Stochastic Models in Experimental Economics

Brian Albert Monroe

January 29, 2018

Abstract

Shortly after the introduction of Expected Utility Theory (EUT), economists and psychologists began publishing results that showed choices made by experimental subjects which apparently violate one or more of the EUT axioms. I discuss economists' responses to this evidence. These vary from developing new theoretical models, models that nest EUT as a special case, such as Rank Dependent Utility (RDU) and Regret Theory, as well as models that do not nest EUT, such as Cumulative Prospect Theory, to critiques of experimental methods and scope, to the promotion of stochastic models of choice. I discuss popular stochastic choice models in depth and evaluate their normative coherence. I find that the "Random Preferences" stochastic model fails to make normatively coherent statements, in contrast to the "Random Error" and "Tremble" models, which do so. I demonstrate a method to calculate the unconditional likelihood of choice errors for populations of EUT-compliant and RDU-compliant agents, and show how certain characteristics of the population relate to the likelihood of these choice errors and their costliness in terms of forgone welfare. I find that elements of the stochastic model that are not related to preference relations tend to have a greater influence on unconditional welfare estimates than the preference parameters themselves. Finally, I conduct a power analysis of the ability of a lottery battery instrument to correctly classify experimental subjects as employing either EUT or RDU, and the effect of this classification on the accuracy of the estimates of welfare surplus for the subjects. For large ranges of parameter values for these models, I find that the probability of type I and type II errors in the classification process are not trivial,

and can be very costly in terms of welfare surplus. Additionally I show that for a hypothetical population comprising subjects employing EUT or RDU, we can arrive at more accurate welfare surplus estimates on average by assuming that every subject employs the RDU functional, rather than by first trying to differentiate RDU subjects from EUT subjects.

Acknowledgements

I would like to thank my supervisors, Dr. Glenn Harrison and Dr. Don Ross, for their tremendous support and guidance throughout this process. I know I would not have wanted to do this thesis with any other supervisors. Their tutelage has given me an invaluable first step into my career as an economist. Also, a strangely debilitating fear of red ink and cursive.

I came to find that writing a thesis focused on such a specific topic as stochastic models can be an intellectually lonely endeavour. I'd like to thank my good friend and economist Dr. Andre Hofmeyr for his friendship throughout this process and always being available to talk economics over beer and hikes. His expertise helped refine some of my better ideas and weed out the weaker ones.

To Dr. Lisa Rutström, for recruiting me to run her experiments at the University of Central Florida, convincing me to continue my academic career by pursuing my masters at Georgia State University and finally encouraging me to pursue a PhD at the University of Cape Town. Indeed, this thesis would never have been written had she not encouraged me to do more and be better at every stage of my academic career. Thank you.

My wife, Quanita, for helping me though the joy and heartbreak of the aforementioned red pen and cursive comments and having the grace to smile and nod when I used her as a sounding board.

Contents

Chapter 1

Choice Anomalies in Experiments, and Economists' Reactions to Them

1.1 "Preference Reversals" and the Grether and Plott (1979) Experiments

Grether and Plott (1979) describes and tests for explanations of "preference reversal" phenomena that had been observed in previous studies by psychologists, in particular Lichtenstein and Slovic (1971, 1973) and Lindman (1971). In these experiments subjects were asked to directly state a preference for one of two bets, termed a P bet and a \$ bet, and then to state a price at which they would be willing to sell each bet. These stated preferences generate an implied preference over the two bets. The observed phenomenon was of subjects stating a preference for either the P or the \$ bet in the direct comparison, and then stating a higher selling price for the opposite bet. This type of behavior was deemed a "preference reversal" because the stated preferences and the implied preferences were inconsistent. This preference reversal was said to be incompatible with Expected

Utility Theory (EUT), which, assuming a deterministic choice process and ignoring indifference for now, requires that a subject state a higher price for the bet which she selected in the direct comparison. Grether and Plott (1979, p. 623) set about to conduct "a series of experiments designed to discredit the psychologists' works as applied to economics" and ended up "as perplexed as the reader who has just been introduced to the problem." (1979, p. 624) after failing to substantially reduce the observed inconsistencies in their own controlled experiment.

Grether and Plott (1979) identified 13 possible theoretical criticisms and/or explanations of this phenomenon, of which 3 related to economic theory, 6 were psychological in nature, and 4 were artifacts of experimental method. Of greatest concern to experimental economists should be the explanations involving experimental method and economic theory. The possible explanations concerning economic theory included misspecified incentives, income effects, and indifference. The possible explanations involving experimental method included confusion and misunderstanding, low frequency of errors/small sample sizes, unsophisticated subjects, and, my favorite, that the experimenters were psychologists. Grether and Plott then detail the ways in which each of these possibilities could potentially lead to the observed seemingly theory-inconsistent data, discuss how the previous literature by psychologists inadequately control for these various possibilities, and how their experiment attempts to impose adequate controls.

In identifying these possibilities, Grether and Plott (1979) touch on aspects of conducting economic laboratory experiments which are later codified as necessary precepts for valid controlled experiments by Smith (1982). These precepts will be elaborated upon later.

To better understand the nature of the preference reversal phenomenon as

described by Grether and Plott (1979), the subsequent critiques of their method, and potential accommodations of these seemingly inconsistent data, a description of the relevant details of the experiment is needed. Subjects were students recruited from economics and political science classes, promised a minimum of $5 for participation and told that the experiment would take no longer than one hour.

The experiment entailed the subjects making 2 types of choices. The first asked them to state either a strict preference for a P bet or a $ bet, where the terms are defined below, or stating that they "Don't care" across 6 pairs of P and $ bets. The second asked subjects to state "the smallest price for which you would sell a ticket to the following bet" (1979, p. 630) for each P bet and $ bet across the same 6 pairs. Though each pair of P bets and $ bets were for different amounts, the relationship between the P bets and $ bets remained similar across all 6 pairs. The P bet, named for being a "probability" bet, is a lottery between winning a small but positive amount of money, P_1 with a high probability (never less than 29/36), and incurring a small loss, P_0 on the subject's initial endowment with a low probability (never more than 7/36). The $ bet, named for being a "money" bet, is a lottery between winning a large amount of money, $\$_1$, with a low probability (never more than 18/36), and incurring a small loss, $\$_0$, on the subject's initial endowment with a relatively high probability (never less than 18/36).

Grether and Plott (1979) conducted two different experiments eliciting the above choices which varied slightly in order to test one of the psychological theories for the preference reversal phenomenon. In the first experiment the subjects were split into two groups in which one group was asked to make a series of hypothetical decisions, for which they would be paid a guaranteed $7, while the other group was initially endowed $7, but told that their final earnings would depend on one of

their choices chosen at random and being played out. Both groups were asked their preferences for pairs 1, 2, and 3, then asked their minimum selling price for all 12 gambles, then asked their preferences for pairs 4, 5, and 6. The second experiment was also split into two groups but both groups were paid based on their decisions and they were also asked for "the exact dollar amount such that you are indifferent between the bet and the amount of money" in addition to being asked for a selling price directly. The first group was asked for a "selling price" first while the second group was asked for the "dollar equivalent" amount first. This additional set of questions were implemented to control for potential strategic behavior that might be associated with the term "selling."

All of the groups that were incentivized with real monetary rewards were told that they would only be paid for one of their choices. This was meant to combat the income effect of accumulating earnings across many questions. If a selling price question was selected for payment, the experimenters would use the method detailed in Becker, DeGroot and Marschak (1964) (BDM). In this method, a subject is asked to report the lowest price she is willing to accept to give up her right to play a certain lottery. The experimenter then selects a "buying" price from a uniform distribution between two feasible price intervals and if the "selling" price reported by the subject is less than the selected buying price the subject receives the buying price, otherwise the subject plays out the lottery; in this experiment the random distribution was between \$0.00 and \$9.99. BDM explain how this type of auction mechanism leads to the subject's true selling price being the weakly dominant response, at least in theory.

Despite having conducted this experiment in expectation of refuting the results of psychologists, Grether and Plott (1979) end up confirming these results with their

own experiment. The lack of a substantial reduction in the proportion of subjects displaying the preference reversal phenomenon, particularly in the groups which had monetary incentives attached to their choices, led Grether and Plott (1979) to suggest that certain assumptions economists had held concerning the structure of preferences may not be valid. Grether and Plott (1979, p. 623) remarked concerning the preference reversal phenomenon "The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind even the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior may result from principles which are of a completely different sort from those generally accepted."

The totals of the different choices for the groups of incentivized subjects are presented in Table 1.1 below. Experiment 1 in Table 1.1 only includes the group which was paid with real money, while Experiment 2-1 and Experiment 2-2 in Table 1.1 represent experiment 2 groups 1 and 2, respectively, where the differences between the groups are detailed above. The mean difference in reported prices for inconsistent choices from experiment 1 is reported in Table 1.2, Grether and Plott (1979) did not report these statistics for experiment 2. In Table 1.2 the "Predicted" preference reversal is that of selecting the P bet in the direct comparison and stating a higher selling price for the \$ bet, and the "Unpredicted" preference reversal is selecting the \$ bet in the direct comparison and stating a higher selling price for the P bet.

In Table 1.1 two things are apparent. The first is the degree of inconsistent choices from subjects who chose the P bet in a binary choice option and then reported a higher price for the \$ bet from exactly the same pair in the BDM

Table 1.1: Grether and Plott (1979) - Results for Incentivized Experiments

Bet	Choices	Reservation Prices: Consistent	Inconsistent	Equal	%Consistent	%Inconsistent
Experiment 1						
P	99	26	69	4	26.26%	69.70%
$	174	145	22	7	83.33%	12.64%
Indifferent	3					
Experiment 2-1						
Selling Price						
P	44	8	30	6	18.18%	68.18%
$	72	54	15	3	75.00%	20.83%
Indifferent	4					
Equivalent Price						
P	44	4	34	6	9.09%	68.18%
$	72	59	11	2	81.94%	15.28%
Indifferent	0					
Experiment 2-2						
Selling Price						
P	44	16	27	1	36.36%	61.36%
$	64	54	9	1	84.38%	14.06%
Indifferent	0					
Equivalent Price						
P	44	19	22	3	43.18%	50.00%
$	72	51	10	3	79.69%	15.63%
Indifferent	0					

Table 1.2: Grether and Plott (1979) - Experiment 1:
Mean Values of Reversals (in Dollars)

	Predicted		Unpredicted	
Bet	Incentives	No Incentives	Incentives	No Incentives
1	1.71	2.49	0.40	0.79
2	1.45	2.65	0.51	0.90
3	1.48	1.29	1.00	0.25
4	3.31	5.59	3.00	0.02
5	1.52	1.79	0.38	0.01
6	0.92	1.18	0.33	0.31

task. Of all the groups with incentivized choices, no fewer than 50% of those who chose the P bet reported a selling price for the P bet that was at least 1 cent below the selling price reported for the $ best. Table 1.2 shows that the average difference between the elicited selling price and the expected value of the lottery in question ranged from 1 cent to $5.59. The magnitude of these differences formed an important part of the critique of these experiments by Harrison (1989, 1992), to be discussed later.

The second implication from Table 1.1 is the asymmetry of the frequency of choice inconsistencies. While the maximum proportion of inconsistency for subjects selecting a P bet choice in the binary comparison was about 77%, the maximum proportion of inconsistency for selecting a $ bet was only about 21%. Also, the mean value of the reversal is larger for the "Predicted" reversal for every bet. Subsequent critiques of these experiments ignore this asymmetry, suggest that it is meaningless due to a failure to satisfy either the saliency or dominance precepts initially proposed by Smith (1982), to be discussed later, or suggest that

this asymmetry is predicted by alternative theories to EUT, such as the "Regret Theory" of Loomes & Sugden (1982).

1.2 Theoretical Critiques of the Grether and Plott (1979) Experiments

Holt (1986) offers an explanation of the preference reversal phenomenon which does not require the forfeiture of transitivity. Instead, Holt (1986) proposes that should the independence axiom not hold, subjects are not making choices which are separable, but instead are making choices between compound lotteries. Choices over compounded lotteries may display the preference reversal phenomenon without being a violation of transitivity.

Take for example a three question scenario which represents the Grether and Plott (1979) instrument, where a subject must make a choice between a P bet and a \$ bet, and the subject is asked to reveal her selling price for both the P bet and the \$ bet using the BDM mechanism. Suppose also that only one of these three choices will be played out for real earnings. The expected utility of such a scenario would be the following:

$$\frac{1}{3}\mathrm{E}\{u(w+\tilde{x})\}+\frac{1}{3}\mathrm{E}\{u(w+\tilde{b}(r_{\$};X_{\$}))\}+\frac{1}{3}\mathrm{E}\{u(w+\tilde{b}(r_P;X_P))\} \tag{1.1}$$

where w is the initial wealth of the subject, $\tilde{x}$ is the random payment determined by the chosen lottery, $X_{\$}$ or X_P, and $\tilde{b}$ is the random payment of the BDM mechanism given the elicited selling price r for each respective lottery, $X_{\$}$ or X_P.

Given equation (1.1) and the validity of the independence axiom, the subject should choose the lottery in the binary comparison which she prefers most, and reveal the minimum price for which she is willing to sell each lottery in order to

maximize her utility. The ranking of the minimum reservation prices should also correspond to the selection of lotteries in the binary choice. However, suppose that the binary choice came after the selling price elicitations. Then the subject can be said to be making a choice between two compound lotteries:

$$\left[\frac{1}{3}, X_P; \frac{1}{3}, B(\hat{r}_\$; X_\$); \frac{1}{3}, B(\hat{r}_P; X_P)\right] \textit{ and } \left[\frac{1}{3}, X_\$; \frac{1}{3}, B(\hat{r}_\$; X_\$); \frac{1}{3}, B(\hat{r}_P; X_P)\right] \quad (1.2)$$

where $B(\hat{r}; X)$ is the lottery resulting from the BDM mechanism between the reservation price $\hat{r}$ and either the $X_\$$ or X_P lotteries. We can rewrite the last part of both these compound lotteries as follows:

$$Z \equiv \left[\frac{1}{2}, B(\hat{(R)}_\$; X_\$); \frac{1}{2}, B(\hat{r}_P; X_P)\right] \quad (1.3)$$

and (1.2) can be rewritten as

$$\left(\frac{1}{3}, X_\$; \frac{2}{3}, Z\right) \textit{ and } \left(\frac{1}{3}, X_P; \frac{2}{3}, Z\right) \quad (1.4)$$

If independence holds, the subject would choose the left option if, and only if, she preferred $X_\$$ to X_P, and vise versa. Should the independence axiom fail, then this relationship also fails and the subject may choose the left option even if she doesn't prefer $X_\$$ to X_P. If such a choice is made, then the price revealed for the chosen lottery using the BDM mechanism may be lower than the price solicited for the alternative lottery because the impact of such a choice has been diluted by the compound lottery.

Holt (1986) offers an explanation for the preference reversal phenomenon, but makes no comment about the asymmetry of the phenomenon, nor does he suggest an alternative theory to EUT other than to state that "...any theory of rational choice in such contexts must be derived from a set of axioms that does not include

or imply the independence axiom..." (1986 p.514) Holt (1986) does state that since the choices are diluted by compounding the opportunity cost of an apparent preference reversal is very low, thus the asymmetry may not be so interesting. This point is elaborated by Harrison (1989, 1992) and will be discussed later.

Karni and Safra (1987) also state that the preference reversal phenomenon can be explained by an alternative to EUT which doesn't include the independence axiom. Differing from Holt (1986) however, they propose a model that would explain the preference reversal phenomenon which they call "Expected Utility with Rank Dependent Utilities" (EURDU). EURDU requires that the independence axiom not hold, that the possible outcomes of a given lottery be ranked, the probabilities of each of the outcomes be transformed, and weights of the outcomes generated by these transformed probabilities be dependent on the rank of the outcome. The first axiomatization of a EURDU model was by Quiggin (1982), though it was called "Anticipated Utility" at the time, and now is more commonly called "Rank Dependent Utility" (RDU). Yaari (1987) independently proposed a utility theory which replaces the independence axiom with a dual independence axiom, resulting in a lottery valuation functional that is linear in prizes but nonlinear in probabilities, called "Dual Theory."

Assuming some discrete lottery, $(x_1, p_1; \ldots; x_n, p_n)$, that the probabilities associated with the outcomes of the lotteries sum to 1, and that $x_1 \leq x_2 \leq \ldots \leq x_n$, then the following function due to Karni and Safra (1987) represents the EURDU of such a lottery:

$$V(x_1, p_1; \ldots; x_n, p_n) = \sum_{i=1}^{n} u(x_i) \left[f\left(\sum_{j=i}^{n} p_j \right) - f\left(\sum_{j=i+1}^{n} p_j \right) \right] \tag{1.5}$$

The above utility structure combined with the compound lottery structure,

necessitated by the lack of the independence axiom and the BDM mechanism, is sufficient to generate utility maximizing choices which appear to be preference reversals. To demonstrate this, Karni and Safra (1987) proposed a simplified version of the BDM mechanism in which a subject must select a selling price, s, from the set $\{1, 2, 3, 4, 5\}$, and the experimenter randomly selects an buying price from the set $\{1, 2, 3, 4, 5\}$ to determine the outcome. Recall that if the buying price is greater than the selling price, the subject receives the buying price, otherwise she plays out the lottery. In addition, they proposed the following candidate probability weighting function and utility function :

$$f(p) = \begin{cases} 1.1564p, & 0 \leq p \leq 0.1833 \\ 0.9p + 0.047, & 0.1833 \leq p \leq 0.7 \\ 0.5p + 0.327, & 0.7 \leq p \leq 0.98 \\ p^{10}, & 0.98 \leq 1 \end{cases} \tag{1.6}$$

$$u(x) = \begin{cases} 30x + 30, & x \leq -1 \\ 10x + 10, & -1 \leq x \leq 12 \\ 6.75x + 49, & 12 \leq x \end{cases} \tag{1.7}$$

Given equations (1.6) and (1.7), an outcome 10 with a probability of .9 would have a probability weight of $f(0.9) = 0.5 \times 0.9 + 0.327 = 0.77$, and a utility of $u(10) = 10 \times 10 + 10 = 110$.

Using two lotteries from the Grether and Plott (1979) experiments (A and B

below), they set up the value function:

$$A = (-1, \frac{1}{36}\,;\, 4, \frac{35}{36}) \quad \textit{and} \quad B = (-1.5, \frac{25}{36}\,;\, 16, \frac{11}{36}) \tag{1.8}$$

$$\begin{aligned} V(A|\Pi(A)) &\equiv V\left(A, \frac{s}{5}\,;\, s+1, \frac{1}{5}\,;\, \ldots\,;\, 5\frac{1}{5}\right) \\ &= V\left(-1, \frac{s}{5} \times \frac{1}{36}\,;\, 4, \frac{s}{5} \times \frac{35}{36}\,;\, s+1, \frac{1}{5}\,;\, \ldots\,;\, 5, \frac{1}{5}\right) \end{aligned} \tag{1.9}$$

where lottery A is a P bet, lottery B is a \$ bet, and $\Pi(A)$ returns the selling price of lottery A elicited from the BDM mechanism. Solving equations (1.8) and (1.9) for the possible values of s produces the following Table from Karni and Safra (1987, p. 679):

Table 1.3: Karni and Safra (1987) - Values for Different Choices of $\Pi(k), k = A, B$

	$V(A\|\Pi(A))$	$V(B\|\Pi(B))$	Value of $\Pi(k)$
1	$40.65 = V(A)$	$40.38 = V(B)$	$5 < \Pi(k)$
2	43.06	44.42	$4 \leq \Pi(k) \leq 5$
3	44.53	46.66^*	$3 \leq \Pi(k) \leq 4$
4	45.25^*	45.06	$2 \leq \Pi(k) \leq 3$
5	43.70	39.61	$1 \leq \Pi(k) \leq 2$
6	39.48	39.48	$\Pi(k) \leq 1$
7	$3.065 = C(A)$	$3.038 = C(A)$	$3.038 = C(B)$

In row 1 of Table 1.3 the conditional values of A and B are equivalent to the unconditional values of A and B because if a subject could have chosen a price greater than 5 in this example the subject would have played out either A or B with 100% probability. The asterisks indicate at which values of s the conditional value of A or B is maximized. Row 7 indicates the certainty equivalents A and B which are a direct result of the monotonicity of $u(\cdot)$ and by definition

$u(C(H)) = V(C(H), 1)$ for every H which is an element of the set of lotteries offered.

From row 1 we can see that the subject prefers lottery A to B and should choose A in the direct lottery choice. From row 7 we can see that the true certainty equivalent of lottery A is greater than lottery B. From rows 3 and 4 we see that when the subject is asked for a selling price for each lottery she would choose a price between 2 and 3 for lottery A, and a price between 3 and 4 for lottery B. The true certainty equivalent of lottery A is out of the range of selling prices which would maximize this subject's utility for this compound lottery, and the utility maximizing selling price for B is greater than that of A. This subject would therefore display a "preference reversal" by selecting A in the direct comparison and selecting a higher price for B with the BDM mechanism.

Using the penny grid employed in the Grether and Plott (1979) experiments, Karni and Safra (1987, p. 680) calculated the selling price that would have been elicited from the same subject for lottery A as \$3.43, and for lottery B as \$4.33. Because the unconditional values of lottery A and lottery B along with their respective certainty equivalents are the same as in the first row and seventh row of Table 1.3, these elicited selling prices also display a "preference reversal." The elicited selling prices using this penny grid are also significantly different from the certainty equivalents as calculated in row 7.

In contrast to the critiques of Holt (1986) and Karni and Safra (1987), Loomes, Starmer and Sugden (1989) suggest that the apparent preference reversals are in fact violations of transitivity as Grether and Plott (1979, p. 623) had stated and offer "Regret Theory" as a potential explanation, which predicts the apparent preference reversal and lacks the transitivity axiom. Regret Theory was originally

and independently developed by Loomes and Sugden (1982) and Bell (1982), then axiomized by Fishburn (1987) and modified by Loomes and Sugden (1987) to include a "convexity" axiom to be described below.

Regret Theory assumes that a subject has a "choiceless" utility function, which is unique up to a linear transformation and assigns a real-valued utility number to every conceivable outcome of an action. This utility function represents the utility the subject would derive from the outcome of an action if she experienced it without having chosen it (Loomes and Sugden 1982, p. 807). The "choiceless" utility of an outcome that would occur in a particular state of the world given an action is compared with the "choiceless" utility of an outcome that would occur in the same state of the world given another feasible action using a "modified" utility function:

$$m_{ij}^{k} = M(c_{ij}, c_{kj}) \tag{1.10}$$

where M is the modified utility function, c_{ij} and c_{kj} are the "choiceless" utilities of the outcome of actions i and k, respectively, in the event that state of the world j occurs, and m_{ij}^{k} is the resulting modified utility of having chosen action i instead of action k.

Loomes and Sugden (1982, p. 809) assume that the degree of regret only depends on the difference in the "choiceless" utilities which would occur in the same state of the world but given different actions. Thus (1.10) can be re-written as follows:

$$m_{ij}^{k} = c_{ij} + R(c_{ij} - c_{kj}) \tag{1.11}$$

where $R(\cdot)$ is a "regret-rejoice" function. A subject would select an action which has the greatest expected modified utility; the sum of the probability weighted

modified utilities across all potential states of the world. Keeping with the notation of Loomes and Sugden (1982), m_{ij}^k will be rewritten as $\psi(x_{ij}, x_{kj})$, where x_{ij} and x_{kj} are the outcomes of actions i and k, respectively, should state of the world j occur. Thus $\psi(\cdot)$ incorporates the transformation of the outcomes into "choiceless" utilities, and those "choiceless" utilities into the modified utilities.

$$A_i \succcurlyeq A_k \Leftrightarrow \sum_j p_j \, \psi(x_{ij}, x_{kj}) \geq 0 \tag{1.12}$$

where A_i and A_k are actions i and k, respectively. Two important assumptions are made about $\psi(\cdot)$: first, $\psi(\cdot)$ is skew-symmetric, i.e. $\psi(x_{ij}, x_{kj}) = -\psi(x_{ij}, x_{kj})$, and second it is convex, i.e. if $x_3 > x_2 > x_1$, then $\psi(x_3, x_1) > \psi(x_3, x_2) + \psi(x_2, x_1)$. This convexity allows subjects to be "regret" averse and is the basis for the prediction of the preference reversal phenomenon.

Using this notation, Loomes, Starmer and Sugden (1989) denote the P bets and \$ bets as actions, and modify the choice of a selling price for the BDM mechanism as a binary choice between an action which returns a constant amount of money for certain, C, and either the P bet or the \$ bet. where the columns represent

Table 1.4: Loomes, Starmer and Sugden (1989, p. 141)
Actions, States, and Outcomes

Action	S_1 p_1	S_2 p_2	S_3 p_3	S_4 p_4
\$	a	a	d	d
P	b	e	b	e
C	c	c	c	c

4 states of the world, $S_1, \ldots, S_4$, with associated probabilities $p_1, \ldots, p_4$, and the rows represent the different actions, \$ bet, P bet, and a certainty C, with potential

outcomes a, b, c, d and e such that $a > b > c > d, e$. If $p_1 + p_3 > 0.5 > p_1 + p_2$ then the first 2 rows correspond to the \$ bet and P bet from Grether and Plott (1979). These actions can be made into 3 pairwise choices, $\{P, \$\}, \{P, C\}, \{\$, C\}$, called a "triple." Applying the formula from (1.12) to the outcomes and probabilities of Table 1.4 generates the following preference relations:

$$P \succcurlyeq \$ \Leftrightarrow p_1\psi(b,a) + p_2\psi(e,a) + p_3\psi(b,d) + p_4\psi(e,d) \geq 0 \tag{1.13}$$

$$\$ \succcurlyeq C \Leftrightarrow p_1\psi(a,c) + p_2\psi(a,c) + p_3\psi(d,c) + p_4\psi(d,c) \geq 0 \tag{1.14}$$

$$C \succcurlyeq P \Leftrightarrow p_1\psi(c,b) + p_2\psi(c,e) + p_3\psi(c,b) + p_4\psi(c,e) \geq 0 \tag{1.15}$$

In the above pairwise choices, the most common preference reversal phenomenon of the Grether and Plott (1979) experiments is observed if the P bet is chosen over the \$ bet, the \$ bet chosen over the C money certainty, and the C money certainty chosen over the P bet. The less common preference reversal is observed if the \$ bet is chosen over the P bet, the P bet is chosen over the C money certainty, and the C money certainty is chosen over the \$ bet. The first of these two preference reversals is predicted by Regret Theory when $p_2 = 0$ and $d \geq e$ in equations (1.13), (1.14), and (1.15), while the second one is not. Loomes, Starmer and Sugden (1989, p. 143) call the first of these preference reversals the "predicted" preference reversal, and the second the "unpredicted" preference reversal.

Loomes, Starmer and Sugden (1989) design three experiments to test whether subjects who display the preference reversal phenomenon follow Regret Theory instead of EUT by utilizing the special case of $p_2 = 0$ and $d \geq e$. The first two experiments confront subjects with different sets of triples, called the "choice-only" design, while the third experiment had some subjects face the BDM mechanism

to elicit selling prices for each lottery as in Grether and Plott (1979), called the "standard" design, and some subjects use the "choice-only" design. Loomes, Starmer and Sugden (1989, p. 142) state that their null hypothesis is that subjects make choices in accordance with EUT but make mistakes randomly such that there should be an equal proportion of subjects who display the "predicted" preference reversal and subjects who display the "unpredicted" preference reversal for any given triple. They state that they will reject this EUT null hypothesis in favor of the alternative of Regret Theory if the frequency of subjects displaying the "predicted" preference reversal is significantly higher than subjects displaying the "unpredicted" preference reversal to an extent that can't be attributed to chance.

Experiment 1 had 283 subjects, 120 of which also participated in Experiment 2 which was held a few days later and had some chance of loosing money. Experiment 3 had 186 subjects. All subjects were randomly assigned to different subsamples, with each subsample assigned to a unique set of triples. An equal number of subjects were assigned to the "choice-only" and "standard" designs in Experiment 3, and the responses from the group that participated in the "standard" design were imputed into choices as if they had conducted the "choice-only" design. The bets used in the pairwise comparisons are presented in Table 1.5.

The subjects in Experiment 1 were split into 4 subsamples, A, B, C, and D, each of which made pairwise choices across one triple, $(\$_1, P_1, C)$, where C varied for each subsample as shown in the "Triple" column of Table 1.6 below. The subjects in Experiment 2 were split into 2 subsamples, E and F, each of which faced 4 triples, with each triple containing the $\$_2$ bet and either the P_2 bet or P_3 bet and unique values of C for each triple. The subsamples for Experiment 2 differed only by the values of C in their triples. This is also shown in Table

Table 1.5: Loomes, Starmer and Sugden (1989)
Bets Used in Three Experiments

\$ bets	P bets
$\$_1 = (12.00, 0.4 \; ; \; 0, 0.6)$	$P_1 = (8.00, 0.6 \; ; \; 0, 0.4)$
$\$_1 = (12.00, 0.4 \; ; \; 0, 0.6)$	$P_1 = (8.00, 0.6 \; ; \; 0, 0.4)$
$\$_1 = (12.00, 0.4 \; ; \; 0, 0.6)$	$P_1 = (8.00, 0.6 \; ; \; 0, 0.4)$
$\$_1 = (12.00, 0.4 \; ; \; 0, 0.6)$	$P_1 = (8.00, 0.6 \; ; \; 0, 0.4)$
$\$_1 = (12.00, 0.4 \; ; \; 0, 0.6)$	$P_1 = (8.00, 0.6 \; ; \; 0, 0.4)$

1.6, along with the number of subjects who responded with particular patterns of choice for each triple.

The subjects in Experiment 3 were split into 6 subsamples, G_1, H_1, I_1 for the "standard" design and G_2, H_2, I_2 for the "choice-only" design. Each subsample of the same letter designation faced the same \$ bet and P bet, and after imputing the elicited selling prices from the "standard" design subsamples into a choice of C, all subsamples of Experiment 3 faced the same value of C. The purpose of imputing these choices was to provide a direct comparison of the "standard" design and the "choice-only" design. The number of subjects who displayed each possible choice pattern for Experiment 3 is shown in Table 1.7 below. Some subjects in the "standard" design reported a selling price for a bet which was greater than the largest possible outcome of the bet, or smaller than the smallest possible outcome of the bet. The number of subjects who did not display this "perverse" behavior in the "standard" design is reported in parenthesis next to the total number of subjects displaying a particular choice pattern in Table 1.7, where $r_\$$ and r_P are the elicited selling prices for the \$ bet and P bet, respectively, and $\succ_c$ stands for "chosen over."

Table 1.6: Loomes, Starmer and Sugden (1989)
Results of Experiments 1 & 2

			Pattern of choice							
Triple	Subsample	n	\$ \$ C	\$ \$ P	\$ C C	\$† C† P†	P^* \$* C^*	P \$ P	P C C	P C P
						Experiment 1				
$\$_t, P_t, 2.50(C)$	A	71	2	9	1	1	3	27	19	9
$\$_t, P_t, 3.50(C)$	B	70	1	4	1	2	12	15	32	3
$\$_t, P_t, 4.50(C)$	C	72	1	4	4	0	7	6	44	6
$\$_t, P_t, 5.50(C)$	D	70	5	4	4	0	4	4	50	2
Total		283	9	21	10	3	26	40	145	20
						Experiment 2				
$\$_2, P_2, 2.25(C)$	E	60	14	17	3	1	7	8	8	2
$\$_2, P_2, 3.75(C)$	E	60	17	8	10	0	8	4	13	0
$\$_2, P_3, 2.25(C)$	E	60	7	22	0	1	2	15	9	4
$\$_2, P_3, 3.75(C)$	E	60	7	18	3	2	7	5	14	4
$\$_2, P_2, 3.00(C)$	F	60	19	11	3	2	11	1	9	4
$\$_2, P_2, 4.50(C)$	F	60	20	3	11	1	6	0	18	1
$\$_2, P_3, 3.00(C)$	F	60	9	21	2	1	2	10	7	8
$\$_2, P_3, 4.50(C)$	F	60	17	5	7	4	6	1	14	6

* indicates the "predicted" preference reversal
† indicates the "unpredicted" preference reversal

Table 1.7: Loomes, Starmer and Sugden (1989)
Results of Experiment 3

Actions	Subsample	n	$\$ \succ_c P$ $\$ \succ_c P$	$\$ \succ_c P$ $\$ \succ_c P$	$\$ \succ_c P$ $\$ \succ_c P$	$\$ \succ_c P$ $\$ \succ_c P$	$\$ \succ_c P$ $\$ \succ_c P$	$\$ \succ_c P$ $\$ \succ_c P$
			Group 1 "Standard" Design: Choice and Valuations					
$\$_3, P_4$	G_1	31	4(3)	0(0)	5(2)	11(11)	3(3)	8(6)
$\$_4, P_5$	H_1	31	14(11)	1(1)	5(1)	6(6)	2(1)	3(2)
$\$_5, P_6$	I_1	31	10(9)	2(1)	5(4)	11(10)	0(0)	3(3)
	Total	93	28(23)	3(2)	15(7)	28(25)	5(4)	14(11)

			Pattern of choice							
Triple	Subsample	n	\$ \$ C	\$ \$ P	\$ C C	\$† C† P†	P^* \$* C^*	P \$ P	P C C	P C P
			Group 1 "Standard" Design: Imputed Choice and Valuations							
$\$_3, P_4, 4.5$	G_1	31	1	1	5	1	2.75	6.75	10.25	2.25
$\$_4, P_5, 4.5$	H_1	31	4	4.5	10.5	1	3	2	5	1
$\$_5, P_6, 4.5$	I_1	31	4	3.5	6.5	3	6	2.5	5	0.5
Total		93	9	9	22	6	11.75	11.25	20.25	3.75
			Group 2 "Choice-Only" Design: Imputed Choice and Valuations							
$\$_3, P_4, 4.5$	G_2	31	2	0	4	0	4	2	12	7
$\$_4, P_5, 4.5$	H_2	31	10	2	3	3	6	1	6	0
$\$_5, P_6, 4.5$	I_2	31	3	5	8	1	4	1	8	1
Total		93	15	7	15	4	14	4	26	8

* indicates the "predicted" preference reversal
† indicates the "unpredicted" preference reversal

In every experiment and across every subsample there were statistically significantly more subjects displaying the "predicted" preference reversal than displaying the "unpredicted" preference reversal, leading Loomes, Starmer and Sugden (1989) to reject the null hypothesis of EUT in favor of the alternative of Regret Theory for every experiment.

Regret Theory is unique in its explanation of the preference reversal phenomenon in that, unlike the alternative proposed by Karni and Safra (1987), Regret Theory predicts both the preference reversal phenomenon's existence as well as the asymmetrical distribution between the two possible patterns of preference reversal.

1.3 Necessary Precepts for Valid Inferences from Economic Experiments, and the Violation of these Precepts

Smith (1982) lays out a conceptual framework for modeling microeconomic systems, such as an economic experiment, in terms of an interactive relationship between an environment, an institution, and agent behavior. In this framework agents send messages to the institution which then maps those messages according to pre-set rules into commodity outcomes. This framework provides the valuable insight that "[...] agents do not choose direct commodity allocations. *Agents choose messages, and institutions determine allocations via the rules that carry messages into allocation*" [emphasis in the original] (Smith 1982, p. 926). Thus the data which we observe from an experiment is derived from message producing behavior which is said to be a function of the environment for that agent and the institution. This mapping of messages into allocations does provide the experimenter with valuable information, but only if four sufficient conditions are met for a controlled

microeconomic experiment as discussed in Smith (1982) and Harrison (1989).

1.3.1 Salience and Potential Violations

Of the four sufficient conditions proposed by Smith (1982), the first, non-satiation, is equivalent to the common requirement of most theories of utility that there exists a reward mechanism such that the subject should not be satiated in it, and should be interpreted in the same way. The second and third are of primary importance in an experiment attempting to elicit what Smith refers to as "home-grown" preferences; preferences that are not induced by the experimenter in a laboratory, but instead are the subject's own latent preferences from outside the laboratory.

The second sufficient condition is Saliency. Smith (1982, p. 931) defines this as the condition according to which "Individuals are guaranteed the right to claim a reward which is increasing (decreasing) in the goods (bads) outcomes, x_i, of an experiment." Harrison (1994, p. 223) notes that the above definition combined with the non-satiation requirement leads to a mapping of a message m' to a reward v' instead of m'' that maps into v'', whenever $v' > v''$. This has also been interpreted by Bruner (2011) to mean that the manner in which messages are mapped to rewards by the institution is understood by the subject, even if it is stochastic, otherwise there exists an institutional failure to induce values. In the syntax of Smith (1982):

$$I^i = \left(M^i, h^i(m), c^i(m), g^i(t_0, t, T)\right) \equiv I^i_p = \left(M^i_p, h^i_p(m), c^i_p(m), g^i_p(t_0, t, T)\right) \quad (1.16)$$

where I represents the property rights of agent i, and the subscript p on the right side of the equality indicates the perceived property rights of agent i. M represents

the language imposed by the institution, i.e. the set of all messages that can be sent, h represents the process which maps messages to rewards, c represents the processes which maps messages to costs, and g is the governing process which indicates at which point events, including the elicitation of messages, will occur from the beginning of the experiment, $t0$, to the end of the experiment, T.

Should any of the elements of I_p^i not be equivalent to their corresponding element in I^i, subjects may believe that they are guaranteed the right to claim a reward which differs from the reward that will be granted by the institution given their message. Or in the Harrison (1994) definition, it could lead to a message m'' mapping to v'' when $v' > v''$ due to the subject believing that m'' maps to v' or that $v'' > v'$.

An example of this apparent mis-mapping of messages and rewards could be that a simple lack of understanding by the agents about their own property rights leads to a lack of salience. The degree to which the equivalence of equation (1.16) holds could depend on both the complexity of the property right endowment and the ability of the agent to comprehend her own endowment. Various tax incentives, for instance, endow a proportion of a population with potentially large savings via a tax credit or deduction conditional on citizens behaving in a certain way. However, these incentives may be too complex to comprehend and thus do not motivate citizens to make the choices that the policy intended even if the citizens would otherwise be willing. Cason and Plott (2014, p. 1237) report an experiment which describes how the complexity of the BDM mechanism results in potentially misreported preferences: "If the individual fails to understand the connection between acts and outcomes, the choice of acts can be misleading about the preferences over outcomes. In such cases, choice cannot be equated with

preference over consequences."

A more complex way in which the equivalence of equation (1.16) would fail is a fundamental mistrust of the experimenter. That is, the subject may fully understand the institution's communication of the elements of the property right endowment, but doesn't believe that the institution will uphold these rights. This in a way alludes to one of the experimental theories proposed by Grether and Plott (1979, p. 629), that a potential cause of preference reversals in experiments conducted by psychologists was that subjects were in experiments conducted by psychologists: "Subjects nearly always speculate about the purposes of experiments and psychologists have the reputation for deceiving subjects. It is also well known that subjects' choices are often influenced by what they perceive to be the purpose of the experiment."

Economic experiments designed to incentivize subjects to reveal their preferences generally rely on the assumption that the subject views the experimenter to be indifferent to the outcome of the experiment. Schneeweiss (1973) explores this view as a alternative explanation of the Ellsberg (1961) paradox critique of the axioms of Savage (1954). The subjects of the Ellsberg (1961) experiments, rather than adopting different preferences for events which were "ambiguous" as opposed to simply uncertain, could view the experiment as a zero-sum two-person game between the experimenter and themselves. Schneeweiss (1973) shows that if the subjects assume that the experimenter strategically wants to minimize the expected payout to subjects, the seemingly paradoxical behavior of the subjects can be explained as game theoretic optimal choice behavior.

Kadane (1992) also notes that if subjects in experiments are skeptical of the intentions of the experimenter, then the seemingly paradoxical behavior described

in Ellsberg (1961) and Allais (1953) can be explained as a rational response to the possibility of being cheated. It can be shown that should the agent assign any positive probability to the possibility of the experimenter selecting an outcome that would lead to the lowest expected payout given the subject's choice, both "paradoxes" fail to violate the axioms of Savage (1954).

Though both the above examples require subjects to assume a profit maximizing experimenter, an agent does not need to believe the experimenter has selfish interests for there to be a disconnect between the property rights, as perceived by the agent, and the property rights intended to be induced by the institution. Harrison and Johnson (2006, p. 178) note that in experiments attempting to describe behavior"variously labeled 'cooperation,' 'altruism,' 'reciprocity' or 'confusion'," the experimental methods used are often confounded by the manner in which the experimenter deals with the "residual" money left on the table after the experiment is over. If subjects incorporate the residual claimant into their preference structure, subjects displaying "altruistic" behavior may in fact be attempting to manipulate the residual to the claimant.

In the bulk of these kinds of experiments the experimenter is the implied residual claimant, though not always. Harrison and Johnson (2006) conducted an experiment utilizing the popular "Dictator" game with four treatments which allowed for variation in both the "peasant" (the recipient of the money from the Dictator) and in the recipient of the residual funds. In treatment "O," the "peasant" was another subject paired with the Dictator, and in treatment "C," the "peasant" was an unspecified charity. In both treatments O and C the residual claimant was implied to be the experimenter (nothing specific was said about the recipient of the residual funds in the instructions to the subjects). In the "O(C)" treatment,

the "peasant" was another subject paired with the Dictator with the residual going to an unspecified charity, while in the "C(O)" treatment the "peasant" was an unspecified charity with the residual going to another subject randomly selected at the end of the experiment.

Harrison and Johnson (2006, p. 196) find greater giving to the "peasant" in the C treatment compared to the O treatment, a greater giving to the "peasant" in the C(O) treatment compared to the C treatment, and a reduction in giving to the "peasant" in the O(C) treatment compared to the O treatment. Each of these differences was statistically different from zero. These results imply that the subjects in this experiment preferred money from the experiment to go to a charity more than to the experimenter, and preferred the money to go to the experimenter more than to the other subjects. The primary importance of this study is to demonstrate that subjects may incorporate the residual claimant of funds of an experiment into their utility functions.

Generally, if an agent views any "third party" (be this the experimenter, a charity, or even Nature) as another agent in the system, while the experimenter views this "third party" as being outside of the economic system, this could conceivably cause any element in I_p^i to differ from its corresponding element in I^i. Most apparent however is the potential for $h_i(m) \neq h_i^p(m)$.

Though the above examples require the agent to view the set of agents in a system differently than the experimenter views the set of agents, this isn't necessary for the agent to believe the institution will not uphold the agent's property right as endowed. For instance, the agent may believe she has superior knowledge of the mechanism which maps messages to rewards thus causing $h_i(m) \neq h_i^p(m)$. In many experiments, subjects are asked to make a choice between lotteries, with their

chosen lottery being played out for a real reward. The mechanism used to select the outcome of a lottery is almost always some physical device, such as a coin flip, a dice roll, or a bingo cage, the physics of which may be well known by the subject to result in certain outcomes with different likelihoods than the institution suggests; the outcome of flipping a US quarter is not precisely a 50/50 gamble between heads and tails for instance. In this instance, the choice between lotteries is the message, the outcome of the lottery is the reward, and the mechanism which selects the outcome of the lottery selected is the mechanism which maps the message to the reward. Asymmetrical beliefs about the reward mechanism can lead to a failure to induce salience as $h_i(m) \neq h_i^p(m)$.

Should either the subject or the experimenter not understand the other or should the subject mistrust the institution, there can be a failure of salience. However, the latter can be said to be the result of a certain preference structure,[1] while the former is usually an experimental artifact. The institution could fail to properly communicate the endowed property rights, there could be a lack of technical ability on behalf of the subject to comprehend her property rights as endowed, or the actual institutional endowment could be misspecified by the experimenter when observing messages, for instance, by ignoring an agent's perception of additional agents.

Both misunderstanding and mistrust could potentially affect the messaging behavior of subjects in an experiment. In the syntax of Smith (1982), the degree of (mis)understanding could be reflected by the technology element, T^i, of equation (1.17) below, while mistrust would be reflected in the utility element, u^i, of equa-

[1]For instance, the greater the potential loss to the experimenter (or gain to the agent), the more the agent may prefer to discount fortuitous events and overweight bad events. This could be interpreted as mistrust being a determinant of the probability weighting function in RDU.

tion (1.17) below. Both elements ultimately shape the behavioral process which determines messages, defined in equation (1.18):

$$e^i = (u^i, T^i, \omega^i) \tag{1.17}$$

$$m^i \simeq \beta^i(e^i|I) \tag{1.18}$$

In equations (1.17) and (1.18), e^i is called the environment of agent i which is determined by the agents' utility structure, u^i, technology endowment, T^i, and wealth endowment, ω^i. An agent's behavior, β^i, then maps the subject's environment conditional on the institutionally granted property right, I, to the message space, m^i. It is this message space which is observed by the institution and ultimately mapped to a reward.

These two ideas are of course not the only ways in which salience could be experimentally violated. A much used experimental method is that of asking subjects for responses to hypothetical questions. There is a general disagreement, even a "gentle aggression,"[2] between experimental economists and experimental psychologists about the use of hypothetical rewards and whether a subject's intrinsic motivation to complete a task proficiently is sufficient to produce a salient mapping of messages to rewards in an experiment.

Grether and Plott (1979, p. 624) state their case against the use of hypothetical rewards in exploring economic theory: "No attempt is made to expand the theory to cover choices from options which yield consequences of no importance.[...] Thus the results of experiments where subjects may be bored, playing games, or otherwise not motivated, present no immediate challenges to theory." This is later

[2]This is how Hertwig and Ortmann (2001) characterize the methodological critique by Smith (1982) of both experimental economics and experimental psychology.

discussed specifically in the context of using of imaginary money as a means to provide salience.

Camerer and Hogarth (1999, p. 31) note that from 1970-97 not a single experimental study was published in the *American Economic Review* in which all subjects face only hypothetical rewards, indicating that economists are typically hostile to the idea of intrinsic motivation being sufficient to produce saliency. Camerer and Hogarth (1999) show in their analysis of a non-random sample of the experimental literature that increasing financial incentives from zero to positive but low stakes typically improves performance over some domain of experimental tasks, in particular tasks which are effort-responsive like judgment, problem-solving, or clerical tasks, but they find that increasing stakes from some low level to a relatively higher level does little to improve performance and sometimes hinders it. With respect to tasks for which there is no normative level of performance to be measured, such as games, auctions or choices between risky lotteries as in the Grether and Plott (1979) experiments, Camerer and Hogarth (1999, p. 34) state tha"the most typical result is that incentives do not affect mean performance, but incentives often reduce variance in responses." A reduction of variance in responses could lead to a reduction in apparent violations of EUT such as the preference reversals of Grether and Plott (1979).

Camerer and Hogarth (1999, p. 8) state however, that "The extreme positions, that [material] incentives make no difference at all, or always eliminate persistent irrationalities, are false," and in no uncertain terms state "*There is no replicated study in which a theory of rational choice was rejected at low stakes in favor or of a well-specified behavioral alternative, and accepted at high stakes* [...] and nothing in any sensible understanding of human psychology suggests that it would [be]."

[emphasis in the original] (1999, pp. 33-34). This echoes earlier statements by Smith and Walker (1993, p. 246) that "[...] rewards matter, and that neither of the polar views - only reward matters, or reward does not matter - are sustainable across the range of experimental economics" and that this view is "common sense."

Should a subject have no intrinsic motivation to respond to a task but values money, the introduction of monetary rewards for responses can potentially induce saliency by changing the mapping of messages from non-valued hypothetical rewards to valued monetary rewards. If a subject does have some degree of intrinsic motivation to respond to a task "correctly" but also values money, the introduction of monetary rewards for responses can potentially make the gross reward of sending a certain message "dominate" the subjective costs of sending that message when intrinsic motivation alone wouldn't have sufficed.

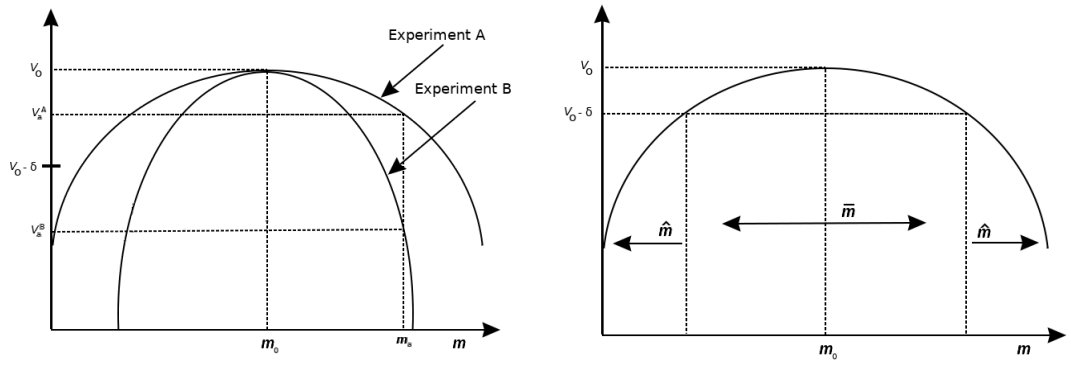
1.3.2 Dominance and Potential Violations

Dominance is the third necessary precept for conducting a valid experiment proposed by Smith (1982) and elaborated on by Harrison (1989, 1992). Smith (1982, p. 934) defines dominance as selecting rewards such that "The reward structure dominates any subjective costs (or values) associated with participation in the activities of an experiment." Harrison (1992, p. 1426) refines this definition and states that dominance "requires that the reward of sending a message corresponding to a null hypothesis be 'perceptibly and motivationally greater' than the reward of sending an alternative hypothesis."

If the messages corresponding to the null and alternative hypotheses are m_0 and m_a, respectively, then dominance requires that the values associated with sending each of these messages, v_0 and v_a respectively, be such that $v_0 > v_a + \delta$, where δ

is the subjective cost to the agent of sending message m_0 instead of message m_a (Harrison 1992, p. 1427). This concept is illustrated in Figure (1.1a) below from Harrison (1992, p. 1427):

Figure 1.1: Flat Maximum Critique - Harrison (1992)



(a) (b)

The potential experiments in Figure (1.1a) have been constructed such that the value to the subject of sending the message corresponding to the null hypothesis is equivalent for both experiments. However, only Experiment B has a value associated with sending the alternative hypothesis that is less than $v_0 - \delta$. Thus only Experiment B satisfies dominance for this particular set of null and alternative hypotheses.

The difference between salience and dominance in the reward medium is obvious in this example: the subject does in fact value the medium and is not satiated in the medium that is being returned to her for sending messages m_0 and m_a in both experiments, satisfying salience. But, only in Experiment B is there a sufficient difference in her valuation of the reward medium from sending m_0 instead of m_a for her to send the message m_0 instead of m_a.

It is easy to extend this idea to a scenario in which one or more composite alternative hypotheses are being tested against a single point-null hypothesis. In this case there exists a set of messages "close" to m_0, $\overline{m}$, which provide rewards that are not sufficiently different from v_0 to warrant sending m_0. Similarly there will be a set of messages "far" enough from m_0, $\hat{m}$, such that the value of sending any of these messages is sufficiently different from v_0 that the subject would be motivated to send messages from the set of m_0 instead of $\hat{m}$. This can be seen in of Figure (1.1b) from Harrison (1992) above.

There is no reason to believe that δ shouldn't vary from subject to subject or from task to task. Looking at Figure (1.1a), Experiment B clearly satisfies dominance for the subject in question and the task requiring the sending of either m_0 or m_a, but a different subject might have a larger subjective cost of sending m_0 instead of m_a making $v_a^B > v_0 - \delta$ and causing a failure of dominance. Looking at Figure (1.1b), the set of messages in $\hat{m}$ correspond to those messages which provide rewards which are valued sufficiently differently from m_0 to dominate the subjective cost of sending message m_0, but it cannot be said that *any* message in $\hat{m}$ satisfies dominance for *any* message in $\bar{m}$.

Harrison (1989) argues that if message m_0 is the optimal choice for a subject conforming to EUT in a particular experiment, an observance of the choice of m_a is only relevant as a critique of EUT if dominance is satisfied for that task. Harrison (1994) replicated the experiments ofGrether and Plott (1979) with only minor differences and observed roughly the same proportion of subjects displaying the apparent preference reversal phenomenon. However, assuming the subject correctly reported their direct preference for either the P bet of the \$ bet, but mis-reported their true selling price with the BDM mechanism, the difference in

expected income for a subject of displaying a preference reversal versus the expected income if they have reported a consistent selling price averaged only \$0.006. For such a small value of $v_0 - v_a$ "one must nihilistically insist that subjects have a sufficiently low threshold δ, perhaps even claiming $\delta = 0$, in order to conclude that such observations allow one to reject the null hypothesis"(Harrison 1992, p. 1428). Harrison (1994, p. 237) concludes"that the subjects in these preference reversal experiments had virtually no incentive to behave any more consistently than they did."

1.4 Holt and Laury (2002) Multiple Price List and Apparent Inconsistencies

The preference reversal phenomenon of Grether and Plott (1979) is not the only instance of apparent violations of EUT to be replicated by experimentalists. The Ellsberg (1961) paradox is a popular early example of an apparent violation of EUT, as well as repeated over-bidding with respect to the Nash predicted bids in laboratory auctions by Cox, Roberson and Smith (1982), Cox, Smith and Walker (1983a,b, 1988).

Cox, Smith and Walker (1985, p. 160) and Harrison (1989, p. 749) both note that this overbidding behavior is consistent with risk-averse subjects, but that the bid deviations are too heterogeneous to be consistent with subjects who are uniformly risk-averse. Cox, Smith and Walker (1985) conduct an experiment which attempts to test for heterogeneous risk preferences of subjects and conclude that their experiments provide "evidence against the compound lottery axiom of EUT" (Cox, Smith and Walker 1985, p. 165). Harrison (1989), though accepting that deviations in bids from Nash predicted outcomes could be caused by risk-averse

subjects, argues that the experiments performed did not meet the dominance criteria for a valid experiment. Thus, the (empirical) question of the degree of heterogeneity in risk preferences among subjects and its influence on bidding behavior remains unanswered.

Hey and Orme (1994, p. 1291) note that the "experimentally observed violations of expected utility theory (EUT) have stimulated a deluge of generalized preference functions, almost all containing EUT as a special case." Hey and Orme (1994) conduct a series of experiments on 80 subjects requiring the subjects to state their preference for one lottery across each of 100 lottery pairs to test if subjects conform to EUT (or the "Dual Theory" of Yaari (1987)) in favor of risk-neutrality, and whether subjects conform to any of 8 various generalizations of EUT in favor of EUT. They report substantial evidence against risk-neutrality, and it is clear from their dataset that there is a great deal of heterogeneity in subjects deviating from risk neutrality. Hey and Orme (1994, p. 1322) state that "we are tempted to conclude by saying that our study indicates that behavior can be reasonably well modeled [...] as 'EU plus noise.'"

Holt and Laury (2002, p. 1644) note that bidding behavior in auctions can be used to elicit risk attitudes and that the over-bidding with respect to the Nash predicted outcomes had been attributed to risk aversion. They propose using a multiple price list (MPL) as a tool for experimentalists to control for individual heterogeneity in risk preferences and conduct an experiment to test whether the extent of risk aversion in subjects is dependent on the stakes of the tasks with which the subjects are presented.

The MPL has been widely used to elicit "homegrown" preferences for risk for several decades. The earliest use of the MPL method is considered to be Miller,

Meyer and Lanzetta (1969), but Binswanger (1980, 1981) is regarded as the first experimental economist to identify risk attitudes using an MPL with real payoffs. The MPL was further developed by Schubert, Brown, Gysler and Brachinger (1999) and Holt and Laury (2002).

The instrument employed by HL requires subjects to make a series of binary choices between two lotteries, A and B, across ten lottery pairs. The instrument used in the "low-stakes" treatment of HL is as follows:

Table 1.8: Holt and Laury (2002)
The Ten Paired Lottery-Choice Decisions with Low Payoffs

Row #	Option A	Option B	Expected Payoff Difference
1	1/10 of \$2.00, 9/10 of \$1.60	1/10 of \$3.85, 9/10 of \$0.10	\$1.17
2	2/10 of \$2.00, 8/10 of \$1.60	2/10 of \$3.85, 8/10 of \$0.10	\$0.83
3	3/10 of \$2.00, 7/10 of \$1.60	3/10 of \$3.85, 7/10 of \$0.10	\$0.50
4	4/10 of \$2.00, 6/10 of \$1.60	4/10 of \$3.85, 6/10 of \$0.10	\$0.16
5	5/10 of \$2.00, 5/10 of \$1.60	5/10 of \$3.85, 5/10 of \$0.10	-\$0.18
6	6/10 of \$2.00, 4/10 of \$1.60	6/10 of \$3.85, 4/10 of \$0.10	-\$0.51
7	7/10 of \$2.00, 3/10 of \$1.60	7/10 of \$3.85, 3/10 of \$0.10	-\$0.85
8	8/10 of \$2.00, 2/10 of \$1.60	8/10 of \$3.85, 2/10 of \$0.10	-\$1.18
9	9/10 of \$2.00, 1/10 of \$1.60	9/10 of \$3.85, 1/10 of \$0.10	-\$1.52
10	10/10 of \$2.00, 0/10 of \$1.60	10/10 of \$3.85, 0/10 of \$0.10	-\$1.85

The logic of the HL MPL is straightforward. In all ten lottery pairs, the monetary outcomes of all A lotteries are the same, and similarly with the B lotteries. Every lottery is comprised of two possible outcomes, with the higher outcome in the B lotteries being greater than the higher outcome in the A lotteries, and the lower outcome in the B lotteries being less than the lower outcome in

the A lotteries. The probability of receiving these outcomes changes from pair to pair. At the top of the list, the probability of receiving the high amount from each lottery is only 0.1, while the probability of receiving the lower amount is 0.9. Moving from the top to the bottom of the list, the probability of receiving the higher amount in each lottery increases by 0.1 for each row until at the bottom of the list, row 10, the probability of receiving the higher amount is equal to 1, collapsing the decision to a choice between two certain outcomes.

If a subject has strictly monotonic utility for money, an EUT maximizer should either prefer option A initially and then working down the rows eventually prefer option B for the remaining rows, or the subject should prefer B for all rows. However, it is often observed that subjects will "switch" back and forth between selecting lottery A and selecting lottery B. This is commonly called "multiple switching behavior" (MSB) (Bruner 2011). Data displaying MSB has often been described as "inconsistent" with an EUT maximization. Holt and Laury (2002, p. 1645) describes a subject switching once from A to B when the B lottery becomes sufficiently attractive as what a risk averse subject "should" do in the HL MPL, implying that MSB is contrary to theory.

Harrison, Lau and Rutström (2007, p. 347) note that a subject could be indifferent between the lotteries of certain pairs, which would explain why a subject displayed MSB. In fact, under EUT with a deterministic theory of choice specification and some mild assumptions, a subject displaying MSB must be indifferent to the lotteries in lottery pairs between the first switch and the last switch. This is shown in Appendix A. Harrison, Lau and Rutström (2007) conduct an experiment in which some subjects are offered an "indifferent" option and note that the proportion of subjects who are not offered the indifference option expressing MSB almost

equals the proportion of subjects offered the indifference option selecting the indifference option. While this is very suggestive that MSB is caused by indifference, it should be noted that the indifference option was played out by the experimenter randomly selecting either option A or option B for payment. The selection of the indifference option could therefore represent a preference for a compound lottery of A and B, and not indifference between A and B.

Choosing option A in row 10 is also an apparent violation of EUT that is not so easily believed to be caused by indifference. Since there is no uncertainty in the outcomes of either lottery A or lottery B, a choice of A apparently violates the axiom of monotonicity. It is possible that a subject who chooses A in row 10 has a very good motivation to do so, perhaps because of some influence from outside of the experiment, and could potentially still be making choices in accordance with EUT. As noted by Smith (1982, p. 930): "It is hard to find an experimentalist who regards anything as self evident, including the proposition that people prefer more money to less." This is mentioned merely as a caveat applied to proclaiming a general absence of rationality on the part of the subjects making such choices. It is more believable that such subjects are making a mistake in choosing A in row ten.

The extent of this type of behavior in economic experiments is not entirely clear. Experimental procedures often vary from experimenter to experimenter and form experiment to experiment. For instance, Holt and Laury (2002) note that some subjects "crossed out and changed" their responses to choices near their switch point. If the experimental design hadn't allowed subjects to change their responses, or if the design made it cumbersome to do so, then there might have been more observed "inconsistent" responses than reported.

While many experimenters report the number or proportion of subjects switch-

ing multiple times, some experimenters exclude subjects who display this behavior from their analysis or only report the number of "safe" choices (the lottery A choices). Holt and Laury (2002, p. 1648) tested an analysis without the"inconsistent" subjects but note that "The average number of safe choices increases slightly in some treatments when we restrict our attention to those who never switch back, but typically by less than 0.2 choices" and ultimately left "inconsistent" subjects in their final analysis. When experimenters only report safe choices, they often don't discriminate between subjects who switched once or multiple times or if one of those "safe" choices was a choice of A in row 10. However, if 10 safe choices are reported, then clearly one of them was a choice of lottery A in row 10.

Filippin and Crosetto (2016, p. 9) collected datasets of 54 published replications of HL to examine gender differences in estimates for risk attitudes. Filippin and Crosetto (2016, pp. 10-11, 17-18) also briefly discuss the number of "inconsistent" subjects in the aggregated dataset they collected. Their Tables 4 and 7 are combined and reproduced in Table 1.9.

Table 1.9 shows to some extent the potential reporting bias of "inconsistent" behavior in experiments. Of the 6707 subjects, there was insufficient data to tell if any type of "inconsistent" behavior occurred with 772 subjects, those with "Summary" detail, and a further 699 subjects in which it was only possible to tell if there was an apparent violation of monotonicity, those with "Partial" detail and only reporting the number of safe choices. About 21.5% of the data reports insufficient information to determine the extent of inconsistent behavior.

In the first part of Table 1.9, if we consider only the "Full" detail data and the "Partial" detail data which indicates whether or not subjects were "consistent," there were 1075 out of 5236 subjects who displayed some sort of inconsistency,

Table 1.9: Filippin and Crosetto (2016)
Prevalence and Type of "Inconsistent" Behavior

	Detail	Consistent Subjects			Inconsistent Subjects		
		Males	Females	Total	Males	Females	Total
Microdata	Full	2119	2205	4324	411	502	913
# of safe choices + consistency	Partial	504	408	912	64	98	162
# of safe choices only		375	324	699	3	1	4
Summary Statistics	Summary	413	359	772			
Total		3411	3296	6707	478	601	1079

	Inconsistent Choices		% Inconsistent Subjects		
	Number	Out Of	Males	Females	Total
Multiple or inverse switching	703	6825	8.8	11.8	10.3
Dominated choices	102	6882	1.8	1.2	1.5
Switch and dominated	270	6825	3.6	4.3	4.0
Total	1075		14.1	17.3	15.8

about 20.5% of subjects. Holt and Laury (2002, p. 1647) reported 28 of 212 (13.2%) of their subjects exhibited MSB.

While this is not as substantial a proportion of subjects acting in apparent violation of EUT as in the Grether and Plott (1979) preference reversal phenomenon, it is not a trivial proportion. Additionally most of the potential explanations for the preference reversal phenomenon proposed by Grether and Plott (1979) should be considered resolved given the many replications of HL. Almost all of the proposed experimental method explanations can be generally rejected: the observance of this behavior is not a low frequency event, nor are sample sizes small; most of the subjects in these experiments (including all of the original HL subjects) were either university students or faculty who can hardly be considered unsophisticated subjects; almost all the experimenters were economists. The question of confusion or misunderstanding, however, remains open, along with many of the theoretical critiques.

Two of the theoretical critiques of the Grether and Plott (1979) experiments deserve particular note. In the HL experiments, subjects were presented with several MPLs and were told that one of their responses would be chosen at random and played out for real earnings. Most experimenters take advantage of this "pay one" mechanism in order to increase the stakes for each individual question without breaking their budget. This "pay one" mechanism, however, requires that an independence axiom hold over for a pattern of choices, with a single switch point as a condition for utility maximization across all preferences. This is no different from the critique noted by Holt (1986) and Karni and Safra (1987) that should a subject have preferences which are in line with RDU, then there is no apparent violation of economic theory by MSB.

Similarly, the use of a "pay one" mechanism may dilute the payoffs of outcomes to the point where the difference in the value of option A versus option B fails to satisfy the dominance criteria of Smith (1982). The HL MPL is structured in such a way that a subject will confront a lottery pair in which she will be nearly (or entirely) indifferent between the two options. The "pay one" mechanism decreases the expected difference in the value of the options by in effect multiplying the probability of each outcome in the lotteries by 0.10.

The two lottery pairs that the subject is closest to indifference over will always be the two on either side of the switch point. If MSB is party due to a failure to satisfy dominance, it should occur with greater frequency near this point. Holt and Laury (2002, p. 1648) comment that "Even for those who switched back and forth, there is typically a clear division point between clusters of A and B choices, with few 'errors' on each side" and that responses that were crossed out and changed generally were around the switch point (2002, p. 1646). Holt and Laury (2002,

pp. 1647-1648) further note that the rate of MSB was lowest for their high stakes treatments and highest for their hypothetical stakes treatment. The way the instrument is built and the frequency of MSB across different treatments suggest that a failure of dominance may be a large factor in explaining the frequency of MSB.

While it may be more believable that dominance is at play for MSB than the non-monotonic choice of lottery A in row 10, dominance failure shouldn't be ruled out a priori for non-monotonic choices. Take for example, a choice between \$0.01 and \$0.02. Even if the outcomes of both options are guaranteed, the value difference is likely to be very small, and thus more likely to fail the dominance criteria. Similarly, because of the "pay one" mechanism, the expected cost of choosing A in row 10 of the HL MPL is only \$0.185. This may be high in comparison to the cost associated with the generally observed MSB, but it is not unfathomably high.

A failure to induce salience could also help explain some "inconsistent" choices. If subjects don't comprehend the two-part payment mechanism (selecting one lottery from the list for payment, then playing the lottery out to determine the reward), or more generally how the probabilities associated with outcomes are mapped to (presumably valued) monetary rewards, then there is a failure to induce salience. A failure to induce salience seems less likely than a failure of dominance when explaining MSB given that most MSB is clustered around what could be called a "true" switching point and that the remaining choices seem "consistent" with a salient reward mechanism. It does, however, seem more likely that there is a failure of salience for subject who selected lottery A in row 10.

1.5 Stochastic Choice as an Explanation of "Inconsistent" Choices

It is easy to imagine the theoretical possibility that a subject's preference ordering among a set of alternatives is mapped perfectly without error to the messages which will realize the optimal outcome for that subject. In this case, the various models make specific predictions as to which elements belong in the chosen set given any set of preferences consistent with that model in question. Any occurrence of some alternative within the chosen set which isn't predicted is therefore an apparent violation of the utility theory in question given this mapping assumption. As has been discussed with respect to the Grether and Plott (1979) and Holt and Laury (2002) replications, the mass of empirical data collected over the past few decades has shown that such apparent violations are common, and that there is a need to "attempt to modify the theory to account for [these] exception[s] without simultaneously making the theory vacuous" Grether and Plott (1979, p. 634). Such a modification that potentially explains observed choices of subjects which appear to be inconsistent with some predetermined utility theory is that observed choices by subjects are a product of a choice process that is at least in part stochastic, and not wholly deterministic.

The first description of choice as a stochastic process appears to be by Edwards (1954), with notable early contributions by Luce (1958), Debreu (1958), Davidson and Marschak (1959), Becker, DeGroot and Marschak (1963), Luce and Suppes (1965). Stochastic choice models add elements of randomness to utility models, which allow for a degree of error during the evaluation of various alternatives, a degree of randomness of the preference relation used in the evaluation of alternatives,

and/or randomness in the selection of an alternative to belong to the chosen set. These stochastic models are used as complements to, rather than substitutes for, deterministic theories of utility. As such, they generally (though not always) seek to link deterministic preference relations, $A \succeq B$, to probabilities of choice, $Pr(A) \geq Pr(B)$.

There are various ways to accomplish this linking of ideas, the most common of which are discussed by Wilcox (2008). Generally, these models fall into one of two groups: Random Preference (RP) models, or deterministic preference with a random error models. The most common deterministic preference with random error models consist of Strong Utility (SU) models, Strict Utility models (a subset of SU models), and Strong Utility's superior derivatives Moderate Utility (MU) models. I lump SU and MU models into the same group because although they differ on several key points, they both implement stochasticity by assuming some error in the evaluation of alternatives. They differ in their treatment of this error, with SU models assuming it is homoscedastic and MU models requiring it to be heteroscedastic in a particular fashion over the domain of potential outcomes. RP models differ from the rest of these by imposing randomness in the preference relation used to evaluate the alternatives.

A notable third group, with only one member I am aware of, could be considered deterministic preferences with a stochastic choice strategy. The sole member model in this group was proposed by Machina (1985) in which randomness is explained as subjects having convex indifference curves in the Machina Triangle space, Machina (1987), and seeking mixtures of alternative lotteries in order to maximize a deterministic preference. The stochastic mixture is said to be deterministically more preferred to any of the "pure" lottery options which make up the mixture for such

subjects. This theory implies that there is no error in the evaluation of alternatives, no error in the choice of the stochastic mixture, and no randomness of preferences, while still predicting noisy observed choices. This theory however has fallen out of favor, and Hey and Carbone (1995) provide strong experimental evidence against it. Because of the large degree of determinant preference behavior in this model, it is not considered in the rest of this text when discussing stochastic choice models.

Another concept called "trembles," a term derived from the notion of a "trembling hand" equilibrium developed by Selten (1975), suggests that some choices are made completely at random with no consideration for the underlying values of the alternatives. Trembles can be implemented by assuming that there is no error in the evaluation of alternatives, no randomness of preferences, and that all of the apparently inconsistent choices are mistakes. Trembles can also be imposed on top of other stochastic models to transform the "true" choice probabilities derived from the stochastic models into observed probabilities of choice. Because in either case the probability of a tremble does not depend on any preference relation, trembles will be left out of the discussion of stochastic models unless otherwise specifically noted.

In order to discuss stochastic models in more depth, some notation will be borrowed from Wilcox (2008). Let S_m and R_m be two lotteries in lottery pair m in which discrete probability distributions $s_{m0}, \ldots, s_{m(I-1)}$ and $r_{m0}, \ldots, r_{m(I-1)}$ apply respectively to a set of I outcomes in $Z = \{z_0, \ldots, z_{(I-1)}\}$. Let the context of lottery pair m, c_m, be the set of outcomes in Z with non-zero probabilities applied to them by any lottery in pair m. Assume that for any lottery pair m, $z_0 > \ldots > z_{(I-1)}$. Finally, let $P_m^n = Pr(y_m^n = 1)$ represent the probability that subject n chooses lottery S in pair m, and let $1 - P_m^n = Pr(y_m^n = 0)$ equal the probability that subject

n chooses lottery R in pair m. It is this concept of probability of choice which is linked to the deterministic concept of the preference relation $\succeq$.

With this notation in place we can define the manner in which the preference relation is most commonly linked to a probability of choice. Assuming for now that a subject has an EUT structure:

$$V(S_m|\beta^n) \equiv \sum_{z=0}^{I-1} s_{mz} u_z^n \quad , \quad V(R_m|\beta^n) \equiv \sum_{z=0}^{I-1} r_{mz} u_z^n \tag{1.19}$$

where u_z^n is the utility of prize z given some elements of the vector of structural parameters β^n, and $V(X_m|\beta)$ is a value function determined by properties of lottery X_m and the vector of structural parameters. The structural parameters β^n can be the utilities of the outcomes themselves, or the determining parameters of some parametric function of utility. Equation (1.19) can be easily transformed to be represented by rank dependent utility with no loss of generalization. The β^n vector would simply have to additionally include parameters defining the probability weights. Thus, for any transitive structure:

$$S_m \succeq^n R_m \Leftrightarrow V(S_m|\beta^n) \geq V(R_m|\beta^n) \Leftrightarrow Pr(y_m^n = 1) \geq Pr(y_m^n = 0) \tag{1.20}$$

RP models posit that for every choice task faced by the subject, a preference relation is drawn randomly from some distribution of preference relations and then the subject makes a choice in accordance with the randomly drawn preference relation. Econometrically it is possible to model choice such that once the preference relation is drawn, further randomness is added by having the evaluation of the alternatives involve some error process, as in SU and MU models. However, this is almost never done. Usually subjects conforming to RP models are said to draw the preference relation randomly from a distribution, and then make a choice

deterministically with respect to the preference relation. Thus the probability of choosing any lottery conditional on this randomly drawn preference is either 0 or 1:

$$Pr(y_m^n = 1|\beta^{n*}) = \{0, 1\} \tag{1.21}$$

where β^{n*} is the vector of parameters randomly drawn by subject n. Given this relationship, the unconditional choice probability is just the probability of observing β^{n*} given some joint distribution of the elements of β, $G_\beta(x|\alpha^n)$, where α^n is a vector of parameters which defines the shape of the distribution. Let $B_m = \{B|V(S_m|\beta) \geq V(R_m|\beta)\}$. Then the unconditional probability that a subject chooses lottery S_m is:

$$P_m^n = \int_{\beta \in B} dG_\beta(x|\alpha^n) \tag{1.22}$$

that is, the probability of the choice is simply the probability of observing a β vector which deterministically conforms to that choice.

The collection of SU and MU models represent preferences as stable across choice tasks but with an error of some kind when the utilities of the lotteries are evaluated. These models are very similar to commonly used latent variable models, with SU models assuming that the latent variable is homoscedastic and MU models assuming that the latent variable is heteroscedastic. Much in the same way as a standard logit model, SU and MU models state that this latent variable, y_m^{n*}, relates to the observed choice such that $y_m^n = 1 \Leftrightarrow y_m^{n*} \geq 0$, thus $P_m^n = Pr(y_m^{n*} \geq 0)$. The latent variable takes the form:

$$y_m^{n*} = V(S_m|\beta^n) - V(R_m|\beta^n) - \frac{\epsilon}{\lambda^n} \tag{1.23}$$

where ϵ is a random variable with a mean of 0, some standard variance and a symmetric c.d.f $F(\cdot)$ where $F(0) = 0.5$. Together with $\frac{1}{\lambda^n}$, this term represents the degree of noise in the evaluation of alternatives. Equation (1.23) is transformed into a choice probability by applying some c.d.f $F(\cdot)$:

$$P_m^n = F\left(\lambda^n[V(S_m|\beta^n) - V(R_m|\beta^n)]\right) \tag{1.24}$$

As λ^n approaches infinity, equation (1.24) will approach either 0 or 1, while as λ^n approaches 0, equation (1.24) will approach 0.5. Should the function $F(\cdot)$ take the form of the logistic c.d.f. then the choice probabilities can be calculated by:

$$P_m^n = \frac{\exp\left[\lambda^n \times V(S_m|\beta^n)\right]}{\exp\left[\lambda^n \times V(S_m|\beta^n)\right] + \exp\left[\lambda^n \times V(R_m|\beta^n)\right]} \tag{1.25}$$

Because utility structures such as EUT and RDU are unique up to positive affine transformations, λ^n can be any arbitrarily chosen constant and choice probabilities would still be preference order preserving. The choice of λ_m is however a defining distinction between SU and MU models. Wilcox (2008) proposes and Wilcox (2011) expands upon a MU model called "contextual utility" (CU) model. A "context" in this model refers to the set of outcomes in a lottery or collection of lotteries that have non-zero probabilities. This expands equation (1.25) by setting λ^n to the following:

$$\lambda^n = \frac{1}{\lambda^{n*}\left[u_{m0}^n(z) - u_{m(I-1)}^n(z)\right]} \tag{1.26}$$

where $\left[u_{m0}^n(z) - u_{m(I-1)}^n(z)\right]$ is the difference in utility of the greatest utility outcome and the least utility outcome in the context of pair m, and λ^{n*} can continue to be adjusted with the same effects as λ^n in equation (1.24). It is this property which makes the latent random variable defined in equation (1.23) heteroscedastic

with respect to the context of the lottery pair. It has several appealing implications.

First and foremost, CU allows the "more risk averse than" relation derived by Pratt (1964) for deterministic risky choice to be extended to the"stochastically more risk averse than" (SMRA) relation across multiple contexts. Wilcox (2011, p. 89) defines it thus: "Agent a is stochastically more risk averse than agent b [...] iff $P^a > P^b$ for every [mean preserving spread] pair $\{S, T\}$." A mean preserving spread (MPS) pair is simply a pair of lotteries with equal expected values. SU models only allow the SMRA relation arcoss pairs that share a context, while CU allows for this relation across contexts. Second, CU only conforms to moderate stochastic transitivity, hence its inclusion as a MU model. This allows CU to be descriptively more appealing as potential choice patterns which violate strong stochastic transitivity are acceptable with moderate stochastic transitivity.

Stochastic choice models can add a great deal of traction to utility theories which would otherwise falter when applied to apparently inconsistent choice data. As the difference in value between any two alternatives approaches 0, the choice probabilities of the two alternatives approach each other. The HL MPL for example is structured in such a way to confront the subject with a list of lottery pairs in which the lotteries get closer and closer in value before diverging in value. With a deterministic choice process, this structure leads to a choice pattern with a single switch point. With a stochastic choice process, a choice pattern with a single switch point is only the most likely choice pattern for a subject with preferences in accordance with EUT.

For any given set of preferences and a modest degree of noise, a choice pattern displaying MSB clustered around the switch point that would be produced by a deterministic choice process is often only marginally less likely than a choice

pattern with one switch point. As noted previously, Holt and Laury (2002, p. 1648) observed MSB with this kind of clustering. In this way, stochastic choice models can explain behavior which may otherwise be taken as to imply a failure of dominance. Also, as noted previously, when the stakes were raised in the HL experiments, the extent of MSB was reduced. With SU models, this could be potentially explained as the increase in stakes causing an increase in the difference in values of the two alternative lotteries, which would lead to choice probabilities moving closer to 0 or 1. This however, is not the case with CU models which would normalize the difference to be the same for any scaling of outcomes.

SU and MU models do incorporate an often overlooked idea about what the choices by subjects ultimately amount to. In particular, it is often suggested, but seldom explored, that occasionally choices by a subject must in fact be inconsistent with the subject's own underlying latent preferences. As stated by Holt (1986): "Each subject must be making some error or mistake or whatever when answering the questions."

A choice error can be defined as the selection by a subject of an option among a set of alternatives which does not provide the greatest utility among the set of alternatives. More generally, to include scenarios where subjects are asked to make multiple selections among alternatives, it is the selection by a subject of an option among a set of alternatives which doesn't belong to the set predicted by a deterministic choice process and some predetermined theory of utility. As such, errors can only arise conditional on some predetermined theory.

A choice error in this context should not be interpreted as a violation of imposed utility theory. Any choice error requires that the manner in which the various alternatives are evaluated and ranked be consistent with the subject's latent

preference structure, but in the mapping of the subject's preference to the choice space, some noise is introduced that leads to a sub-optimal choice.

1.6 Concluding Remarks

Economic orthodoxy over the past half century has been presented with several challenges to the way it characterizes how an agent makes a choice between alternatives in her choice set. One of these challenges in the form of experiments initially conducted by psychologists, but later replicated by economists, spearheaded the discussion by observing a high frequency "preference reversals" which seemingly contradicted Expected Utility Theory. As stated by Grether and Plott (1979): "Taken at face value the data are simply inconsistent with preference theory and have broad implications about research priorities in economics."

The challenge to explain the mounting apparent violations of contemporary economic theory, however, did not go unheeded. Theorists noted that it wasn't necessary to forego the transitivity axiom as suggested by Grether and Plott (1979, p. 623), which along with the completeness axiom forms the basis for what is sometimes labelled as "rationality" in economics. Instead, it was shown by Holt (1986) and later by Karni and Safra (1987) that should the independence axiom be rejected, there exist preferences which conform to the observed apparent violation of choice. The "Anticipated Utility" theory of Quiggin (1982), now referred to as "Rank Dependent Utility," provides an axiomization of such a theory of utility without the independence axiom. Another departure from Expected Utility Theory was "Regret Theory," initially proposed independently by Bell (1982) and Loomes and Sugden (1982), then axiomized by Fishburn (1987), which abandons the transitivity axiom altogether. Loomes, Starmer and Sugden (1989) test this theory

on a replication of the Grether and Plott (1979) experiments and conclude that it fits the entire dataset better than Expected Utility Theory.

Economics as an experimental science has also progressed greatly over the past half century, notably with Smith (1982) defining the necessary precepts of conducting a valid controlled economic experiment Using the framework of these precepts, Harrison (1994) addresses the Grether and Plott (1979) experiments directly by stating that the dominance criteria put forth by Smith (1982) wasn't met in the original experiment and that several of the subsequent replications which purport to address the dominance issue do not do so. Some experimental modifications employed in replications in fact may decrease the saliency of the rewards offered. In this line of thought, the theory is less to blame than the experiments employed. Harrison (1994, pp. 236-239) replicated the Grether and Plott (1979) experiments with some modifications to increase the cost of reporting a selling price inconsistent with a subject's own latent preferences and observed a precipitous drop in the proportion of subjects displaying an apparent preference reversal.

A powerful supplement to current economic theory is the idea of choice as a stochastic processes as opposed to a deterministic one. Some stochastic models regard apparent violations of economic theory as the result of some randomness of latent preferences while others regard them as "mistakes" or "choice errors" from stochastic noise in the evaluation of alternatives. Such models are powerful as they make only very mild requirements on the structure of underlying preferences if any at all, and create a framework for these preferences to be mapped to choices. As such most stochastic models are equally applicable to Expected Utility Theory, Rank Dependent Utility, or Prospect Theory (Kahneman and Tversky 1979; Tversky and

Kahneman 1992), and can encompass and enhance the explanatory power of these utility theories.

Appendix - MSB as Deterministic Indifference

Lemma (1): If independence holds, and we have $L_0 \succeq L_1$ and $L_2 \succeq L_3$, then we must also have $aL_0 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_3 \forall \alpha \in (0,1)$

Proof. Given $L_0 \succeq L_1$ and $L_2 \succeq L_3$

$$\alpha L_0 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_2 \qquad \text{by independence}$$

$$\alpha L_1 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_3 \qquad \text{by independence}$$

$$\alpha L_0 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_3 \qquad \text{by transitivity of } \succeq$$

$$\text{therefore}$$

$$aL_0 + (1-\alpha)L_2 \succeq \alpha L_1 + (1-\alpha)L_3$$

□

Any three rows of lottery pairs in the Grether and Plott (1979, p. 623) MPL conform to the following

$$A_1 = \alpha(A_0) + (1-\alpha)A_2, B_1 = \alpha(B_0) + (1-\alpha)B_2 \;\; \textit{for some } \alpha \in (0,1)$$

An example of a multiple switch from a MPL is the following:

$$A_0 \succeq B_0, B_1 \succeq A_1, A_2 \succeq B_2$$

A multiple switch in the MPL implies indifference between lotteries of the interior

lottery pair:

Since $A_0 \succeq B_0, A_2 \succeq B_2$

then $A_1 \succeq B_1$, by Lemma (1) and reduction of compound lotteries (ROCL)

therefore $A_1 \sim B_1$, by completeness of $\succeq$

If any two lottery pairs in an MPL style instrument display preference relations in the same direction, then every lottery pair which is a linear combination of those two lottery pairs must have the same preference direction if we hold on to EUT and ROCL, otherwise the subject must be indifferent.

Chapter 2

The Normative Promise of Stochastic Models

The recent interest in the applicability of stochastic choice models for explaining choice behavior that seemingly violates Expected Utility Theory (EUT) has led to a variety of calls for, and attempts to seek out, a "true" stochastic model or combination of models which best describe observed choice behavior in experiments. Mosteller and Nogee (1951) may be the first to conduct an experiment that mapped choice frequencies to utility functions. Edwards (1954), who accused economics of having become "exceedingly elaborate, mathematical, and voluminous," continued to call for a research effort to link choice probabilities to utility functions. He criticized economists as "mak[ing] assumptions, and from these assumptions [...] deduce[ing] theorems which presumably can be tested, though it seems unlikely that the testing will ever occur."(1954, p. 380) and proposed that individuals may make choices stochastically as opposed to deterministically, which married some of the extent findings of psychology at the time with economics to help explain the data.

A more recent call to study stochastic models was issued by Hey and Orme

(1994, p. 1321) after they conducted rigorous subject-by-subject tests of some of the popular proposed alternatives to EUT, alongside EUT. They conclude that "possibly the overriding feature of our analysis is the importance of error [...] Perhaps we should now spend some time on thinking about the noise, rather than about even more alternatives to EU?" A wealth of stochastic models has resulted from economists and psychologists taking up the project proposed by Hey and Orme (1994). This chapter describes the results of this effort and introduces discussion of the normative promise of some of these models.

2.1 The Specification of Stochastic Models

I begin by distinguishing three classes of stochastic models. Hey and Orme (1994, p. 1301) proposed a model which incorporates a random error term into the evaluation of lotteries by subjects. The roots of this type of model date back to Fechner (1966) and Luce (1959), which has subsequently been called a "Strong Utility" (SU) or "Fechnerian" model. There are, however, a large number of models that have been derived from the SU model, and so I will refer generally to this class of models as "Random Error" (RE) models. Harless and Camerer (1994) had undertaken their own analysis of various alternatives to EUT and suggested a stochastic model which allows for subjects to potentially disregard their underlying preferences and choose between the available options with equal probability. The models in this class are called the "Constant Error" or "Tremble" (TR) models. Loomes and Sugden (1995) reconsidered and generalized a model initially proposed by Becker, DeGroot and Marschak (1963) called the "Random Preference" (RP) model which, in its most popular form, allows for subjects to have some distribution of preference relations from which they randomly choose every time they evaluate a

choice situation. Generally, any model that calls for an agent to have a distribution of preference relations belongs to the RP class.[1]

A less popular class of stochastic choice models proposed by Machina (1985) and Chew, Epstein and Segal (1991), suggests that subjects have deterministic preferences for "stochastic options" and thus deliberately engage in adding randomness to their choices. That is, subjects have convex indifference curves in the Marshack-Machina Triangle space,[2] and therefore a probabilistic (linear) mixture of two lotteries lies on a higher indifference curve than any two lines which lie on the same curve.

Hey and Carbone (1995) tested the "stochastic options" theory of Machina (1985) using the "Quadratic Mixture" model of Chew, Epstein and Segal (1991) and find strong evidence against it. The model itself has some restrictive aspects for estimation: "First, the likelihood function, although continuous everywhere, is not smoothly so; there are kinks in the function with resulting discontinuities in the derivatives. Second, for certain parameter sets, certain observations are *impossible*" [emphasis in the original] (1995, p. 164). Out of a sample of 45 subjects, Hey and Carbone (1995) find only 4 subjects to whom they can fit the model at all, and of

[1]There are other members of the RP model class that are less commonly utilized, one of which will be discussed later. For now, when referring to the RP model, I refer to the formulation specified by Loomes and Sugden (1995), where one preference relation is drawn per decision situation.

[2]The Marshack-Machina Triangle was developed by Machina (1987) as a way to represent the relation of lotteries with up to 3 outcomes and preferences for those lotteries. Each vertex of the triangle represents an outcome, and any point in the triangle represents a lottery. Any point on a vertex of the triangle represents a lottery with a 100% composition of one outcome. Any interior point represents a lottery composed of a mixture of outcomes, with the relative proportion of any outcome in the lottery defined by its geometric distance from its corresponding vertex. If the independence axiom holds, a straight line between any two points in the triangle space indicates all the lotteries amongst which the agent is indifferent. Parallel lines thus indicate either an increase or decrease in preference. A strictly convex curve connecting 2 lottery points represents a violation of the independence axiom.

these 4, for only 2 does the "Quadratic Mixture" model fit better than a RE type model, and of these 2, for only 1 subject are the estimated coefficients plausible (1995, p. 167). Continued investigation of this class of models has largely ceased since these and additional results by Camerer and Ho (1994).[3] Keeping with this pattern, when referring to stochastic models for the remainder of this text I do not include this class of models in our definition.

These general classes of stochastic choice constitute the bulk of the research on stochastic models, with the RE models possibly being the most widely employed models when estimating utility parameters from choice data. I will begin by defining some notation to characterize how these models operate.

For each option a in a set of alternatives t, stochastic models generate a probability that an agent will select that option from the set. These probabilities are referred to as "choice probabilities" and are generally related to an underlying determinisitic relation of preference. First, I define the Rank Dependent Utility (RDU) structure as formulated by Quiggin (1982), which nests EUT as a special case, as the deterministic structure of preference. Second I define the manner in which RP, TR, and RE models relate this deterministic structure to the stochastic specification of probabilities of choice.

RDU is characterized by the following function:

$$RDU = \sum_{c=1}^{C} [w_c(p) \times u(x_c)] \tag{2.1}$$

where c indexes the outcomes, x_c, from $\{1, \ldots, C\}$ with $c = 1$ being the smallest

[3]Sopher and Narramore (2000) provide the only experimental evidence I am aware of subsequent to the results of Camerer and Ho (1994) that are favorable to the "stochastic options" theory. However, this study does not employ the same kind of statistical tests as Hey and Carbone (1995).

outcome in the lottery and $c = C$ being the largest outcome in the lottery, $u(\cdot)$ returns the utility of its argument, $w_c(\cdot)$ returns the decision weight applied to outcome c given the distribution of probabilities ranked by outcome, p.

The utility function $u(\cdot)$ can take many functional forms due to it being unique up to an affine transformation, and can be normalized in various was, as illustrated by Hey and Orme (1994). It will sometimes be useful to use a specific functional form to make certain concepts clearer, and in such cases the constant relative risk aversion (CRRA) function will be employed:

$$u(x) = \frac{x^{(1-r)}}{(1-r)} \tag{2.2}$$

where r is the coefficient of relative risk aversion (Pratt 1964). Other popular functions such as the constant absolute risk aversion (CARA) function, or the Expo-Power function due to Saha (1993) can alternatively be employed without loss of generality.

The decision weight function, $w_c(\cdot)$, takes the form:

$$w_c(p) = \begin{cases} \omega\left(\sum_{a=c}^{C} p_a\right) - \omega\left(\sum_{b=c+1}^{C} p_b\right) & \text{for } c < C \\ \omega(p_c) & \text{for } c = C \end{cases} \tag{2.3}$$

where the probability weighting function, $\omega(\cdot)$, can take a variety of parametric or non-parametric forms. Many functions have been proposed for $\omega(\cdot)$. One is the "Inverse-S" shaped function popularized by Tversky and Kahneman (1992):

$$\omega(p_c) = \frac{p_c^{\gamma}}{\left(p_b^{\gamma} + (1-p_b)^{\gamma}\right)^{\frac{1}{\gamma}}} \tag{2.4}$$

Another is the power function used by Quiggin (1982):

$$\omega(p_c) = p_c^{\gamma} \tag{2.5}$$

The flexible function proposed by Prelec (1998) is also popular:

$$\omega(p_c) = \exp(-\beta(-\ln(p_c))^{\alpha}) \tag{2.6}$$

where $\alpha > 0$ and $\beta > 0$. In all cases there exist values for the shaping parameters which allow $w_c(p) = p_c$, the special case of EUT. When both a decision weight function applied to probabilities and a utility function applied to outcomes are defined, we have what is called a utility structure.

To make a general point about notation, consider lottery a with $C_a = 2$ possible outcomes, and lottery b with $C_b = 3$ possible outcomes. Suppose that $x_c^a \neq x_c^b \; \forall c$. A lottery a^* can be constructed with $C_{a^*} = C_a + C_b = 5$ such that it is equivalent to lottery a by adding the outcomes in lottery b to lottery a and setting the probabilities associated with each of these added outcomes equal to 0. Similarly for lottery b. This equivalence property holds for all EUT and RDU functional forms as zero probability outcomes are given no weight in either structure. The common set of combined outcomes is what Wilcox (2008) refers to as the choice scenario's "context." Throughout this chapter, the a^* form of lotteries will be assumed whenever notation concerning the composition of individual lotteries is used. This allows for identical notation to be utilized when comparing the probabilities of ranked outcomes across lotteries.

We now define a single choice scenario or task, t, as a discrete set of A mutually exclusive options from which a subject is asked to select one for payment. The most common form of such a task is a binary choice problem where subjects are

presented with 2 alternatives and asked to select one for payment, $t = \{X_1, X_2\}$. Each element of t, X_a, is a vector of the option's observable characteristics, such as various outcomes and the associated probability of those outcomes in a lottery. When presented with such a task, should the subject have a deterministic choice process and preference structure, the subject is assumed to choose option a over the alternative b if and only if the utility of option a was at least as great the utility of option b:

$$y_t = a \;\Leftrightarrow\; X_{at} \succcurlyeq X_{bt} \;\Leftrightarrow\; G(\beta_n, X_a) \geq G(\beta_n, X_b) \tag{2.7}$$

where $G(\cdot)$ is some utility structure such as RDU in equation (2.1), β_n is the unobservable vector of parameters of the utility structure for subject n, such as probability weights and utilities of outcomes, X is as defined above, and $y_t = a$ is a function that records which option a is chosen. The function y_t indicates the *chosen* alternative in task t, not the option *most preferred* by the agent. In the case of a stochastic choice process, y_t does not *necessarily* indicate which option in a set of alternatives is most preferred by the agent.

In the deterministic case presented in equation (2.7), the preference relation $\succcurlyeq$ provides all the necessary conditions for the creation of a utility function $G(\cdot)$, meaning it is complete, transitive, and continuous. Consequently, $\succcurlyeq$ provides a complete ranking of all the available alternatives in task t. It will be convenient to denote this ranking explicitly throughout this chapter by setting the a subscript of option X_{at} equal to its ranking in task t:

$$X_{1t} \succcurlyeq^n X_{2t} \succcurlyeq^n \ldots \succcurlyeq^n X_{at} \succcurlyeq^n \ldots \succcurlyeq^n X_{At} \tag{2.8}$$

This allows us to rank the utility of the A options in task t likewise:

$$G(\beta_n, X_{1t}) \geq G(\beta_n, X_{2t}) \geq \ldots \geq G(\beta_n, X_{at}) \geq \ldots \geq G(\beta_n, X_{At}) \tag{2.9}$$

It is worth reiterating that y_t indicates the option *chosen* in task t, while the value of the a subscript indicates the rank of the option in terms of the agent's *preference* in task t. With y_t defined and the options in task t ranked in terms of the subscript a, we can also define the set of options in task t not selected by subject n as follows:

$$Z = t \setminus y = \{z \in t \mid z \notin y\}. \tag{2.10}$$

As in equation (2.9), we can set the a subscript of the $A-1$ unchosen options in Z equal to their respective rankings in Z:

$$G(\beta_n, X^Z_{1t}) \geq G(\beta_n, X^Z_{2t}) \geq \ldots \geq G(\beta_n, X^Z_{at}) \geq \ldots \geq G(\beta_n, X^Z_{(A-1)t}) \tag{2.11}$$

With these base definitions in place, generalized formulations of the classes of stochastic models can be specified. As stated above, the RP model characterizes each observed choice made by an agent as conforming to a deterministic preference relation which is drawn at random from a set of such relations whenever the agent is confronted with a choice scenario. While the set preference relations can have a discrete distribution, the RP model is most commonly discussed in terms of utility functions with the relevant parameter vector, β_n, being continuously distributed according to some density function, $F_n(\beta|\alpha)$. In this definition, α is a vector containing the necessary parameters to define the shape of the distribution. Thus the probability of choice in a RP model is simply the probability that a vector β^*_n is drawn from the distribution governed by $F_n(\beta|\alpha)$ that would deterministically

satisfy that choice:

$$Pr(y_t = a) = Pr(a = 1) = Pr\left(\beta_n^* \,|\, G(\beta_n^*, X_{at}) \geq G(\beta_n^*, X_{1t}^Z)\right) \tag{2.12}$$

If we let $B_t = \left\{\beta_n^* \,|\, G(\beta_n^*, X_{at}) \geq G(\beta_n^*, X_{1t}^Z)\right\}$, given the density function $F_n(\beta|\alpha)$:

$$Pr(y_t = a) = \int_{\beta \in B_t} dF_n(\beta|\alpha). \tag{2.13}$$

Note the implication concerning first order stochastic dominance (FOSD)[4] with the above specification. Should X_a FOSD X_b, there is no monotonic preference relation β^* that will allow $G(\beta^*, X_{bt}) \geq G(\beta^*, X_{at})$. Thus, if we observe $y_t = b$ in such a scenario, the RP model collapses econometrically.

RE models are consistent with a model of the latent choice process that assumes that the utility of each option is evaluated with some error term, which is assumed to be homoscedastic with the SU model and generally assumed to be heteroscedastic with the derivative models of SU, but with a mean of 0 in either case. A choice is characterized as incorporating this error. Assuming a binary choice scenario, the error terms and utility functions must satisfy the following given the choice of option a:

$$\begin{aligned} G(\beta_n, X_{at}) + \epsilon_{at} \;&\geq\; G(\beta_n, X_{bt}) + \epsilon_{bt} \\ [G(\beta_n, X_{at}) + \epsilon_{at}] \;&-\; [G(\beta_n, X_{bt}) + \epsilon_{bt}] \geq 0 \end{aligned} \tag{2.14}$$

Setting $\epsilon_{at} - \epsilon_{bt} = \epsilon_t \lambda_n$, where λ_n is proportional to the standard deviation of ϵ_t,[5]

[4]Lottery a is said to FOSD lottery b iff:

$$\forall\, x_c,\ \sum_c^C p_c^a \;\geq\; \sum_c^C p_c^b \quad \textit{and} \quad \exists\, x_c,\ \sum_c^C p_c^a \;>\; \sum_c^C p_c^b$$

where c ranks the outcomes of lotteries a and b as described in equation (2.1). All deterministic theories of utility require the dominant option to be chosen over the dominated option.

[5]It is useful to recognize that what is described as "noise" in the data is determined by the

we can rewrite equation (2.14) as:

$$\begin{aligned} G(\beta_n, X_{at}) - G(\beta_n, X_{bt}) + \epsilon_t \lambda_n &\geq 0 \\ \epsilon_t \geq \frac{1}{\lambda_n} \left[G(\beta_n, X_{at}) - G(\beta_n, X_{bt}) \right] & \end{aligned} \tag{2.15}$$

Thus for RE models, the probability option a is chosen is given by:

$$\begin{aligned} Pr(y_t = a) &= Pr\left(\epsilon_t \geq \frac{1}{\lambda_n} \left[G(\beta_n, X_{bt}) - G(\beta_n, X_{at}) \right] \right) \\ &= 1 - F\left(\frac{G(\beta_n, X_{bt}) - G(\beta_n, X_{at})}{D(\beta_n, X_t)\lambda_n^*} \right) \end{aligned} \tag{2.16}$$

where λ^* is a precision parameter that remains after λ_n is adjusted by $D(\beta_n, X)$ for heteroscedastic models. As λ^* approaches 0, choice probabilities approach 0 or 1, while as λ^* approaches ∞, choice probabilities approach 0.5. The asterisk will be dropped from the remaining formulae to save space. $F(\cdot)$ is some cumulative distribution function (cdf) such that $F(0) = 0.5$ and $F(x) = 1 - F(-x)$. Usually $F(\cdot)$ is taken to be either the normal or logistic function, but any distribution function satisfying the previous conditions is acceptable. When discussing the SU model throughout this chapter, what is referred to is the model specified in equation (2.16) with $D(\beta_n, X_t) = 1$. This results in the SU model being said to be homoscedastic.

If utilizing the logistic function, equation (2.16) resembles a latent index model popularly used in a variety of econometric applications, but with a non-linear latent index. While equation (2.16) represents the common 2-option case, using

variance (or standard deviation) of the error term, not its mean. If the sign and magnitude of the mean of the error were anything but 0, choices would reveal a biased preference, but if the variance is sufficiently small, the choices are unlikely to reveal apparent deviations from a utility theory.

the logistic cdf, the RE model can be rewritten to accommodate A alternatives:

$$Pr(y_t = a) = \frac{\exp\left(\frac{G(\beta_n, X_{at})}{D(\beta_n, X_t)\lambda_n}\right)}{\sum_{c=1}^{J}\left[\exp\left(\frac{G(\beta_n, X_{ct})}{D(\beta_n, X_t)\lambda_n}\right)\right]} \tag{2.17}$$

With the TR model the "observed" probability of choice, $Pr(y_n = a)$, needs to be distinguished from the choice probability which would be modeled should the tremble not exist, $Pr_0(y_n = a)$. The agent is said to "tremble" with probability ϕ_n and select among the available options with equal probability, and with probability $(1 - \phi_n)$ select an option according to the underlying process:

$$Pr(y_t = a) = (1 - \phi)Pr(y_t = a) + \frac{\phi}{A}. \tag{2.18}$$

When Harless and Camerer (1994) proposed the TR model, the underlying choice process was made deterministic so that the option with the greatest utility has a choice probability of 1: $Pr_0(y_t = 1) = 1$. Loomes, Moffatt and Sugden (2002) however, proposed that $Pr_0(y_t = a)$ is generated from the RP model as specified in equation (2.13).

2.2 The Empirical Support for Stochastic Models

The previous section provided the econometric specification of the classes of stochastic models typically utilized in the literature to estimate parameters from choice data. The choice of model to utilize however, has not been *ad hoc*; because each stochastic model assigns very specific assumptions to the nature of the "noise" or randomness in choice data, these assumptions "amount to identifying restrictions

which may affect the relative performance of the theories under scrutiny" (Ballinger and Wilcox 1997, p. 1091). In much the same way as the various alternatives to EUT were proposed and then received rigorous testing, stochastic models have also been rigorously tested on the basis of their identifying restrictions.

Ballinger and Wilcox (1997) engaged in detailed tests of the SU and TR models, along with various assumptions about the heterogeneity of subjects, and find generally mixed results. The various models must be combined with somewhat unsavory assumptions about the nature of the heterogeneity of the population to make them statistically plausible. The TR model performs the worst, and requires the most assumptions. Ultimately Ballinger and Wilcox (1997, p. 1104) conclude by continuing to call for development and testing of the stochastic component of choice.

Carbone (1997) investigates the RP model, in addition to the TR and SU models, by estimating each model for each subject in an experiment, and finds that the SU model performs the best of the three, with RP a close second. Carbone (1997, p. 307) notes that if the set of alternative lotteries contains a lottery that stochastically dominates the others, the RP model requires the subject to always select the stochastically dominant lottery from the set, and that "this feature is of importance as it seems to capture well the experimental evidence." This feature of RP models, however, presents a problem when estimating preferences from choice data because, although violations of FOSD typically constitute a relatively small fraction of choices observed in experiments, this small fraction is reliably replicated in experiments.

Loomes and Sugden (1998) (LS) performed a similar investigation of the RP, TR, and SU models and strongly reject the TR model and to a lesser extent the

SU model. The SU model over-predicts violations of FOSD; however, they note, as Carbone (1997) did, that even one violation of stochastic dominance in a dataset is sufficient to cause the *stand-alone* RP model to collapse econometrically, and thus LS reject the model due to the few observations where stochastic dominance was violated. LS note, however, that the RP model can potentially accommodate these violations if it is combined with another stochastic choice model, such as the TR model. This point will be discussed in more detail subsequently. LS also report systematic deviations from EUT near the edges of the Marschak-Machina triangle,[6] and suggest that these cannot be fully accommodated by any of the stochastic models. In these instances, they suggest it is EUT that fails, not the stochastic models they consider.

Loomes, Moffatt and Sugden (2002) also test the RP, TR, and SU models. They recognize that "[t]here is no obvious reason to assume that only one of these forms of randomness is present" (2002, p. 106).[7] When faced with a choice situation, an agent may be best characterized as randomly drawing a preference from some set of preferences (RP model), evaluating the choice situation given that randomly drawn preference with some error (SU model), and then, with some positive probability, selecting an option irrespective of the agent's evaluation of the choice probability (TR model). Practically, this means estimating additional

[6]Recall that the vertices of the Marschak-Machina triangle represent potential outcomes in a lottery, given as a point in the triangle space. The probability of any given outcome is proportional to the geometric distance between the point representing the lottery and the vertex representing the outcome. Thus, if a lottery point lies on an edge of the triangle, the outcome associated with the opposite vertex has 0 probability. If a lottery point is minimally interior of an edge, the outcome associated with vertex opposite of the edge has a small, but positive, probability.

[7]This is a point with which I only partly agree. While I agree that there isn't an obvious reason why some models cannot be jointly present, there *is* a restriction on combining stochastic models: normative coherence. As discussed later, the RP model fails to satisfy this restriction.

parameters and making clear the identifying restrictions of any combination of these models, but mathematically these models are not mutually exclusive.[8]

Loomes, Moffatt and Sugden (2002) report that the best fitting stochastic model was RP plus TR paired with RDU. In addition, the estimated probability of TR diminished with the number of questions answered by the subjects, as did apparent deviations from EUT. When pairs with one lottery that stochastically dominates the other are removed from estimation, it is no longer clear that the RP model is superior to the SU model. This result suggests that "trembles" can be un-learned. Hey (2001) and Moffatt and Peters (2002) report similar results from experiments where it appears that noise is reduced with repetition. Hey (2001) also reports diminishing deviations from EUT with repetition.

Other than the prominent TR, SU, and RP models, there are a variety of alternatives that receive less attention, most of which are derivatives of the SU model with heteroscedastic error terms as opposed to the homoscedastic error of SU.[9] While TR and RP models can be manipulated in different ways to possibly explain more of the observed choice behavior,[10] such attempts often leave underlying, core problems with these models unattended to. For instance, if TR models predict that

[8]To illustrate the combination of RP, RE, and TR models derived in equations (2.13), (2.17), and (2.18):

$$Pr(y_t = a) = \frac{\phi}{A} + (1-\phi)\int_{\beta\in B_t} Pr\left(\epsilon_t \geq \frac{1}{\lambda}\left[G(\beta_n, X_{bt}) - G(\beta_n, X_{at})\right] \mid \beta\right) dF_n(\beta|\alpha)$$

This model would require the estimation of α, λ and ϕ.

[9]The SU model posits an error term with a variance that is independent of the domain of the utility function it is added to. Thus it is said to be homoscedastic. If the error term is correlated with some part of the domain of the utility function to which it is added, it is said to be heteroscedastic. There are many ways in which this correlation may occur, leading to a large variety of heteroscedastic models.

[10]For example, trembles could apply differently to FOSD pairs and non-FOSD pairs, and RP models could assume flexible distributions of the agent's preferences such as the Logit-Normal or Gamma distributions.

with some probability choices will be made irrespective of underlying preferences, why then should this probability vary depending on certain special cases of choice scenarios faced by the subject? Even with flexible distributions of RP models, there is no underlying utility theory which allows violations of FOSD, hence standard, stand-alone RP models can never accommodate such observed violations.[11] In contrast, the ability to manipulate the error term of the SU model in a tractable manner is part of what makes the SU model and its derivatives such popular models of stochastic choice.

To understand the motivations for developing the differing models that are derivatives of SU models, I briefly describe the concept of stochastic transitivity. Borrowing from Wilcox (2008, p. 210), consider three pairs of lotteries: $\{C, D\}$, $\{D, E\}$, and $\{C, E\}$, designated pairs 1, 2, 3, respectively. P_1 is the probability of choosing C in pair 1, and P_2 and P_3, are the probabilities of choosing D and C from pairs 2 and 3, respectively. We can define three forms of stochastic transitivity as follows:

$$\text{Strong Stochastic Transitivity (SST):} \quad min(P_1, P_2) \geq 0.5 \Rightarrow P_3 \geq max(P_1, P_2)$$

$$\text{Moderate Stochastic Transitivity (MST):} \quad min(P_1, P_2) \geq 0.5 \Rightarrow P_3 \geq min(P_1, P_2)$$

$$\text{Weak Stochastic Transitivity (WST):} \quad min(P_1, P_2) \geq 0.5 \Rightarrow P_3 \geq 0.5$$

Stochastic transitivity (ST) enforces a probabilistic form of transitivity for the same reasons that the non-stochastic transitivity axiom is employed for determin-

[11] However, non-standard RP models can sometimes accommodate violations of FOSD. An example of this is a model which specifies agents as randomly drawing a new preference relation for each option in a set of alternatives instead of randomly drawing one preference relation per set of alternatives. I explore the implications of this type of model, which I call the Random Preference Per Option (RPPO) model, later. For now, I will note that the RPPO model is very rarely utilized in the literature on stochastic models.

istic theories of choice: it is a mathematically convenient, and normatively useful way to model the choice process of a viable economic agent. Its consequence is that agents must make choices in a way that are at least stochastically consistent with economic success in an incentivized environment. Each version of ST makes descriptive and normative predictions which are operationalized by the stochastic model that incorporates them. For each proposed model described below, a particular version of ST is utilized because of its perceived superiority, either descriptively or normatively. The SU model, for instance, requires SST, whereas most of the proposed derivatives of SU attempt to relax this requirement in favor of either MST or WST.

Hey (1995) proposed three RE models with heteroscedastic error terms and conducted an experiment with 80 subjects to compare the heteroscedastic models to the homoscedastic SU model, denoted (H.1). In all three models, the Normal distribution was utilized for $F(\cdot)$. The first heteroscedastic model, (H.2), modeled the variance of the error as an exponential function of the time taken by the subject to give an answer to the question m and a corresponding coefficient, α:

$$D(\beta_n, X) = \exp(\alpha \times m) \tag{H.2}$$
$$\lambda_n = 1$$

Thus if $\alpha > 0$, the longer (shorter) it takes a subject to answer the question, the more (less) "noisy" the subject's responses should be. There is, however, no reason to expect $\alpha > 0$ *a priori*.[12]

The second heteroscedastic model, (H.3), modeled the error as an exponential

[12]Additionally, one might suspect that the time spent on a decision problem might itself be determined by the properties of the choice scenario, creating an endogeneity problem for this formulation.

function of the absolute value of the difference in utility of the alternatives multiplied by a coefficient, α:

$$D(\beta_n, X) = \exp\left(\alpha \times |G(\beta_n, X_{bt}) - G(\beta_n, X_{at})|\right) \tag{H.3}$$
$$\lambda_n = 1$$

Thus, if $\alpha < 0$, the larger the difference in utility of the alternatives, the smaller the noise.

The third heteroscedastic model, (H.4), modeled the error as an exponential function of the "difficulty" of the question, d, multiplied by a coefficient, α. In this specification, d is the average number of outcomes per option in the set of alternatives.

$$D(\beta_n, X) = \exp(\alpha \times d) \tag{H.4}$$
$$\lambda_n = 1$$

Thus, if $\alpha > 0$, the greater (smaller) the average number of outcomes per option, the greater (smaller) the noise. All of the heteroscedastic models nest the homoscedastic SU model as a special case (when $\alpha = 0$). While (H.3) and (H.4) did not perform particularly well, (H.1) was rejected in favor of (H.2) for 27 of 80 subjects at the 1% level, and 36 subjects at the 5% level (1995, p. 639).

Another derivative called the "Wandering Vector" (WV) model was proposed initially by Carroll (1980) and expanded on by Carroll and De Soete (1991). It makes the standard deviation of the error proportional to the Euclidean distance

between the probability vectors of the two alternative lotteries, a and b.

$$D(\beta_n, X) = \left[\sum_{c=1}^{C}(p_c^a - p_c^b)^2\right]^{1/2}$$

The original rationale for this model was to incorporate MST into a stochastic model: "for many realistic multidimensional stimulus domains, SST seems *too* strong. On the other hand, WST seems too weak" [emphasis in the original] (1991, p. 343). This model was proposed in the psychology literature to accommodate noisy perceptions of multidimensional stimuli, but this can be utilized as an economic model by reinterpreting the noisy perception of stimuli as noisy measurement of utility.

Wilcox (2011) expands on the "Contextual Utility" (CU) model initially proposed in Wilcox (2008). It makes the standard deviation of the error proportional to the difference in utility of the greatest non-zero probability outcome and the utility of the least non-zero probability outcome:

$$D(\beta_n, X_t) = max[u(x_{ct})] - min[u(x_{ct})]$$
$$st.\ w_c(x_{ct}) \neq 0$$

This model also satisfies MST and additionally allows for the "more risk averse than" relation of Pratt (1964) to be extended to the stronger "stochastically more risk averse than" relation. The "stochastically more risk averse than" relation of the CU model allows for interpersonal comparisons of risk-aversion in a way that is potentially more meaningful than with the SU model (2008, p. 221).

Busemeyer and Townsend (1993) propose "Decision Field Theory" (DFT), which when considering a pair of alternatives where one alternative is a certainty and the other is a lottery of only 2 outcomes, can be formulated in terms of the

$D(\cdot)$ function. If we define the set of outcomes that belong to the lottery as $H = \{x \in X_a \,|\, p_x < 1\} = \{h_1, h_2\}$ where $h_2 > h_1$ we have:

$$D(\beta_n, X) = [u(h_2) - u(h_1)] \sqrt{w_{h_1}(p)[1 - w_{h_2}(p)]}$$

The reasoning behind DFT is ultimately psychological. Busemeyer and Townsend (1993) posit that in cases where objective probabilities of outcomes are unknown, an agent may sample from past experiences with the same decision problem to estimate the objective probabilities. This kind of decision problem is deemed choice under "uncertainty" rather than choice under "risk." "Decision field theory was developed for this more natural type of uncertain decision problem" (1993, p. 436).

Blavatskyy (2014) proposes a model (BF) deemed "Stronger Utility," which shares properties with the "incremental EU advantage model" initially proposed by Fishburn (1987). There exist stochastic choice specifications which can be classified as both a "Stronger Utility" model and an "incremental EU advantage" model. The BF model makes the standard deviation of the error proportional to the difference in the utility of two abstract lotteries. The first abstract lottery is constructed to stochastically dominate all lotteries in the proposed decision scenario, but can itself be stochastically dominated by any other lottery which also scholastically dominates the proposed lotteries. The second abstract lottery is constructed to be stochastically dominated by both lotteries proposed in the decision scenario, while stochastically dominant any other lottery which is also stochastically dominated by both of the proposed lotteries. For FOSD pairs, this specification attaches a probability of 1 to the dominating option and 0 to the dominated option.

The three models proposed by Hey (1995) don't make any special predictions about the likelihood of choosing options in particular scenarios, such as with FOSD

pairs, other than to hypothesize that they improve on the explanatory fit of the SU model. They also require the estimation of additional parameters. The (H.2) and (H.4) models of Hey (1995) are very similar to the method utilized by Harrison and Rutström (2009, p. 142) in which observable characteristics of subjects are modeled as linear covariates of the core parameters to be estimated. However, Hey (1995) models the standard deviation to be exponential functions of these observable characteristics instead of linear functions, making the heteroscedasticity multiplicative rather than additive. The linear specification utilized by Harrison and Rutström (2009) is used to control for observable heterogeneity of subjects, not aspects of the choice scenario. The (H.3) model of Hey (1995), however, seems to reflect the same line of thinking as the subsequent derivatives of the SU model.

The other RE models mentioned above, however, often add new implications which generally help ease some of the shortcomings of the SU model. Both the CU model and the "sure thing vs two-outcome" DFT model discussed previously have the benefit of extending the "more risk averse than" relation of Pratt (1964) to the "stochastically more risk averse than" relation. DFT additionally requires that as the lottery becomes closer to first order stochastically dominating the CE, the probability of selecting the dominant option approaches 1. The BF model also requires that the probability of selecting the dominant option is always 1. Both the WV and the CU models enforce MST as opposed to the more restrictive strong SST, required by SU models. The CU, WV, DFT, and BF models don't require the estimation of any additional parameters on top of those required by the SU model, so any improvement of explanatory fit by them is free in terms of degrees of freedom used in classical estimation.

Wilcox (2008) provides a detailed discussion of the necessary implications of

the TR, RP, and SU models, along with some of the SU model's derivatives. He also discusses various well known events which can sometimes be attributed to stochasticity in choice: the Common Ratio Effect, low-frequency, but persistent, violations of FOSD, and changes in choice probabilities when lotteries are simply scaled.[13] Treatment is also given to whether models hold to various degrees of stochastic transitivity, whether they are descriptive of the "more risk averse than" relation, and how well the various models perform at predicting in-sample and out-of-sample choices.[14]

Wilcox (2008) estimates WV, CU, RP, SU, and an early variation of SU proposed by Luce (1958) called "Strict Utility" on the dataset from Hey (2001). He employs the method developed by Vuong (1989) to test if the log-likelihoods of the fitted models are significantly different from each other. Wilcox (2008, p. 273) finds that given an RDU structure, the CU model fits significantly better ($p = 0.013, p < 0.0001$) than the WV and "Strict Utility" models in-sample, and significantly better ($p < 0.0001, p = 0.0005, p < 0.0001, p < 0.0001$) than the SU, RP, WV, and "Strict Utility" models out-of-sample. Given an RDU structure, the RP model fits significantly better ($p = 0.0388, p < 0.0001$) than the WV and "Strict Utility" models in-sample, and significantly better ($p = 0.049, p < 0.0191, p < 0.0001$) than the SU, WV, and "Strict Utility" models out-of-sample. Neither the CU or RP models fit significantly better or worse than the SU model in-sample with a RDU structure. In all cases where the CU and RP models both fit better than a third

[13]Borrowing again from Wilcox (2008, p. 249): Consider four pairs of lotteries, $\{C, E\}$, $\{D, E\}$,$\{C, E'\}$, $\{D, E'\}$, and make the probability of selecting the first lottery in pair $\{C, E\}$, P_{ce}, and similarly for the remaining pairs. Simple Scalability requires that $P_{ce} > P_{de} \iff P_{ce'} > P_{de'}$. This requirement is only met by transitive utility structures, such as EUT and RDU, combined with stochastic models that satisfy SST.

[14]This particular topic was given extensive attention in Wilcox (2007).

model, the CU model fits better by a greater margin than the RP model.

With an EUT structure, the CU model significantly improves on every model except the SU model in-sample, and on every model out-of-sample, while the RP model fits significantly worse ($p < 0.058$) than the SU model and only improves on the "Strict Utility" significantly ($p < 0.0001$) in-sample, but fits significantly better ($p = 0.051, p = 0.016, p = 0.078$) than the SU, WV, and "Strict Utility" models out-of-sample. In addition to fitting better than the RP model in a direct comparison, the CU model fits better than every other model that the RP model also improves on, and by a greater margin than the RP model.

From these results it is clear that "Strict Utility" is a poor model in terms of goodness of fit given the alternatives: it doesn't significantly fit better than any alternative model, regardless of the utility theory, either in-sample or out-of-sample. The WV model only does marginally better: the only model it significantly improves upon is the "Strict Utility" model.

The more interesting story in light of the literature up to this point is the comparison of the SU, CU, and RP models. Considering in-sample fit, the SU model fits significantly better than the RP model with EUT, and with RDU there is no significant difference in goodness of fit. Out-of-sample the RP model fits significantly better than the SU model under both EUT and RDU. This echoes some of the mixed evidence up to this point concerning which of these two models is superior. New to the competition are the various SU derivatives. Two of these, the WV and "Strict Utility" models, perform relatively poorly in goodness of fit compared to the alternatives, but CU is shown to have generally superior performance compared to all the proposed models. The CU model has a greater log-likelihood than than all of the other models in all of the various test conditions,

in-sample or out-of-sample, with EUT or RDU. This difference is statistically significant for many of the comparisons, as noted above.

This discussion is not meant to be an exhaustive list of every proposed derivative of the SU model and their implications. Such a list would be very long and many of these derivative models deserve detailed discussion in their own right. This discussion simply serves to demonstrate that, as put by Wilcox (2008, p. 277), "we are witnessing a fertile period for stochastic model innovation now." Nearly all of this innovation has come from the RE class of models by changing the error term in the SU model from being homoscedastic to heteroscedastic, in very particular ways that have testable implications. The only apparent non-SU derived innovation was to combine the TR model with the RP model to help explain the low-frequency, but persistent, violations of FOSD observed in economic experiments.[15] Given the variety of theoretical implications of the various stochastic models and the repeatedly demonstrated sensitivity of goodness of fit measures to alternative stochastic models, it is no wonder that Wilcox (2008, p. 275) concludes: "It is hard to escape the conclusion that decision research could benefit strongly from more work on stochastic models."

While such a conclusion is undoubtedly true, there is a question relevant to economics concerning these models which has been sidelined in the continuing effort to find the "true" or "best" stochastic model: "What are the likely welfare implications of an economic agent's choices in an incentivized environment given an assumed stochastic model of choice?" This is the primary question of this chapter. Answering this question helps to draw the distinction between economics and

[15]This combination of models however is not an innovation to the RP model in itself as such a combination is just as possible with the SU model and any of its derivatives. As stated earlier, such a combination fails to address a core problem with the RP model: normative coherence.

decision theory or psychology. I argue that answering this question puts reasonable, restricting conditions on the econometric question "What is the best stochastic model to employ?"

2.3 Utility and its Relation to Welfare

With the RDU structure defined, and the stochastic models specified, we can define the basis of what will become our proposed metrics for individual welfare, the certainty equivalent. For any salient lottery X_a, and any vector β_n, there exists some certain outcome, CE_a, such that subject n is indifferent between the lottery and the certainty equivalent:

$$X_a \sim^n CE_a \;\Leftrightarrow\; G(\beta_n, X_a) = G(\beta_n, CE_a) \tag{2.19}$$

Combining the RDU structure from equation (2.1) with the utility function defined in equation (2.2), we can define the CE as follows:

$$\begin{aligned} \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{(1-r)}}{(1-r)} &= \frac{CE_a^{(1-r)}}{(1-r)} \\ CE_a &= \left((1-r) \times \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{1-r}}{(1-r)} \right)^{1/(1-r)} \end{aligned} \tag{2.20}$$

Thus if we assume some vector of β_n, of which r and the parameters governing $w_c(p)$ are elements, we can easily calculate the CE of lottery X_a.[16] If the utility function employed is monotonically increasing in the domain, as the CRRA function

[16] In general, the CE of any lottery can easily be calculated with numerical methods even if an analytical solution doesn't exist. This is because the CE must lie in the interval between the lowest outcome and the highest outcome. Numerically, one can just iterate through this interval until equation (2.20) is satisfied, or employ an optimization routine to look for the CE directly.

is, then this leads to the logical corollary of equation (2.9):

$$CE_{n1} \geq CE_{n2} \geq \ldots \geq CE_{na} \geq \ldots \geq CE_{nA} \tag{2.21}$$

With equations (2.20) and (2.21), we can also see that the *CE* can itself be considered a utility function, it is complete, transitive, and continuous, which is all that is required for a utility function to be well defined. Utilizing the *CE*s of options in a task is useful because of its potential to be considered a utility function, but also because it allows utility to be normalized to the units of the outcomes. This "*CE*" approach is similar to the "money-metric utility" function Samuelson (1974) employed to calculate welfare surplus. The "money-metric" utility function is used as a normalized utility function by, for instance, Diewert (1983) and King (1983).

We can employ the notation used for the set of unchosen alternatives, Z, derived in equation (2.10), to rank the *CE*s of each unchosen alternative just as in equation (2.11):

$$CE^{Z}_{n1} \geq CE^{Z}_{n2} \geq \ldots \geq CE^{Z}_{na} \geq \ldots \geq CE^{Z}_{n(A-1)} \tag{2.22}$$

Utilizing these *CE*s, four metrics are proposed to help measure welfare. If an agent with deterministic preferences and a deterministic choice process is presented with two lotteries to chose from, X_1 and X_2, she would choose X_1 and receive a welfare change of $CE_1 - CE_2$. This is the rational behind the first metric. With this metric, a change in welfare is measured as the difference between the *CE* of the option chosen and the *CE* of the highest ranked alternative option:

$$\Delta W_{nt} = CE_{nyt} - CE^{Z}_{n1t} = CE^{R}_{nt} \tag{2.23}$$

This welfare metric is similar to the notion of compensating equivalence in standard consumer theory. If equation (2.23) is positive, it calculates the minimum amount of money the agent would need in compensation in order to change her choice. If this metric is negative, it calculates the maximum the agent should be willing to pay in order to change her choice.

Another metric, which also utilizes the *CE*s, characterizes welfare received by choosing an option as a proportion of the *CE* of the option chosen and the *CE* that was ranked highest in the task:

$$\%W_{nt} = \frac{CE_{ny}}{CE_{n1}} \tag{2.24}$$

A variation of (2.23) which can be used to make statements about welfare across tasks could be:

$$\Delta W_{nT} = \sum_{t=1}^{T} \left(CE_{nyt} - CE_{n1t}^{Z} \right) \tag{2.25}$$

A similar variation of (2.24) could be:

$$\%W_{nT} = \frac{\sum_{t=1}^{T} CE_{ny}}{\sum_{t=1}^{T} CE_{n1}} \tag{2.26}$$

All of these metrics have strengths and weaknesses. Metric (2.23) is more relevant to the case of deterministic choice as it can change from task to task, while metrics (2.24) and (2.26) will become more relevant when discussing stochastic choice models and modest claims about inter-subject welfare. Any agent would be best off by making a choice which maximizes either of these metrics.

2.3.1 Special Case of Random Preferences: The Random Preference Per Option Model

Before beginning the discussion of welfare, a final model belonging to the RP class will be noted which requires a more involved explanation. It could be the case that an agent's choices are best characterized by a version of the RP model where a different β^*_{na} is drawn to evaluate every option in the set of alternatives. I refer to this type of RP model as the "Random Preference Per Option" (RPPO) model.

When evaluating option a, a standard preference relation is drawn from a distribution of preference relations that ranks each of the A alternatives, including option a, according to this preference relation:

$$X_{1t} \succcurlyeq^{na} X_{2t} \succcurlyeq^{na} \ldots \succcurlyeq^{na} X_{at} \succcurlyeq^{na} \ldots \succcurlyeq^{na} X_{At} \tag{2.27}$$

A utility function exists which represents these preference relations is shaped by β^a_n:

$$G(\beta^a_n, X_{1t}) \geq G(\beta^a_n, X_{2t}) \geq \ldots \geq G(\beta^a_n, X_{at}) \geq \ldots \geq G(\beta^a_n, X_{At}) \tag{2.28}$$

As was shown in equation (2.20), a CE can be calculated for each of these β^a_n conditional utility functions such that:

$$CE^a_{n1} \geq CE^a_{n2} \geq \ldots \geq CE^a_{na} \geq \ldots \geq CE^a_{nA} \tag{2.29}$$

In equations (2.27), (2.28), and (2.29), the ordinal ranking of the set of alternatives is the same. As stated previously, the CE of an option has the same utility function properties as $G(\cdot)$, and is additionally normalized by the units of the options themselves. If only one β_n vector is drawn to derive cardinal values for the set of alternatives, we have the RP model discussed previously.

But, with RPPO models, when evaluating another option b, such that $b \neq a$, another, potentially different, preference relation is drawn and each of the A alternatives, including both a and b, are ranked according to this preference relation:

$$X_{1t} \succcurlyeq^{nb} X_{2t} \succcurlyeq^{nb} \ldots \succcurlyeq^{nb} X_{at} \succcurlyeq^{nb} \ldots \succcurlyeq^{nb} X_{At} \tag{2.30}$$

where $\succcurlyeq^{nb}$ represents the preference relation drawn by subject n when considering option b across all A outcomes. A utility function exists which represents these preference relations shaped by β_n^b:

$$G(\beta_n^b, X_{1t}) \geq G(\beta_n^b, X_{2t}) \geq \ldots \geq G(\beta_n^b, X_{at}) \geq \ldots \geq G(\beta_n^b, X_{At}) \tag{2.31}$$

And again, a CE can be calculated for each of these β_n^b conditional utility functions such that:

$$CE_{n1}^b \geq CE_{n2}^b \geq \ldots \geq CE_{na}^b \geq \ldots \geq CE_{nA}^b \tag{2.32}$$

The difference between the realizations of equations (2.29) and (2.32) is twofold. First, the different β_n^* vectors may lead to different ordinal rankings of the same set of alternatives. Second, if $\beta_n^a \neq \beta_n^b$, then there may be two different cardinal evaluations for the same option.

The RPPO model takes the cardinal value of each option, evaluated using its own β_n^* vector, and constructs an ordinal ranking of the set of alternatives based on these individual evaluations. To deal with the potential issue of comparing different utility functions cardinally, the utility functions $G(\cdot)$ should be normalized somehow.[17] In line with the previous discussion, we will utilize the CE for each

[17]Without the normalization, the RPPO model requires an unusual interpretation of preference relations to accommodate particular aspects of relatively common $G(\cdot)$ functions. An example will be presented later that will make this clearer.

option based on its individually drawn β_n^* vector.

Adding the superscripts $\{1, 2, \ldots, x, \ldots, X\}$ to the CE to indicate the option that it is associated with, and the subscripts $\{1, 2, \ldots, a, \ldots, A\}$ to the CE to represent its rank in the set of alternatives conditional on its individual β_n^x vector, we could have the following ordinal ranking:

$$CE_{n1}^1 \geq CE_{n2}^2 \geq \ldots \geq CE_{na}^x \geq \ldots \geq CE_{nA}^X \tag{2.33}$$

Notice that the superscripts match the subscripts in this example, but this need not always be the case. The superscripts represent the unranked options, while the subscripts represent the RPPO ranked options. The following could also represent an ordinal ranking from the RPPO model:

$$CE_{n1}^3 \geq CE_{n2}^1 \geq \ldots \geq CE_{na}^x \geq \ldots \geq CE_{nA}^X \tag{2.34}$$

The RPPO model, as discussed here, is characterized as having random preference parameters, but is otherwise deterministic in characterizing choice. That is, this RPPO model characterizes an agent as choosing the option with the highest individually evaluated CE. Just as with the RP model, additional stochasticity can be imposed by including measurement error as with RE models and/or a tremble event as with TR models. These additional elements create unnecessary mathematical complexity for the current discussion, but they will be touched on briefly later. Just as in equation (2.10), the non-chosen options can be expressed as the set of $A - 1$ alternatives in task t that doesn't include the chosen option:

$$Z = t \setminus y = \{z \in t \mid z \notin y\} \tag{2.35}$$

The RPPO model therefore constructs a choice function as follows:

$$y_t = x \Leftrightarrow CE^x_{na} \geq CE^z_{nb} \quad \forall z \in Z \tag{2.36}$$

That is, y_t is a function that indicates that option x is chosen in task t if and only if the CE associated with option x is greater than or equal to the CE of any other option z. The probability of $y_t = x$ is determined by the joint probability of observing the set $\left\{\beta^x_n, \{\beta^Z_n\}\right\}$, such that:

$$CE(\beta^x_n, X_{1t}) \geq CE(\beta^z_n, X_{at}) \forall z \in Z \tag{2.37}$$

Call such a set $\mathbf{B^t_n}$. The probability of $y_t = x$ is therefore:

$$Pr(y_t = x) = \int_{\beta^x_n \in \mathbf{B^t_n}} \int_{\beta^Z_n \in \mathbf{B^t_n}} f_{\mathbf{B^t_n}}\left(\beta^x_n, \{\beta^Z_n\}|\alpha\right)\, d\beta^x_n\, d\beta^Z_n \tag{2.38}$$

where $f_{\mathbf{B^t_n}}(\beta^x_n, \{\beta^Z_n\}|\alpha)$ is the joint density of the elements of $\mathbf{B^t_n}$, the shape of which is governed by the vector α.

2.4 The Stochastic Money Pump: A Tool for Describing Welfare Accumulation

With the various classes of stochastic models defined, I begin the discussion of the welfare implications of these models by first introducing a decision problem which resembles the "Money Pump" argument against intransitive structures in deterministic choice theory, though with several important distinctions that will be made clear later. I will refer to this relatively simple thought experiment as a "Stochastic Money Pump" (SMP). Assume that there is an experimental economist who, through cleverly selected choice problems, is able to correctly identify the utility structure and stochastic process governing the choices of subjects with perfect

knowledge. That is, the utility structure and relevant parameters β_n, as well as the correct stochastic model and relevant parameters that completely characterize some subject n are all known by the experimenter.

The experimenter then offers to sell a lottery ticket to the subject for some amount of money, and should the subject agree to buy the ticket, the experimenter offers to buy the ticket back from the subject for a lower amount of money. The subject can refuse the initial purchase, buy the ticket and refuse to sell the ticket back, or buy the ticket and sell it back to the experimenter. How often can the experimenter expect to be successful in extracting the difference between the buying and selling price of the ticket from the subject without giving the subject anything, and what are the welfare implications of this pair of transactions? The various classes of stochastic models can all have different welfare implications while predicting similar observed choice behavior.

To make this example concrete, we can work this problem out numerically assuming 3 different subjects, Amy, Beth, and Cate. Amy operates with a random preference model of choice, Beth operates with a contextual utility model of choice, and Cate makes choices deterministically but with a tremble. Amy, Beth, and Cate all have the same utility structure of the RDU special case where $w_c(p) = p_c$ for all p_c and incorporate the CRRA utility function from equation (2.2). Amy's distribution function $F_n(\beta|\alpha)$ is $N(\mu, \sigma^2) = N(0, 0.01)$, thus normal with α featuring a mean equal to 0 and a standard deviation of 0.1. Beth operates with a β vector that is composed of $r = 0$, and a $\lambda_n = 0.015$. Cate operates with a β vector that is composed of $r = 0$ and a probability of trembling of $\phi = 0.816$. All values picked in this example are for ease of calculation, but the implications hold when generalized to different parameter values.

The lottery ticket has a 0.5 probability of an outcome of 10 and a 0.5 probability of an outcome of 100, and thus an expected value of 55. The experimenter offers to sell each of the subjects the lottery for 55.50, and to buy the lottery back at 54.50. These values are 0.50 above and below the *CE* of the lottery for Beth and Cate, and 0.50 above and below the mean *CE* of the lottery for Amy. The probability of the experimenter successfully extracting money costlessly is approximately equal for all three subjects:

$$Pr(Extraction) = Pr(Buy) \times Pr(Sell) \approx 0.167 \tag{2.39}$$

The manner in which the this probability is reached is different for each subject: For Amy,

$$\begin{aligned}
B_{Buy} &= \{\beta_A |\ G(\beta_A, X_{Lottery}) \geq G(\beta_A, X_{Buyprice})\} \\
&= \{r_A \Big|\ r \leq -0.0232\} \\
B_{Sell} &= \{\beta_A |\ G(\beta_A, X_{Sellprice}) \geq G(\beta_A, X_{Lottery})\} \\
&= \{r_A \Big|\ r \geq 0.0232\} \\
Pr(y_{Buy} = Buy) &= \int_{\beta \in B_{Buy}} dF_A(\beta|\alpha) = \phi(B_{Buy}, 0, 0.01) \\
&\approx 0.408 \\
Pr(y_{Sell} = Sell) &= \int_{\beta \in B_{Buy}} dF_A(\beta|\alpha) = \phi(B_{Sell}, 0, 0.01) \\
&\approx 0.408 \\
Pr(Extraction_A) &= Pr(y_{Buy} = Buy) \times Pr(y_{Sell} = Sell) \\
&\approx .167
\end{aligned} \tag{2.40}$$

where ϕ is the cumulative normal distribution. For Beth,

$$
\begin{aligned}
D(\beta_B, X) \times \lambda_B &= 0.015[u(100) - u(10)] \times 1 \\
&= 1.35 \\
Pr(y_{Buy} = Buy) &= \frac{\exp\left(\frac{G(\beta_B, X_{Lottery})}{D(\beta_B, X)\lambda_B}\right)}{\exp\left(\frac{G(\beta_B, X_{Lottery})}{D(\beta_B, X)\lambda_B}\right) + \exp\left(\frac{G(\beta_B, X_{Buy\ price})}{D(\beta_B, X)\lambda_B}\right)} \\
&\approx 0.408 \\
Pr(y_{Sell} = Sell) &= \frac{\exp\left(\frac{G(\beta_B, X_{Sell\ price})}{D(\beta_B, X)\lambda_B}\right)}{\exp\left(\frac{G(\beta_B, X_{Lottery})}{D(\beta_B, X)\lambda_B}\right) + \exp\left(\frac{G(\beta_B, X_{Sell\ price})}{D(\beta_B, X)\lambda_B}\right)} \\
&\approx 0.408 \\
Pr(Extraction_B) &= Pr(y_{Buy} = Buy) \times Pr(y_{Sell} = Sell) \\
&\approx .167
\end{aligned}
\tag{2.41}
$$

For Cate,

$$
\begin{aligned}
Pr_0(y_{Buy}) &= \begin{cases} 1 & , \quad G(\beta_C, X_{Lottery}) > G(\beta_C, X_{Buy\ price}) \\ 0.5 & , \quad G(\beta_C, X_{Lottery}) = G(\beta_C, X_{Buy\ price}) \\ 0 & , \quad G(\beta_C, X_{Lottery}) < G(\beta_C, X_{Buy\ price}) \end{cases} \\
Pr_0(y_{Sell}) &= \begin{cases} 1 & , \quad G(\beta_C, X_{Sell\ price}) > G(\beta_C, X_{Lottery}) \\ 0.5 & , \quad G(\beta_C, X_{Sell\ price}) = G(\beta_C, X_{Lottery}) \\ 0 & , \quad G(\beta_C, X_{Sell\ price}) < G(\beta_C, X_{Lottery}) \end{cases}
\end{aligned} \tag{2.42}
$$

$$
\begin{aligned}
Pr(y_{Buy} = Buy) &= (1-\phi)Pr_0(y_{Buy}) + \frac{\phi}{A} \\
&= (1 - 0.816)(0) + (0.816)/2 \\
&= 0.408 \\
Pr(y_{Sell} = Sell) &= (1-\phi)Pr_0(y_{Sell}) + \frac{\phi}{A} \\
&= (1 - 0.816)(0) + (0.816)/2 \\
&= 0.408 \\
Pr(Extraction_C) &= Pr(y_{Buy} = Buy) \times Pr(y_{Sell} = Sell) \\
&\approx .167
\end{aligned}
$$

While the observed choice behavior is identical for all three subjects, the welfare implications are not. According to the metrics defined by equations (2.23) and

(2.24), this decision problem has the same welfare implications for Beth and Cate:

$$\begin{aligned}
\Delta W_{(B,C),Buy} &= CE_{(B,C),Lottery} - CE_{(B,C),Buy\ \ Price} = 55 - 55.5 = -0.50 \\
\Delta W_{(B,C),Sell} &= CE_{(B,C),Sell\ \ Price} - CE_{(B,C),Lottery} = 54.5 - 55 = -0.50 \\
\%W_{(B,C),Buy} &= \frac{CE_{(B,C),Lottery}}{CE_{(B,C),Buy\ \ Price}} = \frac{55}{55.5} \approx 0.99 \\
\%W_{(B,C),Sell} &= \frac{CE_{(B,C),Sell\ \ Price}}{CE_{(B,C),Lottery}} = \frac{54.5}{55} \approx 0.99
\end{aligned} \tag{2.43}$$

In this case, the welfare implications of such decision problem are clear for both the TR and RE models: with a roughly 0.167 probability, Beth and Cate will make 2 consecutive "mistakes" or "choice errors," which results in 1 unit of money and 1 unit of CE being extracted from each of them. The other potential results are easy to calculate. With a $1 - Pr(y_{Buy} = Buy) = 0.592$ probability, Beth and Cate make no mistakes and experience a welfare gain:

$$\begin{aligned}
\Delta W_{(B,C),Buy} &= CE_{(B,C),Buy\ \ Price} - CE_{(B,C),Lottery} = 55.5 - 55 = 0.50 \\
\%W_{(B,C),Buy} &= \frac{CE_{(B,C),Buy\ \ Price}}{CE_{(B,C),Buy\ \ Price}} = \frac{55.5}{55.5} = 1
\end{aligned} \tag{2.44}$$

With a $Pr(y_{Buy} = Buy)(1 - Pr(y_{Sell} = Sell)) \approx 0.242$ probability, Beth and Cate make the mistake of buying the lottery ticket, but not the mistake of selling it back for less:

$$\begin{aligned}
\%W_{(B,C),T} &= \frac{\sum_{t=1}^{T} CE_{(B,C),y,t}}{\sum_{t=1}^{T} CE_{(B,C),1,t}} = \frac{55 + 55}{55.5 + 55} \approx 0.995 \\
\Delta W_{(B,C),T} &= \sum_{t=1}^{T} \left(CE_{(B,C),y,t} - CE^{Z}_{(B,C)1,T} \right) = 55 - 55.5 + 55.5 - 55 = 0
\end{aligned} \tag{2.45}$$

The RP model characterizing Amy's choices, however, does not provide an intuitive understanding of the welfare implications of this decision problem. The

RP model discussed here, the stand-alone RP model, requires that every choice by Amy be characterized by a deterministic preference relation according to some vector β_A randomly drawn from a distribution. Thus:

$$\begin{aligned}
\Delta W_{(A),Buy} &= CE_{(A),Lottery} - CE_{(A),Buy\ Price} \geq 0 \\
\Delta W_{(A),Sell} &= CE_{(A),Sell\ Price} - CE_{(A),Lottery} \geq 0 \\
\%W_{(A),Buy} &= \frac{CE_{(A),Lottery}}{CE_{(B,C),Lottery}} = 1 \\
\%W_{(A),Sell} &= \frac{CE_{(A),Sell\ Price}}{CE_{(B,C),Sell\ Price}} = 1
\end{aligned} \tag{2.46}$$

According to the metric definition in (2.23) and the decision process for Amy defined in (2.40), the ΔW welfare evaluations in (2.46) must be weak inequalities. However, the only situation where $\Delta W_{(B,C),Buy} = 0$ is when $r = -0.0232$, which has a probability of 0 given that $F_A(\beta|\alpha)$ is continuous. Similarly for the choice between selling or not selling. With a probability approaching 1, the RP model predicts that should an extraction occur the choices causing the extraction leave the subject with positive welfare.

Before moving on to the normative implications of these welfare characterizations, we revisit the SMP thought experiment with two new entrants, Dana and Emma. Suppose that Dana's choices are characterized as being in accordance with the RPPO model discussed previously, and Emma operates a tremble model as defined by Loomes, Moffatt and Sugden (2002), where Pr_0 is defined by the RP model. We can pose the same questions concerning Dana and Emma's choices that we asked of Amy, Beth, and Cate's choices: "How often can the experimenter expect to be successful in extracting the difference between the buying and selling price of the ticket from the subject without giving the subject anything and what

are the welfare implications of this pair of transactions?" As we will see below, the answers to these questions for Dana are exactly the same as for Amy, and though the math involved with Emma is slightly more complicated, the counterintuitive interpretation of welfare caused by the RP model remains.

Suppose that for Dana, the marginal distributions of each option's β_A^* vector used to construct $f_{\mathbf{B_n^t}}(\beta_n^x, \beta_n^z|\alpha)$ are identical and uncorrelated.[18] Also, Dana's marginal distribution functions are all $N(\mu, \sigma^2) = N(0, 0.01)$, thus normal with α consisting of a mean equal to 0 and a standard deviation of 0.1. The joint density function, $f_{\mathbf{B_n^t}}(\beta_n^x, \beta_n^z|\alpha)$, is therefore:

$$\begin{gathered} N_2(\mu, \Sigma) \\ \mu = \begin{Bmatrix} \mu_x \\ \mu_z \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \\ \Sigma = \begin{Bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_z \\ \rho\sigma_x\sigma_z & \sigma_z^2 \end{Bmatrix} = \begin{Bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{Bmatrix} \end{gathered} \tag{2.47}$$

where μ is the vector of means, Σ is the covariance matrix for the joint density function $f_{\mathbf{B_n^t}}(\beta_n^x, \beta_n^z|\alpha)$. The choice behavior for Dana is as follows:

[18] These assumptions are made for mathematical simplicity in the following example. There is no obvious reason why it should be necessary that these distributions be identical or uncorrelated, though the conceptual result would be the same even if they were not.

For Dana,

$$
\begin{aligned}
CE_{Buy\ Price} &= Buy\ Price\ \forall\ \beta_D \\
CE_{Sell\ Price} &= Sell\ Price\ \forall\ \beta_D \\
\mathbf{B_D^{Buy}} &= \{\beta_D^x, \beta_D^z | CE^x_{Lottery} \geq CE^Z_{Buy\ Price}\} \\
&= \{r_D^x, r_D^Z | r_D^x \leq -0.0232, r_D^Z \in \Re\} \\
\mathbf{B_D^{Sell}} &= \{\beta_D^x, \beta_D^z | CE^x_{Sell\ Price} \geq CE^Z_{Lottery}\} \\
&= \{r_D^x, r_D^Z | r_D^x \geq 0.0232, r_D^Z \in \Re\} \\
Pr(y_{Buy} = Buy) &= \int_{\beta_D^x \in \mathbf{B_D^t}} \int_{\beta_D^Z \in \mathbf{B_D^t}} f_{\mathbf{B_D^t}}\left(\beta_D^x, \{\beta_D^Z\} | \alpha\right)\ d\beta_D^x\ d\beta_D^Z \\
&= \phi(r_D^Z \leq -0.0232, 0, 0.01) \times \phi(r_D^x \in \Re, 0, 0.01) \\
&= 0.408 \times 1 \\
Pr(y_{Sell} = Sell) &= \int_{\beta_D^x \in \mathbf{B_D^t}} \int_{\beta_D^Z \in \mathbf{B_D^t}} f_{\mathbf{B_D^t}}\left(\beta_D^x, \{\beta_D^Z\} | \alpha\right)\ d\beta_D^x\ d\beta_D^Z \\
&= \phi(r_D^x \in \Re, 0, 0.01) \times \phi(r_D^Z \geq 0.0232, 0, 0.01) \\
&= 1 \times 0.408 \\
Pr(Extraction_D) &= Pr(y_{Buy} = Buy) \times Pr(y_{Sell} = Sell) \\
&\approx .167
\end{aligned}
\tag{2.48}
$$

This thought experiment presents a special case where the RPPO model essentially reduces to the standard RP model. This is due to the fact that the *CE* of any certain amount of money is equal to that amount of money. Because this is true, the distributions of the β_D vectors associated with the buying and selling prices are irrelevant.[19]

[19]Take any degenerate lottery X comprised of a single outcome x with a probability of $p_x = 1$. The utility of this lottery is $U_X = w_x(p_x)u_x(x)$, where w_x is any probability weighting function

The reason why the utility functions are normalized can also be made clear with this example. The CRRA function described in (2.2) and utilized in the above example has some interesting properties around $r = 1$: $u(x|r) = \ln(x)$ at $r = 1$, $u(x|r) \to \infty$ as $r \to 1$ from the left, and $u(x|r) \to -\infty$ as $r \to 1$ from the right.[20] If the RPPO model didn't normalize the CRRA function to its *CE*, the set of $\{\beta_n^x, \{\beta_n^Z\}\}$ would be contradictory in its elements due to the properties of the CRRA function around 1. To demonstrate, assume that the utility of the lottery is evaluated with $r_n^x = -0.0232$: what values of r_n^Z satisfy equation (2.20) such that Dana would decide to purchase the lottery? We might expect that since $r_n^x = -0.0232$ is the value of r_n that sets the utility of the lottery and the utility of the buy price equal to each other, should the buy price be evaluated with greater risk aversion than the lottery, equation (2.37) will hold. That is, we might expect that should $r_n^Z > -0.0232$ Dana would prefer the lottery over the buy price.

Intuitively this makes sense: an increase in risk aversion corresponds to an increase in the concavity of the utility function under EUT (or holding a probability weighting function constant under RDU) which implies lower utility for a given outcome as risk-aversion increases. While this is true when the CRRA function

---

and u_x is any utility function. Since $w_x(p_x = 1) = 1$ for every probability weighting function, $U_X = u_x(x) = u_x(CE_x)$. Since u_x is a well-defined utility function, $CE_x = x$ is a solution for equation (2.19) for every u_x and every x when $p_x = 1$. This solution is unique when u_x is monotonic, as with the CRRA function employed in the example.

[20]The use of $\ln(x)$ for $r = 1$ is not as *ad hoc* as it may seem. Wakker (2008, p. 1333) shows that if equation (2.2) is normalized, it can be seen "that the normalized logarithmic function is the limit of the normalized power functions for r tending to 0 [1], both from above ($r > 0$) [$r > 1$] and from below ($r < 0$) [$r < 1$]:"

$$\lim_{r \to 0} \frac{x^r - c^r}{d^r - c^r} = \frac{\ln(x) - \ln(c)}{\ln(d) - \ln(c)} \quad \forall\, x > 0, d > c > 0$$

While Wakker (2008) uses the single exponent version of the power function, the same limit applies to the formulation of the CRRA function used in equation (2.2), with the bracketed values of r in the above quote representing the revised limits.

is normalized to its CE, this isn't true without the normalization. Without the normalization, β_n^Z consists of $-0.023 \leq r_n^Z \leq 0.9814$ and $r_n^Z > 1$.

The gap of $(0.9814, 1)$ is due to the properties of the CRRA function around 1. The un-normalized RPPO model allows for a relatively risky option to be chosen over a relatively less risky option should the less risky option be evaluated using a preference relation indicating intense risk aversion, but not when the less risky option is evaluated using a preference relation indicating only moderate risk aversion. Intuitively we might expect the reverse to be true, but it is mathematically possible. The possibility of this kind of gap is not removed even if the marginal distributions which make up $f_{\mathbf{B_n^t}}(\beta_n^x, \{\beta_n^Z\}|\alpha)$ are correlated and can be exacerbated if they are not identical.

Now, suppose Emma operates a TR model with the Pr_0 choice probabilities generated by the RP model described for Amy and the tremble parameter described for Cate. Emma's choice behavior is defined as follows:

For Emma,

$$
\begin{aligned}
B_{Buy} &= \{\beta_E |\ G(\beta_E, X_{Lottery}) \geq G(\beta_E, X_{Buyprice})\} \\
&= \{r_E \Big|\ r \leq -0.0232\} \\
B_{Sell} &= \{\beta_E |\ G(\beta_E, X_{Sellprice}) \geq G(\beta_E, X_{Lottery})\} \\
&= \{r_E \Big|\ r \geq 0.0232\} \\
Pr_0(y_{Buy}) &= \int_{\beta \in B_{Buy}} dF_E(\beta|\alpha) = \phi(B_{Buy}, 0, 0.01) \\
&\approx 0.408 \\
Pr_0(y_{Sell}) &= \int_{\beta \in B_{Buy}} dF_E(\beta|\alpha) = \phi(B_{Sell}, 0, 0.01) \\
&\approx 0.408 \\
Pr(y_{Buy} = Buy) &= (1-\phi)Pr_0(y_{Buy}) + \frac{\phi}{A} \\
&= (1 - 0.816)(0.408) + (0.816)/2 \\
&= 0.483072 \\
Pr(y_{Sell} = Sell) &= (1-\phi)Pr_0(y_{Sell}) + \frac{\phi}{A} \\
&= (1 - 0.816)(0.408) + (0.816)/2 \\
&= 0.483072 \\
Pr(Extraction_E) &= Pr(y_{Buy} = Buy) \times Pr(y_{Sell} = Sell) \\
&\approx 0.233
\end{aligned}
\tag{2.49}
$$

We have a more complicated result with Emma when we attempt to describe her welfare. The preferences in this model are provided by the aspects that belong to the RP model. This means that not only is $Pr_0(y_{Buy})$ the probability that Emma would chose to buy the lottery should she not experience a "tremble,"

but it is also the probability that the choice to buy the lottery is the result of greater utility being accumulated from the lottery ticket than is associated with the buying price. Similarly for $Pr_0(y_{Sell})$ and the choice to sell the ticket. Thus, given $Pr(Extraction_E)$, $Pr_0(y_{Buy})$, and $Pr_0(y_{Sell})$, we have:

$$\begin{aligned} Pr\left(\%W_{(E),Extraction} = 1\right) &= Pr(Extraction_E) \times Pr_0(y_{Buy}) \times Pr_0(y_{Sell}) \\ &= 0.233 \times 0.408 \times 0.408 \\ &\approx 0.0388 \end{aligned} \tag{2.50}$$

That is, we characterize Emma as having the 1 unit of money extracted by the SMP, *and* this extraction as having resulted in optimal welfare for Emma 3.88% of the time. Similarly, this probability can be interpreted as a lower bound on $Pr(\Delta W_{(E),Extraction} \geq 0)$.

The probability that the welfare surplus metric is positive in the event of an extraction is equivalent to the probability of an extraction occuring times the probability that the welfare change from buying the ticket plus the welfare change from selling the ticket is positive:

$$Pr(Extraction_E) \times Pr(\beta_{Buy}, \beta_{Sell} \Big| \ CE^{Buy}_{Lottery} - CE^{Buy}_{BuyPrice} + CE^{Sell}_{SellPrice} - CE^{Sell}_{Lottery})$$

$$Pr(Extraction_E) \times Pr(\beta_{Buy}, \beta_{Sell} \Big| \ CE^{Buy}_{Lottery} - CE^{Sell}_{Lottery} \geq 1)$$

If we set $\beta_{B,S} = \{\beta_{Buy}, \beta_{Sell} \Big| \ CE^{Buy}_{Lottery} - CE^{Sell}_{Lottery} \geq 1\}$,[21] we have:

$$0.233 \times \int_{\beta_{Buy} \in \beta_{B,S}} \int_{\beta_{Sell} \in \beta_{B,S}} F(\beta_{Buy}|\alpha) F(\beta_{Sell}|\alpha) d\beta_{Buy} d\beta_{Sell} \geq 0.0388 \tag{2.51}$$

This complication arises because Emma's choice function has both TR and RP

[21] Recall that the *CE* of an certain amount is always that certain amount. So the *CE* of the "Buy Price" is 55.5, and the *CE* of the sell price is 54.5 for any utility function. Therefore $CE_{BuyPrice} - CE_{SellPrice} = 1$ for all utility functions.

elements. Sometimes Emma will value the prospect of buying the lottery ticket extremely highly, not tremble, and choose to buy, and then value the prospect of selling the ticket back only somewhat negatively, tremble, and then sell it back. The welfare gained in the buying choice can therefore sometimes outweigh the welfare lost in the selling choice, resulting in a net positive change of welfare. Note that every time Emma values the buying of the ticket *and* the selling of the ticket positively, the net change in welfare will also be positive. Thus, 0.0388 constitutes a lower bound on $Pr(\Delta W_{(E),Extraction} \geq 0)$.

2.5 The Normative Coherence of Stochastic Models

Having clarified the welfare implications of several representative stochastic models, I can discuss the normative implications of these models. When a variety of stochastic models were detailed by Becker, DeGroot and Marschak (1963), they were intended as a way to "circumvent the difficulty" associated with the problem that "the preference choices of the chooser are often inconsistent with each other." In laying out the implications of these models, Becker, DeGroot and Marschak (1963) detail descriptive implications about the frequency with which certain types of choices would be observed, but do not make any normative claims. The intent behind developing these models was to provide greater *descriptive* veracity to EUT, apparently while maintaining EUT as the *normative* force behind these models.

The historical course that lead to EUT becoming the dominant orthodox theory appears to have started with a descriptive justification, then a normative justification. Moscati (2016) details the correspondences between Paul Samuelson, Leonard Savage, Jacob Marschak, and Milton Friedman concerning the axiomatization of

EUT by Von Neumann and Morgenstern (1944) and its burgeoning acceptance as the orthodox theory of choice involving risky outcomes. The correspondence highlights Samuelson's strong initial reluctance to accept EUT based on his dissatisfaction with what later became known as the independence axiom, which he called a "gratuitously-arbitrary-special-implausible hypothesis" (Moscati 2016, p. 225). Savage, and to a lesser extent Marschak and Friedman, advocated for EUT as being descriptively accurate, theoretically simple, and, eventually, normatively robust. Samuelson had strong reservations about the descriptive veracity of EUT, advocated by both Savage and Friedman, stating that the phenomena associated with gambling are "infinitely richer" than EUT permitted (Moscati 2016, p. 227). Friedman also admitted that in order to be able to explain certain gambling phenomena, EUT would "need complication" (Moscati 2016, p. 229). Eventually, however, Samuelson was persuaded of EUT's normative force by Savage's discussion of what would become known as the "Sure-Thing Principle," without necessarily relenting on the descriptive claims.

What is meant by "normative force" in these discussions is the specification of what a "rational" agent "ought" to do. Marschak, in correspondence with Samuelson, makes this distinction: "It may be *usual* for village carpenters [...] to deviate from the advice of Euclidian geometers [...] All the same, they would be better advised to behave rationally by following Euclid" (Moscati 2016, p. 229). In relenting to Savage's normative arguments, Samuelson concedes that the normative value of the Independence Axiom, and by extension EUT, makes it useful as an assumption "defining 'rational' behavior" (Moscati 2016, p. 231) despite maintaining that EUT doesn't provide "a very illuminating explanation" of gambling or investment behavior even "as a first approximation" (Moscati 2016, p. 232).

The appeal of normative arguments, and their apparent superiority to descriptive veracity arguments in the case of accepting EUT, is based on their adding to a theory's potential to generate statements about the welfare of agents in incentivized environments. Pigou (1929, preface to the third edition) prefaces his third edition with a comment to future students of economics:

> The complicated analyses which economists endeavour to carry through are not mere gymnastic. They are instruments for the bettering of human life. The misery and squalor that surround us, the injurious luxury of some wealthy families, the terrible uncertainty overshadowing many families of the poor - these are evils too plain to be ignored. By the knowledge that our science seeks it is possible that they may be restrained. Out of the darkness light! To search for it is the task, to find it perhaps the prize, which the "dismal science of Political Economy" offers to those who face its discipline.

Varian (1996, p. 238) writes: "economics is a policy science and, as such, the contribution of economic theory to economics should be measured on how well economic theory contributes to the understanding and conduct of economic policy." Leamer (2012, p. 30) echoes this sentiment: "The primary goal [of economics] should not be to amuse each other with mathematical complexities [...] [it] should be to design policy interventions - policies that are intended to help achieve social objectives, notably the highest level of well-being for the largest number of people."[22]

It is with these sentiments, both historical and contemporary, that I interrogate the extent to which stochastic models presented in this chapter provide useful methodologies for the design and interpretation of policy interventions. Critical to this objective is the ability to make statements on how changes in stocks of assets

[22]This utilitarian conceptualization of the primary goal of economics can be relaxed somewhat without losing force. Policy interventions can be, and often are, crafted to improve the welfare of specific sub-populations. For instance, the kind of individuals referenced by Pigou may have a disproportionate number of policies crafted to improve their lot relative to their numbers in the total population without this being at odds with the goals of economics.

affect the welfare of individuals. The various criteria used to assess the normative validity of stochastic models will largely be interpreted from the works of Grüne-Yanoff, Marchionni and Moscati (2014), Berg (2014) and Hands (2014). Normative criteria are relations of particular means and ends that describe what an agent "ought" do or not do in certain circumstances. Economic theories purporting to be normatively justified must provide the mechanisms that satisfy these relationships. In this framework, I will primarily discuss the capacity of stochastic models provide to generate outcomes that have commonly been employed as normative criteria.

2.5.1 Economic Existence and Objective Betterness Criteria

For an economic theory to gain normative force, it should satisfy the constraint that an agent ought not to make choices systematically in such a way as to drive herself out of economic existence. For example, this could mean that an economic theory should not normatively justify an agent's choice to deliberately put themselves into bankruptcy.[23]

The popular, often informal, way EUT is justified in light of this criteria is through its invulnerability to money pumps. The traditional money pump is defined as a series of trades that will succeed in extracting the entirety of an agent's stock of assets, say some amount of good A, if the agent has intransitive preferences, such as $B \succcurlyeq A$, $C \succcurlyeq B$, $A \succcurlyeq C$ with at least one relation being strict. This occurs when the extractor offers to trade his B for the agent's A, then his C for her B, and finally his $A - \epsilon$ for her C, where ϵ is a sufficiently small, but positive amount of

[23]Bankruptcy here is in meant in the abstract sense of the loss of all assets without recourse to recover them. In the United States for instance, given what are called "Chapter 9" and "Chapter 11" bankruptcy provisions, it could be quite rational to engage in institutionally controlled bankruptcy under certain circumstances.

good A such that $A-\epsilon \succcurlyeq C$, *and those trades are accepted*. This process is repeated until the agent has no remaining quantity of good A, and is thus economically eliminated.

Hands (2014, pp. 402-403) refers to this argument as "empirical elimination:"

> This is the argument that agents who act in ways that violate [rational choice theory] will (in fact) cease to exist or at least cease to play an active role among the relevant class of decision-makers. The two most common forms of this argument are the money pump (for agents who have intransitive preferences and thus make choice mistakes) and the Dutch book[...]

Grüne-Yanoff, Marchionni and Moscati (2014, p. 336) refer to "universal loss-avoidance considerations:"

> If an agent violates the transitivity condition on preferences, then that individual can be "money pumped:" all wealth can be taken from her, simply by trading goods with her in a way that exploits her preference intransitivity[...] Consequently, to the extent that any one wants to avoid such sure losses, one must satisfy the corresponding internal consistency criteria.

Cubitt and Sugden (2001) however, methodically decompose the argument that failure to satisfy consistency axioms, in particular transitivity, results necessarily in vulnerability to money pumps. They develop a detailed methodology for describing decision problems without the need to specify an underlying theory of value that traditionally denotes rewards at nodes in decision trees, and provide several examples of how an agent could have preferences that violate consistency axioms and yet remain invulnerable to money pumps.[24] Cubitt and Sugden (2001, p. 154) conclude: "Thus, in relation to what we take to be their original objectives, money

[24]See Cubitt and Sugden (2001) for a novel methodology on atheoretic representations of decision problems, as well as the examples indicated.

pump arguments are a failure."

The critiques of Cubitt and Sugden (2001) show that adherence to EUT is not a necessary condition for invulnerability to a money pump, only a sufficient one. Indeed, it is possible for an agent to conform to standard consistency axioms of completeness and transitivity and still be economically eliminated.[25] However, the claim that an agent should not make choices in such a way as to have her stock of assets stripped away from her is still largely seen as a necessary condition for a normative theory. More generally, we claim that for a theory to be useful in guiding policy it should accord with the intuition that being stripped of her stock of assets renders an agent worse off.[26] I call this the economic sustainability (ES) criterion.

Cubitt and Sugden (2001, p. 141) define a relation communicating this necessity of not valuing less assets to more, which they state "coincides with *weak statewise dominance*" and that "Many theories of choice under uncertainty generate choice functions which respect statewise dominance and hence, within this setup, objective betterness. Obviously, this is true of expected utility theory. But it is also true of, for example, Quiggin (1982) rank-dependent expected utility theory [RDU]."

Marschak (1950, p. 112) seems to endorse this notion of ES as a normative criterion, describing agents who choose dominated offers as worse off: "In dealing

[25]The poem "Smart" by Shell Silverstein (1974) exhibits such an agent. A son receives a dollar from his father and gleefully trades it for two quarters, and those two quarters for three dimes, and those three dimes for four nickles, and those four nickels for five pennies. It is clear that the son's preferences for these objects are both complete and transitive, satisfying the axioms for consistency. Perhaps the son has not been economically eliminated, since the five pennies do still have some value, but should there exist infinitely divisible denominations of hard currency, the son would approach elimination. The implied reaction of the father to his son's trades, lost on the son, suggests that such a set of preferences is normatively unacceptable.

[26]The requirement of "just" compensation in eminent domain provisions in the United States implies that policy crafters agreed that stripping a citizen of her assets, even for the public good, leaves her worse off.

with his environment ('nature' which includes 'society') a man who often makes mistakes in his inferences and his sums is, in the long run, apt to fare less well than if he had been a better logician and arithmetician."

With this interpretation of ES as a necessary criterion for a normative theory of choice, we return to the SMP thought experiment presented earlier. In the SMP, at no point does an agent face a single choice that incorporates FOSD. Thus, taken in isolation, a choice to buy or sell the lottery ticket cannot be said to necessarily reveal anything about changes in the agent's welfare. Should an agent buy the ticket and then proceed to sell the ticket back for less money, the agent's choices consider as a set lead directly to her stock of assets being reduced by 1 unit. For a considered theory to be a normatively acceptable theory of choice it must, at a minimum, describe the agent as having become worse off than if she had never engaged in the trades.

This concept of describing objective betterment across aggregated choices is no different to previous uses of the standard money pump argument.[27] Utilizing the above example of a standard money pump, it isn't until the extractor seeks to trade back a reduced quantity of the agent's original endowment, $A - \epsilon$, for her current stock of assets, C, that the agent is said to be made worse off. Furthermore, all trades prior to the final trade cannot be said to necessarily make the agent worse off, but are simply transfers of different stocks of goods. With all the trades aggregated together however, the final trade is what completes the extraction that leaves the agent with a smaller stock of assets.

All the stochastic models examined in the SMP thought experiment allow for the

[27]Note that Cubitt and Sugden (2001) do not dispute that the strict loss of a stock of assets, as happens when a subject is "money pumped" over a series of choices, is normatively unacceptable. They argue that EUT is sufficient, but not necessary, to prevent such a loss of assets.

choices resulting in an extraction to occur. All of the examples except for Emma even predict the extraction will occur with the same probability. Additionally, every model allows for a mechanism to describe the agent's welfare at every choice. However, it is only the TR and CU models associated with the Beth and Cate examples that satisfy the normative requirement described in this section.

Equations (2.41) and (2.42) show that the CU and TR models allow for the observed extraction to occur, while equation (2.43) states that these agents are worse off than they would have been had they not engaged in the trades. The TR and CU models would make similar allowances for the selection of dominated lotteries in FOSD lottery pairs, and also correctly align the subjective welfare assessment with the ES criterion. Even considering alternative RE models that assign a probability of 0 to dominated lotteries in FOSD lottery pairs, and thus econometrically collapse with every proposed utility function having equal, 0, likelihood, all proposed utility functions would *still* respect ES. Thus, it is a general result that stand-alone RE and TR models successfully adhere to the normative criteria described in this section.

The RP, RPPO, and the TR-RP models of Amy, Dana, and Emma however, all share the property that the extraction of 1 unit of wealth from the agents can be described subjectively as an increase in welfare. For the RP and RPPO models, this description of an extraction as welfare improvement is guaranteed, while for the TR-RP model this description is applicable under some conditions. Thus, I argue that this property of the RP class of stochastic models violates the normative criteria described in this section.

2.5.2 Willingness to "Correct" Choices Criterion

The willingness to correct choices (WCC) criteria is often interpreted in multiple ways. A general version of WCC requires that should an agent deviate from the requirements of a theory of choice, and should she then be confronted with the deviation, she will willingly "correct" her choices to conform with the theory. Sometimes this argument states that the theory in question is normatively justified if such a willingness to correct choices be observed empirically.

This criteria seems to require multiple moving pieces. First, an agent must make choices that apparently violate the economic theory in question, such as the kind of choices described in the money pump and SMP examples. Secondly, someone, often characterized as an "expert" in decision problems, must confront the agent with the theory and a prescription of how the agent should make choices in light of this theory. Finally, the agent, presumably having been convinced about the validity of the theory, chooses again and selects a set of choices that conform to the theory. An alternate version of the criterion only requires that the agent in question be an "expert" or some kind of exemplary decision maker herself.

Such a confrontation of experts is famously said to have occurred at a conference on decision theory held in Paris in May 1952 between Leonard Savage and Maurice Allais (Allais 1953, p. 1; Moscati 2016, p. 221). Savage, having presented arguments for EUT during the conference, was asked by Allais to choose a lottery ticket he would prefer to own from two pairs of lottery tickets. Savage made choices over these two lottery pairs that violated the Independence Axiom of EUT, and was confronted by Allais with proof of the violation. Allais argued that if even Savage did not make choices in accordance with EUT, it could not be accepted as a

normative theory. Savage replied that he had made a mistake in his choices, and having been confronted with the error was willing to correct them. He argued that his willingness to change his choices upon having been confronted with his error was evidence in favor of the normative validity of EUT. In the terms of his private correspondence with Samuleson, he would have nothing to "reproach" himself for in having changed his choices to be in accordance with EUT (Moscati 2016, p. 230).

Most often, agents engaging in economically salient choices are not confronted by other agents directly with suggestions about how their choices could better conform to some normative theory. Instead, agents are generally only confronted with the way their choices deviate from an economic theory indirectly, through consequences of the operation of market forces. We can reformulate the WCC criterion with market forces as the confronter of agents and pose the following question: "When confronted with salient market outcomes resulting from choices that are discordant with some theory, do agents willingly change their choices to be in accordance with the theory in similar subsequent market interactions?"

Chu and Chu (1990) deliver evidence in favor of EUT responding to exactly this question. They conduct an experiment replicating the design of Grether and Plott (1979), which found frequent deviations from EUT, specifically, apparent violations of transitivity.[28] Chu and Chu (1990) differ from the previous replications of Grether and Plott (1979) by actively engaging in arbitrage with subjects who committed apparent violations of transitivity, thus experimentally simulating the kind of market forces that would operate outside of the laboratory. Chu and Chu (1990, p. 910) find that incidences of apparent violations of transitivity were

[28]There is an extensive literature, reviewed in Chapter 1, concerning the Grether and Plott (1979) experiments.

eliminated from all subjects' choices after an average of 1.71 arbitrage transactions, and that "subjects displayed substantially fewer reversals [apparently intransitive choices] after they were exposed to a marketlike environment in previous rounds of games." The largest number of arbitrage transactions needed to induce conformity to the transitivity axiom was 3, and this number of transactions occurred for only one subject across all of their treatments. On the other hand, Braga, Humphrey and Starmer (2009), in a similar experiment, find that while most anomalies are eliminated with market exposure, others arise as rounds progress and that there may exist a "just enough" amount of market exposure to induce conformity to EUT.

Our revised WCC criterion requires that exposure to market forces must induce a "correction" to choices that do not consistently lead to worse welfare outcomes, such as the arbitrage transactions implemented by Chu and Chu (1990). TR and RE models handle this requirement in 2 ways. First, both classes of models are structured such that the most likely choice in every scenario is one that will leave the agent at least as well off as she currently is. Thus there is a built-in correcting mechanism in these models. Secondly, one could model the relevant stochastic parameters, ϕ in the case of the TR model and $D(\beta, X)$ or λ in the case of RE models, as being determined in part by the number of choice problems (or arbitrage transactions) experienced with the assumption of a negative coefficient. With this specification, increased market interaction would lead to lower probabilities of mistakes, and in the limit would result in choices conforming with EUT, just as was observed by Chu and Chu (1990).

However, it isn't clear which choices should be "corrected" given an RP model and a set of choices which in aggregate do not conform to EUT. Each choice is

the optimal choice given the preference relation drawn for that task. Changing a choice in the RP model to make an aggregate choice pattern better conform to EUT therefore implies that an option with lower utility should be selected over an option with higher utility for certain tasks. This then, paradoxically, reduces the expected welfare of the corrected choice pattern. Thus, the RP class of models do not provide normatively useful statements with respect to the WCC criterion.

2.6 Concluding Remarks

In this chapter I have shown that stochastic models of economic agents have been given an increasing amount of attention, and have discussed three classes of models at length: the "Tremble" (TR) model, the "Random Error" (RE) model, and the "Random Preference" (RP) models. The primary purpose for which these models were developed was to account descriptively for the apparent deviations from EUT frequently observed in experimental data. To further this descriptive purpose, these models were formulated in such a way as to make specific predictions about observed choice probabilities in particular choice scenarios. The TR model requires that all deviations from EUT are equally likely; the SU model requires Strong Stochastic Transitivity; and the RP model requires that violations of FOSD have a zero probability of occurring.

Additionally, these models are formulated mathematically to be parsimonious and modular, meaning that their stochastic elements effectively never interfere with the elements concerned with utility. The TR model and most of the RE models only require the estimation of one parameter in addition to whatever model of utility is employed, whereas common interpretations of the RP model only require the additional estimation of two parameters. The modularity of these models also

allows researchers such as Loomes, Moffatt and Sugden (2002) to combine them to address potential descriptive shortfalls arising from their individual application.

However, the main purpose of this chapter is not to illustrate the descriptive capabilities of these models, but to draw attention to their normative implications and potential justifications. I propose a simple thought experiment involving a contrived decision problem, the "Stochastic Money Pump" (SMP), and several hypothetical agents who individually operate as if using differing classes of stochastic models when making choices. The SMP is structured in such a way as to demonstrate the possibility of an agent entering into a decision problem and then leaving with a strictly lower stock of assets, as in the traditional "money pump." With the SMP in hand, we show that, at least for this decision problem, each of the major classes of stochastic models can be parameterized in such a way that they produce exactly the same descriptive choice probabilities. This descriptive equality, however, does not in any way imply that the welfare implications of these models are always equivalent or even coherent.

I show that for the examples of the models given, the TR and RE models make equivalent implications concerning the welfare of agents, in particular, that agents who have some of their assets extracted from them are modeled as strictly worse off than if this had not happened. The same cannot be said about the RP model or any of its derivatives, such as the proposed "Random Preference Per Option" (RPPO) model or the RP-TR combination model. The mathematical descriptions of welfare from these models beg the larger question of how or whether the stochastic models can be normatively justified.

I attend to the discussion of normative coherence by first noting the historical and contemporary emphasis on the potential of an economic theory to make

statements about welfare that are useful for assessing policy. I attempt to limit the discussion by focusing on two prominent criteria used in the literature on normative justifications: "Economic Sustainability" (ES) and the "Willingness to Correct Choices" (WCC).

I argue that the ES criterion requires adherence to a notion that strictly greater stocks of assets are objectively better than smaller stocks of assets, and normative models must require agents to subjectively conform to this valuation. I conclude that only the RE and TR models make coherent statements with respect to ES; the SMP thought experiment demonstrates how the RP model allows for the implication that agents who have had assets extracted from them are subjectively better off than if they had not suffered extraction.

I argue that to effectively posit the WCC criterion as a condition for normative coherence, the role of the "confronter" must be delegated to the market. In this interpretation, the WCC builds on the notion of "objective betterness" described by ES by requiring that having been confronted with the market outcomes of a choice, should the agent face the same decision again, the most likely choice she will make will leave her at least as well off as her previous choice. I conclude that, again, the RE and TR models make coherent statements with respect to this criterion, but the RP models do not.

I find that the motivations underpinning the development of the RP model as a descriptive model of choice probabilities are relatively sound. There has been some evidence showing the RP model does statistically fit choice data better than many alternative models, particularly when combined with the TR model, though there has been a significant amount of evidence showing that heteroscedastic RE models outperform RP. The RPPO model presents the possibility of still better

statistical fit, particularly since it allows for the violation of FOSD in certain choice scenarios.[29] Whether the RPPO model does in fact perform statistically better than other stochastic models is an empirical question that hasn't been given much attention.

I argue that the RPPO model shouldn't be given much attention. The RP and RPPO models both exist in a territory of economic modeling that concerns itself with statistical fit and predictive quality, which are indeed things economists should be concerned about, but cannot be used to make persuasive arguments about how an agent maintains economic welfare through choices. It is the latter of these two concerns which constitute the economic, as opposed to the technical, content of the inquiry. The exercise presented in this chapter helps to inform the econometric question proposed by analysts of "what is the 'best' stochastic model?" by suggesting that the "best" model is the one which has the greatest "fit" among the models *that make normatively coherent statements about the welfare of the modeled agents.*

I conclude from this experiment that the RP model's failure to provide coherent statements on how the choice mechanism relates to a useful interpretation of welfare renders it unsuitable for modeling the kind of choices over risky alternatives that it was initially developed to describe. There may very well be other economic environments in which the RP model provides useful normative statements, but when explaining "noise" in binary lottery choices there are models that can make sense of the noise and additionally provide useful normative statements. When

[29]The normalized RPPO model doesn't allow an option with a single certain outcome to be chosen over an option with a larger single certain outcome, but the un-normalized RPPO model might, depending on the particular utility functions being utilized and distributions of parameters.

attempting to describe choices over risky alternatives, these models should be preferred to the RP model and its derivatives, regardless of any evidence that suggests it is a better fitting model. I recognize that rejecting a model that potentially fits choice data from economic experiments better than its alternative seems counter-empiricist, but if estimates from these models are only useful for describing choice probabilities, and not the welfare implications of the choices made, the model is not useful in economic applications.

Chapter 3

The Welfare Implications of Stochastic Models

Given the discussion about how the various stochastic models generally support incorporation of the normative notion of welfare, I reintroduce the question asked earlier in Chapter 2, section 2.2: "What are the likely welfare implications of an economic agent's choices in an incentivized risky environment given an assumed stochastic model of risky choice?" The conclusion for the Random Preference (RP) model and its derivative, the Random Preference Per Option (RPPO) model, is "no perfectly coherent statements can be made." As stated in the conclusion of Chapter 2, the Random Error (RE) and Tremble (TR) models do not suffer from this inadequacy, and will be referred to as "coherent models."

Many stochastic models make specific restrictions on the probabilities associated with certain special choice scenarios. For instance, options in a choice scenario which are first order stochastically dominated (FOSD) by another option are prohibited under the RP model and severely restricted under many heteroscedastic RE models. For any choice scenario, an option which is FOSD by another option can also be said to provide the subject with less expected utility, and thus less

expected welfare. Thus, at the individual choice level there is a perfect relationship between likelihood and welfare realization. In this chapter, I make clear that this relationship between likelihood and welfare realization does not hold for aggregated choice patterns. I also show that choice patterns which display behavior that cannot be rationalized by a utility function can often result in greater welfare than choice patterns which can be rationalized.[1]

How often this divergence between welfare realization and likelihood occurs, and the extent to which this divergence in welfare terms is meaningful for an individual agent, depends on the nature of the choice scenarios presented to the agent and the agent's preferences. Additionally, the likelihood of an agent holding preferences in a population of interest will determine how likely we are to observe choice patterns that are able to be rationalized, but are in fact suboptimal. To help understand the relationship between the likelihood of observing a choice pattern and its potential welfare consequences, I conduct a numerical exercise with methods related to Maximum Simulated Likelihood (MSL) and utilizing the Multiple Price List (MPL) proposed by Holt and Laury (2002) (HL) for a given hypothetical population of agents. To understand how the distribution of preferences in a population influence the expected welfare realization of a population, I repeat this analysis for many different populations. First however, I revisit some notation from Chapter 2, briefly describe some econometric methods for identification, and then propose some further notation to make concepts cleaner.

[1] A similar point is made by Wilcox (1993, p. 1402) concerning the relationship between the expected value of choice patterns and their conformity to axioms.

3.1 Notation and Estimation

For any salient lottery X_a, and any vector of parameters β_i, there exists some certain outcome, CE_a, such that subject i is indifferent between the lottery and the certainty equivalent:

$$X_a \sim^i CE_a \;\Leftrightarrow\; G(\beta_i, X_a) = G(\beta_i, CE_a) \tag{3.1}$$

where $G(\cdot)$ is some utility function with all the usual properties. For our purposes throughout this chapter, we will assume some variation of the Rank Dependent Utility (RDU) structure defined as follows:

$$RDU = \sum_{c=1}^{C} [w_c(p) \times u(x_c)] \tag{3.2}$$

where $u(\cdot)$ is the CRRA utility function throughout this chapter,

$$u(x) = \frac{x^{(1-r)}}{(1-r)}, \tag{3.3}$$

and $w_i(p)$ is the decision weight applied to option a defined as

$$w_c(p) = \begin{cases} \omega\left(\sum_{k=c}^{C} p_k\right) - \omega\left(\sum_{k=c+1}^{C} p_k\right) & \text{for } c < C \\ \omega(p_c) & \text{for } c = C \end{cases} \tag{3.4}$$

where $\omega(\cdot)$ is a probability weighting function and $w(\cdot)$ are decision weights. In cases where $\omega(p_c) = p_c$, the RDU structure is equivalent to Expected Utility Theory (EUT) as the decision weights for each option will equal their objective probabilities, p. Many parametrized probability weighting functions allow for this special case to occur.

Combining the RDU structure with a CRRA utility function, we can define the CE as follows:

$$\begin{aligned} G(\beta_i, X_a) &= \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{(1-r)}}{(1-r)} = \frac{CE_a^{(1-r)}}{(1-r)} \\ CE_a &= \left((1-r) \times \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{1-r}}{(1-r)} \right)^{1/(1-r)} \end{aligned} \tag{3.5}$$

where c indexes the C outcomes of option a in task t.

I continue the notation from Chapter 2 where the value of a also represents each option's ordinal rank among the alternative options in task t. Thus $X_1 \succcurlyeq X_2$ and $X_a \succcurlyeq X_b$, where $b \geq a$. Similarly, we define the set of unchosen options from the full set of alternatives as $Z = t \setminus y = \{z \in t \mid z \notin y\}$, with the subscript on the elements of Z indicating their ordinal rank in the set of Z. Thus $X_1^Z \succcurlyeq X_2^Z$ and $X_a^Z \succcurlyeq X_b^Z$, where $b \geq a$.

The probability of any choice a by some subject i, given some vector of parameters β, being observed for a task t, is denoted by $\Pr(y_t = a)$, where $y_t = a$ is an indicator function that records option a as being chosen in task t. To make explicit the dependency of this probability on the option in question, the subject, the task, and the β vector, this relationship will be re-framed as follows:

$$P_{iat}(\beta_i) = \Pr(y_t = a) \tag{3.6}$$

The likelihood of observing a series of choices is the product of the probability of observing the option chosen for each task across all tasks, T :

$$P_{iT}(\beta_i) = \prod_{t}^{T} P_{iat}(\beta_i) \tag{3.7}$$

This is the standard likelihood function applied to binary choice data, which

assumes choices are statistically independent between tasks. We could take the log of equation (3.7) and conduct standard maximum likelihood estimation (MLE) by searching for the vector $\hat{\beta}_i$ which maximizes the log-likelihood function:

$$LP_{iT}(\beta_i) = \sum_{t}^{T} \ln\left(P_{it}(\beta_i)\right) \tag{3.8}$$

Thus, the maximum likelihood estimator $\hat{\beta}_i$ for subject i is:

$$\hat{\beta}_i = \arg\max_{x} \sum_{t}^{T} \ln\left(P_{it}(\beta_i)\right) \tag{3.9}$$

We can utilize this estimator to recover the *CE* for every option in every task, and then utilize these *CE*s to recover our best estimate of the proportion of welfare the subject obtained. While conducting welfare analysis given individually estimated parameter vectors is rare in the economics literature,[2] the recovery of parameter vectors through MLE is relatively common. Hey and Orme (1994), Wilcox (2015) and Hey (2001) provide several examples of parameter estimation. These particular examples, however, are distinctly different from other uses of MLE in experimental economics, primarily because equation (3.9) is estimated for every subject individually, as opposed to pooling all subject data together and estimating a parameter vector for one, representative agent (RA), as proposed in the pioneering Camerer and Ho (1994).

There are legitimate methodological (and practical) reasons for modeling choices across subjects as the choices of a single RA. For instance, the analyst could be primarily concerned with the economic characteristics of the whole sample, rather than with the individuals composing the sample. As shown in Harrison and

[2]Examples of this kind of analysis are Harrison and Ng (2016, 2018) and Harrison, Martínez-Correa, Ng and Swarthout (2017).

Rutström (2008, p. 63), it is easy to allow the $\hat{\beta}$ to be determined by a linear combination of observable characteristics of the subjects and/or experimental treatments. For instance, if the race, gender and age of each of the subjects were known, we could estimate:

$$\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 \times race + \hat{\beta}_2 \times gender + \hat{\beta}_3 \times age \tag{3.10}$$

where $\hat{\beta}_1$ through $\hat{\beta}_3$ represent the mean marginal effects[3] of race through age respectively on the vector $\hat{\boldsymbol{\beta}}$.

Another useful technology demonstrated by Harrison and Rutström (2009) for RA modeling is the use of finite mixture modeling. This is when a finite mixture of stochastic specifications are estimated jointly on the same data along with mixture parameters. For instance,

$$\begin{aligned} \boldsymbol{P_T} = \prod_t^T \left[\sum_m^M \pi_m \times L_T^m(\beta^m) \right] \\ st. \sum_m^M \pi_m = 1 \end{aligned} \tag{3.11}$$

where π_m is the proportion of model m in the mixture, β is the vector of parameters to be estimated in model m and L_T^m is the likelihood of the choice data across the T tasks explained by model m given the vector β^m.[4] Similarly, the log-likelihood for finite mixture models is defined as:

$$\begin{aligned} \boldsymbol{LP_T} = \sum_t^T \left[\ln \left(\sum_m^M \pi_m \times L_T^m(\beta^m) \right) \right] \\ st. \sum_m^M \pi_m = 1 \end{aligned} \tag{3.12}$$

[3] "Marginal" with reference to the default set of characteristics captured by the constant $\hat{\beta}_0$.

[4] In this formulation, each observation can be generated by any of the models in the mixture, as opposed to each subject having all observations being generated by one of the models in the mixture.

Thus M β^m vectors and $M-1$ π_m scalars need to be estimated. These parameters can additionally each be determined by observed characteristics, as in equation (3.10). This method can be useful if the analyst wishes to estimate the proportion of a sample which more closely adheres to RDU versus EUT for instance, or if the analyst wants to determine if there is some heterogeneity in the sample that is revealed by choice, but unobservable otherwise. Harrison and Rutström (2009, p. 141) use this method to jointly estimate a specification composed of Prospect Theory (PT) and EUT. They employ a Strong Utility (SU) stochastic model to generate the probabilities. Although there does not appear to be any literature doing so, it is possible to estimate a mixture of two differing stochastic models. For instance, an analyst could use a mixture model to determine what proportion of subjects in a dataset are better characterized by the SU or TR models.[5]

There are also some methodological problems, or at least limitations, when conducting estimation on pooled data. The estimates represent the means of the relevant parameters in the sample, but the distributions of these parameters and whether these distributions are correlated can potentially provide more important information to analysts.[6] While the methods described in equations (3.10) and

[5]This process could be used to help with the econometric limitations of the pure RP model, since those subjects who violate FOSD can be picked up by an alternative model which permitted such violations. This process, of course, doesn't resolve the RP model's normative failures discussed in chapter 2.

[6]For an example of why it could be problematic to make inferences about a population from an estimate which represents the mean of a distribution of preferences consider a population that has preferences distributed as *Logit-Normal* $\sim \mathcal{N}(0,5)$. Logit-Normal is a distribution in which the logistic function, Λ, is applied to the realization of a Normal distribution $N(\mu, \sigma^2)$. See Figure 2 of Andersen, Harrison, Hole, Lau and Rutström (2012, p. 83). This distribution is highly bi-modal, and the area around the mean of the distribution has very low density. Thus, if a single stochastic specification is estimated on a sample from this population, the estimated parameters representing their distributional means give highly misleading information about the choice behavior we would expect from individual agents sampled from this population. In this case a mixture model of two models could potentially identify the modes, thus providing more, but still limited, information about the population. A similar approach is utilized by Conte, Hey

(3.11) provide some insight into the heterogeneity of a pooled sample, this is mostly limited to estimating average deviations from the mean due to observable heterogeneity. While it is theoretically possible to have a mixture model with greater than two underlying stochastic specifications, in reality this is computationally demanding and thus the mixture model presented in (3.11) is often only utilized with two mixtures.

Estimating parameter vectors for every subject in a sample helps to improve on this limitation, as can "random coefficients," discussed below. If every subject has an individually estimated parameter vector, then an analyst can use the distribution of these estimates to approximate the distribution of parameter vectors of the population from which this sample was drawn. However, the individually estimated parameters are still estimates, and thus they all have associated standard errors and positive probabilities of misidentification. The likelihood of misidentification typically decreases with the number of choice tasks presented to subjects, just as standard errors are negatively correlated with sample size. Hey and Orme (1994) estimate parameters for individual subjects utilizing 100 choice tasks per subject in order minimize the potential for misidentification. Hey (2001) utilized 500 choice problems per subject.

However, conducting experiments where subjects are required to give responses to a large number of tasks has practical problems, which then spill over and generate theoretical problems. Subjects can become bored or tired, which may make the tasks less salient or cause them to fail to satisfy the dominance criteria described by Smith (1982) and Harrison (1992). Often experimenters utilize a random lottery incentive mechanism (RLIM) in experiments, selecting one choice by the subject at

and Moffatt (2011).

random for payment. While in theory this is incentive compatible with EUT, it is not necessarily incentive compatible with any utility theory that doesn't require the independence axiom (IA), such as RDU (Cox, Sadiraj and Schmidt 2015; Harrison and Swarthout 2014). Furthermore, each additional choice task presented to the subject dilutes the expected outcomes of the other choice tasks. This means that the task could fail the dominance criteria unless the outcomes are sufficiently scaled up, even if the outcomes and the payment mechanism are salient. Thus, when the experimenter implements the RLIM for practical reasons, such as not needing to resolve and then compensate a subject for all of potentially hundreds of choices, he potentially introduces a serious theoretical concern.

These qualifications to estimation of individual parameter vectors should not be considered fatal for this method, but they should be noted when conducting this kind of estimation. Hey (2001) split the 500 choice tasks over 5 days to help mitigate the potential for subjects to become bored. Other experimenters split the T lottery tasks into smaller sets of tasks which are split by other, potentially unrelated, tasks. These kinds of designs help mitigate the procedural problems with such estimation, though sometimes they may introduce other concerns. While subjects may be less bored by doing choice tasks over 5 days rather than all on 1 day, subjects may experience events in between sessions that change their beliefs about the lottery pairs presented during the sessions.

An alternative method to recover greater information on entire samples of agents is to estimate the distributions of the parameter vectors describing individual preferences directly from pooled data.[7] Instead of estimating preference parameters,

[7]Andersen, Harrison, Hole, Lau and Rutström (2012) discuss the application of these well-known econometric methods to the estimation of standard models of risk (and time) preferences.

the parameters which shape the distributions of preferences in the population are estimated. This is often called a "random coefficients model". We can call equation (3.6), which is at the heart of equations (3.7) through (3.12), a conditional probability, because the probability is conditional on a particular β vector. We can however weight this function by the likelihood of observing the β vector from a given distribution.[8] We call this weighted probability the unconditional probability:

$$P_{it}(\theta) = \int P_{it}(\beta_i) f(\beta|\theta) d\beta \tag{3.13}$$

where $f(\beta|\theta)$ is the density function of the β vector given some vector of hyper-parameters θ shaping the distribution of the β.

This unconditional probability can be substituted for the conditional probability used in equations (3.7) and (3.8) to give us the unconditional likelihood equation:

$$L_{iT}(\theta) = \prod_{t}^{T} P_{it}(\theta) \tag{3.14}$$

and its counterpart, the unconditional log-likelihood equation:

$$LL_{iT}(\theta) = \sum_{t}^{T} \ln\left(P_{it}(\theta)\right) \tag{3.15}$$

Equations (3.13) through (3.15) are computationally impossible to estimate directly due to the general "inability of computers to perform integration" for non-trivial distributions in a closed-form (Train 2002, p. 2). However, equation

[8]It is worth noting the relation of these statements to a Bayesian approach. Having knowledge of the distribution of preferences in a population is akin to holding a prior in a Bayesian approach. This prior could then be incorporated to condition individual level estimates and produce an individual level choice probability. This Bayesian technique is different from the two approaches discussed here. The individual level approach discussed here does not incorporate a distributional prior in its estimation process, while the unconditional approach generates choice probabilities directly from pooled data, not individual data.

(3.13) can be approximated by simulation as follows:

$$SP_{it}(\theta) = \sum_{h}^{H} \frac{P_{it}(\beta^h)}{H} \tag{3.16}$$

Equation (3.16) needs some explanation. The integration involved in equation (3.13) is approximated by taking H random draws of β^h from the distribution governed by θ, evaluating equation (3.6) with each of these H randomly drawn β^h, and taking a simple average across these H evaluations. Only a simple average is needed because if the β^h vectors are drawn at random from the distribution governed by θ, then the likelihood of their occurrence is already weighted by the distribution's density.

The use of H as the term characterizing draws from a distribution is not arbitrary. It indicates that the random draws will often be approximated by a Halton sequence of numbers. The Halton routine is a numerical method to produce a sequence of numbers which efficiently approximate random draws from a uniform distribution bounded between 0 and 1, and which has been shown to provide better coverage of the distribution than other pseudo-random[9] number generators.[10]

The Halton sequence of uniformly distributed numbers can be transformed into a sequence of randomly drawn numbers from any invertible, univariate distribution.

[9]All "random" numbers generated by computers are in fact "pseudo-random" numbers produced algorithmically. Train (2002, p. 234) describes these numerical routines as follows: "The intent in [the] design [of pseudo-random routines] is to produce numbers that exhibit the properties of random draws. The extent to which this intent is realized depends, of course, on how one defines the properties of 'random' draws. These properties are difficult to define precisely since randomness is a theoretical concept that has no operational counterpart in the real world." Because of the non-existence of truly "random" number generators, the term "random" will be used in place of "pseudo-random" throughout this text.

[10]See the remainder of Train (2002, Chapter 9) for an in-depth discussion and derivation of why Halton sequences are widely viewed as being superior to many other pseudo-random number generators for the purposes of simulating estimators, at least when the dimensionality of the estimation problem is small.

That is, if μ is taken to be a random variable indicating a draw from a uniform distribution, and $F(\epsilon)$ is an invertible, univariate, cumulative distribution, then given μ, draws of ϵ from this distribution can be obtained by solving $\epsilon = F^{-1}(\mu)$. Train (2002, p. 236) discusses this method for obtaining random draws from invertible, univariate distributions, as well as using Choleski transformations to obtain draws from multivariate normal distributions.

With this simulated unconditional probability, we can obtain the simulated unconditional likelihood by substituting equation (3.16) for equation (3.13) in equation (3.14):

$$SL_{iT}(\theta) = \prod_{t}^{T} \left[\sum_{h}^{H} \frac{P_{it}(\beta^h)}{H} \right] \tag{3.17}$$

Equation (3.17) is limited in terms of identifying θ because, as indicated by the i subscript, this metric is defined for a single agent. Since the normatively coherent stochastic models discussed in Chapter 2 have non-random elements composing β_i, there is no distribution of β_i to be estimated from a single agent's choices. The real power of this method is realized, however, when sample data are pooled together and the distribution of β_i vectors is estimated from this pooled data. This is an easy extension of equation (3.17), which is logged for numerical reasons:

$$SLL_{NT}(\theta) = \sum_{i=1}^{N} \left(\sum_{t}^{T} \left[\ln \left(\sum_{h}^{H} \frac{P_{it}(\beta^h)}{H} \right) \right] \right) \tag{3.18}$$

We call equation (3.18) the unconditional simulated log-likelihood function, or just the simulated log-likelihood function (SLL). Maximum simulated likelihood (MSL) methods can be applied to this equation to return the MSL estimator $\hat{\theta}$ which maximizes this function. The characteristics of simulated estimators are reviewed in depth by Train (2002, Chapter 10), and the critical insight is that the estimator

$\hat{\theta}$ derived from equation (3.18) approaches the estimator from equation (3.15) with a sufficiently large, H, number of draws from the distribution governed by θ.

Estimating the distribution of preferences for a sample with MSL may improve the analyst's position over RA models with pooled data. The limitation of estimating only the conditional mean preference parameter for pooled data with standard MLE is no longer binding. Flexible distributions such as the Logit-Normal[11] can be employed to estimate higher moments of the distribution such as the variance, skewness and kurtosis. Additionally, the individual elements of θ can be modeled as linear functions of observable covariates, as was done in equation (3.10) for pooled MLE. This added flexibility allows the analyst to have greater information about preferences at the sample level, and can also be used to make characterizations of welfare at the individual level. On the other hand, MSL routines are computationally intensive, and become even more so when MSL mixture models are estimated.

3.2 The Holt and Laury (2002) MPL and the Unconditional Assessment of Expected Welfare

Each of the econometric methods detailed above provide some information about economics agents, either at the individual or collective level, and each have their own strengths and limitations. Issues concerning statistical power and identification for individual level estimation will be discussed in more depth in Chapter 4. In

[11] Andersen, Harrison, Hole, Lau and Rutström (2012, p. 82) utilize the Logit-Normal distribution because of its high degree of flexibility and because "MSL algorithms developed for univariate or multivariate Normal distributions can be applied directly." The figures they present (2012, p. 83) display some of the flexible forms this distribution can take.

the discussion below, I detail the usefulness of knowledge of the distributions of preferences in a population, given by θ, but not the methods and limitations of estimating $\hat{\theta}$ via MSL or some other estimation procedure. The results presented below are therefore numerical approximations of statistics for candidate values of θ, not estimates of statistics given an estimated $\hat{\theta}$. The formulae presented below for θ could be extended to incorporate standard errors associated with he elements of $\hat{\theta}$, but additional assumptions about the sampled population would need to be made. In this section I demonstrate that knowledge of the distribution of preferences in a population, given by θ, can provide useful information about the expected welfare of individual agents from this population for any given pattern of choices.

To make this discussion more concrete, we can utilize one of the HL-MPL instruments alluded to earlier and displayed in Table 3.1. In the HL experiment subjects were presented with this table, without the "Expected Payoff Difference" and "CRRA for Indifference" columns, and asked to select one option from each row. The "Option A" column indicates the outcomes and associated probabilities for option A in each of 10 tasks, and similarly for the "Option B" column. The "CRRA for Indifference" column indicates the CRRA value that would make an EUT agent indifferent between option A and option B. Thus, an agent with a CRRA value of 0.5 would theoretically select option A for rows 1-6, and then "switch" to selecting option B for the remaining rows.

The popularity of this approach is in part due to its straightforward logic: if a subject conforms to a deterministic EUT specification, then she should start off selecting option A, then at some point switch once, and only once, to selecting option B for the remaining rows or she should select B for every row. The point at which the subject switches reveals an interval in which preference for risk must lie,

Table 3.1: The Ten Paired Lottery-Choice Decisions with Low Payoffs
Holt and Laury (2002, p. 1645)

Row #	Option A	Option B	Expected Value Difference	CRRA for Indifference
1	1/10 of \$2.00 , 9/10 of \$1.60	1/10 of \$3.85 , 9/10 of \$0.10	\$1.17	-1.7134
2	2/10 of \$2.00 , 8/10 of \$1.60	2/10 of \$3.85 , 8/10 of \$0.10	\$0.83	-0.9468
3	3/10 of \$2.00 , 7/10 of \$1.60	3/10 of \$3.85 , 7/10 of \$0.10	\$0.50	-0.4866
4	4/10 of \$2.00 , 6/10 of \$1.60	4/10 of \$3.85 , 6/10 of \$0.10	\$0.16	-0.1426
5	5/10 of \$2.00 , 5/10 of \$1.60	5/10 of \$3.85 , 5/10 of \$0.10	-\$0.18	0.1464
6	6/10 of \$2.00 , 4/10 of \$1.60	6/10 of \$3.85 , 4/10 of \$0.10	-\$0.51	0.4115
7	7/10 of \$2.00 , 3/10 of \$1.60	7/10 of \$3.85 , 3/10 of \$0.10	-\$0.85	0.6762
8	8/10 of \$2.00 , 2/10 of \$1.60	8/10 of \$3.85 , 2/10 of \$0.10	-\$1.18	0.9706
9	9/10 of \$2.00 , 1/10 of \$1.60	9/10 of \$3.85 , 1/10 of \$0.10	-\$1.52	1.3684
10	10/10 of \$2.00 , 0/10 of \$1.60	10/10 of \$3.85 , 0/10 of \$0.10	-\$1.85	N/A

at least under EUT.

However, this pattern need not necessarily occur given stochastic specifications. Subjects may, and sometimes do, switch multiple times between option A and option B as they work their way down the rows. Some subjects even select option A in row 10, despite it being dominated by option B. The first of these observed choice behaviors is often referred to as multiple switching behavior (MSB), while the second is a form of FOSD since there is no risk involved in row 10. Holt and Laury (2002, p. 1647) observe that 28 of their 212 subjects exhibited MSB. Rather than discussing all of the potential reasons why a subject would exhibit MSB, we will assume a normatively coherent stochastic model and discuss the implications of MSB within it.

The HL-MPL instrument is a useful instrument to discuss the welfare implications of stochastic models because the observed MSB is an apparent violation of EUT that is easy to notice visually without estimation. As discussed in Chapter 2, when utilizing normatively coherent stochastic models, observed violations of

EUT necessarily imply that some welfare has been forgone by the agent because of the violation according to EUT. In Savage's terms, one would have something to "reproach" oneself for by violating the theory (Moscati 2016, p. 230). Since there is no deterministic EUT utility function which allows either the switching back and forth from option A to option B or the selection of a guaranteed, lower outcome over a guaranteed, higher outcome, it must be the case that the observance of MSB implies that some welfare has been forgone, at least under an EUT framework.

An important and often overlooked reality of stochastic models is that even if a subject doesn't display MSB, the subject may still have made choice errors and therefore have foregone some amount of welfare. This may not seem obvious at first, since any non-MSB choice pattern can be rationalized by some preference relation. Cases such as these arise when a subject makes a choice error with respect to the utility function they employ, but this choice error results in a choice pattern that is still rationalizable, or "consistent." When we incorporate knowledge of a sample's distribution of preferences governed by θ, we can see that many observed, apparently "consistent" choice patterns contain more choice errors and are often more costly in terms of foregone welfare than apparently "inconsistent" choice patterns. This will be made clear in the discussion below, but first we must define some notation.

Utilizing notation from the beginning of Section 3.1, an option in a set of alternatives t is represented as X_{at}, where a indicates the option's ordinal rank among the set of alternatives given the agent's utility parameter vector, β_i, and $y_t = a$ indicates that option a was chosen by the agent in task t. We can define a "choice error" as any choice where the option chosen was not ordinally ranked the highest among the set of alternatives with respect to the agent's preferences,

given by β_i. Therefore a choice error in task t is when $y_t \neq 1$ (recall the description of subscripts given for equation (3.5)), and an indicator function for choice errors given some vector of assumed utility parameters β_i is given by:

$$K_t(\beta_i) = \begin{cases} 1 & y_t \neq 1 \\ \\ 0 & y_t = 1 \end{cases} \tag{3.19}$$

The frequency of choice errors by agent i in the choice pattern $y_t \times T$ is:

$$M_T(\beta_i) = \sum_{t}^{T} K(\beta_i) \tag{3.20}$$

Given the distribution parameter vector θ, we can define the expected frequency of choice errors in the choice pattern $y_t \times T$ as:

$$\mathrm{E}(M|\theta) = \int M(\beta_i) f(\beta|\theta) d\beta \tag{3.21}$$

where, just as in equation (3.13), $f(\beta|\theta)$ is the density function of the β vector given the vector of hyper-parameters θ shaping the distribution of the β. Equation (3.21) is just the mean of the discrete distribution of choice errors in the choice pattern $y_t \times T$, given the distribution parameter vector θ. Because the distribution of choice errors is discrete, $M(\beta_i) \in [0, T] \subset \mathbb{N}^0$, we can define the probability mass function of choice errors as follows:[12]

$$P_E(e|\theta) = \int N[M(\beta), e] f(\beta|\theta) d\beta \tag{3.22}$$

[12] $\mathbb{N}^0$ indicates the set of natural numbers, inclusive of 0. $\mathbb{N}^1$ or $\mathbb{N}^+$ would indicate the set of natural numbers not inclusive of 0.

where

$$N[M(\beta), e] = \begin{cases} 1 & M(\beta) = e \\ 0 & M(\beta) \neq e \end{cases} \tag{3.23}$$

and e indicates the number of choice errors for the given choice pattern and θ vector. Equation (3.22) provides useful information about whether an observed pattern deviates from a deterministic choice model, but is limited since it assigns equal weight to errors which are very costly in terms of welfare and errors that are not so costly.

We can incorporate two of the metrics developed in Chapter 2 for welfare assessment into this sample framework. The first metric, calculated for a choice pattern $y_t \times T$, is equivalent to a standard consumer surplus calculation:

$$\Delta W_{iT} = \sum_{t=1}^{T} \left(CE_{iyt} - CE_{i1t}^{Z} \right) \tag{3.24}$$

where CE_{iyt} is the CE of the option chosen, indicated by the subscript y, by agent i in task t, and CE_{i1t}^{Z} is the CE of the option that is ordinally ranked the highest among the set of *unchosen* alternatives, Z, with respect to the agent's preferences, given by β_i, in task t. Throughout this chapter, we will refer to the metric in equation (3.24) as the "welfare surplus" metric. The second metric we propose to characterize the welfare implications of choices is similar to the concept of auction and market "efficiency" proposed by Plott and Smith (1978):

$$\%W_{iT} = \frac{\sum_{t=1}^{T} CE_{iyt}}{\sum_{t=1}^{T} CE_{i1t}} \tag{3.25}$$

In the metric defined in equation (3.25), the CE's of the options chosen by the

agent across all tasks T are summed, and then divided by the CE's of the options that were ordinally ranked the highest with respect to the agent's preferences across all the tasks. Therefore, should an agent never make a choice error, this metric would take on the value of 1, and should the agent make at least one choice error, it would take on a value between 0 and 1.[13] Throughout this chapter, we will refer to the metric in equation (3.25) as the "welfare efficiency" metric.

The welfare surplus and welfare efficiency metrics from equations (3.24) and (3.25) can be used in place of equation (3.20) in equation (3.21) to gather useful information for a given choice pattern and θ vector:

$$E(\Delta W_T|\theta) = \int \Delta W_T(\beta) f(\beta|\theta) d\beta \tag{3.26}$$

$$E(\%W_T|\theta) = \int \%W_T(\beta) f(\beta|\theta) d\beta \tag{3.27}$$

Given equation (3.23), we can denote the expected welfare surplus and the expected welfare efficiency obtained by agents who have committed $e \in [0, T]$ errors by making choices $y_t \times T$ as follows:

$$E(\Delta W_T|\theta, e) = \int \Big(\Delta W_T(\beta) \times N[M(\beta), e]\Big) f(\beta|\theta) d\beta \tag{3.28}$$

$$E(\%W_T|\theta, e) = \int \Big(\%W_T(\beta) \times N[M(\beta), e]\Big) f(\beta|\theta) d\beta \tag{3.29}$$

The same limitation mentioned about MSL concerning a computer's inability to perform closed-form integration in general applies to equations (3.21), (3.22), and (3.26) through (3.27). However, these equations can be approximated in the

[13] There are a few mathematical peculiarities with this metric. This metric can lose its $(0, 1)$ bounds if any of the T tasks has a mixed frame, that is, a task that has both positive and negative outcomes. This metric would become negative if the CE of a chosen option is negative and the CE of the highest ranked option is positive. Also, this metric becomes undefined if the CE of the highest ranked alternative is 0. These general issues will not be of concern in this chapter because all examples of lotteries have outcomes in the strictly positive domain.

manner described for MSL in equation (3.16): the terms in these equations between the integrand and the density function will be evaluated with β vectors randomly drawn H times from the distribution governed by θ, and then averaged. As H gets sufficiently large, the simulated statistics approach the true statistics.

3.2.1 Sample Level Analysis with an EUT Population

The simulation methods described here and for the remainder of this chapter characterize an individual agent as having a single β_i vector representing her preferences, and making choices in an economic environment that satisfies the Smith (1982) precepts for valid economic experiments. An individual agent i generates an *observed* choice pattern $y_t \times T$ by resolving the stochastic process defined by her preferences. In Chapter 2 we described normatively coherent stochastic models as those models that characterize agents as having non-random preferences, thus an agent's preferences do not change from choice to choice.[14] Individual β_i parameter vectors are themselves drawn from a population of β vectors. This distribution of β vectors in the population is characterized by the parameter vector θ. Throughout the following discussion, we will refer to a choice pattern's likelihood of being *observed*, by which we mean the choice pattern's simulated likelihood as calculated in equation (3.17). This is the probability of observing a choice pattern given that it has been generated by an agent randomly drawn from the population defined by θ. Likewise, when we discuss the expected welfare implications of a choice pattern

[14]Not only are we assuming that agents do not have random preferences, we're also assuming that an agent's preferences are the same across choices generally. We could, as Hey (2001) does, model some or all of the parameters in an agent's utility function as being partly determined by the number of choices that the agent has encountered. Because preferences modeled in this way change from choice to choice in a non-random manner, the welfare analysis discussed in this chapter could be extended in a normatively coherent manner to incorporate this "learning," potentially with interesting implications. This would involve specifying additional marginal distributions which characterize the parameters defining the "learning" process.

for a given population, we are discussing the expected welfare implications for an agent from that population who generated that choice pattern.

To construct an explicit numerical example, we first define the models characterizing an individual agent's choice probabilities, and then the marginal distributions of the elements of β which together define the population characterized by θ. For the sake of simplicity, we first consider a population entirely composed of agents conforming to an EUT utility model with a Contextual Utility (CU) stochastic model due to Wilcox (2008). Thus, choice probabilities for an individual agent are defined as follows:

$$\begin{aligned} P_{iat}(\beta_i) &= Pr\left(\epsilon_t \geq \frac{1}{D(\beta_i, X_t)\lambda_i}\left[G(\beta_i, X_{kt}) - G(\beta_i, X_{jt})\right]\right) \\ &= 1 - F\left(\frac{G(\beta_i, X_{kt}) - G(\beta_i, X_{jt})}{D(\beta_i, X_t)\lambda_i}\right) \\ &= Pr(y_t = a) \end{aligned} \tag{3.30}$$

where ϵ_t defines the random error associated with the measurement of utility, the functional form of the utility function, $G(\cdot)$, is the CRRA function of the form $u(x) = \frac{x^{1-r}}{(1-r)}$, $F(\cdot)$ is the logistic cumulative distribution function (cdf), and the adjusting function $D(\cdot)$ is as follows:

$$\begin{aligned} &D(\beta_i, X_t) = max[u(x_{it})] - min[u(x_{it})] \\ &st.\ w_i(x_{it}) \neq 0 \end{aligned} \tag{3.31}$$

Thus the β_i vector for each agent is said to consist of only two parameters, r and λ. The joint distribution of these two parameters characterizes the population of agents and is characterized by the parameter vector θ. We assume the marginal distributions of the r and λ parameters to be independent and uncorrelated in

the population.[15] The r parameter can conceivably take any value, but to make the bulk of the density lie in the familiar range of the literature employing the HL-MPL instrument, we assume it to be distributed normal, with mean of 0.65 and a standard deviation of 0.3, thus $r \sim \mathcal{N}(0.65, 0.3^2)$. The λ parameter must be strictly positive, so it will be assumed to be distributed as gamma with a mean of 0.35 and a standard deviation of 0.3. This is equivalent to a gamma distribution with a shape parameter of $k \approx 1.36$ and a scale parameter of $t \approx 0.26$, thus $\lambda \sim \Gamma(1.36, 0.26)$. Together these 4 parameters make the joint distribution-shaping parameter $\theta = \{0.65, 0.3^2, 1.36, 0.26\}$.

The metrics described in equations (3.17) and (3.20) through (3.29) rely on a given choice pattern, $y_t \times T$. In the HL-MPL instrument there are a total of $2^{10} = 1024$ choice patterns that can be observed. To begin the discussion of the welfare implications of stochastic choice models, we calculate the values for equations (3.17), and (3.20) through (3.29) for all $TT = 1024$ choice patterns and all $e \in [0, T]$ for the given θ, with $H = 2.5 \times 10^6$.

To make clear how the result of these equations are arrived at, we can work through the calculations step by step. First, we select a choice pattern from one of the 1024 choice patterns possible with the HL-MPL instrument, for example, the choice of option A for the first five rows and option B for rows 6 through 10. Next a β_i vector is drawn from the joint distribution defined by θ. As an example, we assume $\beta_i = \{r = 0.65, \lambda = .35\}$ was drawn; recall that we are also assuming EUT with a CU stochastic model. Utilizing this β_i and choice pattern, we can evaluate

[15] This is done for convenience; adding correlation among the marginal distributions would require the specification of a covariance matrix. In samples of real populations, we might expect there to be correlation among these marginal distributions, and this analysis can be easily extended to accommodate it. This additional step is not difficult, but introduces more parameters to keep track of and doesn't significantly add to the narrative.

the various metrics proposed. First, we evaluate equation (3.7), the likelihood that agent i would produce this choice pattern, utilizing equation (3.30) to calculate choice probabilities for the individual tasks:

$$\begin{aligned}
P_{i,a,1} &= Pr(y_1 = A \,|\, \beta_i) = 0.82 \\
P_{i,a,2} &= Pr(y_2 = A \,|\, \beta_i) = 0.78 \\
P_{i,a,3} &= Pr(y_3 = A \,|\, \beta_i) = 0.74 \\
P_{i,a,4} &= Pr(y_4 = A \,|\, \beta_i) = 0.68 \\
P_{i,a,5} &= Pr(y_5 = A \,|\, \beta_i) = 0.62 \\
P_{i,a,6} &= Pr(y_6 = B \,|\, \beta_i) = 0.44 \\
P_{i,a,7} &= Pr(y_7 = B \,|\, \beta_i) = 0.51 \\
P_{i,a,8} &= Pr(y_8 = B \,|\, \beta_i) = 0.57 \\
P_{i,a,9} &= Pr(y_9 = B \,|\, \beta_i) = 0.63 \\
P_{i,a,10} &= Pr(y_{10} = B \,|\, \beta_i) = 0.95 \\
P_{iT} &= \prod_{t=1}^{T=10} P_{iat}(\beta_i) = 0.0154
\end{aligned} \tag{3.32}$$

Note that $P_{i,a,6} = 0.44 < 0.50$. With the CU stochastic model, given in equation (3.30), the option with the greatest utility, expected or otherwise, will always have the greatest probability of being chosen. Since there are only two alternatives in row #6, and the choice probability of option B is less than that of option A, it must be the case that option B in this row had a lower expected utility than option A. If B has a lower expected utility than A, the choice of B in row #6 is a choice error. Using the notation defined in Section 3.1, $y_6 = 2$, and for all $t \in \{T \setminus 6\}$, $y_t = 1$. This information allows us to evaluate equation (3.20), the frequency of

choice errors in a given choice pattern, utilizing equation (3.19):

$$
\begin{aligned}
P_{i,a,1} = Pr(y_1 = A \,|\, \beta_i) = 0.82 \;\Rightarrow\; K_1(\beta_i) = 0 \\
P_{i,a,2} = Pr(y_2 = A \,|\, \beta_i) = 0.78 \;\Rightarrow\; K_2(\beta_i) = 0 \\
P_{i,a,3} = Pr(y_3 = A \,|\, \beta_i) = 0.74 \;\Rightarrow\; K_3(\beta_i) = 0 \\
P_{i,a,4} = Pr(y_4 = A \,|\, \beta_i) = 0.68 \;\Rightarrow\; K_4(\beta_i) = 0 \\
P_{i,a,5} = Pr(y_5 = A \,|\, \beta_i) = 0.62 \;\Rightarrow\; K_5(\beta_i) = 0 \\
P_{i,a,6} = Pr(y_6 = B \,|\, \beta_i) = 0.44 \;\Rightarrow\; K_6(\beta_i) = 1 \\
P_{i,a,7} = Pr(y_7 = B \,|\, \beta_i) = 0.51 \;\Rightarrow\; K_7(\beta_i) = 0 \\
P_{i,a,8} = Pr(y_8 = B \,|\, \beta_i) = 0.57 \;\Rightarrow\; K_8(\beta_i) = 0 \\
P_{i,a,9} = Pr(y_9 = B \,|\, \beta_i) = 0.63 \;\Rightarrow\; K_9(\beta_i) = 0 \\
P_{i,a,10} = Pr(y_{10} = B \,|\, \beta_i) = 0.95 \;\Rightarrow\; K_{10}(\beta_i) = 0 \\
M(\beta_i) = \sum_{t=1}^{T=10} K_t(\beta_i) = 1
\end{aligned}
\tag{3.33}
$$

Thus we see that our subject i has committed one choice error across the T tasks, the choice error in row #6. This result allows us to calculate equation (3.23) for values of $e \in [0, T]$. Equation (3.23) is just an indicator function that signals if the frequency of choice errors in the choice pattern, given by equation (3.19), is equal to some scalar e. Thus, if we know the result of equation (3.19), which for the example shown above is 1, we know that the result of equation (3.23) will be 1

for $e = 1$, and 0 for all other values of e.

$$
\begin{aligned}
N(M_T(\beta_i) = 1, e = 0) &= 0 \\
N(M_T(\beta_i) = 1, e = 1) &= 1 \\
N(M_T(\beta_i) = 1, e = 2) &= 0 \\
N(M_T(\beta_i) = 1, e = 3) &= 0 \\
N(M_T(\beta_i) = 1, e = 4) &= 0 \\
N(M_T(\beta_i) = 1, e = 5) &= 0 \\
N(M_T(\beta_i) = 1, e = 6) &= 0 \\
N(M_T(\beta_i) = 1, e = 7) &= 0 \\
N(M_T(\beta_i) = 1, e = 8) &= 0 \\
N(M_T(\beta_i) = 1, e = 9) &= 0 \\
N(M_T(\beta_i) = 1, e = 10) &= 0
\end{aligned}
\tag{3.34}
$$

Next we can calculate the two welfare metrics from equations equations (3.24) and (3.25), indicating welfare surplus and welfare efficiency respectively. First we calculate the *CE* of option A and option B for all $T = 10$ tasks using equation (3.5). We note the *CE* of the chosen and unchosen options for the given choice pattern, the difference between the two, and the greatest *CE* of the two options for each task:

With the *CE*'s calculated, we can substitute them in to equations (3.24) and (3.25). For equation equation (3.24), we take the sum of column 6 in Table 3.2:

$$\Delta W_{iT} = \sum_{t=1}^{T} \left(CE_{iyt} - CE_{i1t}^{Z} \right) = 8.92 \tag{3.35}$$

and for equation (3.25), we take the sum of column 4 and divide it by the sum of

Table 3.2: Example *CE*'s of EUT Agent with HL-MPL

Task	*CE* of A	*CE* of B	*CE* of Chosen Option	*CE* of Unchosen Option	*CE* of Chosen - *CE* of Unchosen	Greatest *CE*
1	1.64	0.19	1.64	0.19	1.44	1.64
2	1.68	0.33	1.68	0.33	1.35	1.68
3	1.71	0.52	1.71	0.52	1.20	1.71
4	1.75	0.76	1.75	0.76	0.99	1.75
5	1.79	1.07	1.79	1.07	0.72	1.79
6	1.83	1.46	1.46	1.83	−0.38	1.83
7	1.87	1.92	1.92	1.87	0.04	1.92
8	1.92	2.47	2.47	1.92	0.55	2.47
9	1.96	3.11	3.11	1.96	1.15	3.11
10	2.00	3.85	3.85	2.00	1.85	3.85

column 7:

$$\%W_{iT} = \frac{\sum_{t=1}^{T} CE_{iyt}}{\sum_{t=1}^{T} CE_{i1t}} = \frac{21.37}{21.74} = .983 \tag{3.36}$$

Finally, we multiply the welfare metrics derived in equations (3.35) and (3.36) by the indicator functions derived for $e \in [0, T]$ in equation (3.23):

$$\begin{aligned} N(M_T(\beta_i) = 1, e = 1) \times \Delta W_{iT} &= 1 \times 8.92 = 8.92 \\ N(M_T(\beta_i) = 1, e \neq 1) \times \Delta W_{iT} &= 0 \times 8.92 = 0 \end{aligned} \tag{3.37}$$

$$\begin{aligned} N(M_T(\beta_i) = 1, e = 1) \times \%W_{iT} &= 1 \times .983 = .983 \\ N(M_T(\beta_i) = 1, e \neq 1) \times \%W_{iT} &= 0 \times .983 = 0 \end{aligned} \tag{3.38}$$

The indicator functions in equation (3.34) are mutually exclusive, therefore the product of the indicator functions and the welfare metrics will be 0 for all but one value of e, and equal to the welfare metrics for the remaining e, in this example,

for $e = 1$.

Having derived the results of these equations for one given choice pattern, we iterate through the remaining 1023 choice patterns for this particular agent, repeating the numerical exercise described above for each choice pattern. With metrics defined for this particular agent across all $TT = 1024$ possible choice patterns, a new β_i vector is drawn from θ, and the entire process repeated. For the calculations described below, we repeat the process of drawing a β_i from θ and calculating the results of these metrics for all choice patterns $S = 2.5 \times 10^6$ times. This process results in a 3 dimensional array with $(\#\ of\ metrics \times \#\ of\ choice\ patterns \times S) = 33 \times 1024 \times (2.5 \times 10^6) = 8.448 \times 10^{10}$ elements.

To arrive at the population level metrics, we take the average of each metric defined in equations (3.32) through (3.38) across all S simulated agents for each choice pattern. Since each β_i was drawn randomly from the distribution governed by θ, only a simple average is needed. This averaging leaves us with a dataset that has $33 \times 1024 = 33,792$ elements.

This resulting dataset, however, is too large to be usefully displayed in full, so for now we restrict attention to the 10 choice patterns most likely to be observed in a population governed by θ, and discuss the metrics calculated in equations (3.21), (3.26), (3.27), and (3.22) with $e = (0, 1)$. The results of these equations for the 10 most likely choice patterns to be observed in a population governed by θ are given in Table 3.3.

For the "Choice in Row" column in Table 3.3, 0 indicates a choice of A for the row, and 1 indicates a choice of B. Note that the choice pattern that is mostly likely to be observed from a sample drawn from the specified population governed by θ, shown in the first row where *Rank* is 1, is the choice pattern we would observe from

Table 3.3: HL-MPL Welfare and Error Expectations for Choice Patterns with Top Ten Simulated Likelihoods, EUT

	Choice in Row										Simulated Likelihood	Expected Errors	Welfare Efficiency	Welfare Surplus	$P_E(e=0)$	$P_E(e=1)$
Rank	1	2	3	4	5	6	7	8	9	10						
1	0	0	0	0	0	0	1	1	1	1	0.0360	0.880	0.9861	9.33	0.322	0.489
2	0	0	0	0	0	0	0	1	1	1	0.0341	0.950	0.9859	9.32	0.323	0.456
3	0	0	0	0	0	1	1	1	1	1	0.0244	1.454	0.9658	8.56	0.167	0.364
4	0	0	0	0	0	1	0	1	1	1	0.0231	1.523	0.9656	8.55	0	0.489
5	0	0	0	0	0	0	1	0	1	1	0.0219	1.595	0.9663	8.36	0	0.456
6	0	0	0	0	0	0	0	0	1	1	0.0207	1.665	0.9661	8.35	0.134	0.331
7	0	0	0	0	1	0	1	1	1	1	0.0166	1.787	0.9474	7.93	0	0.364
8	0	0	0	0	1	0	0	1	1	1	0.0158	1.857	0.9472	7.92	0	0.323
9	0	0	0	0	0	0	1	1	0	1	0.0150	1.864	0.9506	7.16	0	0.322
10	0	0	0	0	0	1	1	0	1	1	0.0148	2.168	0.9460	7.59	0	0.167

an agent described by a deterministic choice process with preferences at the mean of the distribution of r. The next two most likely choice patterns, where *Rank* is 2 and 3, correspond to the choice pattern we would observe from agents described by a deterministic choice process with preferences one standard deviation either side of the mean of the distribution of r.

Interestingly, for each of the three most likely choice patterns, it is far more likely than not that an agent displaying these choice patterns made at least one choice error, and thus did not obtain maximal welfare from her choices. This is shown by the values in column $P_E(e=0)$, which reference equation (3.22), all being less than 0.50, indicating that less than 50% of subjects will commit 0 choice errors for these choice patterns. Note that only 32.2% of agents who display the most likely choice pattern in row 1 are expected to have *not* made any choice errors and therefore obtain maximal welfare. This is despite the fact that any of these choice patterns can be rationalized by some set of preferences for our assumed model.

These patterns do, however, produce relatively high expected welfare efficiency and surplus. The welfare surplus metric is less informative in this comparison: it is more useful in making absolute rather than relative statements about welfare.

The relatively large values of $1 - P_E(e = 0)$, which imply that most choice patterns contain at least 1 choice error, is mainly due to the shape and location of the distribution of r. The mean of r, at 0.65, lies just next to the indifference boundary between rows 6 and 7 of the HL-MPL instrument, as indicated in the column "CRRA for Indifference" of Table 3.1. That means that the bulk of the r values drawn from this distribution define utility values that indicate near indifference between the A and B lotteries in row 7 of the HL-MPL instrument. All RE models increase the probability of a choice error the closer an agent is to being indifferent between 2 options, so it should not be a surprise that with this particular choice of distribution for r we have a large proportion of choice errors. Similarly, since many agents in this population would be nearly indifferent between the A and B lotteries in row 7 of the HL-MPL, the expected cost of the choice errors in row 1 in welfare terms is realtively low, as can be seen by the value in the "Welfare Efficiency" being very close to 1.

The fourth and fifth most likely choice patterns in Table 3.3, where *Rank* is 4 and 5, are not consistent with any deterministic EUT preferences. These patterns display what we will call "Light MSB": not including the choice made in row 10, the agent has "switched" between choosing A and B three times.[16] Because MSB is not consistent with any deterministic EUT preferences, $P_E(e = 0) = 0$ for these patterns. In fact, the only choice patterns in which $P_E(e = 0) > 0$ will be those

[16]The reason that row 10 is not included in this definition is because we are making a distinction between patterns which do and do not include a choice of A in row 10 later.

which are "Consistent": displaying a choice pattern that can be rationalized by some deterministic EUT preferences.

Despite the patterns in rows 4 and 5 of Table 3.3 being obviously inconsistent with a deterministic EUT process, they both are more likely to be observed from agents drawn from a population defined by θ, and obtain greater welfare surplus than the sixth most likely choice pattern which is "Consistent." The likelihood of the "Light MSB" choice patterns in rows 4 and 5, displayed in the "Simulated Likelihood" column, are greater than the likelihood of the choice pattern in row 6, which is consistent. The welfare efficiency metric for row 5, displayed in the "Welfare Efficiency" column, is greater than that of row 6, and the welfare surplus metrics for both rows 4 and 5 are greater than for row 6. Since metrics for all $TT = 1024$ choice patterns were calculated, we will see in the discussion below that these two Light MSB patterns are both more likely to be observed and to be less costly in terms of welfare surplus than 6 out of 10 "Consistent" patterns.

Another interesting aspect of this analysis is the correlation of welfare and the likelihood of observing a choice pattern. The correlation between the simulated likelihood of the choice patterns and their expected welfare efficiency is 0.62 across the whole dataset, while the simulated likelihood and expect welfare surplus has a correlation of 0.68. These are positive but far from 1. That is, as the likelihood of observing a choice pattern increases, the expected welfare efficiency and surplus of the choice pattern generally increases as well, but not always. This is apparent in rows 8 and 9 of Table 3.3. The choice pattern described in row 8 is more likely to be observed than the pattern in row 9, but the pattern in row 9 has a higher expected welfare efficiency than row 8, though not by much. The very large number of draws employed in these calculations rules out the possibility that this is a statistical

fluke caused by the random way these statistics were calculated.

This example illustrates how stochastic models are not "welfare ranking" models, but instead incorporate aspects of the choice process that are considered to be normatively desirable, while maintaining descriptive power. The example of row 8 and 9 only depicts the most common occurrence where the expected welfare efficiency of a pattern and its likelihood diverge in this hypothetical population. The most drastic divergence occurs between the patterns which have violated FOSD by selecting option A in row 10, and those that have not.

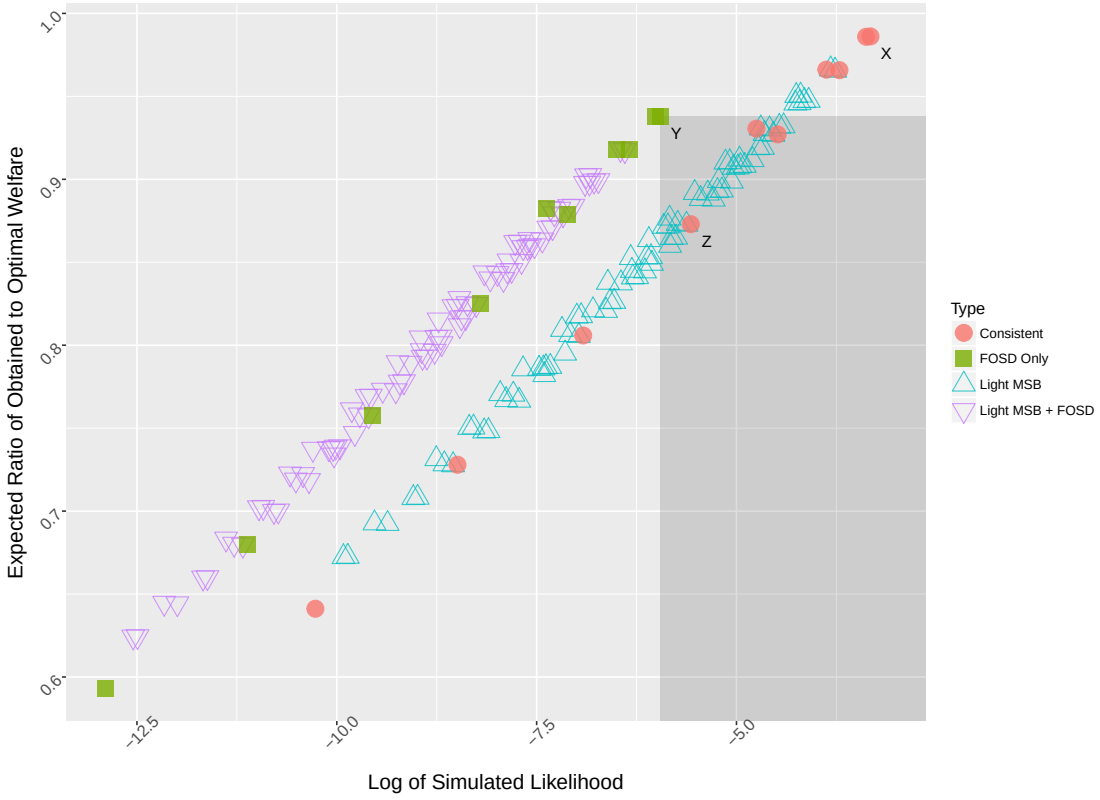
To make this distinction clear, Figure 3.1 plots the log of the SL (SLL) against the expected welfare efficiency of the choice patterns that:

- are Consistent with deterministic EUT,
- are Consistent other than the choice of A in row 10 (FOSD Only),
- display Light MSB, the agent has "switched" between choosing A and B three times, with a choice of B in row 10,
- display Light MSB with a choice of A in row 10 (Light MSB + FOSD).

In Figure 3.1 each point represents a unique choice pattern. For any given point plotted, any other point to the Southeast of that point indicates a pattern that *is both more likely to be produced by an agent drawn randomly from this population and provides lower expected welfare efficiency.* For instance, any point in the shaded region of Figure 3.1 represents a choice pattern that is both more likely to be observed and has a lower expected welfare efficiency than pattern Y.

Figure 3.1 shows that the choice of A in row 10 greatly decreases the SLL of the pattern, but barely decreases the expected ratio of obtained welfare to maximal welfare, all else being equal. For example, the most likely consistent choice pattern

Figure 3.1: Consistent and Light MSB, With and Without Row 10 Error



is the top right-most red dot in Figure 3.1, labeled "X," which corresponds to row 1 of Table 2 and has a welfare efficiency of 0.986 and a SL of 0.036. The most likely choice pattern with a choice of A in row 10 is the top right-most green dot, labeled "Y." This pattern is identical to the "X" pattern other than the selection of A in row 10 and has a welfare efficiency of 0.938 and a SL of 0.00246. The ratio of welfare obtained to maximum welfare differs only by 0.0483, but pattern X is about 14.63 times more likely to be observed than pattern Y.[17] The seventh most likely consistent pattern, not displayed in Table 1, can be seen as the red dot in Figure 3.1 labeled "Z" and is about 1.55 times more likely to be observed than

[17]Probability of X ÷ Probability of Y = 0.036 ÷ 0.00246 = 14.63

pattern Y and has an expected welfare efficiency that is about 0.065 *lower* than pattern Y.

The general implication of this exercise is to make it clear that stochastic models do not reliably link the likelihood of a choice pattern with its realized welfare as consumer surplus or efficiency. This is due to the way in which heteroscedastic RE models disproportionately "punish" FOSD in welfare terms by assigning occurrences of it a very low likelihood. The choice of A in row 10 is punished in welfare terms even more by the fact that there is no risk involved.

Empirically, experimental economists rarely observe behavior such as the choice of A in row 10 because the agents they study are in environments that incentivize them to reject dominated offers. There is, by definition, no extra benefit to actively choosing a dominated offer. But just because there is no extra benefit to be had, it shouldn't be inferred that the agent doesn't value the dominated option positively. Any agent who selected option A in row 10 still receives a $2 benefit from having had the choice problem presented to her if that choice is selected for payment.

There is a disconnect between how stochastic models map welfare and probability when considering individual choices versus patterns of choice. When considering an individual choice RE models create a perfect mapping of welfare and probability; an option that is more likely to be chosen from a set of alternatives than another option always also provides greater welfare. However, as we can see from Figure 3.1 and Table 3.3, when considering patterns of choice, this mapping breaks down, and it is no longer the case that a more likely pattern of choices also provides greater welfare.

3.2.2 Sample Level Analysis with a Mixed EUT-RDU Population

The above discussion focuses on a population that is entirely composed of EUT conforming agents. Individual level estimates from Hey and Orme (1994) and the mixture model estimates from Harrison and Rutström (2009) show that many populations are likely not composed entirely of EUT agents. We can extend the example above, defining the population as being composed of some mixture of EUT agents and RDU agents. By "mixture," we mean that there will be two subpopulations of a grand population, with agents employing either the EUT or RDU functions.

Before beginning the analysis of this mixture population, we can extend the metrics utilized in equations (3.17) and (3.20) through (3.29) to be defined for mixed populations. This is implemented in much the same way as mixture models were defined in equation (3.11); each metric, Q_m, for subpopulation m is weighted by the proportion of the subpopulation in the grand population, M.

$$\begin{aligned} \boldsymbol{Q^M} &= \sum_m^M \pi_m \times Q^m \\ st. \sum_m^M \pi_m &= 1 \end{aligned} \tag{3.39}$$

where π_m is the proportion of subpopulation m in the grand population. For example, the probability of observing any given choice pattern $y \times T$ for a grand population made of M subpopulations is:

$$\begin{aligned} \boldsymbol{L_{iT}^M} &= \sum_m^M \pi_m \times L_{iT}^m(\theta^m) \\ st. \sum_m^M \pi_m &= 1 \end{aligned} \tag{3.40}$$

where L_{iT}^{m} is as described in equation (3.14) for some subpopulation m defined by θ^{m}.

A final metric before considering the example of the mixed population is the probability that any given choice pattern was produced by population m. Utilizing equation (3.40) we define the probability that a pattern was produced by population m as the ratio of the weighted simulated likelihood of observing the pattern from subpopulation m to the likelihood of observing the pattern in the grand population:

$$Prop_{T}^{m} = \frac{\pi_{m} \times L_{iT}^{m}(\theta^{m})}{\boldsymbol{L_{iT}^{M}}} \tag{3.41}$$

With this mixing framework in mind, we can define our grand population. We assume that 70% of agents in the grand population conform to EUT, while the remaining 30% conform to RDU. Given that the previous example thoroughly examined an EUT population, rather than duplicate the analysis, we assume that the EUT subpopulation is the same as the previous EUT-only example. Thus, the EUT subpopulation is defined as using a CU stochastic model and CRRA function with the r parameter normally distributed $r \sim \mathcal{N}(0.65, 0.3^{2})$ and the λ parameter following a gamma distribution $\lambda \sim \Gamma(1.36, 0.26)$. This results in $\theta^{EUT} = \{0.65, 0.3^{2}, 1.36, 0.26\}$.

For the RDU subpopulation, we employ the flexible 2 parameter decision weighting function defined by Prelec (1998) as the probability weighting function to be substituted into equation (3.4):

$$\omega(p) = \exp(-\beta(-\ln(p))^{\alpha}) \tag{3.42}$$

where $\alpha > 0$ and $\beta > 0$. We continue to use the CRRA utility function and CU

stochastic model for the RDU subpopulation.

The r parameter is assumed to be distributed identically to the r parameter in the EUT population $r \sim \mathcal{N}(0.65, 0.3^2)$, and the λ parameter still uses the gamma distribution, but is distributed $\lambda \sim \Gamma(0.563, 0.26)$, which results in the mean of the λ distribution at 0.15 and a standard deviation of 0.2.[18] Both the α and β parameters for the decision weight function must be greater than 0, so they will also be assumed to be distributed with a gamma distribution, $\alpha \sim \Gamma(169, 7.69 \times 10^{-3})$ and $\beta \sim \Gamma(144, 8.33 \times 10^{-3})$. Thus the mean of α is ≈ 1.3 and its standard deviation is ≈ 0.1, and the mean of β is ≈ 1.2 and its standard deviation is ≈ 0.1. This results in $\theta^{RDU} = \{0.65, 0.3^2, 0.563, 0.26, 169, 7.69 \times 10^{-3}, 144, 8.33 \times 10^{-3}\}$.

Once again, we employ 2.5×10^6 draws from this joint distribution times and calculate the values for equations (3.17) and (3.20) through (3.29) for all $TT = 1024$ choice patterns and all $e \in [0, T]$ for the RDU subpopulation. With the results of the calculations for the EUT subpopulation calculated previously, and the results of the same calculations for the RDU population, we can mix each of these metrics as described in equation (3.39) with $\pi_{EUT} = 0.7$ and $\pi_{RDU} = 0.3$. Again, it is impractical to display the results of all metrics for all 1024 choice patterns, so first, we recreate Table 3.3 with the results of the RDU metrics.

There is a great deal of similarity between Table 3.4 and Table 3.3. In particular, the two subpopulations share the same 3 most likely choice patterns, though with different simulated likelihood, welfare, and error metrics. Again we note that the choice patterns which display Light MSB, rows 4, 5, 6, 8, 9, and 10, have 0 probability of containing 0 choice errors, and that there are a few examples of the

[18] With the mass of the λ distribution closer to 0, *a priori* we should expect fewer choice errors among the RDU population than the EUT population.

Table 3.4: HL-MPL Welfare and Error Expectations for Top Ten Choice Patterns, RDU

	Choice in Row										Simulated Likelihood	Expected Errors	Welfare Proportion	Welfare Surplus	$P_E(e=0)$	$P_E(e=1)$
Rank	1	2	3	4	5	6	7	8	9	10						
1	0	0	0	0	0	0	1	1	1	1	0.1617	0.702	0.9896	1.0403	0.4013	0.4979
2	0	0	0	0	0	0	0	1	1	1	0.1135	0.98	0.9831	1.012	0.2941	0.467
3	0	0	0	0	0	1	1	1	1	1	0.0813	1.227	0.9707	0.9687	0.2038	0.434
4	0	0	0	0	0	1	0	1	1	1	0.0571	1.505	0.9642	0.9404	0	0.4979
5	0	0	0	0	0	0	1	0	1	1	0.0445	1.568	0.9626	0.896	0	0.467
6	0	0	0	0	1	0	1	1	1	1	0.033	1.635	0.947	0.8881	0	0.434
7	0	0	0	0	0	0	0	0	1	1	0.0312	1.845	0.9561	0.8678	0.0656	0.2955
8	0	0	0	0	0	0	1	1	0	1	0.0252	1.699	0.9491	0.7724	0	0.4013
9	0	0	0	0	1	0	0	1	1	1	0.0232	1.912	0.9405	0.8599	0	0.2941
10	0	0	0	0	0	1	1	0	1	1	0.0224	2.093	0.9437	0.8244	0	0.2038

disconnect between likelihood and welfare. The Light MSB choice pattern in row 8 is expected to contain fewer choice errors than the Consistent pattern in row 7. The Light MSB choice pattern in row 9 is expected to provide greater welfare surplus than the Consistent pattern in row 7. Additionally, we observe that going from row 6 to 7 the Simulated Likelihood decreases, but the Welfare Proportion metric increases.

A major difference between the two subpopulations is that the RDU subpopulation's most likely choice pattern has a much greater likelihood (0.1617) than the EUT subpopulation's most likely choice pattern (0.0360). Much of this is due to the greater mass of the λ distribution close to 0 in the RDU subpopulation compared to the EUT subpopulation, but it is also because the distributions chosen for the decision weighting parameters imply greater risk aversion. This means that although the CRRA coefficients lie near the boundary of row 6 and 7 of the HL-MPL instrument, the way the RDU subpopulation weights probabilities makes them more risk averse, and therefore more likely to switch at row 7 than

if they did not weight probabilities. If instead of utilizing the mixture of EUT and RDU shown here, we utilized a mixture of two EUT subpopulations, with one subpopulation being more risk averse than the other, we would see similar results. Agents from a more risk averse population of EUT agents would switch at later rows than agents from the less risk averse EUT population, just as agents from the RDU population discussed here switch at later rows than agents from the less risk averse EUT population. These differences are important when we look at the grand population metrics.

Table 3.5: HL-MPL Welfare and Error Expectations for Top Ten Choice Patterns, EUT-RDU Mixture

	Choice in Row										Proportion EUT	Simulated Likelihood	Expected Errors	Welfare Proportion	Welfare Surplus	$P_E(e=0)$
Rank	1	2	3	4	5	6	7	8	9	10						
1	0	0	0	0	0	0	1	1	1	1	0.41	0.0822	0.827	0.9872	0.9648	0.3455
2	0	0	0	0	0	0	0	1	1	1	0.48	0.0656	0.959	0.9851	0.9557	0.314
3	0	0	0	0	0	1	1	1	1	1	0.46	0.0451	1.386	0.9673	0.8898	0.1778
4	0	0	0	0	0	1	0	1	1	1	0.53	0.0365	1.518	0.9651	0.8807	0
5	0	0	0	0	0	0	1	0	1	1	0.58	0.0314	1.587	0.9652	0.8539	0
6	0	0	0	0	0	0	0	0	1	1	0.64	0.0263	1.719	0.9631	0.8448	0.1137
7	0	0	0	0	1	0	1	1	1	1	0.57	0.0232	1.741	0.9473	0.8217	0
8	0	0	0	0	1	0	0	1	1	1	0.64	0.0194	1.873	0.9452	0.8126	0
9	0	0	0	0	0	0	1	1	0	1	0.61	0.0193	1.814	0.9501	0.733	0
10	0	0	0	0	0	1	1	0	1	1	0.62	0.0178	2.146	0.9453	0.7788	0

The grand population metrics displayed in Table 3.5 are barely noteworthy by themselves. They easily could have been generated by a population composed entirely of EUT agents with a distribution of λ somewhat closer to 0 than the EUT subpopulation that actually composes 70% of the agents in this population. Many of the same features of the two subpopulations are apparent in the mixed grand population; choice patterns displaying any form of MSB have 0 likelihood of 0 choice errors, and there are some disconnects between simulated likelihood and

welfare as observed in rows 5-6, and 7-8 for the welfare efficiency metric and rows 9-10 for the welfare surplus metric.

Of greater interest is the "Proportion EUT" column of Table 3.5, defined by equation (3.41). This metric calculates the unconditional likelihood that a subject displaying a particular choice pattern belongs to the EUT subpopulation we defined. For every choice pattern in the top ten most likely to be observed choice patterns, we observe that the proportion of the agents belonging to the EUT subpopulation that generated the choice pattern is smaller than the proportion of EUT agents in the grand population. For the top 3 choice patterns, the difference between the proportion of EUT agents *in the total population* and the proportion of EUT agents *that generated the choice pattern* is greater than 20 percentage points. In fact, it is more likely than not that these choice patterns are generated by the RDU subpopulation. This is despite the fact that the EUT subpopulation makes up 70% of the grand population, and that the top three most likely to be observed choice patterns in the grand population all correspond to the same top three choice patterns in the EUT subpopulation.

It should also be clear from Table 2 and Figure 3.1 that not all choice patterns that are consistent with EUT should be judged as superior to choice patterns which are apparently inconsistent with EUT from the perspective of welfare realization as consumer surplus or welfare efficiency.

3.3 Population Level Analysis of Welfare: Preferences, Noise, and the Instrument

The proposed characterizations of the welfare of a sample, including the degree to which certain consistent choice patterns are expected to be more costly in

welfare terms than inconsistent choice patterns, are ultimately determined by the distribution of preferences and stochastic parameters in the sample. To analyze how the welfare characterizations change as the distribution of preferences change in the sample, we could repeat the computational exercise that led to Table 3.3 for a few different distributions and discuss implications pattern by pattern. This exercise, however, will produce data only for the populations chosen, and will be less informative about how expectations of welfare change as the population changes. Instead, it will be useful to define a few population-level metrics that allow us to look at the data at the aggregate level. For instance, for each $y \times T$ choice pattern, we can weigh the expected welfare efficiency resulting from equation (3.27) by the simulated likelihood resulting from equation (3.17) and then sum across all TT choice patterns to retrieve the sample expected welfare efficiency:

$$\mathrm{E}(\%W_T(\theta)) = \sum_{tt=1}^{TT} SL_{Ntt}(\theta) \times \mathrm{E}(\%W_{tt}|\theta) \tag{3.43}$$

Similar expectations can be derived for any of the per-choice pattern statistics defined previously, but we pay particular interest to the statistics derived from equations (3.21), and (3.22) where $e = 0$. We are not limited to looking at expectations however: we can utilize equation (3.43) to derive higher moments of these statistics, such as the variance:

$$\mathrm{Var}(\%W_T(\theta)) = \sum_{tt=1}^{TT} SL_{Ntt}(\theta) \times [\mathrm{E}(\%W_{tt}|\theta) - \mathrm{E}(\%W_T|\theta)]^2 \tag{3.44}$$

Having the means and variances of the statistics described allows us to make high-level inferences about the welfare implications of an instrument like the HL-MPL instrument on different populations for a given stochastic model. That is, we can contribute to answering of our primary question of "what are the welfare

implications of stochastic models" by solving equations (3.43) and (3.44) for various values of θ and relating the elements of θ to these results. We can substitute any $y \times T$ statistic derived from equations (3.21), (3.22), (3.27), and (3.29) in place of $\%W_T$ in equations (3.43) and (3.44) to describe these statistics on a population by population basis.

While equations (3.43) and (3.44) may in fact have analytical solutions to determine these relationships, meaning we could attempt to solve for the partial derivative of equations (3.43) and (3.44) with respect to each element of θ, any analytical solution will be unique with respect to so many idiosyncratic factors that this becomes infeasible and potentially uninformative. These factors include:

- the stochastic model;
- the utility model;
- the location, dispersion and shape of the joint distribution governing the complete stochastic specification;
- the number of Halton draws used to simulate the probabilities;
- the base prime number used for the Halton sequences; and
- the specific tasks faced by the sample population.

Given these limitations, we instead examine the relationship of the parameters making up the stochastic specification, i.e. the elements of θ, with the associated results of equations (3.43) and (3.44) visually and with the use of locally weighted polynomial regression (LOESS) developed by Cleveland, Grosse, Shyu, Chambers & Hastie (1992). To conduct this examination, we generate 500,000 unique population parameter sets, θ_i, the elements of which are assumed to be uncorrelated, and solve equations (3.43) and (3.44) for the statistics derived in equations (3.21), (3.22)

and (3.27) with $e = 0$ and with each equation solved with $10,000$ draws from each unique population.

Each population has normally distributed preference parameters and gamma distributed stochastic error parameters. Thus, each θ_i is comprised of 4 elements: the mean of the CRRA parameter and the λ term, μ_r and μ_λ, and standard deviation of the CRRA parameter and the λ term, σ_r and σ_λ. Each of the candidate θ_i vectors was randomly drawn from a joint uniform distribution of these elements. The bounds of the marginal distributions of these elements are as follows: $\mu_r \in [-1.9, 1.55]$, $\sigma_r \in [0, 1]$, $\mu_\lambda \in [.05, 2.25]$, $\sigma_\lambda \in [.01, .75]$. These bounds are almost arbitrary; the bounds for μ_r were chosen to be just outside the indifference bounds of the HL-MPL instrument, but the remaining marginal distributions were chosen to be broad enough to yield some interesting patterns.

This exercise results in 8 statistics for each θ_i: the means and variances of the expected proportion of welfare to the maximum attainable welfare, the expected welfare surplus, the expected number of choice errors, and the expected proportion of agents who make no errors. Each statistic can be plotted against the 4 elements of θ_i. The result is 32 plots of the raw data and 32 charts of the LOESS lines associated with the raw data plots. All LOESS lines are plotted along with 95% confidence intervals.

Each plot and chart also attempts to give information about another parameter not plotted on the x or y axes by color coding the plotted data with respect to different values of this "z" parameter. For μ_r this "z" parameter is σ_r, for σ_r it is μ_r, for μ_λ it is σ_λ and for σ_λ it is μ_λ. For each of the charts, the "z" parameter is split into quartiles and for the LOESS line charts, LOESS lines are calculated for the "x" and "y" parameter values that belong to each quartile. Additionally, in

the raw data plots, each point has been given a large degree of transparency. This means that the density of points in the plot is represented by the density of color in the plot.

I examine the LOESS line charts of these data. First I discuss the effect on welfare expectations of the parameters governing the stochastic model, and then discuss the parameters governing the utility model. Thus, we first look at Figures 3.2, 3.3, 3.4 and 3.5. Figures 3.2 and 3.3 demonstrate the effect of the mean of the distribution of the λ term on welfare and the error frequencies, while Figures 3.4 and 3.5 demonstrate the effect of the standard deviation of the λ term on the same statistics.

Figure 3.2: Mean of λ Compared to Welfare

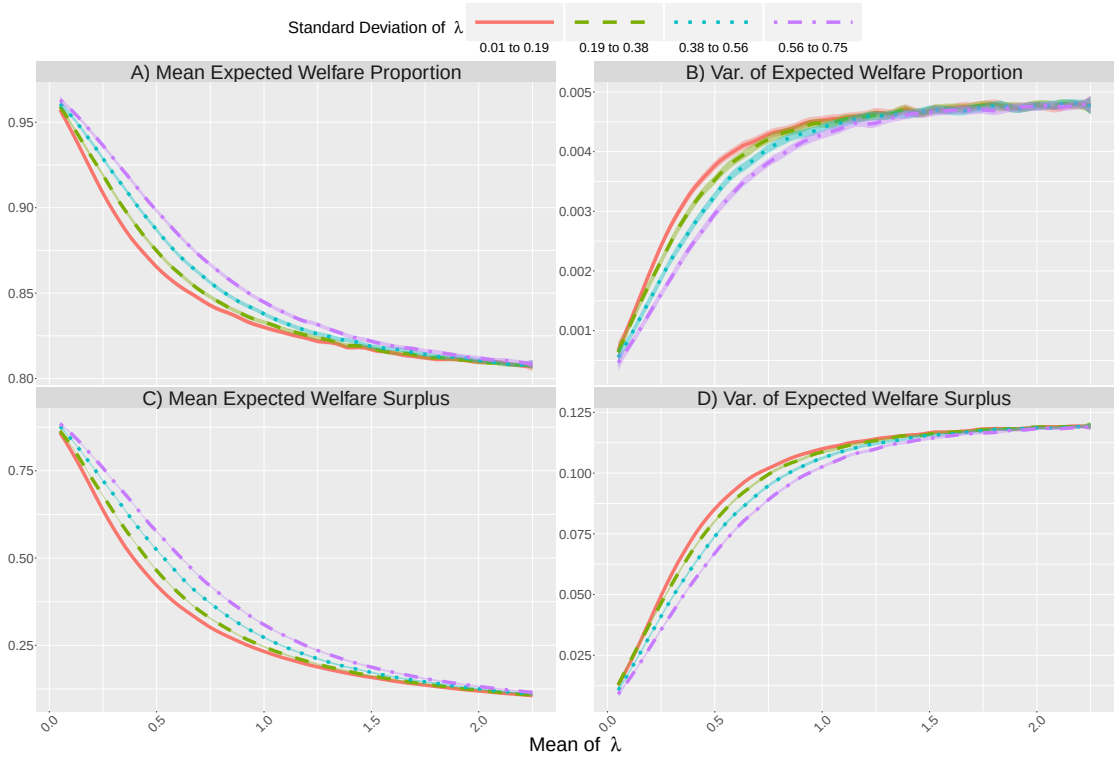
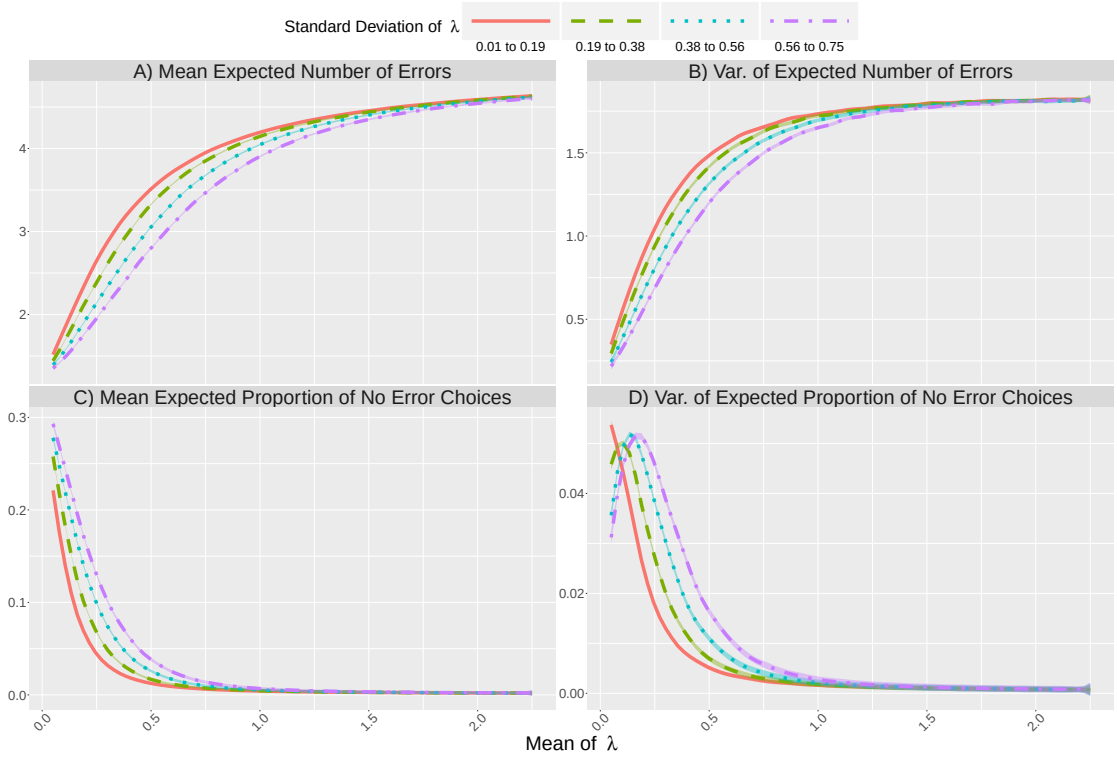


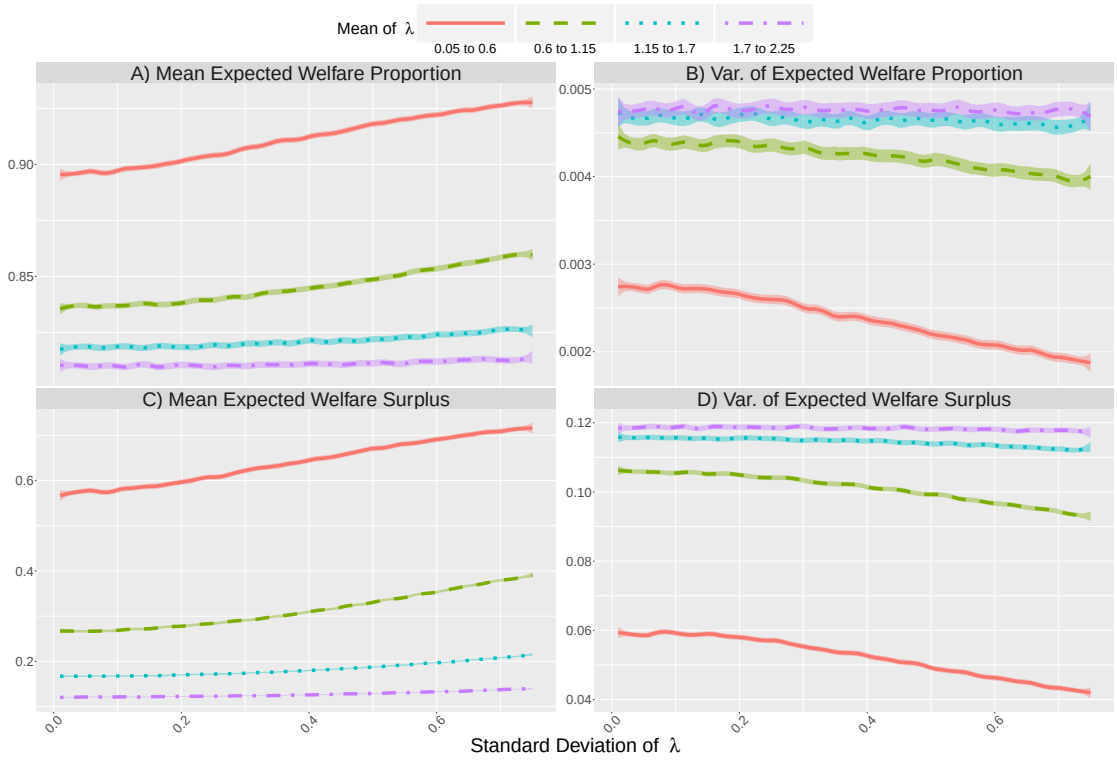
Figure 3.3: Mean of λ Compared to Errors



The results of the plots of stochastic model parameters are mostly intuitive and unsurprising. Looking at Figures 3.2 and 3.3, as the mean of the distribution increases, the expected welfare and expected proportion of 0-error choice patterns monotonically decreases, while the expected number of choice errors monotonically increases. Because λ has a gamma distribution, for any given mean, a higher standard deviation implies that the mass of the distribution shifts closer towards 0. Thus, it is unsurprising that those populations with high standard deviations of λ tend to exhibit choice patterns with fewer expected choice errors and greater expected proportions of no error choice patterns. This is because for any given choice problem, a lower value of λ implies a lower probability of committing a

choice error.[19] This directly translates into greater expected welfare than those populations with lower standard deviations holding the mean constant.

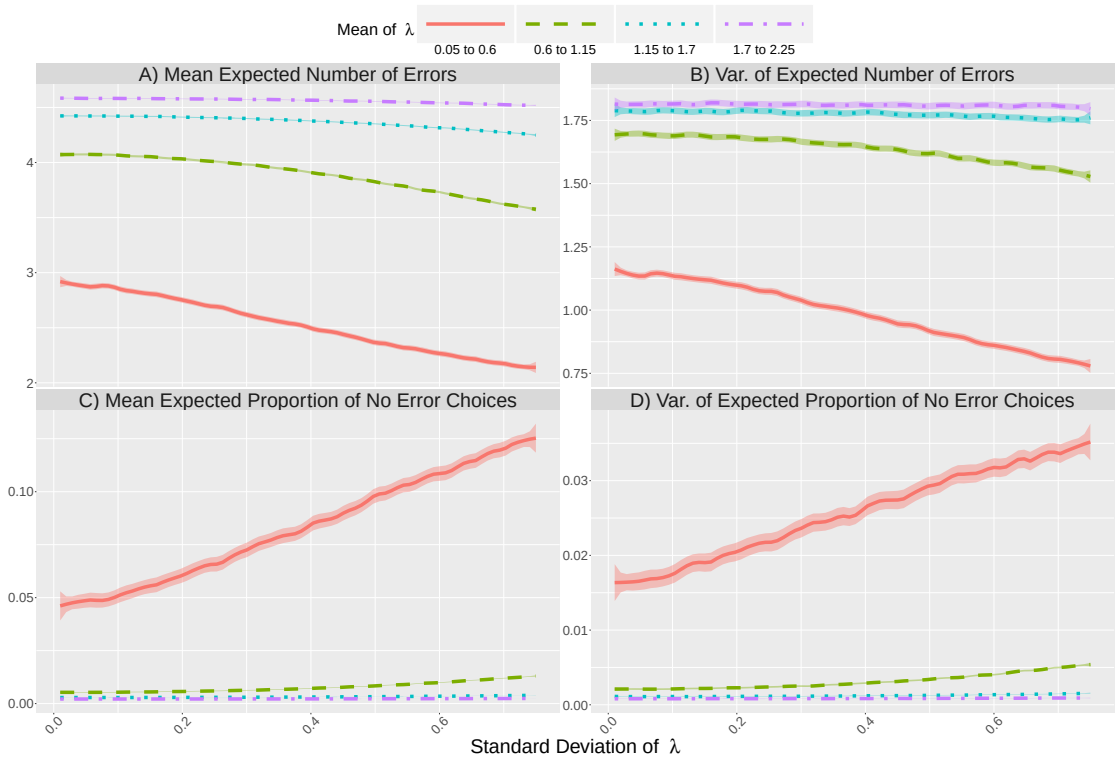
Figure 3.4: Standard Deviation of λ Compared to Welfare



Looking at Figures 3.4 and 3.5, we can see that σ_λ is far less influential than μ_λ. In the (A) and (C) charts of Figure 3.4, the slopes of the LOESS lines are slightly positive, but mostly flat other than the line for the lowest quartile of μ_λ. In the (A) and (C) charts of Figure 3.5, we see much the same mostly flat lines indicating very little variation across the parameter space. Again the exception is the line for the lowest quartile of μ_λ. This should not be surprising given that the populations were generated with a CU stochastic model. The third quartile of μ_λ begins at

[19] Since λ is in the denominator of each exponential transformation, as $\lambda \to 0$, $Pr(y_t = a) \to 1$ for $a = 1$ and $Pr(y_t = a) \to 0$ for $a \neq 1$ regardless of the other parameters.

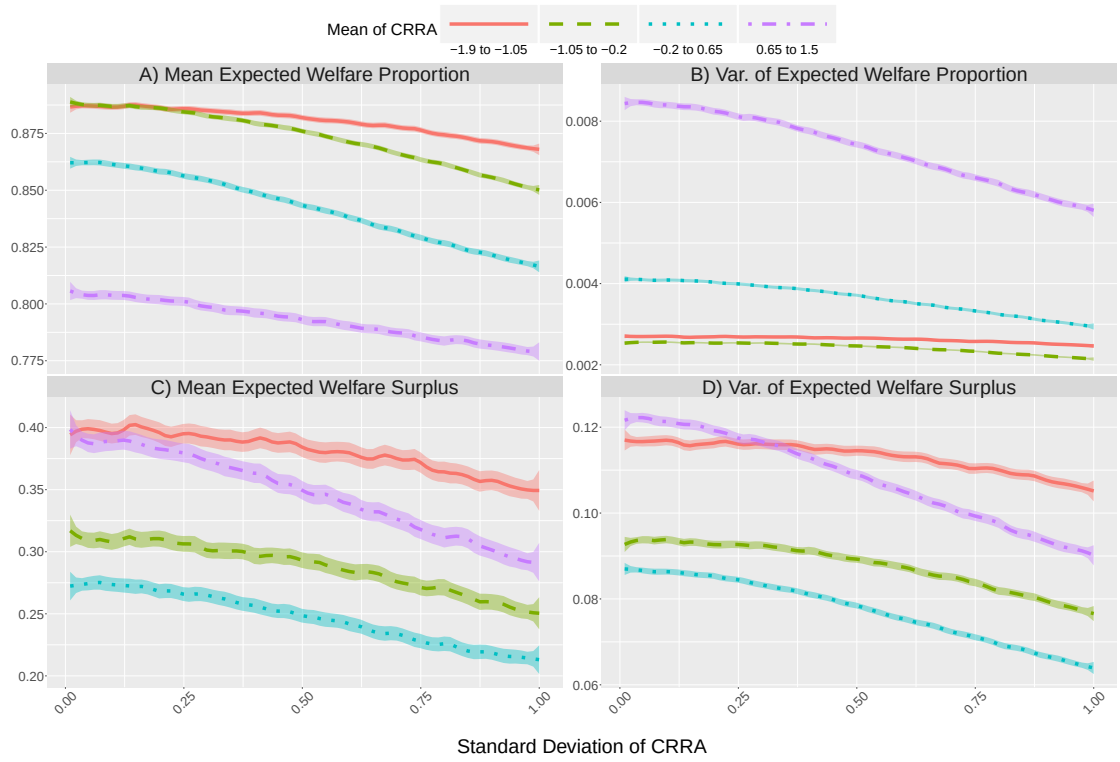
Figure 3.5: Standard Deviation of λ Compared to Errors



1.15, which means that the majority of the mass of the distribution of lambda in any population will lie above 1 for any value of σ_λ in the range explored. At these high levels of lambda, most choice probabilities will converge to something close to $\Pr(y_t = a) \to 0.5$.

In contrast to the monotonic relations of the lambda distribution, the effect of the CRRA parameters on the expected welfare and expected error statistics displays influences of the idiosyncratic aspects of the HL-MPL instrument. This is most apparent in the plots of μ_r. In interpreting these plots, it is important to keep in mind that the CRRA parameters used in each population are normally distributed. Thus, the mean of the distribution always represents the point of the

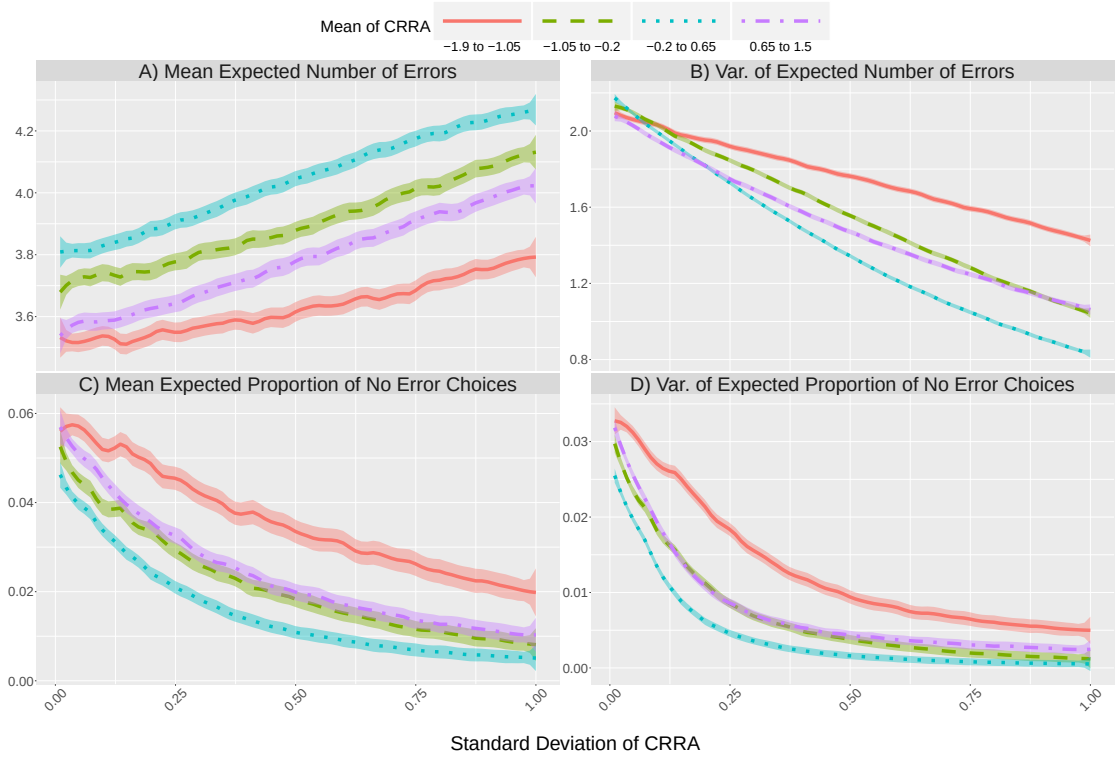
Figure 3.6: Standard Deviation of CRRA Compared to Welfare



distribution with the greatest density, with smaller standard deviations leading to greater concentration of the mass of the distribution around the mean and larger standard deviations leading to the reverse.

In Figures 3.8 and 3.9, each tick mark on the x-axis represent the values of the CRRA parameter at which an agent would be indifferent between lotteries for some row of the HL-MPL instrument. From left to right, the first tick mark corresponds to the value of the CRRA parameter that would make an agent indifferent between the lotteries in the first row of the instrument, the second tick mark corresponds the second row of the instrument, and so on. There are only 9 ticks because the there does not exist any CRRA parameter which would set an agent to be indifferent

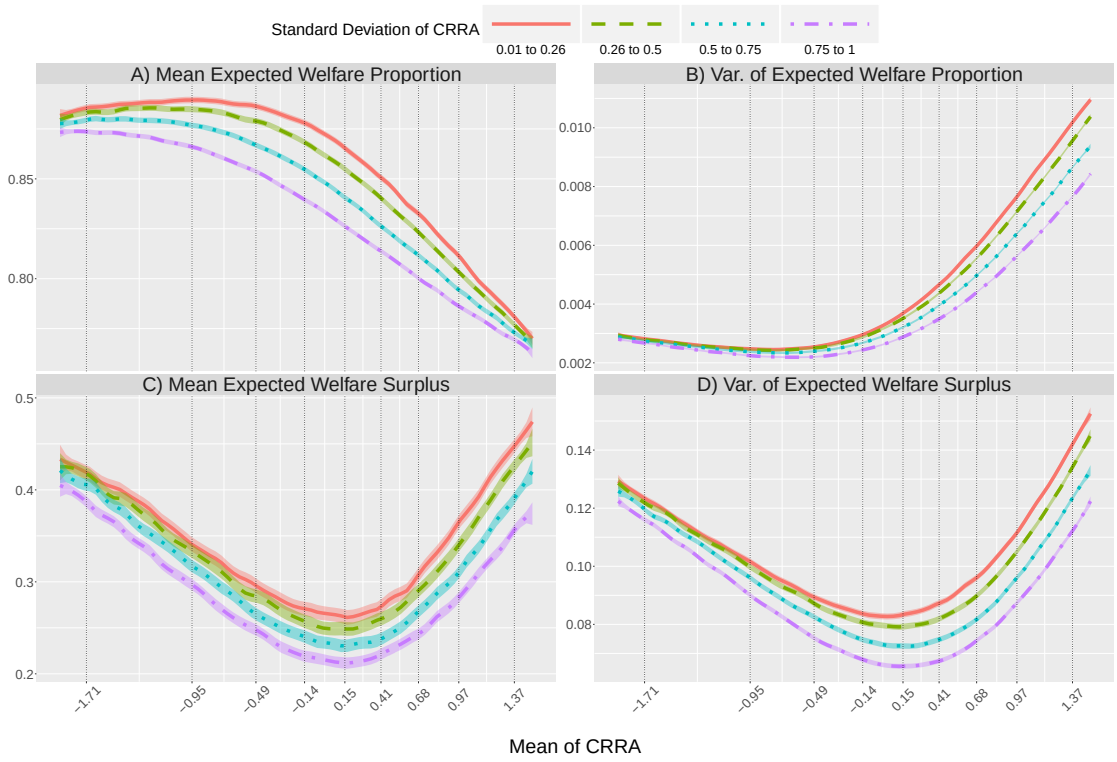
Figure 3.7: Standard Deviation of CRRA Compared to Errors



between the lotteries in row 10 of the instrument.

I begin by first discussing the effect of μ_r on choice errors as displayed in Figure 3.9. Something that is immediately apparent is that the orange LOESS line, depicting populations with low standard deviations of CRRA parameters, is much more volatile than the other quartile lines. Interestingly, the orange line dips downward in plots (B), (C) and (D) and peaks upward in plot (A) at the values of μ_r that correspond to the indifference values described previously. From plots (A) and (C), we draw the conclusion that as the mass of the distribution of preferences grows around parameter values which correspond to values which imply indifference in a choice scenario, we see an increase in the number of choice

Figure 3.8: Mean of CRRA Compared to Welfare

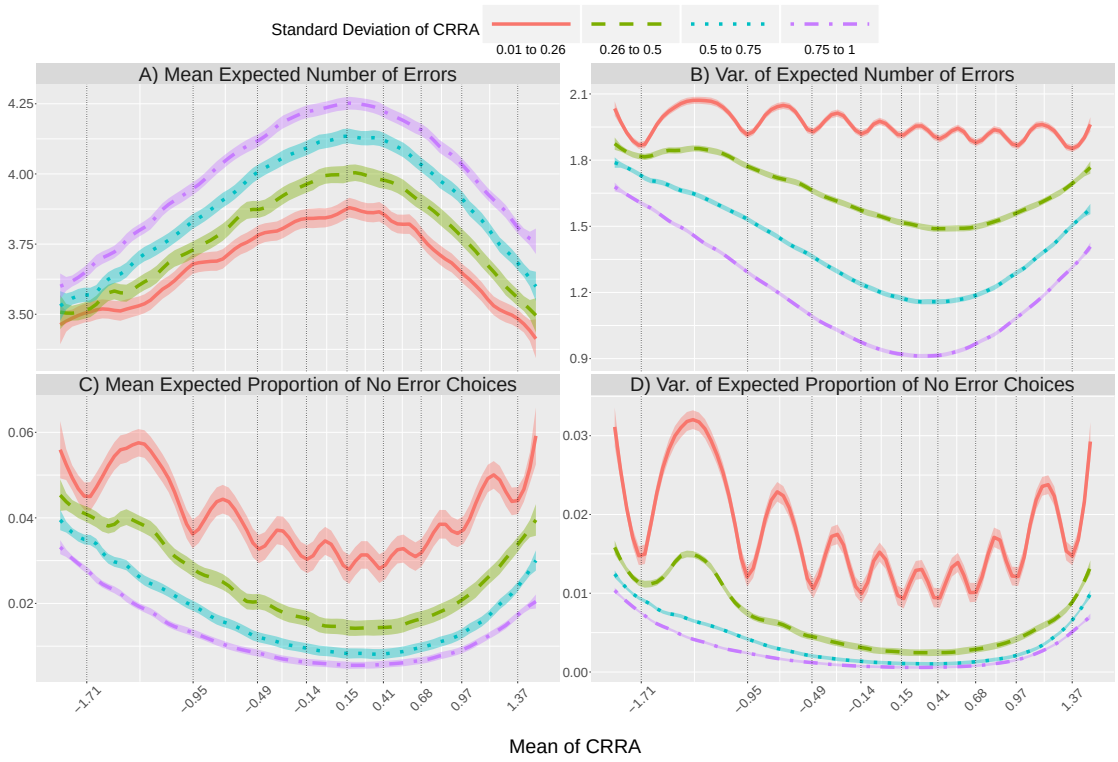


errors in the population.

In the case of the quartile described by the orange line, the idiosyncratic relationship between μ_r and the points that represent indifference also holds for the variance of expected choice errors, as depicted in plots (B) and (D) of Figure 3.9. That is, the increase in the average number of expected errors at these points is largely driven by a sharp reduction in the probability of observing a choice pattern with few expected errors relative to the probability of observing a choice pattern with a large number of errors.[20]

[20]Because $\Pr(y_t = a) = \Pr(y_t = b) \forall a, b$ as $\lambda \to \infty$, the maximum expected number of errors that can ever be observed is $\sum_{t=1}^{T} \frac{A_t - 1}{A_t}$. That is, since every option is given equal probability in the limit, and only one option is not an error, the sum of ratio of choice errors to options across all tasks is the maximum expected number of choice errors in the limit. The maximum in the

Figure 3.9: Mean of CRRA Compared to Errors



The remainder of the quartiles however do not follow this general pattern of heightened influence around the indifference points. Instead, for plots (B),(C) and (D) of Figure 3.9, the lines generally decrease until $\mu_r = 0.15$ and plot (A) increases until just about the same point. This less volatile pattern is because the 3 highest quartiles all indicate populations with high standard deviations. Consider the 3 upper quartile lines around $\mu_r = 0.15$. The distances between this point and the two closest indifference points are 0.26 and 0.29. The second lowest quartile's lower bound of σ_r is 0.26, which means that the density of the preference relation distribution at these points of indifference is much larger than for the

case of the HL-MPL instrument where $A_t = 2\ \forall t$ and $T = 10$ is therefore 5.

lowest quartile, relatively. It should be apparent from observing the lowest quartile line that as the density of the preference distribution increases around these points of indifference, the frequency of errors will increase. We can attempt to see this more formally by creating a metric that characterizes how much the distribution of preferences "sits" on these points of indifference:

$$D = \sum_{r}^{R} \frac{f(r)}{\max f(x)} \tag{3.45}$$

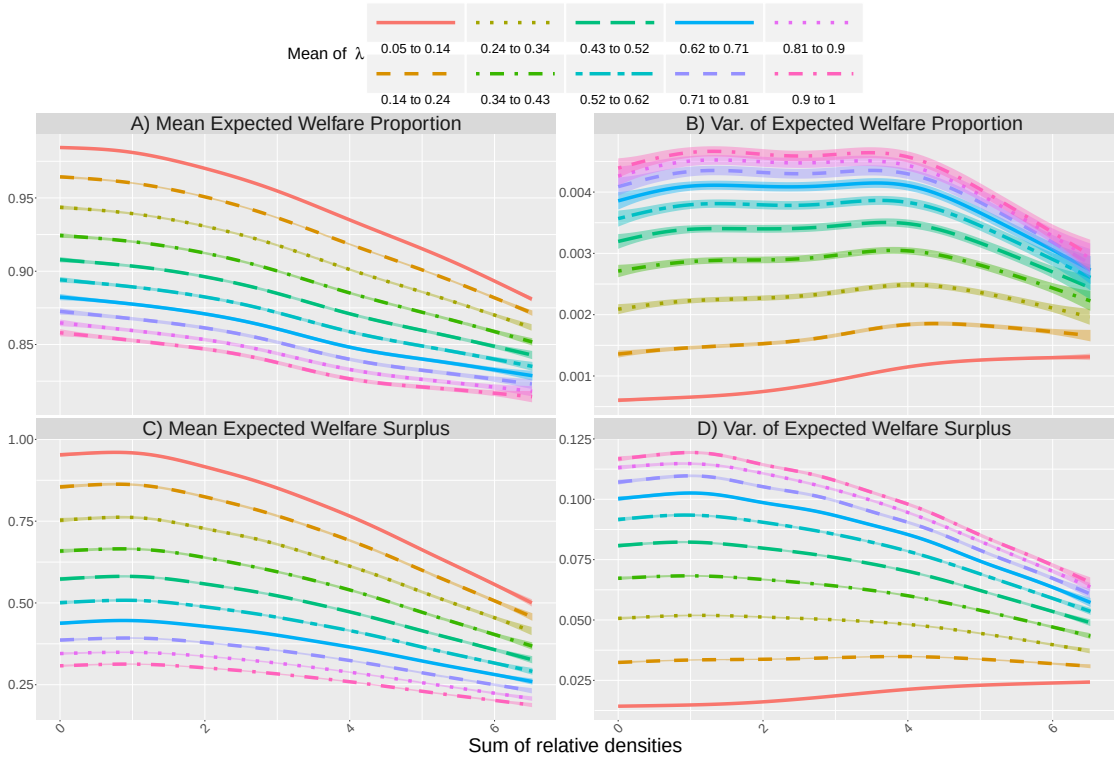
where $f(r)$ is the density of the distribution of CRRA parameters for the population at point r and R is the set of values for the CRRA parameters at which an agent would be indifferent between the two options in each choice problem. The denominator of the ratio is the maximum density of the distribution $f(\cdot)$ for the population. Since the CRRA parameters were distributed normally, this value is always equivalent to the density at the mean, μ_r. The set of R in for the HL-MPL instrument is:

$$R \equiv \{-1.71, -0.95, -0.49, -0.14, 0.15, 0.41, 0.68, 0.97, 1.37\} \tag{3.46}$$

We evaluate the metric from equation (3.45) against the 8 statistics utilized in Figures 3.8 through 3.5. Since it can be seen in Figures 3.2 and 3.3 that the effect of the μ_λ term asymptotes rapidly as $\mu_\lambda > 1$, we restrict our plots to populations for which $\mu_\lambda < 1$. This leaves us with about 150k observations. These 150k observations are first split into deciles of μ_λ and then the LOESS lines are calculated for each decile. This splitting of the data helps to make clear the large effect of the stochastic elements on the statistics explored and also the large amount of heterogeneity in the effect of preference parameters caused by the stochastic parameters.

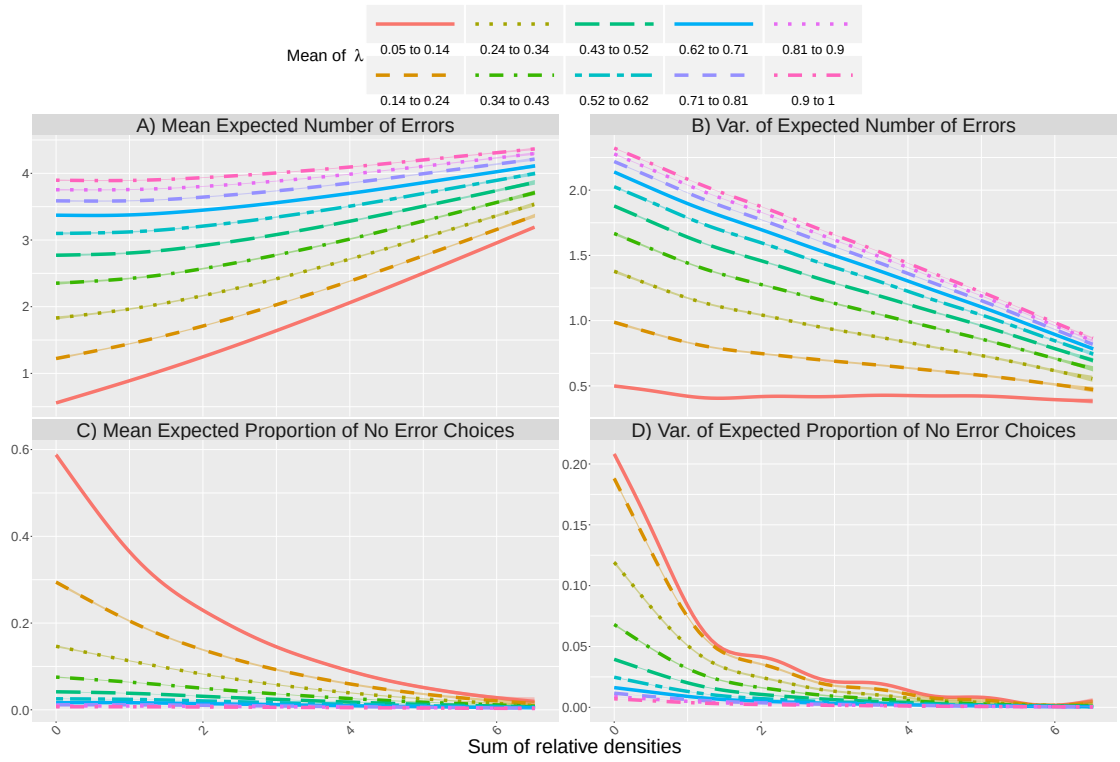
The metric developed in equation (3.45) is not perfect, we should expect to see clumping of data points around 0 and 1 where populations will be wholly sitting on one point or wholly between points, but it does provide a generally good description of the phenomenon we are concerned with. Looking at Figure 3.10, plots (A) and (C), we can confirm what was suspected to be driving the shape of the plots in Figure 3.9. As D increases, and more of the density of the CRRA distribution is shifted onto the points describing indifference, the greater the expected number of errors we should observe.

Figure 3.10: D Statistic Compared to Welfare



This effect is remarkably monotonic across every decile of μ_λ, though the effect is strongest for lower deciles. What should be no surprise is that for the highest

Figure 3.11: D Statistic Compared to Errors



3 deciles of μ_λ we observe close to 0% of the populations considered to produce choice patterns with no choice errors, as can be seen in plot (C). The variance statistics in plots (B) and (D) are generally monotonic, but not universally so. In general, the variance in the number of expected errors across populations tends to decrease as D increases. This is in line with the populations becoming increasingly error prone.

In Figure 3.11 we see the story of Figure 3.10 interpreted into welfare terms, but with an interesting and important difference: the expected welfare metrics in plots (A) and (C) are effectively equal around $D = 0$ and $D = 1$. There doesn't exist an equality in the error metrics around these values of D in Figure 3.10, nor

should there be. $D = 0$ corresponds to populations which have a μ_r and σ_r such that the entire population sits between the indifference points in R. $D = 1$ will generally[21] represent the opposite; such a population will have a μ_r and σ_r such that the entire population sits on top of one of the indifference points in R. If the entire population sits far from an indifference point, holding the stochastic element constant, we expect there to be fewer errors compared to a population that sits on top of an indifference point because the average agent will not be close to indifference for the lottery pair in question. But, this is also precisely why the welfare metrics are close to equivalent: if a population sits on an indifference point, it means that agents are mostly indifferent between the options in the lottery pair, and therefore any errors made for this lottery pair will be relatively less costly in terms of welfare.

Other than the particular case where $D = 0$ and $D = 1$, in Figure 3.11 we see the general trend that we might expect from looking at Figure 3.10: as the D metric increases and the relative density of the CRRA distribution increases around points of indifference, expected welfare decreases monotonically. This is because, other than the case of $D = 1$ where errors should be relatively frequent but not costly in welfare terms, an increasing D not only means that a greater proportion of agents lie on the indifference points, but also lie around it. It is this greater proportion of agents lying sufficiently near an indifference point to make an error relatively likely, but sufficiently far to make it relatively costly which drives down expected welfare. Similarly to what was seen in Figure 3.10, in Figure 3.11 we see that the effect of D is stronger with populations with μ_λ in the lower deciles

[21]Generally because there are multiple ways to get $D = 1$. A population with μ_r close to one and a σ_r such that there is some density on $r_j \in R$ *s.t.* $i \neq j$ can potentially make $D \to 1$. However, $\mu_r \to r_j \in R$ and $\sigma_r \to 0$ is the most frequent scenario.

and weakest with populations with μ_λ in the higher deciles.

What is also clear from Figures 3.10 and 3.11 is that the preference aspect of the utility model, represented by D, contributes far less to expected choice errors and, more importantly, to expected welfare than is contributed by the stochastic aspects of the model. Looking at the lowest decile lines in Figures 3.10 and 3.11, we can see that a relatively large increase in D is needed to cause the same effect as moving to the next lowest decile. Comparing the lowest decile with the highest decile reveals tremendous changes in expected errors and expected welfare while holding D constant for the populations analyzed.

The general analysis of the population level data reveals several somewhat expected results, and several somewhat unexpected results. Firstly it is clear, and unsurprising, that the means of both the CRRA and λ distributions individually drive a great deal of the variation in the number of expected choice errors and the expected welfare of a population. Specifically, the finding that the effect of the mean of λ on the expected number of choice errors was large should have been obvious *a priori*. The λ parameter directly influences choice probabilities regardless of the underlying instrument. Similarly, that populations with CRRA parameters tightly distributed around a point of indifference would have greater expected number of choice errors is intuitive. That larger numbers of expected choice errors generally lead to lower welfare was already apparent from previous analyses.

Somewhat more surprising is just how dominant the stochastic elements of utility functions are over the preference aspects when deriving expectations around welfare. Figures 3.10 and 3.11 make clear, despite the potential flaws with the D metric, that the way the preference parameters interact with the idiosyncratic

aspects of the instrument matter a great deal, but the stochastic parameters unambiguously matter more. This result should be important to economists and policy makers concerned with estimating the potential welfare implications of new policy instruments.

3.4 Summary of Analyses

In this chapter I analyze the relationship between an experimental instrument and the preferences of populations of agents. I demonstrate a method for calculating the unconditional welfare surplus and efficiency for a given pattern of choices from an experimental instrument and for a given population of agents.

When considering a single hypothetical population of EUT agents several surprising results emerge. First, it becomes clear that many choice patterns which are able to be rationalized by either EUT or RDU can be more likely to contain a choice error than not. Given the hypothetical population chosen for analysis however, most of the choice errors for the choice patterns most likely to be observed can be considered to be not costly in welfare terms. Second, there are several choice patterns that contain obvious violations of EUT that nonetheless are more likely to be observed *and* are expected to produce greater welfare surplus than many choice patterns that do not contain any such apparent violations. Third, and most interestingly, there is a less than perfect correlation between the likelihood of a choice pattern being observed and the expected welfare surplus of that choice pattern. That is, there are many choice patterns that are less likely to be observed than other choice patterns, but nonetheless provide greater expected welfare.

Of particular note, shown in Figure 3.1, are choice patterns which include a choice of an option that is dominated by another option. This figure shows that

patterns of choice which contain FOSD choices are many times less likely to be observed from this hypothetical population than the equivalent choice pattern without the dominated choice, but only provide somewhat less expected welfare. Additionally, choice patterns with FOSD choices can provide greater expected welfare than consistent choice patterns that are much more likely to be observed. This is seen in Figure 3.1 by comparing the choice pattern designated "Y," which contains a FOSD choice, and the choice patterns designated by red circles in the shaded area of the same figure. The choice patterns designated by the red circles in the shaded region are consistent with EUT, more likely to be observed than "Y," and provide less welfare efficiency than "Y."

The extent to which the distribution of preferences in a population influences the expected unconditional welfare is also assessed. I simulate populations of agents and map their distributional parameters to expected unconditional welfare, as well as to the frequency of choice errors. Additionally I construct a metric which relates the marginal distribution of risk preferences to the "indifference points" of an instrument called the "D" statistic. These indifference points are the values of the parameters that would cause an individual agent to be indifferent between the options in the lottery pair. From this exercise two interesting results arise.

First, as the density of the distribution of preferences increases around any value that would indicate indifference for a lottery pair in the instrument, the expected welfare surplus decreases and the rate of choice errors increases. As the density of the distribution of preferences increases around multiple such indifference points, and thus the D statistic increases, this increase is even more apparent. This can be seen in Figures 3.8, 3.9, 3.10, and 3.11. Secondly, the value of the parameters governing the stochastic aspects of the model seem to play a larger than expected

role in the expected welfare surplus for any given population. This can be seen in Figure 3.2, and in how the slope of the D statistic compares to moving from one gradient of λ means to another in Figures 3.10, and 3.11.

Theses results should caution economists and policy makers concerned with the design a policy or experimental instrument. In a warning to those who would wish to "nudge" the behavior of agents to particular patterns of choice, this exercise has shown that policy makers should take care about the choice patterns they may which to nudge a agent into. For many agents, it would be to their detriment to nudge them from a pattern of choices which contain obvious errors to one which, on appearance to an informed observer, would contain none. Analyses of experiments involving risky choice that rely on tests which measure differences in choice frequencies around certain lottery pairs should note how the frequency of choice errors and "consistent" choice patterns vary with the distribution of risk preferences in a population.

Chapter 4

Welfare Inferences From Experimental Instruments

In Chapter 1 we described the efforts of economists to account for apparent violations of Expected Utility Theory (EUT) in economic experiments. Some of these efforts were directed at the development of alternative deterministic theories of utility, such as Prospect Theory by Kahneman and Tversky (1979), Rank Dependent Utility (RDU) by Quiggin (1982), and Regret Theory by Bell (1982), Loomes and Sugden (1982). Other efforts were focused on the redevelopment of stochastic models, such as the constant error or "tremble" model by Harless and Camerer (1994), the Strong Utility model by Hey and Orme (1994), the random preference model by Loomes and Sugden (1995), along with many derivatives of the Strong Utility model.

Many of the newly proposed theoretical explanations of the apparent violations of EUT were tested experimentally. A well known example is that of Hey and Orme (1994) (HO), who conduct an experiment to test if any of a variety of generalizations (and one restriction) of EUT can explain experimentally collected data significantly better than EUT while utilizing the Strong Utility model. HO

pick "winning" models for each subject on the basis of their estimates for each model and whether each model can be operationally distinguished different from EUT. They conclude that "our study indicates that behavior can be reasonably well modelled (to what might be termed a 'reasonable approximation') as 'EU plus noise.'" However, HO note:

> The inferences that can be drawn [...] about the adequacy or otherwise of EU are not, however, clear cut - mainly because of the large number of generalizations of EU under consideration. As this research has evolved, and the number of generalizations under consideration has increased, the number of subjects for whom EU emerges as "the winners" has declined. This is inevitable, though it is not clear how one should judge the rate of decline. [...] Monte Carlo work would be needed to shed more accurate light on such issues

The concerns raised by HO can largely be considered as referring to statistical power, and to the weight economists should place on type I versus type II identification errors. That there are asymmetries in the probability of type I and type II errors should be of little surprise to most econometricians, but the degree of asymmetry in the *cost* of these errors, I argue, is more important. In this chapter, I analyze the experimental instruments utilized by Harrison and Ng (2016) (HN) for recovering the utility functions of agents. The HN experiment, detailed in depth below, is utilized for this analysis because it links the econometric classification of individual subjects and the measurement of their risk preferences directly with welfare evaluation for the decision maker. Estimation of the welfare consequences of a subject's choices allows economists to make a judgement about how much the individual gains or loses, in expectation, about any given choice. In the case of the HN experiment, the focus is on how much the subject gains or loses when purchasing, or not purchasing, an insurance policy. Harrison (1989, 1992), in what

has become known as the "Flat Maximum" or "Payoff Dominance" critique, argues that as the difference in utility between two options approaches zero, the subject cares less and less about choosing one option over the other, and so economists should care less and less about the choices over options where the utility difference approaches zero. In a similar vein, I argue that we should care less about classification accuracy if the implied difference in welfare consequences between alternative models is minimal.

The following analysis focuses firstly on the capacity of the HN procedure to correctly classify an agent as employing one of two different utility models, and secondly on the welfare consequences of this characterization. Thus I attempt to remove some uncertainty about the power of the instrument, and propose metrics to address the question of how much economists should care about statistical power issues by linking them directly with welfare evaluations. To begin this analysis, I describe and replicate the classification and welfare calculation exercises of HN. Next I conduct a simulation analysis of the lottery instrument used in HN to determine the frequency of misclassification for two of the four models utilized by HN, and the welfare consequences of this misclassification. I next propose two ways to potentially alleviate the welfare concerns of misidentification.

4.1 Estimating a Benchmark using Harrison and Ng (2016)

HN report the results of an experiment intended to evaluate the welfare consequences of individuals' decisions to purchase insurance. This is in part a response to the large literature cited by HN (2016, p. 92) which evaluates insurance on the basis of "take-up": the rate at which individuals purchase insurance. They argue

that although a take-up metric is transparent and easy to measure, it doesn't allow for behaviorally general statements about whether an individual *should* have taken up the insurance product, and it does not quantify the consumer surplus from making the correct insurance purchase decision. These are, however, precisely the kind of normative welfare statements that economists should be making about the economic choices of agents. They are also the kind of normative welfare statements that can be made from estimating the utility functions of individuals and evaluating their choices with respect to these functions.

HN address the problem of evaluating the welfare consequences of the decision to purchase insurance or not by conducting a 2-part experiment. In the first part each subject is presented with a battery of 80 lottery pairs and asked to select one lottery from each pair that will be played out for payment. This part will be referred to as the "lottery task" throughout. The responses of each subject to the lottery task are used to estimate utility functions for that individual. In the second part each subject is endowed with \$20 and presented with 24 choices where they are asked to choose between a lottery which will result in a loss of \$15 with some probability p or no loss of the initial endowment with probability $(1-p)$, and a certain amount of money between \$15.20 and \$19.80. The choice of the certain amount of money is framed as the purchase of insurance against the risk of loss in the lottery option. This part will be referred to as the "insurance task" throughout. Both of these instruments are detailed in full in Appendix C of HN.

For each individual, HN use the data recovered in the lottery task to estimate four models, 1 Expected Utility Theory (EUT) model and 3 models in the Rank Dependent Utility framework first proposed by Quiggin (1982). Since EUT is a

special case of RDU, we can describe all 4 models in the framework of RDU:

$$RDU = \sum_{c=1}^{C} [w_c(p) \times u(x_c)] \tag{4.1}$$

where c indexes the outcomes, x_c, from $\{1, \ldots, C\}$ with $c = 1$ being the smallest outcome in the lottery and $c = C$ being the greatest outcome in the lottery, $u(\cdot)$ is a standard utility function, $w_c(\cdot)$ decision weight function applied to outcome c given the distribution of probabilities in the lottery ranked by outcome, p. The decision weight function, $w_c(\cdot)$, takes the form:

$$w_c(p) = \begin{cases} \omega\left(\sum_{k=c}^{C} p_k\right) - \omega\left(\sum_{k=c+1}^{C} p_k\right) & \text{for } c < C \\ \omega(p_c) & \text{for } c = C \end{cases} \tag{4.2}$$

where the probability weighting function, $\omega(\cdot)$, can take a variety of parametric or non-parametric forms. In the special case of EUT, the probability weighting function is just the identity of the objective probabilities:

$$\omega(p_c) = p_c \tag{4.3}$$

HN estimate 3 probability weighting functions for the RDU models. The first pwf is the power function (RDU_{Pow}) used by Quiggin (1982):

$$\omega(p_c) = p_c^{\gamma} \tag{4.4}$$

where $\gamma > 0$. The second pwf is the "Inverse-S" shaped function (RDU_{Invs})

popularized by Tversky and Kahneman (1992):

$$\omega(p_c) = \frac{p_c^{\gamma}}{\left(p_c^{\gamma} + (1-p_c)^{\gamma}\right)^{\frac{1}{\gamma}}} \tag{4.5}$$

where $\gamma > 0$. The third pwf is the flexible function proposed by Prelec (1998) (RDU_{Prelec}):

$$\omega(p_c) = \exp(-\beta(-\ln(p_c))^{\alpha}) \tag{4.6}$$

where $\alpha > 0$ and $\beta > 0$.

For all three RDU probability weighting functions there exist values for the probability weighting parameters which allow $w_c(p) = p_c$, the special case of EUT. For all four models HN use the CRRA utility function:

$$u(x) = \frac{x^{(1-r)}}{(1-r)} \tag{4.7}$$

where r is the coefficient of relative risk aversion proposed by Pratt (1964).

I continue to use the notation used in chapters 2 and 3 to describe a choice scenario by a subject, but limit it to a binary choice between two options, a and b. In this framework a choice of option a in task t is indicated by the function $y_t = a$, where $y_t = 1 \geq^i y_t = 2$. The values of a and b do not indicate the order or frame with which the options in task t were presented to the subject, but rather the ordinal rank the subject's utility function assigns to the options, with 1 always being the option of greatest utility. This notation is useful when describing the welfare consequences of choices below.

HN also use Contextual Utility (CU), as defined by Wilcox (2008), as the stochastic model. Thus for the models utilized, the probability that option a is

chosen is given by:

$$\begin{aligned} Pr(y_t = a) &= Pr\left(\epsilon_t \geq \frac{1}{\lambda_i}\left[G(\beta_i, X_{bt}) - G(\beta_i, X_{at})\right]\right) \\ &= 1 - F\left(\frac{G(\beta_i, X_{bt}) - G(\beta_i, X_{at})}{D(\beta_i, X_t)\lambda_i}\right) \end{aligned} \tag{4.8}$$

where ϵ_t is a mean 0 error term, F is a symmetric cumulative distribution function (cdf), meaning $1 - F(x) = F(-x)$, $G(\cdot)$ is the RDU utility model that takes the parameters β_i to calculate the utility of lottery a or b in task t comprised of outcomes and probabilities X_{at}, and λ_i is a precision parameter. The function $D(\cdot)$ separates contextual utility from a Strong Utility model:

$$\begin{aligned} &D(\beta_i, X_t) = max[u(x_{ct})] - min[u(x_{ct})] \\ &st.\ w_c(x_{ct}) \neq 0 \end{aligned} \tag{4.9}$$

Usually, the Normal or Logistic cdf is chosen for F. HN utilized the Logistic cdf and I employ the Logistic cdf for all calculations throughout. Given that each choice considered here only involves two lottery options, we can define the probability of choosing option a given a particular model, parameter set β_i, precision parameter λ_i, and outcomes and probabilities of option a, X_{at}, as

$$Pr(y_t aj) = \frac{\exp\left(\frac{G(\beta_i, X_{at})}{D(\beta_i, X_t)\lambda_i}\right)}{\exp\left(\frac{G(\beta_i, X_{at})}{D(\beta_i, X_t)\lambda_i}\right) + \exp\left(\frac{G(\beta_i, X_{bt})}{D(\beta_i, X_t)\lambda_i}\right)} \tag{4.10}$$

These choice probabilities in turn are logged and summed to produce a log-likelihood function for each of the four different models:

$$LL_i = \sum_t^T \ln\left[Pr(y_t)\right] \tag{4.11}$$

As a metric of welfare, HN primarily use the consumer surplus (CS) of each

choice. The CS of each choice is defined as the difference between the certainty equivalent (CE) of the chosen option and the certainty equivalent of the unchosen option. Since the CRRA utility function defined in equation (4.7) is used for all models discussed, we can define the CE as:

$$\begin{aligned} \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{(1-r)}}{(1-r)} &= \frac{CE_a^{(1-r)}}{(1-r)} \\ CE_a &= \left((1-r) \times \sum_{c=1}^{C} w_c(p) \frac{x_{ca}^{1-r}}{(1-r)} \right)^{1/(1-r)}, \end{aligned} \tag{4.12}$$

and the welfare surplus metric derived from this CE for any choice as:

$$\Delta W_{it} = CE_{iyt} - CE_{i1t}^{Z}, \tag{4.13}$$

and the accumulated welfare surplus as:

$$\Delta W_{iT} = \sum_{t=1}^{T} \left(CE_{iyt} - CE_{i1t}^{Z} \right) \tag{4.14}$$

where the y subscript indicates the option chosen (either 1 or 2 in the binary scenario we consider here), and the Z superscript indicates the remaining, unchosen options, of which the CE with the greatest value, designated by the subscript 1, is considered the foregone opportunity. HN (2016, p. 106) consider an additional metric of forgone welfare surplus as the difference between the maximal CE for every choice and the CE of the option actually chosen by the subject

$$\Delta F_{it} = -1 \times \left(CE_{i1t} - CE_{iyt} \right), \tag{4.15}$$

and the accumulated forgone welfare surplus

$$\Delta F_{iT} = \sum_{t=1}^{T} -1 \times \left(CE_{i1t} - CE_{iyt} \right) \tag{4.16}$$

With these metrics, the best possible value for any subject is 0, which would indicate that all choices made were optimal, whereas any positive value indicates the amount of welfare surplus forgone by the subject due to choice errors. These metrics line up easily with the metrics defined in equations (4.13) and (4.14), as should $y_t \neq 1$, $CE^Z_{i1t} = CE_{i1t}$.

HN estimate values of the CRRA utility parameter, r, the probability weighting parameters, γ, α, β, and the stochastic parameter λ, for each of the models presented above via maximum likelihood estimation (MLE) using the choices made by the subjects in the lottery task. HN (2016, pp. 107,110) initially calculate the welfare consequences of the choices made by each subject by using only the point estimates from the MLE, and then employ a bootstrap method which incorporates the covariance matrix of the standard errors.

For the bootstrap method, a multivariate normal distribution of parameter sets is bootstrapped from the estimates using the point estimates of these parameters as the means of the marginal distributions, and the covariance matrix of standard errors used as the covariance matrix of standard deviations. For each subject's parameter estimates 500 draws of parameter sets were taken, the welfare metrics calculated for each set of parameters, and then the values of the metrics averaged across the 500 draws. Since the covariance matrix used in the bootstrap method draws parameters from the joint distribution with respect to their density in the joint distribution, only a simple average is needed.

The experimental subjects consisted of 111 undergraduate students enrolled in several different colleges at Georgia State University, USA. Every subject received, and expected to receive, a guaranteed \$5 show up fee, but no specific information about the experiment or expected earnings was communicated to the subjects

before the experiment HN (2016, p. 98). The full set of instructions delivered to the subjects is available in Appendix C of HN.

4.1.1 Individual Level Estimation

HN employ a multi-step process for picking a "winning" model for each subject. First, all four models models cited in equations (4.3), (4.4), (4.5), (4.6) are estimated for each subject. Next, data are dropped from analysis on the basis of an *ex ante* defined set of "exclusionary rules" applied to every model estimated on the subjects. Finally, a "classification process" is employed to choose a model with which to categorize the subject.

HN propose 4 exclusionary rules:

- Any estimate for which the optimizer did not return a convergence code indicating both a gradient near 0 and a negative definite Hessian.
- Any model with a CRRA coefficient estimated to be greater than 15 or less than -15.
- RDU_{Pow} and RDU_{Invs} models where the γ parameter was estimated to be greater than 5.
- Any model with a CRRA coefficient estimated to be greater than .99 and less than 1.01.

The gradient and Hessian conditions indicate that the estimates are at a local maximum of a concave portion of the likelihood function. The next two rules indicate parameter values that, although mathematically possible for the given functionals, are nonetheless considered to be extreme to the point of not being reliable. Wakker (2008) details how the CRRA utility function has certain asymptotic

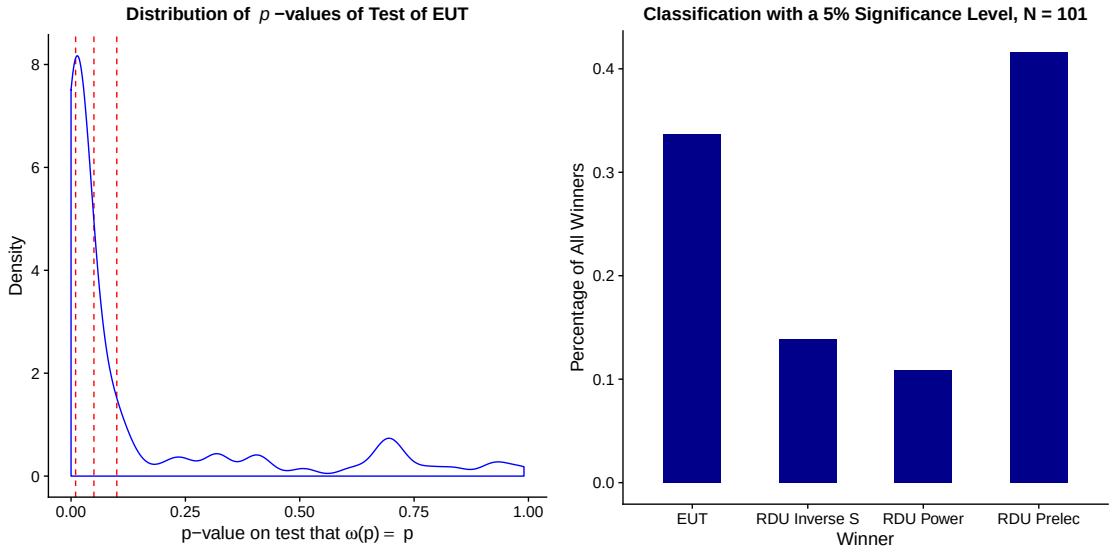
unattractive properties around 1. These properties may create numerical issues for the optimizer, and so estimated values very near 1 are viewed as less credible and are excluded from the analysis.

The classification process proposed by HN applies to all the remaining, non-excluded data. For 9 subjects, no model passed the exclusionary rules. The log-likelihood function given in equation (4.11) is equally applicable to all four models considered by HN, and seems a natural metric to declare a "winning" model among the 4 alternatives proposed. However, since the RDU models nest EUT as a special case (noted in equation 4.3), *a priori* we would expect RDU models to produce greater log-likelihoods than an EUT model on any given dataset, numerical issues aside. HN (2016, p. 102) note this issue and propose the additional qualification on RDU models that the probability weighting function implied by the estimated model must be statistically significantly different from a linear function, the special case of EUT, at the 10, 5, or 1 percent significance levels.

The EUT null hypothesis for the RDU_{Pow} and RDU_{Invs} models is $H_0 : \gamma = 1$, and the null hypothesis for the RDU_{Prelec} model is $H_0 : \alpha = \beta = 1$. Non-linear Wald tests are used to test these hypotheses. Any RDU model that fails to reject the null hypothesis is removed from consideration as a "winning" model. If the EUT model did not converge for the subject in question, the models considered will only consist of the RDU models which tested as different to EUT. If the EUT model did not converge *and* no RDU model tested as different to EUT, then all of the converged RDU models will be considered. The "winning" model for each subject is chosen from among the models which have met criteria derived from the Wald test. The winning model is then used to calculate the welfare consequences of the subject's choices on the insurance task.

When I utilize the same classification processes employed by HN on their data, we see a somewhat different distribution of subjects classified to the four models in Figure 4.1. These differences are relatively minor, showing somewhat more RDU subjects and fewer EUT subjects than reported by HN. I do however, replicate in Figure 4.2 the distribution of per-choice consumer surplus presented in Figure 10 of HN (Harrison and Ng 2016, p. 108). Figure 10 of HN and Figure 4.2 are not visually distinguishable, and the mean welfare surplus metric is the same.

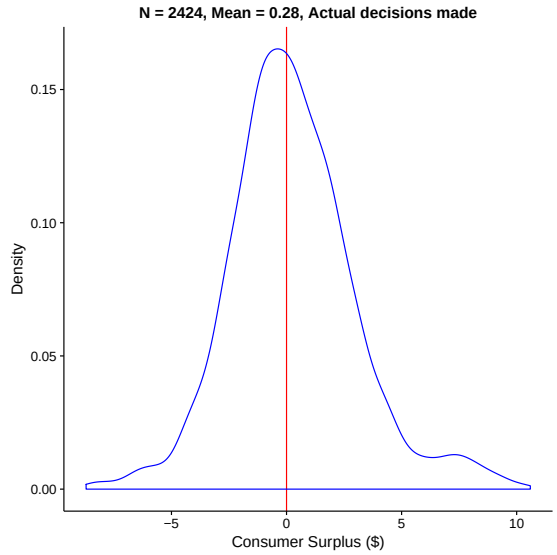
Figure 4.1: Classifying Subjects as EUT or RDU



4.2 Individual Classification and Welfare Estimation Accuracy

Whether the results presented in Figure 4.1 provide an accurate estimation of the proportions of subjects belonging to those models depends on our confidence in the classification process to correctly classify a subject as one of these four models,

Figure 4.2: Distribution of Consumer Surplus, Using Data from Harrison and Ng (2016)



as well as our confidence that the subjects in the experiments actually belong to one of the four models we test for. Our confidence that the classification process can correctly classify a subject in turn depends on the nature of the experimental instrument presented to the subject.

The degree of confidence in the classification process, and indeed most statistical tests in the economics literature, can be assessed through power analysis.[1] However, power analyses are rarely conducted in parallel with econometric estimations. McCloskey and Ziliak (1996) find that only 4.4% of the 181 papers published in *The American Economic Review* reported the power of the test they were performing. Zhang and Ortmann (2013, p. 6) review all papers published in the journal *Experimental Economics* for the years 2010-2012, and find that no paper

[1] A "power analysis" is a process for assessing the probability of type II errors for a given econometric test on data. This usually involves simulating independent variables, specifying an effect size, simulating a dependent variable given the independent variables and the effect size, and then testing how frequently the effect size can be recovered from tests on multiple repetitions of the simulated data.

stated the optimal sample size for their analyses, and only one paper mentions power as an issue.

There are some examples of experimental economists utilizing power calculations to inform their analysis or experimental designs. Rutström and Wilcox (2009) conduct a power analysis by simulating agent behavior in a Matching Pennies games and choosing payoffs that would result in the best chance of identifying the effect they sought to identify if it were there. Brown and Healy (2016, p. 2) conduct a power analysis to inform the choice of sample size for their experiments. In this instance, Rutström and Wilcox (2009) and Brown and Healy (2016) conduct a power analysis in order to influence the design of their experiment.

Wilcox (2015, p. 8) conducts Monte Carlo simulations of agents responding to a lottery battery, all of which employ the CRRA utility function, the RDU_{Prelec} probability weighting function, and the CU stochastic model. Wilcox (2015) designates four data generating processes (DGP) by specifying four parametrizations of these models and uses them to generate choice data, with each DGP making choices on the instrument 1000 times. He then estimates non-parametric RDU models for each of the 1000 choice realizations per DGP and classifies the resulting estimates into one of 5 categories, one for each of the DGPs and an additional "unclassified" category. This is an example of using power analysis to lend support to a methods and conclusions of the research. Both *a priori* power analysis, as done by Rutström and Wilcox (2009) and Brown and Healy (2016), and *ex post* power analysis, as done by Wilcox (2015), are useful for understanding the statistical support for experimental research as well as its limitations.

There are also theoretical aspects of experimental design that may increase statistical power. Loomes and Sugden (1998) (LS) utilize multiple Marshack

Machina (MM) triangles to construct lottery pairs that "provide good coverage of the space within each triangle, and also span a range of gradients sufficiently wide to accommodate most subjects' risk attitudes." Since a lottery is a point in the MM triangle, if an agent conforms to EUT and is indifferent between two lotteries, a straight line can be connected between the two lotteries in the triangle with every point on the line indicating a lottery that the agent would also be indifferent to. Thus by varying the *gradient* of the lines connecting lottery pairs, a wide range of risk attitudes, at least for agents employing the EUT functional, can potentially measured. Additionally, the use of lottery pairs on the "bottom-edge" of the MM triangle can theoretically increase the statistical power of an instrument to discriminate between subjects employing the EUT or RDU functional. Lotteries on the edges of a MM triangle space indicate that one or more outcomes have a low probability. LS (1998, p. 595) note the conclusion of Harless and Camerer (1994, p. 1285) that "nonlinear weighting of small probabilities is an important factor in explaining observed choices." These techniques of varying the gradient of lottery pairs in the MM triangle and constructing lottery pairs closer to the edges of the triangle were adopted by HN to inform the construction of their lottery battery (2016, p. 99), both to increase the precision of estimates of risk aversion parameters, and to help discern between agents employing the EUT or RDU functionals.

I interrogate the statistical power of the instrument and classification process to correctly classify subjects in HN, and the accuracy of the welfare calculations given classifications. I conduct this analysis via simulation methods similar to those defined by Feiveson (2002), which resemble an extension of the Monte Carlo analysis performed by Wilcox (2015). Feiveson (2002, p. 108) briefly describes a simulation method for determining the power of an experiment:

> [W]e contemplate a hypothetical scenario in which the identically sized experiment could be run over and over, each time collecting new data and doing a new hypothesis test. If this scenario can be adequately modeled, we may thus estimate power by simulating data from multiple replications of the experiment and simply calculate the proportion of rejections [of the null hypothesis] as an estimate of the power.

Feiveson (2002, p. 109) outlines this method in more detail, and concludes by noting " The estimated power for a [specified significance level] test is simply the proportion of observations (out of [some large number of replications]) for which the p-value is less than [the specified significance level]. "

In this framework, and given the nature of the classification process defined previously, should an EUT subject be classified as employing an RDU_{Prelec} model, this would constitute a type I error (a "false positive" of probability weighting), and should an RDU_{Prelec} subject be classified as employing an EUT model this would constitute a type II error (a "false negative" of no probability weighting). The probability of a type II error is called the "power" of the test and when researchers engage in *ex ante* power analysis, they typically aim for a power of 80% (Cohen 1988; Gelman and Loken 2014), and significance level ("p-value") of either 1, 5, or 10%. These values are based on convention, though Ronald Fisher and others disagreed with picking the same level significance for every analysis: "[...] no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." (Fisher 1956)

I simulate subjects conforming to the EUT and RDU_{Prelec} models, have these simulated subjects respond to both the lottery and insurance task, estimate the subjects' parameter sets given their responses to the lottery task, classify each

subject based on the classification process employed by HN as described in the previous section, and calculate the welfare surplus for each subject based on the winning model.[2] A simulated subject is represented by a single parameter set and an assigned model. For each model, we employ the CRRA utility function defined in (4.7) and the CU stochastic model defined in equations (4.8) and (4.9). For EUT subjects, the parameter set consists of $\{r, \lambda\}$, and for RDU_{Prelec} subjects $\{r, \alpha, \beta, \lambda\}$. The r parameter in every set is the CRRA parameter from equation (4.7) and λ is the precision parameter defined in equation (4.8). The remaining α, and β parameters relate to the probability weighting parameters of the RDU_{Prelec} model defined in equation (4.6).

For each model, we draw parameter sets from a joint uniform distribution over the parameters needed for that model, where the marginal distributions are uncorrelated.[3] For both models, the marginal distribution for r is where $r \in [-1, 0.95]$ and for λ is $\lambda \in [0.01, 0.30]$. For the RDU_{Prelec} model the marginal distribution for α and β is where $\alpha \in [0.10, 2]$ and $\beta \in [0.10, 2]$.

I draw 250,000 parameter sets for each model for a total of 500,000 simulated subjects. The number of draws from these joint distributions was chosen in an attempt to fill as much of the relevant parameter space as possible.[4] Each simulated

[2]I restrict the analysis and discussion in this chapter to only EUT and RDU_{Prelec} subjects and estimated models to improve the clarity of the discussion. However, the analysis can easily be extended to all four models considered by HN.

[3]To create uncorrelated joint uniform distributions, uncorrelated normal distributions were generated using a Gaussian copula process. The inverse normal cumulative distribution function was then applied to each marginal distribution to get uncorrelated uniformly distributed variables in the $[0, 1]$ space. These uniformly distributed variables were then stretched and shifted to fit the uniform spaces described here while retaining the 0 correlation coefficient. This process was employed to ensure that the (admittedly low) probability of accidental correlation that might occur from simply drawing from a uniform distribution directly was minimized.

[4]A limitation of choosing the same number of draws for each model is that the square uniform space for the EUT model will have smaller gaps than the hypercubic space of the RDU_{Prelec} model. The smaller the gaps between parameter sets in their joint space, the better the prediction

subject uses the parameter set and model assigned to it to calculate the choice probabilities for each option in each lottery pair of the lottery task and the insurance task. A random number is drawn from a univariate uniform distribution, and if the choice probability calculated for the A option was greater than the random number, the subject chooses A, otherwise they choose B. This process ensures that subjects' choices are made probabilistically with respect to the subjects' model and parameter set.[5]

After the subjects have made choices, each of the models we consider is estimated for each subject on the choices made in the lottery task. Any model which didn't converge with a gradient close to 0 and a negative definite Hessian matrix or converged on parameters outside of exclusionary rules defined in the previous section was dropped from consideration. Each subject was then classified based on the classification process defined in the previous section using a 5% significance level. If no model met the consideration criteria, the subject was classified "NA". The welfare surplus of the choices made on the insurance task are then calculated using the parameters of the winning model.

This process of classification simulation differs from that employed by Wilcox (2015) in that I simulate a total of 500,000 DGP, each producing a single set of choice data, whereas Wilcox (2015) simulates 4 DGP, each producing 1000 sets of choice data. The approach of Wilcox (2015) allows for individual DGP to be characterized by multiple sets of choices, while the approach I employ allows us to see how the power of the instrument changes with resect to a wide range of DGP.

accuracy of classifying subjects for the parameter sets that exist in the empty space.

[5]Consider a choice probability for option A calculated to be 0.90, and therefore the choice probability for option B is 0.10. A random number drawn from a univariate uniform distribution has a 90% chance of being below or equal to 0.90, so option A would be chosen 90% of the time by the simulated subject.

The limitation of the approach I employ of only generating one choice data set per DGP is mitigated by the large number of simulated subjects and the statistical methods employed to predict classification probabilities described below.

4.2.1 Harrison and Ng (2016) Classification Power

Consider a simulated subject X which employs the EUT model with a CRRA parameter of 0.5 and a λ value of 0.1. Additionally consider the 2-dimensional parameter space Z, where CRRA $\in (0.475, 0.525)$, $\lambda \in (0.09, 0.11)$ and the parameters are uncorrelated in the space. There is only one choice dataset for subject X, but there are 430 datasets in the space Z given the number of simulations conducted. We could calculate the average number of subjects in space Z that are classified as employing the EUT model, and use this statistic as an approximation of the probability of correctly classifying subject X. This approach to approximating the classification probabilities for a single set of parameter values based on an average of some number of "nearest neighbor" parameter values is useful if the range of parameter values chosen to average over is small and there are many data points in the range. We could potentially improve this approach by fitting a probit or logit model to the data in Z with the classification as EUT being the dependent variable, and the parameter values as the independent variables, and predict a classification probability for subject X. We could then consider a different subject Y with a new Z space distributed around its parameters and repeat the process.

These approaches are naïve versions of other "smoothing" approaches such as local regressions (LOESS) due to Cleveland (1979) and Cleveland, Grosse, Shyu, Chambers and Hastie (1992), and generalized additive models (GAM) due to Hastie and Tibshirani (1986). These approaches account for certain edge cases which

makes them more attractive than the naïve approaches.[6]

Both the LOESS and GAM approaches would allow for predictions of classification probabilities for any set of parameters covered by the range of parameters simulated, but the GAM approach is utilized throughout. The GAM approach allows us to predict the probability that a subject employing a particular model is classified as EUT, RDU_{Prelec}, or unable to be classified.[7] First we separate the data into subsets based on the models the simulated subjects actually employ, either EUT or RDU_{Prelec}. For each pooled group we fit a GAM model predicting whether the subjects were classified as EUT, RDU_{Prelec}, or unclassified.

$$\begin{aligned} (winner = N|A = EUT) &= s(r) + s(\lambda) \\ (winner = N|A = RDU_{Prelec}) &= s(r) + s(\alpha) + s(\beta) + s(\lambda) \end{aligned} \tag{4.17}$$

where N is one of EUT, RDU_{Prelec}, or "NA" and $s(\cdot)$ indicates some smooth, potentially non-linear function of its arguments.

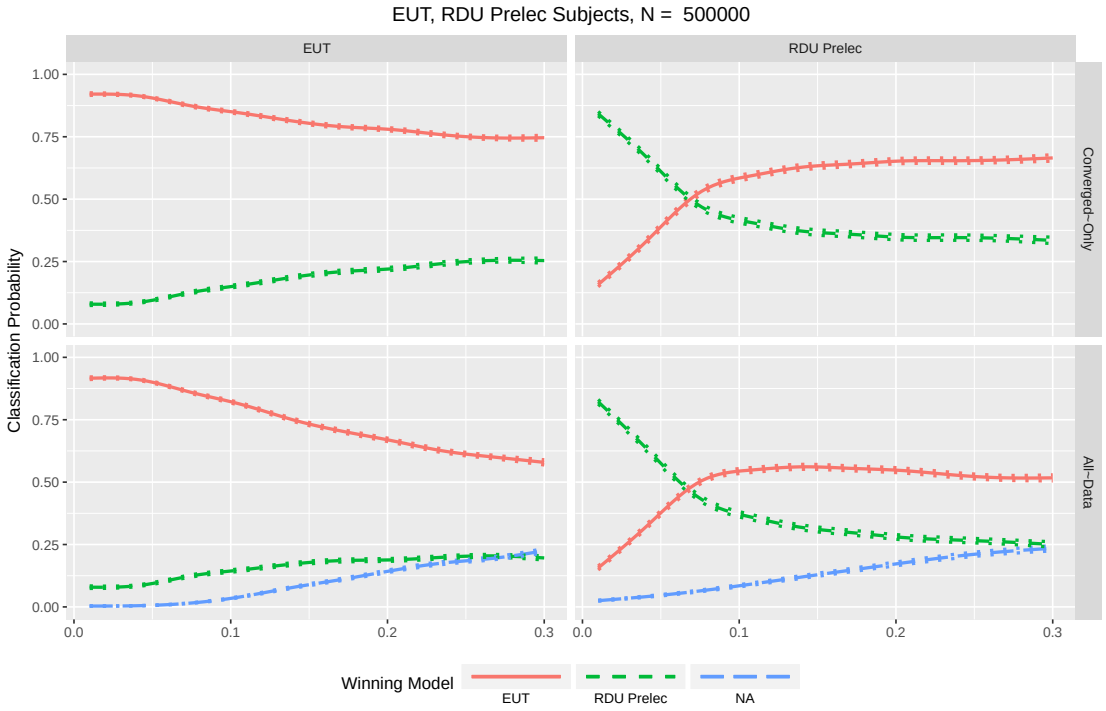
The dependent variable in each of the GAM models in equation (4.17) is either 1 if the subject was classified as model N, or 0 if the subject was not. The independent variables in each model are smooth functions of the actual parameter

[6]Given that the parameter space is very dense, with 250,000 points per model, and that the parameters are uncorrelated in the space, the naïve averaging approach is likely to make predictions similar to that of the LOESS and GAM approaches for most of the data. However, there are several reasons to prefer either the LOESS or GAM approaches over the naïve simple averaging approaches for our purposes. The properties of "smoothers" at the edges of the parameter space are of particular interest. The naïve averaging approach works well when the point of interest X is close to the midpoint of the range Z, but as X approaches the edge of the full parameter space, a Z space can no longer be constructed with X as the midpoint, leading to estimates for X being biased towards (or equal to) the estimates of the actual midpoint of Z. The LOESS and GAM approaches handle these cases, in part by weighting parameters based on their distance to the point X. Since the λ, α and β parameters must all be greater than 0, and we simulate parameter sets close to these limits, we need an approach that is useful at the edges of parameter spaces.

[7]If neither the EUT nor the RDU_{Prelec} model passed the exclusionary rules defined in the previous section, no model would be declared the winner and the subject would be classified as "NA."

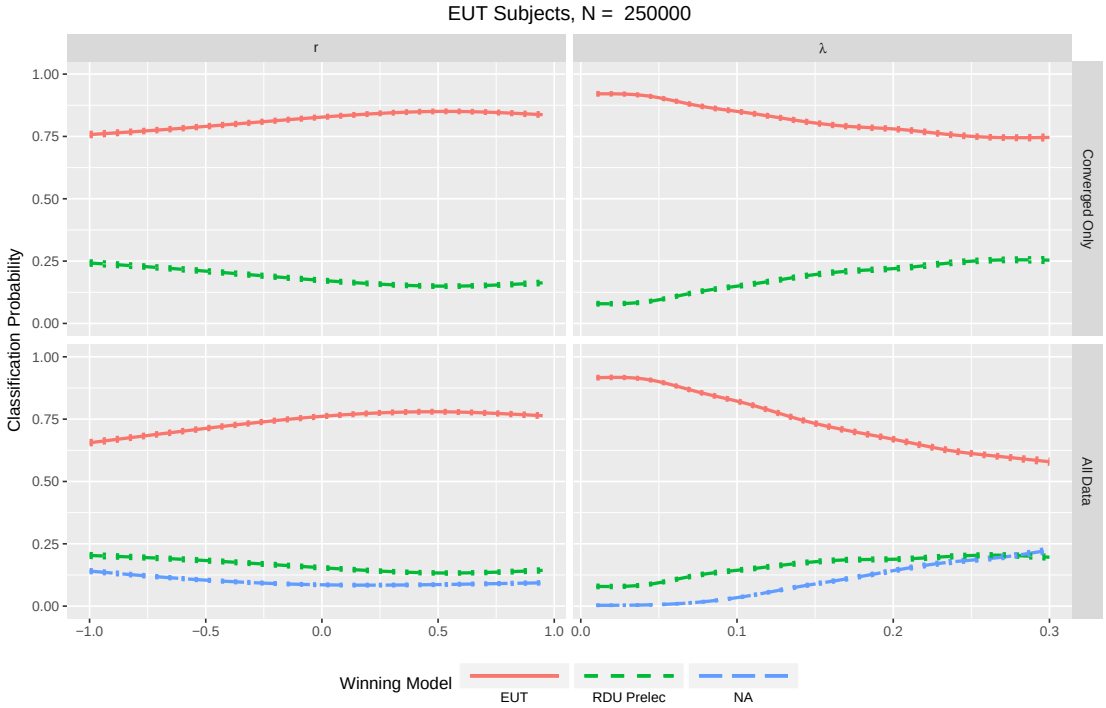
values the subject employs. For every model, each parameter gets its own smooth function. Thus, 3 GAM models are fitted for each of the two model types in the population, resulting in 6 fitted models in total. I then repeat this process but drop subjects that were unclassified from the data before fitting the models. This results in 4 additional models. Given the fitted models and a parameter set for a model type, we can use a fitted GAM model to predict the probability that a subject with the given parameter set will be classified as any of the N models. The results of this fitting process are presented in Figures 4.3, 4.4, and 4.5.

Figure 4.3: Probability of "Winning" for Given λ Values



In Figures 4.3, 4.4, and 4.5 the X-axis is the simulated subjects' values of the parameter for that plot, and the Y-axis is the probability that a given model was declared the winner. In each Figure the solid red line indicates the estimates for the

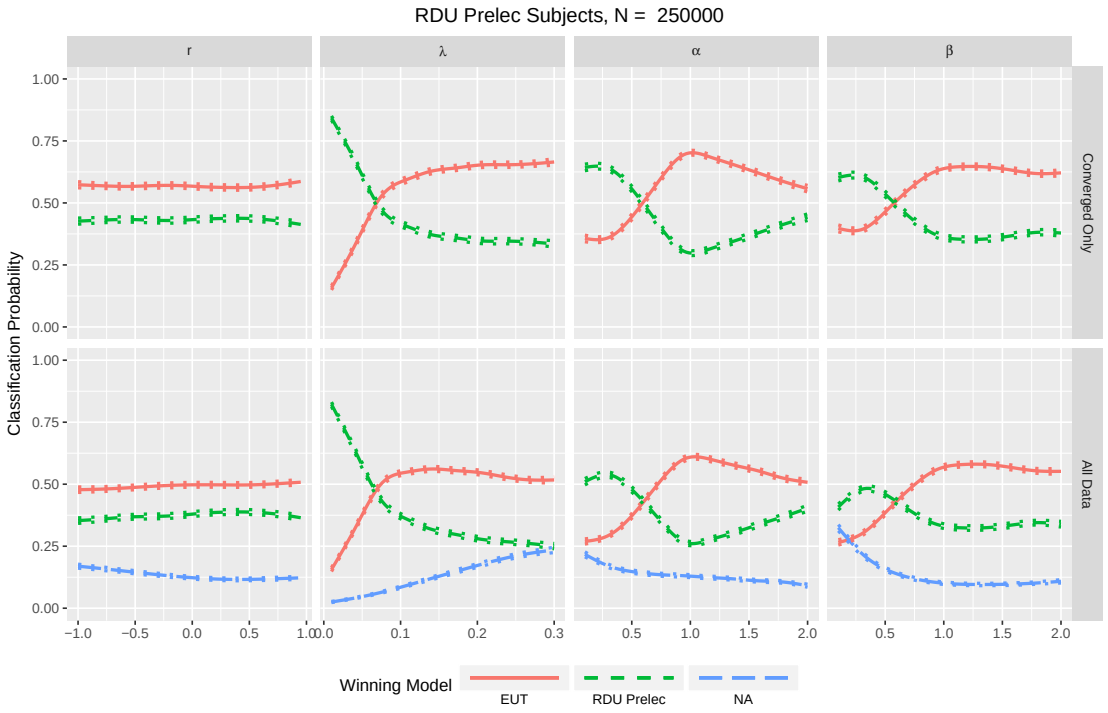
Figure 4.4: Probability of "Winning" for EUT subjects



EUT model, the short dashed green line indicates the estimates for the RDU_{Prelec} model, and the long dashed blue line indicates the estimates for non-convergence or exclusion. In all figures, the 95% confidence interval is given by the dotted lines surrounding the lines given above. In each Figure, the second row contains estimates derived from all the subjects, while the first row only contains estimates derived from subjects that were classified as either EUT or RDU_{Prelec}.

In Figure 4.3 the first column contains estimates for EUT subjects, while the second column contains estimates for RDU_{Prelec} subjects. The X-axis of this figure is the value of the λ parameter. Thus, the top-left plot shows the probability of an EUT subject with a converged model being classified as either EUT or RDU_{Prelec} for a range of λ values. In Figures 4.4 and 4.5, the column titles indicate the

Figure 4.5: Probability of "Winning" for RDU_{Prelec} subjects



parameter given on the X-axis. Thus, in Figure 4.4, the top-left plot shows how the probability of an EUT subject with a converged model is classified as either EUT or RDU_{Prelec} for a range of CRRA values. In Figure 4.5, the bottom-right plot shows how the probability of an RDU_{Prelec} subject is classified as either EUT, RDU_{Prelec}, or "NA" for a range of β values.

In Figure 4.4 and in the left column of Figure 4.3, the red solid line shows the probability of EUT subjects being correctly classified as EUT. In Figure 4.5 and in the right column of Figure 4.3, the green dotted line shows the probability of RDU_{Prelec} subjects being correctly classified as RDU_{Prelec}.

The results presented in Figures 4.3, 4.4 and 4.5 offer both surprising and intuitive results. In Figure 4.3 we see that the probability of EUT subjects being

misclassified as RDU_{Prelec}, RDU_{Prelec} subjects being misclassified as EUT, and either type of subject being unclassified, increases with λ. This is intuitively reasonable. As the λ parameter increases, the likelihood that a subject makes a choice error increases. For EUT subjects, these choices errors can present as probability weighting when there is none, and for RDU_{Prelec} subjects these choice errors can present as linear probability weighting. Another way to characterize this effect is to say that as λ increases, the noise in the data increases. Indeed, as $\lambda \to \infty$ choice probabilities for every option are equal, resulting in totally random data. The more noise there is in the data, the lower the likelihood of the optimizer converging on reasonable, or any, estimates, and the greater the likelihood that any latent process will be identified as another.

In the third and fourth columns of Figure 4.5 we have additional intuitive results. We can see in these columns that the probability of an RDU_{Prelec} subject being classified as EUT peaks when the probability weighting parameters approach the value of 1, and diminishes as these parameter values move away from 1. Since the RDU_{Prelec} model nests EUT when $\alpha = \beta = 1$, we should expect the likelihood of misclassification to increase around these values. It appears the α parameter plays a more decisive role in the classification probability for the range of parameter values we consider; the probability of a RDU_{Prelec} subject being classified as RDU_{Prelec} drops at a greater rate as the α parameter approaches 1 than as the β parameter approaches 1 from either the left or the right.

In Figure 4.4 we see that the probability of an EUT subject being classified as EUT is greater for values of CRRA > 0 than for values of CRRA < 0, though only modestly so. Values of CRRA > 0 indicate risk aversion in an EUT model, and the design of the HN lottery instrument placed more emphasis on identifying

degrees of risk aversion than identifying degrees of risk seeking (CRRA values < 0) in EUT subjects. Similarly, in Figure 4.5 we see that the CRRA parameter has very little effect on the probability of a RDU_{Prelec} subject being correctly classified as RDU_{Prelec}. Since it is the probability weighing function that defines RDU_{Prelec} as being different from EUT, it should not be surprising that the utility parameter has little effect on on the probability of RDU_{Prelec} subjects being correctly classified.

However, the relatively low probability with which RDU_{Prelec} subjects are correctly classified as RDU_{Prelec} over a wide range of parameters is surprising. Looking at Figure 4.5 we see that for most of the parameter values considered, the probability of an RDU_{Prelec} subject being correctly classified as RDU_{Prelec} is below 50% and that for most of these values, it is more likely that an RDU_{Prelec} subject is classified as EUT than as RDU_{Prelec}.

The statistical results presented here generally show wide variation of the power of the HN instrument to correctly classify subjects as employing either the EUT or RDU_{Prelec} functionals across parameter spaces. This wide variation in power for different DGP suggests that power analysis conducted on only several DGP, as is done in Wilcox (2015) for example, should be extended to incorporate more DGP across the ranges of parameters an experimenter may expect real subjects to employ.

4.2.2 Harrison and Ng (2016) Insurance Task Welfare Expectations

The probabilities provided in Figures 4.3, 4.4 and 4.5 are useful for describing the degree of success of the classification process has in correctly identifying the model employed by a subject if the subject employs one of the two models considered

here. The classification process itself, however, is only useful to economists insofar as it provides us with a model that allows us to make normative characterizations of subjects' choices. Given our simulation process, we can measure the success of the classification process in normative terms by calculating the difference in the estimated welfare surplus of the choices made in the HN insurance task against actual welfare surplus for each subject.

Utilizing the definition of accumulated welfare surplus given by equation (4.14), we follow HN (2016, pp. 110-111) and bootstrap the estimated welfare surplus of the subjects. We generate 500 random draws from a multivariate normal distribution using the point estimates of the parameters of the winning model as the means of the marginal distributions, and the inverse of the estimated Hessian matrix as the covariance matrix.[8] With each draw we calculate equation (4.14) and define the estimated welfare surplus as the average of these 500 calculations. Therefore the difference between the estimated welfare surplus, given by this bootstrap method, and real welfare surplus, which we can observe directly for the simulated subjects, is given by:

$$\text{WSD}_N = \Delta W_{iT}(\hat{\Omega}_N) - \Delta W_{iT}(\Omega) \tag{4.18}$$

where N is the model the subject has been classified as employing, Ω is the set of parameters that define the utility function actually employed by subject i, and $\hat{\Omega}_N$ is the set of estimated parameters for model N for subject i. Values of 0 for equation (4.18) indicate that the subject's estimated welfare surplus equals the

[8]All the probability weighing parameters and the λ parameter are restricted mathematically to be greater than 0. In the estimation process, this was accomplished by exponentiating the raw parameter values passed by the optimizer to the likelihood function. When generating the multivariate normal distribution described, we use the raw parameter estimates to generate the distribution and exponentiated the marginal distributions of the parameters that are restricted to be greater than 0. Thus, these resulting marginal distributions are actually log-normal distributions.

subject's real welfare surplus, and thus the welfare estimates are "accurate" in terms of their approximation of the real welfare surplus. Since equation (4.18) is based on equation (4.14), and this equation is based on the *CE* of the lotteries in a lottery pair, the units of the welfare surplus difference (WSD) are the same monetary units as the *CE*s of the lotteries. If the subject has been misclassified, Ω and $\hat{\Omega}_N$ will not represent the same set of parameters.

Just as we predicted probabilities of classification in equation (4.17), we can predict the difference in estimated welfare surplus and real welfare surplus given by equation (4.18). GAM models are utilized once more to allow predictions across the range of parameter values simulated. The data are first separated into subsets based on the models the simulated subjects actually employ, either EUT or RDU_{Prelec}, and then for each pooled group we fit a GAM model predicting the WSD given by equation (4.18) as a function of the parameters the subject actually employs.

$$\begin{aligned}(\text{WSD}_{N,M}|M = EUT) &= s(r) + s(\lambda) \\ (\text{WSD}_{N,M}|M = RDU_{Prelec}) &= s(r) + s(\alpha) + s(\beta) + s(\lambda)\end{aligned} \tag{4.19}$$

where N indicates the model that the subject was classified as, M indicates the model the subject actually employs, and $s(\cdot)$ indicates some smooth, potentially non-linear function of its arguments. A model is fitted for each combination of the 2 M models and 2 N models, and so 4 models are fitted in total. Additionally, given our estimates of the probability of EUT and RDU_{Prelec} subjects being classified as employing EUT or RDU_{Prelec} respectively, we can calculate point estimates for the expected WSD by multiplying the probabilities presented in the top row of Figures 4.4 and 4.5 with the predicted WSD given by equation (4.19). The WSD predictions

for subjects that were classified as either EUT or RDU_{Prelec}, as well as the expected WSD, are presented in Figures 4.6, 4.7, and 4.8. Since the subjects which didn't converge on either EUT or RDU_{Prelec} (labeled "NA" previously) didn't produce estimates with which we can make welfare calculations, they are not plotted.

Figure 4.6: Welfare Surplus Difference of "Winning" Models for Given λ Values

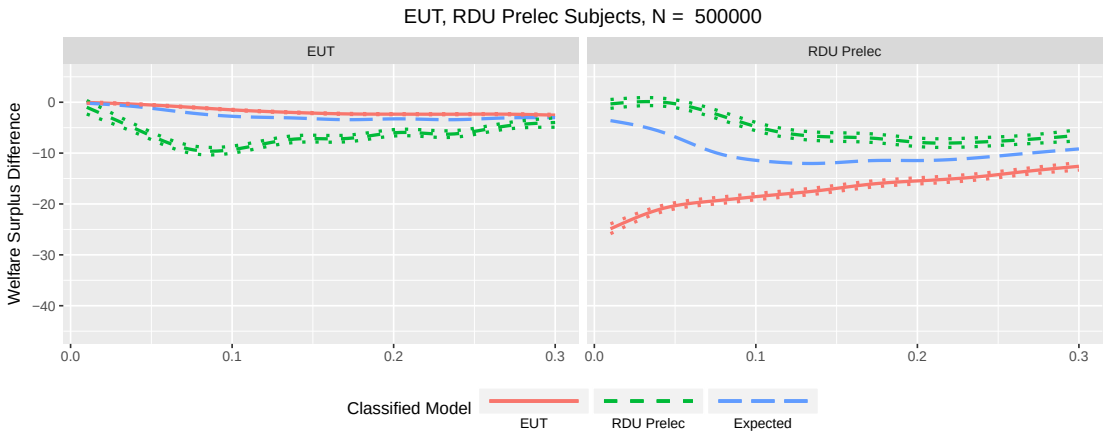
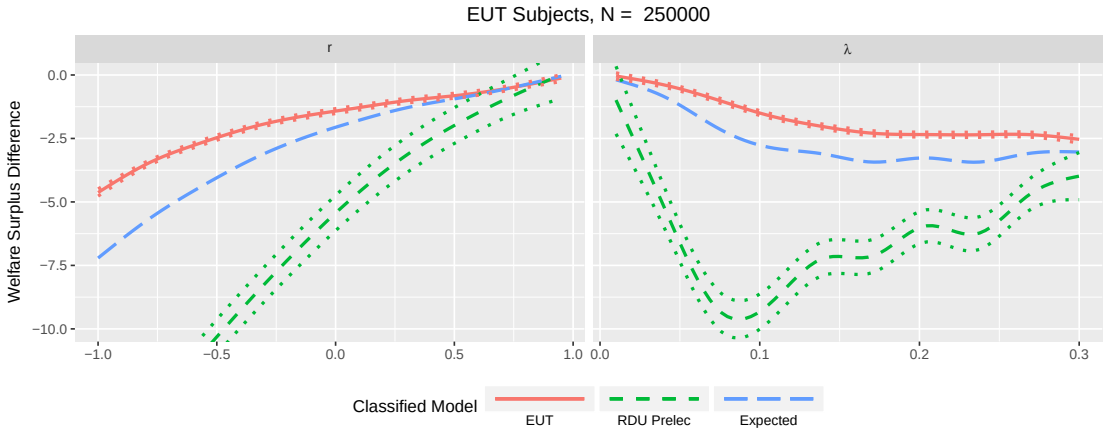
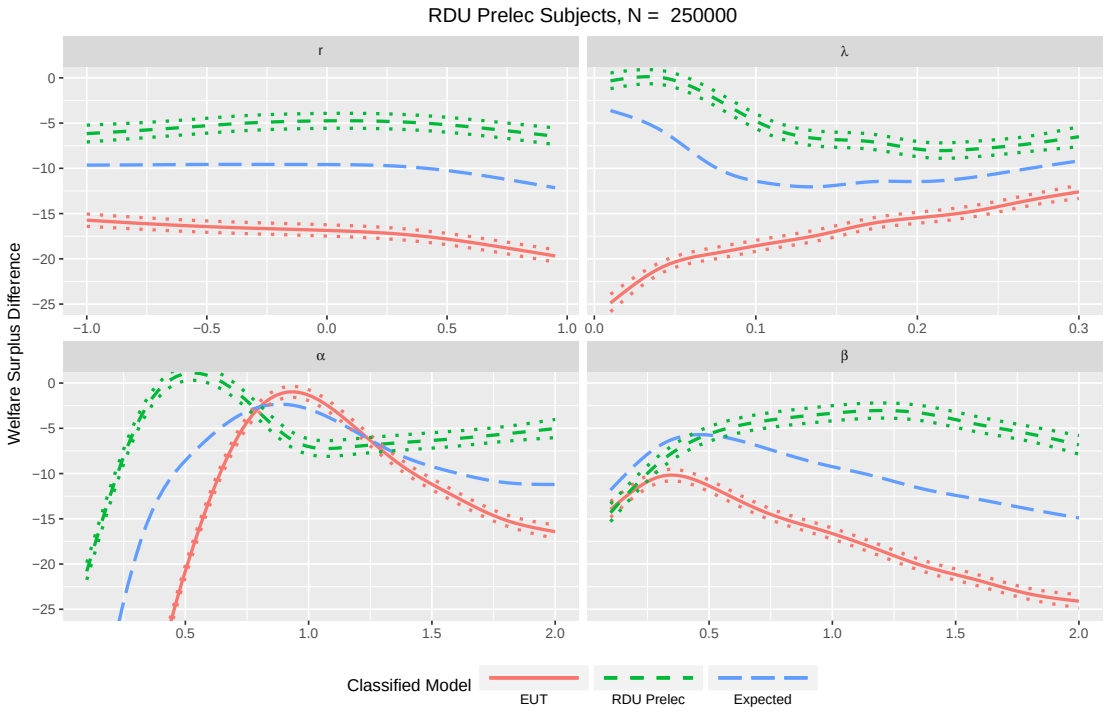


Figure 4.7: Welfare Surplus Difference of "Winning" Models for EUT subjects



The solid red line again represents subjects that were classified as EUT, and the short dashed green line represents subjects that were classified as RDU_{Prelec}.

Figure 4.8: Welfare Surplus Difference of "Winning" Models for RDU_{Prelec} subjects



In these figures, however, the long dashed blue line represents the expected WSD.

Assessment of how the classification process relates to the welfare surplus of the subject being classified is in many ways more important than the accuracy of the process itself. This is because economists distinguish themselves from decision theorists by making normative statements about how an individual's choices relate to their economic well-being. The accuracy of the classification process is valuable only because it can aid in the accuracy of the normative statements we can construct using this process. Leamer (2012, p. 25) makes a similar statement when discussing the general fallibility of macroeconomic models: "[O]ur goal as economists is not soundness, but usefulness."

Looking at Figure 4.7, which depicts EUT subjects, for values of CRRA greater

than 0.5, shown in the left plot, the confidence intervals of subjects classified as EUT or RDU_{Prelec} overlap, indicating that there isn't a noticeable difference in WSD between correctly classified and misclassified subjects in this range. In Figure 4.8, which depicts RDU_{Prelec} subjects, we see that for values of α just greater than 1 or just less than 1, shown in the bottom left plot, the lines showing the predicted WSD of subjects classified as either EUT or RDU_{Prelec} overlap briefly. These two cases indicate that for some parameter values employed by subjects of either model, misclassification is costless in terms of WSD. Additionally, in Figure 4.8 we can see that of the RDU_{Prelec} subjects that have α values very close to 1, shown in the bottom left plot, the subjects that have been classified as EUT instead of RDU_{Prelec} have WSD that are somewhat closer to 0 than the subjects that had been classified as RDU_{Prelec}. The finding that misclassified subjects in these cases have WSD relatively close to 0, or closer to 0 than for correctly classified subjects, indicates that even though the classification process has *not* been accurate for these subjects, it nonetheless *can* be useful when used to characterize the welfare surplus of subjects' choices in the insurance task. This conclusion will be revisited later.

However, we can also see that for wide ranges of parameter values, misclassified subjects have a WSD that is significantly different from 0 and is farther from 0 than for correctly classified subjects. The cost of misclassification is particularly great for RDU_{Prelec} subjects. In Figure 4.6, comparing RDU_{Prelec} subjects misclassified as EUT, shown in the right plot as the solid red line, to EUT subjects misclassified as RDU_{Prelec}, shown in the left plot as the dotted green line, we can see that the WSD is more negative for misclassified RDU_{Prelec} subjects than for misclassified EUT subjects across the entire range of λ. Looking at the bottom left plot of Figure 4.8, we can see that as the α parameter approaches 0, RDU_{Prelec} subjects

that have been misclassified as EUT have WSD values that increasingly diverge from 0, indicating an increasing cost of misclassification. Looking at the bottom right plot of Figure 4.8, we see generally that as β increases past 1, the subjects that have been incorrectly classified as employing an EUT model also have WSD values that increasingly differ from 0, but this divergence is of roughly the same magnitude seen bottom left plot of Figure 4.8 as α increases above 1.

That subjects actually employing a RDU_{Prelec} model are badly characterized by an EUT model when they have probability weighting parameters that differ greatly from 1 should not be surprising. Probability weighting is what distinguishes RDU models from EUT models, and so when this is ignored by classifying an RDU_{Prelec} subject as EUT, the consequences in terms of welfare surplus estimates can be meaningful. On the other hand, RDU models nest EUT as a special case, and so when EUT subjects are misclassified as RDU_{Prelec} there is the possibility that even though the estimated probability weighting parameters are statistically significantly different from 1, the magnitude of this difference is small enough to not matter as much in terms of welfare surplus.

4.3 Alternative Approaches for Welfare Prediction

The analyses thus far constitute *ex post* power analyses of the experimental instrument and classification process employed by HN, and an analysis of the expected welfare characterizations that can be made with this classification process. The power analysis aspect of this process constitutes a statistical inquiry into an experimental protocol and is similar to other *ex ante* and *ex post* power analyses. The welfare characterization aspect of this analysis constitutes the economic inquiry

into this experimental protocol. Both inquires are important, but making accurate predictions or characterizations about the welfare consequences of choices by economic agents should be of greater importance to economists than the descriptive accuracy of the model used to derive these calculations.[9]

The two inquiries are related as noted by HN (2016, p. 105). A model is needed in order to make calculations of consumer surplus and thus we need a reasonable method for selecting a model on which to base these calculations. However, if our objective is to generate accurate welfare characterizations, and not *necessarily* accurate model classifications, then we should explore how different experimental designs, model specifications, and model selection processes influence the accuracy of welfare characterizations. For instance, the selection of the number and type of lottery pairs should be influenced by how they result in more accurate welfare predictions in the choice domain that is welfare relevant to the experimenter; the insurance task in the case of HN.

These kind of enquires into how differing experimental methods affect the accuracy of welfare characterizations are themselves experiments of a kind. In this section we propose two modifications to the experimental protocol employed by HN and investigate how they differ in terms of expected welfare surplus predictions. The first of these proposals is a recommendation that would be familiar to any statistician: increase the sample size per subject by increasing the number of lottery pairs in the lottery instrument used in estimation. The second proposal is to forego any attempt to accurately classify subjects as EUT or RDU and instead use the fitted RDU_{Prelec} models when they have passed the exclusionary rules set by HN, and use non-excluded EUT models otherwise.

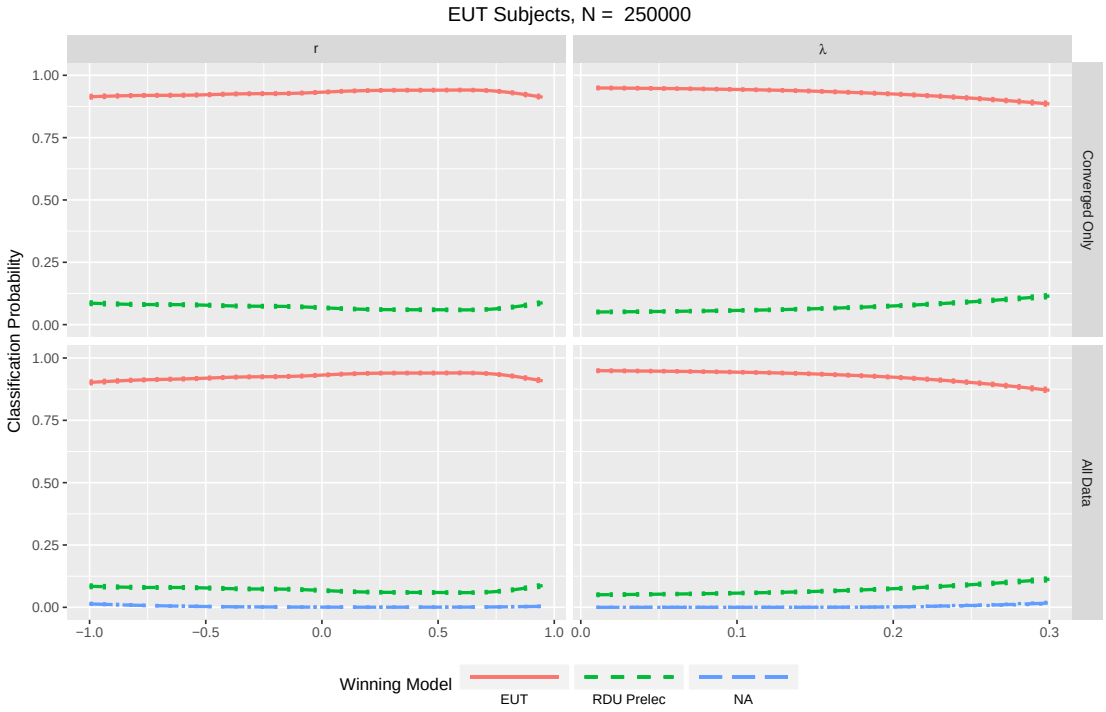
[9] I make this argument against the Random Preferences stochastic model in Chapter 2.

For the second proposal, we utilize the choice data and model estimations from the simulation process described previously and simply change the critical value for the non-linear Wald test of linear probability from 0.05 to 1 so that the null hypothesis of equivalence with EUT is rejected in every case. This proposal will be referred to as the "Default" approach. For the first proposal, however, we use the same simulated subjects used in all the analyses thus far, but have them each respond to the HN lottery instrument 3, 5, 7, 9, 11, and 13 times instead of once. This results in 240, 400, 560, 720, 880, and 1040 choices per subject, respectively. The estimation procedure, application of exclusionary rules, and classification process is then applied to these new choice data to select a winning model for each subject. This proposal will be referred to as the HN_C approach, and when referring to individual repetitions the "C" will be replaced by the number of lottery pairs for the given repetition. Thus HN_{240} refers to the instrument where the subject made 240 choices, HN_{400} to the instrument where the subject made 400 choices, and so on. I additionally refer to the original lottery instrument proposed by HN as the HN_{80} instrument, as it has 80 choices per subject.

The parameter estimates of the winning model from each approach are used to calculate the welfare surplus of the subject in the insurance task as before. Thus, the HN_C approach changes the experimental instrument used to estimate models, leaving the exclusionary rules and classification process unchanged, while the Default approach leaves the experimental instrument and exclusionary rules unchanged and alters the classification process. The results of the classification process for the HN_{1040} instrument are presented in Figures 4.9, 4.10, and the estimated WSD results are presented in Figures 4.13 and 4.14. The probability of correctly classifying subjects for each of the HN_C instruments is presented in

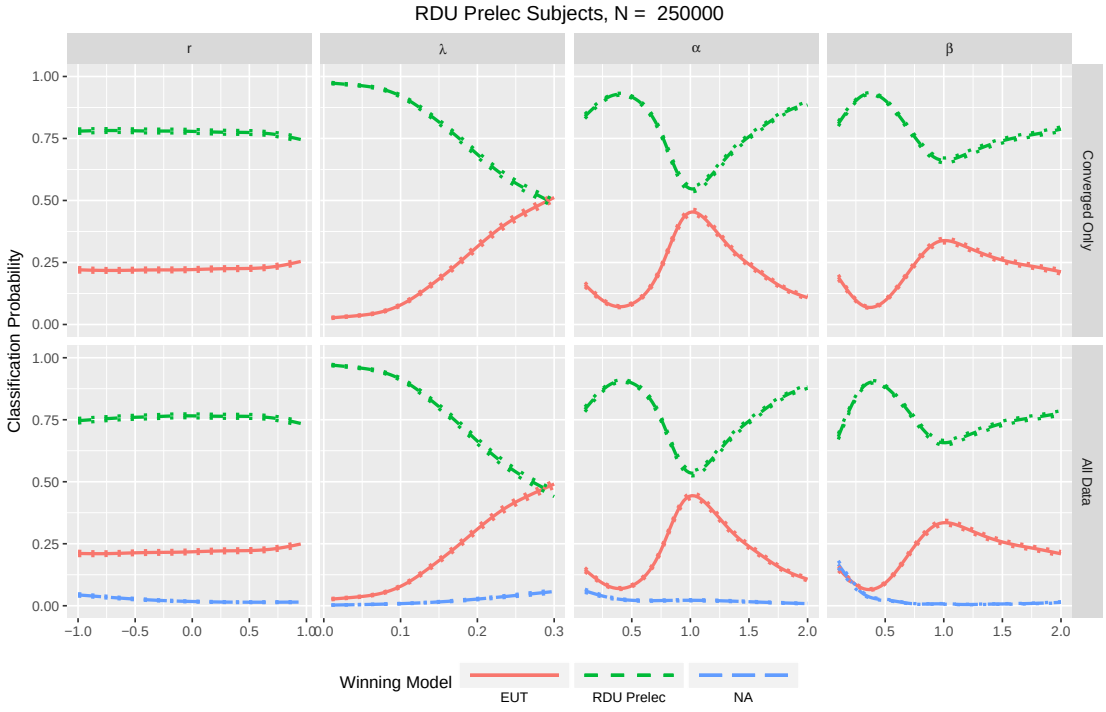
Figures 4.11 for EUT subjects and 4.12 for RDU_{Prelec} subjects. The estimated WSD for the Default approach are given in Figures 4.15 and 4.16. The plots of the expected WSD for the Default approach, and the HN_C for all C are given in Figures 4.17 and 4.18.

Figure 4.9: Probability of "Winning" for EUT subjects
HN_{1040} Approach



I initially present the HN_{1040} instrument as a potential limiting case of a sample size increase. Clearly, 1040 choices per subject lies beyond the feasible number of lottery pairs to present to subjects in any one session, but this large number of lottery pairs has attractive statistical properties. Looking first at the classification power of the HN_{1040} instrument in Figures 4.9 and 4.10, we see that this instrument has significantly improved power overall, and that the variation of power over the

Figure 4.10: Probability of "Winning" for RDU_{Prelec} subjects
HN_{1040} Approach



range of parameter values follows much the same pattern as the original HN_{80} instrument. In Figure 4.9 we see that classification power is largely uniform across the entire range of parameters considered, with some small increase in the probability of EUT subjects classified as RDU_{Prelec} as λ values increase, as seen in the right tail of the lines in the right column plots, and small increase in the probability of EUT subjects being classified as EUT as the CRRA value increases, as seen in the right tail of the left column plots. The probability of correctly classifying EUT subjects as EUT is greater under the HN_{1040} instrument than the HN_{80} instrument across the entire range of parameters considered. The rate of non-convergence in the HN_{1040} instrument, however, is also noticeably different in

the HN_{1040} instrument; it is not perceptibly different from 0 across the entire range of parameters considered.

In Figure 4.10 we see that classification power of the HN_{1040} instrument follows the patterns of the original HN instrument, but more rapid changes in the slopes of the lines for each parameter except the CRRA parameter. The probability of an RDU_{Prelec} subject being misclassified as EUT increases rapidly as the λ parameter increases, as seen in the second column, and as either of the probability weighting parameters approach 1 from either side, as seen in the third and fourth columns. The probability of correctly classifying RDU_{Prelec} subjects however is again universally higher under the HN_{1040} instrument, and the probability of non-convergence is nearly 0 for almost the entire range of parameters considered. The probability of non-convergence increases somewhat as λ increases, as α and the CRRA parameters decrease, and increases rapidly as the β parameter goes below 0.5.

It should not be a surprise that we should generally see the same patterns as before, but with significantly greater probabilities of correctly classifying subjects across the whole ranges of parameter values considered. The probabilities of type I and type II errors generally decrease with sample size in econometric tests with consistent estimators, and so we should expect this result when we increase the per-subject sample size 13-fold. The patterns of how the probabilities change with parameters values are much the same as before is due to the lottery pairs, considered models, and classification process being identical. With a different composition of the type of lottery pairs, we would expect to see different probability patterns, perhaps even seeing increasing power in the parameter ranges we would expect to see from real subjects.

Figure 4.11: Probability of Correct Classification for EUT subjects HN_C Approaches

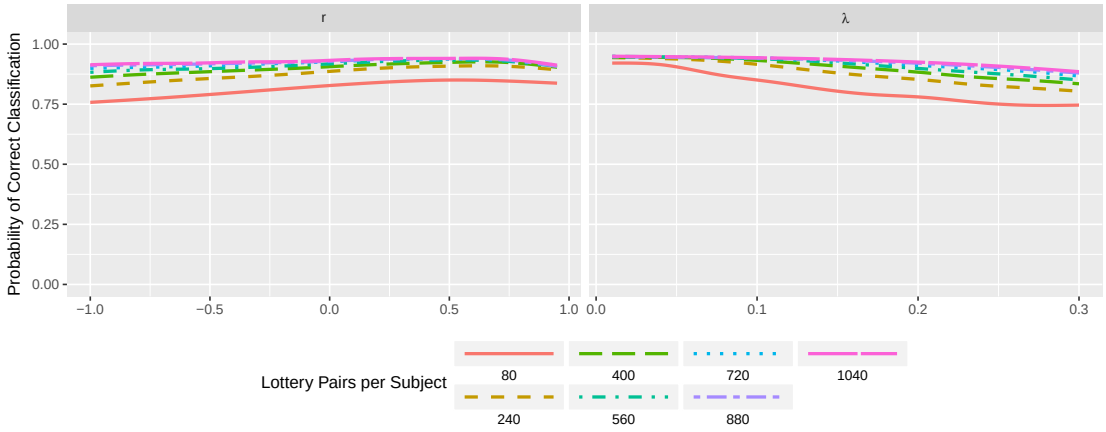
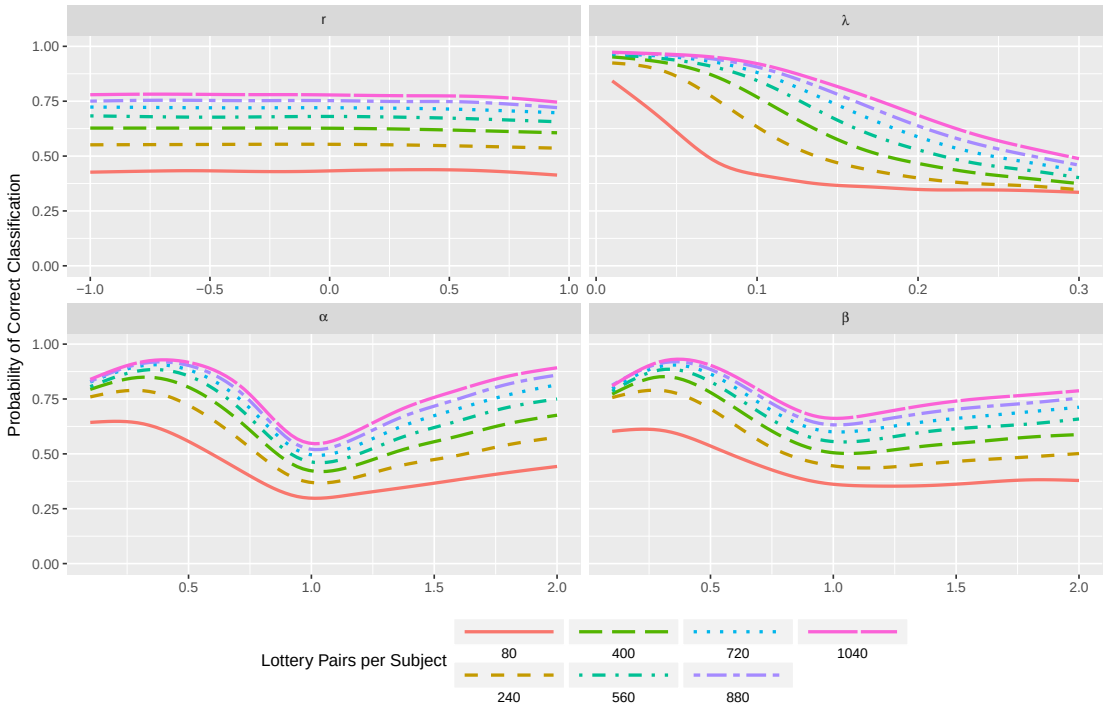


Figure 4.12: Probability of Correct Classification for RDU_{Prelec} subjects HN_C Approaches



In Figures 4.11 and 4.12 we see the point estimates of the predicted probability of correctly classifying EUT subjects and RDU_{Prelec} subjects, respectively, for each of $C \in \{80, 240, 400, 560, 720, 880, 1040\}$. Figures 4.11 and 4.12 show correct classification probabilities (CCP) among converged subjects only. In Figure 4.11 we can see that as the number of lottery pairs in the instrument increases, given by the different colored lines in the plots, the CCP for EUT subjects increases across the entire range of parameters considered. Just as we saw for the HN_{1040} instrument in Figure 4.9, and the HN_{80} instrument in Figure 4.4, for all C instruments, the CCP decreases with the λ parameter, shown in the right plot, and increases with the CRRA parameter, shown in the left plot. Interestingly, for instruments with $C \geq 240$, given by the lines that are not solid and red, there does not appear to be much difference in the CCP for values of λ less than 0.1, as seen in the right plot, and values of CRRA greater than 0.75, in the left plot.

In Figure 4.12 we again see that as the number of lottery pairs in the instrument increases, given by the different colored lines in the plots, the probability of correctly classifying RDU_{Prelec} subjects increases across the entire range of parameters considered. However, the differences across the C instruments in the CCP for RDU_{Prelec} subjects is more exaggerated than for EUT subjects. The difference in CCP across the C instruments is particularly pronounced for values of $\lambda < 0.15$, given in the top right plot, and increasingly pronounced as values of α and β diverge from 1, as shown in the bottom left and right plots, respectively. Although, as the α and β parameters approach the value of 0, the limit of the RDU_{Prelec} function, the CCP begins to converge for $C \geq 240$.

The value of this increase in classification power, as stated earlier, lies in the superior leverage we gain for making better welfare characterizations. We can

Figure 4.13: Welfare Surplus Difference of "Winning" Models for EUT subjects HN_{1040} Instrument

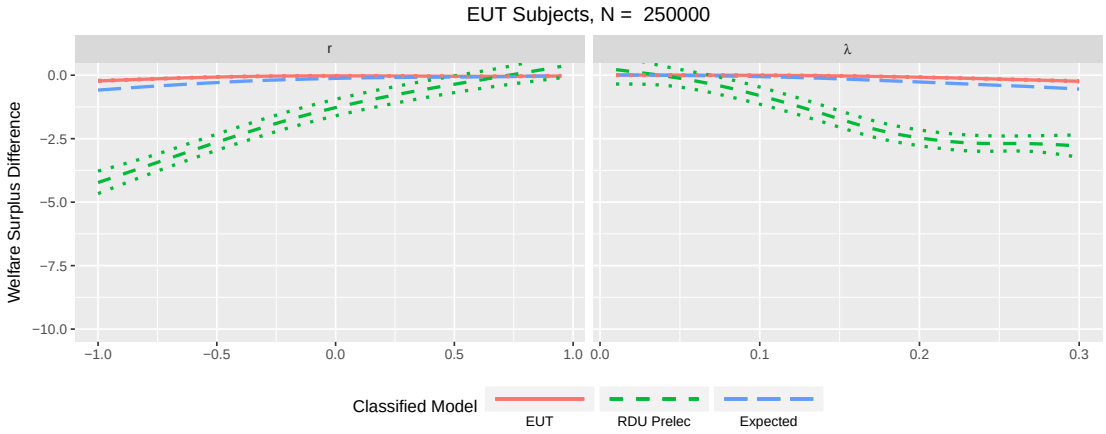
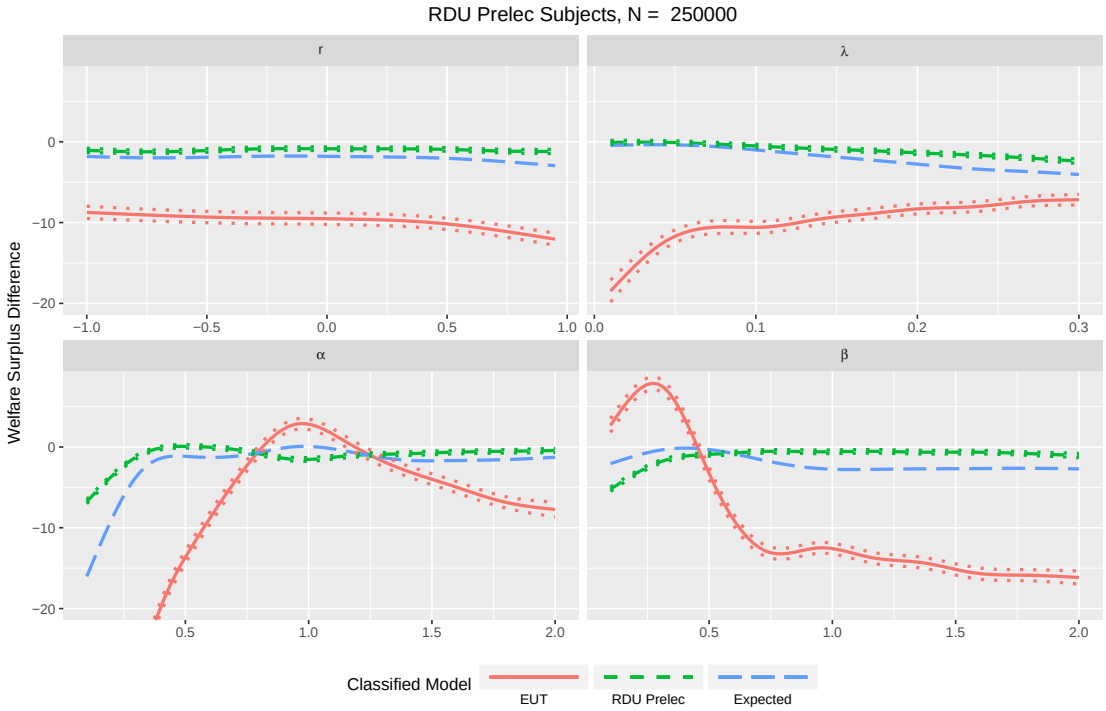


Figure 4.14: Welfare Surplus Difference RDU_{Prelec} subjects HN_{1040} Instrument



see the estimates of welfare surplus given the classification based on the HN_{1040} instrument in Figures 4.13 and 4.14. In Figure 4.13 we see that for EUT subjects classified correctly as EUT, given by the solid red line, the expected WSD is imperceptibly different from 0 across much of the range of parameters considered. In addition, even though EUT subjects classified as RDU_{Prelec} have generally worse WSD estimates, given the high likelihood of EUT subjects being correctly classified, the expected WSD is also very close to 0 for much of the parameter ranges considered. This indicates that not only is the classification process much more accurate, but the parameter estimates for the models are likely to be more accurate as well. We see that as the CRRA value goes below -0.5 and the λ value increases, the WSD becomes more negative for subjects classified as either model. In Figure 4.14 we see that for RDU_{Prelec} subjects classified correctly the expected WSD is also very close to 0 across much of the range of parameters considered. As α parameter gets close to 0, seen in the bottom left plot, we see the WSD deviates more from 0 than for the rest of the range.

Figure 4.15: Welfare Surplus Difference for EUT subjects
Default Approach

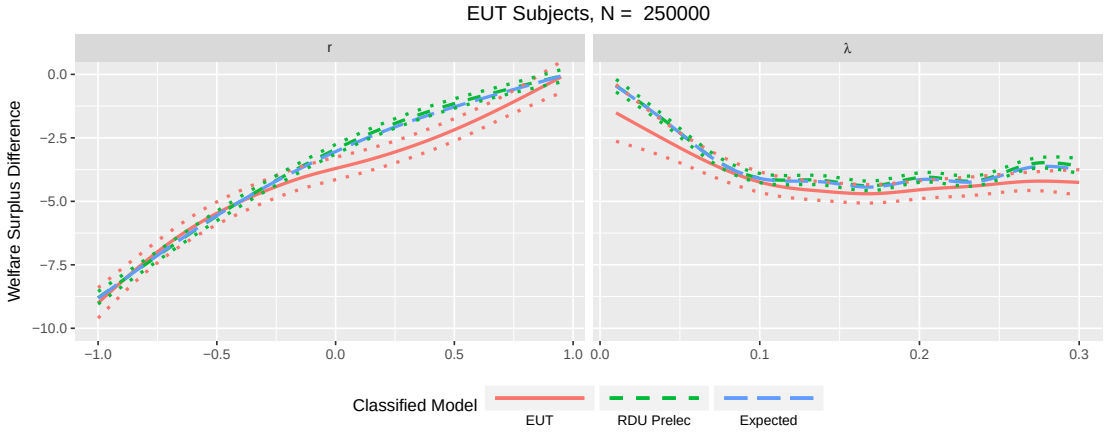
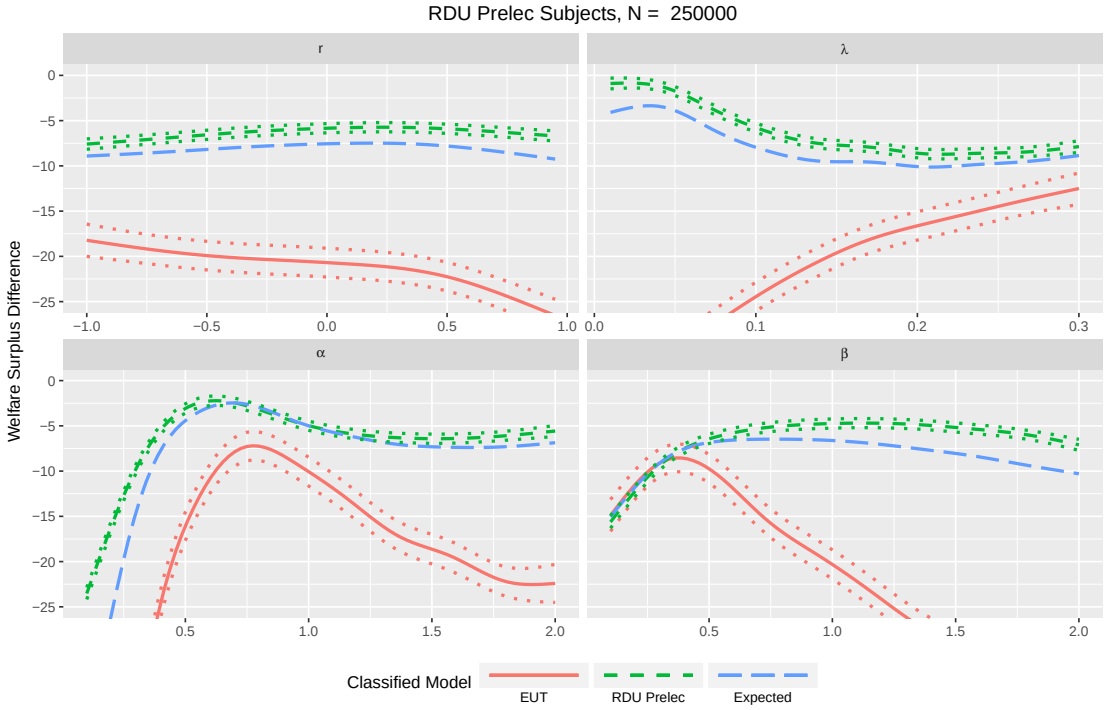


Figure 4.16: Welfare Surplus Difference for RDU_{Prelec} subjects Default Approach



In Figures 4.15 and 4.16 we can see the WSD estimates for the Default approach, where subjects are classified as employing an RDU_{Prelec} model if it hasn't been excluded, and EUT otherwise. In 4.15 we see that the WSD for EUT subjects classified as either model approaches 0 as the CRRA parameter increases, and the WSD generally becomes more negative as λ increases. For both the CRRA and λ parameters, there is little difference in the welfare estimates of subjects classified as either EUT or RDU_{Prelec}. In 4.16, on the other hand, we see there is generally a large gap between RDU_{Prelec} subjects classified as either model, with subjects classified as RDU_{Prelec}, given by the green line, being significantly better characterized than those classified as EUT, given by the red line. However, since

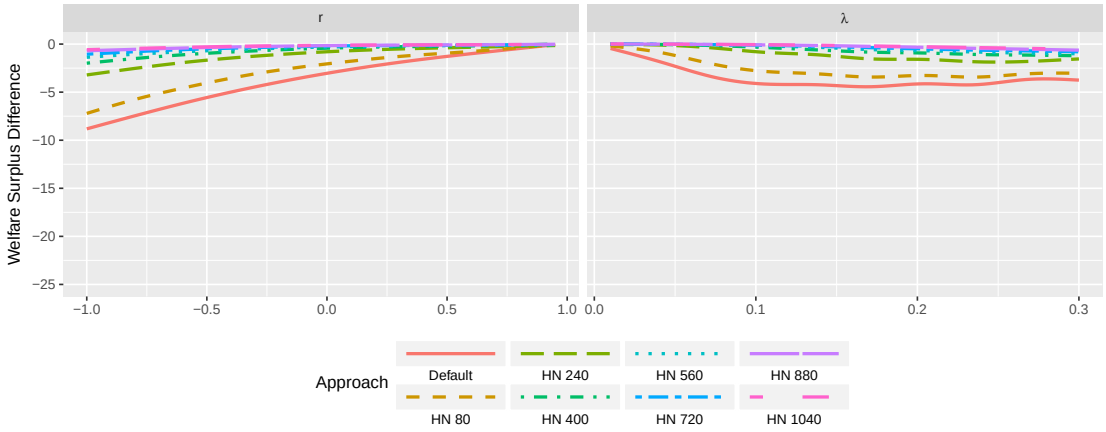
the probability of an RDU_{Prelec} subject being correctly classified is so great under this approach, the expected WSD does not deviate very much from the WSD given by the RDU_{Prelec} model.

The results presented in Figures 4.15 and 4.16 should not be surprising. The RDU_{Prelec} model nests the EUT model, thus EUT subjects can be accurately represented by an RDU_{Prelec} model by setting $\alpha = \beta = 1$. However, the EUT model does not allow for probability weighting, and thus RDU_{Prelec} subjects that undertake significant probability weighting and are classified as EUT will have their welfare surplus significantly mischaracterized.

In Figures 4.17 and 4.18 we can see the expected WSD for the Default approach and for the HN_C approach for all $C \in \{80, 240, 400, 560, 720, 880, 1040\}$. The Default approach is given by the solid green line, the original HN_{80} instrument is given by dotted yellow line, and remaining lines indicate the remaining C replications of the HN instrument. In the left plot of Figure 4.17, we see that there is very little difference in the expected WSD between any of the approaches or instruments when the CRRA parameter is greater than 0.4. In particular, there is almost no difference at all between using the classification process employed by HN vs a classification process that never rejects the RDU_{Prelec} model in this parameter range. In the right plot of Figure 4.17, depicting λ values, the WSD for the for the HN_C approaches with $C \geq 240$ are noticeably closer to 0 than for either the original HN_{80} or the Default approach, though the magnitude of this difference is still relatively small.

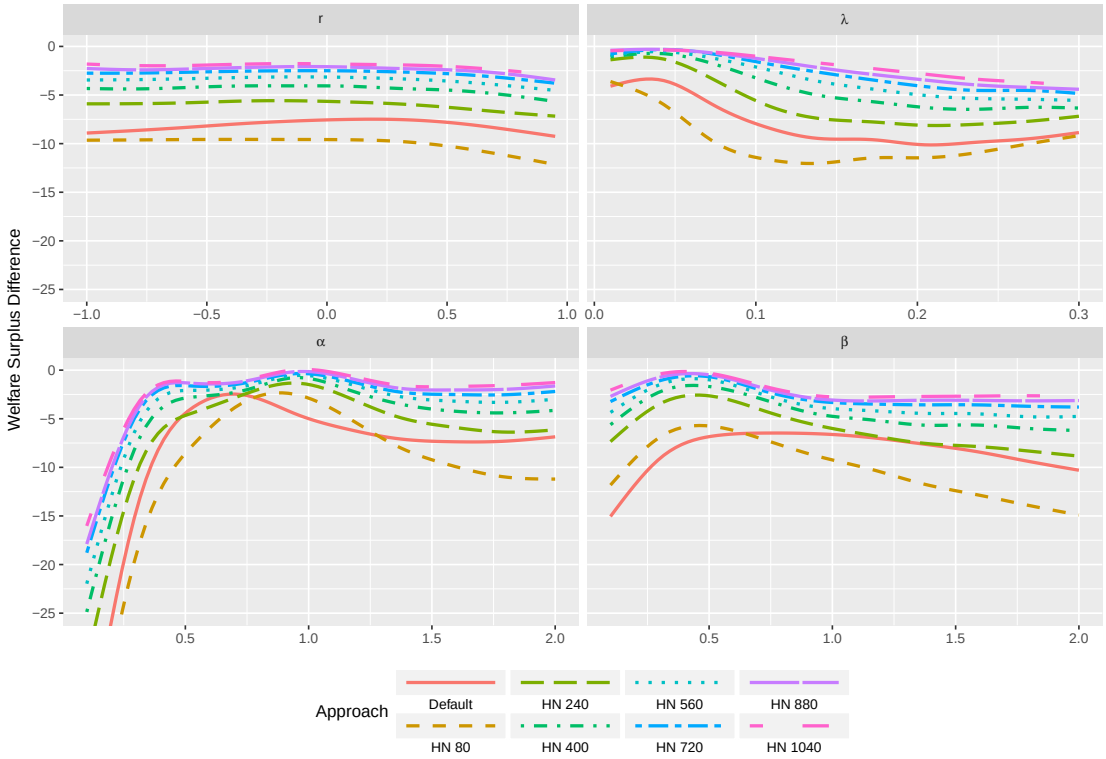
In Figure 4.18 on the other hand, we see a noticeable difference between the different approaches for RDU_{Prelec} subjects. The WSD for the Default approach, given by the solid red line, is closer to 0 than for the HN_{80} approach, given by

Figure 4.17: Welfare Surplus Difference for EUT subjects
All Approaches



the dotted yellow line, for the entire range of CRRA parameters considered, top left plot, the entire range of λ parameters considered, top right plot, and for most of the ranges of the α and β parameters considered, bottom left and right plots respectively. In particular the Default approach generally performs better when the α parameter is far from 1, and the β parameter is greater than 0.7. These differences are also much greater than the differences for the EUT subjects over any parameter values shown in Figure 4.17. Interestingly, the Default approach performs as well as the HN_{240} approach for values of β near 1.25, shown in the bottom right plot, and better than the HN_{240} approach for α greater than 0.5 and less than 0.75. Its clear, however, though that over a wide range of parameters the HN_{C} approaches for $C \geq 240$ provide more accurate WSD estimates. This suggests that the increased number of lottery pairs not only provides a greater likelihood of correctly classifying a subject, but also provides more accurate parameter estimates, which lead to more accurate estimates of the subjects' welfare surplus.

Figure 4.18: Expected Welfare Surplus Difference for RDU_{Prelec} subjects
All Approaches



4.3.1 How Much Does This Matter?

These differences should matter to researchers as a matter of methodological principle. How much they should matter depends on the population of subjects the experimenter expects to encounter and how much inaccuracy is tolerable in the characterization of welfare. If we consider a world that is made up only of agents employing some parameterization of either the EUT or RDU_{Prelec} models we consider here, the proportion of the population belonging to either model should influence how much we care about these differences. If most of the EUT agents in the population employ a CRRA parameter greater than 0.4, we might not care

which approach is used to classify EUT subjects. But if a significant proportion of them are risk seeking (CRRA < 0), we may care. Likewise, if the RDU_{Prelec} agents in the population don't undertake significant probability weighting, choosing between the various approaches presented may not matter a great deal in terms of welfare characterizations.

We can observe this more cleanly by considering a hypothetical population of EUT and RDU_{Prelec} agents, and predicting the expected WSD for these populations. As a basis for the hypothetical population, I first classify the real subjects from the HN experiments as either EUT or RDU_{Prelec} using the HN classification process, then fit a pooled EUT model to the subjects classified as EUT and a pooled RDU_{Prelec} model to the subjects classified as RDU_{Prelec}. I classify 52 subjects as employing the EUT model, 44 subjects as employing the RDU_{Prelec} model, and 15 subjects remain unclassified. The point estimate of the CRRA parameter for the EUT subjects is 0.49, and the point estimate of the λ value is 0.10. For the RDU_{Prelec} subjects, the point estimates of the CRRA parameter is 0.52, the α parameter is 1.48, the β parameter is 0.74, and the λ parameter is 0.12. Although these are estimates, and not the real values of parameters which we have been discussing, they allow us to construct a useful hypothetical scenario.

Consider a population comprised of EUT and RDU_{Prelec} agents that employ the EUT and RDU_{Prelec} models that have been specified. In this population, suppose that for both EUT and RDU_{Prelec} agents, the CRRA parameter is distributed normally with a mean of 0.5 and a standard deviation of 0.11, and the λ parameter is distributed log-normal with a mean of 0.1 and a standard deviation of 0.02. For the RDU_{Prelec} agents the α parameter is distributed log-normal with a mean of 1.50 and a standard deviation of 0.1, and the β parameter is also distributed

log-normal with a mean of .7 and a standard deviation of 0.1. Assume that for each model, none of the marginal distributions are correlated. Utilizing the fitted models from equations (4.17) and (4.19), for each of the two model populations specified above, we draw 10,000 agents from the hypothetical population and predict classification probabilities and WSD for the Default approach and HN_{C} for $C \in \{80, 240, 400, 560, 720, 880, 1040\}$. Figures 4.19 and 4.20 show the kernel density plots for the real welfare surplus of these populations, the welfare surplus estimates of those subjects that are classified as EUT or RDU_{Prelec}, and the expected welfare surplus estimates given the classification probabilities by population. These figures show welfare surplus estimates, *not* the WSD metric, which is shown in the tables below.

Figure 4.19: Welfare Surplus, HN_{80} Instrument
Default Approach

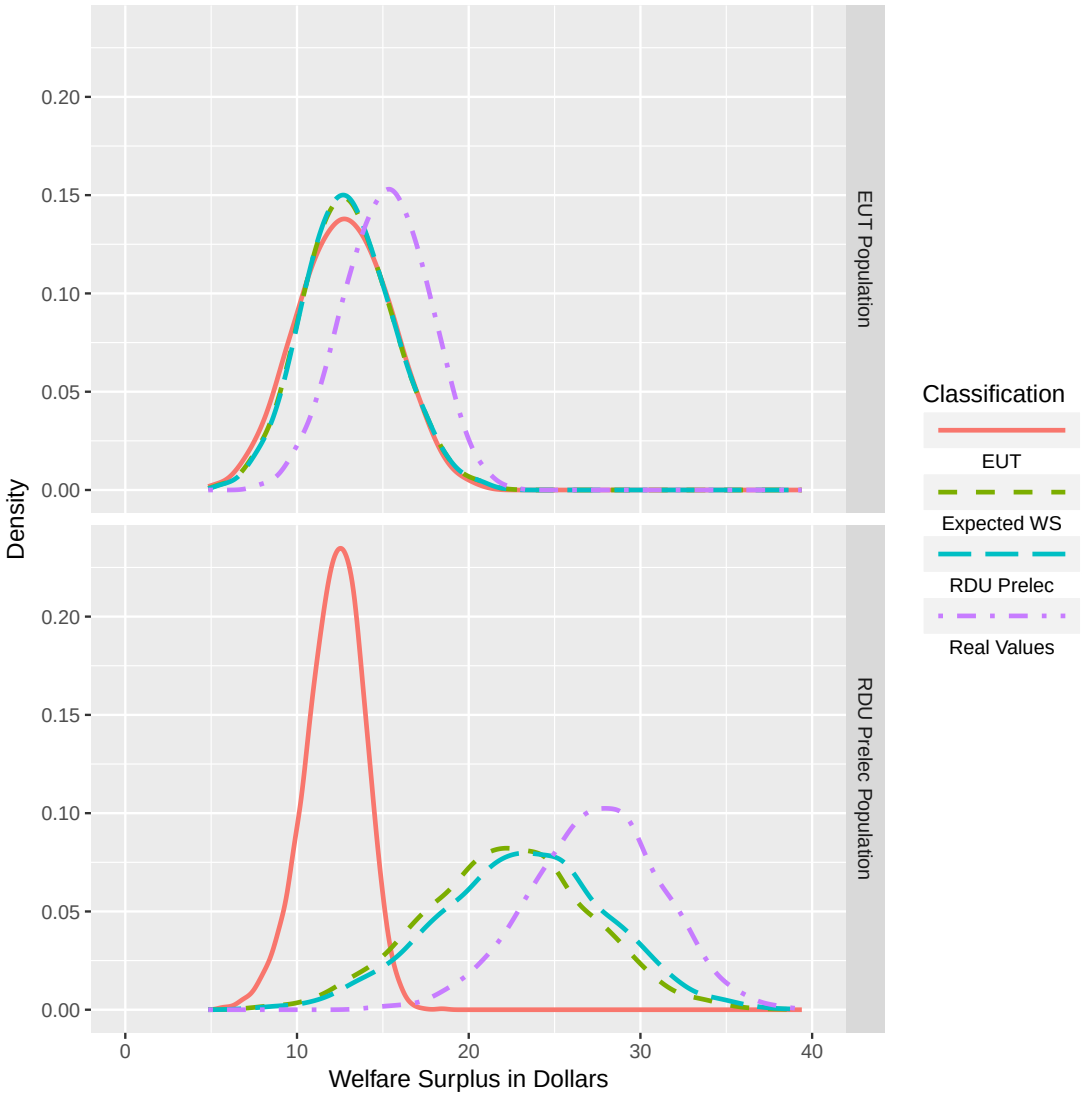


Figure 4.20: Welfare Surplus, HN_{80} Instrument
HN Approach

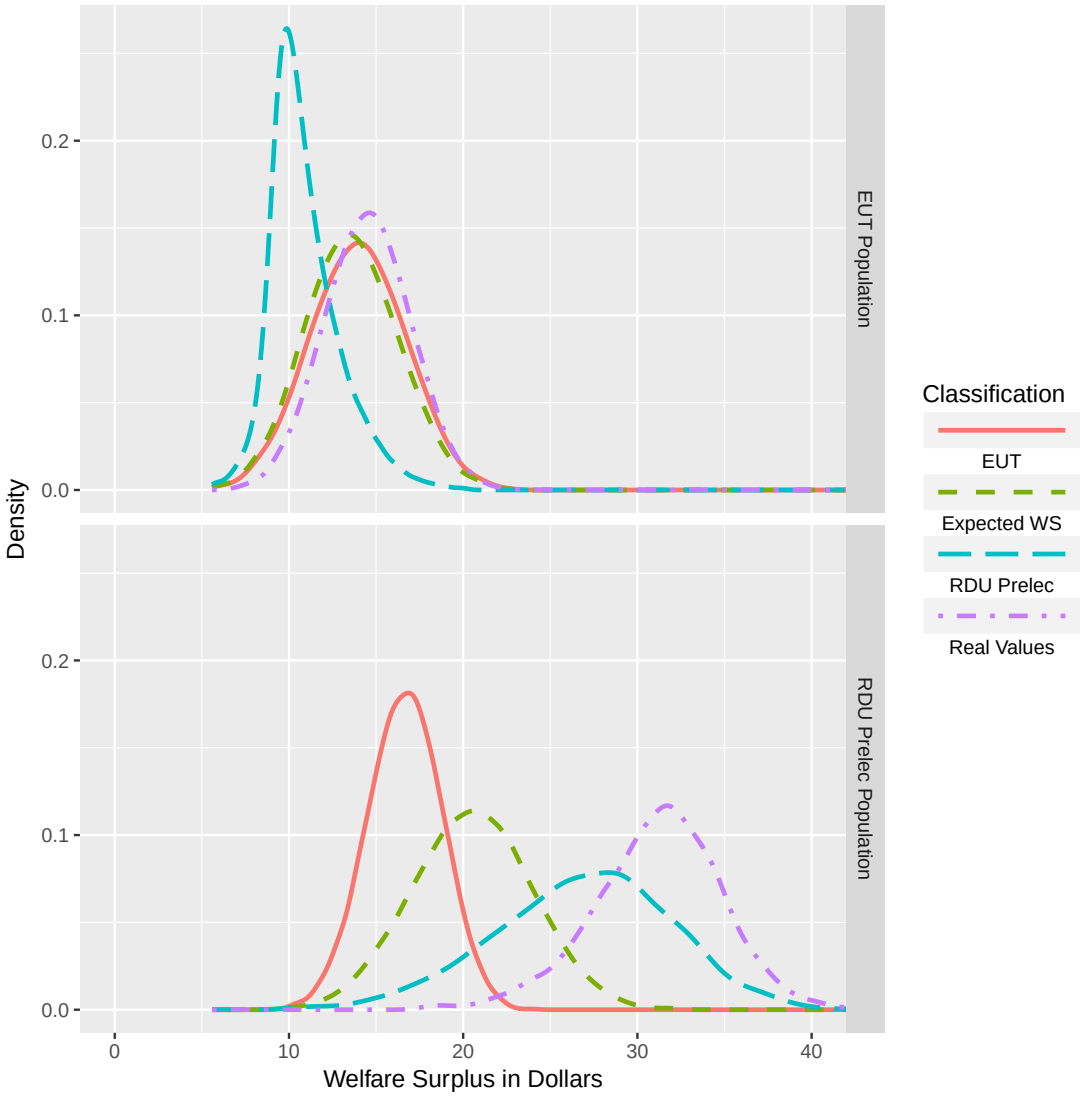


Table 4.1: Expected Welfare Surplus Difference, Default Approach

	p(EUT)	p(Prelec)	WSD_{EUT}	WSD_{Prelec}	Expected WSD
EUT Subjects	0.10	0.91	−2.30	−1.54	−1.61
Prelec Subjects	0.09	0.91	−16.51	−4.29	−5.30

Table 4.2: Expected Welfare Surplus Difference, HN_{80} Approach

	p(EUT)	p(Prelec)	WSD_{EUT}	WSD_{Prelec}	Expected WSD
EUT Subjects	0.88	0.12	−0.52	−4.50	−0.99
Prelec Subjects	0.63	0.38	−10.09	−4.40	−7.91

Table 4.3: Expected Welfare Surplus Difference, HN_{400} Approach

	p(EUT)	p(Prelec)	WSD_{EUT}	WSD_{Prelec}	Expected WSD
EUT Subjects	0.95	0.05	0.07	−0.71	0.02
Prelec Subjects	0.25	0.75	−7.34	−0.19	−1.96

Table 4.4: Expected Welfare Surplus Difference, HN_{560} Approach

	p(EUT)	p(Prelec)	WSD_{EUT}	WSD_{Prelec}	Expected WSD
EUT Subjects	0.95	0.05	0.06	−0.04	0.06
Prelec Subjects	0.17	0.84	−6.76	0.05	−1.08

Table 4.5: Expected Welfare Surplus Difference, HN_{1040} Approach

	p(EUT)	p(Prelec)	WSD_{EUT}	WSD_{Prelec}	Expected WSD
EUT Subjects	0.95	0.05	0.02	0.40	0.04
Prelec Subjects	0.07	0.93	−6.85	0.20	−0.31

In Tables 4.1 through 4.5 we see the average of the predictions for the hypothetical population for the Default approach and the $\mathrm{HN_C}$ approach for $C \in \{80, 400, 560, 1040\}$. The names of the rows in these tables give the model that the agents actually employ. In the first two columns of each table, we see the average probability of an agent employing a model given by the name of the row being classified as the model given in the column. In the third and fourth columns of each table, we see the average WSD should the agent be classified as the model given in the column name. In the fifth column of each table, we see the average expected WSD for the row population.

In Tables 4.1 through 4.5 we see a snapshot of the patterns depicted in the Figures presented throughout this chapter. Correctly classified EUT subjects are better characterized under any of the $\mathrm{HN_C}$ approaches than under the Default Approach, as seen by comparing the first row, third column of each table. Correctly classified RDU_{Prelec} subjects are better characterized under the Default approach than under the $\mathrm{HN_{80}}$ approach, though just barely so, as seen by comparing the second row, fourth columns of Tables 4.1 and 4.2. All subjects are more likely to be correctly classified, and have better a expected WSD for $\mathrm{HN_C}$ approaches with $C \in \{400, 560, 1040\}$. What these tables show more clearly, however, is the cost in terms of welfare surplus of choosing between these approaches given populations of agents we might readily encounter in experiments with real subjects.

In terms of correctly classifying subjects, we can see that for this population the average probability of correctly classifying RDU_{Prelec} subjects in the original $\mathrm{HN_{80}}$ approach is a surprisingly low 38%. The probability of correctly classifying EUT subjects with the $\mathrm{HN_{80}}$ is much greater than for RDU_{Prelec} subjects, at 88%, and rapidly approaches the 95% limit. The RDU_{Prelec} subjects however, are not

correctly classified 95% of the time for any of the repetitions, as seen in the second row second column of each plot, and only reach a correct classification probability of 80% with more than 400 lottery pairs per subject.

In Figures 4.19 and 4.20 we see the differences between the estimated welfare surplus and the real welfare surplus for the original HN approach and the Default Approach for the HN_{80} instrument. The estimated welfare surplus for subjects classified as EUT is given by the solid red line, the estimated welfare surplus for subjects classified as RDU_{Prelec} is given by the log-dashed blue line, the expected welfare surplus given the probabilities of classification for this population is given by the short-dashed green line, and the real welfare surplus for these subjects is given by the dot-dashed purple line. These displays provide some distributional information, as well as show the raw welfare surplus estimates for these two approaches, whereas Tables 4.1 and 4.2 provide metrics for the average and expected WSD, and classification probabilities for the same approaches. The raw figures show that the average real welfare surplus for EUT subjects is roughly \$15, and for RDU_{Prelec} is roughly \$28. This means that the expected WSD for RDU_{Prelec} subjects in the HN_{80} approach of -\$7.90, shown in the second row, fifth column of Table 4.2, is particularly large with respect to the RDU_{Prelec} subjects' average real welfare surplus.

Looking at the fifth column of Tables 4.2 and 4.1, we see that going from the HN_{80} approach to the Default approach, RDU_{Prelec} subjects have an improvement in the accuracy of their expected WSD of \$2.60, while EUT subjects only have a decrease of \$0.62. That is, the average RDU_{Prelec} subject will have a welfare surplus estimate that is \$2.66 closer to their real welfare surplus under the Default approach, while the average EUT subject has welfare surplus estimates that are \$0.62 farther

away from their real welfare surplus. This difference is roughly 10% of the real welfare surplus for RDU_{Prelec} subjects, and only about 4% of the EUT subjects' real welfare surplus. To put it another way, assume a grand population made up of the two populations of EUT and RDU_{Prelec} subjects we've posited here. The proportion of EUT subjects in this grand population would have to be greater than 80.7% for the loss of the WSD for EUT subjects to outweigh the gain to RDU_{Prelec} subjects.[10] Thus, if we expect real subjects to employ parameters similar to those assumed here, and if we expect the proportion of EUT subjects to be lower than 80.7% of the population, the Default approach will produce more accurate welfare surplus estimates than the HN_{80} approach. Given the results presented in Figure 4.1 and the power calculations presented throughout this chapter, the evidence would weigh against a population of real subjects with such a high proportion of EUT subjects.

Choosing between the Default approach or one of the HN_C approaches, however, requires consideration of more than just the classification or welfare surplus accuracy in real experiments. In both the HN_{80} and Default approaches, the experimental protocol and instruments were identical, and thus comparing the expected WSD of the two approaches is an appropriate way of choosing between the approaches. The HN_{1040} approach, however, changes the size of the experimental instrument dramatically and would therefore require changes in the experimental protocol. Even a more modest increase in the number of lottery pairs, to 400 for intance, to increase the probability of correctly classifying RDU_{Prelec} subjects, can create new methodological concerns. Unlike our simulated subjects, real subjects may experience boredom or fatigue should the experiment be conducted in one sitting,

[10]For $p = 0.807$, $p \times 0.62 \approx (1-p) \times 2.6$. For $p > 0.807$, $p \times 0.62 > (1-p) \times 2.6$.

and they may experience changes in the background wealth, risks, or beliefs should the experiment be conducted over several days. Any of these factors may plausibly result in a subject employing one functional at the beginning of the experiment and another functional by the end. Indeed, Hey (2001) test the hypothesis that subjects change the functional they use when lottery tasks are repeated 5 times by presenting subjects with a 100 lottery pair battery over 5 days. He concludes that "Across the repetitions the variablility of responses declines for some subjects but stays constant for others (and indeed actually increases for a small number of subjects.)" Experimenters need to weigh these methodological concerns against their ability to provide more accurate estimates of welfare, and better classification accuracy.

4.4 Conclusions

This chapter demonstrates a method for conducting a power analysis over a wide range of potential DGP, shows that conducting individual level classification with subjects responding to fewer than several hundred lottery pairs is likely to lead to frequent misclassifications, and that these misclassifications can be costly in terms of the measurement of subjective welfare surplus. Though inferential objectives vary greatly across the experimental literature, many researchers estimate multiple structural models of utility and classify individual subjects as employing one based on their estimates.

Given the inferential objective of HN of assessing the subjective welfare consequences of the decision to purchase, or not to purchase, a particular insurance product, I present mixed evidence. The capacity of the classification process to correctly classify a subject as employing either the EUT or RDU_{Prelec} model is rela-

tively low for parameterizations of these models we expect real subjects to employ. For a hypothetical population parameterized by the point estimates of real subject data, the average probability of correctly classifying RDU_{Prelec} subjects is shown to be less than 40%. However, although misclassification results in negative welfare consequences for the subjects, these negative consequences are not particularly massive, and the gain of the alternative "Default" approach, in which subjects are classified as RDU_{Prelec} if feasible and EUT otherwise, averages only several dollars across the hypothetical population of RDU_{Prelec} subjects. Nonetheless, I conclude that utilizing the proposed Default approach of classification or increasing the sample size by several hundred lottery pairs per subject *would* result in more accurate subjective welfare estimates in aggregate for populations of EUT and RDU_{Prelec} subjects we may expect to encounter in experiments.

These two approaches, increasing the sample size and disregarding classification altogether, are not the only options available to increase the accuracy of the classification process or the accuracy of welfare surplus estimates. There exists the possibility of alternative experimental designs and/or econometric procedures. Econometrically, one can imagine a Baysian approach in which small groups of subjects are grouped together based on observable characteristics, such as age, sex and education, and then pooled estimations from these subgroups being used as priors to inform the individual level estimates of the members of the groups. Additionally, non-parametric or semi-parametric estimation techniques may fare better in terms of classification accuracy. These econometric approaches could be performed on existing data, although power analyses should be performed to test if they improve classification accuracy or the accuracy of welfare surplus estimates.

In terms of instrument design, there are more than 1 septillion ($10^{23} < 2^{80}$)

possible choice patterns in the HN lottery instrument! Reducing this choice space while maintaining the same number of lottery pairs would require a different experimental procedure, but this could reduce the chance of choice errors causing a misidentification, perhaps by explicitly prohibiting subjects from selecting certain choice patterns which are likely to lead to misclassification. Of course, prohibiting certain choice patterns would require additional econometric restrictions since choices across individual lottery pairs could no longer be said to be independent. The task of modifying the experimental design and econometric procedure, while guarding against other concerns proposed by experimental methodology, is Herculean, and this chapter cannot provide much guidance with respect to this task beyond increasing the sample size.

One of the more difficult questions this chapter hoped to help address is that of "how much does this matter?" By representing the cost of misclassification as a function of the difference between estimates of welfare surplus and the known, "real" welfare surplus of our simulated subjects, we bring the question of "how much?" into a normative domain that economists are familiar with. However, it remains unclear *by how much* estimates of welfare surplus need to deviate from real welfare surplus before they truly "matter." Harrison (1989, 1992) argues specifically that when differences in consumer surplus between choices amount to fractions of a penny in First Price Auction experiments, the choices presumably didn't matter to the subjects, and so conclusions drawn from these choices should not matter much to economists either. Generally, he argues that the dominance precept of Smith (1982) needs to be taken into serious consideration when drawing conclusions from the choice behavior of economic agents. Hey (2001, p. 21) raises similar concerns about assessing the *economic* significance of results showing that subjects may

employ different functionals when faced with the same choice task over 5 days:

> The problem with these analyses is that they are essentially statistical in nature. We as economists, might be more interested in the *economic* significance of the results. Given that the EU preference function is much easier to apply to the economic analysis of behaviour, we might want to know how far wrong we might be if we use the EU functional rather than the alternatives in such applications. It is not obvious how we might answer this question as it depends upon the particular application. But we could ask how often we would make mistakes in the prediction of behaviour using the various preference functions. This depends upon the predictions we are wanting to make. One possibility is to use the specific questions asked in this experiment — though it should be noted that the results of this analysis does depend on the specific questions. It might be better to use some kind of generally-accepted set of questions — which can be used to test the various functions — but such a set is not available and is not clear how such a set could be constructed (and then made generally-acceptable).

This chapter addresses the first of these concerns by demonstrating how much the cost is in economic terms of using the EUT functional instead of some alternative. I conclude that the cost can be very high for those subjects who undertake significant deviations from EUT. I also conclude that doing the reverse, employing an RDU function when available, results in relatively *little* cost to EUT subjects and improves the accuracy of estimates of welfare surplus for subjects employing an RDU_{Prelec} functional. As for the second point raised by Hey (2001) of using "some kind of generally-accepted set of questions" to assess the economic significance of a classification process, I take the insurance policy task of HN as an example to conduct such an economic analysis. Although this instrument usefully characterizes the choice domain of interest to HN given their inferential objective, it isn't clear that this particular instrument is suitable when assessing the welfare consequences of misclassification given different inferential objectives, or even that any instrument

could be suitably constructed to be generally applicable to many different inferential objectives.

I conclude by agreeing with Gelman and Loken (2013, p. 14): "Criticism is easy, doing research is hard." The simulation analysis performed in this chapter provides valuable insight into the power of a given instrument, but does not make recommendations on how to design an instrument to achieve a particular level of statistical power beyond the unsurprising result that power increases with sample size. It is incredibly difficult to develop an experimental design that allows for the identification of a model that subjects actually employ, or even to identify if the subject engages in probability weighting at all. The lottery instrument utilized by HN (2016, pp. 98-99) is designed to incorporate the experimental findings of Camerer (1989), Harless (1992) and Loomes and Sugden (1998), among others, that offer design elements specifically introduced to help identify probability weighting. The relatively low probability of correctly identifying probability weighting using this instrument speaks to the difficulty of conducting research in this domain.

Chapter 5

Conclusions

5.1 Review of Chapters

I focus broadly on the interpretation of choice behavior that seemingly violates Expected Utility Theory (EUT). Chapter 1 discusses economists' responses to the experimental evidence presented by Grether and Plott (1979), which investigated apparent violations of transitivity. These responses vary from developing new theoretical models, to critiques of experimental method and scope, to the promotion of stochastic models of choice. The remainder of this thesis examines on how stochastic elements of choice models influence normative statements of welfare.

Chapter 2 discusses the normative coherence of three classes of stochastic models: the "Tremble" (TR) model developed by Harless and Camerer (1994), the "Random Error" (RE) model developed by Hey and Orme (1994), and the "Random Preferences" (RP) model developed by Loomes and Sugden (1998). TR models require that with some probability, a choice is made as if it was selected entirely at

random from the set of alternatives. RP models require that subjects choose as if they had randomly picked a preference relation from some distribution of preference relations and made a choice deterministically with respect to that preference relation. RE models generally require that as the difference in expected utility of the options grows, the choice probability of the option with the greatest expected utility will approach 1, while the choice probabilities of the other options will approach 0.[1] I propose an extension of the RP model, called the Random Preference Per Option (RPPO) model, which requires an agent to choose a preference parameter from a distribution of preferences for *each* option in the set of alternatives instead of a single preference relation for the entire set of alternatives. This extension allows for options which are First Order Stochastically Dominated (FOSD) by another option to have a positive choice probability, which is prohibited by the stand-alone RP model. The RE model proposed by Hey and Orme (1994) is similar to a homoscedastic latent index model, and can be modified in useful ways by making the latent index heteroscedastic, several examples of which are detailed in Chapter 2. Chapter 2 specifically investigates the Contextual Utility (CU) model proposed by Wilcox (2008), and the remainder of the thesis utilizes this stochastic model.

Chapter 2 proposes a thought experiment, the Stochastic Money Pump (SMP), to explore the capacity of stochastic models to support coherent normative statements according to two criteria. The first criterion stipulates that should an agent be left with strictly fewer assets after a resource allocation, that agent would be said to be worse off than had she had her previous, greater, stock of assets. The second criterion stipulates that exposure to market forces should incentivize the agent to

[1]Some RE models assign a probability of 0 to options which are First Order Stochastically Dominated by another option, regardless of how great the difference in expected utility is between the two options.

behave as if conforming to the theory in question. The SMP allows for a choice pattern that would leave agents with a strictly smaller stock of assets, deemed an "extraction." The probability off an extraction and the welfare consequences of the extraction are calculated for each of the TR, CU, RP, and RPPO models, as well as a combination of RP and TR models (RP+TR). The various models can be parameterized in such a way to produce identical probabilities of extraction.

Of particular concern is the first normative criterion relating strict stocks of assets to statements about welfare. The TR and RE models usefully characterize a strict loss of assets as a loss of consumer surplus. However, the RP model, and the related RPPO and RP+TR models, allow for an extraction event to result in *greater* consumer surplus. This is despite the fact that the RP models strictly prohibit the choice of a lower stock of assets over a greater stock of assets at the individual choice level. I conclude that the RP model and its derivatives do not support coherent normative statements, and caution against its use in domains where individual level welfare is being assessed.

Chapter 3 adopts an unconditional probability and welfare framework to continue to discuss the relationship between choice probabilities and welfare. The multiple price list popularized by Holt and Laury (2002) (HL-MPL) is utilized to illustrate a disconnect between the probability of a pattern of choices and the expected welfare realization of those choices. The HL-MPL is utilized because it contains only 10 pairs of lotteries, and thus there are only 1024 possible patterns of choices, and one of the lottery pairs is a behaviorally obvious case of FOSD. For a hypothetical, simulated population, the correlation between likelihood and welfare realization is positive, but not particularly close to 1. As the unconditional likelihood of a choice pattern increases, the welfare realization of the choice pattern

generally also increases, but this is not the case across all choice patterns. In particular, it is shown that most choice patterns which do not exhibit FOSD are many times more likely to be observed, but which nonetheless provide less welfare than many choice patterns which do exhibit FOSD. This disconnect between likelihood and welfare realization is due to the manner in which the CU model is formulated to make individual choices which exhibit FOSD particularly unlikely, even if the cost of violating FOSD is relatively small in welfare terms.

Chapter 3 also discusses how the unconditional likelihood of "choice errors" and the expected unconditional welfare surplus of EUT populations relate to the distribution of preferences in the population and the instrument on which choices are made. As the density of preferences in a population increases around a "point of indifference," a parameter value for the CRRA function which would indicate indifference indifferent between the options of some lottery pair, the likelihood of choice errors increases. Generally, as the density of preferences in a population increases around multiple points of indifference, the cost of choice errors in a population also increases. However, the stochastic "noise" parameters employed by the population seem to drive the welfare costs of choice errors than the preferences representing the "deterministic core" themselves.

Chapter 4 conducts a power analysis on individual level estimation utilizing the experimental design and protocol of Harrison and Ng (2016) (HN). HN critique the "take-up" metric used in the insurance literature to judge the "success" of an insurance product. They conduct an experiment to demonstrate how the structural estimation of a utility function at the individual level can be used to calculate the consumer surplus of decisions to purchase, or not purchase, insurance products. They presented subjects with two instruments, a lottery task used to estimate the

structural model of risk preferences, and an insurance policy choice task used to measure the consumer surplus of the same subject's choices. This process requires that a model be selected in order to calculate the consumer surplus. I call this process the "classification" process, and assess the power of this process, paired with the lottery task, to correctly identify agents employing either the EUT model or a Rank Dependent Utility (RDU) model with the flexible probability weighting function given by Prelec (1998) (RDU_{Prelec}).

I find that the accuracy of the classification process depends on both the model employed by the simulated subject and the values of the parameters of that model. The probability of correctly classifying subjects that employed the RDU_{Prelec} model was found to generally be lower than 50% across the (*a priori* plausible) parameter space explored, and was noticeably lower than for EUT subjects, who were generally correctly classified between 80% and 90% of the time. The cost of misclassification in terms of the difference between estimated and actual welfare surplus was much larger for subjects that employed the RDU_{Prelec} model than for EUT subjects.

Given the asymmetry of the accuracy of estimates of welfare surplus between EUT and RDU_{Prelec} subjects, I propose an approach which classifies every subject as employing an RDU_{Prelec} model if feasible, and EUT otherwise. This "Default" approach results in greater accuracy of welfare surplus estimates for RDU_{Prelec} subjects and slightly worse accuracy for EUT subjects. For a given hypothetical population, the proportion of subjects employing the EUT model would have to be greater than 81% for the improvement in average welfare accuracy for RDU_{Prelec} subjects to be outweighed by the average loss of accuracy for EUT subjects. I additionally show that increasing the number of lottery pairs per subject from the 80 used in HN to up to 1040 results in greater classification accuracy, and more

accurate welfare surplus estimates. Even small increases in the number of lottery pairs results in greater classification accuracy, but more than 400 lottery pairs per subject would be needed to increase the probability of correctly classifying RDU_{Prelec} subjects to greater than 80% for the hypothetical population considered.

5.2 Limitations

This thesis has several limitations. In Chapter 2 the example used to make the argument that RP models do not make perfectly coherent statements about welfare relies on a parameterization of the RP model that makes the "extraction" event relatively rare and small in expected value terms compared to the expected value of the lotteries concerned. If the difference between the RP and RE models in terms of expected welfare realization is small for choice domains that economists are concerned about, then descriptive concerns may outweigh the lack of coherence of the RP model and its derivatives. Additionally, Chapter 2 assesses the normative coherence of the stochastic models on the basis of the two normative criteria referenced above. Economists may find other criteria to be of greater value in making normative prescriptions. However, any gain in normative coherence for the RP model on the basis of additional or different criteria would still have to be weighed against the criteria considered in Chapter 2.

In Chapter 3 the analysis conducted is a numerical approximation of proposed statistics for given distributions of risk preferences. These analyses do not address the difficulty of identifying these distributions of preferences experimentally. These analyses also center on the HL-MPL, which has fewer lottery pairs than many modern experimental instruments, and may not constitute a choice domain economists

are presently concerned with. Additionally, the "D statistic," proposed as a way to describe how the distribution of risk preferences interacts with the instrument, would have to be extended to include probability weighting parameters in order to be applicable to RDU populations.

In Chapter 4 the power analysis is constrained to only two different types of utility models, the EUT and RDU_{Prelec} model. Additionally, both models utilized the same risk response function, the CRRA function, and the same stochastic model, the CU model. There are many possible response functions and stochastic models that can be employed with either the EUT or RDU_{Prelec} framework. Real subjects may employ one of these different response functions or stochastic models, while still conforming to EUT or RDU_{Prelec} in general. Subjects may also employ a probability weighting function that does not closely resemble any parameterization of the RDU_{Prelec} model, but nonetheless is permissible under the general RDU framework. It may be the case that if subjects employ these different response functions, probability weighting functions, and stochastic models, their welfare is more accurately characterized under the classification process proposed by HN than under the "Default" approach proposed in Chapter 4. One should therefore be cautious when extending the conclusions drawn here to more general inferential objectives. Power analyses utilizing a wider range of EUT and RDU models can help ease this concern.

Additionally, although the "Default" approach proposed in Chapter 4 achieves greater accuracy of welfare surplus estimates for populations made up of both EUT and RDU_{Prelec} subjects in aggregate, it does so at the cost of the accuracy of classifying subjects as employing one model or the other. There are reasons why increased likelihood of correct classification would be useful normatively. Economists

should be concerned if agents routinely violate the axioms they ascribe to rational behavior. In proposing RDU as an extension of EUT, behavior which was previously considered normatively unacceptable can be rationalized as resulting in gains in welfare. Having experimental and econometric methods that can accurately identify whether or not probability weighting is a real phenomena guides economists in determining whether choices made by agents are welfare optimal, or made in error. The HN_{C} approach proposed in Chapter 4 suggests that one way to improve classification accuracy is to increase the number of lottery pairs presented to subjects. However, there are experimental methodology concerns about increasing the number of task presented to subjects that need to be weighed against the increase in statistical power.

It should also be clear that the results of these power analyses are limited by the scope of the objective under investigation. While these results are useful for experiments where the objective is to classify subjects as employing either the EUT or RDU_{Prelec} models at the individual level, these results should not be construed to suggest that similar experiments with different inferential objectives have the same strengths or weaknesses. For instance, these analyses cannot make any claims concerning inferences at the sample level from pooled data, even if the experimental protocol was identical.

This thesis offers cautions and insights for the experimental economics literature, as that literature starts to contribute rigorously to normative evaluations. It cautions that some models of choice may be useful in facilitating description, but less useful for supporting the kind of normative assessments that economists care about. It offers insights into the power of economic experiments to identify whether subjects employ probability weighting.

Bibliography

Allais, M. (1953). "Le Comportement de l ' Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l ' Ecole Americaine." *Econometrica* 21.4, pp. 503–546.

Andersen, Steffen, Glenn W. Harrison, Arne Risa Hole, Morten Lau and E. Elisabet Rutström (2012). "Non-linear mixed logit." *Theory and Decision* 73, pp. 77–96.

Ballinger, T. Parker and Nathaniel T. Wilcox (1997). "Decisions, Error and Heterogeneity." *Economic Journal* 107.443, pp. 1090–1105.

Becker, G. M., M. H. DeGroot and Jacob Marschak (1963). "Stochastic models of choice behavior." *Behavioral Science* 8.1, pp. 41–55.

— (1964). "Measuring utility by a single-response sequential method." *Behavioral Science* 9, pp. 226–232.

Bell, David E. (1982). "Regret in Decision Making under Uncertainty." *Operations Research* 30.5, pp. 961–981.

Berg, Nathan (2014). "The consistency and ecological rationality approaches to normative bounded rationality." *Journal of Economic Methodology* 21.4, pp. 375–395.

Binswanger, Hans P. (1980). "Attitudes toward risk: Experimental measurement in rural India." *American Journal of Agricultural Economics* 62, pp. 395–407.

Binswanger, Hans P. (1981). "Attitudes Toward Risk : Theoretical Implications of an Experiment in Rural India." *Economic Journal* 91, pp. 867–890.

Blavatskyy, Pavlo R. (2014). "Stronger utility." *Theory and Decision* 76.2, pp. 265–286.

Braga, Jacinto, Steven J. Humphrey and Chris Starmer (2009). "Market experience eliminates some anomalies-and creates new ones." *European Economic Review* 53.4, pp. 401–416.

Brown, Alexander L. and Paul J. Healy (2016). "Separated Decisions." *Working Paper 2016-02*. Ohio State University, Department of Economics.

Bruner, David M. (2011). "Multiple switching behaviour in multiple price lists." *Applied Economics Letters* 18.5, pp. 417–420.

Busemeyer, Jerome R. and James T. Townsend (1993). "Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment." *Psychological Review* 100.3, pp. 432–459.

Camerer, Colin F. (1989). "An experimental test of several generalized utility theories." *Journal of Risk and Uncertainty* 2.1, pp. 61–104.

Camerer, Colin F. and Teck-hua Ho (1994). "Violations of the Betweenness Axiom and Nonlinearity in Probability." *Journal of Risk and Uncertainty* 8.2, pp. 167–196.

Camerer, Colin F. and R. M. Hogarth (1999). "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Journal of Risk and Uncertainty* 19.1-3, pp. 7–42.

Carbone, Enrica (1997). "Investigation of stochastic preference theory using experimental data." *Economics Letters* 57, pp. 305–311.

Carroll, J. Douglas (1980). "Models and methods for multidimensional analysis of preferential choice (or other dominance) data." *Similarity and Choice*. Ed. by E. D. Lantermann and H. Feger. Bern, Switzerland: Huber, pp. 234–289.

Carroll, J. Douglas and Geert De Soete (1991). "Toward a new paradigm for the study of multiattribute choice behavior: Spatial and discrete modeling of pairwise preferences." *American Psychologist* 46.4, pp. 342–351.

Cason, Timothy N. and Charles R. Plott (2014). "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy* 122.6, pp. 1235–1270.

Chew, S. H., L. G. Epstein and U. Segal (1991). "Mixture Symmetry and Quadratic Utility." *Econometrica* 59.1, pp. 139–163.

Chu, Yun Peng and Ruey Ling Chu (1990). "The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note." *American Economic Review* 80.4, pp. 902–911.

Cleveland, William S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74.368, pp. 829–836.

Cleveland, William S., E. Grosse, W. Shyu, J. Chambers and Trevor Hastie (1992). "Local regression models." *Statistical Models in S*, pp. 309–376.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*. Vol. 2. New York: Academic Press.

Conte, Anna, John D. Hey and Peter G. Moffatt (2011). "Mixture models of choice under risk." *Journal of Econometrics* 162.1, pp. 79–88.

Cox, James C., Bruce Roberson and Vernon L. Smith (1982). "Theory and Behavior of Single Object Auctions." *Research in Experimental Economics*. Ed. by Vernon L. Smith. Vol. 2. Greenwich: JAI Press Inc., pp. 1–43.

Cox, James C., Vjollca Sadiraj and Ulrich Schmidt (2015). "Paradoxes and mechanisms for choice under risk." *Experimental Economics* 18.2, pp. 215–250.

Cox, James C., Vernon L. Smith and James M. Walker (1983a). "A Test that Discriminates Between Two Models of the Dutch-First Auction Non-Isomorphism." *Journal of Economic Behavior & Organization* 4, pp. 205–219.

— (1983b). "Tests of a Heterogeneous Bidders Theory of First Price Auctions." *Economics Letters* 12, pp. 207–212.

— (1985). "Experimental Development of Sealed-Bid Auction Theory; Calibrating Controls for Risk Aversion." *American Economic Review* 75.2, pp. 160–165.

— (1988). "Theory and Individual Behavior of First-Price Auctions." *Journal of Risk and Uncertainty* 1, pp. 61–99.

Cubitt, Robin P. and Robert Sugden (2001). "On Money Pumps." *Games and Economic Behavior* 37.1, pp. 121–160.

Davidson, Donald and Jacob Marschak (1959). "Experimental Tests of a Stochastic Decision Theory." *Measurement: Definitions and Theories* 17, p. 274.

Debreu, Gerard (1958). "Stochastic Choice and Cardinal Utility." *Econometrica* 26, pp. 440–444.

Diewert, W.E. (1983). "Cost-benefit analysis and project evaluation." *Journal of Public Economics* 22.3, pp. 265–302.

Edwards, Ward (1954). "The theory of decision making." *Psychological Bulletin* 51.4, pp. 380–417.

Ellsberg, Daniel (1961). "Risk, ambiguity, and the Savage axioms." *Quarterly Journal of Economics* 75.4, pp. 643–669.

Fechner, Gustav (1966). *Elements of psychophysics. Vol. I.* Ed. by Davis Humphrey Howes and Edwin Garrigues Boring. New York: Holt, Rinehart and Winston, p. 286.

Feiveson, Alan H. (2002). "Power by simulation." *Stata Journal* 2.2, pp. 107–124.

Filippin, Antonio and Paolo Crosetto (2016). "A Reconsideration of Gender Differences in Risk Attitudes." *Management Science* 62.11, pp. 1–31.

Fishburn, Peter C. (1987). "Reconsiderations in the foundations of decision under uncertainty." *Economic Journal* 97.388, pp. 825–841.

Fisher, Ronald (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd, p. 175.

Gelman, Andrew and Eric Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time." *Working Paper*. Department of Statistics, Columbia University.

— (2014). "The Statistical Crisis in Science." *American Scientist* 102, pp. 460–465.

Grether, David M. and Charles R. Plott (1979). "Economic theory of choice and the preference reversal phenomenon." *American Economic Review*, pp. 623–638.

Grüne-Yanoff, Till, Caterina Marchionni and Ivan Moscati (2014). "Introduction: methodologies of bounded rationality." *Journal of Economic Methodology* 21.4, pp. 325–342.

Hands, D. Wade (2014). "Normative ecological rationality: normative rationality in the fast-and-frugal-heuristics research program." *Journal of Economic Methodology* 21.4, pp. 396–410.

Harless, David W. (1992). "Predictions about indifference curves inside the unit triangle. A test of variants of expected utility theory." *Journal of Economic Behavior & Organization* 18.3, pp. 391–414.

Harless, David W. and Colin F. Camerer (1994). "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62.6, pp. 1251–1289.

Harrison, Glenn W. (1989). "Theory and Misbehavior of First-Price Auctions." *American Economic Review* 79.4, pp. 749–762.

— (1992). "Theory and misbehavior of first price auctions: Reply." *American Economic Review* 79.4, pp. 1426–1443.

— (1994). "Expected utility theory and the experimentalists." *Empirical Economics* 19, pp. 223–253.

Harrison, Glenn W. and Laurie T. Johnson (2006). "Identifying Altruism in the Laboratory." *Experiments Investigating Fundraising and Charitable Contributors*. Ed. by R. Mark Isaac and Douglas D. Davis. Vol. 11. Research in Experimental Economics. Bingley: Emerald Group Publishing Limited, pp. 177–223.

Harrison, Glenn W., Morten I. Lau and E. Elisabet Rutström (2007). "Estimating Risk Attitudes in Denmark: A Field Experiment." *Scandinavian Journal of Economics* 109.2, pp. 341–368.

Harrison, Glenn W., Jimmy Martínez-Correa, Jia Min Ng and J. Todd Swarthout (2017). "Evaluating the Welfare of Index Insurance." *Working Paper 2016-07*. Center for the Economic Analysis of Risk, Georgia State University.

Harrison, Glenn W. and Jia Min Ng (2016). "Evaluating the Expected Welfare Gain From Insurance." *Journal of Risk and Insurance* 83.1, pp. 91–120.

Harrison, Glenn W. and Jia Min Ng (2018). "Welfare Effects of Insurance Contract Non-Performance." *forthcoming, Geneva Risk and Insurance Review*. Center for the Economic Analysis of Risk, Georgia State University.

Harrison, Glenn W. and E. Elisabet Rutström (2008). "Risk aversion in the laboratory." *Research in Experimental Economics*. Ed. by James C Cox and Glenn W Harrison. Vol. 12. Bingley: Emerald Group Publishing Limited, pp. 41–196.

— (2009). "Expected utility theory and prospect theory: one wedding and a decent funeral." *Experimental Economics* 12.2, pp. 133–158.

Harrison, Glenn W. and J. Todd Swarthout (2014). "Experimental payment protocols and the Bipolar Behaviorist." *Theory and Decision* 77.3, pp. 423–438.

Hastie, Trevor and Robert Tibshirani (1986). "Generalized Additive Models." *Statistical Science* 1.3, pp. 297–318.

Hertwig, Ralph and Andreas Ortmann (2001). "Experimental practices in economics: A methodological challenge for psychologists?" *Behavioral and Brain Sciences* 24.3, pp. 383–451.

Hey, John D. (1995). "Experimental Investigations of Errors in Decision Making Under Risk." *European Economic Review* 39, pp. 633–640.

— (2001). "Does Repetition Improve Consistency?" *Experimental Economics* 4, pp. 5–54.

Hey, John D. and Enrica Carbone (1995). "Stochastic choice with deterministic preferences: An experimental investigation." *Economics Letters* 47, pp. 161–167.

Hey, John D. and C. Orme (1994). "Investigating generalizations of expected utility theory using experimental data." *Econometrica* 62.6, pp. 1291–1326.

Holt, CA Charles A. (1986). "Preference Reversals and the Independence Axiom." *American Economic Review* 76.3, pp. 508–515.

Holt, Charles A. and Susan K. Laury (2002). "Risk Aversion and Incentive Effects." *American Economic Review* 92.5, pp. 1644–1655.

Kadane, Joseph B. (1992). "Healthy Scepticism as an EU Explanation of the Phenomena of Allais and Ellsberg." *Decision Making Under Risk and Uncertainty*. Ed. by John Geweke. London: Kluwer Academic Publishers, pp. 11–16.

Kahneman, Daniel and Amos Tversky (1979). "Prospect theory: An analysis of decision under risk." *Econometrica* 47.2, pp. 263–292.

Karni, E. and Z. Safra (1987). "Preference reversal and the observability of preferences by experimental methods." *Econometrica* 55.3, pp. 675–685.

King, Mervyn A. (1983). "Welfare analysis of tax reforms using household data." *Journal of Public Economics* 21.2, pp. 183–214.

Leamer, Edward E. (2012). *The Craft of Economics: Lessons from the Heckscher-Ohlin Framework*. Cambridge, Massachusetts: MIT Press.

Lichtenstein, Sarah and Paul Slovic (1971). "Reversals of preference between bids and choices in gambling decisions." *Journal of Experimental Psychology* 89.1, pp. 46–55.

— (1973). "Response-induced reversals of preference in gambling: An extended replication in Las Vegas." *Journal of Experimental Psychology* 101.1, pp. 16–20.

Lindman, Harold R. (1971). "Inconsistent preferences among gambles." *Journal of Experimental Psychology* 89.2, pp. 390–397.

Loomes, Graham, Peter G. Moffatt and Robert Sugden (2002). "A microeconometric test of alternative stochastic theories of risky choice." *Journal of Risk and Uncertainty* 24.2, pp. 103–130.

Loomes, Graham, Chris Starmer and Robert Sugden (1989). "Preference Reversal: Information-Processing Effect or Rational Non-Transitive Choice?" *Economic Journal* 99.395, pp. 140–151.

Loomes, Graham and Robert Sugden (1982). "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *Economic Journal* 92.368, pp. 805–824.

— (1987). "Some implications of a more general form of regret theory." *Journal of Economic Theory* 41.2, pp. 270–287.

— (1995). "Incorporating a stochastic element into decision theories." *European Economic Review* 39.3-4, pp. 641–648.

— (1998). "Testing different stochastic specifications of risky choice." *Economica* 65, pp. 581–598.

Luce, Robert Duncan (1958). "A Probabilistic Theory of Utility." *Econometrica* 26.2, pp. 193–224.

— (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Luce, Robert Duncan and Patrick Suppes (1965). "Preference, Utility, and Subjective Probability." *Handbook of Mathematical Psychology*. Ed. by Robert Duncan Luce, R. R. Bush and E. Galanter. 3. New York: Wiley. Chap. 19, pp. 249–410.

Machina, M. J. (1985). "Stochastic choice functions generated from deterministic preferences over lotteries." *Economic Journal* 95, pp. 575–594.

— (1987). "Decision-making in the presence of risk." *Science* 236, pp. 537–543.

Marschak, Jacob (1950). "Rational Behavior, Uncertain Prospects, and Measurable Utility." *Econometrica* 18.2, pp. 111–141.

McCloskey, Deirdre N. and Stephen T. Ziliak (1996). "The Standard Error of Regressions." *Journal of Economic Literature* 34, pp. 97–114.

Miller, Louis, David Edward Meyer and John T. Lanzetta (1969). "Choice among equal expected value alternatives: Sequential effects of winning probability level on risk preferences." *Journal of Experimental Psychology* 79.3, pp. 419–423.

Moffatt, Peter G. and Simon A. Peters (2002). "Testing for the Presence of a Tremble in Economic Experiments." *Experimental Economics* 4, pp. 221–228.

Moscati, Ivan (2016). "How Economists Came to Accept Expected Utility Theory: The Case of Samuelson and Savage." *Journal of Economic Perspectives* 30.2, pp. 219–236.

Mosteller, Frederick and Philip Nogee (1951). "An Experimental Measurement of Utility." *Journal of Political Economy* 59.5, pp. 371–404.

Pigou, Arthur C. (1929). *The Economics of Welfare*. 3rd ed. London: Macmillan.

Plott, Charles R. and Vernon L. Smith (1978). "An Experimental Examination of Two Exchange Institutions." *Review of Economic Studies* 45.1, pp. 133–153.

Pratt, John W. (1964). "Risk Aversion in the Small and in the Large." *Econometrica* 32.1/2, pp. 122–136.

Prelec, Drazen (1998). "The Probability Weighting Function." *Econometrica* 66.3, pp. 497–527.

Quiggin, John (1982). "A Theory of Anticipated Utility." *Journal of Economic Behavior & Organization* 3, pp. 323–343.

Rutström, E. Elisabet and Nathaniel T. Wilcox (2009). "Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test." *Games and Economic Behavior* 67.2, pp. 616–632.

Saha, Atanu (1993). "Expo-Power Utility: A 'Flexible' Form for Absolute and Relative Risk Aversion." *American Journal of Agricultural Economics* 75.4, pp. 905–913.

Samuelson, Paul A. (1974). "Complementarity: An Essay on The 40th Anniversary of the Hicks-Allen Revolution in Demand Theory." *Journal of Economic Literature* 12.4, pp. 1255–1289.

Savage, L. J. (1954). *Foundations of Statistics*. New York: J. Wiley and Sons Inc.

Schneeweiss, H. (1973). "The Ellsberg paradox from the point of view of game theory." *Selecta Statistica Canadiana* 1, pp. 65–78.

Schubert, Renate, Martin Brown, Matthias Gysler and Hans Wolfgang Brachinger (1999). "Financial decision-making: Are women really more risk-averse?" *American Economic Review* 89, pp. 381–385.

Selten, R. (1975). "Reexamination of the perfectness concept for equilibrium points in extensive games." *International Journal of Game Theory* 4.1, pp. 25–55.

Silverstein, Shell (1974). "Smart." *Where the Sidewalk Ends*. New York: Harper Collins Publishers, p. 35.

Smith, Vernon L. (1982). "Microeconomic systems as an experimental science." *American Economic Review* 72.5, pp. 923–955.

Smith, Vernon L. and James M. Walker (1993). "Monetary rewards and decision cost in experimental economics." *Economic Inquiry* 31, pp. 245–261.

Sopher, Barry and J Mattison Narramore (2000). "Stochastic Choice and Consistency in Decision Making Under Risk: An Experimental Study." *Theory and Decision* 48.4, pp. 323–350.

Train, Kenneth (2002). *Discrete Choice Methods with Simulation*. Berkeley, California: Cambridge University Press.

Tversky, Amos and Daniel Kahneman (1992). "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty* 5.4, pp. 297–323.

Varian, Hal (1996). "What use is economic theory?" *Foundations of Research in Economics: How Do Economists Do Economics?* Ed. by Steven G. Medema and Warren J. Samuels. Cheltenham, UK: Elgar, pp. 238–247.

Von Neumann, J. and Oskar Morgenstern (1944). *Theory of Games and Economic Behavior*. Vol. 2. Princeton, New Jersey: Princeton University Press.

Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57.2, pp. 307–333.

Wakker, Peter P. (2008). "Explaining the characteristics of the power (CRRA) utility family." *Health Economics* 17.12, pp. 1329–44.

Wilcox, Nathaniel T. (1993). "Lottery Choice: Incentives, Complexity and Decision Time." *Economic Journal* 103.421, pp. 1397–1417.

— (2007). "Predicting Risky Choices Out-of-Context: A Monte Carlo Study." *Working Paper*. University of Houston, Department of Economics.

— (2008). "Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison." *Research in Experimental Economics*. Ed. by James C. Cox and Glenn W. Harrison. Vol. 12. Bingley, U.K.: Emerald Group Publishing Limited, pp. 197–292.

— (2011). "'Stochastically more risk averse:' A contextual theory of stochastic discrete choice under risk." *Journal of Econometrics* 162.1, pp. 89–104.

— (2015). "Error and Generalization in Discrete Choice." *Working Paper*. Economic Science Institute, Chapman University.

Yaari, Menahem E . (1987). "The Dual Theory of Choice under Risk Author." *Econometrica* 55.1, pp. 95–115.

Zhang, Le and Andreas Ortmann (2013). "Exploring the Meaning of Significance in Experimental Economics." *Working Paper*. Australian School of Business, University of New South Wales.