m is the safe lottery implies $s_{ml} < r_{ml}$, and $w(q \mid \gamma^n)$ increasing then implies that $w(s_{ml} \mid \gamma^n) - w(r_{ml} \mid \gamma^n) < 0$. So we use the second version of equation 4.14 above for the Hey and Orme data.

Equations 4.0 and 4.14 may be used to build a likelihood function for a subject's choices on context Z_m. This is an ingenious econometric implementation of random preferences RDEU for a single context. But when contexts vary, as they usually do in most experiments and as they

[13] Notice that when $(s_{mk} + s_{ml}) - (r_{mk} + r_{ml})$ and $r_{ml} - s_{ml}$ have opposite signs, the argument of H_{vm} is negative, so that $H_{vm} = 0$. This is the case for FOSD pairs, where one lottery dominates the other. As discussed shortly, the random preference model requires considered choice probabilities to be zero or one in this case, so this is sensible.

definitionally do when we wish to predict behavior out of context, the method quickly becomes intractable except for very special cases. I now work out one such special case. Note that the constraints of this special case are central to the remainder of this chapter and the empirical work undertaken in it: It dictates limits on the data we may take from existing experiments for the purpose of comparing the out of context predictive success of the various structural/stochastic model combinations.

Consider those experiments, such as Hey and Orme (1994), with outcome vector Z = (0,1,2,3) and random utility vector $U^n = (0,1,u_2^n,u_3^n)$ on Z, where $u_2^n \geq 1$ and $u_3^n \geq u_2^n$. Under random preferences, we can let $g_1^n \equiv u_2^n - 1 \in \mathbb{R}^+$ and $g_2^n \equiv u_3^n - u_2^n \in \mathbb{R}^+$ be two underlying random variates determining the random utilities, according to $u_2^n = 1 + g_1^n$ and $u_3^n = 1 + g_1^n + g_2^n$. Therefore, lottery pairs on contexts (0,1,2) and (0,1,3) have outcome utility vectors $(0,1,1+g_1^n)$ and $(0,1,1+g_1^n+g_2^n)$, respectively. Apply the affine transformation $\dot{u} = (u-1)/(u_2^n - 1)$ to the utility vector $(1,u_2^n,u_3^n)$ on context (1,2,3), transforming it into $(0,1,(u_3^n-1)/(u_2^n-1))$. Since $(u_3^n-1)/(u_2^n-1) = 1 + (u_3^n-u_2^n)/(u_2^n-1) = 1 + g_2^n/g_1^n$, lottery pairs on context (1,2,3) may then be regarded as having the outcome utility vector $(0,1,1+g_2^n/g_1^n)$. We have now put the outcome utility vectors of three contexts into the form $(0,1,1+v_m^n)$ that is convenient for estimating the RDEU model with random preferences, in terms of two underlying random variables g_1^n and g_2^n.

The crucial step, though, is choosing the random variates g_1^n and g_2^n so that g_1^n, $g_1^n + g_2^n$ and g_2^n/g_1^n have tractable parametric distributions. The only choice I am aware of is two independent gamma variates, each with the gamma distribution's c.d.f. $G(x|\phi,\kappa)$, with identical "scale

parameter" κ^n but possibly different shape parameters ϕ_1^n and ϕ_2^n.[14] Under this choice, g_1^n will have c.d.f. $G(x|\phi_1^n, \kappa^n)$, $g_1^n + g_2^n$ will have c.d.f. $G(x|\phi_1^n + \phi_2^n, \kappa^n)$, and g_2^n / g_1^n will have the c.d.f. $B'(x \mid \phi_2^n, \phi_1^n)$ of the "beta-prime" distribution on $\mathbb{R}^+$, also known as a "beta distribution of the second kind" (Aitchison 1963).[15] These assumptions imply a joint distribution of $u_2^n - 1$ and $u_3^n - 1$ known as "McKay's bivariate gamma distribution" with a correlation coefficient $\sqrt{\phi_1^n / (\phi_1^n + \phi_2^n)}$ between u_2^n and u_3^n in subject *n*'s "urn" of random preferences (McKay 1934; Hutchinson and Lai 1990).

An acquaintance with the literature on estimation of random utility models may make these assumptions seem very special and unnecessary. They <u>are</u> very special, but this is partly because theories of risk preferences over money outcomes are very special relative to the kinds of preferences that typically get treated in that literature. Consider the classic example of transportation choice well-known from Domencich and McFadden (1975). Certainly, most post-Becker economists expect the value of time and money to be correlated across the population of commuters. But for a single commuter making a specific choice between car and bus on a specific day, we do not require a specific relationship between the disutility of commuting time and the marginal utility of income she happens to "draw" from her random utility urn on any particular morning at the moment of choice. This gives us fairly wide latitude when we choose a distribution for the unobserved parts of her utilities of various alternatives.

---

[14] The restriction that g_1 and g_2 have common scale resembles a restriction that we will see later in strong utility models (to the effect that the variance of errors of computation, perception or discrimination does not vary across contexts). At any rate, for random preferences using gamma variates on contexts (0,1,2) and (0,1,3), I examined this restriction across those contexts and find no cause for worrying about it, for both the EU and RDEU structures.

[15] The ratio relationship here is a generalization of the well-known fact that the ratio of independent chi-square variates follows an F distribution. Chi-square variates are gamma variates with common scale parameter $\kappa = 2$. In fact, a beta-prime variate can be transformed into an F variate: If x is a beta-prime variate with parameters a and b, then bx/a is an F variate with degrees of freedom $2a$ and $2b$. This is convenient because almost all statistics software packages contain thorough call routines for F variates, but not necessarily any call routines for beta-prime variates.

This is definitely not true of any specific trial of her choice of a lottery pair *m*: We demand, for instance, that she "draw" a vector of outcome utilities that respects monotonicity in *z*, since that is a maintained assumption about all preference orderings "in her urn." Moreover, we face the problem of making assumptions about unobserved random utilities probabilistically consistent across pair contexts. If we wish to treat utilities of money in a general way by picking a joint distribution of u_2^n and u_3^n (and not commit to something like the CRRA utility of money), our pick immediately implies exact commitments regarding the distribution of any and all functions of u_2^n and u_3^n. The issue does not arise in a data set where subjects make choices from pairs on just one context, as in Loomes, Moffatt and Sugden (2002): In this simplest of cases, <u>any</u> distribution of v_m^n on $\mathbb{R}^+$, including the lognormal choice they make, is a wholly legitimate hypothesis. But as soon as each subject makes choices from pairs on several different overlapping contexts, staying true to the demands of random preferences is much more exacting. Unless we can specify a joint distribution of g_1^n and g_2^n that implies it, we are <u>not</u> entitled to assume that v_m^n follows lognormal distributions in all of three overlapping contexts for a single subject.[16] Our choice of a joint distribution for g_1^n and g_2^n, or the joint distribution of the v_m^n in contexts (0,1,2) and (0,1,3), has exact and inescapable implications for the distribution of v_m^n in context (1,2,3). Carbone (1997) correctly saw this in her random preference treatment of the EU structure. Under these circumstances, a cagey choice of the joint distribution of g_1^n and g_2^n is necessary to stay true to the formal meaning of random preferences.

There is a strong implication of my specific "independent gamma" version of the random preference model that requires serious notice, as well as a weaker implication of <u>any</u> random

[16] Although ratios of lognormal variates are lognormal, there is no similar simple parametric family for sums of lognormal variates. The independent gammas with common scale are the only workable choice I am aware of.

preference approach to this particular out-of-context prediction situation. Notice that while the common scale parameter κ^n will have descriptive content in contexts (0,1,2) and (0,1,3), it is wholly irrelevant in context (1,2,3); only the estimates of ϕ_1^n and ϕ_2^n from pairs $m \in m23$ will be relevant to our predictions about pairs $m \in m0$. Econometrically, the random preferences model as specified above "drops sample information" from Y_n^{m23}, specifically the information embodied in the estimate of κ^n, as it moves to out-of-context prediction of observations in Y_n^{m0}. In small samples, this may be a significant handicap for out-of-context prediction.

Generally, while the absolute scale <u>levels</u> of the random variates g_1^n and g_2^n may be descriptively relevant in contexts (0,1,2) and (0,1,3), only the <u>ratio</u> of their scales (and not the absolute levels of their scales) can matter for random preference prediction and explanation on the context (1,2,3) since in this context $v_m^n \equiv g_2^n / g_1^n$ is the ratio of the two underlying random variates. This is true regardless of the joint distribution of g_1^n and g_2^n. This is a somewhat weaker implication than the strong implication of the specific "independent gamma model" described above, but it nevertheless implies that certain relevant sample information in the estimation contexts (scale levels) cannot matter in the prediction context. Therefore, I conjecture that <u>any</u> random preference approach will have relatively poor small-sample predictive performance in the particular tasks I examine here. To my knowledge, no one has addressed these specific or general issues in discussions of random preferences. Theorists typically do not consider estimation and out-of-context prediction issues, and Loomes and Sugden, the chief proponents of the random preference approach (1995; 1998; Loomes, Moffatt and Sugden 2002), have not (as far as I know) considered experiments with data from multiple outcome contexts.

The random preference model can be quite limiting in practical applications. For instance, you should by now wonder what happened to the "lost context" (0,2,3). In data sets such as Hey and Orme (1994), a quarter of observations are, after all, on that context; I have only developed a unified treatment of three-fourths of such data, the choices from pairs on contexts (0,1,2), (0,1,3) and (1,2,3). On the context (0,2,3), we have $v_m^n \equiv g_2^n/(1+g_1^n)$. Is there a choice of the joint distribution of g_1^n and g_2^n that gives us convenient parametric distributions of g_1^n, $g_1^n + g_2^n$, g_2^n/g_1^n and $g_2^n/(1+g_1^n)$, so that those four distributions collectively depend on just three parameters? I have found no answer to this question. But it is worth stating it, since it is a good question to think about. To have a name, call it the random preferences RDEU four-context, three-parameter problem. I look forward to seeing a solution. For other stochastic choice models, such as "strong utility" (considered next), the similar problem is trivially simple.

Specifications that adopt some parametric form for the utility of money, and then regard the randomness of preference as arising from the randomness of a utility function parameter, offer no obvious escape from these difficulties, at least for the RDEU structure. For instance, if we adopt the CRRA form, it is fairly simple to show that this implies $v_m^n \equiv (l^{\varphi^n} - k^{\varphi^n})/(k^{\varphi^n} - j^{\varphi^n})$, where (*j*,*k*,*l*) is the context of pair *m*. Substituting into equation 4.13 and noting our convention that S_m is the safe lottery so that $w(s_{ml}) - w(r_{ml}) < 0$, we then have

$$(4.15) \qquad S_m \succeq_n R_m \Leftrightarrow \frac{l^{\varphi^n} - k^{\varphi^n}}{k^{\varphi^n} - j^{\varphi^n}} \leq \frac{w(s_{mk} + s_{ml}) - w(r_{mk} + r_{ml})}{w(r_{ml}) - w(s_{ml})}.$$

There are two possible routes for implementing this when φ^n is a random variable. The first is to solve the inequality in (4.15) for φ^n as a function of *j*, *k* and *l*, the pair characteristics, and

whatever parameters of $w(q)$ we have. We could then choose a distribution for φ^n and be done. I invite readers to try it: Contexts (0,1,2) and (0,1,3) are simple but context (1,2,3) is intractable.

A second route is suggested by an approach that works well for the EU structure where $w(q) \equiv q$. In this case, although we still cannot analytically solve 4.15 for all contexts, we can use numerical methods to find φ_m^* (to any desired degree of accuracy) prior to estimation, for each pair m on whatever context, such that

$$(4.16)\quad \frac{l^{\varphi_m^*} - k^{\varphi_m^*}}{k^{\varphi_m^*} - j^{\varphi_m^*}} = \frac{s_{mk} + s_{ml} - (r_{mk} + r_{ml})}{r_{ml} - s_{ml}}.$$

Then we can choose a distribution $H_\varphi(x \mid \alpha^n)$ for φ^n and use $P_{nm}^c = 1 - H_\varphi(\varphi_m^* \mid \alpha^n)$ as our model of considered choice probabilities under the EU structure and random preference. For RDEU, however, φ_m^* is a <u>function</u> of any parameters of the weighting function $w(q)$. Therefore we cannot simply provide a constant φ_m^* to our model; we need the function $\varphi_m^*(\gamma)$ (in the case of the Prelec weighting function with parameter γ) so that we can write $P_{nm}^c = 1 - H_\varphi[\varphi_m^*(\gamma) \mid \alpha^n]$. But we have been here before: We cannot solve (4.15) for this function, so we would have to approximate this function numerically, on the fly, for each pair m, within our estimation. Numerical methods probably exist for such tasks, but they are beyond my current knowledge. On the basis of this discussion, though, I think it fair to say that in the case of theories of choice under risk more general than expected utility theory, random preferences can be extremely difficult to work with in practical senses that matter to empirical economists.

Random preference models have a feature that makes them seductive identifying restrictions in the realm of <u>structural tests</u>, that is, tests concerning the descriptive truth of an assumed structure V (Loomes and Sugden 1995). Consider a set Ω of pairs with some common property,

or such that pairs in it stand in some relationship to one another, such that there is a theorem about the structure V (or an axiomatic property of V) to the effect that

(4.15) $V(S_l \mid \beta) \geq V(R_l \mid \beta) \Leftrightarrow V(S_m \mid \beta) \geq V(R_m \mid \beta) \ \forall \beta, \forall l$ and $m \in \Omega$.

Call such theorems structural preference equivalences or SPEs. Sets of lottery pairs used to reveal "common ratio effects" [17] (e.g. Kahneman and Tversky 1979; Loomes and Sugden 1998) are an example of such sets Ω; it is precisely because EU implies an SPE on those sets that researchers regard common ratio effects as a problem for the EU structure. If an SPE is descriptively correct, then $B_l \equiv B_m$. The random preference model then immediately predicts that $P^c_{nl} = P^c_{nm}$ and therefore $P_{nl} = P_{nm}$, independently of the form of G_β or the value of η^n. And just as obviously, population-level expectations of these choice probabilities must be equal as well. Therefore, the theory-testing enterprise (but not the β and η estimation enterprise) becomes simple and distribution-free if random preferences are the correct stochastic model. All SPEs may be tested with a simple test of the equality of all sample proportions of choices of $S_m \in \Omega$.

There are few SPEs in decision under risk which are common to a broad collection of structures, so it is difficult to test random preferences independently of the structural theory under scrutiny. One that stands out, however, is first-order stochastic dominance or FOSD. Recall my convention that, in FOSD pairs m, the dominating lottery is denoted by S_m; and let Ω_{fosd} be the set of all such pairs. Almost all structures (including EU and RDEU) predict that $V(S_m \mid \beta) > V(R_m \mid \beta) \ \forall \beta, \forall m \in \Omega_{\text{fosd}}$, so that $P^c_{mn} = 1$ and $P_{mn} = 1 - \omega_n / 2 \ \forall m \in \Omega_{\text{fosd}}$.

---

[17] Let $\{(z_1, \tau s),(z_2, \tau r)\}$ denote a lottery pair where $z_2 > z_1$ are the only nonzero outcomes possible, and $\tau s > \tau r$ are probabilities of receiving z_1 and z_2, respectively, $s > r$ and $\tau \in [0,1]$. Typically, preference directions switch from the relatively safe lottery $(z_1, \tau s)$ to the relatively risky lottery $(z_2, \tau r)$ as τ falls; this is called the common ratio effect and it is contrary to the EU structure. For any given z_1, z_2, s and r, sets of pairs for all $\tau \in [0,1]$ are perhaps the best-known example of what I mean by "an expected utility SPE."

Therefore, we expect violations of FOSD to occur at a very modest rate that is common to all $m \in \Omega_{fosd}$, and this seems to be observed in various studies (e.g. Loomes and Sugden 1998). Loomes, Moffatt and Sugden (2002, p. 126) argue that "the low rate of dominance violations…must count as evidence [favoring random preferences]" because they feel that other stochastic models, such as strong utility, do not predict this.

As a pure question of econometric modeling, I find this argument unpersuasive. It is simple to add the restriction $P_{mn} = 1 - \omega^n / 2 \ \forall \ m \in \Omega_{fosd}$ to any stochastic model that already contains a tremble[18] with no additional parameters, so there is no loss of parsimony associated with such a modification. As a purely theoretical question, it depends on how you feel about information processing arguments. If you sometimes find them sensible, as I do, then this simple econometric modification can make excellent theoretical sense. A first stage of processing where transparent dominance[19] is first detected can proceed on the basis of computationally cheap ordinal comparisons and avoid compensatory judgments, function evaluations and weighting operations that are likely both more computationally costly and more error-prone (indeed, only a very poorly designed information processor would not exploit such efficient and normatively sound short-cuts, where available). Finally, the fact that random preferences correctly describe FOSD pair choices says nothing about their descriptive or predictive adequacy in pairs where interesting tradeoffs are at stake which, of course, is what really matters to the bulk of applied microeconomic theory and empirical microeconomics.

---

[18] Even in data sets containing no FOSD pairs like Hey and Orme, the hypothesis $\omega^n = 0 \ \forall \ n$ appears to be rejected (Moffatt and Peters 2001). Therefore, trembles seem empirically necessary before we even broach the subject of modeling violations of FOSD; so extending the tremble to account for FOSD entails no loss of parsimony.

[19] Pairs where there is a "transparent dominance" relation are actually a proper subset of all FOSD pairs; for a suggested definition, see Blavatskyy (2007). Dominance relations are sometimes "nontransparent" and FOSD violations for these situations do occur at rates far too high to be properly described as tremble events (Tversky and Kahneman 1986, Birnbaum 2004), though this existing evidence comes from hypothetical or "near-hypothetical" (very small likelihoods of decisions actually counting) designs.

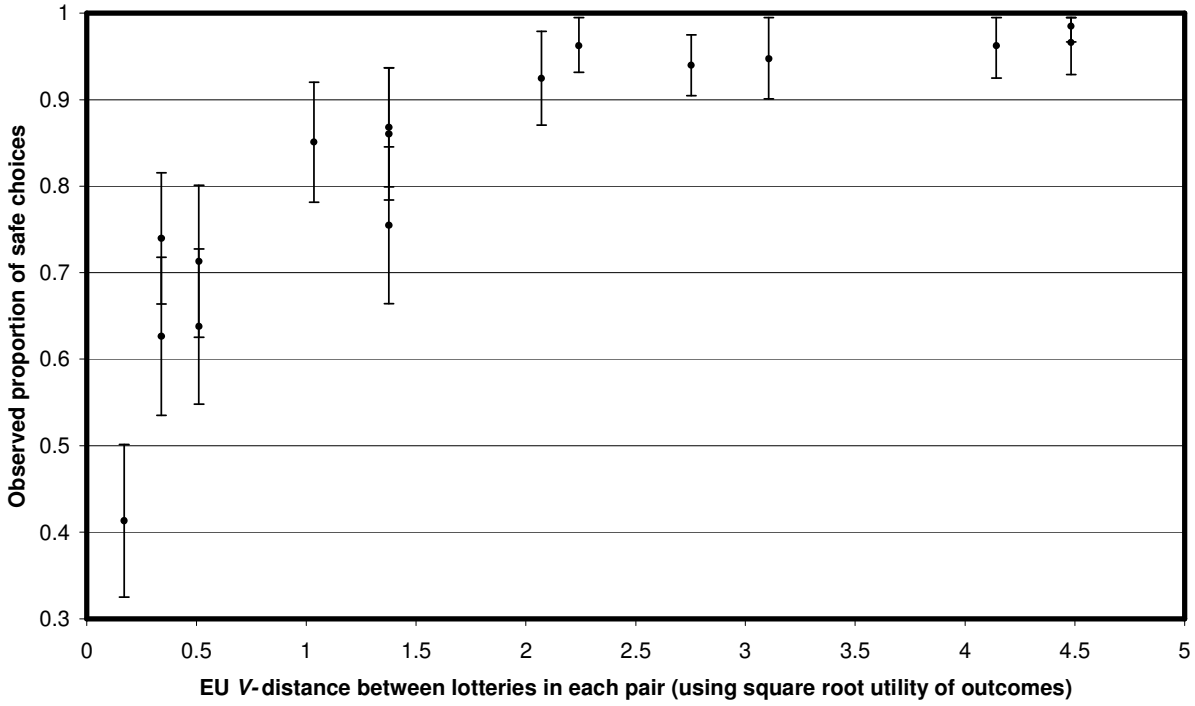
The random preference model also makes predictions, in concert with expected utility theory and typical minor assumptions, which are nonintuitive and easily rejected. Strong utility models do not make those predictions and in fact correctly predict the observed patterns in question. For one example, consider a set Ω_{mps} of mean-preserving spread pairs, defined so that R_m is a mean-preserving spread of $S_m \; \forall \; m \in \Omega_{\text{mps}}$. As is well-known (Rothschild and Stiglitz 1970), under expected utility any weakly concave $u(z)$ implies that $S_m \succeq R_m$, while any weakly convex $u(z)$ implies the opposite. If we assume that all utility functions in a random preferences urn are either weakly concave, weakly convex or both (linear utility), then, mean-preserving spread pairs are an SPE: $B_l \equiv B_m \; \forall \; l$ and $m \in \Omega_{\text{mps}}$, so that $P_{nl} = P_{nm} \; \forall \; l$ and $m \in \Omega_{\text{mps}}$ under random preferences, expected utility and the assumption that all utility functions are either weakly concave, weakly convex or both—a common assumption in both empirical and applied theoretical work.

Figure 2 shows observed proportions of safe choices (and conservative 95% confidence intervals around them) in sixteen mean-preserving spread pairs from Hey (2001). It is clear, first, that a test against the equality of these sixteen choice proportions would easily be rejected. Additionally, notice that the proportions have been ordered according to the EU structure *V*-distance between the lotteries in each pair, using the square root CRRA utility function $u(z) = 2z^{0.5}$ to calculate this distance.[20] Notice that proportions of safe choices generally rise with this distance, as if this distance had stochastic behavioral meaning. Random preference models have the property that *V*-distance between lotteries in a pair is irrelevant to the choice probability in the pair. By contrast, *V*-distance is central to considered choice probabilities in both the strong and moderate utility models discussed subsequently. Yet unique *V*-distance does not flow

[20] The 0.5 coefficient of relative risk aversion is chosen because it is similar to values estimated in many studies; the general thrust of Figure 2 is unchanged for other plausible choices in the unit interval.

directly from affine structures; the next section on strong utility opens with a discussion of this point.

Figure 2. Proportion of safe lottery choices in 16 mean-preserving spread pairs in Hey (2001): Black dots are the proportions; whiskers are conservative 95 percent confidence intervals.



A similar example involving lottery triples, rather than lottery pairs, is moderate violations of the "betweenness" property of expected utility (and other structures that produce linear lottery indifference curves in a Machina-Marschak triangle). These are expected under strong utility EU models, while they should happen only at a low tremble rate under a random preference EU model. Becker, DeGroot and Marschak (1963b) showed rates of violation of betweenness in about 30% of all choices from suitably constructed lottery triples. This is far too high a rate to be a tremble event, but just about right for strong utility models, given contemporary knowledge about retest reliability of lottery choice (e.g. Ballinger and Wilcox 1997). Blavatskyy (2006a)

discusses how patterns of betweenness violations observed within and across many, many studies are predicted by strong utility models.

A proponent of random preferences would now reply that RDEU structures violate betweenness, so that such violations can be explained by a random preference RDEU model. She would also say that RDEU structures do not imply an SPE for all $m \in \Omega_{\text{mps}}$; therefore, Figure 1 does not tell decisively against random preferences in general, but only when combined with the EU structure. And she would be correct. Yet the case for random preferences made by Loomes and Sugden (1998) is made on the back of (i) the first-order stochastic dominance SPE, and (ii) the fact that both random preferences and strong utility, when combined with the EU structure, are inconsistent with the full evidence of common ratio effect. I have already explained why I view the former as an unconvincing point; and the two examples above show that the latter claim, while true of the common ratio effect that Loomes and Sugden highlight, is false for both mean-preserving spreads and betweenness, where strong utility clearly outperforms random preferences from the viewpoint of the EU structure.

Finally, without restrictions on the domain of allowable preference orderings, random preference models do not imply any sort of stochastic transitivity—not even weak stochastic transitivity—even if all preference orderings in the subject's "preference urn" are structurally transitive (Loomes and Sugden 1995; Fishburn 1999). The reason for this is identical to the well-known "voting paradox" of public choice theory (Black 1948): Unless all preference orderings in the urn have the property of single-peakedness, there need be no Condorcet winner.[21] Those who regard tests of transitivity as central to empirical decision research may find this aspect of

[21] Briefly, let there be just three equally likely linear (and hence transitive) orderings in subject n's urn of orderings of lotteries C, D and E, denoted *CDE, DEC* and *ECD*, where each ordering is from best to worst. Then $P_{nj} = 2/3$, $P_{nk} = 2/3$ and $P_{nl} = 1/3$, violating all stochastic transitivities.

random preferences deeply troubling. Strong and moderate utility do not have this difficulty, but they do not have the highly convenient (I say seductive) SPE property of random preferences.

4.2 Strong utility.

Because EU and RDEU are affine structures, the *V*-distance between lotteries is not unique, but instead given by $\lambda \cdot [V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]$ where λ is an arbitrary positive constant. Strong utility models attach behavioral meaning to *V*-distance. In particular, the strong utility model is the assumption that there exists an increasing function $F:\mathbb{R} \to [0,1]$, with $F(0) = 0.5$ and $F(x) = 1 - F(-x)$ (i.e. skew-symmetry about zero), such that

(4.21) $$P_{nm}^c = F\left(\lambda^n [V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]\right).$$

From one theoretical viewpoint, then, a strong utility model for expected utility or RDEU might seem ill-conceived since the *V*-distance which is its argument is theoretically of arbitrary magnitude. We need to keep in mind, though, that the representation theorems of expected utility, RDEU and others are representations of an underlying system of preference directions, not an underlying system of choice probabilities. A more positive way to view the matter is that from the viewpoint of the structure *V*, λ^n is a free parameter that may be chosen to do stochastic descriptive work that is not the structure's original descriptive purpose. No preference direction represented by the structure *V* is changed, for any fixed β^n, as λ^n varies.

There is a virtually one-to-one relationship between strong utility models and homoscedastic "latent variable models" widely employed in empirical microeconomics for modeling discrete dependent variables. In general, such models assume that there is an underlying but unobserved

continuous random latent variable y^*_{nm} such that $y_{nm} = 1 \Leftrightarrow y^*_{nm} \geq 0$; then we have $P^c_{nm} = \Pr(y^*_{nm} \geq 0)$. In our context, the latent variable takes the form

(4.22) $$y^*_{nm} = V(S_m \mid \beta^n) - V(R_m \mid \beta^n) - \sigma^n \varepsilon,$$

where ε is a mean zero random variable with unit variance and c.d.f. $F(x)$ such that $F(0) = 0.5$ and $F(x) = 1 - F(-x)$, usually assumed to be the standard normal or logistic c.d.f.[22] The resulting latent variable model of a considered choice probability is then

(4.23) $$P^c_{nm} = F\left([V(S_m \mid \beta^n) - V(R_m \mid \beta^n)] / \sigma^n\right).$$

In latent variable models, the random variable $\sigma^n \varepsilon$ may be thought of as random computational, perceptual or evaluative error in $V(S_m \mid \beta^n) - V(R_m \mid \beta^n)$, with σ^n being proportional to the standard deviation of this noise. As σ^n approaches zero, choice probabilities converge on either zero or one, depending on the sign of $V(S_m \mid \beta^n) - V(R_m \mid \beta^n)$; in other words, the observed choice becomes increasingly likely to express the underlying preference direction. To complete the analogy with a strong utility model, one may interpret λ^n as equivalent to $1/\sigma^n$. In keeping with common (but not universal) parlance, I will call λ^n subject *n*'s precision parameter. In strong utility models, the stochastic parameter vector is $\eta^n = (\lambda^n, \omega^n)$.

The strong utility model was first axiomatized by Debreu (1958), but it has a very old pedigree going back to Thurstone (1927) and Fechner (1966/1860) and many writers call it the Fechnerian model. Strong utility models imply a testable condition known as strong stochastic transitivity.[23] Consider any three pairs *j*, *k* and *l*, denoting choices between lotteries *C* and *D*, *D* and *E*, and *C* and *E*, respectively; then if $P_{nj} \geq 0.5$ and $P_{nk} \geq 0.5$, it must also be true that $P_{nl} \geq$

[22] We may not wish to make these standard assumptions, however, for reasons having to do with restrictions on the largest 'plausible' computational errors of evaluation subjects might make. See Blavatskyy (2007) for a discussion.
[23] This is true only for transitive structures, but this is still a very broad class of structural theories of choice under risk.

$max[P_{nj},P_{nk}]$. Although strong stochastic transitivity holds much of the time, evidence against its general descriptive adequacy in many judgment and choice contexts is common and longstanding in economic and psychological literature (Block and Marschak 1960; Luce and Suppes 1965; Tversky and Russo 1969; Tversky 1972; Luce 1977), and that evidence coincides with theoretical reasoning based on similarity and/or dominance relations (Debreu 1960); this same evidence generally supports moderate stochastic transitivity instead, which characterizes the moderate utility model considered subsequently. In the specific case of lottery choice, some evidence against strong stochastic transitivity comes from experiments where lottery outcome probabilities are made uncertain or imperfectly discriminable by experimental manipulation (e.g. Chipman 1963; Tversky 1969), but there are occasional violations with standard lotteries too (Ballinger and Wilcox 1997).

As noted earlier, the special case of first-order stochastic dominance can be handled quite easily as an add-on to either a strong or moderate utility model. Let $\delta_m = 1 \ \forall \ m \in \Omega_{\text{fosd}}$, $\delta_m = 0$ otherwise; then the overall choice probability under the strong utility model would be

$$P_{nm} = (1-\delta_m)[(1-\omega^n)P_{nm}^c + \omega^n/2] + \delta_m(1-\omega^n/2), \text{ or}$$

$$(4.24) \quad P_{nm} = (1-\delta_m)\left\{(1-\omega^n)F\left(\lambda^n[V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]\right) + \omega^n/2\right\} + \delta_m(1-\omega^n/2).$$

This simply says that the probability of a dominance violation is to be treated solely as a result of tremble events, and not of noisy utility comparisons which an efficient information processor would in any case not perform after detecting dominance. Yet objections to strong utility models based on similarity effects are not so easily brushed aside; and in any case the strong utility model's supposition that the error variance of noise σ^n is independent of the characteristics of lottery pairs has no obvious computational justification. Interestingly, these two criticisms can be treated in a theoretically unified manner, and that is just what moderate utility models do.

One of Luce's (1959) stochastic choice models, known as the strict utility model, may be thought of as a logarithmic special case of strong utility models, and it appears in contemporary applied work such as Holt and Laury (2002). Considered choice probabilities in this model are

$$(4.25) \qquad P_{nm}^{c} = V(S_m \mid \beta^n)^{\lambda^n} / [V(S_m \mid \beta^n)^{\lambda^n} + V(R_m \mid \beta^n)^{\lambda^n}].$$

A little bit of algebra shows that this is equivalent to

$$(4.26) \qquad P_{nm}^{c} = \Lambda\left[\lambda^n \left(\ln[V(S_m \mid \beta^n)] - \ln[V(R_m \mid \beta^n)]\right)\right],$$

where $\Lambda(x) = [1 + e^{-x}]^{-1}$ is the Logistic c.d.f. This closely resembles equation 4.21 except that natural logarithms of *V* are differenced to create the latent variable, rather than differencing *V* itself. Note that strict utility requires strictly positive values of *V*, and since we have already expressed outcome utilities so that the minimum outcome has a utility of zero this is satisfied for all lotteries except a sure zero outcome.[24]

If we wish to examine how risk aversion in the EU structure changes with a proportional change in outcomes, as say Holt and Laury (2002) do, it is worth noticing that strict utility imposes an extremely strong restriction on EU structures with CRRA utility functions. Let π be a positive constant, consider the outcome vector $\pi Z = (0, \pi, 2\pi, ..., \pi(I-1))$, and let $\{\pi S_m, \pi R_m\}$ be a pair on this outcome vector. The linearity of the EU structure in probability, and CRRA utility of outcomes, then imply that $V(\pi S_m \mid \varphi^n) = \pi^{1-\varphi^n} V(S_m \mid \varphi^n)$, and hence that

$$(4.27) \qquad \ln[V(\pi S_m \mid \varphi^n)] - \ln[V(\pi R_m \mid \varphi^n)] = \ln[V(S_m \mid \varphi^n)] - \ln[V(R_m \mid \varphi^n)] \ \forall \ \pi.$$

This immediately implies that, under strict utility, a CRRA EU structure has the property that considered choice probabilities are independent of a proportional increase in outcomes. This is

[24] Formally, there is something peculiar about the marriage of affine structures like EU and RDEU and strict utility models. The axiom systems behind affine structures imply only that *V* is unique up to an affine transformation. Yet formally speaking, the axiom systems that produce strict utility models imply that the *V* within the stochastic specification is the stronger kind of scale known as a ratio scale, in which *V* must be strictly positive and is unique only up to a ratio transformation.

clearly not true of strong utility; in this case, the argument of F would be $\lambda^n \pi^{1-\varphi^n}[V(S_m \mid \varphi^n) - V(R_m \mid \varphi^n)]]$, so that where $V(S_m \mid \varphi^n) - V(R_m \mid \varphi^n) > 0$, the considered choice probability is an increasing function of the scale of outcomes π for all $\varphi^n < 1$. It follows that Holt and Laury's inference of increasing relative risk aversion depends on strict utility as a stochastic identifying restriction; the broad patterns of their data (increased proportions of safe choices with increased outcome scale) are also consistent with strong utility and constant relative risk aversion. My point here, though, is not to criticize Holt and Laury's preferred interpretation of their data (which may well be correct), but rather to note how stochastic assumptions can be crucial identifying restrictions for fairly important inferences.[25] The example also motivates an empirical comparison between strict and strong utility.

4.3 Moderate utility.

Moderate utility models get their name from their property of moderate stochastic transitivity. Using the same three pairs j, k and l used to define strong stochastic transitivity in the previous section, moderate stochastic transitivity is: If $P_{nj} \geq 0.5$ and $P_{nk} \geq 0.5$, then $P_{nl} \geq min[P_{nj}, P_{nk}]$. As mentioned in the previous section, moderate stochastic transitivity characterizes judgment and choice behavior fairly well in many psychological studies (though there are interesting exceptions, e.g. Tversky 1969) and frequently better than strong stochastic transitivity. Given this and the common use of strong utility models in contemporary work, it seems sensible to consider a moderate utility alternative here. Econometrically, moderate utility

[25] In fact, any common proportional change in the V of both lotteries in a pair will leave choice probabilities constant under strict utility—including of course a "common ratio" change (see footnote 15 above) in both lotteries for EU structures. In this respect strict utility resembles a random preference model. Yet it is not a random preference model for either EU or RDEU: For instance, under strict utility, considered choice probabilities of choosing stochastically dominated lotteries are not identically zero, as they are for true random preference models.

is, first, a heteroscedastic latent variable model, that is one where the standard deviation of judgmental noise is conditioned on pairs *m* so that we write σ_m^n, and considered choice probabilities become

(4.31) $$P_{nm}^c = F\left([V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]/\sigma_m^n\right).$$

However, moderate stochastic transitivity also requires that standard deviations σ_m^n behave like a distance norm and satisfy the triangle inequality $\sigma_j^n + \sigma_k^n \geq \sigma_l^n$ (Halff 1976). To see this, let $V_{nS} \equiv V(S \mid \beta^n)$ be shorthand for subject *n*'s structural *V* of any lottery *S*. Moderate stochastic transitivity fails if we have both $P_{nl} < P_{nj}$ and $P_{nl} < P_{nk}$ when P_{nj} and P_{nk} exceed one half. With equation (4.31), this would be both $(V_{nC} - V_{nE})/\sigma_l^n < (V_{nC} - V_{nD})/\sigma_j^n$ and $(V_{nC} - V_{nE})/\sigma_l^n < (V_{nD} - V_{nE})/\sigma_k^n$. For affine structures we can choose $V_{nC} = 1$ and $V_{nE} = 0$, so that this failure becomes both $1/\sigma_l^n < (1 - V_{nD})/\sigma_j^n$ and $1/\sigma_l^n < V_{nD}/\sigma_k^n$, or both $\sigma_j^n < \sigma_l^n(1 - V_{nD})$ and $\sigma_k^n < \sigma_l^n V_{nD}$; adding these yields $\sigma_j^n + \sigma_k^n < \sigma_l^n$, the opposite of the triangle inequality. Letting d_m be the distance between the lotteries in pair *m* according to some distance norm, then, this suggests rewriting the moderate utility model in the form

(4.32) $$P_{nm}^c = F\left(\lambda^n d_m^{-1}[V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]\right).$$

Conveniently, the moderate utility model allows one to treat similarity without any extra parameters, say with the Euclidean norm $\left(\sum_{i=1}^{I}(s_{im} - r_{im})^2\right)^{0.5}$ for d_m. Or if adding another parameter is acceptable, one may specify the Minkowski norm $\left(\sum_{i=1}^{I} \mid s_{im} - r_{im} \mid^a\right)^{1/a}$ and estimate *a* ≥ 1 instead, or use any norm obeying the triangle inequality. In general, equation (4.32) implies that for <u>given</u> *V*-distance $\lambda^n[V(S_m \mid \beta^n) - V(R_m \mid \beta^n)]$, choice becomes less noisy as the lotteries

in pair *m* become more similar according to a distance norm (as d_m becomes small). Intuitively, moderate utility asserts that more similar lotteries are compared with less noise.[26]

Carroll (1980) pioneered a simple computational underpinning for this intuition called the wandering vector or WV model. It implies the Euclidean norm as the proper choice for d_m, so that the WV model is a moderate utility model with no extra parameters, and hence an attractive competitor to random preferences and strong utility. Therefore, I illustrate it very briefly here for the expected utility structure. Suppose that subject *n* has a noisy perception of her utilities of each outcome *z*; in particular, suppose this utility is a random variable $\breve{u}_z^n = u_z^n - \xi_z^n$, where $\xi_z^n \sim N[0,(\sigma_{\breve{u}}^n)^2] \ \forall\, z$ and u_z^n is her mean utility of outcome *z*. At each new trial of any pair *m*, assume that a new vector of noisy utility perceptions occurs, so that there is a new realization of the vector $(\xi_0^n, \xi_1^n, ..., \xi_{I-1}^n)$, the "wandering" part of the utility vector. Then by definition, $V(S_m \mid \beta^n) - V(R_m \mid \beta^n) - \sigma^n \varepsilon_m \equiv \sum_{z=0}^{I-1}(s_{mz} - r_{mz})\breve{u}_z^n \equiv \sum_{z=0}^{I-1}(s_{mz} - r_{mz})u_z^n - \sum_{z=0}^{I-1}(s_{mz} - r_{mz})\xi_z^n$, so that $\sigma^n \varepsilon_m \equiv \sum_{z=0}^{I-1}(s_{mi} - r_{mi})\xi_z^n$. Since $\sum_{z=0}^{I-1}(s_{mi} - r_{mi})\xi_z^n$ is a linear combination of normally distributed random variables, it is normally distributed too. If we further assume that $\mathrm{cov}(\xi_z, \xi_{z'}) = 0 \ \forall\, z \neq z'$—what Carroll and De Soete (1991) call the "standard" wandering vector model—the variance of $\sum_{z=0}^{I-1}(s_{mi} - r_{mi})\xi_z^n$ is then $(\sigma_{\breve{u}}^n)^2 \sum_{z=0}^{I-1}(s_{mz} - r_{mz})^2$. Therefore, the standard deviation of the latent variable's error term $\sigma_n \varepsilon_m$ becomes $\sigma_{\breve{u}}^n \left(\sum_{z=0}^{I-1}(s_{mz} - r_{mz})^2\right)^{0.5}$. Thus

[26] Note that since the distances d_m are distances between entire lotteries, this is a measure of the similarity of two lotteries. One may also ask questions about the similarity of individual dimensions of lotteries, e.g. are these two probabilities of receiving outcome z_i very close, and hence so similar that the outcome z_i can safely be ignored as approximately irrelevant to making a decision? This "dimension-level similarity" is a different kind of similarity not dealt with by d_m, but it also has decision-theoretic force: It implies a different structure, usually an intransitive one with a *D* representation rather than a *V* one. See Tversky (1969), Rubinstein (1988) or Leland (1994).

the standard wandering vector or WV model is a moderate utility model of the equation (4.32) form, where $\lambda^n = 1/\sigma_{\vec{u}}^n$, d_m is the Euclidean distance norm, and $F(x)$ is the standard normal c.d.f.

Moderate utility models are just one variety of heteroscedastic latent variable models. The wandering vector model nicely illustrates how such models can be given plausible computational underpinnings that connect to perennial decision-theoretic puzzles such as the role of similarity, and produce testable implications such as moderate stochastic transitivity. We ought also to expect stochastic consequences of pair complexity, and there is evidence of this (e.g., Sonsino, Benzion and Mador 2002), but I refrain from any development of this here. Perhaps this kind of stochastic modeling deserves wider examination by experimentalists. Again, strong utility models are simple and powerful workhorses, and probably equal to many estimation tasks; but they have had descriptive problems that moderate utility models largely solved, at least in psychologists' experiments. While many of those experiments did not follow the methodological dicta of experimental economics, they still give us a good reason to examine the WV model alongside strong utility and random preference.

Finally, I suggest another type of moderate utility model that, to my knowledge, has not been discussed elsewhere: Call it contextual utility. Let $z_m^{\min}$ and $z_m^{\max}$ denote the minimum and maximum possible outcomes in the context Z_m of pair m, and let $V(z \mid \beta^n)$ be subject n's structural value of a "degenerate" lottery that pays z with certainty. Contextual utility is a moderate utility model in which the distance norm is $d_m^n = [V(z_m^{\max} \mid \beta^n) - V(z_m^{\min} \mid \beta^n)]$,[27] so that considered choice probabilities are

$$(4.33) \qquad P_{nm}^c = F\left(\lambda_n [V(S_m \mid \beta^n) - V(R_m \mid \beta^n)] / [V(z_m^{\max} \mid \beta^n) - V(z_m^{\min} \mid \beta^n)]\right).$$

[27] This will satisfy the triangle inequality for all triples of choice pairs with overlapping outcome contexts. This rules out only some FOSD pairs, which would in any case be taken care of in the manner of equation 4.24.

Contextual utility essentially asserts that the stochastic perceptual impact of *V*-distance in a pair is mediated by the range of possible outcome utilities in a pair, a notion which has a ring of psychophysical plausibility about it. Econometrically, it is the assumption that the standard deviation of computational error σ_m^n is proportional to the range of outcome utilities in pair *m*.

There is a serious theoretical reason for considering contextual utility as an alternative to other stochastic choice models in the specific domain of choice under risk. Contextual utility makes the stochastic implications of structural definitions of the "more risk averse than" relation sensible, while strong utility does not. For instance, let *n* and *o* be two subjects with CRRA EU structural preferences, and suppose that their coefficients of relative risk aversion are $\varphi^n > \varphi^o$, so that in Pratt's (1964) structural sense we would say that "*n* is (structurally) locally and globally more risk averse than *o*." Suppose also that stochastic choice is governed by strong utility and that the two subjects have equivalent precision parameters $\lambda^n = \lambda^o$. Then it will be true that $P_{nm}^c > P_{om}^c \ \forall\, m \in (\Omega_{\text{mps}} \cap m3)$. That is, for mean-preserving spread pairs on context (0,1,2), subject *n* will be more likely to choose the safe lottery from those pairs than subject *o*. Put differently, the stochastic implications of greater structural risk aversion will "make sense" on context (0,1,2), holding precision constant in a strong utility model. But it will <u>not</u> necessarily be true that $P_{nm}^c > P_{om}^c \ \forall\, m \in (\Omega_{\text{mps}} \cap m0)$: That is, for mean-preserving spread pairs on context (1,2,3) it can happen that *n* is more risk averse than *o* in Pratt's sense, but that *o* is more likely to choose the safe lottery from those pairs than is *n* under strong utility (given equivalent precision):[28] Holding precision constant, greater structural risk aversion will <u>not</u> necessarily have

[28] The intuition is quite simple. As the coefficient of relative risk aversion gets larger, the utility of outcomes function becomes progressively "flatter" over any given range of outcomes above zero. As a result, *V*-distance between lotteries on contexts where the minimum possible outcome exceeds zero approach zero as the coefficient of relative risk aversion becomes arbitrarily large. Thus, the strong utility assumption that the standard deviation of evaluative error is independent of context implies that all choice probabilities approach 0.5 for sufficiently great

a "sensible" stochastic implication. In Wilcox (2007) I discuss this problem with strong utility (and strict utility) and several stochastic modeling remedies for it. In this chapter, where we will not be imposing a parametric utility function such as the CRRA form, the contextual utility model is the most straightforward way of restoring the congruence between structural definitions of "more risk averse than" and the stochastic implications of that relation.[29]

5. The data.

The Hey and Orme (1994) experiment provides the data for my analysis. The unit outcome in Hey and Orme's design is U.K. £10, so that the outcome vector $Z = (0,1,2,3)$ corresponds to currency outcomes of 0, £10, £20 and £30. There are $M = 100$ lottery pairs in the experiment: 25 pairs on each of the contexts (0,1,2), (0,1,3), (0,2,3) and (1,2,3). A common set of 25 pairs of probability vectors was used to create the 25 lottery pairs on all four contexts. All subjects made all one hundred choices twice (once on a first day, and again a few days later) so there are $T = 2$ trials of every lottery pair in this experiment. Thus, there are a total of 50 choice observations on each of the four contexts, for every subject. For reasons discussed in section 4.1 on random preferences, I will not be using any of the data from the context (0,2,3) anywhere in this study. The 50 pairs $m \in m23$ are the estimation context—the pairs on contexts (0,1,2) and (0,1,3); the 100 choices coded by the dummy variables $y_{nmt} \in Y_n^{m23}$ are the data for the estimation. The 25

---

relative risk aversion on contexts not involving outcome zero. This specific problem is noticed by Loomes, Moffatt and Sugden (2002) and Moffatt (2005), but they do not notice the relationship between the problem and definitions of "more risk averse than;" as a result, or so I argue in Wilcox (2007), their "fixes" for the problem are not wholly satisfying. The contextual utility model fixes the problem by "normalizing" V-distances by the range of outcome utilities in any particular context: This range shrinks fast enough as risk aversion grows to fix the problem and to make received definitions of "more risk averse than" match plausible stochastic definitions of that relation.

[29] Another approach is to make λ^n in the strong utility model covary with coefficients of risk aversion in a very specific but theoretically sensible way; see Wilcox (2007) for details.

pairs $m \in m0$ are the prediction context—the pairs on context (1,2,3); these 50 choices are coded by the $y_{nmt} \in Y_n^{m0}$.

Hey and Orme (1994) had eighty subjects in their experiment. I will be using the data from just sixty-eight of them, so that here $N = 68$. If there is very little or no variation in choices for a subject, estimating even a three-parameter model (such as EU with strong utility and no trembles) for that subject becomes perilous or impossible.[30] My rule for including subjects (chosen before viewing any real or Monte Carlo results of out-of-context prediction) is that there must be at least three choices of both safe and risky lotteries by a subject in the estimation context. The binding constraint turns out to be at least three risky choices in all cases: there are twelve subjects of Hey and Orme's eighty who violated this (making 98 or more choices of safe lotteries in the 100 estimation context choices). The remaining sixty-eight provide my data.

Hey and Orme (1994) allow subjects to express indifference between lotteries. One may actually model indifference responses with the help of an extra threshold parameter within a strong or moderate utility framework, an approach examined by Hey and Orme. An alternative parameter-free approach, and the one I take here, is to treat indifference in a direct way suggested by decision theory, where the indifference relation $S_m \sim_n R_m$ is the intersection of two weak preference relations, i.e. "$S_m \succeq_n R_m \cap R_m \succeq_n S_m$." This suggests treating indifference responses as two responses in the likelihood function—one of S_m being chosen from m, and another of R_m being chosen from m—but dividing that total log likelihood by two since it is really based on just one independent observation. Formally, the definite choice of S_m adds $\ln(P_{mn})$ to the total log likelihood function; the definite choice of R_m adds $\ln(1-P_{mn})$ to that

[30] Of course, a random parameters estimation approach can include these subjects. Later, Tables 10, 11 and 12 will provide a comparison between random parameters estimation results using just the sixty eight selected subjects versus all eighty of Hey and Orme's subjects.

total; and indifference adds $[\ln(P_{mn}) + \ln(1-P_{nm})]/2$ to that total. This is algebraically equivalent to letting $y_{nmt} = 1/2$ denote indifference responses and letting $y_{nmt}\ln(P_{nm}) + (1-y_{nmt})\ln(1-P_{nm})$ be the generic log likelihood term for all three possible responses. From here on, indifference responses are coded this way and all computations and statistics involving y_{nmt} use that coding (1 for safe choice, 0.5 for indifference and 0 for risky choice).[31]

Table 1. Descriptive Statistics from the Hey and Orme (1994) Data

Pair Context	Switching rate	Proportion of safe choices (pooled over subjects)	Distribution of subject proportions of safe choices		
			10^{th} centile	median	90^{th} centile
Estimation context *m*23…(0,1,2) and (0,1,3)	0.199	0.724	0.396	0.760	0.952
Prediction context *m*0…(1,2,3)	0.334	0.343	0.136	0.330	0.542
		mean	10^{th} centile	median	90^{th} centile
Distribution of $\rho_{w^*,n}^{m0}$ across subjects (benchmark for task II performance)		0.387	0.197	0.380	0.595

The first two rows of Table 1 shows basic descriptive statistics in the estimation and prediction contexts. The second column shows the switching rate, which is the mean of $|y_{nm1} - y_{nm2}|$ across all subjects and pairs in each context; it is a measure of the retest reliability of choice in such experiments.[32] The switching rate is greater in the prediction context than the

[31] See also Papke and Wooldridge (1996) and Andersen et al. (2007) for related justifications of this approach.

[32] Actually, as shown by Ballinger and Wilcox (1997), switching rates are determined not just by choice reliability, but also by heterogeneity of choice probabilities across subjects and pairs, which depresses switching rates. For instance, if there were no such heterogeneity at all in context *m*23, the aggregate safe choice proportion 0.724 would generate a switching rate equal to 2·0.724·(1–0.724) = 0.3966, much higher than the actual switching rate of 0.1994 in context *m*23. The difference 0.3966–0.1994=0.1972 is one measure of heterogeneity of choice probabilities in the

estimation context. The third column shows the sample proportions $\bar{y}^{=m23}$ and $\bar{y}^{=m0}$ of safe choices across all subjects and pairs in contexts *m*23 and *m*0: Safe choices predominate in the estimation context, but risky choices predominate in the prediction context. Finally, let $\bar{y}_{cx}^{set}$ denote the x^{th} centile of the distribution of individual subject choice proportions $\bar{y}_{n}^{set}$ for pairs $m \in set$, where *set* denotes some context. The rightmost three columns show $\bar{y}_{c10}^{set}$, $\bar{y}_{c50}^{set}$ and $\bar{y}_{c90}^{set}$ for pairs in *set* = *m*23 (first row) and *set* = *m*0 (second row). The $10^{\text{th}}/90^{\text{th}}$ centile range of these proportions is noticeably larger in the estimation context than the prediction context. These simple differences between the contexts illustrate the hard work of task I (out-of-context prediction) that our models face. From choices that largely tend towards safety, we must predict choices that mostly favor risk; from relatively reliable choices, we must generate relatively unreliable choices; and from relatively heterogeneous behavior, we must predict noticeably less heterogeneous behavior.

Let E(*S*) and Var(*S*) denote the mean and variance of outcomes in any lottery *S*. To provide a benchmark result for task II (explaining within-subject, between-pair variation)—a "position to beat," so to speak, for our models—I will use Pratt's (1964) result that the risk premium of a lottery with small risk is approximately Var(*S*)·*clra*[E(*S*)]/2, where *clra*(*x*) is Pratt's coefficient of local risk aversion at *x*. Estimation of CRRA utility functions is common in the experimental literature, and a 0.5 coefficient of relative risk aversion, implying a coefficient of local risk aversion at *x* of $(2x)^{-1}$, resembles many estimates we find there. Therefore, the CRRA assumption and this particular coefficient give a risk premium Var(*S*)/[4·E(*S*)] and, therefore, an approximate certainty equivalent CE(*S*) of lottery *S* equal to E(*S*) −Var(*S*)/[4·E(*S*)].

---

estimation context *m*23, and this roughly twice the similar calculation for the prediction context *m*0, which is 0.1170.

Therefore, let $W_m^* = \mathrm{E}(S_m) - \mathrm{Var}(S_m)/[4\cdot\mathrm{E}(S_m)] - \mathrm{E}(R_m) + \mathrm{Var}(R_m)/[4\cdot\mathrm{E}(R_m)]$, a rough and ready approximation of the difference between the certainty equivalents of the lotteries in pair *m*. Call this the "Pratt covariate:" It is an estimation-free alternative to any covariate $\hat{W}_{nm}(Y_n^{m23})$ we might use for task II (though it is based on an out-of-sample stylized fact from other researchers' estimations). It wholly ignores subject heterogeneity of risk attitudes; on the other hand, it has no sampling error. Let $\bar{y}_{nm} = (y_{nm1} + y_{nm2})/2$, subject *n*'s mean choice across the two trials of pair *m*; and let $\rho_{w^*,n}^{m0}$ be the Pearson correlation between $\bar{y}_{nm}$ and W_m^* for $m \in m0$, for subject *n*. The third row of Table 1 shows the mean and three deciles of this correlation across the sixty-eight Hey and Orme subjects: These are benchmarks for evaluating the success of any task II covariate $\hat{W}_{nm}(Y_n^{m23})$.

Although Table 1 makes it clear that we would fail miserably if we used $\bar{y}_n^{m23}$ as a point predictor of $\bar{y}_n^{m0}$, $\bar{y}_n^{m23}$ is perhaps the most obvious choice of a model-free covariate for doing task III (explaining between-subject variation). The Pearson correlation between $\bar{y}_n^{m23}$ and $\bar{y}_n^{m0}$, across the sixty-eight subjects, is $\rho_{b^*} = 0.462$. This is the benchmark against which I will compare estimation-based covariates $\hat{A}_n(Y_n^{m23})$ for task III. This implies that a linear regression of $\bar{y}_n^{m0}$ on $\bar{y}_n^{m23}$ would yield an R^2 of $(0.462)^2 = 0.21$, an uninspiring figure; yet we will see that this is a tough benchmark for estimation-based covariates $\hat{A}_n(Y_n^{m23})$.

6. The Monte Carlo study

Let H denote the null hypothesis that a certain pairing of a structure and stochastic model is true of our sample. We will examine six such pairings in detail. The two structures are EU and

RDEU, and they are paired with three stochastic models, Strong, CU and RP (strong utility, contextual utility and random preferences, respectively). For instance, H = (EU,Strong) is the null hypothesis that expected utility is the true structural model, and strong utility the true stochastic choice model, for all subjects in the sample. Later, in section 6.6, random parameters estimations will suggest that strict utility and the wandering vector model are dominated by the other stochastic models. Moreover they are not as widely used as either strong utility or random preferences. However, contextual utility will turn out to be a powerful competitor to strong utility and random preferences. Therefore, the detailed Monte Carlo study focuses on strong utility, contextual utility and random preferences. The strict utility and wandering vector models will be revisited later in section 6.6.

Any relevant Monte Carlo analysis requires a relevant distribution of structural and stochastic parameters to sample from. Therefore, the first step is to characterize parameter heterogeneity under each null H in a reasonably complete manner. For this step, we will use all of the data: To be more precise, at this step we will combine the data from the estimation and prediction contexts, and characterize parameter heterogeneity under the assumption that the null H is true across both contexts by temporarily treating all of the data as the estimation context and none of it as prediction context. This combined estimation will allow the Monte Carlo data sets and estimations on them to generate what are known as "parametric bootstrap" (Efron and Tibshirani 1998) estimates of confidence intervals of any moment in the data, any estimator or any measure of predictive or explanatory success we wish to look at, under the null H.

We begin with the assumption that each subject has a structural and stochastic parameter vector $\psi_H^n = (\beta_H^n, \eta_H^n)$, specified by null H, that is constant across the estimation and prediction contexts. We combine the data from the estimation and prediction contexts, and from this data

we estimate the joint distribution $J_H(\psi_H \mid \theta_H)$ of these vectors in the population from which our real subjects were sampled, where θ_H are the parameters of this distribution under null H. Let $\tilde{\theta}_H$ be our estimate of θ_H: Then a Hey and Orme-sized sample of "simulated subjects" can be generated by drawing sixty-eight vectors ψ_H from the estimated joint distribution $J_H(\psi_H \mid \tilde{\theta}_H)$. Those vectors will, when combined with lottery pair characteristics, the null H, and binomial random draws, create simulated choice data for the sample of simulated subjects—100 observations of y_{nmt} in the estimation context and 50 observations of y_{nmt} in the prediction context. We may then estimate any models we wish on the simulated estimation context data, and use those estimates to predict choices and/or explain variation in choices on the simulated prediction context data. By doing this on a lot of simulated samples we build a distribution, and hence confidence intervals or *p*-values, for any measure of predictive or explanatory success in the prediction context we wish to examine, under whatever null H we wish to consider. We can then compare the real explanatory and predictive success of our models (estimated on the real estimation context data) in the real prediction context data to what we expect to see on the basis of the "parametric bootstrap" confidence intervals we created through Monte Carlo sampling, estimation and prediction.

The next section illustrates the details for the null (EU,Strong); after this, the details will be skipped and the focus will be on results under the various nulls. In particular, I mostly relegate the details and results of estimations of the joint distributions $J_H(\psi_H \mid \theta_H)$ to the Appendix, where readers can find random parameters treatments of parameter heterogeneity, and resulting Hey and Orme (1994) sample log likelihoods under the various nulls. My focus in the text will be out-of-context predictive and explanatory performance at the level of the individual.

6.1 The detailed example of expected utility with strong utility

At the level of individual subjects, the full specification of the (EU,Strong) model is

(6.11) $$P_{nmt} = (1-\omega^n)\Lambda\left(\lambda^n[(s_{m1}-r_{m1})+(s_{m2}-r_{m2})u_2^n+(s_{m3}-r_{m3})u_3^n]\right)+\omega^n/2,$$

where $\Lambda(x)=[1+\exp(-x)]^{-1}$ is the Logistic c.d.f. (which will be consistently employed as the function $F(x)$ for the strong utility, strict utility, contextual utility and wandering vector models). There are no FOSD pairs in Hey and Orme's data, so we need no special gizmo as in equation 4.24 to account for them. I begin by estimating the parameters of a simplified version of equation 6.11 individually, for each of the sixty-eight subjects, using all 150 observations y_{nmt} in the combined estimation and prediction contexts.

The purpose of this initial subject-by-subject estimation is to get an initial impression of what the joint distribution of parameter vectors $\psi=(u_2,u_3,\lambda,\omega)$ looks like across subjects and how we might choose $J(\psi\mid\theta)$ to represent that distribution. At this initial step I do not actually estimate ω^n for any subject, but rather assume it is constant across subjects and equal to 0.04. There are two reasons for this, both stemming from out-of-sample lessons learned working in similar fashion with Hey's (2001) still larger data set. In that data set, where there are 125 observations on each of the four contexts, I have estimated subject-specific trembles ω^n separately on all four contexts, and find that there is no significant correlation of these subject-specific estimates of ω^n across contexts. This suggests that there is little or no reliable between-subjects variance in tremble probabilities—that $\omega^n=\omega$ for all *n*—and I will henceforth assume that this is true of the population for all models. The second reason is that estimation of ω^n is a nuisance at the individual level: Doing so substantially increases the sampling variability of estimates of the other parameters, and may therefore obscure the more important variances and

covariances we need to see at this step. Later, we will see that even when the population has a nonzero tremble parameter, trying to estimate it at the level of individual subjects greatly decreases the out-of-context predictive performance of many models.

Therefore, I begin by estimating u_2, u_3 and λ in the model

$$P_m = 0.96\Lambda\left(\lambda[(s_{m1} - r_{m1}) + (s_{m2} - r_{m2})u_2 + (s_{m3} - r_{m3})u_3]\right) + 0.02, \tag{6.12}$$

which temporarily fixes ω at 0.04, for each subject. The log likelihood function for subject n is

$$LL^n(u_2, u_3, \lambda) = \sum_{m \in \{m23 \cup m0\}} y_{nmt} \ln(P_m) + (1 - y_{nmt}) \ln(1 - P_m) \tag{6.13}$$

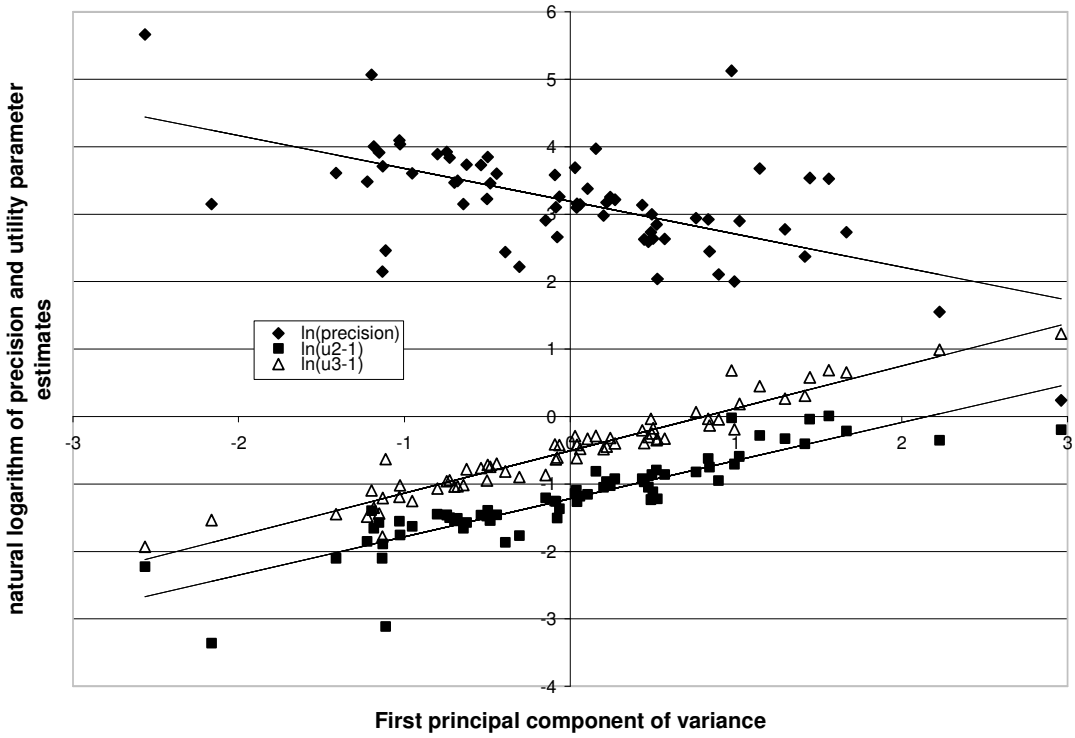
with the specification of P_m in equation 6.12. This is maximized for each subject n, yielding $\tilde{\psi}^n = (\tilde{u}_2^n, \tilde{u}_3^n, \tilde{\lambda}^n, 0.04)$, the initial estimates from the combined contexts for each subject n.

Figure 3 graphs $\ln(\tilde{u}_2^n - 1)$, $\ln(\tilde{u}_3^n - 1)$ and $\ln(\tilde{\lambda}^n)$ against their first principal component, which accounts for about 69 percent of their collective variance.[33] The figure also shows regression lines on the first principal component. The Pearson correlation between $\ln(\tilde{u}_2^n - 1)$ and $\ln(\tilde{u}_3^n - 1)$ is fairly high (0.848). Given that these are both estimates and hence contain some pure sampling error, it appears that an assumption of perfect correlation between them in the underlying population may not do too much violence to truth. Therefore, I will make that assumption about the joint distribution of ψ in the population. While $\ln(\tilde{\lambda}^n)$ does appear to share variance with $\ln(\tilde{u}_2^n - 1)$ and $\ln(\tilde{u}_3^n - 1)$ (Pearson correlations of –0.22 and –0.45, respectively), it obviously either has independent variance of its own or is estimated with relatively low precision.

These observations suggest modeling the joint distribution $J(\psi \mid \theta)$ of $\psi = (u_2, u_3, \lambda, \omega)$ as being generated by two independent standard normal deviates x_u and x_λ, as follows:

[33] Two hugely obvious outliers have been removed for both the principal components extraction and the graph.

Figure 3. Shared variance of initial individual parameter estimates in the (EU,Strong) model



(6.14) $u_2(x_u,\theta) = 1+\exp(a_2+b_2x_u)$, $u_3(x_u,\theta) = 1+\exp(a_3+b_3x_u)$,

$\lambda(x_u,x_\lambda,\theta) = \exp(a_\lambda+b_\lambda x_u+c_\lambda x_\lambda)$ and ω a constant,

where $\theta = (a_2,b_2,a_3,b_3,a_\lambda,b_\lambda,c_\lambda,\omega)$.

Then the (EU,Strong) model, conditional on x_u, x_λ and θ, becomes

(6.15) $P_m(x_u,x_\lambda,\theta) =$

$(1-\omega)\Lambda\big(\lambda(x_u,x_\lambda,\theta)[(s_{m1}-r_{m1})+(s_{m2}-r_{m2})u_2(x_u,\theta)+(s_{m3}-r_{m3})u_3(x_u,\theta)]\big)+\omega/2$.

Now, estimate θ by maximizing the following random parameters log likelihood function in θ:

(6.16) $LL(\theta) = \sum_n \ln\Big(\iint\Big(\prod_{m\in\{m0\cup m23\}} P_m(x_u,x_\lambda,\theta)^{y_{nmt}}[1-P_m(x_u,x_\lambda,\theta)]^{1-y_{nmt}}\Big)d\Phi(x_u)d\Phi(x_\lambda)\Big)$,

where Φ is the standard normal c.d.f. and $P_m(x_u, x_\lambda, \theta)$ is as given in equation 6.15.[34] This is a difficult maximization problem, but fortunately the linear regression lines in Figure 3 may provide reasonable starting values for the parameter vector θ. That is, initial estimates of the *a* and *b* coefficients in θ are the intercepts and slopes from the linear regressions of $\ln(\tilde{u}_2^n - 1)$, $\ln(\tilde{u}_3^n - 1)$ and $\ln(\tilde{\lambda}^n)$ on their first principal component; and the root mean squared error of the regression of $\ln(\tilde{\lambda}^n)$ on the first principal component provides an initial estimate of c_λ.

Table 2. Random parameters estimates of the (EU,Strong) model, using choice data from the contexts (0,1,2), (0,1,3) and (1,2,3) of the Hey and Orme (1994) sample.

Structural and stochastic parameter models	Distributional parameter	Initial estimate	Final estimate	Asymptotic standard error	Asymptotic t-statistic
$u_2 = 1 + \exp(a_2 + b_2 x_u)$	a_2	−1.22	−1.2679	0.046	−27.8
	b_2	0.57	0.5151	0.032	16.1
$u_3 = 1 + \exp(a_3 + b_3 x_u)$	a_3	−0.51	−0.6305	0.037	−16.9
	b_3	0.63	0.6266	0.034	18.5
$\lambda = \exp(a_\lambda + b_\lambda x_u + c_\lambda x_\lambda)$	a_λ	3.19	3.3615	0.11	29.5
	b_λ	−0.49	−0.5004	0.12	−4.20
	c_λ	0.66	0.6242	0.074	8.48
ω constant	ω	0.04	0.0652	0.015	4.49
Log likelihood = −4841.50					

Notes: x_u and x_λ are independent standard normal variates. Standard errors are calculated using the "sandwich estimator" (see Wooldridge 2002 pages) and treating all of each subject's choices as a single "super-observation," that is, using degrees of freedom equal to the number of subjects rather than the number of subjects times the number of choices made.

Table 2 shows the results of maximizing 6.16 in θ. As can be seen, the initial parameter estimates are good starting values, though a couple of the final estimates are significantly different from the initial estimates (judging from the asymptotic standard errors of the final

[34] Such integrations must be performed numerically in some manner for estimation. I use gauss-hermite quadratures, which are practical up to two or three integrals; for integrals of higher dimension, simulated maximum likelihood is more practical. Judd (1998) and Train (2003) are good sources for these methods.

estimates). The final parameter estimates in Table 2 are the basis for creating simulated data sets for Monte Carlo study. To create a simulated subject *k*, draw two standard normal deviates x_u and x_λ and use the final estimates $\tilde{\theta}$ of the vector θ, in conjunction with the structural and stochastic parameter models in equation 6.14 (also shown in the first column of Table 2), to transform these draws into a vector $\psi^k = (u_2^k, u_3^k, \lambda^k, \omega)$. To create simulated choice data for this simulated subject, substitute these parameters into equation 6.11 and put the lotteries (probability vectors) of all 75 pairs $m \in m0 \cup m23$ into equation 6.11 too, creating 75 simulated "true choice probabilities" P_{km}, 50 on the estimation context $m23$ and 25 on the prediction context $m0$. Finally, draw two independent binomial random outcomes, given each probability P_{km}, to create observations y_{kmt} for two trials $t = 1$ and 2 of each pair *m*, as in Hey and Orme's data. Follow these steps sixty-eight times, and we have a simulated sample equivalent in size to the original sample. Do all that one hundred times, and we have one hundred such simulated data sets, containing in all sixty-eight hundred simulated subjects.

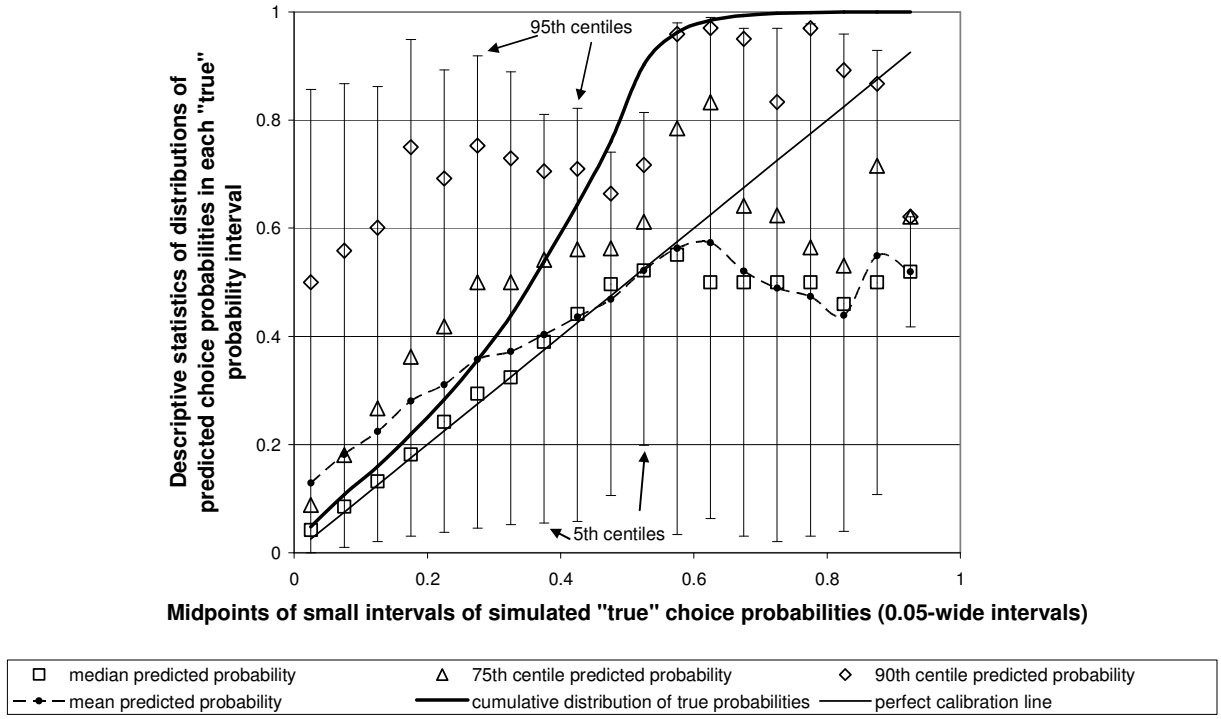
We can now study the sampling characteristics of out-of-context prediction and explanation of individual choice under risk, under the null hypothesis (EU,Strong). Return to estimation of the equation 6.11 model one subject at a time. However, now estimate the model using only the 100 observations in the estimation context $m23$, and initially let us estimate the entire parameter vector $\psi = (u_2, u_3, \lambda, \omega)$ (that is, now include estimation of the tremble probability ω). Begin by doing this for the 6800 simulated subjects. Let $\hat{\psi}^k = (\hat{u}_2^k, \hat{u}_3^k, \hat{\lambda}^k, \hat{\omega}^k)$ be the resulting parameter estimates for simulated subject *k*: These estimates may now be substituted into equation 6.11, along with the characteristics of the lottery pairs $m \in m0$, the prediction context, to get predicted out-of-context probabilities $\hat{P}_{km}$ for each simulated subject *k*.

The simulated data sets contain a "true" choice probability P_{km} in the prediction context—something we do not have in the real data set—corresponding to every predicted probability $\hat{P}_{km}$ we have just estimated. We can therefore look at the calibration of predicted probabilities to true probabilities using a calibration graph, much in the same manner we can look at the calibration of observed choice proportions to predicted probabilities. But we can do much more: We can characterize the entire distribution of predicted probabilities conditional on true probabilities. This is done in Figure 4. Define subscript sets $km(j) = \{km \mid j - 0.025 < P_{km} \leq j + 0.025, m \in m0\}$, denoting subject/pair combinations *km* in the simulated data for which true probabilities in the prediction context fall in 0.05-wide intervals around midpoints $j \in \{0.025, 0.075, 0.125 \ldots, 0.975\}$. The horizontal axis plots these midpoints *j* of small (0.05-wide) intervals of true probabilities. The vertical axis plots various descriptive statistics of the predicted out-of-context probabilities $\hat{P}_{km}$ associated with the true probabilities in each of these intervals. These statistics are the means (small black dots), medians (open squares), 90 percent confidence interval (5^{th} and 95^{th} centile whiskers), 75^{th} centiles (open triangles) and 90^{th} centiles (open diamonds). The lower distribution centiles (the 10^{th} and 25^{th} centiles) are omitted to minimize unnecessary visual clutter (the important news is mostly communicated by the upper centiles of the distributions).

The smooth and heavy black line in Figure 4 represents the cumulative distribution of "true" probabilities, which tells us what parts of the figure are of practical importance. In particular, this line tells us that only about 4% of true probabilities lie to the left of the interval *j* = 0.575. It is clear from the interval medians (open squares) that the calibration of median predicted probabilities to true probabilities is excellent up to and including the *j* = 0.575 interval, only becoming poor for true probabilities greater than 0.6; even the interval means (black dots) are not too badly calibrated up through true probabilities of about 0.6 (though they are upward-

biased and this will be of some importance). Calibration of median predicted probabilities to true probabilities is poor for true probabilities greater than 0.6, but since only 4% of true probabilities exceed 0.6, this is not of any great practical concern.

Figure 4. Calibration and resolution of predicted out-of-context choice probabilities (EU,Strong,Trembles) to "true" probabilities in simulated data sets (EU,Strong,Trembles)



The important news in Figure 4, then, is not news of any bias of median predicted probabilities conditional on true probabilities. Rather, it is news of high variance and long, fat upside tails of predicted probabilities conditional on true probabilities. The 95th, 90th and 75th centiles (upper whiskers, open diamonds and open triangles) make this phenomenon crystal clear. The true probability interval around $j = 0.275$ provides a revealing example: Fully one quarter of the predicted probabilities in this interval are 0.5 or larger, and ten percent of them exceed 0.75. If results like this were limited to a handful of intervals or confined to relatively rare true probability intervals, they would not be very important. But Figure 4 makes it very clear

that long, fat upside tails of predicted probabilities are the rule, not the exception: The smooth black line representing the c.d.f. of true probabilities shows that this occurs heavily over the densest parts of the distribution of true probabilities.

There is terminology for this phenomenon: We say that predicted probabilities are poorly resolved or have poor resolution. One may have decent calibration (a good match of median or mean predicted probabilities to underlying true probabilities in an interval) but still have poor resolution (wide variance of predicted probabilities conditional on true probabilities), and this is what we see in Figure 4. Because most of the distributional mass of true probabilities is on low probabilities, the poor resolution mostly takes the form of long and fat upside tails of predicted probabilities conditional on true probabilities. This translates into a general upward bias in mean predicted probabilities (even though median predicted probabilities are remarkably well-calibrated). It also implies that a disproportionate number of predicted probabilities will be upward-biased misclassifications of the true probabilities of safe choices. Such common misclassification is the essence of poor resolution. This is why the introductory Figure 1 looks like it does: There are way too many predicted probabilities greater than 0.5, relative to the distribution of true probabilities; these are in fact disproportionately upward-biased misclassifications of true probabilities; and therefore, the observed sample proportions of choices of safe lotteries in those higher intervals of predicted probabilities fall far short of the midpoints of those intervals.

The out-of-context predicted probability intervals and actual sample proportions of safe choices in each interval shown in Figure 1 come from the actual Hey and Orme (1994) data. The equation 6.11 model was estimated for each of the actual subjects *n* in Hey and Orme's data, using the 100 observations in the estimation context *m*23, producing estimated parameter vectors

$\hat{\psi}^n = (\hat{u}_2^n, \hat{u}_3^n, \hat{\lambda}^n, \hat{\omega}^n)$. These were substituted into equation 6.11, along with the characteristics of the lottery pairs $m \in m0$ (the prediction context) to get predicted out-of-context probabilities $\hat{P}_{nm}$ for each subject *n* in each pair $m \in m0$. Subscript sets $nm(\hat{j}) = \{nm \mid \hat{j} - 0.025 < \hat{P}_{nm} \leq \hat{j} + 0.025\}$ contain subject/pair combinations *nm* for which the predicted probabilities fall into 0.05-wide intervals around the midpoints $\hat{j} \in \{0.025, 0.075, 0.125\ldots, 0.975\}$; and the actual proportions of safe choices for predicted probabilities in each of those intervals is $\overline{\overline{y}}_{\hat{j}} = \sum_{nm \in nm(\hat{j})} y_{nmt} / \#nm(\hat{j})$. By doing exactly the same thing with each of the one hundred simulated data sets, we can create simulated confidence intervals for each $\overline{\overline{y}}_{\hat{j}}$ in each interval $\hat{j}$. For instance, if we want a 90 percent confidence interval, we simply take the 5th highest and 5th lowest values of $\overline{\overline{y}}_{\hat{j}}$ from the 100 values of $\overline{\overline{y}}_{\hat{j}}$ computed in the 100 simulated data sets.[35]

Figure 5 repeats Figure 1, but superimposes these 90 percent confidence intervals for the observed choice proportions in each interval. Additionally, the cumulative distribution of simulated true probabilities is added as a smooth grey line, for comparison with the smooth black line representing the cumulative distribution of predicted out-of-context probabilities in the actual Hey and Orme data. By doing this, we can clearly see how the poor resolution of predicted probabilities to true probabilities, shown in Figure 4, becomes poor calibration on the right half of Figure 1 and Figure 5. The 90 percent confidence intervals on the right half of Figure 5 show that this is an expected consequence of maximum likelihood-based estimation and prediction at this sample size under the (EU,Strong) null, given this particular out-of-context prediction

[35] In fact these are only estimates of the true 5th and 95th centiles; using 500 or 1000 simulated data sets will result in better estimates. Since I am mostly concerned with the general shape and pattern of a large number of interrelated confidence intervals, rather than their precise width, I am content with the results of 100 simulated samples under each null, especially because of the large number of nulls I will consider.

situation. Both the progressive overprediction (true choice proportions below predicted probabilities) and the erratic choice proportions on the right side are mirrored by the flattening of the path taken by median simulated choice proportions and the widening of the simulated confidence intervals around them. The cumulative distribution of predicted probabilities is well to the right of the cumulative distribution of simulated true probabilities, illustrating how the long, fat upside tails of predicted probabilities conditional on true probabilities results in too many high predicted probabilities.

Figure 5. Calibration of actual choice proportions to out-of-context predictions (EU,Strong,Trembles) in Hey and Orme (1994), with 90% confidence intervals estimated from simulated data sets (EU,Strong,Trembles).

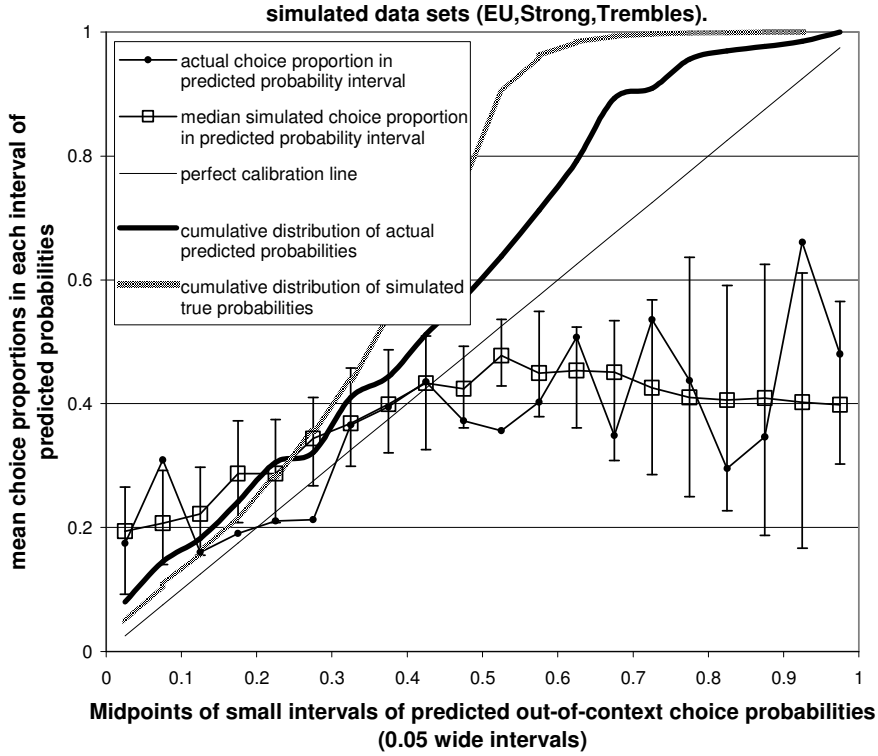


Table 3 shows the mean sample value, and the value in the 10th, 50th and 90th centile samples, of a collection of statistics based on "true" and predicted out-of-context probabilities P_{km} and $\hat{P}_{km}$, and observations y_{kmt}, in the simulated data. Corresponding values of the same statistics calculated using the predicted out-of-context probabilities $\hat{P}_{nm}$ and observations y_{nmt} in

the actual Hey and Orme data appear at the far right. Here are definitions of all the statistics reported on rows of the table. All definitions below are couched in terms of simulated subjects *k* and true probabilities P_{km}; substitute *n* for *k* (for actual Hey and Orme subjects) and/or $\hat{P}_{km}$ for P_{km} (for predicted rather than true probabilities) as appropriate.

$\overline{P}_{c10}^{m0}$, $\overline{P}_{c50}^{m0}$, $\overline{P}_{c90}^{m0}$ and $\overline{\overline{P}}^{m0}$: Let $\overline{P}_{k}^{m0} = \sum_{m \in m0} P_{km} / (\#m0)$ be the mean probability for subject *k* over pairs $m \in m0$. These are the values of $\overline{P}_{k}^{m0}$ for the 10th, 50th and 90th centile subject, and its mean across subjects, in a sample.

ll: The log likelihood $\sum_{k \in sample} \sum_{m \in m0} y_{kmt} \ln(P_{km}) + (1 - y_{kmt})[\ln(1 - P_{km})]$ in a sample, in the prediction context. A measure of task I (prediction) performance.

sse: The sum of squared errors $\sum_{k \in sample} \sum_{m \in m0} (P_{km} - y_{kmt})^2$ in a sample, in the prediction context. An alternative measure of task I (prediction) performance.

$\rho_{w,c10}^{m0}$, $\rho_{w,c50}^{m0}$, $\rho_{w,c90}^{m0}$ and $\rho_{w,mean}^{m0}$: Let $\rho_{w,k}^{m0}$ be the correlation between $\overline{y}_{km} = (y_{km1} + y_{km2})/2$ and P_{km} over pairs $m \in m0$, for subject *k*. These are the values of $\rho_{w,k}^{m0}$ for the 10th, 50th and 90th centile subject and its mean over subjects in a sample. These are task II (explaining within-subject, between-pair variation) performance measures, using true or estimated probabilities as the covariate.

$\rho^{m0}_{w*,c10}$, $\rho^{m0}_{w*,c50}$, $\rho^{m0}_{w*,c90}$ and $\rho^{m0}_{w*,mean}$: Same as above, except based on $\rho^{m0}_{w*,k}$, the correlation between $\bar{y}_{km} = (y_{km1} + y_{km2})/2$ and the "benchmark" estimation-free Pratt covariate $W^*_m =$ E(S_m) −Var(S_m)/[4·E(S_m)] − E(R_m) + Var(R_m)/[4·E(R_m)] for task II, discussed in section 5.

ρ_{bP}: Let $\bar{y}^{m0}_k = \sum_{m \in m0} y_{kmt}/(\#m0)$ be the proportion of safe choices made by subject k in the prediction context. This is the correlation between $\bar{y}^{m0}_k$ and $\bar{P}^{m0}_k$ in a sample. A measure of task III (explaining between-subject variation) performance using mean true or predicted probabilities as the covariate.

ρ_{b*}: Let $\bar{y}^{m23}_k = \sum_{m \in m23} y_{kmt}/(\#m0)$ be the proportion of safe choices made by subject k in the estimation context, that is for $m \in m23$. This is the correlation between $\bar{y}^{m0}_k$ and $\bar{y}^{m23}_k$ in a sample: It is the performance of a "benchmark" estimation-free covariate for task III, as discussed in section 5.

ρ_P: The overall correlation between $\bar{y}_{km}$ and P_{km}, pooling over all pairs $m \in m0$ and all subjects k in a sample.

Table 3 shows the distributions of these statistics across simulated samples, based on the "true" simulated probabilities (first group of four columns), and then on the predicted out-of-context probabilities (second group of four columns). The first group of four columns, based on the "true" simulated probabilities, are provided to give upper bounds on performance: These can be thought of as asymptotic values of the predicted probability statistics (second group of four

columns) as the number of observations in the estimation context gets very large and predicted probabilities converge to true probabilities. The second group of four columns provide us with expectations for out-of-context prediction and explanation results in the actual Hey and Orme data (the far right column). In particular the sixth and eighth columns, which are the 10^{th} and 90^{th} centiles of simulated sample values of each statistic, are bootstrapped 80% confidence intervals for statistics in the far right column. I will refer to these as "simulated $10^{th}/90^{th}$ ranges" below.

The first four rows (1 to 4) affirm two obvious things. First, the distribution of out-of-context predicted probabilities will be more diffuse than the distribution of true probabilities is; mean predicted out-of-context choice probabilities will, for instance, be more variable than they truly are in the population. This is simply because estimates $\hat{\psi}^k$ contain sampling variance not present in the true ψ^k. Second, there will be an overall upward finite sample bias of the predicted probabilities, in accord with the discussion of Figures 4 and 5. This is because true probabilities in the prediction context are mostly less than 0.5 and because of the long, fat upside tails of predicted probabilities relative to true probabilities. The far right column shows that the distribution of predicted out-of-context probabilities across subjects in the actual Hey and Orme data set are a bit on the high side of what we would expect, judging by the simulated $10^{th}/90^{th}$ ranges. Recall from Table 1 that the actual proportion of safe choices in the estimation context is 0.343; the (EU,Strong) model predicts 0.42 in the Hey and Orme data set. We will see that this overprediction is common to many of the models, and unacceptably large for a few of them.

The next two rows (5 and 6), those displaying the log likelihoods and sums of squared errors, show us how far our predictive performance will be from what is asymptotically possible. They also indicate that the variability of out-of-context predictive performance will be much larger in the finite samples than is the underlying variability of fit given the true probabilities.

Table 3. Various statistics constructed from true and predicted probabilities in simulated (EU,Strong,Trembles) data sets and the Hey and Orme data. Predicted probabilities based on (EU,Strong,Trembles) model.

Statistic		Using true out-of-context probabilities P_{kn} in the 100 simulated samples				Using predicted out-of-context probabilities $\hat{P}_{kn}$ …				
						…in the 100 simulated samples				…in the actual Hey and Orme sample
		mean over samples	*10^{th} centile of samples*	*50^{th} centile of samples*	*90^{th} centile of samples*	*mean over samples*	*10^{th} centile of samples*	*50^{th} centile of samples*	*90^{th} centile of samples*	
1	$\overline{P}_{c10}^{m0}$	0.21	0.18	0.20	0.24	0.14	0.10	0.15	0.18	0.168
2	$\overline{P}_{c50}^{m0}$	0.36	0.34	0.36	0.38	0.35	0.31	0.34	0.39	0.390
3	$\overline{P}_{c90}^{m0}$	0.50	0.47	0.50	0.52	0.72	0.64	0.72	0.82	0.687
4	$\overline{\overline{P}}^{m0}$	0.36	0.34	0.36	0.38	0.39	0.36	0.39	0.42	0.424
5	*ll*	–1981.91	–2045.57	–1983.61	–1915.69	–2711.29	–3048.12	–2667.7	–2444.52	–2491.30
6	*sse*	170.94	162.54	170.63	179.92	265.07	242.43	263.45	290.24	275.83
7	$\rho_{w,c10}^{m0}$	0.10	0.04	0.10	0.16	–0.03	–0.10	–0.02	0.01	0.00
8	$\rho_{w,c50}^{m0}$	0.41	0.36	0.40	0.44	0.33	0.28	0.32	0.37	0.362
9	$\rho_{w,c90}^{m0}$	0.67	0.63	0.66	0.71	0.62	0.58	0.62	0.67	0.643
10	$\rho_{w,mean}^{m0}$	0.39	0.36	0.39	0.43	0.30	0.27	0.30	0.35	0.351
11	$\rho_{w^*,c10}^{m0}$	0.10	0.04	0.10	0.17	0.10	0.04	0.10	0.17	0.197
12	$\rho_{w^*,c50}^{m0}$	0.38	0.34	0.38	0.41	0.38	0.34	0.38	0.41	0.380
13	$\rho_{w^*,c90}^{m0}$	0.56	0.53	0.55	0.59	0.56	0.53	0.55	0.59	0.595
14	$\rho_{w^*,mean}^{m0}$	0.35	0.32	0.35	0.38	0.35	0.32	0.35	0.38	0.387
15	ρ_{bP}	0.82	0.76	0.82	0.87	0.27	0.10	0.28	0.41	0.172
16	ρ_{b^*}	0.48	0.33	0.49	0.61	0.48	0.33	0.49	0.61	0.462
17	ρ_{P}	0.47	0.44	0.47	0.50	0.22	0.17	0.21	0.28	0.227

These out-of-context fit statistics in the actual Hey and Orme data set (far right column) are well within the simulated 10^{th}/90^{th} ranges.

The next eight rows (7 to 14) report some interesting news. Here, we are looking at task II performance, the explanation of within-subject, between-pair variance. Concentrate on a comparison of rows 10 and 14, which report $\rho^{m0}_{w,mean}$ and $\rho^{m0}_{w*,mean}$, respectively. The first four columns in these rows show that "true" choice probabilities in the prediction context—the P_{km} in the simulated data sets that actually generated choice draws y_{nmt}—would (on average) be a better covariate for this task than the estimation-free "Pratt" covariate suggested in section 5, as indeed they must be. The proper way to interpret this is as follows: If we had a very large number of observations in the estimation context, so that estimated parameters, and hence estimated prediction context probabilities $\hat{P}_{nm}$, had converged to true parameters and true prediction context probabilities, we could do better in task II than the estimation-free Pratt covariate does. But not much: Notice that the difference between the correlations $\rho^{m0}_{w,mean}$ and $\rho^{m0}_{w*,mean}$, although positive, never exceeds 0.05 on the left side of Table 3.

Of course, we never observe the true probabilities in practice: We must estimate them, and our estimation context sample is finite. When we move to the second group of four columns, the comparison is reversed: Predicted out-of-context probabilities $\hat{P}_{km}$ are actually expected to perform somewhat worse as a task II covariate than the seat-of-the-pants Pratt covariate, though the difference is a small 0.03 to 0.05, depending on the centiles of simulated samples compared. And in fact, this same reversal occurs in the actual Hey and Orme data, where the mean value of $\rho^{m0}_{w,n}$ over subjects *n* is 0.351 while the mean value of $\rho^{m0}_{w*,n}$ over subjects *n* is 0.387. Most of

these task II correlations in the Hey and Orme data are on the upper edge of their simulated $10^{th}/90^{th}$ ranges.

The next two rows (15 and 16) are the headline news. Looking first at the first four columns and row 15, we can clearly see that the true probabilities would be a fine covariate for task III (explaining between-subject variation in mean choice proportions). This must be so, since the only thing keeping this correlation below 1 is the finite sample size of the prediction context. Comparison with row 16 shows that the observed choice proportions in the estimation context don't even come close. Yet when we move to the next four columns, we see that this is decisively reversed: When we only have predicted out-of-context probabilities, their mean is clearly inferior to observed choice proportions in the estimation context as a task III covariate. The explanatory gap between the true and predicted probabilities is really quite shocking. The mean correlation 0.82 between mean true probabilities and choice proportions in the prediction context implies an $R^2 = (0.82)^2 = 0.67$ in a regression of the latter on the former. By contrast, the mean correlation 0.27 between predicted out-of-context probabilities and choice proportions in the prediction context implies an $R^2 = (0.27)^2 = 0.073$ in a regression of the latter on the former. The near order-of-magnitude shrinkage in variance explained here is astonishing and sobering. In any case, the right column shows the same reversal in the actual Hey and Orme data, where we have $\rho_{bP} = 0.172$ and $\rho_{b^*} = 0.462$, both well within their simulated $10^{th}/90^{th}$ ranges: This dramatic shrinkage is <u>expected</u>.

The message coming from this analysis is very clear. For point prediction in new contexts, we have little recourse other than the estimation of models and prediction based on the estimates. But we do have choices about covariates for explaining variation in cross-sections, whether the cross-section in question is choices in varying pairs for a given subject, or average choices over

pairs for varying subjects. Model estimation is not necessarily the best strategy for these explanatory tasks: In both of the cases examined here, simple estimation-free covariates can do a better job of explaining variation than model estimates can. For our models of stochastic choice under risk, this seems to be the case even when we have 100 observations per subject for estimating those models. The highly nonlinear nature of these models, the discrete nature of the data and the small sample size probably conspire to give these estimations statistical properties that are quite distant from those promised by the consistency of ML estimation.

Perhaps demographic, personality and cognitive covariates can help to improve a model-based covariate for explaining between-subjects variation. I strongly suspect this is true, and believe we should pursue this possibility. One model-based approach would be to use a "conditional random parameters" characterization: We estimate a joint distribution $J(\psi \mid \theta, \ , X)$, where X^n is a vector of demographic, personality and/or cognitive measures for subject *n*, and is a vector of parameters describing the conditional dependence of ψ on X^n in the population. With estimates $\tilde{\theta}$ and $\tilde{}$ in hand based only on estimation context choices and subject covariate vectors X^n, we may then recover a conditional estimate $\hat{\psi}^n(\tilde{\theta}, \tilde{}, Y_n^{m23}, X^n)$ for any subject *n* by using Bayes' Theorem as in Moffatt (2005). This in turn may be used to construct predicted out-of-context probabilities for subject *n* that use the additional information about subject *n* in X^n. Call these estimates $\hat{P}_{nm}(X^n)$, which could be averaged over the pairs of the prediction context to yield a conditional model-based regressor for explaining cross-sectional variation in $\bar{y}_n^{m0}$.

Yet this sort of estimation begs its own questions, similar to those treated here. For instance, we are always entitled to ask similar questions about the relative performance of less theory-

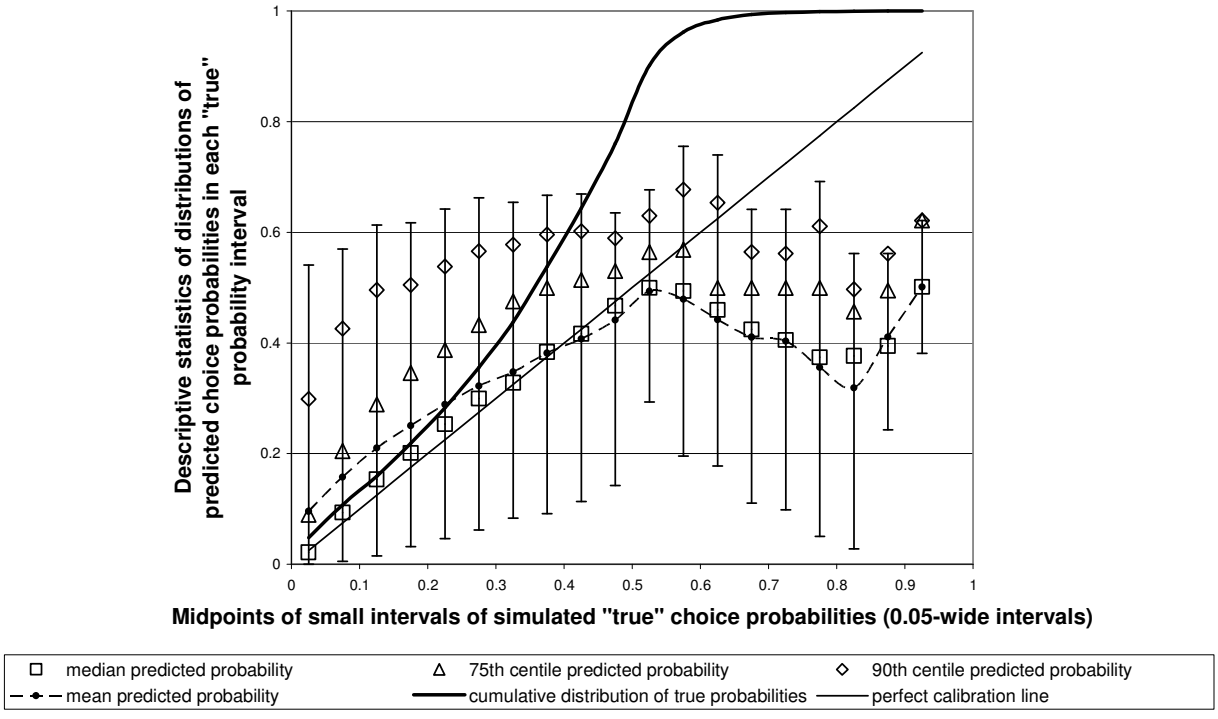
laden "benchmark" approaches that uses the same (or less) data. Does such a complex model-based approach to explaining variation in $\bar{y}_n^{m0}$ work any better than the relatively atheoretic approach where we simply regress $\bar{y}_n^{m0}$ on $\bar{y}_n^{m23}$ and X^n? Too, we add an extra layer of potential specification error when we include covariates in a model, since we will usually be taking a position on the parametric form taken by the conditional dependence of *J* on X^n and .[36] Nothing new there: We already have potential for specification errors (in the choice of structure *V*, stochastic model *P*, their parameterizations and the form of the joint distribution *J*), but we do add to this potential when we parametrically condition on X^n. That may be worth the candle, even under minor specification errors. But we don't really know right now whether it is or not, especially relative to regressing $\bar{y}_n^{m0}$ on $\bar{y}_n^{m23}$ and X^n in the case of task II (explaining between-subject variation), but even in the case of task I (prediction). And finally, suppose we include covariates in a model-based fashion and the results are "disappointing" or are "unstable" across samples. What exactly do we make of that? My over-arching theme is that, given existing sample sizes, we should expect such disappointments to be frequent. It will be nearly impossible to tell what to make of such "disappointments" or "instabilities" without careful Monte Carlo study.

The comparisons between the left and right sides of Table 3 make it plain that finite-sample estimators are quite noisy at the individual subject level. Therefore, it is quite possible that for out-of-context prediction at the sample sizes we typically have, "lean and mean" models—even misspecified lean and mean models—may perform better than their correctly specified counterparts. The tremble probability is a prime candidate for parametric liposuction.

[36] There is of course the possibility of less parametric approaches here and elsewhere. Given the size of samples we have, or can likely get within the limits of subjects' attention and patience, I doubt that data-hungry nonparametric and semiparametric approaches are worth thinking about. And they can be very data-hungry, relative to parametric alternatives: For a classic Monte Carlo demonstration of the poverty of a nonparametric approach relative to an asymptotically equivalent parametric one in any practical-sized finite sample, see Schwert (1989).

Remembering that the data-generating process used to create the simulated (EU,Strong) data sets actually contains a tremble probability $\omega = 0.0652$, let us nevertheless estimate an (EU,Strong) model stripped of the tremble probability on that simulated data and the Hey and Orme data, producing a restricted parameter vector estimate $\widehat{\psi}^k = (\widehat{u}_2^k, \widehat{u}_3^k, \widehat{\lambda}^k, 0)$ from estimation context observations, *n* replacing *k* for the actual Hey and Orme data. Figures 6 and 7 are identical to Figures 4 and 5 except that the predicted out-of-context probabilities $\widehat{P}_{km}$ and $\widehat{P}_{nm}$ used to construct them are based on these restricted estimates $\widehat{\psi}^k$ and $\widehat{\psi}^n$, respectively.

Figure 6. Calibration and resolution of predicted out-of-context choice probabilities (EU,Strong,No Trembles) to "true" probabilities in simulated data sets (EU,Strong,Trembles)



The differences between Figures 4 and 6 are obvious and dramatic. Predicted out-of-context probabilities $\widehat{P}_{km}$ are much better resolved than are their correctly specified counterparts $\hat{P}_{km}$. A close comparison of the figures reveals that this comes at a small cost: The calibration of median

values of $\widehat{P}_{km}$ to true P_{km} is slightly poorer than the calibration of median values of $\hat{P}_{km}$ to true P_{km} in the left half of the figures, and begins to seriously break down as we move rightward a little bit sooner in Figure 6 than in Figure 4. Figures 5 and 7 show that in an overall sense, the distribution of predicted out-of-context probabilities is much closer to the distribution of simulated true probabilities for the misspecified $\widehat{P}_{km}$ than the correctly specified $\hat{P}_{km}$ (compare the distance between the common smooth grey curve and the black lines across the figures). The smooth black line in Figure 7 indicates that about 22% of the $\widehat{P}_{km}$ lie in the left-side region of poor simulated median calibration (here, predicted probabilities greater than about 0.55), compared to about 29% of the $\hat{P}_{km}$ lying in the similar region of Figure 5 (predicted probabilities greater than about 0.60).

Figure 7. Calibration of actual choice proportions to out-of-context predictions (EU,Strong,No Trembles) in Hey and Orme (1994), with 90% confidence intervals estimated from simulated data (EU,Strong,Trembles)

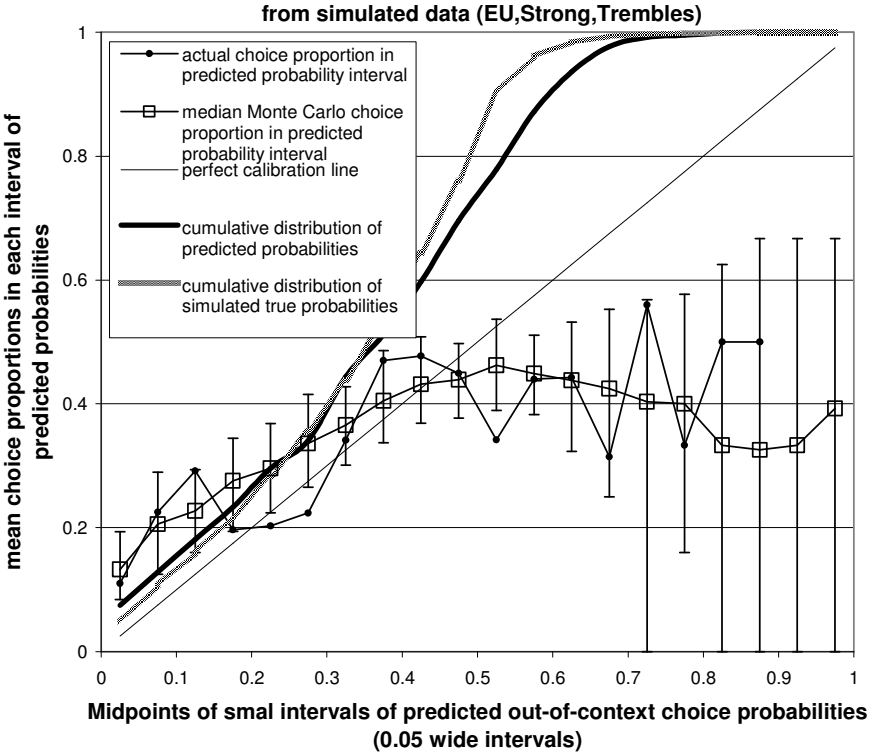


Table 4. Statistics constructed from correctly specified models (with trembles) and misspecified models (no trembles) of predicted out-of-context probabilities in simulated data sets (EU,Strong,Trembles), with the same comparison in the actual Hey and Orme data.

		Correctly specified (with trembles) predicted out-of-context probabilities $\hat{P}_{kn}$ in simulated samples				Misspecified (no trembles) predicted out-of-context probabilities $\widehat{P}_{kn}$ in simulated samples				Actual Hey and Orme sample	
		mean over samples	*10th cent. of samples*	*50th cent. of samples*	*90th cent. of samples*	*mean over samples*	*10th cent. of samples*	*50th cent. of samples*	*90th cent. of samples*	Tremble estimated ($\hat{P}_{nm}$)	No tremble ($\widehat{P}_{nm}$)
1	$\overline{P}^{m0}_{c10}$	0.14	0.10	0.15	0.18	0.17	0.14	0.17	0.19	0.168	0.179
2	$\overline{P}^{m0}_{c50}$	0.35	0.31	0.34	0.39	0.33	0.30	0.33	0.37	0.390	0.354
3	$\overline{P}^{m0}_{c90}$	0.72	0.64	0.72	0.82	0.57	0.53	0.57	0.62	0.687	0.593
4	$\overline{\overline{P}}^{m0}$	0.39	0.36	0.39	0.42	0.36	0.34	0.36	0.38	0.424	0.374
5	*ll*	–2711.29	–3048.12	–2667.7	–2444.52	–2413.06	–2804.13	–2326.38	–2210.64	–2491.30	–2265.83
6	*sse*	265.07	242.43	263.45	290.24	225.59	211.27	225.24	243.45	275.83	238.15
7	$\rho^{m0}_{w,c10}$	–0.03	–0.10	–0.02	0.01	–0.01	–0.08	0.00	0.05	0.00	0.149
8	$\rho^{m0}_{w,c50}$	0.33	0.28	0.32	0.37	0.36	0.31	0.36	0.41	0.362	0.405
9	$\rho^{m0}_{w,c90}$	0.62	0.58	0.62	0.67	0.63	0.59	0.63	0.67	0.643	0.641
10	$\rho^{m0}_{w,mean}$	0.30	0.27	0.30	0.35	0.33	0.30	0.33	0.37	0.351	0.400
11	$\rho^{m0}_{w*,c10}$	0.10	0.04	0.10	0.17	0.10	0.04	0.10	0.17	0.197	0.197
12	$\rho^{m0}_{w*,c50}$	0.38	0.34	0.38	0.41	0.38	0.34	0.38	0.41	0.380	0.380
13	$\rho^{m0}_{w*,c90}$	0.56	0.53	0.55	0.59	0.56	0.53	0.55	0.59	0.595	0.595
14	$\rho^{m0}_{w*,mean}$	0.35	0.32	0.35	0.38	0.35	0.32	0.35	0.38	0.387	0.387
15	ρ_{bP}	0.27	0.10	0.28	0.41	0.25	0.08	0.26	0.35	0.172	0.121
16	ρ_{b*}	0.48	0.33	0.49	0.61	0.48	0.33	0.49	0.61	0.462	0.462
17	ρ_{P}	0.22	0.17	0.21	0.28	0.25	0.20	0.26	0.29	0.227	0.250

Table 4 shows that the improvement in resolution that comes from estimating one fewer parameters pays for the small decline in median calibration and more. The first group of four columns in Table 4 are a simple copy of the second group of four columns in Table 3—the set of out-of-context prediction and explanation statistics generated by the correctly specified $\hat{P}_{km}$. The second group of four columns reports the same statistics, but using the misspecified $\widehat{P}_{km}$. The two right-most columns report the same statistics using the actual Hey and Orme data, with the statistics based on $\widehat{P}_{km}$ reported in the far right column. Comparison of these two right-most columns against one another, as well as with Table 1, show that the actual out-of-context performance of estimates without trembles beats performance with trembles in almost every way. The only exception is task III (explaining between-subject variance); but since both of these performances are easily bested by the estimation-free covariate $\overline{y}_k^{m23}$, it is hardly a relevant criterion for choosing between them. Particularly in the case of statistics of out-of-context fit, that is task I (point prediction) performance, the (EU,Strong) model without trembles improves quite dramatically on the (EU,Strong) model with trembles. Comparison of the first and second groups of four columns shows that even when the true data-generating process contains a nonzero tremble probability, this superiority of the misspecified tremble-free model for out-of-context prediction is expected in most cases.

6.2 A comparison of models on out-of-context point prediction success

Table 5 gathers together out-of-context fit statistics (sum of log likelihoods *ll* and sum of squared errors *sse*) for the various models. These are our measures of task I (point prediction) performance. The first four columns show distributions of "true fits" in the prediction context,

obtained by calculating these fits using the "true" probabilities in the simulated samples.[37] Notice that there is a good deal of overlap in the ranges of these "true" log likelihoods and sums of squared errors across all of the models. This is not an intentional result, but it is a convenient and important fact to bear in mind. If we observe large differences in the simulated out-of-context predictive performance of the models according to these two criteria, we cannot really attribute it to large differences in their asymptotic fit performance (since the "true fits" are the asymptotic fits): It must instead be because the finite sample behavior of estimates of the various models are very different. Having said this, the "true fits" of the contextual utility stochastic model, whether combined with the EU or the RDEU structure, are estimated to have a somewhat better distribution than other models do.

Yet this difference appears small and uninteresting when compared to differences between the models in their finite-sample out-of-context predictive performance. The left side of Table 5 is composed of two groups of three columns each. The first group reports results obtained when models are estimated with a tremble parameter, while the second (right-most) group reports results obtained when estimation proceeds without a tremble parameter (the tremble probability is constrained to be zero). Within each of these groups of three columns, the first two columns report the simulated $10^{th}/90^{th}$ range for the fit statistics, while the third column reports the fit statistics in the actual Hey and Orme data. In both cases, of course, these are fits in the prediction context using parameters estimated in the estimation context.

Consider first the results in the actual Hey and Orme data. There, (EU,Strong) is the point prediction winner amongst all models—whether the estimation uses trembles or not—according to both the *ll* and *sse* criteria. As pointed out previously, the leaner (EU,Strong) version without

[37] Refer to the Appendix for details of the random parameters estimation of the joint distributions of model parameters that are used to create simulated samples under each null.

trembles is quite a bit better than the fat version with them: This result extends to all strong utility and contextual utility results, whether combined with EU or RDEU, for both the *ll* and *sse* criteria. The opposite usually holds for random preference models: Their point prediction success is improved by estimating a tremble parameter. The simulated $10^{th}/90^{th}$ ranges show that this pattern of results is expected for strong utility and random preferences, but not for contextual utility (which, according to the simulated confidence intervals, might be expected to actually fit somewhat worse when estimated without a tremble parameter).

What really jumps out from Table 5, however, is the relatively awful out-of-context predictive success of the contextual utility models and especially random preference models—whether they are estimated with trembles or not—in finite samples. This is true both of the fit statistics using the actual Hey and Orme data and of the Monte Carlo $10^{th}/90^{th}$ ranges of the fit statistics emerging from the simulated data. Again, this is <u>not</u> because either contextual utility or random preference models actually do fit worse with "true" (asymptotic) probabilities. The right side of Table 5 shows us this. For instance, in the simulated data, the (EU,RP) model's true simulated $10^{th}/90^{th}$ range of asymptotic log likelihoods is [−2047.08,−1913.30], and we have a nearly identical range for the (EU,Strong) model, namely [−2045.57,−1915.69]. Therefore, if we had an arbitrarily large collection of lottery choices in the estimation context, and so could expect that our parameter estimates had essentially converged to the true parameters, predicted probabilities would equal true probabilities in the prediction context, and we could expect to observe log likelihoods in this essentially common range regardless of which model we estimate (when that model is the true model). With just 100 observations in the estimation context, however, the $10^{th}/90^{th}$ range for the (EU,RP) model estimated with trembles is [−5163.67,−3247.51] while the same range for the (EU,Strong) model estimated with trembles is

[−3048.12,−2444.52]. Since both of these ranges were generated by estimating each model on simulated data where each model was the true data generating process, and since their asymptotic fit ranges are nearly identical, a conclusion is forced on us: The small sample behavior of (this version of) the random preference model, and the contextual utility model, for out-of-context prediction tasks at the level of individuals, is very much poorer than the small sample behavior of the strong utility model.

What accounts for these differences? In the case of random preferences, Section 4.1 noted that the independent gamma version of the random preference model might well find this particular out-of-context prediction problem relatively difficult. The reason for this is that the model's scale parameter κ, which is empirically relevant in the estimation context, is irrelevant in the prediction context, so that there is a sense in which some of the sample information in the estimation context—whatever is embodied in an estimate of κ—gets "dropped" when estimated parameters are taken into the prediction context. Neither the strong utility nor the contextual utility model have this property; all three of their parameters estimated in the estimation context (u_2,u_3 and λ) are put to work in the prediction context. As an asymptotic matter, "dropping" κ should be irrelevant to the performance of random preferences in the prediction context (when it is the true model). In a finite sample, however, the explanatory roles of parameters have not been fully separated from one another, and this appears to be very important for random preferences.

Table 5. Point prediction (task I) performance: Out-of-context measures of fit under various nulls.

			Distribution across samples using simulated "true" probabilities				Simulated 10th/90th ranges and actual Hey and Orme sample results using predicted out-of-context probabilities					
							With trembles			Without Trembles		
Struc-tural model	Sto-chastic model	Stat	Mean over samples	10th cent. of samples	50th cent. of samples	90th cent. of samples	simulated 10th/90th range		Actual Hey and Orme	simulated 10th/90th range		Actual Hey and Orme
							10th cent.	90th cent.		10th cent.	90th cent.	
EU	Strong Util.	*ll*	−1981.91	−2045.57	−1983.61	−1915.69	−3048.12	−2444.52	−2491.30	−2804.13	−2210.64	−2265.83
		sse	170.94	162.54	170.63	179.92	242.43	290.24	275.83	211.27	243.45	238.15
	Context. Util.	*ll*	−1785.59	−1844.95	−1790.36	−1717.24	−3857.17	−2969.04	−3369.24	−4109.59	−3043.02	−3327.32
		sse	150.25	141.35	150.96	159.01	296.16	361.03	354.66	281.69	348.59	322.57
	Random Prefs	*ll*	−1982.17	−2047.08	−1978.21	−1913.3	−5163.67	−3247.51	−4776.81	−8467.37	−4933.93	−6744.99
		sse	171.14	162.6	170.44	179.38	272.16	347.00	394.51	292.62	369.93	381.44
RDEU	Strong	*ll*	−1978.58	−2031.47	−1977.82	−1920.98	−3139.61	−2441.03	−2667.19	−2985.02	−2196.65	−2308.51
		sse	170.78	163.24	170.91	179.25	235.97	284.52	291.94	207.65	240.91	243.13
	Context. Util.	*ll*	−1747.13	−1808.25	−1750.91	−1679.25	−4269.86	−3052.53	−3893.93	−5102.52	−3101.46	−3627.16
		sse	146.7	138.61	146.07	156.56	294.74	380.50	388.65	280.88	355.86	333.35
	Random Prefs	*ll*	−1985.44	−2038.43	−1985.4	−1936.73	−4768.38	−2892.42	−5837.11	−6254.47	−3714.78	−7192.01
		sse	170.28	162.8	170.32	176.9	259.93	316.07	379.40	254.53	314.04	390.44

6.3 A methodological digression

There is an important methodological corollary of these findings. Comparing these hypotheses (say strong utility versus random preferences) by comparing individually estimated out-of-context fit can be an econometric blunder in small enough samples, and is indeed a blunder for a design such as Hey and Orme (1994). This is because the different models can have asymptotically similar out-of-context fit (that is, fit in the prediction context using true probabilities, as in the left side of Table 5), but their out-of-context fits can nevertheless converge to those asymptotically similar fits at very, very different rates. The large-looking differences on the right side of Table 5 are overwhelmingly due to strong differences in the finite sample behavior of estimators of these models that are quite unrelated to the very small differences we would observe with "very large" estimation context samples.

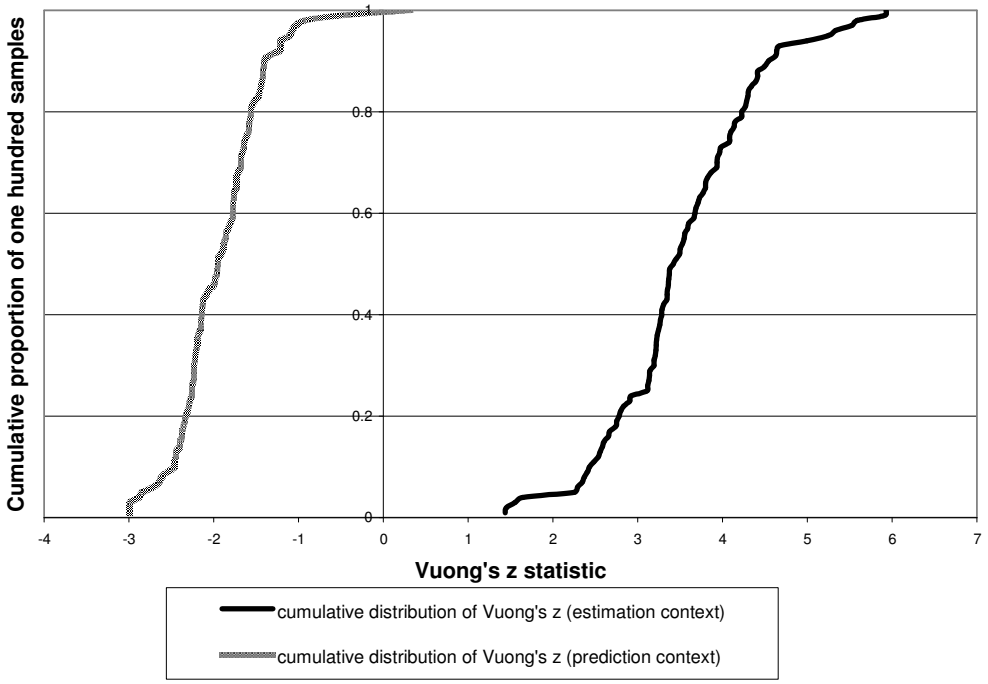
To illustrate this in convincing fashion, consider what happens when we estimate both the (EU,Strong) and (EU,RP) model, one subject *k* at a time, using the simulated (EU,RP) subject samples' estimation context data. Let Δ_k^{set} be the difference between the log likelihoods of the (EU,RP) an (EU,Strong) model from this estimation, either in the estimation context (*set* = *m*23) or the prediction context (*set* = *m*0). We know that if the estimation context had a sufficiently large number of observations, (EU,RP) must fit better on these simulated samples than (EU,Strong) for each and every subject *k*, both in the estimation context and in the prediction context. After all, (EU,RP) is the true data-generating process for each and every one of these simulated subjects *k*, and each subject's parameter vector is constant across the estimation and prediction context. Therefore, as an asymptotic matter, we expect Δ_k^{set} to be positive in both the estimation and prediction context.

Vuong (1989) provides an asymptotic justification for treating a z-score based on the Δ_k^{set} as following a normal distribution under the hypothesis that two non-nested models are equally good, computed as $z^{set} = \sum_{k=1}^{N} \Delta_k^{set} / (s_\Delta^{set} \sqrt{N})$, where s_Δ^{set} is the sample standard deviation of Δ_k^{set} computed without the usual adjustment for a degree of freedom and *N* is the number of subjects in a sample. Figure 8 shows the cumulative distributions of "Vuong's *z*" across the 100 simulated (EU,RP) samples, both in the estimation context (*set* = *m*23) and the prediction context (*set* = *m*0). Virtually all of the distribution of z^{m23} exceeds 2; we would almost always correctly reject equally good fit in favor of (EU,RP) having better fit in the estimation context by Vuong's *z*. Yet virtually all of the distribution of z^{m0} is negative, and actually less than −2 in 45 of the 100 simulated samples. That is, if we apply Vuong's test to compare out-of-context fits of (EU,RP) and (EU,Strong) when (EU,RP) is the true data-generating process, we will basically never correctly reject equal fit in favor of (EU,RP) and will incorrectly reject equal fit in favor of (EU,Strong) about half the time, in a design like Hey and Orme (1994).

Table 5 says this, and nothing more: If your estimation context sample is small (100 choices) for an individual; and if all you care about is predicting out-of-context choices of that individual so as to maximize your log likelihood or minimize your sum of squared errors; then the expected utility structure, combined with a "strong utility sans trembles" stochastic model, is your best bet. The previous example shows that this is not because individual out-of-context fit comparisons gave us any confidence that this is the true data-generating process: The finite sample behavior of individual out-of-context fit varies too much across models to draw such a conclusion. If the question "which model fits best out-of-context" is the wrong question from the viewpoint of hypothesis-testing, then how do we evaluate the plausibility of each model as a true data-generating process? The previous example also suggested that we can safely confine

ourselves to comparing fits within the estimation context. But we won't learn much about the degree to which models can predict and explain in new contexts by doing that.

Figure 8. Distribution of Vuong's z statistic in the estimation and prediction context, comparing fit of (EU,RP) and (EU,Strong) when (EU,RP) is the true data-generating process (positive values indicate correct inference)



I suggest two approaches, which I view as complements rather than substitutes. First, we can instead ask which model behaves as expected out of context, under the assumption that it is the true data-generating process and given our sample sizes, at the level of individuals. That is the next section's topic. Second, we can back off a bit from the desire to treat each individual individually, and instead approach the out-of-context prediction problem using a random parameters approach. In a later section, we will see that this is a much better approach to out-of-context predictive power.

6.4 Expected versus actual behavior of out-of-context predictions of the models

Figures 9 through 14 are actual Hey and Orme sample calibration curves, showing the relationship between observed choice proportions and predicted probabilities out-of-context, for each of the six models. As with Figure 5, all of these figures have simulation-based medians, and 90 percent confidence intervals around those medians, for the observed choice proportions superimposed on them: We expect most observed choice proportions to fall within these intervals, and we also expect observed choice proportions not to fall predominantly above or below the interval medians. Again as with Figure 5, a smooth black line in each figure represents the cumulative distribution of a model's predicted out-of-context probabilities in the actual Hey and Orme sample. There is one difference between these figures and earlier figures, though: The smooth grey line is now the simulated cumulative distribution of predicted probabilities, rather than "true" simulated probabilities. Thus the smooth grey line is the expected cumulative distribution of predicted out-of-context probabilities under the model null: We expect the smooth black line to be close (in a statistical sense) to this smooth grey line. All predicted probabilities used to construct these figures are based on models estimated with trembles, and the data-generating process underlying the simulated data contains a nonzero tremble probability as well (see Table 2 and Tables A1 and A2 in the Appendix).

Figures 9 to 14 show that all of the models produce the same pattern of poor calibration discussed in detail for the (EU,Strong) model in section 6.1: Middling calibration on the left, and strong overprediction on the right. The explanation for this is also identical for all of the models: Predicted probabilities are poorly resolved relative to true probabilities, and since the bulk of true probabilities are below 0.5, long fat upside tails produce a preponderance of upside misclassifications of subject/pair choice probabilities. These show up on the left side of these

Figure 9. Distribution and calibration of out-of-context predictions (EU,Strong,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions.

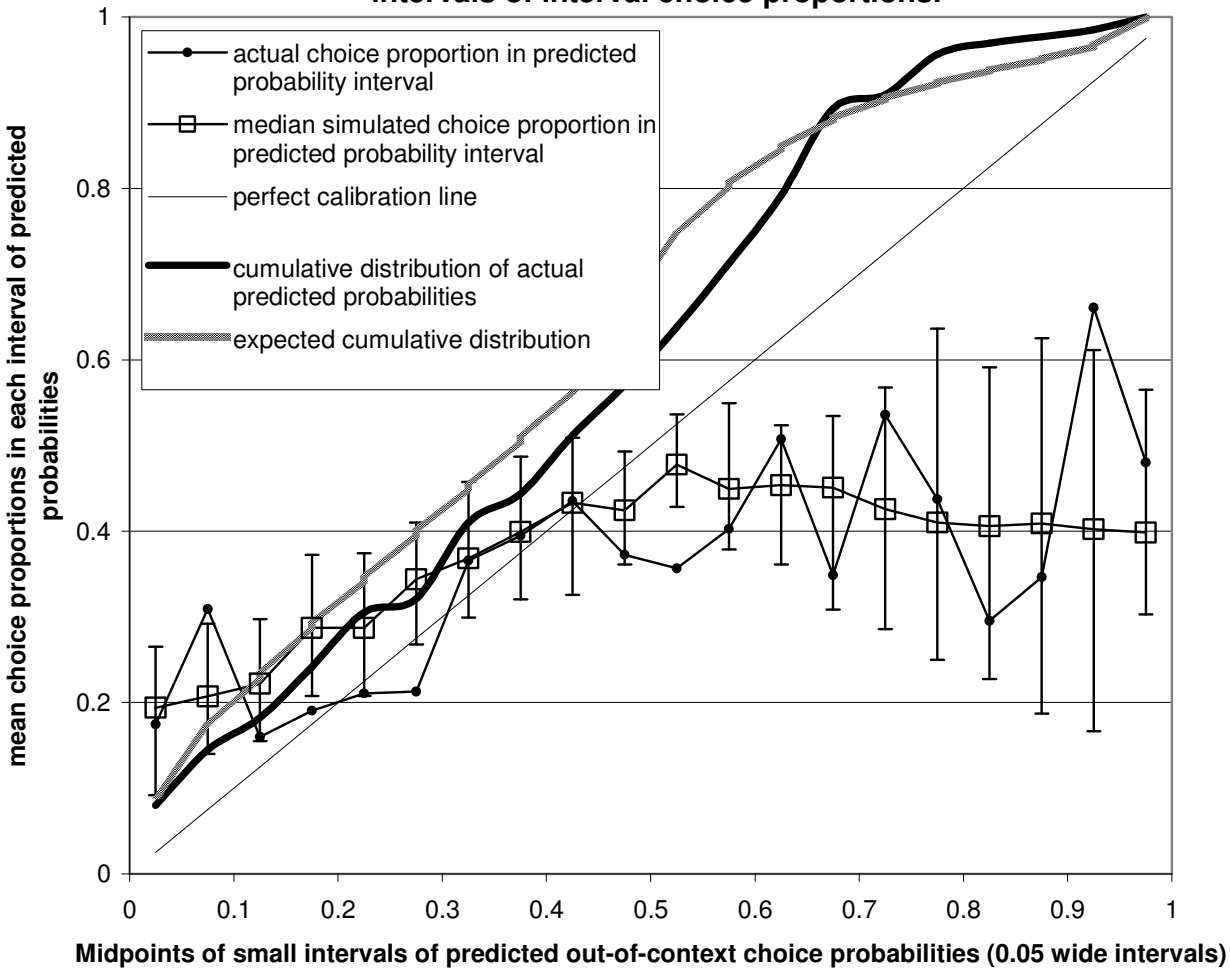


Figure 10. Distribution and calibration of out-of-context predictions (EU,CU,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions.

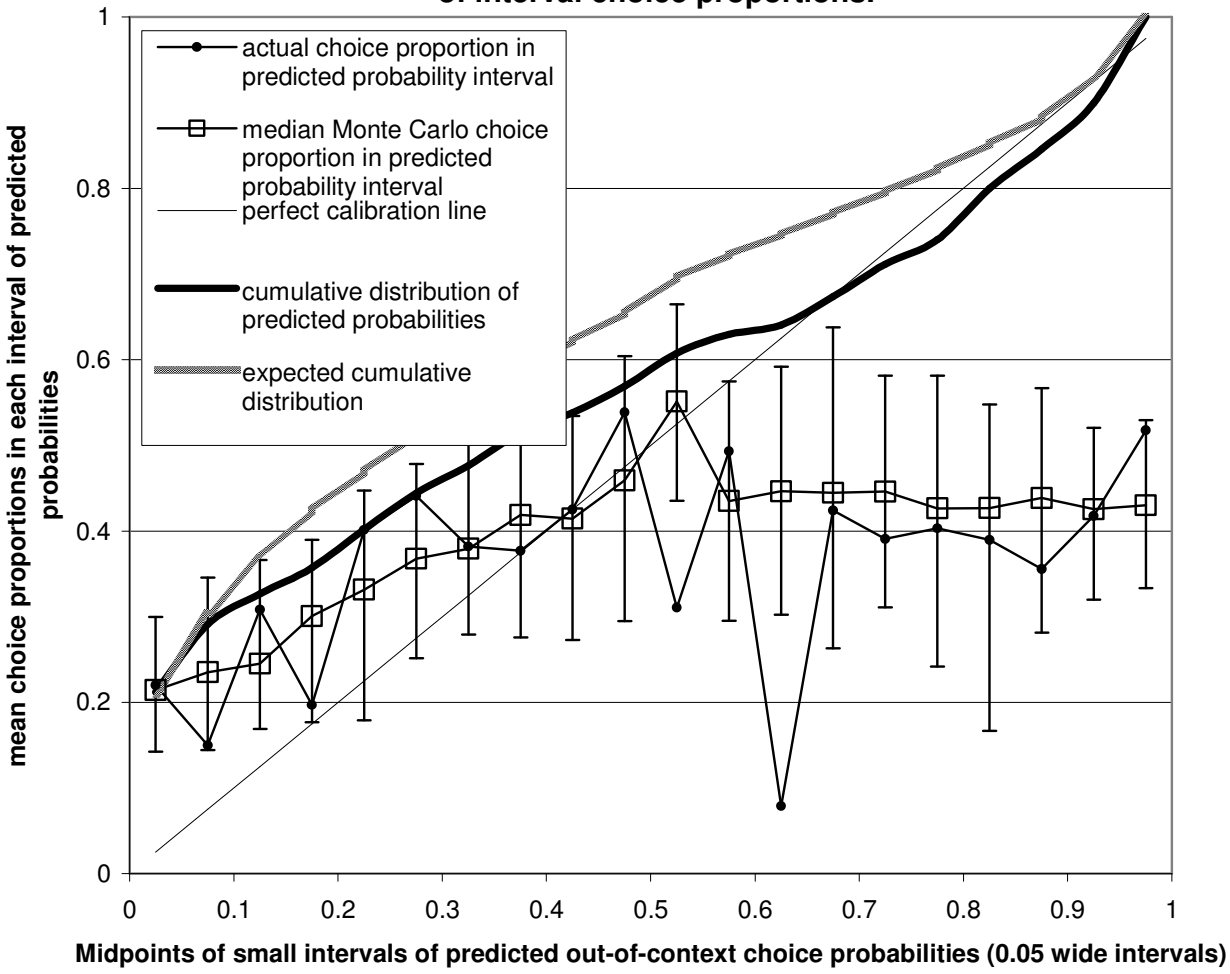


Figure 11. Distribution and calibration of out-of-context predictions (EU,RP,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions

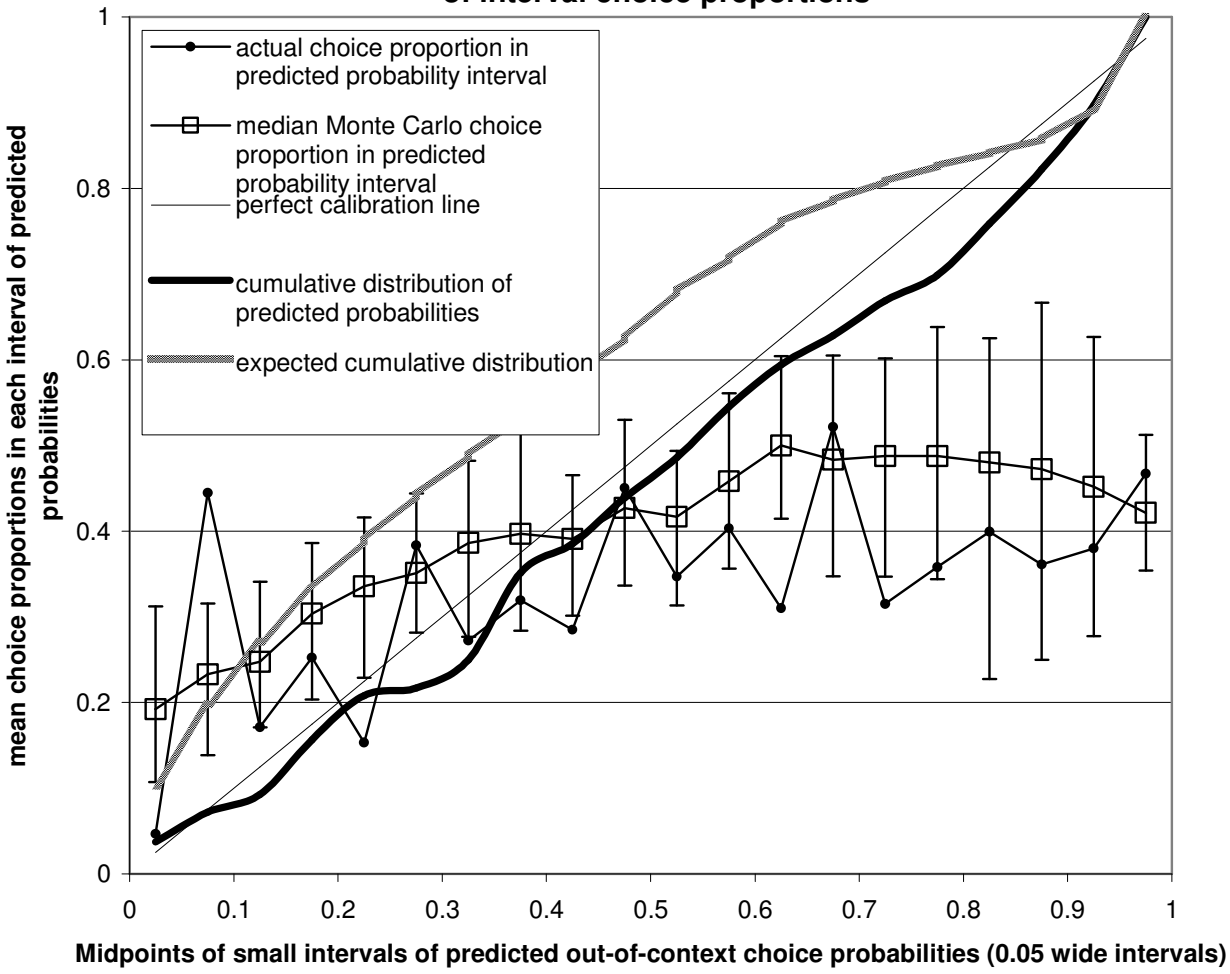


Figure 12. Distribution and calibration of out-of-context predictions (RDEU,SU,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions

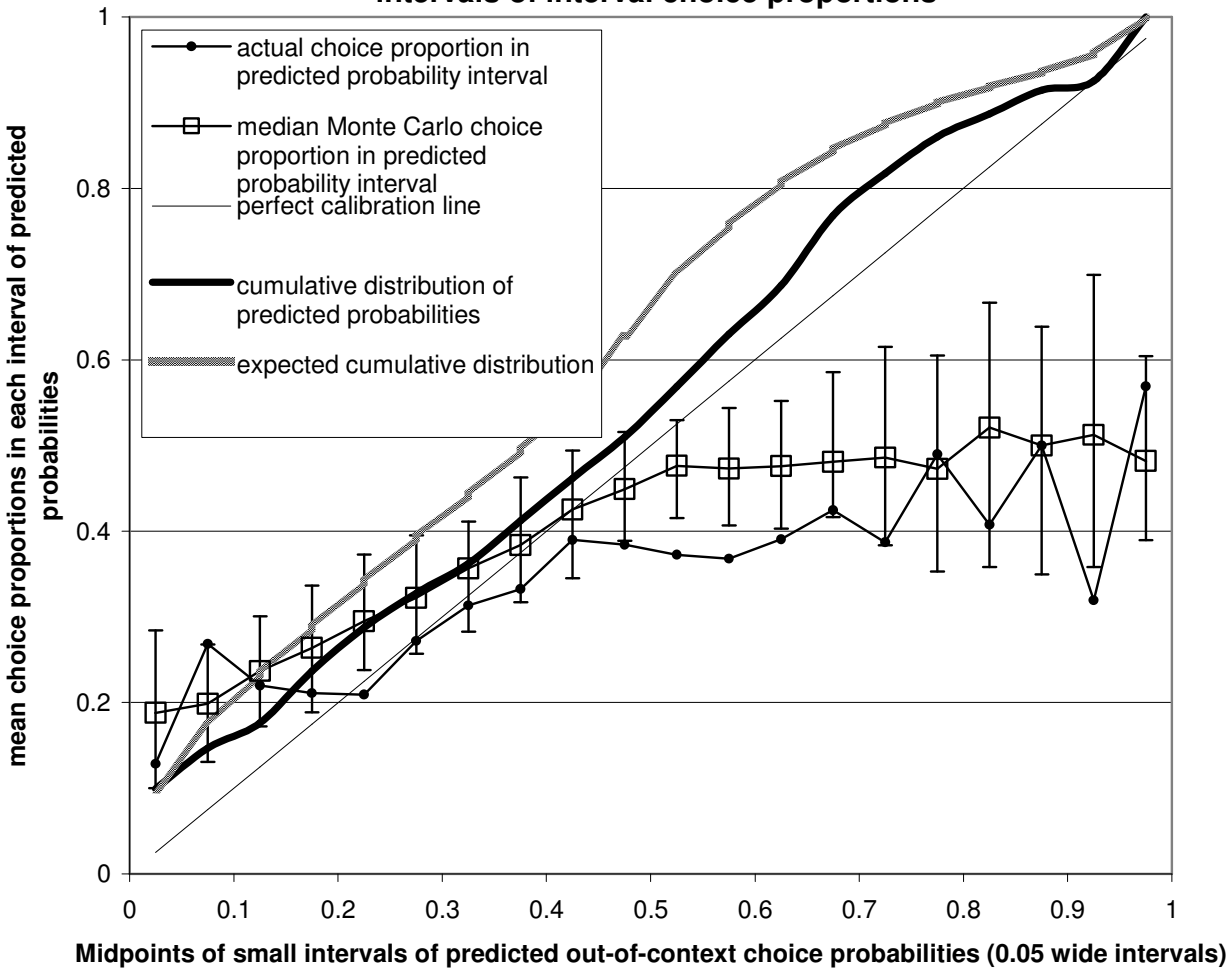


Figure 13. Distribution and calibration of out-of-context predictions (RDEU,CU,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions

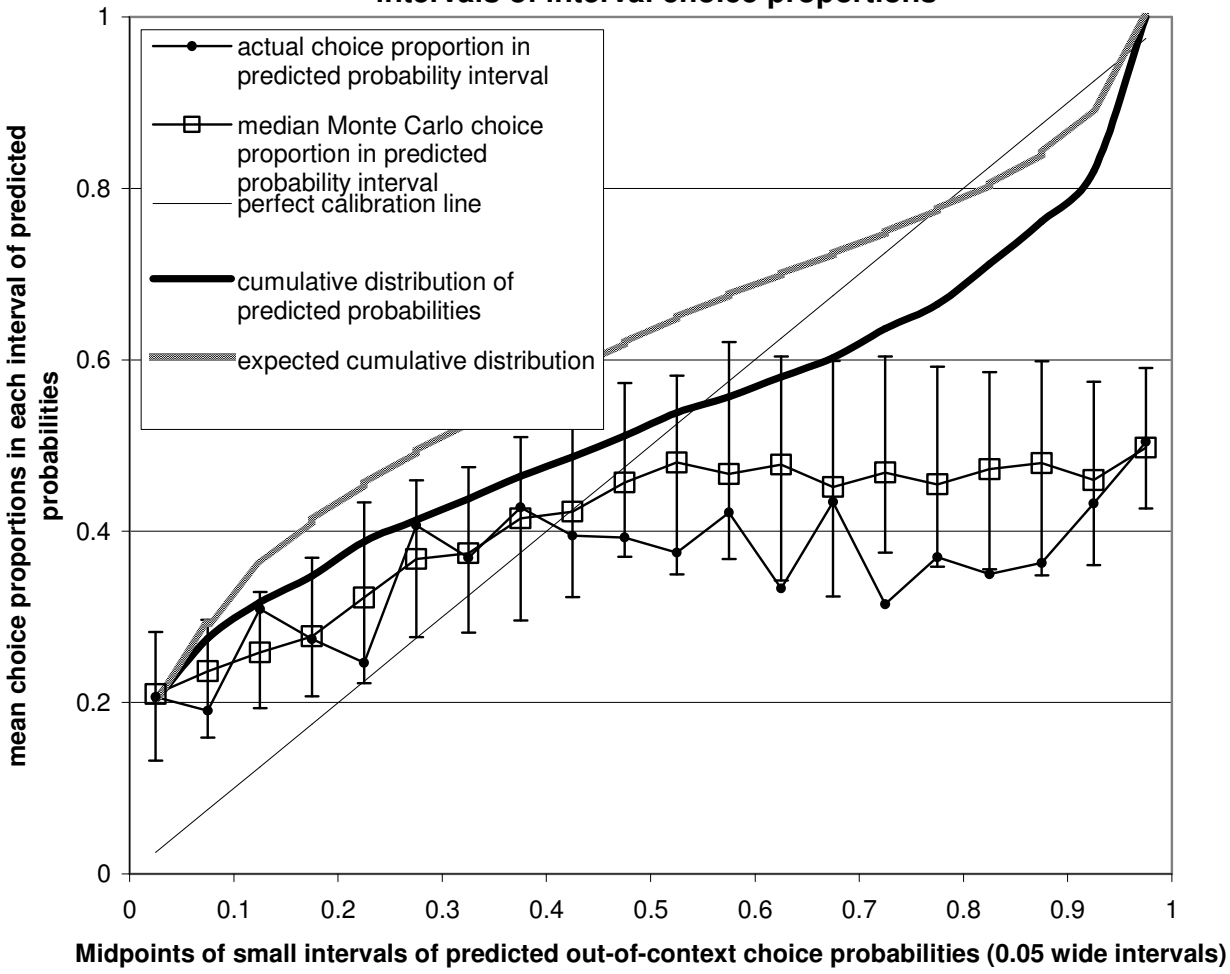
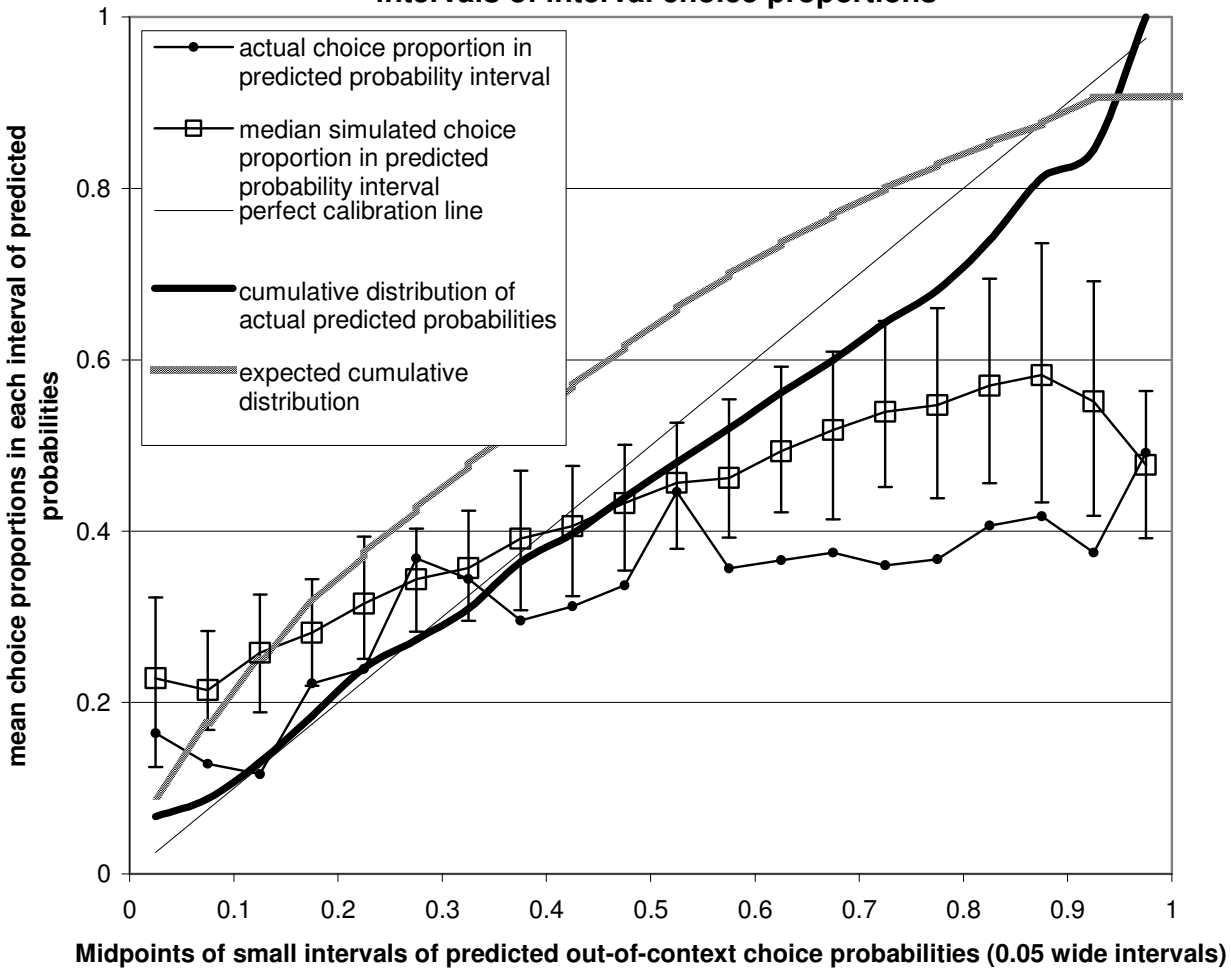


Figure 14. Distribution and calibration of out-of-context predictions (RDEU,RP,Trembles) in Hey and Orme (1994), with simulation-based expected distribution and 90% confidence intervals of interval choice proportions



figures as high predicted probability intervals with observed choice proportions systematically below interval midpoints. To repeat, this is all expected behavior of maximum likelihood estimation and out-of-sample prediction at these sample sizes.

Table 6 summarizes the unexpected outcomes in Figures 9 to 14 and reports simulation-based *p*-values for each of them. There are five different varieties of these outcomes to consider. Let $med_{\hat{j}}$ and $CI_{\hat{j}}$ be the simulated data median and 90 percent confidence interval for observed choice proportions in predicted probability interval $\hat{j}$. The first column is #CI_{out}, where $CI_{out} = \{\hat{j} \mid \bar{\bar{y}}_{\hat{j}} \notin CI_{\hat{j}}\}$, the simple count of observed choice proportions in each figure falling outside their simulated 90 percent confidence intervals. Such events can represent relatively small numbers of predictions, so the second column reports the fraction of observations associated with these "confidence interval failures," that is $\sum_{\hat{j} \in CI_{out}} \#nm(\hat{j}) / 3400$ (25 pairs in the prediction context times 2 trials per pair times 68 subjects = 3400 total observations in the prediction context). The third column is #$med_{<}$, where $med_{<} = \{\hat{j} \mid \bar{\bar{y}}_{\hat{j}} < med_{\hat{j}}\}$, the simple count of observed choice proportions in each figure falling below their simulated median values; as there are twenty intervals in each figure, we might expect this count to be around 10 or 11. It is also interesting to know what fraction of observations are associated with these events. Therefore, the fourth column reports this fraction, that is $\sum_{\hat{j} \in med_{<}} \#nm(\hat{j}) / 3400$; we might expect this to be about 0.50. Finally, the fifth column is the maximum vertical absolute difference between the actual, and simulated expected, cumulative distributions of predicted probabilities—the Kolmogorov-Smirnov "D" statistic for comparing an observed and null distribution.

Table 6. Distribution and calibration failures under the various model nulls
(statistical analysis of information in Figures 9 to 14), with simulation-based *p*-values

		observed choice proportions falling outside simulated 90 percent confidence intervals		observed choice proportions falling below simulated median choice proportion		maximum absolute vertical distance between expected and actual cumulative distribution of predicted probabilities (Kolmogorov-Smirnov one-sample goodness of fit statistic)
		Number of intervals out of the twenty intervals…	…and total fraction of predicted probabilities involved	Number of intervals out of the twenty intervals…	…and total fraction of predicted probabilities involved	
EU	Strong Utility	5 (*p*=0.03)	0.216 (*p*=0.05)	13 (*p*=0.16)	0.701 (*p*=0.08)	0.107 (*p*=0.03)
	Contextual Utility	2 (*p*=0.56)	0.050 (*p*=0.52)	11 (*p*=0.41)	0.457 (*p*=0.57)	0.106 (*p*=0.09)
	Random Preferences	7 (*p*=0.01)	0.280 (*p*=0.01)	15 (*p*=0.03)	0.766 (*p*=0.05)	0.238 (*p*=0.00)
RDEU	Strong Utility	7 (*p*=0.00)	0.334 (*p*=0.00)	16 (*p*=0.03)	0.808 (*p*=0.02)	0.131 (*p*=0.01)
	Contextual Utility	3 (*p*=0.33)	0.104 (*p*=0.29)	16 (*p*=0.04)	0.726 (*p*=0.06)	0.121 (*p*=0.01)
	Random Pref.	14 (*p*=0.00)	0.615 (*p*=0.00)	18 (*p*=0.00)	0.812 (*p*=0.02)	0.180 (*p*=0.00)

Table 6 reports *p*-values for all five of these statistics. In all cases, the *p*-value is obtained by simulation—that is, by computing each statistic in the 100 simulated samples and comparing the observed values in Table 5 to this simulated distribution. The *p*-value is simply the fraction of the 100 simulated samples in which the statistic is greater than or equal to the observed value. Table 6 makes it very clear that only the (EU,CU) model comes through these tests relatively unscathed. The fourth column of the table particularly illuminates the general predictive failure of all six of the models: There is a general overprediction of safe choice in the prediction context by all of the models, though clearly the (EU,CU) model is least prone to this, at least in a directional sense. In section 5, I noted that the switch from generally safe choices to generally risky choices between the estimation and prediction contexts in the Hey and Orme data (see Table 1) would likely be a key challenge for these models, and this general overprediction of safe choices illustrates this.

Table 7 provides a subject-level comparison between the expected behavior of the models and actual choices in the prediction context. Let $\bar{e}_n^{m0} = \sum_{nm \in m0} (\bar{y}_{nm} - \hat{P}_{nm}) / \# m0$ be the average prediction error of a model for subject *n*. Table 7 shows the 10th, 50th and 90th centile subject values of this average error, as well as its mean value across subjects, for each of the models, estimated both with and without trembles. The fourth column, the mean errors when models are estimated with trembles, are all negative, again illustrating the general overprediction of safe choices made by all of the models in the prediction context; the eighth column, the mean errors when the models are estimated without trembles, show a part of the reason why the fit of the strong and contextual utility models improve when they are estimated without trembles: Their upward prediction bias is reduced by doing that. The error statistics all have a simulation-based 90 percent confidence intervals just below them; and the error statistics falling outside these

Table 7. Distribution of errors in predicting subject proportions of safe choices in the prediction context, under the various model nulls: 10th, 50th and 90th centile subjects, and mean subject, with simulated 90 percent confidence intervals for prediction errors

Model		Models estimated with trembles				Models estimated without trembles			
		10th centile prediction error	50th centile prediction error	90th centile prediction error	mean prediction error	10th centile prediction error	50th centile prediction error	90th centile prediction error	mean prediction error
EU	Strong Util.	–0.429 [–.47,–.22]	–0.051 [–.07,0.03]	0.252 [0.15,0.30]	–0.081* [–.08,0.01]	–0.290 [–.29,–.15]	–0.009 [–.04,.03]	0.285* [.15,.28]	–0.030 [–.04,.03]
	Contextual Util.	–0.59* [–.58,–.38]	–0.053 [–.06,.07]	0.237* [.24,.45]	–0.095 [–.10,.04]	–0.452 [–.54,–.29]	0.050 [–.03,.11]	0.336 [.28,.48]	–0.013 [–.06,.08]
	Random Pref.	–0.620* [–.58,–.30]	–0.209* [–.08,.04]	0.143* [0.18,0.37]	–0.217* [–.12,0.01]	–0.580 [–.62,–.43]	–0.205* [–.14,–.01]	0.101* [.12,.30]	–0.230* [–.18,.07]
RDEU	Strong Util.	–0.463* [–.44,–.23]	–0.107* [–.07,0.03]	0.114* [0.13,0.26]	–0.126* [–.09,0.01]	–0.344* [–.32,–.15]	–0.042* [–.04,.04]	0.236 [.13,.25]	–0.054 [–.06,.03]
	Contextual Util.	–0.612* [–.57,–.38]	–0.110* [–.07,.06]	0.212* [.22,.42]	–0.144* [–.11,0]	–0.526 [–.54,–.33]	–0.005 [–.03,.09]	0.310 [.27,.47]	–0.053 [–.07,.04]
	Random Pref.	–0.563* [–.53,–.29]	–0.182* [–.07,.04]	0.146* [0.17,0.34]	–0.215* [–.10,0]	–0.620* [–.55,–.33]	–0.209* [–.10,0]	0.117* [.12,.26]	–0.238* [–.13,–.04]

intervals are marked with an asterisk. It is clear in Table 7 that both random preference models do very poorly with this particular set of tests. On balance, the contextual utility models survive these tests best, but none of the models do very well with them and contextual utility certainly does not look impressively superior to strong utility in Table 7.

6.5 Explaining variation out-of-context

In many applied situations, we do not care so much about point prediction: Instead, we care much more about explaining variation in some cross-sectional dimension. A model could be an inferior point predictor but still explain variation better than its competitors. This is why we should also examine the performance of the models in tasks II and III. Table 8 does this for task II, that is explaining within-subject, between-pair variation. Recall that we have a benchmark for this task, the "Pratt covariate" W_m^*. The first row of Table 8 shows its performance, in terms of its correlation with subjects' mean choices $\bar{y}_{nm}$ in each pair *m*, in the prediction context. For reference, simulated 90 percent confidence intervals for the Pratt covariate are shown as well, under the assumption that (RDEU,CU) is the true data-generating process. The remaining rows use the predicted out-of-context probabilities of each model as the covariate, estimated both with and without trembles ($\hat{P}_{nm}$ and $\widehat{P}_{nm}$, respectively). In all cases, correlations are shown for the 10th, 50th and 90th centile subject, and for the mean correlation across subjects; and simulated 90 percent confidence intervals are shown for each statistic under the null of each model.

The estimation-free Pratt covariate actually turns in a remarkably good performance relative to all of the estimated covariates. Though its mean correlation with subject choices (0.39) is occasionally bested by tremble-free estimations, the improvement is never larger than a small 0.04 difference observed for the (RDEU,Strong) model estimated without trembles. Moreover,

the Pratt covariate has a better-behaved lower tail than any of the estimated task II covariates: The 10th centile subject correlation for the Pratt covariate is 0.20, better (usually much better) than any of the estimated covariates. Still, the strong and contextual utility models estimated without trembles produce 10th centile subject correlations from 0.15 to 0.17, nearly as good. Nevertheless, no model-based task II covariate in Table 8 is a convincing winner.

When we turn to task III, explaining between-subjects variation, Table 9 makes it crystal clear that estimating models at the level of the individual is unnecessary and counterproductive at these sample sizes. As in Table 8, the top row of Table 9 provides a benchmark result—here, the correlation ρ_{b^*} between observed proportions of safe lottery choices $\bar{y}_n^{m23}$ and $\bar{y}_n^{m0}$ in the estimation and prediction contexts, respectively. The alternatives to the estimation-free covariate $\bar{y}_n^{m23}$ are the mean predicted out-of-context probabilities $\hat{\bar{P}}_n^{m0}$ and $\widehat{\bar{P}}_n^{m0}$ (from estimation with and without trembles, respectively) produced by the various models. The other rows of Table 9 show the correlations between these statistics and $\bar{y}_n^{m0}$, along with simulated 90 percent confidence intervals.

Briefly, none of these estimation-based covariates even come close to the simple estimation-free covariate $\bar{y}_n^{m23}$. The right column shows simulated 90 percent confidence intervals for the correlation between $\bar{y}_n^{m23}$ and $\bar{y}_n^{m0}$ under each model. Notice that comparisons between these and the simulated 90 percent confidence intervals for the correlation between the estimation-based covariates and $\bar{y}_n^{m0}$ shows that we should generally expect the estimation-free covariate to perform better than the estimation-based covariates in task III (except when random preference models are the true stochastic model).

Table 8. Performance of task II covariates under the various nulls in the Hey and Orme sample: Distribution of within-subject correlations between predicted out-of-context pair probability $\hat{P}_{nm}$ and mean pair choice $\bar{y}_{nm}$, with expected 90 percent confidence intervals from simulations.

Model		Models estimated with trembles				Models estimated without trembles			
		10th centile subject correlation	50th centile subject correlation	90th centile subject correlation	mean subject correlation	10th centile subject correlation	50th centile subject correlation	90th centile subject correlation	mean subject correlation
"Pratt" covariate W_m^* (benchmark)		0.20 [.17,.32]	0.38 [.42,.49]	0.59 [.56,.64]	0.39 [.41,.47]	0.20 [.14,.26]	0.38 [.38,.45]	0.59 [.53,.61]	0.39 [.37.43]
EU	Strong Util.	0.00 [−.13,.02]	0.36 [.26,.39]	0.64 [.55,.69]	0.35 [.25,36]	0.15 [−.10,.09]	0.40 [.30,.42]	0.64 [.57,.69]	0.40 [.29,.38]
	Contextual Util.	0.11 [−.11,.06]	0.38 [.34,.42]	0.63 [.64,.77]	0.37 [.33,.44]	0.17 [−.07,.17]	0.39 [.39,.51]	0.63 [.66,.77]	0.39 [.37,.47]
	Random Pref.	0.00 [−.01,0]	0.39 [.21,.36]	0.62 [.50,.64]	0.36 [.22,.33]	0.04 [0,0]	0.43 [.25,.37]	0.66 [.51,.63]	0.39 [.25,.34]
RDEU	Strong Util.	0.03 [−.11,.08]	0.41 [.29,.44]	0.69 [.57,.65]	0.38 [.28,.40]	0.15 [−.07,.15]	0.45 [.33,.47]	0.66 [.60,.72]	0.43 [.32,.42]
	Contextual Util.	0.04 [−.06,.19]	0.37 [.36,.49]	0.65 [.64,.75]	0.37 [.35,.44]	0.16 [−.01,.21]	0.42 [.40,.51]	0.65 [.64,.74]	0.42 [.39,.47]
	Random Pref.	0.00 [0,.09]	0.44 [.35,.44]	0.68 [.57,.68]	0.40 [.30,.40]	0.00 [0,.15]	0.45 [.36,.45]	0.64 [.58,.68]	0.41 [.34,.42]

Notes: Expected confidence intervals for the "Pratt" covariate use the simulated data of the (RDEU,CU) model. All other confidence intervals use the simulated data of the row model.

Table 9. Performance of task III covariates under the various nulls in the Hey and Orme sample:
Between-subject correlations of mean predicted out-of-context probabilities $\hat{\bar{P}}_n^{m0}$ and proportion of safe choices $\bar{y}_n^{m0}$.

		Models estimated with trembles	Models estimated without trembles	Simulated confidence interval of $\bar{y}_n^{m23}$ covariate under each null
$\bar{y}_n^{m23}$ covariate (benchmark)		0.46	0.46	
EU	Strong Util.	0.17 [.05,.46]	0.12 [.06,.39]	[.27,.65]
	Contextual Util.	0.22 [−.07,.38]	0.05 [−.11,.31]	[.53,.74]
	Random Pref.	0.15 [.19,.52]	0.24 [.21,.54]	[.24,.58]
RDEU	Strong Util.	0.35 [.13,.51]	0.17 [.10,.48]	[.36,.67]
	Contextual Util.	0.30 [.09,.48]	0.13 [−.03,.41]	[.51,.71]
	Random Pref.	0.26 [.20,.55]	0.22 [.21,.55]	[.18,.52]

I cannot stress the importance of these findings too strongly. We should "expect to be disappointed" by the out-of-sample performance of our individually estimated models at tasks II and III, especially relative to estimation-free benchmarks, even with 100 observations in the estimation context and 50 observations in the prediction context. Seeing such things, we will be tempted to dismiss the models as superfluous mumbo-jumbo. Yet the simulated 90 percent confidence intervals in Tables 8 and 9 make it clear that even when the models are true, we should expect to see these things. Seeing these things is completely uninformative of the truth of the models or the presence of systematic heterogeneity in the underlying population. Yet this particular glass is half-full too: Simple estimation-free approaches to tasks II and III work as well and frequently better than the burdensome task of estimating structural and stochastic parameters for each subject. That is good news for researchers who wish to control for heterogeneous preferences (just use the sample proportion $\bar{y}_n^{m23}$) or decision tasks (just use the "Pratt covariate" or something similar to it) in a simple, relatively model-free manner.

6.6 The random parameters approach

Preceding sections strongly suggest that individual-level estimation with small numbers of observations per individual does not produce good performance in any of our three predictive and explanatory tasks; the simulation results show that this is all expected finite sample behavior of maximum likelihood estimation. In tasks II and III, simple estimation-free covariates perform nearly as well or systematically better than estimation-based covariates in explaining out-of-context variation. Perhaps there is a much better approach to out-of-context prediction at these sample sizes. Since the simulations have been based on random parameters characterizations of

the heterogeneity of structural and stochastic parameters, it is natural to ask whether the random parameters estimation itself outperforms individual estimation in out-of-context prediction.

Table 10 reports log likelihoods based on various random parameters estimations and predictions. The strict utility and wandering vector stochastic models are now added to the mix. The top half of Table 10 reports results using the sixty-eight selected Hey and Orme subjects we have been looking at throughout this study. But since random parameters estimation can include data from subjects with little or no variation in choices, the bottom half of Table 10 also reports the same results using all eighty of Hey and Orme's subjects in the estimations and predictions. The first column of Table 10 shows the log likelihoods on all three contexts, using estimation on all three contexts: In the top half of the table, these are the log likelihoods of the models on which the simulated samples of previous sections were based (see Appendix Tables A1 and A2 for parameter estimates from these estimations). The other two columns show log likelihoods resulting from estimations on choices from lottery pairs in $m23$, the estimation context of previous sections. The center column and right column show the resulting log likelihoods on the estimation and prediction context, respectively, of these random parameters estimates. All of these estimations are carried out using the procedure detailed in section 6.1 for the (EU,Strong) model and summarized for all models in the Appendix.

The right column of Table 10 is the correct one to compare against the "Actual Hey and Orme" columns of Table 5. It is remarkable, but true, that all of these out-of-context log likelihoods improve on those reported in Table 5: The worst of the random parameter model fits is better than the best of the individual model fits. There is, of course, one prosaic reason to expect this. The random parameter model fits are based on at most 11 parameters (RDEU models) for characterizing the entire sample (the parameters in θ), whereas the individual model

fits are based on as many as 340 parameters (RDEU models, 68 subjects times five parameters per subject). We should be unsurprised that an out-of-sample prediction based on 340 parameters fares worse than one based on 11 parameters: Shrinkage associated with liberal burning of degrees of freedom is to be expected, after all.

However, there is some surprise here too. Consider that as an asymptotic matter, the individual model fits must be better than the random parameters fits, even if the random parameters characterization of heterogeneity—that is, the specification of the joint distribution function $J(\psi \mid \theta)$—is exactly right. This is because a random parameters likelihood function takes the expectation of probabilities with respect to *J* before taking logs, while the individual estimations do not (see equation 6.16). Since the log likelihood function is concave in *P*, Jensen's inequality implies that asymptotically (that is, as estimated probabilities converge to true probabilities for both the random parameters and individual estimations) the "expected log likelihoods" of individual estimation must exceed the "log expected likelihoods" of random parameters estimation. That this asymptotic expectation is so clearly reversed in our comparison between Table 5 and Table 10 (even when our specification of *J* in the random parameters models must surely be wrong) just hammers home how far individual estimations are from getting to the promised land of large sample consistency.

Can we draw any serious conclusion about the "best-fitting" model from the comparative log likelihood information in Table 10? Generally, (RDEU,CU) is the "fit winner," both in the estimation and prediction contexts as we have defined them all along and in the three contexts combined. But the advantage of contextual utility over strong utility and random preference is most noticeable in the three contexts combined, and in out-of-context fit on (1,2,3); in general, the stochastic models produce quite similar fits on the estimation context composed of lotteries

on (0,1,2) and (0,1,3). Do note that, at the level of individuals, strong and contextual utility must have identical fit within this estimation context.[38] Therefore, for any given structural model in Table 10, we hope that the random parameters fits of the strong and contextual utility models are quite similar in the center column (and this is mostly true, these being largely the smallest differences in the table): If not, there is suspiciously mixed success in characterizing parameter heterogeneity for the two models. In any case, fit differences between strong and contextual utility in the estimation context should be attributed to mixed success in charactering heterogeneity, rather than any real difference in model performance.

There is another quite interesting result in Table 10. Consider that, as far as out-of-sample fit goes (the left column), the maximum log likelihood improvements associated with moving from EU to RDEU is 50.03 (with random preferences) for the 68 selected subjects and, for all 80 subjects, is just 21.19 (with contextual utility). By contrast, notice that the minimum log likelihood improvements associated with moving from strong utility to contextual utility is 52.57 (with the EU structure) for the 68 selected subjects and, for all 80 subjects, is as large as 113.39 (with the RDEU structure). In a scholarly decision-theoretic research enterprise decidedly dominated by attention to variations in structural models, a very small number of researchers have called for more attention to the stochastic part of decision making under risk (e.g. Hey and Orme 1994; Ballinger and Wilcox 1997). These facts support their call—in spades. Note in particular that while the move from EU to RDEU costs an extra parameter, the move from strong to contextual utility is "free," costing no extra parameters at all.

[38] On just the combined contexts (0,1,2) and (0,1,3), contextual utility is simply a reparameterization of strong utility. Contextual utility is only distinguishable from strong utility when we consider three or more sufficiently distinct contexts, at least one of which has a distinct minimum outcome from the others.

Table 10. Log likelihoods of random parameters characterizations of the models in the Hey and Orme sample

		Selected sixty-eight Hey and Orme subjects used for individual subject estimations		
		Estimated on all three contexts	Estimated on contexts (0,1,2) and (0,1,3)	
Structure	Stochastic Model	Log Likelihood on all three contexts (in-context fit)	Log likelihood on contexts (0,1,2) and (0,1,3) (in-context fit)	Log likelihood on context (1,2,3) (out-of-context fit)
EU	Strong Utility	−4841.50	−2873.84	−1983.05
	Strict Utility	−4963.85	−2961.24	−2000.66
	Wandering Vector	−4871.93	−2891.59	−1997.61
	Contextual Utility	−4819.91	−2882.70	−1930.48
	Random Preferences	−4859.71	−2889.00	−2002.30
RDEU	Strong Utility	−4748.03	−2810.20	−1977.12
	Strict Utility	−4838.96	−2840.78	−2038.75
	Wandering Vector	−4769.84	−2815.19	−1998.44
	Contextual Utility	−4731.77	−2810.28	−1921.81
	Random Preferences	−4752.74	−2815.22	−1952.27
		All eighty Hey and Orme subjects		
		Estimated on all three contexts	Estimated on contexts (0,1,2) and (0,1,3)	
Structure	Stochastic Model	Log Likelihood on all three contexts (in-context fit)	Log likelihood on contexts (0,1,2) and (0,1,3) (in-context fit)	Log likelihood on context (1,2,3) (out-of-context fit)
EU	Strong Utility	−5311.44	−2975.68	−2409.38
	Strict Utility	−5448.50	−3069.85	−2373.12
	Wandering Vector	−5362.61	−3004.08	−2417.76
	Contextual Utility	−5297.08	−2980.13	−2302.55
	Random Preferences	−5348.36	−2993.91	−2356.60
RDEU	Strong Utility	−5207.81	−2912.12	−2394.75
	Strict Utility	−5306.48	−2949.34	−2450.41
	Wandering Vector	−5251.82	−2925.31	−2397.91
	Contextual Utility	−5190.43	−2909.09	−2281.36
	Random Preferences	−5218.00	−2906.06	−2335.55

Table 11. Vuong (1989) non-nested tests between stochastic model pairs (random parameters estimations of EU structure), by subjects considered and in-context versus out-of-context fit.

Selected 68 Hey and Orme subjects									
Estimated on all three contexts, and comparing fit on all three contexts (in-context fit comparison)					Estimated on contexts (0,1,2) and (0,1,3), and comparing fit on context (1,2,3) (out-of-context fit comparison)				
	Random Prefs.	Strong Utility	Wandering Vector	Strict Utility		Random Prefs.	Strong Utility	Wandering Vector	Strict Utility
Contextual Utility	$z = 1.574$ $p = 0.058$	$z = 1.219$ $p = 0.11$	$z = 1.882$ $p = 0.030$	$z = 5.131$ $p < 0.0001$	Contextual Utility	$z = 6.868$ $p < 0.0001$	$z = 2.485$ $p = 0.0065$	$z = 3.291$ $p = 0.0005$	$z = 1.665$ $p = 0.048$
Random Prefs.	—	$z = -0.885$ $p = 0.19$	$z = 0.746$ $p = 0.228$	$z = 4.396$ $p < 0.0001$	Random Prefs.	—	$z = -0.912$ $p = 0.18$	$z = -0.260$ $p = 0.40$	$z = -0.912$ $p = 0.18$
Strong Utility	—	—	$z = 1.512$ $p = 0.065$	$z = 5.170$ $p < 0.0001$	Strong Utility	—	—	$z = 2.212$ $p = 0.013$	$z = -0.390$ $p = 0.35$
Wandering Vector	—	—	—	$z = 4.719$ $p < 0.0001$	Wandering Vector	—	—	—	$z = -1.002$ $p = 0.16$
All 80 Hey and Orme subjects									
Estimated on all three contexts, and comparing fit on all three contexts (in-context fit comparison)					Estimated on contexts (0,1,2) and (0,1,3), and comparing fit on context (1,2,3) (out-of-context fit comparison)				
	Random Prefs.	Strong Utility	Wandering Vector	Strict Utility		Random Prefs.	Strong Utility	Wandering Vector	Strict Utility
Contextual Utility	$z = 1.723$ $p = 0.042$	$z = 0.703$ $p = 0.241$	$z = 2.354$ $p = 0.0093$	$z = 6.067$ $p < 0.0001$	Contextual Utility	$z = 4.387$ $p < 0.0001$	$z = 3.044$ $p = 0.0012$	$z = 3.509$ $p = 0.0002$	$z = 2.739$ $p = 0.0031$
Random Prefs.	—	$z = -1.574$ $p = 0.058$	$z = 0.739$ $p = 0.230$	$z = 5.419$ $p < 0.0001$	Random Prefs.	—	$z = 1.639$ $p = 0.051$	$z = 2.148$ $p = 0.016$	$z = 1.422$ $p = 0.078$
Strong Utility	—	—	$z = 3.236$ $p = 0.0006$	$z = 5.961$ $p < 0.0001$	Strong Utility	—	—	$z = 0.965$ $p = 0.167$	$z = 0.028$ $p = 0.49$
Wandering Vector	—	—	—	$z = 5.079$ $p < 0.0001$	Wandering Vector	—	—	—	$z = -.322$ $p = 0.37$

Notes: Positive z means the row stochastic model fits better than the column stochastic model.

Table 12. Vuong (1989) non-nested tests between stochastic model pairs (random parameters estimations of RDEU structure), by subjects considered and in-context versus out-of-context fit.

Selected 68 Hey and Orme subjects									
Estimated on all three contexts, and comparing fit on all three contexts (in-context fit comparison)					Estimated on contexts (0,1,2) and (0,1,3), and comparing fit on context (1,2,3) (out-of-context fit comparison)				
	Random Prefs.	Strong Utility	Wandering Vector	Strict Utility		Random Prefs.	Strong Utility	Wandering Vector	Strict Utility
Contextual Utility	$z = 0.793$ $p = 0.21$	$z = 0.854$ $p = 0.20$	$z = 1.351$ $p = 0.088$	$z = 4.446$ $p < 0.0001$	Contextual Utility	$z = 2.363$ $p = 0.0091$	$z = 3.040$ $p = 0.0012$	$z = 3.972$ $p < 0.0001$	$z = 5.875$ $p < 0.0001$
Random Prefs.	—	$z = -0.227$ $p = 0.41$	$z = 0.810$ $p = 0.21$	$z = 4.340$ $p < 0.0001$	Random Prefs.	—	$z = 1.155$ $p = 0.124$	$z = 2.266$ $p = 0.0117$	$z = 3.639$ $p = 0.0001$
Strong Utility	—	—	$z = 1.145$ $p = 0.126$	$z = 5.789$ $p < 0.0001$	Strong Utility	—	—	$z = 2.366$ $p < 0.0090$	$z = 7.876$ $p < 0.0001$
Wandering Vector	—	—	—	$z = 2.951$ $p = 0.0016$	Wandering Vector	—	—	—	$z = 3.953$ $p < 0.0001$
All 80 Hey and Orme subjects									
Estimated on all three contexts, and comparing fit on all three contexts (in-context fit comparison)					Estimated on contexts (0,1,2) and (0,1,3), and comparing fit on context (1,2,3) (out-of-context fit comparison)				
	Random Prefs.	Strong Utility	Wandering Vector	Strict Utility		Random Prefs.	Strong Utility	Wandering Vector	Strict Utility
Contextual Utility	$z = 0.981$ $p = 0.163$	$z = 0.877$ $p = 0.190$	$z = 2.239$ $p = 0.013$	$z = 4.352$ $p < 0.0001$	Contextual Utility	$z = 3.879$ $p < 0.0001$	$z = 3.304$ $p = 0.0005$	$z = 4.040$ $p < 0.0001$	$z = 5.978$ $p < 0.0001$
Random Prefs.	—	$z = -0.44$ $p = 0.330$	$z = 1.765$ $p = 0.0388$	$z = 3.808$ $p < 0.0001$	Random Prefs.	—	$z = 1.652$ $p = 0.049$	$z = 2.073$ $p = 0.0191$	$z = 3.831$ $p < 0.0001$
Strong Utility	—	—	$z = 2.697$ $p = 0.0035$	$z = 5.973$ $p < 0.0001$	Strong Utility	—	—	$z = 0.261$ $p = 0.397$	$z = 3.918$ $p < 0.0001$
Wandering Vector	—	—	—	$z = 2.700$ $p < 0.0035$	Wandering Vector	—	—	—	$z = 5.695$ $p < 0.0001$

Notes: Positive z means the row stochastic model fits better than the column stochastic model.

We can use Vuong's (1989) test (as described earlier in section 6.3) to compare either the in-context or out-of-context fits. The problems noted in section 6.3, and illustrated by Figure 8, with applying Vuong's z to individual out-of-context fits are much less likely to be a problem with these random parameter estimations where degrees of freedom are much, much more abundant; but we should keep them in mind.[39] Tables 11 and 12 report the results. In all cases, the reported p-value is against the null of equally good fit with a one-tailed alternative that the directionally better fit is significantly better. Contextual utility is always directionally better than all of its stochastic model competitors: This is true for (a) both the EU or RDEU structures; (b) both in-context for all three contexts combined, and out-of-context; and (c) whether estimation uses just the 68 selected subjects or all 80 subjects. But contextual utility really shines in out-of-context prediction, where it spanks all other stochastic models with strong significance.

More than one research team has argued that the stochastic model we prefer may depend on the structural model we examine (Buschena and Zilberman 2000; Loomes, Moffatt and Sugden 2002), and have offered empirical illustrations of this point. As a general methodological point, I agree with them. Yet the overwhelming impression one gets from Tables 11 and 12 is that contextual utility is generally a better stochastic model, regardless of whether we view the matter from the perspective of the EU or RDEU structures, or from the perspective of in-context or out-of-context fit. This conclusion accords with most (though certainly not all) of the evidence discussed earlier in section 6.4 regarding the expected out-of-context prediction behavior of individual estimation.

[39] A Monte Carlo analysis of the behavior of the random parameters estimator is a mighty undertaking. For instance, even using the method described in the Appendix for choosing good starting values, the three-integral quadratures required to estimate the random parameters RDEU model took about eight hours to estimate for all eighty Hey and Orme subjects on three contexts. Multiply this by one hundred samples, and the variety of models estimated, to get a sense of the time involved.

7. Conclusions

It has become relatively easy to estimate complex nonlinear models using statistical software packages such as SAS and STATA. This can be a good thing, provided that we understand the limitations. We need to remember that each qualitative dependent response carries very little information about any hypothesized continuous latent construct we may wish to estimate, such as the parameters of a *V*-distance or a stochastic parameter. We also need to remember that estimation of *k* parameters of a nonlinear function of parameters is very different from estimating the effects of *k* orthogonal regressors, such as *k* independently varied treatment variables. This is because the first derivatives of a nonlinear function, which play the mathematical role of regressors in nonlinear estimation, are usually correlated with one another (as our orthogonal regressors are not). For both of these reasons (that is, because our data is discrete and our models are nonlinear), estimation of models of discrete choice under risk is a very data-hungry business.

This study illustrates how data-hungry this business is. It is surprising, at least to me, how little we should expect from individual-level estimation of these models, even with 100 observations per subject, in terms of our ability to predict and explain in a new sample with 50 observations per subject. My own prior intuition about this was so poor, that I think it is informative to summarize this study in terms of my own humbling.

I believed that decision theory, being theory of individuals, ought to be tested on an individual-by-individual basis. I also believed that out-of-context prediction, in the form of out-of-context fit, ought to be a sensible criterion for comparing the theories. It turns out that at the sample sizes we typically have, those beliefs are in conflict. The reason is that the finite-sample out-of-context fit behavior of different models of decision under risk (when individually

estimated) differs much more severely across models than do their asymptotic out-of-context fit behavior—and in ways that do not correspond directionally to the differences in their asymptotic behavior (see section 6.3, Table 5 and Figure 8).

The proviso "out-of-context" is also key here: Recall that the "in-context" Voung test was actually surprisingly well-behaved with individual estimation, especially compared to how badly behaved that sample test is out-of-context with individual estimation. Therefore, what I say here does not amount to a criticism of Hey and Orme (1994): They did indeed use individual estimation to compare models, but their comparisons are in-sample comparisons of fit. But, back to my own faulty intuition: Comparing estimated models by their out-of-sample predictive content, on a subject-by-subject basis, seemed compelling. This intuition turned out to be dead wrong when we estimate models of choice under risk, one subject at a time, with 100 or fewer observations. The key to understanding this apparent paradox is remembering that the intuition is based on what is asymptotically true, and remembering that asymptotic truths can be irrelevant or worse in finite samples (in this case, worse).

I also believed that individually estimated models of decision under risk ought to provide us with reasonably good covariates for explaining both variations in each subject's choices across pairs and variations in mean choices across subjects. I believed this must work better than some "unsophisticated" approach, such as explaining these variations by means of the "Pratt covariate" or the sample proportions of safe choices in the estimation context, respectively. Those beliefs, however, were unsophisticated: In fact Monte Carlo simulations show that I was foolish to expect this (Tables 8 and 9). At these sample sizes, individually estimated models explain out-of-context variation in pair choices not much better than the Pratt covariate, and individually estimated models are almost certain to explain out-of-context variations across subjects worse

than the simple sample proportions observed for each subject in the estimation context. This is not because any particular model examined here is a bad model, or because model parameters are unstable across decision contexts or pairs: After all, the models are "made true" in the Monte Carlo data sets, with stable parameters by construction. As it turns out, I should have expected a near order of magnitude shrinkage of the between-subjects explanatory power of the models (when estimated one subject at a time) at these sample sizes, relative to what that power would be asymptotically (with a very large estimation context sample). At typical sample sizes, for the purpose of out-of-context prediction and explanation, estimating models on a subject-by-subject basis appears quixotic at best and at worst utterly misleading. When we instead treat heterogeneity by a random parameters approach, we get nothing like the (inferentially illusory) huge differences in out-of-context predictive power observed with subject-by-subject estimation, and the out-of-context fits are uniformly better as well (Tables 5 and 10).

On the whole, it is fair to say that in its explanatory efforts, the academic community of decision theory spends far more time with the structural part of decision making than the stochastic part. Yet the random parameters estimations reported here suggest that there is perhaps more to be gained by getting the stochastic model right than by proliferating parameters and functions in structural models. These gains are sometimes parameter-free, as in the comparison between strong and contextual utility discussed in section 6.6. I repeat my own call (Ballinger and Wilcox 1997) and Hey's many calls (Hey and Orme 1994; Hey 2001; Hey 2005) for greater theoretical attention to the stochastic part of decision under risk. The random parameters estimations here suggest that my own contextual utility model (Wilcox 2007) beats all competitors in out-of-context prediction.

The original motivation for this study was to examine the out-of-context predictive success of individually estimated models of decision under risk, and in particular the stochastic models that go with them, as a criterion for judging those models. Specifically, I sought no critique of claims of preference instability across institutions and response modes. Perhaps these claims are ultimately correct. But the results of this study do suggest that much of what passes for preference instability is, in part, simply insufficient design power. I say this with some sympathy: My own prior intuitions about what could be expected in this study were, after all, laid low. Yet maybe it is a good thing to have one's intuition thoroughly humbled now and then; It may help lead one away from destruction and toward deliverance.

References

Aitchison, J. (1963). Inverse Distributions and Independent Gamma-Distributed Products of Random Variables. *Biometrika 50*, 505-508.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutstrom, E. E. (2007). Eliciting Risk and Time Preferences. Working paper 5-24, University of Central Florida Department of Economics.

Archibald, G. and Wilcox, N. (2002). A new variant of the winner's curse in a Coasian contracting game. *Experimental Economics 5*,155-172.

Ballinger, T. P. and Wilcox, N. T. (1997). Decisions, Error and Heterogeneity. *Economic Journal 107*, 1090-1105.

Ballinger, T. P., Hudson, E., Karkoviata, L., and Wilcox, N. (2006). Saving Behavior and Cognitive Abilities. Working paper, University of Houston Department of Economics.

Barsky, B., Juster, F. T., Kimball, M. and Shapiro, M. (1997). Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study. *Quarterly Journal of Economics 112*, 537-579.

Becker, G. M., DeGroot, M. H. and Marschak, J. (1963a). Stochastic Models of Choice Behavior. *Behavioral Science 8*, 41–55.

_____ (1963b). An Experimental Study of Some Stochastic Models for Wagers. *Behavioral Science 8*, 199-202.

Benjamin, D., Brown, S. and Shapiro, J. (2006). Who is "Behavioral"? Cognitive Ability and Anomalous Preferences. Working paper, Harvard University Department of Economics.

Berg, J., Dickhaut, J. and McCabe, K. (2005). Risk Preference Instability across Institutions: A Dilemma. *Proceedings of the National Academy of Sciences 102*, 4209-4214.

Birnbaum, M. (2004). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes 95*, 40-65.

Black, D. (1948). On the Rationale of Group Decision Making. *Journal of Political Economy 56*, 23-34.

Blackburn, M., Harrison, G., and Rutström, E. (1994). Statistical Bias Functions and Informative Hypothetical Surveys. *American Journal of Agricultural Economics 76,* 1084-1088.

Blavatskyy, P. R. (2006a) Violations of Betweenness or Random Errors? *Economics Letters 91*, 34-38

Blavatskyy, P. R. (2007). Stochastic Choice Under Risk. *Journal of Risk and Uncertainty* (forthcoming).

Block, H. D. and Marschak, J. (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* (pp. 97-132). Stanford CA: Stanford University Press.

Buschena, D. E. and Zilberman, D. (2000). Generalized Expected Utility, Heteroscedastic Error, and Path Dependence in Risky Choice. *Journal of Risk and Uncertainty 20*, 67-88.

Camerer, C. (1989). An Experimental Test of Several Generalized Expected Utility Theories. *Journal of Risk and Uncertainty 2*, 61-104.

Carbone, E. (1997). Investigation of Stochastic Preference Theory Using Experimental Data. *Economics Letters 57*, 305-311.

Carroll, J. D. (1980). Models and Methods for Multidimensional Analysis of Preferential Choice (or Other Dominance) Data. In E. D. Lantermann and H. Feger (Eds.), *Similarity and Choice* (pp. 234-289). Bern, Switzerland: Huber.

Carroll, J. D., and De Soete, G. (1991). Toward a New Paradigm for the Study of Multiattribute Choice Behavior. *American Psychologist 46*, 342-351.

Chew, S. H. (1983). A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox. *Econometrica 51*, 1065-1092.

Chipman, J. (1963). Stochastic Choice and Subjective Probability. In D. Willner (Ed.), *Decisions, Values and Groups* (pp. 70-95). New York: Pergamon.

Debreu, G. (1958). Stochastic Choice and Cardinal Utility. *Econometrica 26*, 440-444.

Debreu, G. (1960). Review of R. D. Luce, "Individual Choice Behavior: A Theoretical Analysis." *American Economic Review 50*, 186-188.

Domencich, T., and McFadden, D. (1975). *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.

Edwards, W. (1954). A Theory of Decision Making. *Psychological Bulletin 51*, 380-417.

Efron, B. and Tibshirani, R. (1998). *An Introduction to the Bootstrap*. Boca Raton, Florida: CRC Press (First published in 1993 by Chapman and Hall; reprinted in 1998 by CRC Press).

Fechner, G. (1966/1860). *Elements of Psychophysics Vol. 1*. New York: Holt, Rinehart and Winston.

Fishburn, P. (1999). Stochastic Utility. In S. Barbara, P. Hammond and C. Seidl (Eds.), *Handbook of Utility Theory Vol. 1* (pp. 273-320). Berlin: Springer.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 24-42.

Halff, H. M. (1976). Choice Theories for Differentially Comparable Alternatives. *Journal of Mathematical Psychology 14*, 244-246.

Harrison, G. W., and Rutström, E. E. (2005). Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral. Working paper, University of Central Florida Department of Economics.

Harrison, G. W., Johnson, E., McInnes, M. and Rutström, E. E. (2005). Risk Aversion and Incentive Effects: Comment. *American Economic Review 95*, 897-891.

Hershey, J. C., and Schoemaker, P. J. H. (1985). Probability versus Certainty Equivalence Methods in Utility Measurement: Are They Equivalent? *Management Science 31* 1213-1231.

Hey, J. D. (1995). Experimental Investigations of Errors in Decision Making Under Risk. *European Economic Review 39*, 633-640.

Hey, J. D. (2001). Does Repetition Improve Consistency? *Experimental Economics 4* 5-54.

Hey, J. D. (2005). Why We Should Not Be Silent About Noise. *Experimental Economics 8*, 325-345.

Hey, J. D. and Carbone, E. (1995). Stochastic Choice with Deterministic Preferences: An Experimental Investigation. *Economics Letters 47*, 161–167.

Hey, J. D. and Orme, C. (1994). Investigating Parsimonious Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica 62*, 1291-1329.

Hryshko, D., Luengo-Prado, M., and Sorensen, B. (2006). Childhood Determinants of Risk Aversion: The Long Shadow of Compulsory Education. Working paper, University of Houston Department of Economics.

Hutchinson, T. P. and Lai, C. D. (1990). *Continuous Bivariate Distributions, Emphasizing Applications*. Adelaide, Australia: Rumsby Scientific Publishers.

Judd, K. L. (1998). *Numerical Methods in Economics*. Cambridge, MA: MIT Press.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica 47*, 263-291.

Kimball, M., Sahm, C., and Shapiro, M. (2007). Imputing Risk Tolerance from Survey Responses. Working Paper, University of Michigan Department of Economics.

Leland, J. (1994). Generalized Similarity Judgments: An Alternative Explanation for Choice Anomalies. *Journal of Risk and Uncertainty 9*, 151-172.

Loomes, G. and Sugden, R. (1995). Incorporating a Stochastic Element into Decision Theories. *European Economic Review 39*, 641-648.

Loomes, G. and Sugden, R. (1998). Testing Different Stochastic Specifications of Risky Choice. *Economica 65*, 581-598.

Loomes, G., Moffatt, P. and Sugden, R. (2002). A Microeconometric Test of Alternative Stochastic Theories of Risky Choice. *Journal of Risk and Uncertainty 24*, 103-130.

Luce, R. D. (1977). The Choice Axiom after Twenty Years. *Journal of Mathematical Psychology 15*, 215-233.

Luce, R. D. and Suppes, P. (1965). Preference, Utility and Subjective Probability. In R. D. Luce, R. R. Bush and E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. III* (pp. 249-410). New York: Wiley.

Machina, M. (1985). Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries. *Economic Journal 95*, 575-594.

McKay, A. T. (1934). Sampling from Batches. *Supplement to the Journal of the Royal Statistical Society 1*, 207-216.

Moffatt, P. (2005). Stochastic Choice and the Allocation of Cognitive Effort. *Experimental Economics 8*, 369-388.

Moffatt, P. and Peters, S. (2001). Testing for the Presence of a Tremble in Economics Experiments. *Experimental Economics 4*, 221-228.

Quiggin, J. (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization 3*, 323-343.

Papke, L. E., and Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics 11*, 619-632.

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.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing Risk I: A Definition. *Journal of Economic Theory 2*, 225-243.

Rubinstein, A. (1988). Similarity and Decision Making Under Risk (Is There a Utility Theory Resolution to the Allais Paradox?). *Journal of Economic Theory 46*, 145-53.

Schwert, G. W. (1989). Test for Unit Roots: A Monte Carlo Investigation. *Journal of Business and Economic Statistics 7*, 147-159.

Sonsino, D., Benzion, U. and Mador, G. (2002). The Complexity Effects on Choice With Uncertainty—Experimental Evidence. *Economic Journal 112*, 936-965.

Starmer, C. and Sugden, R. (1989). Probability and Juxtaposition Effects: An Experimental Investigation of the Common Ratio Effect. *Journal of Risk and Uncertainty 2*, 159-78.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review 76*, 31-48.

Train, K. (2003). *Discrete Choice Methods With Simulation*. Cambridge, U.K.: Cambridge University Press.

Tversky, A. (1969). Intransitivity of Preferences. *Psychological Review 76*, 31-48.

Tversky, A. (1972). Elimination by Aspects: A Theory of Choice. *Psychological Review 79*, 281-299.

Tversky, A. and Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin 76*, 105-110.

Tversky, A. and Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business 59*, S251-S278.

Tversky, A. and Russo, J. E. (1969). Substitutability and Similarity in Binary Choices. *Journal of Mathematical Psychology 6*, 1-12.

Vuong, Q. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*, 307–333.

Wilcox, N. T. (1993). Lottery Choice: Incentives, Complexity and Decision Time. *Economic Journal 103*, 1397-1417.

Wilcox, N. T. (2006). Theories of Learning in Games and Heterogeneity Bias. *Econometrica 74*, 1271-1292.

Wilcox, N. T. (2007). 'Stochastically More Risk Averse:' A Theory of Contextual Utility for Stochastic Choice Under Risk. Working paper, University of Houston Department of Economics.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge MS: MIT Press.

Appendix: Random Parameters Estimates of the Models

A uniform procedure was used to select and then estimate random parameters characterizations of heterogeneity for all models. The estimated parameters θ then become the basis for creating simulated data sets under the null of each model, for bootstrapping *p*-values and confidence intervals of various statistics, as described in detail for the (EU,Strong) model in section 6.1. In each case, a model's parameter vector ψ^n is first estimated separately for each subject with a fixed tremble probability $\omega = 0.04$. let these estimates be $\tilde{\psi}^n$. The correlation matrix of the parameters is then computed, and the vectors $\tilde{\psi}^n$ are also subjected to a principal components analysis, with particular attention to the first principal component.

As with the detailed example of the (EU,Strong) model, all models with utility parameters u_2^n and u_3^n yield quite high Pearson correlations between $\ln(\tilde{u}_2^n - 1)$ and $\ln(\tilde{u}_3^n - 1)$ across subjects, and heavy loadings of these on first principal components of the estimated parameter vectors $\tilde{\psi}^n$. Therefore, the joint distributions $J(\psi \mid \theta)$, where $\psi = (u_2, u_3, \lambda, \omega)$ (non-random-preference EU models) or $\psi = (u_2, u_3, \gamma, \lambda, \omega)$ (non-random-preference RDEU models) are in all cases modeled as having a perfect correlation between $\ln(u_2^n - 1)$ and $\ln(u_3^n - 1)$ in the population, generated by an underlying standard normal deviate x_u. Quite similarly, individual estimations of the two random preference models with shape parameters ϕ_1^n and ϕ_2^n yield quite high Pearson correlations between $\ln(\tilde{\phi}_1^n)$ and $\ln(\tilde{\phi}_2^n)$ across subjects, and heavy loadings of these on first principal components of estimated parameter vectors $\tilde{\psi}^n$. Therefore, the joint distributions $J(\psi \mid \theta)$, where $\psi = (\phi_1, \phi_2, \kappa, \omega)$ for the (EU,RP) model and $\psi = (\gamma, \phi_1, \phi_2, \kappa, \omega)$ for

the (RDEU,RP) model, are both modeled as having a perfect correlation between $\ln(\phi_1^n)$ and $\ln(\phi_2^n)$ in the population, generated by an underlying standard normal deviate x_ϕ.

In all cases, all other model parameters are characterized as possibly partaking of some of the variance represented by x_u or x_ϕ, but also having independent variance represented by an independent standard normal variate. In essence, all correlations between model parameters are represented as arising from a single underlying first principle component (x_u or x_ϕ) which in all cases accounts for two-thirds (frequently more) of the shared variance of parameters in $\tilde{\psi}^n$ according to the principal components analyses. The correlation is assumed to be a perfect one for $\ln(u_2^n - 1)$ and $\ln(u_3^n - 1)$ (in non-random-preference models) or $\ln(\phi_1^n)$ and $\ln(\phi_2^n)$ (in random preference models), since this seems very nearly characteristic of all individual model estimates; but aside from ω, other model parameters are given their own independent variance since their correlations with $\ln(u_2^n - 1)$ and $\ln(u_3^n - 1)$ are almost always weaker than that observed between $\ln(u_2^n - 1)$ and $\ln(u_3^n - 1)$ (similarly for $\ln(\phi_1^n)$ and $\ln(\phi_2^n)$ in random preference models).

The following equation systems show the characterization for all models, where any subset of x_u, x_ϕ, x_λ, x_κ and x_γ found in each characterization are jointly independent standard normal variates. Tremble probabilities ω are modeled as constant in the population, for reasons discussed in section 6.1, and so there are no equations below for ω. The systems present just the considered choice probabilities P_m^c; the overall choice probabilities $P_m = (1-\omega)P_m^c + \omega/2$ are of course used for estimation. As in the text, Λ, G and B' are the Logistic, Gamma and Beta-prime cumulative distribution functions, respectively.

(A.1) (EU,Strong), (EU,Strict), (EU,WV) and (EU,CU) models: $P_m^c(x_u, x_\lambda, \theta) =$

$$\Lambda\big(\lambda(x_u, x_\lambda, \theta)[f(s_{m1} + s_{m2}u_2(x_u, \theta) + s_{m3}u_3(x_u, \theta)) - f(r_{m1} + r_{m2}u_2(x_u, \theta) + r_{m3}u_3(x_u, \theta))]/d_m(\theta, x_u)\big),$$

where $f(x) = \ln(x)$ for Strict, and $f(x) = x$ for Strong, WV and CU;

$d_m(\theta, x_u) = u_2(x_u, \theta)\ \forall\ m \in m3,\ u_3(x_u, \theta)\ \forall\ m \in m2$, and $u_3(x_u, \theta) - 1\ \forall\ m \in m0$ for CU,

$d_m(\theta, x_u) = \left(\sum_{z=0}^{3}(s_{mz} - r_{mz})^2\right)^{0.5}\ \forall\ m$ for WV, and $d_m(\theta, x_u) \equiv 1\ \forall\ m$ for Strong and Strict; and

$u_2(x_u, \theta) = 1 + \exp(a_2 + b_2 x_u)$, $u_3(x_u, \theta) = 1 + \exp(a_3 + b_3 x_u)$, and

$\lambda(x_u, x_\lambda, \theta) = \exp(a_\lambda + b_\lambda x_u + c_\lambda x_\lambda)$, where $\theta = (a_2, b_2, a_3, b_3, a_\lambda, b_\lambda, c_\lambda, \omega)$.

(A.2) (EU,RP) model: $P_m^c(x_\phi, x_\kappa, \theta) =$

$$G\left(\frac{s_{m1} + s_{m2} - r_{m1} - r_{m2}}{r_{m2} - s_{m2}} \mid \phi_1(x_\phi, \theta), \kappa(x_\phi, x_\kappa, \theta)\right)\ \forall\ m \in m3,$$

$$G\left(\frac{s_{m1} + s_{m3} - r_{m1} - r_{m3}}{r_{m3} - s_{m3}} \mid \phi_1(x_\phi, \theta) + \phi_2(x_\phi, \theta), \kappa(x_\phi, x_\kappa, \theta)\right)\ \forall\ m \in m2\ \text{ and}$$

$$B'\left(\frac{s_{m2} + s_{m3} - r_{m2} - r_{m3}}{r_{m3} - s_{m3}} \mid \phi_2(x_\phi, \theta), \phi_1(x_\phi, \theta)\right)\ \forall\ m \in m0\ ;\ \text{and}$$

$\phi_1(x_\phi, \theta) = \exp(a_1 + b_1 x_\phi)$, $\phi_2(x_\phi, \theta) = \exp(a_2 + b_2 x_\phi)$, and $\kappa(x_\phi, x_\kappa, \theta) = \exp(a_\kappa + b_\kappa x_\phi + c_\kappa x_\kappa)$,

where $\theta = (a_1, b_1, a_2, b_2, a_\kappa, b_\kappa, c_\kappa, \omega)$.

(A.3) (RDEU,Strong), (RDEU,Strict), (RDEU,WV) and (RDEU,CU) models: $P_m^c(x_u, x_\lambda, x_\gamma, \theta) =$

$$\Lambda\big(\lambda(x_u, x_\lambda, \theta)[f(ws_{m1}(x_u, x_\gamma, \theta) + ws_{m2}(x_u, x_\gamma, \theta)u_2(x_u, \theta) + ws_{m3}(x_u, x_\gamma, \theta)u_3(x_u, \theta)) -$$

$$f(wr_{m1}(x_u, x_\gamma, \theta) + wr_{m2}(x_u, x_\gamma, \theta)u_2(x_u, \theta) + wr_{m3}(x_u, x_\gamma, \theta)u_3(x_u, \theta))]/d_m(\theta, x_u)\big),$$

where $f(x) = \ln(x)$ for Strict, and $f(x) = x$ for Strong, WV and CU;

$d_m(\theta, x_u) = u_2(x_u, \theta)\ \forall\ m \in m3,\ u_3(x_u, \theta)\ \forall\ m \in m2$, and $u_3(x_u, \theta) - 1\ \forall\ m \in m0$ for CU,

$d_m(\theta, x_u) = \left(\sum_{z=0}^{3} (s_{mz} - r_{mz})^2\right)^{0.5} \forall\ m$ for WV, and $d_m(\theta, x_u) \equiv 1 \forall\ m$ for Strong and Strict;

$ws_{mi}(x_u, x_\gamma, \theta) = \quad w\left(\sum_{z \geq z_{mi}} s_{mz} \mid \gamma(x_u, x_\gamma, \theta)\right) - \ w\left(\sum_{z > z_{mi}} s_{mz} \mid \gamma(x_u, x_\gamma, \theta)\right)$ and

$wr_{mi}(x_u, x_\gamma, \theta) = \quad w\left(\sum_{z \geq z_{mi}} r_{mz} \mid \gamma(x_u, x_\gamma, \theta)\right) - w\left(\sum_{z > z_{mi}} r_{mz} \mid \gamma(x_u, x_\gamma, \theta)\right)$, where

$w(q \mid \gamma(x_u, x_\gamma, \theta) = \exp\left(-[-\ln(q)]^{\gamma(x_u, x_\gamma, \theta)}\right)$; and

$u_2(x_u, \theta) = 1 + \exp(a_2 + b_2 x_u)$, $u_3(x_u, \theta) = 1 + \exp(a_3 + b_3 x_u)$, $\gamma(x_u, x_\gamma, \theta) = \exp(a_\gamma + b_\gamma x_u + c_\gamma x_\gamma)$,

and $\lambda(x_u, x_\lambda, \theta) = \exp(a_\lambda + b_\lambda x_u + c_\lambda x_\lambda)$, where $\theta = (a_2, b_2, a_3, b_3, a_\gamma, b_\gamma, c_\gamma, a_\lambda, b_\lambda, c_\lambda, \omega)$.

(A.4) (RDEU,RP) models: $P_m^c(x_\phi, x_\kappa, x_\gamma, \theta) =$

$$G\left(\frac{w[s_{m1} + s_{m2} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[r_{m1} + r_{m2} \mid \gamma(x_\phi, x_\gamma, \theta)]}{w[r_{m2} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[s_{m2} \mid \gamma(x_\phi, x_\gamma, \theta)]} \mid \phi_1(x_\phi, \theta), \kappa(x_\phi, x_\kappa, \theta)\right) \forall\ m \in m3\ ,$$

$$G\left(\frac{w[s_{m1} + s_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[r_{m1} + r_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)]}{w[r_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[s_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)]} \mid \phi_1(x_\phi, \theta) + \phi_2(x_\phi, \theta), \kappa(x_\phi, x_\kappa, \theta)\right) \forall\ m \in m2$$

and $B'\left(\frac{w[s_{m2} + s_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[r_{m2} + r_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)]}{w[r_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)] - w[s_{m3} \mid \gamma(x_\phi, x_\gamma, \theta)]} \mid \phi_2(x_\phi, \theta), \phi_1(x_\phi, \theta)\right)\ \forall\ m \in m0$;

$w(q \mid \gamma(x_\phi, x_\gamma, \theta) = \exp\left(-[-\ln(q)]^{\gamma(x_\phi, x_\gamma, \theta)}\right)$; and

$\phi_1(x_\phi, \theta) = \exp(a_1 + b_1 x_\phi)$, $\phi_2(x_\phi, \theta) = \exp(a_2 + b_2 x_\phi)$, $\gamma(x_\phi, x_\gamma, \theta) = \exp(a_\gamma + b_\gamma x_\phi + c_\gamma x_\gamma)$, and

$\kappa(x_\phi, x_\kappa, \theta) = \exp(a_\kappa + b_\kappa x_\phi + c_\kappa x_\kappa)$, where $\theta = (a_1, b_1, a_2, b_2, a_\gamma, b_\gamma, c_\gamma, a_\kappa, b_\kappa, c_\kappa, \omega)$.

For models with the EU structure, a likelihood function nearly identical to equation 6.16 is maximized in θ; for instance, for the (EU,RP) model simply replace $P_m(x_u, x_\lambda, \theta)$ with $P_m(x_\phi, x_\kappa, \theta)$, and replace $d\Phi(x_u)d\Phi(x_\lambda)$ with $d\Phi(x_\phi)d\Phi(x_\kappa)$. For models with the RDEU structure, a third integration appears since these models allow for independent variance in γ (the Prelec weighting function parameter) through the addition of a third standard normal variate x_γ. In all cases, the integrations are carried out by gauss-hermite quadrature. For models with the EU structure, where there are two nested integrations, 14 nodes are used for each nested quadrature of an integral. For models with the RDEU structure, 10 nodes are used for each nested quadrature. In all cases, starting values for these numerical maximizations are computed in the manner described in section 6.1 for the (EU,Strong) model: Parameters in $\tilde{\psi}^n$ are regressed on their first principle component, and the intercepts and slopes of these regressions are the starting values for the *a* and *b* coefficients in the models, while the root mean squared errors of these regressions are the starting values for the *c* coefficients found in the equations for λ, κ and/or γ. Appendix Tables A1 and A2 show the final parameter estimates for the Strong, CU and RP stochastic models with the EU and RDEU structures, estimated on the selected 68 Hey and Orme (1994) subjects. These are the parameters used to create simulated data sets in section 6 for Monte Carlo analysis and parametric bootstraps. Estimates for the WV and Strict models are not shown since they tend to fare more poorly and receive no Monte Carlo analysis in section 6.

Appendix Table A1. Random parameters estimates on the Hey and Orme data (68 selected subjects): Strong utility and contextual utility models on contexts (0,1,2), (0,1,3) and (1,2,3) combined (data-generating processes for creating simulated data used for parametric bootstraps).

		Strong Utility				Contextual Utility			
		EU structure		RDEU structure		EU structure		RDEU structure	
Structural and stochastic parameter models	Distributional parameter	estimate	asymptotic standard error	estimate	asymptotic standard error	estimate	asymptotic standard error	estimate	asymptotic standard error
$u_2 = 1+\exp(a_2 + b_2 x_u)$	a_2	–1.268	0.046	–1.215	0.055	–1.417	0.064	–1.398	0.056
	b_2	0.515	0.032	0.412	0.030	0.999	0.078	0.832	0.056
$u_3 = 1+\exp(a_3 + b_3 x_u)$	a_3	–0.631	0.037	–0.596	0.064	–0.813	0.068	–0.800	0.055
	b_3	0.627	0.034	0.512	0.054	1.135	0.079	0.954	0.054
$\gamma = \exp(a_\gamma + b_\gamma x_u + c_\gamma x_\gamma)$	a_γ	—	—	–0.081	0.020	—	—	–0.077	0.026
	b_γ	—	—	–0.008	0.008	—	—	–0.023	0.016
	c_γ	—	—	0.181	0.016	—	—	0.177	0.022
$\lambda = \exp(a_\lambda + b_\lambda x_u + c_\lambda x_\lambda)$	a_λ	3.362	0.114	3.342	0.368	3.170	0.084	3.253	0.053
	b_λ	–0.500	0.119	–0.335	0.354	0.077	0.085	0.045	0.063
	c_λ	0.624	0.074	0.466	0.221	0.514	0.071	0.424	0.049
ω constant	ω	0.065	0.015	0.049	0.047	0.0332	0.009	0.028	0.009
log likelihood		–4841.50		–4748.03		–4819.91		–4731.77	

Appendix Table A2. Random parameters estimates on the Hey and Orme data (68 selected subjects): Random preference models on contexts (0,1,2), (0,1,3) and (1,2,3) combined (data-generating processes for creating simulated data used for parametric bootstraps).

		Random Preferences			
		EU structure		RDEU structure	
Structural and stochastic parameter models	Distributional parameter	estimate	asymptotic standard error	estimate	asymptotic standard error
$\phi_1 = \exp(a_1 + b_1 x_\phi)$	a_1	0.647	0.164	0.860	0.144
	b_1	0.778	0.169	0.214	0.193
$\phi_2 = \exp(a_2 + b_2 x_\phi)$	a_2	0.558	0.151	0.716	0.154
	b_2	1.010	0.166	0.474	0.131
$\gamma = \exp(a_\gamma + b_\gamma x_\phi + c_\gamma x_\gamma)$	a_γ	—	—	-0.096	0.024
	b_γ	—	—	0.018	0.026
	c_γ	—	—	0.145	0.025
$\kappa = \exp(a_\kappa + b_\kappa x_\phi + c_\kappa x_\kappa)$	a_κ	−1.858	0.173	−1.940	0.163
	b_κ	–0.453	0.165	–0.061	0.203
	c_κ	0.639	0.038	0.444	0.107
ω constant	ω	0.067	0.017	0.050	0.015
log likelihood		–4859.71		–4752.74	