The Explanatory and Predictive Power of Non Two-Stage-Probability Theories of Decision Making Under Ambiguity

By JOHN D HEY AND NOEMI PACE*

Representing ambiguity in the laboratory using a Bingo Blower (which is transparent and not manipulable) and asking the subjects a series of allocation questions (which are more efficient than pairwise choice questions), we obtain data from which we can estimate by maximum likelihood methods (with explicit assumptions about the errors made by the subjects) a significant subset of the empirically relevant models of behaviour under ambiguity, and compare their relative explanatory and predictive abilities. Our results suggest that not all recent models of behaviour represent a major improvement in explanatory and predictive power, particularly the more theoretically sophisticated ones.

JEL: D81, C91

Keywords: Alpha Model, Ambiguity, Bingo Blower, Choquet Expected Utility, Contraction Model, Rank Dependent Expected Utility, Subjective Expected Utility,Vector Expected Utility.

The past decade has seen an explosion of theoretical work in the modelling of behaviour under ambiguity. Now it is the turn of the experimentalists to investigate the empirical validity of these theories. That is the primary purpose of this paper: to provide experimental evidence which investigates the empirical performance of theories of behaviour under ambiguity. Specifically, we complement a growing experimental literature, and, in particular, add to the work of, Abdellaoui *et al* (2011), Halevy (2007), Ahn *et al* (2010) and Hey *et al* (2010), though our detailed objectives, methods and results

* Hey: University of York, UK, john.hey@york.ac.uk. Pace: Ca' Foscari University of Venice, Italy, n.pace@unive.it. Website: There is an associated website with additional material at http://www-users.york.ac.uk/~jdh1/hey%20and%20pace/

differ in many respects substantially from theirs.

In essence, all these papers (and others) are aimed at the same fundamental objective: to discover which of the many theories of behaviour under ambiguity are empirically most appealing. However our work differs from these earlier works in terms of: (1) the representation of ambiguity (except for Hey *et al*); (2) in terms of the experimental design (except for Ahn *et al*); (3) in terms of the theories being explored; and (4) the econometric methods (except for Hey *et al*).

Ambiguity is represented in different ways in the experiments on which these different papers were based. Ambiguity is understood as a situation in which probabilities do not exist or the decision-maker does not know the actual probabilities. Both Halevy and Abdellaoui *et al* use as one of their representations the traditional 'Ellsberg Urn': subjects are told what objects are in the urn but are not told the quantities of each object, so that the probability of drawing any particular object can not be known by the subject. Abdellaoui *et al*, given that their objective is to examine the impact of different *sources* of ambiguity, also consider other sources (changes in the French Stock Index, the temperature in Paris, and the temperature at some randomly drawn remote country - all on a particular day). Ahn *et al*'s representation is simply not to tell the subjects what the precise probability of two of the three possible outcomes was; this is a sort of continuous Ellsberg Urn and inevitably suffers from the usual problem that the subjects may simply consider it as the 'suspicious urn'. In contrast, Hey *et al* used an open and transparent representation: a Bingo Blower. This is also what we used in the experiment reported in this paper. The Blower removes any possible suspicion; moreover it enables us to carry out two treatments which unambiguously have different amounts of ambiguity.

The papers by Hey *et al* and Abdellaoui *et al* use the 'traditional' form of experimental question: pairwise choice, while Halevy uses reservation price questions. In contrast, Ahn *et al* use the *allocation* type of question pioneered originally by Loomes (1991) but forgotten for many years until revived by Andreoni and Miller (2002) in a social choice context and later by Choi *et al* (2007) in a risky choice context. In this paper we use allocation questions, which are more informative than pairwise choice questions and probably more reliable than reservation price questions, and thus more able to detect true preferences. In this respect the comments by Wilcox (2007) as to the informative nature of experimental data in general, and pairwise choice questions in particular, should be noted.

The set of theories of behaviour under ambiguity is now very large. Those theories - such as maximin - which do not incorporate a preference functional - have been largely discredited (partly by Hey *et al*); we do not consider them here. Of the remaining theories, one can make a broad distinction between the set of theories that use second-order probabilities[1] and the set that does not. For example, if there are I possible events $i = 1, 2, ..., I$, but the probabilities of them are not known, those theories that use second-order probabilities assume that the decision-maker works on the basis that there is a set of J possible values for these probabilities, with the j^{th} set taking values $p_{1j}, p_{2j}, ..., p_{Ij}$ with the probability that the j^{th}set being true given by $\pi_j, \pi j = 1, 2, ..., J$. In contrast, the set of theories that do not use second-order probabilities may assume that the p_i may take a range of values in the decision-maker's mind, but he or she does not attach probabilities to these possible values. We restrict attention here to this second set (non two-stage-probability models). This is for three reasons: the way we represent ambiguity in the laboratory (there is no obvious first stage); the complexity of the resulting models in the two-stage-probability set; and problems with identifiability of the underlying preference functionals (because of the large number of parameters). In contrast, Halevy uses two-stage-probability models because his experimental design effectively makes such models appropriate. One could also argue that the same applies to the Ahn *et al* experiment: there they have three possible outcomes 1, 2 and 3. Subjects are told p_2 but they are not told anything about p_1 and p_3 (except that they obey the usual probability rules). However, if subjects had read footnote 4 of their paper[2] then a two-stage-probability representation would have been natural.

[1] Sometimes called Mulitple Prior models.

[2] Which read "In practice, the probability of one of the 'ambiguous' states was drawn from the uniform distribution over [0,2/3]. This distribution was not announced to the subjects." If the distribution had been revealed to the subjects, the decision problem would have involved compound risk rather than ambiguity."

Ahn *et al* make another distinction amongst the various specifications of behaviour under ambiguity: between those specifications that they call *smooth* and those that they call *kinked*. Essentially this distinction depends upon whether preference depends upon the *ordering* of the outcomes: in Expected Utility theory this is not the case and hence this is a smooth specification; in contrast, Choquet Expected Utility and (as one can tell from its name) Rank Dependent Expected Utility (if one considers this as a theory of behaviour under ambiguity) are kinked. Ahn *et al* do not estimate particular preference functionals but rather two general specifications - one smooth and one kinked. They note that the smooth specification "can be derived from" Recursive Expected Utility (REU, which is a two-stage-probability model); while the kinked specification "... can be derived as a special case of a variety of utility models: MEU, CEU, Contraction Expected Utility, and α-MEU[3]". We note two things: first that the smooth specification does not come *only* from REU (indeed it comes from several other models, such as the Variational model of Maccheroni *et al* (2006)); and secondly, but perhaps more importantly, Ahn *et al* do not estimate any preference functionals that come *specifically* from the models that they mention. So they do not test *directly* any of the recent theories of behaviour under ambiguity.

Abdellaoui *et al* effectively investigate only one model - essentially Rank Dependent Expected Utility (RDEU) theory. This has two key elements, a utility function, which they take to be CRRA (Constant Relative Risk Averse), and a (probability) weighting function, which they take, in the 'Ellsberg Urn' part of the experiment, to be of the Prelec form: $w(p) = (\exp(-(-\ln(p))^{\alpha}))^{\beta}$. It may seem a bit odd using probabilities in a study of ambiguity, but these probabilities are the true probabilities, which, of course, the experimenters know, if not the subjects. Using these functional specifications, RDEU is a special case[4] of Choquet Expected Utility, which we estimate. In the 'Natural Uncertainties' part of the experiment they do not assume any particular form for the weighting

[3](Our note) MEU, CEU and α-MEU are respectively MaxMin Expected Utility, Choquet Expected Utility, and Alpha Expected Utility (all of which we consider specifically later).

[4]If they had estimated the weighting function at all points, rather than estimating the parameters of the particular functional form, it would have been precisely Choquet.

function, so RDEU in this context is precisely Choquet - which we estimate.

There are significant econometric differences between these various papers. First, and rather hidden from view, is the fact that we carried out extensive pre-experimental simulations to ensure that we had a sufficient number and an appropriate set of questions to ask the subjects; too many experiments have too few questions and thus lack power to discriminate amongst the theories. Second, the estimation methods vary. Underlying any particular chosen estimation method, there is an assumption about the stochastic specification of the model. Usually this is tacit; it should be explicit, particular as there is an obvious source for the stochastic component of the data - if one is estimating subject by subject (which is the case in all these papers) this comes from either randomness in preferences or errors made by the subjects. We see no mention in any of this literature of randomness in preferences, so the noise, the stochastic component, must come from errors made by the subjects. We explicitly include a story of such mistakes. We estimate the various preference functionals with the stochastic specification specifically built in to the estimation; this does not appear to be the case in the other papers we have mentioned in this section (though it is the case in Anderson *et al* (2009) which we will discuss later), so their assumptions about errors is not clear.

We also go one step further than all these other papers. Believing that economics is all about predicting, rather than just explaining, we compare our different models by seeing how good they are at predicting. For the importance of this, see Wilcox (2007 and 2011).

In summary: we represent ambiguity in the laboratory in an open and non-manipulable manner; we ask a set of allocation questions to the subjects (obviously with an appropriate incentive mechanism) chosen after extensive simulations; we use maximum likelihood estimation, with a carefully-specified stochastic component, to estimate a significant sub-set of the empirically relevant theories of behaviour under ambiguity, and to compare their relative goodness of fit; finally we compare the various theories in terms of their predictive ability. To prepare the reader for what is to come, we should warn that the theorists are going to be disappointed: the recent elegant theories are in general not empirically superior to simpler theories.

The rest of the paper is organised as follows. In the next section we give a brief overview of the theories that we are going to fit to our data. The following section describes the way that our data was generated in our experiment. We then relate what we do to the literature, a part of which we have discussed in this introduction, and give more detail about what others have done. A section describing the technicalities underlying our analysis then follows, after which we present our results. We then conclude.

I. Theories under investigation

This section discusses the theories of decision-making under ambiguity that we investigate. We confine our attention to those theories in which there is an explicit preference functional, and hence we exclude earlier (and largely discredited[5]) theories which proceed directly to a decision rule. In all the theories which we consider the decision maker is perceived, in any decision problem, as maximising the value of some preference functional. As noted above, we include: (1) Subjective Expected Utility (SEU) theory in which the decision-maker is envisaged as working with subjective probabilities; (2) the Choquet Expected Utility (CEU) model, usually nowadays accredited to Schmeidler (1989), which allows the agent's beliefs to be represented by unique but nonadditive "capacities"; (3) the Alpha Expected Utility model (AEU) of Ghirardato *et al* (2004), which models the agent's beliefs as being represented by a set of probabilities (but without attaching probabilities to the members of this set) - and its two special cases, (4) Maxmin Expected Utility of Gilboa and Schmeidler (1989), and (5) Maxmax Expected Utility; (6) Vector Expected Utility (VEU) of Siniscalchi (2009) in which uncertain prospect is assessed according to a baseline expected utility evaluation and an adjustment that reflects the individual's perception of ambiguity and her attitude toward it; and (7) the Contraction Model of Gajdos, Hayashi, Tallon and Vergnaud (2008) which combines Maxmin with Expected Utility at a particular point in the probability set. We note that SEU is a 'smooth' specification in the sense used above, while all the rest are

[5]Such as, for example, MaxMin (in which the decision-maker looks at the worst that can happen and makes that as good as possible) and MaxMax (in which the decision-maker looks at the best that can happen and makes that as good as possible). See Hey, Lotito and Maffioletti (2010) for the empirical evidence against such theories.

kinked specifications. We tried to fit the Variational model of Maccheroni *et al* (2006), which is a smooth specification, but without success (in terms of goodness of fit); this may have been the consequence of the particular context of our experiment.

We give an overview of these theories below.[6] We restrict attention in both the overview and the detail to decision problems with at most three events - which was the case in our experiment. Call these events E_1, E_2 and E_3 To each event there will be associated an outcome to the decision-maker which consists of an amount of money. We denote the utility of the decision-maker for these three outcomes u_1, u_2 and u_3. For some of the theories - those with a 'rank-dependent' flavouring - the ordering of the outcomes will be crucial and we will assume in what follows that $u_1 \geq u_2 \geq u_3$ though it should be noted that it is not necessarily the case that the ordering of the outcomes is the same as the original ordering of the events: this depends upon the decisions that the decision-maker makes. Let us denote the event which leads to the highest outcome by $E_{(1)}$, that to the second highest outcome by $E_{(2)}$ and that to the lowest outcome by $E_{(3)}$. We note that the set $\{E_{(1)}, E_{(2)}, E_{(3)}\}$ consists of the numbers $1, 2$ and 3, though not necessarily in that order.

Subjective Expected Utility Theory

The preference functional for SEU is given by

$$SEU = \sum_{i=1}^{3} p_i u_i \tag{1}$$

where p_i is the subjective probability that event $E_{(i)}$ occurs. In this case $p_i = \text{Prob}(E_{(i)})$ for all i, and, of course $p_1 + p_2 + p_3 = 1$.

[6]More technical detail are available on this web-site: http://www-users.york.ac.uk/~jdh1/hey%20and%20pace/

According to Schmeidler (1989), the Choquet Expected Utility of a lottery is given by

$$CEU = \sum_{i=1}^{3} \bar{w}_i u_i \tag{2}$$

where the $\bar{w}$'s are weights that depends on nonadditive *capacities* w that satisfy the normalisation conditions and monotonicity. In the context of our experiment, a CEU subject works with six nonadditive capacities $w_{E_{(1)}}$, $w_{E_{(2)}}$, $w_{E_{(3)}}$, $w_{E_{(2)} \cup E_{(3)}}$, $w_{E_{(1)} \cup E_{(3)}}$ and $w_{E_{(1)} \cup E_{(2)}}$ referring to the three events and their pairwise unions. Crucially, the weights $\bar{w}$ depend upon the ordering of the outcomes:

$$\begin{aligned} \bar{w}_1 &= w_{E_{(1)}} \\ \bar{w}_2 &= w_{E_{(1)} \cup E_{(2)}} - w_{E_{(1)}} \\ \bar{w}_3 &= 1 - w_{E_{(1)} \cup E_{(2)}} \end{aligned} \tag{3}$$

We note that the main difference between CEU and SEU consists in the finitely additive probability measure being replaced by a nonadditive capacity measure. If the capacities are actually probabilities (that is, if $w_{E_{(2)} \cup E_{(3)}} = w_{E_{(2)}} + w_{E_{(3)}}$, $w_{E_{(1)} \cup E_{(3)}} = w_{E_{(1)}} + w_{E_{(3)}}$, $w_{E_{(1)} \cup E_{(2)}} = w_{E_{(1)}} + w_{E_{(2)}}$ and $w_{E_{(1)}} + w_{E_{(2)}} + w_{E_{(3)}} = 1$) then (2) is equivalent to (1). We note that CEU is the same as Rank Dependent Expected Utility (which is not regarded by all as a theory of behaviour under ambiguity because it uses objective probabilities, but also uses, as we have already noted, to rescue it from that criticism, a weighting function, mapping objective probabilities into subjective probabilities) under an appropriate interpretation of that latter theory[7]. Similarly Cumulative Prospect Theory, with a fixed reference point, can be regarded in the same way as the same as CEU.

[7] In the context of our experiment, where there are three outcomes and hence 6 capacities, then the relationship between the two theories is given by the following, where $p_{(1)}, p_{(2)}, p_{(3)}$ are the objective probabilities and $w(.)$ is the weighting function, and the capacities for CEU are as denoted above:
$w_{E_{(i)}} = w(p_i)$ for $i = 1, 2, 3$ and
$w_{E_{(j)} \cup E_{(j)}} = w(p_j + p_k)$ for $j \neq k \in 1, 2, 3$

ALPHA EXPECTED UTILITY THEORY

Alpha Expected Utility theory (AEU) was proposed by Ghirardato *et al* (2004) as a generalization of the theory proposed in Gilboa and Schmeidler (1989). Ghirardato *et al* (2004)'s model implies that, although the decision maker does not know the true probabilities, she acts as if she believes that the true probabilities lie within a set D of probabilities on different events. We can refer to each prior $p \in D$ as a "possible scenario" that the decision maker envisions. According to Ghirardato *et al*, the set D of probabilities represents formally the ambiguity that the decision maker feels in the decision problem (they introduce the concept of "revealed ambiguity"). In other words, the size of the set D measures the perception of ambiguity. The larger D is, the more ambiguity the decision maker appears to perceive in the decision problem. In particular, no decision maker perceives less ambiguity than one who reveales a singleton set $D = \{p_1, p_2, p_3\}$. In this case the decision maker is a SEU maximiser with subjective probabilities p_1, p_2 and p_3.

According to Alpha Expected Utility Theory, decisions are made on the basis of a weighted average of the minimum expected utility over the set D of probabilities and the maximum expected utility over this set:

$$AEU = \alpha \min_{p \in D} \sum_{i=1}^{3} p_i u_i + (1-\alpha) \max_{p \in D} \sum_{i=1}^{3} p_i u_i \tag{4}$$

The parameter α can be interpreted as an index of the ambiguity aversion of the decision maker. The larger is α the larger is the weight the decision maker gives to the pessimistic evaluation given by $\min_{p \in D} \sum_{i=1}^{3} p_i u_i$.

In order to estimate this model we need to characterise the set D. The theory offers no advice and we chose the simplest: that the set is defined by three lower bounds $\underline{p}_1$, $\underline{p}_2$ and $\underline{p}_3$ and the condition that every element in the set has $p_1 \geq \underline{p}_1$, $p_2 \geq \underline{p}_2$ and $p_3 \geq \underline{p}_3$. In addition, of course $p_1 + p_2 + p_3 = 1$ for each element in the set. These conditions imply that the set D is a triangle properly within the Marschak-Machina Triangle. It reduces to a single point, and hence AEU reduces to SEU, if $\underline{p}_1 + \underline{p}_2 + \underline{p}_3 = 1$.

Maxmin Expected Utility theory (NEU) and Maxmax Expected Utility theory (XEU) are special cases of AEU. If α is 1 then AEU reduces to the Gilboa and Schmeidler (1989) Maxmin Expected Utility model with a non-unique prior, and if α is 0 then AEU reduces to what we might term the "Gilboa and Schmeidler" Maxmax Expected Utility model.

Vector Expected Utility

The Vector Expected Utility (VEU) theory has been recently proposed by Siniscalchi (2009). In this model, an uncertain prospect is assessed according to a baseline expected utility evaluation and an adjustment that reflects the individual's perception of ambiguity and his or her attitude toward it. This adjustment is itself a function of the exposure to distinct sources of ambiguity, and its variability.

The key elements of the VEU model are a baseline probability and a collection of random variables, or adjustment factors, which represent acts exposed to distinct ambiguity sources and also reflect complementarity between ambiguous events.

The VEU model can be formally defined as follows:

$$VEU = \sum_{i=1}^{3} p_i u_i + A\left(\left(\sum_{i=1}^{3} p_i \zeta_{ji} uz_i\right)_{1 \leq j < 3}\right) \tag{5}$$

Here $p = (p_1, p_2, p_3)$ is the baseline prior; for $1 \leq j < 3$, each $\zeta_j = (\zeta_{j1}, .., \zeta_{j3})$ is an adjustment factor that satisfies $E_p[\zeta_j] = \sum_{i=1}^{3} p_i \zeta_{ji} = 0$; and A: $\mathbb{R}^n \to \mathbb{R}$ satisfies $A(0) = 0$ and $A(\phi) = A(-\phi)$. The function A is an adjustment function that reflects attitudes towards ambiguity. We need to specify the function $A(.)$ and also the values of the ζ[8]. After some simplification, we get that the VEU objective function takes the form, under these assumptions:

$$VEU = \sum_{i=1}^{3} p_i u_i - \delta(|u_1 - u_2| + |u_2 - u_3|) \tag{6}$$

[8] More details of the assumptions that we have made are available on the web-site: http://www-users.york.ac.uk/~jdh1/hey%20and%20pace/

This has intuitive appeal: decisions are made on the basis of expected utility 'corrected' for differences between the utilities of the various outcomes, weighted by a parameter δ that reflects the decision-maker's attitude to ambiguity.

THE CONTRACTION MODEL

Gajdos *et al* (2008) proposed a model (the "Contraction Model" or CM) in which it is possible to compare acts under different objective information structures. According to this theory, preferences are given by

$$CM = \alpha \min_{p \in P} \sum_{i=1}^{3} p_i u_i + (1-\alpha) \sum_{i=1}^{3} P_i u_i \quad (7)$$

where α measures imprecision aversion and P_1, P_2, P_3 is a particular probability distribution in the set D of possible distributions. It is what is called the 'Steiner Point' of the set - which is, in a particular sense, the 'centre' of the set. If we take the set D of possible distributions as all points (p_1, p_2, p_3) such that $p_1 + p_2 + p_3 = 1$ and $p_1 \geq \underline{p}_1, p_2 \geq \underline{p}_2, p_3 > \underline{p}_3$then the Steiner point is the point (P_1, P_2, P_3) where $P_i = \underline{p}_i + (1 - \underline{p}_1 - \underline{p}_2 - \underline{p}_3)/3$ for $i = 1, 2, 3$. We note that we have characterised this set D (of possible probabilities) in the same way as we have done for the Alpha Expected Utility model - as a triangle properly within the Marschak-Machina Triangle. The Steiner point is the 'central' point of this triangle.

II. Our Experimental Design

As we have already noted, in our experiment ambiguity was implemented with a Bingo Blower and subjects were presented with a set of allocation problems, which were determined after extensive Monte Carlo simulations.

Subjects completed the experiment individually at screened computer terminals. They were given written instructions and then showed a PowerPoint presentation of the instructions. There was a Bingo Blower in action at the front of the laboratory throughout the experiment.[9] The Bingo Blower is a rectangular-shaped, glass-sided, object some 3 feet high and 2 feet by 2 feet in horizontal section. Inside the glass walls are a set of balls in continuous motion being moved about by a jet of wind from a fan in the base. In addition, images of the Blower in action were projected *via* a video camera onto two big screens in the laboratory. Subjects were free at any stage to go closer to the Blower to examine it as much as they wanted. All the balls inside the Blower can at all times be seen by people outside, but, unless the number of balls in the Blower is low, the number of balls of differing colours can not be counted because they are continually moving around. Hence the balls in the bingo can be seen but not counted (unless the total number of balls is low), and the information available is not sufficient to calculate objective probabilities. This ensures that, while objective probabilities do exist, the decision-makers cannot know them. In this way, we have created a situation of genuine ambiguity which eliminates the problem of suspicion; the problem of using directly a second-order probability distribution; and the problem of using real events, therefore keeping the problem more similar to the original Ellsberg one (Ellsberg, 1961). We note that a further advantage of this way of creating ambiguity in the laboratory is the fact that the information available is the same for all subjects. Hence there is no role for the so called ‘comparative ignorance’ (Fox and Tversky, 1995), and hence we can exclude such a factor as a possible explanation of behaviour.[10]

This Bingo Blower played a crucial role in representing ambiguity and in providing incentives. Inside the Bingo Blower were balls of three different colours: pink, yellow and blue. The number of each colour depended on the treatment:

	Treatment 1	Treatment 2
pink	2	8
yellow	3	12
blue	5	20

[9] On our site http://www-users.york.ac.uk/~jdh1/hey%20and%20pace/ can be found the instructions (both in Word and PowerPoint) and videos of the Bingo Blower, as well as screenshots of the experimental software.

[10] One criticism concerning the implementation of ambiguity in the lab using the Bingo Blower comes from Morone and Ozdemir (2011). The criticism consists of the observation that the ability of getting the right probabilities is subject specific; that is, subjects have different counting skills, or might have problems in the perception of colours. This criticism may be true but it is not clear how this could affect the validity of the Bingo Blower in generating ambiguity in the lab.

In Treatment 1 the balls of each colour could be counted, though one might not be sure of the number of blue balls; this was *the least ambiguous treatment*. In Treatment 2 the balls of each colour could not be counted; this was *the most ambiguous treatment*. Note that in this latter treatment subjects could get some idea of the relative numbers of balls of the different colours but not count the numbers precisely. It was reasonably clear that there were more blue balls than yellow, and more yellow than pink, though precise calculations could not be made.

Sixty six subjects completed Treatment 1 and sixty three completed Treatment 2. In both treatments, subjects were presented with a total of 76 questions. Each of these asked them to allocate a given quantity of tokens between the colours. There were two *types* of question. Type 1 asked then to allocate the tokens between *two* of the colours; Type 2 asked them to allocate the tokens between one of the three colours and the other two. In each problem subjects were told the *exchange rate* between tokens and money for each of the colours in the question. Thus an allocation of tokens implied an allocation of money to two or three of the colours.

We provided an incentive for carefully choosing the allocations with the following payment scheme. We told subjects that, after answering all 76 questions, one of the questions would be chosen at random, and the subject's allocation to the two or three colours for that problem retrieved from the computer. At that point the subject and the experimenter went over to the Bingo Blower, and the subject rotated the tube to expell one ball. The colour of the ball, the question picked at random and their answer to that question determined their payment. To be precise: if the problem chosen was one of Type 1, then they would be paid the money implied by their allocation to the colour of the ball expelled; if it was the colour not mentioned in that question they would be paid nothing; if the problem chosen was one of Type 2, then they would be paid the money implied by their allocation to the colour of the ball expelled. In addition they received a show-up fee of £5. They filled in a brief questionnaire, were paid, signed a receipt and were free to go. A total of 129 subjects participated at the experiments, 40 of them at CESARE at LUISS in Rome (Italy) and the remaining 89 at EXEC at The University of York (UK). In both cases, subjects were recruited using the ORSEE (Greiner 2004) software and the experiment was run using a purpose-written software written in Visual Basic 6.[11]

III. Related Experimental Literature

Having described our experimental implementation and motivation we are now in a position to survey the relevant experimental literature in more detail. We confine ourselves to recent important contributions to the literature; earlier literature is surveyed in Camerer and Weber (1992) and Camerer (1995).

Hey *et al* (2010), using the same implementation of ambiguity in the laboratory as we use here, also with three possible outcomes, but asking a large number (162) of pairwise choice questions, examined the descriptive and predictive ability of twelve theories of behaviour under ambiguity: some very old and not using a preference functional (proceeding directly to a decision rule) such as the original MaxMin and MaxMax; and some very recent, such as the Alpha Expected Utility model. The findings were that the very old simple models (those without a preference function) were largely discredited, but that more modern and rather sophisticated models (such as Choquet) did not perform sufficiently better than simple theories such as Subjective Expected Utility theory. Estimation of the preference functions was done using maximum likelihood techniques with the stochastic specification determined by a model of how subjects made errors in their pairwise choices.

Ahn *et al* (2010) used allocation questions, like we do here, but implemented ambiguity by simply not telling the subjects the true objective probabilities of two of the three possible outcomes of the experiment. They did not look at the predictive ability of any models; neither did they examine the descriptive performance of any specific theory. Instead they examined two broad classes of functionals, smooth and kinked, which are special cases of various theoretical models that we specifically estimate. Econometrically, they estimated, subject by subject, the risk-aversion parameter of an assumed

[11] Which can be found and downloaded from http://www-users.york.ac.uk/~jdh1/hey%20and%20pa.ce/

Constant Absolute Risk Aversion utility function, and a second parameter measuring ambiguity aversion, using Non Linear Least Squares (NLLS), that is by minimising the sum of squared differences between actual allocations and the theoretically optimal allocations for those risk and ambiguity aversion coefficients. Interestingly they comment in a footnote that "...for simplicity, the estimation technique for both specifications is NLLS, rather than a structural model using maximum likelihood (ML). We favor the NLLS approach, because it provides a good fit and offers straightforward interpretation." They do not give sufficient detail to make clear exactly what this rather cryptic comment means.

Halevy (2007) implemented ambiguity in the laboratory using traditional Ellsberg Urns and asked reservation price questions. Because of the way that his 'Ellsberg Urns' were implemented, his set of models includes some models that we do not consider here, particularly two-stage-probability models such as Recursive Nonexpected Utility and Recursive Expected Utility. But we include some that he does not - making the two papers complementary. He used *reservation price* questions; we should describe and discuss these - as they are an alternative to pairwise choice questions and to allocation questions. Essentially he wants to know how much subjects value bets on various events. Let us consider a particular Ellsberg Urn and a particular colour. The subject is asked to imagine that he or she owns a bet which pays a certain amount of money ($2) if that coloured ball is drawn from that particular urn. Halevy wanted to elicit the subject's reservation price for this bet; this reservation price telling us about the subject's preferences. Halevy used the Becker-DeGroot-Marschak mechanism: "the subject was asked to state a minimal price at which she was willing to sell the bet... The subject set the selling price by moving a lever on a scale between $0 and $2. Then a random number between $0 and $2 was generated by the computer. The random number was the “buying price” for the bet. If the buying price was higher than the reservation price that the subject stated, she was paid the buying price (and her payoff did not depend on the outcome of her bet). However, if the buying price was lower than the minimal selling price, the actual payment depended on the outcome of her bet." This BDM technique is well-known in the literature, but is complicated to describe and difficult for subjects to understand. Moreover there are well-known problems, see Karni and Zafra (1979), with using this technique when preferences are *not* expected utility preferences - which, of course, is precisely the concern of the paper. Halevy did not use his data to estimate preference functionals and hence did not compare their descriptive and predictive power; instead he carried out an extensive set of tests of the various theories. Unfortunately this econometric procedure does not help to draw unique conclusions about the 'best' preference functional, even for individual subjects. Indeed Halevy concludes that his "...findings indicate that currently there is no unique theoretical model that universally captures ambiguity preferences".

Abdellaoui *et al* (2011) investigated only Rank Dependent Expected Utility theory. They did not explicitly examine its descriptive (nor predictive) ability, being more concerned with the effect on the estimated utility and weighting functions of different *sources* of ambiguity. As we have already noted, they implemented ambiguity in the laboratory in two ways: in one part of the experiment, using 8-colour 'Ellsberg Urns'; and in the other part using 'natural' events. They elicited certainty equivalents (or reservation prices) in order to infer preferences, not using the BDM mechanism (presumably because of the problems we have alluded to above), but instead using Holt-Laury price lists.[12] This mechanism seems to be a better way of eliciting certainty equivalents, even though the outcome does appear to be sensitive to the elements in the list - the number of them and their range. The resulting certainty equivalents are a valuation, just as Halevy's reservation prices, even though they come from a set of pairwise choice questions. However econometrically it must be the case that the valuation resulting from a list with n elements is less informative than n independent pairwise choice questions. They estimated utility functions (assumed to be power or CRRA) "using nonlinear least squares estimation with the certainty equivalent as dependent variable"; similarly they estimated the weighting function by "minimising the quadratic distance". They do not explain why.

Anderson *et al* (2009) use a technique similar to that used by Ahn *et al* (2010) in estimating two parameters (one a measure of risk aversion and the other a measure of

[12] In which subjects are presented with a set of pairwise choices arranged in a list. In each pair subjects are asked to choose between some ambiguous lottery and some certain amount of money. As one goes down the list, the certain amount increases. The subject's certainty equivalent is revealed by the point at which the subject switches from choosing the lottery to choosing the certain amount. See Holt and Laury (2002).

ambiguity) in a minimalist non-EU model. They comment that this minimalist model comes either from the Source-Dependent Risk Attitude model or the Uncertain Priors model; in our termininology it is a two-stage-probability model[13] that looks exactly like Recursive Expected Utility. The bottom line is the following: suppose that there are I possible outcomes $i = 1, 2, ..., I$ with unknown probabilities. The decision-maker has a set of J possible values for these probabilities; we denote the j'th possible value $p_{1j}, p_{2j}, ..., p_{Ij}$ and the decision-maker considers that the probability that this is the correct set is π_j. The preference function is the maximisation of

$$\sum_{j=1}^{J} \pi_j v[\sum_{i=1}^{I} p_{ij} u(x_i)]$$

Note that there are two functions here: $u(.)$ which can be considered as a normal utility function, capturing attitude to risk; and $v(.)$ which can be considered as an ambiguity function; note that if $v(y) = y$ then this model reduces to Expected Utility theory. It is the non-linearity of $v(.)$ that captures aversion to ambiguity. Anderson et al (2009) assumed that both these functions are power functions - so that $u(x) = x^{\alpha}$ and $v(y) = y^{\beta}$ They estimated the two parameters α and β using maximum likelihood techniques (with careful attention paid to the stochastic specification) and rather heroic assumptions[14] about the π's and p's.

IV. Technical Assumptions

Before proceeding to our estimates we need to make some technical assumptions. In particular we need to specify our stochastic assumptions and those concerning the utility function. As far as the first is concerned we assumed a Fechner error for each subject's decisions. As far as the second is concerned, we chose to use the CARA function. We give more detail below.

[13] Chambers *et al* (2010) also investigate a generic Multiple Priors model.

[14] We note that the authors freely admit this and discuss the serious identification problems with two-stage-probability models.

Stochastic assumptions

We need to specify the stochastic nature of our data. We assume that the subject implements his or her optimal allocation in each decision problem with some error, either because of an error in its computation or an error in its implementation. So, rather than allocate x_i^* to colour i, the subject allocates $x_i^* + \varepsilon$. To complete this story, we have to make an assumption about the distribution of ε. We follow precedence and assume that it is Fechner error: that is, we assume that the ε are independent and identically normally distributed with zero mean and constant variance $1/s^2$. We estimate s along with the other parameters.[15]

Functional form of the utility function

We have to assume a particular form for the utility function of the subjects. We took this to be the CARA form:

$$\begin{aligned} u(x) &= \frac{1-\exp(-rx)}{1-\exp(-75r)} \text{ if } r \neq 0 \\ &= x/75 \text{ if } r = 0 \end{aligned} \tag{8}$$

We assume that the parameter r varies from subject to subject and we estimate it along with the other parameters. We chose this functional specification for a number of reasons: it has a single parameter; it produces good estimates and it leads to a nice functional form for the optimal allocations for the preference functionals under consideration. We can show this latter as follows. For all of the problems and most of the preference functionals the objective is to maximise a function of the form

$$w_1 u(e_1 x_1) + w_2 u(e_2 x_2)$$

subject to the constraint that $x_1 + x_2 = m$. Here m denotes the amount of tokens to

[15] We are aware of the work of Wilcox (2008) and of Blavatskyy (2011) both of which suggest that the distribution might be heteroscedastic. But, in the context of our experimental design, it is not clear how the heteroscedasticity should be specified.

allocate, x_i the tokens allocated to colour i and e_i the exchange rate between tokens allocated to colour i and money ($i = 1, 2$). The weights w_1 and w_2 depend upon the problem and the preference functional. The solution to this problem can be shown to be:

$$x_1^* = \frac{e_2 m + \{\ln[(w_1 e_1)/(w_2 e_2)]\}/r}{e_1 + e_2}$$
$$x_2^* = \frac{e_1 m + \{\ln[(w_2 e_2)/(w_1 e_1)]\}/r}{e_1 + e_2}$$

We note that there is no guarantee that the optimal allocations are positive and less than m. In the experiment subjects were constrained to have all allocations non-negative and their sum equal to m and we took that into account in the estimation.

In partial defence of the assumption to use a CARA specification, we note that if we used instead the Constant Relative Risk Aversion (CRRA) utility function, then the optimisation would imply an allocation of a *proportion* of the initial tokens to each colour, with the optimal proportional being *strictly* between 0 and 1. This clearly contradicts the empirical evidence which frequently showed all-or-nothing allocations to particular colours. The CARA utility function has the advantage that it easily accomodates boundary portfolios, while the CRRA specification does not.

V. Results

We estimated each of the 7 preference functionals for each of the 129 subjects on a subset of the data - precisely a randomly chosen 60 of the 76 questions[16] - using the constrained maximum likelihood procedure in GAUSS. We thus have, for each preference functional and each subject, estimates of the parameters of the functional, of s, the precision, and of r, the risk aversion parameter. In addition, we have the maximised log-likelihood. We then used, for each subject and each preference the estimated parameters to predict behaviour on the remaining 16 questions. This gives us a prediction log-likelihood for each functional and for each subject - this is, of course, a measure

[16]Because the subjects received the 76 questions in different orders (and with the colour on the left and the colour on the right randomly selected) this means that the position of the 60 estimation questions (and hence the 16 prediction questions) varied from subject to subject, but for each subject they were randomly positioned).

of the predictive ability of the theory. All this information is available on our site. In synthesising these results, we first present information from the estimation part and then from the prediction part.

A. Estimation

This section reports our results on estimation. While we are mainly interested in the goodness-of-fit of the various models, we start with some observations about treatment effects. We then turn to goodness-of-fit, as measured by the maximised log-likelihoods, and a comparison of the goodness-of-fit of the various models, necessarily incorporating corrections of the log-likelihoods for different degrees of freedom (numbers of parameters).

We had two treatments: Treament 1 with very little ambiguity; Treatment 2 with considerable ambiguity. We had expected to find treatment effects in our estimates. One place that we did *not* find such effects was in the goodness-of-fit of the various models. The mean log-likelihoods (both fitted and prediction, corrected and uncorrected) and their standard deviations, are shown in Tables 2 and 3. The means of the log-likelihoods are, with one exception, always higher in Treatment 2 than in Treatment 1, but, at the same time, the standard deviations are higher in Treatment 2 than in Treatment 1. An examination of histograms of the data shows the reason - the distributions of the log-likelihoods for Treatment 1 are symmetrical while those for Treatment 2 are skewed to the left.[17] So one can *not* conclude that the fit is generally better in Treatment 2 than in Treatment 1. Indeed this conclusion comes out of the t-tests reported in Tables 2 and 3, which show clearly that there are no treatment effects in the average goodness-of-fit, neither fitted nor prediction.

The same conclusion comes from an examination of the estimates of the precision parameter s. Examine Table 1, and, in particular, the rows reporting the average estimates of s in the various models in the two treatments. Nowhere are the differences significant. So subjects on average are equally noisy, equally likely to make mistakes, in the

[17] The histograms are on the website. We have no explanation for these differently shaped distributions.

two treatments. This is quite reassuring: subjects, on average, are not reacting to the increased ambiguity by becoming less precise.

What they are doing can be seen from the other average parameter estimates presented in Table 1. These are averages over the 66 subjects in Treatment 1 and the 63 subjects in Treatment 2. Starting with the SEU estimates, we note that the average estimates in Treatment 1 of the subjective probabilities attached to pink, yellow and blue were 0.2186, 0.3131, and 0.4683 respectively, while in Treatment 2 they were 0.2300, 0.3481 and 0.4219. Recall that the true probabilities in both treatments were 0.2, 0.3 and 0.5. So in Treatment 1 they were very close on average to the true probabilities, while in Treatment 2 they were on average further away, and closer to equal probabilities. Note also that the average estimated probabilities for the yellow and blue average probabilities are significantly different between the two treatments. So this seems to be the reaction to the increased ambiguity between treatments: not a change in precision, but a change in the estimates of the subjective probabilities. The average risk aversion and precision estimates do not differ significantly across treatments, which is interesting and reassuring: on average we do not have different subjects in the two treatments.

A similar pattern is evident across all theories: for example with Choquet there are significant differences in the average capacities attached to yellow and blue, and significant differences in the average capacities attached to (pink and blue) and to (pink and yellow) between the two treatments. In fact, we see very similar treatment effects on average parameter estimates across all models, with the possible exception of the Contraction Model. With that model the set of possible distributions is slightly larger, and the alpha parameter also slightly larger, in Treatment 2 compared to Treatment 1, but not significantly so.

We now examine the goodness-of-fit of the various models - as measured by the maximised log-likelihoods on the 60 observations used for fitting. Table 2a shows the average log-likelihoods over each of the various theories. In both treatments, CEU comes out as the best, followed by CM and by the AEU. However, this is an unfair comparison as CEU has more parameters than all the other models. Indeed, as the models differ in terms of the numbers of parameters involved in their estimation, we should correct these log-likelihoods for the number of degrees of freedom.[18] We do so using the Bayes Information Criterion[19] and hence we get Table 2b. From this we see that in Treatment 1, all the models perform very similarly, with AEU and CM marginally better (though SEU does perform strongly). while in Treatment 2 CM emerges as the overall winner.

However these average figures obscure lots of detail. Moreover, we are not trying to find one model that is best for everyone. Therefore let us look in more detail at these log-likelihoods, both uncorrected (for degrees of freedom) and corrected, and, in particular, let us look at the *rankings* of the various models.

We start with the uncorrected log-likelihoods for the 7 functionals. In Table 4 the number in brackets after the name of the functional is the number of estimated parameters. The table shows the percentage of subjects for which the uncorrected log-likelihood of each preference functional came in the various possible positions - 1^{st} through to 7^{th}. We also show the average ranking for each functional, obtained assigning a value of 1 to the first position in the ranking and a value of 7 to the 7^{th} position in the ranking. This means that the lower is the average ranking, the better a particular model is performing from the descriptive point of view. We note that not surprisingly, the models with more parameters, CEU, AEU and CM do well on this ranking, while the models with few parameters are penalized. We once again notice very few differences between the two treatments.

If we correct the log-likelihoods for the number of parameters (using the Bayesian Information Criterion) and redo the rankings we get Table 5. On this basis, on average, NEU, SEU and CM do well, and interestingly because they very rarely end up in a very bad position. In contrast, CEU does rather poorly (after being penalised for degrees of freedom) not only because it does not end up in good positions very often but also because it comes out worst rather often.

We follow convention[20] (though not necessarily approving of it) by reporting signifi-

[18]The number of parameters are as follows: SEU 4, CEU 8, AEU 6, NEU 5, XEU 5, VEU 5 and CM 6. Notice that in SEU and VEU the three probabilities add to 1. We include in this count the parameters common to all models, namely the risk aversion parameter and the precision.

[19]Though see our comments later.

[20]However we do not follow what seems to be the convention in the related literature in reporting tests of GARP. This

cance tests. We use different kinds of tests for different pairs of models. If one model is nested inside another[21] we test whether the latter fits significantly better than the former by using standard log-likelihood ratio tests. If neither model is nested inside the other we use Clarke (Clarke 2007) tests. Using a significance level of 5% we get Table 6; at 1% we get Table 7. In each table the entry reports the percentage of subjects for whom the row model is significantly better on these criteria than the column model. We are looking for models where the row numbers (except for the main diagonal) are large and the column numbers are small. On this basis both CEU and the CM look good.

The results using significance tests seem to conflict with the picture painted by the corrected fitted log-likelihoods. This is partly because there is no consensus amongst econometricians as to what is the correct 'correction' to apply to fitted log-likelihoods to correct for degrees of freedom: some econometricians advocate the Akaike Information Criterion (AIC), other the Bayesian Information Criterion (BIC). The former uses $2k - 2LL$ where k is the number of degrees of freedom and LL the log-likelihood; the latter uses $k\ln(n) - 2LL$ where n is the number of observations used in estimation. If n is 8 or more, the BIC punishes the log-likelihood more heavily for its degrees of freedom than does the AIC. In our experiment the n used for fitting was 60 so there is heavy punishment for degrees of freedom. Moreover, these corrections are answering different questions than those that the significance tests attempt to answer.[22] So it is hardly surprising that we get different answers.

The overall impression that emerges from the goodness-of-fit analyses, particularly if we correct for the numbers of parameters involved in the estimation, is that there are not enormous differences between the various models. Tables 2b and 5 make this point particularly clearly. Let us now see if we get a clearer picture from the prediction log-

is for three reasons: (1) to apply a GARP test it must be the case that income is constant, while in our experiment token income varied across the 76 allocation tasks; (2) the GARP analysis is useful when two allocations commodities are symmetric (that is, an individual cares only about the relative price of these goods and does not have any particular taste on commodities themselves). In our paper this does not apply because we never have symmetric situations (we consider an ambiguous setting with unknown objective probabilities); (3) we are not sure what these tests tell us other than that the behaviour of our subjects was not random (which is rather obvious from the results that we do report).

[21] SEU is nested inside all the other models, NEU is nested inside CM, and NEU and XEU are both nested inside AEU.

[22] Note that the critical values of the chi-square statistic do not rise linearly with the number of degrees of freedom, while both the Akaike and Bayesian corrections do.

likelihoods.

B. Prediction

We turn to predictive ability. As we have already noted, we follow the Wilcox recommended procedure (Wilcox, 2011), in which the models are estimated on a subset of problems (60 of them), and then these estimates are used to predict choices on the rest of the problems in the experiment (16 of them).[23] Doing this we can compare the predictions with actual behaviour and compute the appropriate log-likelihoods, and, in principle, come up with recommendations about the use of theories. However, how one should do these comparisons is not clear, and there seems to be no clear message emerging from the data. Or perhaps that **is** the message: there is no clear winner amongst the preference functionals.

We present a number of different analyses of the prediction log-likelihoods. Our first set of analyses assumes that we simply use the models for prediction and do not consider their descriptive ability. We then move on to asking whether prior information (about the best-fitting functional) improves our predictive ability. Within the first set of analyses, we start with simple histograms. We then rank the preference functionals on the basis of their predictive ability, and then see how often each functional comes first in predictive power. Both these measures are solely concerned with the *ordering* of predictive ability, however, and do not consider the magnitude of the differences in prediction. So we then introduce a *cardinal* measure of relative predictive ability. We then move on to our second set of analyses, where we assume that the person making the prediction about a particular individual is aware of the best-fitting functional for that individual. We ask whether this improves prediction, and present two different analyses in an attempt to answer it.

We begin with predictions unconditional on fitting, and present in Figure 1 histograms of the prediction log-likelihoods over all subjects. These are rather remarkable, showing

[23] Stictly speaking Wilcox (2011) advocates what he calls out-of-context predictions, which are possible in his context (pairwise choice) but not in ours. The split of the 76 questions between those used for fitting (60) and those for prediction (16) was chosen rather arbitrarily.

a property that is also true for the fitted log-likelihoods: the distributions across subjects are very similar for the different models. In Figure 2 we add substance to this observation by drawing scatters of the SEU log-likelihoods, against each of the other's log-likelihoods. This makes the point that the prediction log-likelihoods are strongly correlated and almost identical across preference functionals. The overwhelming impression that one gets from these figures is that there is more variability between subjects than between preference functionals. This is also the case with the Bayesian corrected log-likelihoods which we show in Figure 3, and on which we have already commented.

Let us try and delve deeper. There are various ways that we can do this. First, we rank the various preference functionals on the basis of their predictive ability. Doing this gives us Table 8a. We extract from this the average rankings (Table 8b), obtained, as for the estimation analysis, by assigning a value of 1 to the first position in the ranking and a value of 7 to the 7^{th} position in the ranking. Thus the lower is the average ranking, the better a particular model is performing from the prediction point of view. Further we tabulate in Table 8c the number of times each functional was best in predicting. What do we conclude from Tables 8a, 8b and 8c? That, on average, if we know nothing at all about our subjects we would be very little better off using one model rather than any other: CM is slightly better than the others in Treatment 1 and the AEU in Treatment 2, but the differences are trivial. Indeed what is rather remarkable about these three Tables is how similar the various models are in their predictive ability. However, one could argue that this conclusion is misleading as it is based on *rankings* and does not worry about *magnitudes*. Let us therefore present a final analysis, the results of which are presented in Table 9.

To construct the table we have calculated in percentage terms how much higher is the highest prediction log-likelihood (obviously obtained by the best predicting model) relative to each other model's prediction log-likelihood and the averaged this across all subjects. In this way we get a measure of how much poorer each model is compared to the best predicting model. For example SEU has a prediction log-likelihood which is bettered on average by 13.17% by the best-predicting model (which of course varies from subject to subject). The interesting point about Table 9 is the similarity between the numbers - with the sole exception of CM. If one uses CM instead of the best-predicting model then one has a relatively small deterioration of 8.13% with respect to the maximum prediction log-likelihood.

However the differences are small: the conclusion coming from all the above is that, regardless of how one analyses the data, there is very little difference between theories in terms of predictive ability. So if one knows nothing about a particular individual and wants to predict their behaviour, one might as well use any theory. However, this does not mean that certain theories might not come into their own if one *does* know something about the individual whose behaviour you want to predict. Consider Table 10, in which the columns indicate the best fitting model (as determined by the Bayes corrected log-likelihood) and the table entries indicate the average ranking on prediction using the row model. So, for example, if SEU fits the best according to this criterion (it happens for 24 percent of the subjects), then using SEU to predict has an average ranking of 3.75, using CEU the same, using AEU 4.25 and so on. Ideally the diagonal entries should be the smallest in each column - on the grounds that the best fitting should be the best predicting, but this is not always the case. One clear exception is CM - when it is the best fitting it also ranks highest for prediction - in both treatments. Indeed CM generally does rather well on this prediction-ranking criterion. But overall there is again the same message: there is not an enormous difference in predictive ability between theories.

One might object to the analysis above in that it uses *rankings* - which do not worry about magnitudes but just orderings. Accordingly we include Table 11 where once again the columns indicate the best fitting model (using the Bayes corrected log-likelihood) while the rows give the square root of the average sum of squared deviations between the actual allocation and that predicted from the row model; this is a measure of the goodness of fit of the predictions. Once again, we would expect to have the smallest entries in any particular column along the main diagonal, but this is not the case. When SEU or AEU (in Treatment 1) or CM are the best, then it is best respectively to use SEU, AEU and CM for prediction purposes, but when, for example, CEU is the best fitting it seems to be the worst in prediction.The bottom line from Table 11 is that SEU would be the best to use for prediction in Treatment 1 and CM in Treatment 2 if these are the best-fitting models. Similarly AEU (and its subcases NEU and XEU) in Treatment 1; but if CEU or VEU are the best-fitting, one might as well use SEU for prediction. The one thing that sticks out from Treatment 2 is that CM does rather well everywhere.

It seems to be really difficult to come up with an answer to the question: "having fitted all the models to the data for any individual, which is the best model to use for predicting the behaviour of that individual". It does seem to be the case that in Treatment 1 SEU does not do badly, while in Treatment 2, if either AEU or CM fits best then one should use respectively AEU and CM to predict, but apart from that one would just be as well of using any model to predict. As we said earlier there is more noise between subjects than between preference functionals.

VI. Conclusions

In one sense Figures 2 and 3 tell almost everything: there is much more difference between subjects than between preference functionals. This is evidenced by the fact that: (1) the average (across all subjects) of the standard deviation (across all preference functionals) of the corrected log-likelihoods is just 6.46, while (2) the average (across all preference functionals) of the standard deviation (across all subjects) of the corrected log-likelihoods is 85.96. Similarly: (3) the average (across all subjects) of the standard deviation (across all preference functionals) of the prediction log-likelihoods is just 2.16, while (4) the average (across all preference functionals) of the standard deviation (across all subjects) of the prediction log-likelihoods is 20.75.[24] The similarity of the theories in prediction is particularly galling for the theorists. However, there are two models that seem to be potentially better than SEU - namely AEU and, particularly CM. CEU seems to simply have two many parameters. VEU does not seem to be a serious improvement

[24]Treatment 1 and 2 values are as follows:
1: 6.32 and 6.59
2: 76.22 and 95.50
3: 1.86 and 2.47
4: 19.72 and 21.88

on SEU. These results seem to be true both for Treatment 1 and Treatment 2, though we could recommend the use of SEU in Treatment 1 and CM in Treatment 2.

We return to the treatment difference that we noted at the beginning of the Estimation section: namely, the differences in the estimated parameters across treatments. This difference seems to be more important than the differences in the goodness-of-fits of the various models. So, it seems to be more the case that subjects *change their perception* of the ambiguity rather than *change their reaction* to it. Obviously we do not have subjects who did both treatments, so what follows is a conjecture: if they are SEU in Treatment 1 then they remain SEU in Treatment 2, but change their perception of the probabilities. And the same, *mutatis mutandis,* with the other models. This is important for any theorist doing comparative statics.

There remains the large variance across subjects. For some subjects - those with large log-likelihoods (and hence those for whom the models fit well) - we may be capturing their behaviour with one or other of these models. However, for those subjects with small log-likelihoods (those for whom the estimated models fit badly) we seem not to be capturing their behaviour. It follows either that *they are very imprecise in their decision-making, or that none of these preference functionals describe their decision-making correctly.* If we knew which of these was true, we would know what to do next. With the first explanation we should spend more time investigating their errors, rather than creating new preference functionals[25]; while with the second we should do exactly the opposite: produce more empirically valid theories rather than worry about errors. But given the theoretical activity of the past decade, one may well ask what kind of theories these should be.

The many recent theories, not only those in the set that we have been considering, but also, and particularly, those in the two-stage-probability set, can be characterised by the sophistication of the decision-maker. While we understand and appreciate the "as-if" methodology of economics, we suspect that these sophisticated characterisations are moving in the wrong direction. The allocation problem under ambiguity that we

[25] Which is a point with which Wilcox (2008) would agree.

have posed to our subjects is a difficult one; we suspect that many of our subjects were adopting a strategy of simplifying the problem first before applying a simple preference functional. They may well be editing the problem before evaluating it. So perhaps the message is that we experimentalists need to try and observe this editing process, and then ask the theorists to produce a simplified theory (not necessarily acting through preference functionals) to tackle the simplified problem. Sophisticated theory does not seem to work.

References

Abdellaoui, M., Baillon,A. Placido, L. and P. Wakker (2011), "The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation", American Economic Review, 101: 695-723.

Anderson, S., Fountain, J., Harrison, G.W. and E.E. Rutström (2009), "Estimating aversion to uncertainty", Working Paper.

Ahn, D.S., Choi, S., Gale, D., and S. Kariv (2010), "Estimating Ambiguity Aversion in a Portfolio Choice Experiment", Working Paper.

Blavatskyy P. (2011), "A Model of Probabilistic Choice Satisfying First-Order Stochastic Dominance", Management Science, 57, 3: 542-548.

Camerer, C.,and M. Weber (1992), "Recent Development in Modeling Preferences: Uncertainty and Ambiguity", Journal of Risk and Uncertainty, 5,4: 325-370.

Camerer, C. (1995), "Individual Decision Making" in J. Kagel and A. Roth (Eds.), Handbook of Experimental Economics, Princeton University Press, 587-703.

Chambers, R.G., Melkonyan, T. and D. Pick (2010), "Experimental Evidence on Multiple Prior Models in the Presence of Uncertainty", Working Paper.

Choi, S., Fisman R., Gale, D., and S. Kariv (2007), "Consistency and Heterogeneity of Individual Behavior under Uncertainty", The American Economic Review, 97, 4: 1921-1938.

Clarke, K.A. (2007), "A Simple Distribution-Free Test for Non-nested Model Selection", Political Analysis, 15: 347-363.

Ellsberg, D. (1961), “Risk, Ambiguity and the Savage Axiom”, Quaterly Journal of Economics, 75: 643-669.

Fox, C.R. and A. Tvesky (1995), “Ambiguity Aversion and Comparative Ignorance”, Quaterly Journal of Economics, 110, 3: 585-603.

Gajdos, T., Hayashi T., Tallon J-M and J-C Vergnaud (2008), “Attitude toward imprecise information ”, Journal of Economic Theory, 140, 27-65.

Ghirardato, P., Maccheroni F., and M. Marinacci (2004), “Differentiating Ambiguity and Ambiguity Attitude”, Journal of Economic Theory, 118, 2: 133-173.

Gilboa, I. and D. Schmeidler (1989), “Maxmin Expected Utility with Non-Unique Prior”, Journal of Mathematical Economics, 18: 141-153.

Greiner, B (2004), “The Online Recruitment System ORSEE 2.0 – A Guide for the Organization of Experiments in Economics”, University of Cologne Discussion Paper (www.orsee.org).

Halevy, Y. (2007), “Ellsberg Revisited: An Experimental Study”, Econometrica, 75, 2: 503-536.

Hey, J.D., Lotito G. and A. Maffioletti (2010), “The Descriptive and Predictive Adequacy of Theories of Decision Making Under Uncertainty/Ambiguity”, Journal of Risk and Uncertainty, 41, 2: 81-111.

Holt, C.A. and Laury, S.K. (2002). “Risk aversion and incentive effects”, American Economic Review, 92, 5: 1644-1655.

Karni, E., Safra, Z. (1987), "Preference reversals and the observability of preferences by experimental methods", Econometrica, 55: 675–685.

Maccheroni, F., Marinacci, M. and A. Rustichini (2006), “Ambiguity Aversion, Robustness, and the Variational Representation of Preferences”, Econometrica, 74: 1447-1498.

Morone, A. and O. Ozdemir (2011), “Displaying uncertainty information about probability: experimental evidence”, Bulletin of Economic Research, Forthcoming. DOI: 10.1111/j.1467-8586.2010.00380.x

Schmeidler, D. (1989), “Subjective Probability and Expected Utility without Additivity”, Econometrica, 57, 3: 571-587.

Siniscalchi, M. (2009), “Vector Expected Utility and Attitudes Toward Variation”, Econometrica, 77, 3: 801-855.

Wilcox, N. (2007), “Predicting Risky Choices Out-of-Context: A Monte Carlo Study”, University of Houston Working Paper.

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

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

A.

Tables and Figures

Theory	Parameter	All	Treatment1	Treatment2	t-test diff.
SEU	p_1	0.224	0.219	0.230	-0.590
	p_2	0.330	0.313	0.348	-3.531
	p_3	0.446	0.468	0.422	2.319
	r	0.118	0.120	0.116	0.184
	s	0.194	0.193	0.195	-0.059
CEU	w_1	0.203	0.200	0.206	-0.290
	w_2	0.255	0.230	0.282	-3.145
	w_3	0.403	0.429	0.376	2.157
	W_1	0.728	0.741	0.715	1.216
	W_2	0.556	0.576	0.535	1.893
	W_3	0.468	0.437	0.501	-2.304
	r	0.104	0.110	0.097	0.703
	s	0.223	0.224	0.221	0.101
AEU	w_1	0.165	0.161	0.168	-0.424
	w_2	0.268	0.254	0.284	-2.079
	w_3	0.381	0.402	0.358	1.800
	α	0.613	0.637	0.587	1.039
	r	0.123	0.122	0.124	-0.098
	s	0.210	0.212	0.208	0.110
NEU	w_1	0.192	0.186	0.198	-0.677
	w_2	0.290	0.274	0.306	-2.714
	w_3	0.403	0.423	0.382	1.777
	r	0.109	0.109	0.109	0.014
	s	0.204	0.206	0.203	0.081
XEU	w_1	0.191	0.188	0.195	-0.394
	w_2	0.301	0.288	0.315	-2.242
	w_3	0.416	0.441	0.389	2.295
	r	0.131	0.131	0.131	0.000
	s	0.198	0.197	0.199	-0.083

TABLE 1—DESCRIPTIVE STATISTICS ESTIMATED PARAMETERS

Theory	Parameter	All	Treatment 1	Treatment 2	t-diff test
VEU	p_1	0.231	0.226	0.235	-0.411
	p_2	0.328	0.311	0.346	-3.464
	p_3	0.441	0.463	0.419	2.160
	δ	0.027	0.030	0.024	0.618
	r	0.108	0.107	0.110	-0.151
	s	0.198	0.197	0.198	-0.042
CM	w_1	0.116	0.118	0.115	0.201
	w_2	0.216	0.205	0.227	-1.115
	w_3	0.330	0.356	0.303	1.607
	α	0.393	0.428	0.358	0.988
	r	0.107	0.108	0.107	0.066
	s	0.212	0.213	0.210	0.102

TABLE 1 — (CONT.)

A) Uncorrected log-likelihoods								
	SEU	CEU	AEU	NEU	XEU	VEU	CM	Obs.
All	-177.45	-169.12	-172.36	-174.11	-175.71	-175.75	-171.23	129
	(42.62)	(44.19)	(42.39)	(42.84)	(42.51)	(44.06)	(42.25)	
Treatment 1	-179.77	-171.18	-174.40	-176.14	-178.22	-178.26	-173.37	66
	(37.53)	(39.39)	(37.98)	(38.05)	(37.30)	(38.70)	(37.83)	
Treatment 2	-175.02	-166.96	-170.22	-171.97	-173.09	-173.11	-169.00	63
	(47.57)	(48.95)	(46.78)	(47.56)	(47.53)	(49.23)	(46.63)	
t-stat diff.	-0.631	-0.540	-0.558	-0.552	-0.683	-0.663	-0.586	

TABLE 2—A) AVERAGE FITTED LOG-LIKELIHOODS (STANDARD DEVIATION ARE IN PARENTHESIS)

B) Bayesian Information Criterion								
	SEU	CEU	AEU	NEU	XEU	VEU	CM	Obs.
All	371.27	370.99	369.29	368.68	371.89	371.96	367.03	129
	(85.25)	(88.38)	(84.78)	(85.67)	(85.03)	(88.12)	(84.49)	
Treatment 1	375.91	375.11	373.36	372.76	376.90	377.00	371.30	66
	(75.05)	(78.77)	(75.96)	(76.10)	(74.59)	(77.39)	(75.66)	
Treatment 2	366.41	366.68	365.01	364.41	366.65	366.69	362.56	63
	(95.14)	(97.91)	(93.57)	(95.11)	(95.06)	(98.47)	(93.26)	
t-stat diff.	0.631	0.540	0.558	0.552	0.683	0.663	0.586	

TABLE 2 — B) BAYESIAN INFORMATION CRITERION (STANDARD DEVIATION ARE IN PARENTHESIS)

	SEU	CEU	AEU	NEU	XEU	VEU	CM	Obs.
All	-50.46	-52.40	-50.57	-50.93	-50.30	-50.85	-49.97	129
	(18.90)	(25.26)	(20.11)	(20.23)	(18.98)	(21.70)	(20.06)	
Treatment 1	-51.02	-51.93	-51.03	-51.36	-50.79	-51.24	-50.50	66
	(18.50)	(22.58)	(20.06)	(19.62)	(18.63)	(18.87)	(19.77)	
Treatment 2	-49.86	-52.89	-50.08	-50.47	-49.79	-50.43	-49.42	63
	(19.44)	(27.97)	(20.31)	(21.00)	(19.47)	(24.46)	(20.50)	
t-stat diff.	-0.347	0.215	-0.267	-0.247	-0.299	-0.213	-0.304	

TABLE 3—AVERAGE PREDICTION LOG-LIKELIHOODS (STANDARD DEVIATION ARE IN PARENTHESIS)

Treatment 1								
Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
SEU (4)	0	0	0	0	0	15	85	6.85
CEU (8)	71	14	8	6	2	0	0	1.53
AEU (6)	5	50	30	5	5	3	3	2.76
NEU (5)	0	3	27	47	17	3	3	3.98
XEU (5)	0	0	14	9	39	29	9	5.11
VEU (5)	3	15	2	5	30	45	0	4.80
CM (6)	21	18	20	29	8	5	0	2.97
Treatment 2								
Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
SEU (4)	0	0	3	0	2	14	81	6.70
CEU (8)	56	29	8	5	3	0	0	1.71
AEU (6)	14	27	32	16	5	3	3	2.92
NEU (5)	0	3	21	40	24	10	3	4.25
XEU (5)	2	6	22	5	27	33	5	4.68
VEU (5)	2	21	2	11	29	35	2	4.56
CM (6)	27	14	13	24	11	5	6	3.17

TABLE 4—RANKINGS BASED ON UNCORRECTED FITTED LOG-LIKELIHOODS (ALL VALUES REPRESENT PERCENTAGES)

Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
Treatment 1								
EU (4)	24	17	6	11	18	24	0	3.55
CEU (8)	11	14	6	8	8	6	48	5.00
AEU (6)	9	18	24	3	18	23	5	3.89
NEU (5)	18	23	23	27	8	2	0	2.88
XEU (5)	6	6	12	23	23	14	17	4.58
VEU (5)	8	18	17	5	15	12	26	4.41
CM (6)	24	5	12	24	11	20	5	3.70
Treatment 2								
Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
EU (4)	21	16	11	19	17	16	0	3.44
CEU (8)	14	10	10	5	6	8	48	4.94
AEU (6)	10	16	19	11	13	25	6	4.03
NEU (5)	17	17	22	21	17	3	2	3.19
XEU (5)	6	16	14	17	14	19	13	4.25
VEU (5)	8	14	14	14	19	11	19	4.32
CM (6)	24	11	10	13	13	17	13	3.83

TABLE 5—RANKINGS BASED ON FITTED LOG-LIKELIHOODS CORRECTED USING THE BAYESIAN INFORMATION CRITERION (ALL VALUES REPRESENT PERCENTAGES)

Treatment 1

Model	EU	CEU	AEU	NEU	XEU	VEU	CM
EU	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CEU	64	n.a.	6	9	21	33	8
AEU	55	0	n.a.	35	59	8	0
NEU	56	0	n.a.	n.a.	3	11	n.a.
XEU	27	0	n.a.	0	n.a.	2	0
VEU	24	0	2	0	6	n.a.	0
CM	62	0	2	42	6	18	n.a.

Treatment 2

Model	EU	CEU	AEU	NEU	XEU	VEU	CM
EU	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CEU	57	n.a.	14	19	24	21	8
AEU	56	0	n.a.	32	56	3	0
NEU	48	0	n.a.	n.a.	6	3	n.a.
XEU	32	0	n.a.	0	n.a.	2	0
VEU	25	0	5	3	10	n.a.	2
CM	60	0	13	60	16	14	n.a.

TABLE 6—SIGNIFICANCE TESTS AT 5 PERCENT (RESULTS FROM THE CLARKE TEST ARE IN BOLD)

Treatment 1

Model	EU	CEU	AEU	NEU	XEU	VEU	CM
EU	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CEU	48	n.a.	5	6	9	9	5
AEU	42	0	n.a.	20	45	5	0
NEU	42	0	n.a.	n.a.	2	5	n.a
XEU	15	0	n.a.	0	n.a.	0	0
VEU	14	0	2	0	3	n.a.	0
CM	50	0	0	33	0	9	n.a.

Treatment 2

Model	EU	CEU	AEU	NEU	XEU	VEU	CM
EU	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CEU	48	n.a.	11	10	16	10	5
AEU	48	0	n.a.	30	33	0	0
NEU	33	0	n.a.	n.a.	2	2	n.a.
XEU	25	0	n.a.	0	n.a.	0	0
VEU	10	0	5	3	6	n.a.	2
CM	48	0	10	48	8	11	n.a.

TABLE 7—SIGNIFICANCE TESTS AT 1 PERCENT (RESULTS FROM THE CLARKE TEST ARE IN BOLD)

Treatment 1								
Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
EU	8	20	6	20	8	24	15	4.33
CEU	21	20	11	5	15	5	24	3.83
AEU	18	8	20	8	15	23	9	3.98
NEU	8	11	18	26	18	9	11	4.06
XEU	11	12	17	11	20	15	15	4.23
VEU	8	12	17	17	14	14	20	4.36
CM	27	18	12	15	11	11	7	3.20
Treatment 2								
Model	1st	2nd	3rd	4th	5th	6th	7th	Average Ranking
EU	8	17	6	13	24	21	11	4.33
CEU	22	13	10	5	6	10	35	4.29
AEU	14	22	13	14	22	14	0	3.51
NEU	5	17	22	19	17	6	13	3.97
XEU	8	16	14	10	14	19	19	4.40
VEU	17	10	24	11	6	17	14	3.89
CM	25	5	11	29	10	13	8	3.62

TABLE 8—A) RANKING BASED ON PREDICTION LOG-LIKELIHOODS (ALL VALUES REPRESENT PERCENTAGES)

Average Rankings	SEU	CEU	AEU	NEU	XEU	VEU	CM
Treatment 1	4.33	3.83	3.98	4.06	4.23	4.36	3.20
Treatment 2	4.33	4.29	3.51	3.97	4.40	3.89	3.62

TABLE 8 — B) AVERAGE RANKINGS BASED ON PREDICTION LOG-LIKELIHOODS

% of subjects for whom best	SEU	CEU	AEU	NEU	XEU	VEU	CM
Treatment 1	8	21	18	8	11	8	27
Treatment 2	8	22	14	5	8	17	25

TABLE 8 — C) MODELS FIRST IN THE RANKING FOR PREDICTION

	SEU	CEU	AEU	NEU	XEU	VEU	CM
All	14.75	24.00	12.34	14.26	14.18	13.33	8.13
Treatment 1	13.17	23.18	10.89	13.24	12.41	13.53	7.73
Treatment 2	16.42	24.87	13.85	15.32	16.07	13.13	8.54

TABLE 9—AVERAGE DEPARTURE FROM BEST PREDICTION

Treatment 1							
	SEU (24%)	CEU (11%)	AEU (9%)	NEU (18%)	XEU (6%)	VEU (8%)	CM (24%)
SEU	3.75	4.14	5.00	4.67	4.75	4.40	4.38
CEU	3.75	5.57	4.00	3.33	2.75	4.20	3.63
AEU	4.25	2.71	3.00	4.00	4.75	5.00	4.13
NEU	4.19	4.00	4.00	4.08	2.75	3.00	4.63
XEU	4.38	4.00	3.67	4.17	6.50	3.60	4.06
VEU	4.00	4.14	4.50	4.25	5.00	4.20	4.75
CM	3.69	3.43	3.83	3.50	1.50	3.60	2.44
Treatment 2							
	SEU (21%)	CEU (14%)	AEU (10%)	NEU (17%)	XEU (6%)	VEU (8%)	CM (24%)
SEU	4.54	3.89	4.67	3.55	3.75	5.20	4.73
CEU	5.46	4.56	3.17	4.82	3.75	2.80	3.80
AEU	2.54	3.56	3.33	4.27	3.75	4.20	3.53
NEU	3.77	4.11	4.50	4.45	2.75	3.80	3.87
XEU	3.62	4.44	3.33	4.27	5.50	5.60	4.87
VEU	4.23	3.00	4.33	2.82	4.25	3.20	4.87
CM	3.85	4.44	4.67	3.82	4.25	3.20	2.33

TABLE 10—AVERAGE RANKING ON PREDICTION OF THE ROW MODEL WHEN THE COLUMN MODEL FITS THE BEST

Treatment 1							
	SEU (24%)	CEU (11%)	AEU (9%)	NEU (18%)	XEU (6%)	VEU (8%)	CM (24%)
SEU	6.25	6.49	18.12	5.29	7.91	15.33	8.89
CEU	6.96	8.21	17.15	6.44	7.85	25.69	9.54
AEU	6.53	6.22	5.97	5.24	7.94	16.15	8.99
NEU	6.54	6.75	16.99	5.27	7.13	16.11	9.21
XEU	6.30	6.32	18.80	5.21	8.61	15.52	8.70
VEU	6.30	6.43	18.68	5.10	7.95	21.08	9.54
CM	6.30	6.60	6.96	5.17	6.60	16.13	8.15
Treatment 2							
	SEU (21%)	CEU (14%)	AEU (10%)	NEU (17%)	XEU (6%)	VEU (8%)	CM (24%)
SEU	10.39	5.07	25.61	7.19	17.42	6.35	10.76
CEU	15.67	6.94	15.12	10.78	15.54	8.48	10.12
AEU	10.90	5.09	26.65	6.99	17.26	6.20	10.69
NEU	10.52	5.31	26.59	7.02	16.98	6.13	11.60
XEU	10.87	4.94	33.38	7.22	20.66	6.35	10.74
VEU	12.30	4.92	29.51	7.00	17.67	6.02	11.48
CM	10.23	5.21	15.28	6.83	13.71	5.92	9.14

TABLE 11—ACTUAL VS PREDICTED ALLOCATION OF THE ROW MODEL WHEN THE COLUMN MODEL FITS THE BEST

FIGURE 1. HISTOGRAMS OF PREDICTION LOG-LIKELIHOODS

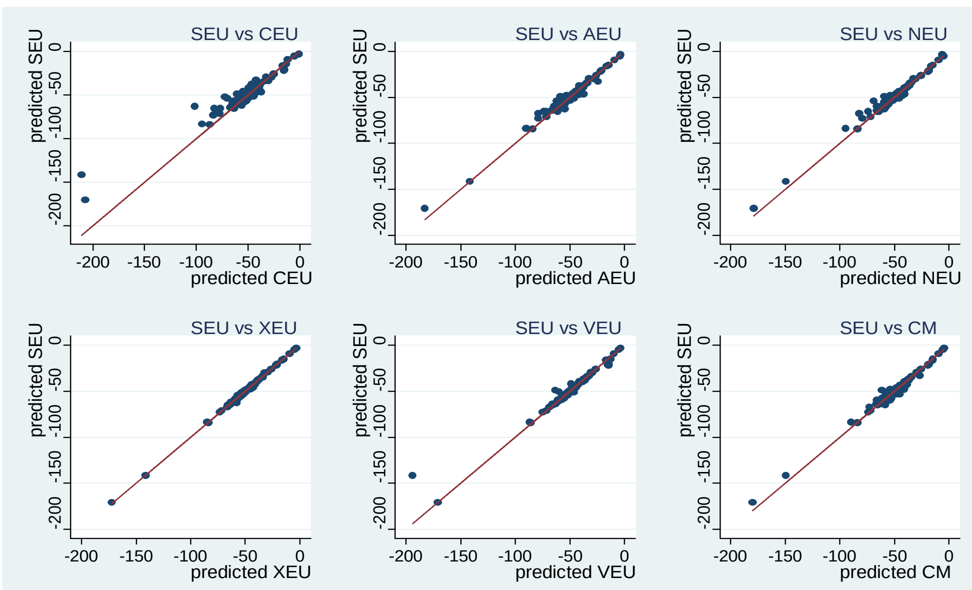

FIGURE 2. PAIRWISE COMPARISON PREDICTED LOG-LIKELIHOODS: SEU VS OTHER THEORIES

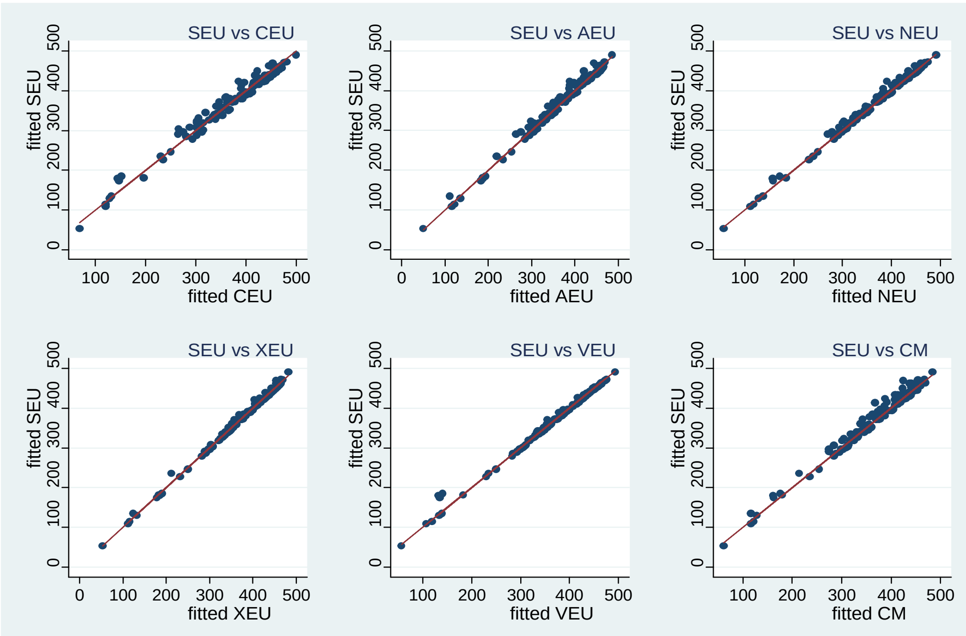

FIGURE 3. PAIRWISE COMPARISON FITTED (CORRECTED BIC) LOG-LIKELIHOODS: SEU VS OTHER THEORIES