Parametric preference functionals under risk in the gain domain: a Bayesian analysis

Article

Accepted Version

www.reading.ac.uk/centaur

CentAUR

Parametric Preference Functionals Under Risk in the Gain Domain: A Bayesian Analysis

Kelvin Balcombe

University of Reading

and

Iain Fraser

University of Kent

Address for correspondence:

Iain Fraser

School of Economics

University of Kent

Canterbury

Kent, CT2 7NP

UK

Tel: +44 (0)1227 823513

i.m.fraser@kent.ac.uk

Parametric Preference Functionals Under Risk in the Gain Domain: A Bayesian Analysis

Abstract

The performance of rank dependent preference functionals under risk is comprehensively evaluated using Bayesian model averaging. Model comparisons are made at three levels of heterogeneity plus three ways of linking deterministic and stochastic models: the differences in utilities, the differences in certainty equivalents and contextual utility. Overall, the "best model", which is conditional on the form of heterogeneity is a form of Rank Dependent Utility or Prospect Theory that captures the majority of behaviour at both the representative agent and individual level. However, the curvature of the probability weighting function for many individuals is S-shaped, or ostensibly concave or convex rather than the inverse S-shape commonly employed. Also contextual utility is broadly supported across all levels of heterogeneity. Finally, the Priority Heuristic model, previously examined within a deterministic setting, is estimated within a stochastic framework, and allowing for endogenous thresholds does improve model performance although it does not compete well with the other specifications considered.

1. Introduction

There is a long history of research questioning the validity of Expected Utility Theory (EUT), with many economists wishing to apply non EUT theories to problems relating to decisions under risk. Within the literature some degree of consensus appears to have emerged that probability weighting models such as Prospect Theory (PT) or Rank Dependent Utility (RDU) offer the best alternative to EUT (Wakker, 2010; Fehr-Duda and Epper, 2012). Yet, while the scope for applications of PT and RDU is increasing (see Barberis, 2013, and Schleifer, 2012), the growth of empirical applications is arguably less than one might expect given their theoretical prominence. One potential reason is that the range of parametric variants of these theories can itself be baffling, and perhaps may inhibit their adoption. Therefore, this paper reconsiders the appropriate selection of parametric specifications of choice under risk within the gain domain.

While there is plenty of evidence that most economic agents do not seem to unerringly use probabilities as summative linear weights to utilities of outcomes, what they actually do remains the subject of debate. Leading critics of EUT include Kahneman and Tversky (1979) and more recently Rabin (2000) and Rabin and Thaler (2001), who consider the weight of evidence against

2

EUT sufficient to label it an '*ex-hypothesis*'. In contrast, Birnbaum (2006) argues a case against probability weighting of the PT form, and more recently, the unfavourable implications for EUT from the concavity-calibration argument of Rabin (2000) have been challenged by Cox and Sadiraj (2006) and Cox et al. (2012) on the grounds that calibration arguments lead to equally problematic implications for nonlinear probability weighting. Furthermore, models in which outcomes are weighted by functions of probabilities have also been challenged at the process level (Fiedler and Glockner, 2012).

The understanding that emerges from the literature is further muddied by the fact that individuals may use different (and multiple) strategies. For example, Bruhin et al. (2010) report results that indicate that at least 20% of respondents in their experiments can be classified as EUT types, while Harless and Camerer (1994), and Hey and Orme (1994) have presented analysis of a range of theories and models suggesting that no one theory clearly outperformed all others. There is also the important question about how best to nest what are ostensibly deterministic theories within a stochastic setting. As Hey and Orme (1994) observed, while the issue of "noise" has often been treated as an ancillary one, it deserves greater attention such as the research presented by Wilcox (2011).

In practice, for applied researchers examining decision making under risk the implications of the above for conducting research come down to a choice of appropriate functional forms. Thus, the choice of functional forms to be employed to operationalise the theory is key and even if researchers narrow the range of candidate models to within the PT or RDU class[1], they face an enormous set of potential models. Furthermore, the literature is still unable to give definitive advice in this regard mainly because there are so many potential combinations of functional forms that are used to model the different aspects such as value (utility), probability transformations and those linking the deterministic models to stochastic outcomes. Each of these model aspects interacts with others to determine overall model performance and there is a need to understand how different model aspects perform in combination.

The data employed here is from Stott (2006), which to date provides one of the most comprehensive studies of the performance of a range of functionals characterising PT in the gain domain. The results reported in Stott are frequently cited for the choice of functionals employed in PT/RDU research (e.g., Bruhin et al., 2010). Unlike Stott (2006) we employ a Bayesian approach to the

---

[1]We use PT to mean its cumulative variant which is sometimes termed Cumulative Prospect Theory.

analysis of this data. The analysis in Stott (2006) and Booij et al. (2010) are typical in that they have been conducted from a classical perspective using maximum likelihood as the estimation method. However, serious issues emerge in deciding on an optimal combination of functional forms when there are so many combinations of competing specifications. Also, as noted by Booij et al. (2010) the wrong choice of a functional form can result in contamination and bias of other estimates of parameters. In this paper, we exploit the advantages of Bayesian Model Averaging (BMA) which provides an internally consistent and coherent approach to this type of modeling problem. Like Stott (2006), we examine a large number of functional forms at the individual as well as the aggregate level. Furtermore, we examine and compare specifications based on the 'contextual utility' approach developed in Wilcox (2011), a generalisation of the Priority Heuristic' (PH) developed in Brandstätter et al. (2006), and the 'Transfer of Attention Exchange' (TAX) weighting function of Birnbaum and Chavez (1997). The TAX and PH models are examined as they are viewed as alternatives to PT/RDU and in both cases positive experimental evidence has been presented.

The difficulty in deciding on an optimal specification for this type of problem stems from the very large model space. Classical pairwise comparison of nested models can be made using a range of standard tests (e.g. Likelihood Ratio, F, Wald) providing appropriate adjustment is made for cases where parameters lie on the edge of the parameter space or alternatives are restricted to a subset of possible values (e.g. Andrews, 1998). Classical non-nested models can also be tested using the methods developed by Vuong (1989) and others.[2] However, when the number of potential models is very large, pairwise testing implies an extremely large number of tests[3], whereby the transitivity of these tests is not assured in finite samples (Findley, 1990). This means that an unambiguous ranking of models is difficult. Information criteria (IC) offer an alternative way to evaluate models. However, while IC are additive over individuals (when models are estimated at the individual level), the formal basis for using them as model weights is through their asymptotic approximation of logged marginal likelihoods. The use of IC in Bayesian Analysis of Classical Estimates has been motivated by the desire to avoid informative priors (e.g. Sala-i-Martin et al., 2004). Yet as shown in Fernandez et al. (2001) the choice of alternative g-priors leads to asymptotically different IC, which rather weakens the claim that using IC means that one is less

---

[2]See Pesaran and Weeks (2007) for an overview.

[3]There are (n-1)n/2 combinations, which for the current paper means that the number of pairwise comparisons are of the order 10^9.

dependent on priors. In contrast, full BMA provides an internally consistent and coherent approach to this problem, providing one can provide priors. In the context of PT, we believe that there is substantive theoretical and previous empirical evidence that give a basis for setting these priors, and we examine this in detail.

We are also concerned here with explicitly recognising that model selection may depend on whether one is seeking an overarching model that explains aggregate behaviour, or whether one is seeking models that allow heterogeneity in behaviour. To follow our clothing analogy, let us imagine that our goal is to select the best style of hat. If we are constrained to dressing all individuals in exactly the same clothes, our choice of hat style may be quite different to when we allow the styles and colours of other items of clothing to differ across individuals. As discussed in Andersen et al. (2008), arguments for and against models have sometimes been implicitly or explicitly based on the idea of a 'representative agent' where it is assumed that there exists a common model of behaviour across all individuals both in the preference functionals and forms and the parameters that characterise those functionals (e.g., Brandstätter et al., 2006). However, what has been insufficiently recognised is that choosing a singular specification that best represents all individuals is a different task from choosing multiple specifications that represent different groups of individuals. Different people may do different things when it comes to making decisions. For example, it is possible that some may employ a heuristic like the PH, and others adhere to PT. In this paper we recognise that optimal model specification may differ depending on whether the researcher seeks a model that performs best when applied to all individuals; or whether one is interested in explaining individuals behaviour. Importantly, there may be models that do extremely well in explaining the behaviour of a subgroup of individuals, but do very badly if applied to all individuals.

When parametric models are being estimated, there are three levels of heterogeneity that are commonly applied. Level 0 is where individuals share functional forms and have the same parameters values (i.e., the representative agents. Level 1is where individuals share functional forms but with potentially heterogenous parameter values. And Level 2 is where individuals need not share functional forms or parameter values.

Heterogeneity in parameters can be introduced, in a limited sense, by allowing the parameters to be conditioned on covariates, but more general models include those that are either a latent class model (or finite mixture of distributions) or a random parameter (or Heirachical Bayes) approach

(e.g. Nilsson et al., 2011). Heterogeneity in models can be introduced using the weighted likelihood approach outlined in Harrison and Rustrom (2009) and related approaches in Bruhin et al. (2010) and Conte et al. (2011). In contrast, a number of papers, (e.g. Hey and Orme, 1994, Birnbaum and Chavez, 1997, and Stott 2006) have estimated multiple models at the individual level. This approach is flexible in terms of model estimation, but also requires large amounts of information to be collected at the individual level. Studies that have pursued this approach typically offer a very large number of choices (e.g. 100 or more) to each person. While an individual specific approach is flexible, it is clearly less than optimal if there is an overarching framework that is able to allow heterogeneity on one hand, but allowing the pooling of information across individuals to estimate parameters that are common to all.

In this paper, we consider the model performance at all three levels which involves estimating models at the representative agent level (Level 0) and the individual level (Level 2). Inference about Level 1 specifications can be examined by using the Level 2 models, and calculating the log marginal likelihoods with the common model restrictions imposed. That is, there is no additional estimation required for Level 1 models, once all Level 2 models have been estimated.

The paper proceeds by describing general framework and specific models in Section 2. Section 3 discusses our approach to model comparison and model estimation. Our results are presented and discussed in Section 4 and Section 5 concludes.

2. Model Descriptions

The choices under risk that are evaluated within this paper were elicited using a gamble format with a discrete number of payoffs (See Stott (2006) for specific details). The prospect (g) is of the form

$$g_i = \left(\{p_{ik}\}_{k=1}^{K_i}, \{x_{ik}\}_{k=1}^{K_i} \right) \tag{1}$$

where $\{p_{ik}\}$ are the probabilities and $\{x_{ik}\}$ are the monetary payoffs where, without loss of generality, it is assumed that they have been ordered $x_{i1} \geq x_{i2}, ... \geq x_{iK_i}$. In the empirical part of the paper $K_i = 2$ for all prospects, but we shall discuss the theory more generally.

Common to all economic models used to examine this type of data is the idea that there is to some degree, compensatory behaviour (i.e. respondents make trade-offs) with regard to both the payoffs in the prospects and the probabilities of obtaining those payoffs. As such we refer to this general class of models as compensatory. In this sense PT/RDU and TAX are compensatory

models.

In contrast within the psychology literature the idea that people apply heuristics (i.e., a decision process) that do not necessarily imply such trade-offs is commonplace. A popular example of a heuristic is the PH introduced by Brandstätter et al. (2006). Although the PH has proven very popular within the literature it has been the subject of criticism as well (see Birnbaum, 2008, and Brandstätter et al., 2008).

Note we will refer to the non-TAX, non-linear compensatory models as being PT models. Also as we are dealing with models in only the gain domain, there is nothing really to distinguish PT models from RDU models, subject to the fact that payoffs are not be evaluated relative to wealth, but around a reference point of zero. However, for simplicity we shall use the term PT only.

2.1. Compensatory Specifications

The compensatory models specified in this paper are defined by four key components. We refer to these as '*aspects*' of the model and they are: i) Value v; ii) Probability weighting (P-weight) w; iii) Inner Link $\tilde{\lambda}$; and, iv) Outer Link $\bar{\lambda}$.

Each aspect may take a number of specific functional 'forms' from a defined set: $v \in V$, $w \in W$; $\tilde{\lambda} \in \tilde{\Lambda}$ and $\bar{\lambda} \in \bar{\Lambda}$.[4] Each form of each aspect has a particular parameter space except for the Inner Link which does not contain any free parameters (i.e. they cannot be defined differently by setting parameter values).

In this paper we will (as we outline below) combine six v with seven w, three $\tilde{\lambda}$ and five $\bar{\lambda}$ and. Therefore, the number of combinations is $6 \times 7 \times 3 \times 5 = 630$. However, because the constant probability $\bar{\lambda}$ is dependent only on the sign (not magnitude) of the signal from the deterministic component, models with this $\bar{\lambda}$ are invariant to the nature of the $\tilde{\lambda}$, so the actual number of models we estimate is slightly smaller, 549 once the three PH models are taken into account. These 549 models need to be estimated at the representative agent level, and for every individual in our sample for Level 2 models, thus, requiring approximately 50,000 models to be estimated in total.

2.1.1. Value Forms (v-forms)

The v aspect evaluates the preference for a monetary amount that will be given with certainty. We employ six forms commonly encountered in the literature:

---

[4]Note that in this paper we employ the term "Link" in a different manner than that used in Stott (2006) who refers to the "choice" function, which corresponds to what we call the outer link.

$$\text{POWER-I} \quad : \quad v(x) = x^{\alpha_1} : \alpha_1 > 0 \tag{2}$$

$$\text{EXPO-I} \quad : \quad v(x) = 1 - e^{-\alpha_2 x} : \alpha_2 > 0$$

$$\text{LOG} \quad : \quad v(x) = \ln(1 + \alpha_3 x) : \alpha_3 > 0$$

$$\text{QUAD} \quad : \quad v(x) = x - \frac{\alpha_4 x^2}{2} : \quad \alpha_4 > 0, \ \alpha_4 < \frac{2}{x_{\max}}$$

$$\text{POWER-II} \quad v(x) = (\alpha_5 + x)^{\alpha_6} : \ \alpha_5 > 0, \ \alpha_6 > 0$$

$$\text{EXPO-II} \quad : \quad v(x) = 1 - e^{-\alpha_7 x^{\alpha_8}} : \alpha_7 > 0, \ 0.5 < \alpha_8 < 1.5$$

For all of our v-forms, the set of parameter restrictions ensures that the value function is always monotonically increasing. This is obvious for POWER-I, EXPO-I and LOG. For the functional form QUAD, $x_{\max}$ is the largest payoff out of all the prospects, and the parameter restrictions ensure that the function is monotonically increasing in value over the range of the data. Also, in our analysis x is normalised by dividing through by $x_{\max}$ prior to estimation such that x only varies between 0 and 1.

We also note that some of the v-forms appear in the literature with different names. For example, the POWER-II is also referred to as the Hyperbolic Absolute Risk Aversion function, and the EXPO-II is the Power-Expo-Utility function. However, since our set of restrictions on these value forms are somewhat more restrictive than those applied in the literature, we use the terms above to signal that they are generalisations of the POWER-I and EXPO-I.

2.1.2. P-Weight Forms (w-forms)

The w aspect transforms the probability of obtaining the monetary amount into some other measure that lies between 0 and 1. All the w-forms operate on the cumulative probability function except the TAX model of Birnbaum and Chavez (1997) which operates on the probabilities of the ranked outcomes. Assuming the prospects have been ordered as $x_{i1} \geq x_{i2}, ... \geq x_{iK_i}$ then the probability weights are constructed directly on the probabilities (rather than the cumulative)

$$\text{TAX:} \quad w(p_i) = \frac{p_i^{\beta_3} - \frac{1}{n+1} \sum_{j=i+1}^{n} p_i^{\beta_3} + \frac{1}{n+1} \sum_{j=1}^{i-1} p_j^{\beta_3}}{\sum_j p_j^{\beta_3}} \tag{3}$$
$$: 0 < \beta_3 < 2$$

For PT w-forms yield the weighting based on the cumulative or decumulative distributions and take the form

$$w_1 = w(p_1)$$

and (4)

$$w_j = w\left(\sum_{i=1}^{j} p_i\right) - w\left(\sum_{i=1}^{j-1} p_i\right) \text{ for } j = 2, ...K_i$$

where

$$
\begin{aligned}
\text{PRELEC-I} \quad &: \quad w(p) = e^{\left(-(-\ln(p))^{\beta_1}\right)} : \ 0 < \beta_1 < 2 \\
\text{K\&T} \quad &: \quad w(p) = \frac{p^{\beta_2}}{\left(p^{\beta_2} + (1-p)^{\beta_2}\right)^{\frac{1}{\beta_2}}} : \ 0.27 < \beta_2 < 1 \\
\text{LINEAR} \quad &: \quad w(p) = p \\
\text{POWER} \quad &: \quad p^{\beta_4} : \ 0 < \beta_4 < 2 \\
\text{PRELEC-II} \quad &: \quad e^{\left(-\beta_5(-\ln(p))^{\beta_6}\right)} : \ 0 < \beta_5 < 2, \ 0 < \beta_6 < 2 \\
\text{G\&E} \quad &: \quad \frac{\beta_8 p^{\beta_7}}{\left(\beta_8 p^{\beta_7} + (1-p)^{\beta_7}\right)} : \ 0 < \beta_7 < 2, \ 0 < \beta_8 < 2
\end{aligned}
\tag{5}
$$

The LINEAR probability form is included in our analysis because of its significance in terms of corresponding to the EUT model. We wish to assess its performance relative to the other models that have proven popular in the literature.

Our K&T specification could have an extended parameter space, but its lower bound ensures that the weight is monotonically increasing in p (Ingersoll, 2008) and the upper bound imposes the inverse-S (IS) behavior restriction (with linearity at the edge of parameter space). We imposed this restriction so we can specifically investigate a w-form with the IS condition imposed. This type of transformation was supported by and Tversky and Kahneman (1992), as being the predominant form of behaviour, but has been challenged by others (e.g. Birnbaum and Chavez, 1997, Harrison et al., 2010) who provide evidence of S-shaped behaviour. A comprehensive overview of the empirical evidence supporting IS probability weighting is provided by Wakker (2010).

The IS condition is not imposed on the other w-forms (e.g., PRELEC-I, II, G&E[5] and POWER), although the former two can be either IS or S-shaped. The lower bounds for these forms are required

---

[5]We take this nomenclature from Stott (2006). G&E is an abbreviation of Goldstein and Einhorn (1987).

so that the weights are monotonically increasing in probability, and upper bounds for these forms are not particularly restrictive in the sense that they allow for a wide variety of behaviour, but are still useful in ensuring that our estimates are convergent.

2.1.3. Inner and Outer Links

The purpose of the links are to determine the probability that one prospect (g_i) will be preferred to another (g_j). We also adopt this approach, but explore some alternative specifications. The literature generally only refers to having one link (or choice) function, but we consider it useful to think of it as being a composite function composed of an Inner and Outer link. The link aspects take the signal determined by the v-form and w-form aspects to yield a probability that an individual will choose a given prospect.[6]

Given common forms of (v,w) these may be combined in various ways to enter the link in different ways. Therefore, (v, w, g_i, g_j) combine to give, $u_i = h_{v,w}(g_i)$ and $u_j = h_{v,w}(g_j)$, which we shall term the 'utilities' of the prospects (as distinct from the values or 'utilities' of the payoffs within the prospects). If one adopts a particular PROBIT or LOGIT link form, there is still a choice as to how to combine the utilities within the 'Outer Links'. Thus, the link is a composite function, $\lambda(v, w, g_i, g_j) = \bar{\lambda}\left(\tilde{\lambda}(u_i, u_j)\right)$, composed of the Inner Link $(\tilde{\lambda})$ and Outer Link $(\bar{\lambda})$.

2.1.3.1. Inner Link Forms $\left(\tilde{\lambda}\text{-forms}\right)$

The majority of studies to date have used the difference between utilities as the Inner Link (i.e. $\tilde{\lambda}(u_i, u_j) = u_i - u_j$). We investigate this approach as well as.the difference in certainty equivalents and the 'contextual utility' approach of Wilcox (2011). Wilcox introduced the contextual utility approach for a number of reasons. First, is the observation that affine transformations of the same Value form do not necessarily lead to the same utility differences. Second, utility differences are not monotonically related to the degree of risk aversion perhaps casting doubt on how well a given model will be identified. Third, if one seeks a stochastic generalisation of the idea that one individual is more risk averse than another, then utility differences do not lead to such a definition, whereas under the conditions outlined in Wilcox (2011) such a definition can be obtained, though this definition requires individuals to have the same w-forms and parameters. Wilcox (2011) provides further evidence that the contextual approach is superior to the difference in utilities, but we are not aware that there has been any study that has compared it to the difference in certainty

---

[6]In the case of PT v and w take different forms in the gain and loss domains (and more generally may be asymmetric around a given reference point). In this study we only consider the gain domain.

equivalents also.

We define the Inner Link to be a latent variable representing one of three quantities where each is calculated as:

- UTILITY: $\Delta_u = u_i - u_j$ are the differences in utility across the two prospects

- C-UTILITY: $\Delta_c = (u_i - u_j)\,\varphi_{ij}^{-1}$ where $\varphi_{ij} = v\left(x_{upper,ij}\right) - v\left(x_{lower,ij}\right)$ and $x_{upper,ij}$ and $x_{lower,ij}$ are the highest and lowest payoffs over prospects i and j

- C-EQUIV: $\Delta_e = e_i - e_j$ are the differences in certainty equivalents where $e_i = v^{-1}\left(u_i\right)$ and $e_j = v^{-1}\left(u_j\right)$

2.1.3.2. Outer Link Forms $\left(\bar{\lambda}\text{-forms}\right)$

If we take $y = 1$ as the indicator function that an individual selects the prospect with the higher UTILITY, C-UTILITY, or C-EQUIV, then the Outer Link is a function $F\left(y = 1|\psi, \Delta\right) = \Pr\left(y = 1|\psi, \Delta\right)$ which can take several forms.

$$
\begin{aligned}
\text{LOGIT} \quad &: \quad F\left(y = 1|\psi_1, \Delta\right) = \left(\frac{e^{\psi_1\Delta}}{1 + e^{\psi_1\Delta}}\right) : \psi_1 \geq 0 \\
\text{PROBIT} \quad &: \quad F\left(y = 1|\psi_2, \Delta\right) = 1 - \Theta\left(\psi_2\Delta\right) \; : \psi_2 \geq 0 \text{ where } \Theta \text{ is a standard normal cdf} \\
\text{CONSTANT} \quad &: \quad F\left(y = 1|\psi_3, \Delta\right) = \psi_3 \; if \; \Delta > 0 \text{ and } 1 - \psi_3 \text{ if } \Delta \leq 0 \text{ where } 0.5 \leq \psi_3 \leq 1 \\
\text{BETA-I} + \text{(BETA-II)} \quad &: \quad F\left(y = 1|\psi_4, \Delta\right) = 1 - Cbeta^*\left(y = 1|\psi_4, \Delta, \bar{u}, \bar{l}\right)
\end{aligned}
$$

$$\text{(where } Cbeta^* \text{ is a cumulative beta distribution)}$$

$$\tag{6}$$

The LOGIT, PROBIT and CONSTANT $\bar{\lambda}$-forms have been commonly used as stochastic links in the literature, but the BETA link has not been investigated, at least in the way that is being used within this paper. The BETA link has two forms with BETA-II being a generalisation of BETA-I. The motivation for these two Outer Links is derived from the fact that the utilities from gambles can, under one rationalisation, be viewed as bounded from above and below. But, being bounded need not matter depending on interpretation. For example, in a pure 'trembles' setting, the individual may nearly always report their non-stochastic preference, except for occasions where they lapse. However, if one views the choices as arising from a subjective distribution of utilities or certainty equivalents, then the subjective distributions of these are bounded by the upper and lower levels in the prospects.

2.2. The Priority Heuristic

The PH can be described as follows. A respondent compares the lowest payoffs between two prospects. If the difference between these is greater than $((\varphi_1 \times 100)\%)$ of the highest payoff over the two prospects, they choose the one with the highest minimum payoff. Otherwise, they compare the probabilities of the two lowest payoffs. If the higher of these probabilities is φ_2 more than the lower, they choose the prospect with the lower probability. Otherwise, they compare the highest payoffs between the two prospects and choose the one with the higher payoff. If they have not made a decision, they choose randomly. The choice of φ_1 and φ_2 in Brandstätter et al. (2006) was $\varphi_1 = 0.1$ (i.e.10%) and $\varphi_2 = 0.1$. These thresholds were set on the basis of what respondents were used to dealing with in a decimal system.

The PH is typically employed in a deterministic setting. However, it can also be used in a stochastic setting by estimating the probability p that the choice indicated by the PH is chosen by the individual ($p > 0.5$). That is, it has a constant probability link as outlined above. Once this has been assigned then the likelihood function for the individual is defined. Additionally, the thresholds used in the standard PH, φ_1 and φ_2, can be treated as estimable parameters.

As well as offering a number of criticisms of the PH, Andersen et al. (2010) argue that such an approach is '*ad hoc*' and that restrictions would need to be placed on the model to allow estimation of the likelihood. However, we found no problems in estimation providing bounds are set relatively narrow for the parameters, which nonetheless represented a considerably more flexible parameterisation than the non-stochastic version.

In this paper, we implement three versions of the PH model. The first (PH-0) is with the thresholds φ_1 and φ_2 being set exogenously at 0.1, the second (PH-I) where both are allowed to vary according to $\varphi_1 = \varphi_2 = \varphi$ where φ is estimated, and the third (PH-II) where φ_1 and φ_2 are estimated and not constrained to be equal, thus nesting both PH-0 and PH-I. φ_1 and/or φ_2 were constrained to lie within the interval (0.01 and 0.20) in the generalised models.

2.3. Model Reparameterisations and Prior Distributions

Within the Bayesian approach prior distributions need to be specified for all parameters in a model. For a model using the marginal likelihood, these priors need to be proper, and to some extent informative. In general, the prior distributions should have mass in regions in a way that reflects prior knowledge and beliefs. However, since prior knowledge and beliefs differ between people, these priors are usually set in a relatively diffuse way. With relatively diffuse

priors, the data will quickly dominate the prior, providing the data is itself informative. Here we parameterise our models by populating them with parameters ϑ with normal priors, where the parameter of interest (θ) are transformations of ϑ (See Appendix A1).

2.3.1. v-Form Parameter Priors

In setting the priors for the v-form parameters, we need to emphasise the way in which the parameter changes the curvature of the v-form. Simply imposing equal prior probability values for the parameters could lead to priors giving high weight to regions that we consider unlikely. Therefore, before considering the priors for each of the v-forms it is useful to examine the Pratt risk coefficient (henceforth pc) (See Appendix A2).

Recalling we normalise x so that it lies within the unit interval, then based on the pc, then curvature at any given point is related to the various coefficients in equation [2] in very different ways. For example, since the utility forms must be monotonically increasing functions in x, which follows from the various constraints we impose on α_i, then for the POWER-I form, pc is bounded from above (at $\frac{1}{x}$) but not from below.

The greatest curvature for any given value of α_2 is found for low values of x, but must be less than one at the largest value of x ($x = 1$). The LOG form has a pc that is bounded between 0 and $\frac{1}{x}$, thus, it shares the same upper limit as the POWER-I form, whereas the EXPO-I form pc does not vary with x. The QUAD on the other hand has a lower bound at $\frac{-1}{1+x}$ and and upper bound $\frac{1}{1-x}$. Thus, it has the greatest convexity at low levels of x, but the highest possible concavity at the upper end of x.

With regard to the parameters α_1, α_2, and α_3 each of these are bounded by zero but have no upper limit. For these parameters, we set the prior distributions with the majority of the mass over regions that we consider plausible, but are relatively diffuse so as not to dominate the influence of the data. In doing so, we set an upper and lower bound for which a specified percentage of the mass that lies above and below these values. The parameters α_2 and α_3 are positively related to concavity, while α_1 is negatively related to concavity. However, with log-normal priors, we can just as easily think in terms of the reciprocals of α_2 and α_3 since these are also log-normal. That is, in setting the distributions for α_2 and α_3 we can immediately deduce the distributions of their reciprocals or vice versa.

The POWER-I form offers a useful starting point because previous studies can be used to infer a prior distribution for α_1 without reference to the scaling of x. Generally, previous studies have

commonly found values of α_1 (See Stott (2006), Table 5) as low as 0.225 and as high as 0.89. However, these are aggregate estimates, whereas for individuals there is likely to be a greater degree of variability. At the individual level there must be scope for some individuals to display convexity. Therefore, we set our prior to have $Pr(\alpha_1 < 0.1) = .10$ and $Pr(\alpha_1 < 2) = .90$. This equates to having 75% of the prior mass in the concave region. Thus, the POWER-I form displays significant concavity, but only at relatively low values of at x (e.g. $\alpha_1 = 0.1$, $x = 0.1$ has $pc = 9$). This distribution also has relatively cumulative high density at points close to zero (representing approximate risk neutrality).[7]

The prior for EXPO-I requires concavity like the POWER-I form at lower values of x. Thus, the distribution for α_2 needed to be more diffuse than for α_1. On the other hand, no finite level of variance for α_2 can make it as concave for sufficiently small values of x. Thus, we set the prior distribution to have $Pr(\alpha_2 < 0.1) = .10$ and $Pr(\alpha_2 < 10) = .90$.

For the QUAD, the parameters must lie between the boundaries -1 and 1 with coefficients having a lower bound at $\frac{-1}{1+x}$ and and upper bound $\frac{1}{1-x}$. Thus, it has greatest convexity at low levels of x, but the highest possible concavity at the upper end of x. The prior we adopted here assigned 75% mass on the concave region with an approximately linearly decreasing mass as we move from concavity to convexity.

For the two parameter v-forms (POWER-II and EXPO-II) the same priors were adopted for α_5 and α_7 as for α_2 and α_3 respectively. Then, for α_6, we note that the POWER-II form increase in curvature decreases rapidly, particularly at lower levels of x. Therefore, for α_6, we specified a log-normal with 50% of the mass below 0.5 and and 10% of the mass above 1. For the EXPO-II form the parameter α_8 was specified a bounded prior between 0.5 and 1.5 but with the highest density at 1.

2.3.2. w-Form Parameter Priors

All parameters for the w-forms were parameterised using the bounded transformation being set to conform to the inequalities presented in Appendix A1. The mean and variance were set so that the implied priors were for the transformed normals and were approximately uniform.

2.3.3. $\bar{\lambda}$-Form Parameter Priors

---

[7]For the LOG and EXPO-I functions if the parameters to be equal ($\alpha_2 = \alpha_3$) and to achieve the same level of concavity (if $\alpha_2 < \frac{1}{x}$) α_3 needs to be higher. In effect, the prior for α_3 should be more diffuse with a higher mean unless the aim was to construct a prior supporting risk neutrality. However, we see that for values of α_3 equal to 100 has a value at r that exceeds 99% of its possible value whereas at 10 it is at least 90% of its possible value. We therefore placed 1% of the mass above 100 and 10% below 0.1. Resulting in a relatively small shift in the mass above 10, at 13% rather than the 10% for EXPO-I function.

For the $\bar{\lambda}$-form priors, the parameters ψ_i $(i = 1, 2)$ were given log-normal priors so that 0.01 percent of the prior mass was below 0.0001 and 99% of the prior mass was below 100. The prior for $\psi_{4,1}$ was also set in this manner and $\psi_{4,2}$ was specified to be approximately uniform on the interval (0,1) and for the constant probability, the prior was set so that it was uniform between 0.5 and 1.[8]

3. Bayesian Model Comparison and Model Estimation

3.1. What do we mean by the word "Model"?

In this paper, the word '*model*' refers to the quadriplet $m_r = \left(v_r, w_r, \tilde{\lambda}_r, \bar{\lambda}_r \right)$ (where $v_r \in V$, $w_r \in W$ and $\tilde{\lambda}_r \in \tilde{\Lambda}$, $\bar{\lambda}_r \in \bar{\Lambda}$) unless it is the PH (of which there are three variants) and where r is a specific model. Therefore, models are indexed by r with the set of all models being contained in the set $\mathcal{R} = (1, ..., \#\mathcal{R})$. However, the model spaces may be limited to subsets of $\mathcal{R}$ which we call R, which contain $\#R$ elements. Therefore, a model is defined when it is populated by a set of aspect forms, but where the parameters need not be set. For each model, there is a set of parameters specific to the model with different parameter supports. We shall denote the collection of all these parameters for a given model as the vector θ_r (or $\theta_{n,r}$ where applied to individual n) where each model generates a parameter support Θ_r. The probability of choosing one prospect relative to another is dependent on the pair (m_r, θ_r). The term 'model' can be ambiguous, since it can sometimes be used to refer to m_r alone, and sometimes to the pair (m_r, θ_r). Here we refer to m_r as the model, since it is useful to be able to say that two individuals have the same model even though they may differ in their parameters.

3.2. Marginal Likelihoods, Bayes Ratios and Model Probabilities

The Marginal Likelihood (ML) is a distinctly Bayesian quantity, the calculation of which provides the basis for model comparison (through Bayes Ratios) and model averaging. The calculation of the ML is common practice in Bayesian econometrics, with a considerable literature devoted to its calculation and use. However, as the purpose of this paper is not to introduce unfamiliar readers to this approach, we relegate a fuller discussion of the construction of the MLs and associated statistics to Appendix A3.

In general, MLs can be defined at Levels 0, 1 and 2, and can be constructed to compare singular models or classes of models. In our model comparisons, we calculate and employ the quantities

---

[8]Some expost sensitivity analysis was performed on these priors. For example, the two parameter probability weigthings were re-estimated by doubling and halving the prior variances. These had no substantive impact on the results herein.

$l_j(\mathcal{N}, R)$ which represents the logged Marginal Likelihood (LML) for all individuals $(\mathcal{N})$ for a given model space R at Level $j(j = 0, 1, 2)$. Within the paper the set R will usually be a model class defined by a particular aspect form (the model class is, therefore, a limited subset of $\mathcal{R}$ called R). For example, R_A could refer to all the models with the POWER-I v-form and R_B the set of models with the QUADRATIC v-form. Alternatively, R_A and R_B can be defined by the absence of these forms. A comparison of $l_j(\mathcal{N}, R_A)$ with $l_j(\mathcal{N}, R_B)$ enables a comparison between these two sets of models (with the larger being preferred) where there has been averaging over the other model aspects (w-forms and links). For example, we can make a determination about how the POWER-I v-form compares to the QUADRATIC v-form, which is not conditional on a specific w-form or link.

The Bayes Ratio supporting R_A over R_B at level j is $exp(l_j(\mathcal{N}, R_A) - l_j(\mathcal{N}, R_B))$ So, for example, a Bayes Ratio of 10 would, under uniform model priors, indicate that the model space with the higher LML was relatively 10 times as likely compared to the model with the lesser LML.

In our empirical section, we report Logged Bayes Ratios (LBR) since the raw Bayes Ratios can be very large. The reported LBRs are the difference between the LML where a given aspect form has either been solely included or excluded and the LML is where all combinations of aspect forms are allowed (the unrestricted model space $\mathcal{M}_\mathcal{R}$).

We also calculate and present the individuals model probabilities $(\pi_{n,R})$. These can be interpreted as the probability that a model class R should be applied to an individual n, and we present $\pi_{n,R}$ in the form of histograms for key model classes. These probabilities are also used to produce model averaged estimates of "quantities of interests" such as Δ, which we use to estimate an individual's probability for the w-form.

3.3. Model Estimation

Adaptive Monte Carlo Markov Chains (MCMC) (see Andrieu and Johannes 2008) were used to estimate all models. This followed from an investigation of a subset of models estimated on a subset of individuals, which initially used a random walk Metropolis Hastings algorithm (e.g. see Koop 2003). While this algorithm converged quickly for most model-individual combinations, the mixing of the sampler was slow for a small proportion of models. While the parameters of interest are non-normal, each of the parameters is expressed as a function of a parameter with a normal prior, which suggested that a multivariate normal proposal density would be an appropriate choice for an MCMC independence chain. Investigation of the output from the random walk samplers also confirmed that the posterior distributions (for the untransformed parameters) could

be approximated by normals. Therefore, our approach to estimation used an initial phase to finding proposal densities, followed by another to estimate the parameters (See.Appendix A4 for details).

3.4. Model Comparison Strategy

Our approach to model comparison takes account of the three different levels of heterogeneity ($j = 0, 1, 2$) discussed in the Introduction. The results reported in Table 1 and 2 are LBR for the model sets defined by the **sole inclusion** (in Table 1) or **exclusion** (in Table 2) of a given aspect form at Levels 0, 1 and 2. Sole inclusion means that all forms within a given aspect have been excluded from the model space other than the one listed in the label column. Exclusion means that a particular aspect form is no longer part of the model space. We have taken this approach to model comparison as the sole inclusion and exclusion comparisons address slightly different questions in relation to model comparison:

- The inclusion approach asks whether a particular aspect form can adequately replace all the forms within that aspect; and,

- The exclusion approach asks whether a particular aspect form can be replaced by the collection of other forms within the aspect.

As part of the model comparison exercise we report the LML values for the unrestricted model space ($\mathcal{M_R}$) for each heterogeneity level at the bottom of Tables 1 and 2. These estimates are $l_0(N, \mathcal{R}) = -4206.09$, $l_1(N, \mathcal{R}) = -3531.26$ and $l_2(N, \mathcal{R}) = -3633.92$, and they can be used in conjunction with the LBRs within the Tables (at the same respective levels) to obtain the "top" or "best" model LMLs for each $\mathcal{M}_R$ defined by the inclusion or exclusion of an aspect form. Notably, the models with the highest LML at Levels 0 and 1 would be the same as if we were to assemble models by choosing each of the highest performing aspect forms based on their model averaged LMLs.[9]

So for example, if we consider Level 0 in Table 1, and then we take the LBR for the best aspect forms, add the LBRs together and then take this value away from $l_0(N, \mathcal{R})$ we arrive at the estimate of the LML for the "Top Model". Thus, a positive LBR in Table 1 means that by imposing a

---

[9]We note, that while this makes complete sense, it is not a formal requirement that the two should equate. A particular aspect form could perform well when averaged across the other aspect forms, yet not actually be part of the model with the very highest LML.

particular aspect form on all individuals has resulted in an improvement of the performance of the model space relative to the unrestricted model space $\mathcal{M}_R$. In contrast, a negative LBR in Table 2 indicates that the exclusion of this model aspect reduces the explanatory power of the model space. However, one important difference in terms of model comparison is that the PH specification is a separate model and it is not combined with other aspects. Also, for the PH specifications there is no distinction between Level 1 and Level 2 for 'sole inclusion' in Table 1, although there is a distinction between Levels 1 and 2 for exclusion in Table 2.

Although not a formal requirement, at Levels 0 and 1, we would generally expect a positive LBR in Table 1 to be associated with a negative LBR in Table 2. However, for Level 2, this is not necessarily to be expected. For example, a particular aspect form may do well in explaining a subset of individuals, yet do badly when applied to everybody. In this case we might obtain a negative LBR in Table 2 and positive LBR in Table 1 associated with this aspect form. Indeed, it can be observed within Table 2, that the LBRs for a number of aspect forms are the same for Levels 0 and 1. This is because these aspects are associated with models with very low LMLs. However, since the model space continues to include the highly performing models, the reduction in the LML is ostensibly due to the relatively small penalty incurred by increasing the dimension of the model space from a large dimension to an even larger one.

4. Results

4.1.1. The Representative Agent (LEVEL 0)

We see from Tables 1 and 2, that at Level 0, the worst performing specification (given the most negative values in Table 1, and slight positive value in Table 2) is the PH, although the most general specification (PH-II) does improve model performance with estimates for the thresholds of $\{E(\varphi_1), stdv(\varphi_1)\} = \{0.0392, 0.00496\}$ and $\{E(\varphi_2), stdv(\varphi_2)\} = \{0.1470.0139\}$. The TAX w-form is the worst performing model within the compensatory class. Interestingly, the second best w-form is the POWER w-form, even though ultimately it is not supported in terms of inclusion or exclusion in Tables 1 and 2. This finding is also reflected by the plot of the w-form for the top performing Level 0 specification, the PRELEC-II, which is presented in Figure 1 For the PRELEC-II the resulting parameter estimates are $\{E(\beta_5), stdv(\beta_5)\} = \{0.629, 0.055\}$ and $\{E(\beta_6), stdv(\beta_6)\} = \{0.829, 0.026\}$).[10]

---

[10]Stott (2006) reports values of exactly 1 for both parameters of the PRELEC-II which is actually Linear, even though the PRELEC-I estimate is not linear. This seems unlikely, though is technically possible as the estimates are derived as medians of individuals, rather than using the representative agent model we are reporting here.

Interestingly, the shape of this w-form shown in Figure 1 is not really of the classic IS shaped form favoured in the literature, but more of a concave function over the entire range. Therefore, our representative agent is estimated to be risk averse in the sense of having a concave v, but counter to this, w overweights low probability large payoffs, though is rather optimistic with respect to high payoffs with high probability also.[11]

Overall the best combination of aspects incorporates a POWER-I v-form[12] and PRELEC-II w-form, a LOGIT $\bar{\lambda}$-form and C-UTILITY $\bar{\lambda}$-form (its LML is reported at the bottom of Tables 1). This conclusion is reached since the sole inclusion of each of these aspect forms is supported in Table 1, and their exclusion is unsupported in Table 2. The LBRs for each of these aspect forms in Table 1, while positive, are moderate or small. The LBRs in Table 2 are somewhat larger in absolute terms. At Level 0 there is only one positive LBR within each aspect in terms of inclusion and only one negative value for the same aspect forms in terms of exclusion.

{Approximate Position of Tables 1 and 2}

4.1.2. Parameter Heterogeneity - Model Homogeneity (LEVEL 1)

Next we consider the heterogeneous parameter specifications (Level 1) with results again reported in Tables 1 and 2. With a Level 1 specification all respondents are endowed with a common model, but allowed to have different parameters.

Dealing first with the PH model, our Table 2 results show that there is not much to be lost or gained by including the PH within the model space with LBRs close to zero at Level 1. However, as we can see in Table 1, the sole inclusion of the PH is clearly outperformed by any of the compensatory models given the very negative values relative to the other LBRs for the compensatory models. The generalisations of the PH, allowing it to have estimated thresholds, significantly improves its relative performance, nonetheless, even with these generalisations there is no basis for arguing that these PH specifications outperform the compensatory models, at least in this context.

As with Level 0, at Level 1 within each aspect there is only one form with a positive LBR in Table 1 for each of the non-PH models, and a negative LBR in Table 2. Thus, the choice of highest

---

[11] We note the observation of Wakker (2010) page 228 about the stability of probability weighting compared to utility curvature.

[12] Although not explicitly reported the estimated parameter value is $E(\alpha_1) = 0.197$ with a standard deviation of 0.013. This results indicates a strongly concave form, which is consistent with Stott (2006) who reports 0.19.

performing aspects forms is unambiguous. As in the Level 0 case, starting with the v-form aspect, we see that the POWER-I v-form is a clear winner with the highest LBR in Table 1 and a negative LBR in Table 2. It is larger than the alternative forms by a considerable margin. It is then followed by the LOG, and EXPO-II in Table 1, though neither are supported by having positive LBRs in Table 1 or negative in Table 2.

Turning to the w-form aspect, the two top performing w-forms are the more general ones (PRELEC-II and G&E) with the G&E being the best performing followed by the PRELEC- II since it is the only one with a positive LBR in Table 1 and a negative LBR in Table 2. As with Level 0, the worst performing w-form in Table 1 is the TAX model. These results are inconsistent with Stott (2006), who concluded that the PRELEC-I is the better w-form. However, Stott (2006) arrives at this conclusion by arguing that, after elimination of other poorly performing aspects, the PRELEC-I performs the best, even though in general the two parameter forms (PRELEC-II and G&E based on rankings and averages of AICs) are the top ranked forms if there is **no** elimination of poorly performing aspect forms.

With respect to the $\tilde{\lambda}$-form it is evident, as at Level 0, that the C-UTILITY $\tilde{\lambda}$-form outperforms both the UTILITY and the C-EQUIV. This result further supports the idea that contextual utility has both empirical support as well as theoretical motivation. However, we note that the difference between C-UTILITY and UTILITY is relatively small, whereas the C-EQUIV model does considerably worse than both these other forms.

With regard to the $\bar{\lambda}$-form, the BETA-II $\bar{\lambda}$-form outperforms the other $\bar{\lambda}$-forms, with the next preferred being the BETA-I. Our results suggest that the PROBIT outperforms the LOGIT specification, a finding that again does not completely accord with that of Stott (2006). The best performing links have changed as a result of moving from Level 0 to Level 1. If we were to give this a structural interpretation, it would be that the treatment of the $\bar{\lambda}$-form in terms of individuals forming a subjective distribution of outcomes which takes account of the bounded nature of that distribution is supported. However, the way that people construct that distribution differs across individuals. The poor performance of the CONSTANT $\bar{\lambda}$-form is noteworthy as the worst link to be imposed on all models.

4.1.3 Heterogeneity in Parameters and Models (LEVEL 2)

We now consider our Level 2 results. In this case, in addition to the results in Tables 1 and 2, we report a "best-worst" analysis in Table 3.

Table 3 reports for each model aspect the number of times a particular form occurs as the top model over all individuals. Particular care needs to be taken in interpreting the numbers in Table 3, and should not be used as an accurate guide to the overall performance of a given specification. In addition to Table 3, we also present Figure 3 that illustrates model probabilities by individuals by general model class: the PH; TAX; the Linear Probability w-form; and the Non-Linear PT w-forms. In Figure 2 these model probabilities have been calculated using a "uniform" 0.5 prior probability on the models within that class and a collective 0.5 prior probability on all other models of a different class, where all models within the classes are considered equally likely. This represents a change in model priors for each case, so there is no reason for model probabilities across model classes to add to one across individuals. Therefore, Figures 3 illustrates the revision of the probability distributed across individuals after observation of the data, when one starts from the position that they are equally likely to come from a particular model class and the class of all other models.

Dealing first with the PH, we have already established that the sole inclusion of the PH performs very poorly at Level 1 or 2. From the top part of Table 3, we can see that for five of these individuals the PH-II is in fact the top performing model and that these individuals have very high posterior probabilities of being PH types. These results highlight the fact while as a model of collective behaviour (as discussed in the preceding section) the PH is a poor performer, the PH-II is a good candidate model for small a number of individuals, which was being reflected in the positive LBR in Table 2. .Turning to Figure 2, we can also see that very low model probabilities are assigned to the collective PH models, with 75% or so of individuals having near zero weight assigned to the PH model class, with the remainder being given non-negligible weight. Importantly, however, this does not mean that the PH model did not perform well for some individuals.

If now consider the compensatory specifications at the Level 2, results in Tables 1 and 2 are different from Levels 0 and 1 in that there are several forms within some of the aspects that are supported. First, the POWER-I v-form enjoys the most support in terms of both inclusion and exclusion. This support is also reflected in the high number of individuals who consider POWER-I the best v-form. We also note, that the LOG v-form also has positive LBRs in both Tables 1 and

2 and clear support in Table 3. Likewise, both the PRELEC-II and G&E w-forms are supported by positive LBRs in Table 1 and negative LBRs in Table 2. However, the removal of the POWER and PRELEC-I w-forms are marginally not supported given their negative values. For the Inner Links both the UTILITY AND C-UTILITY $\tilde{\lambda}$-forms are both supported as being components of the best performing model space, as both have positive values in Table 1and negative in Table 2. The results in Table 1 with regard to the Outer Link also suggest that one $\bar{\lambda}$-form can adequately substituted for the others, though the BETA links do best. In Table 2 we observe that the removal of either BETA $\bar{\lambda}$-form reduces the performance of the model space, and of note is the fact that unlike Levels 0 and 1, the removal of the CONSTANT $\bar{\lambda}$-form is not supported in Table 2. What this suggests is that while the CONSTANT $\bar{\lambda}$-form is a very poor form to ascribe to everybody (given its large negative value in Table 1), it does very well at describing some individuals (given its large negative value in Table 2). Also the last panel of Figure 2 contains the collective model probabilities for the non-linear (or rather potentially non-linear) variants of PT. As can be seen this class of model has considerably more support than the others, but notably, very little support for a few individuals and one individual in particular.

Turning to the TAX model even though its removal was supported within Table 2, an examination of Table 3 indicates that the TAX model is also the top model for 5 individuals. This finding is also observed in Figure 2 with around 50% of individuals with very small posterior probabilities, with only 15 individuals having prior mass above 50%. Thus, the TAX model remains a good candidate model for a small number of individuals, though as a characterisation of behaviour for all or most individuals it is poor as noted discussed in the previous section.

Finally, if we consider the LINEAR w-form we observe little support in Tables 1 and 2 but interestingly this is the top model for 20 individuals as reported in Table 3. Notably, however, it is also in the worst for specification for 68 people. Similarly, in Figure 2 there are a considerable number of individuals that have relatively large prior probabilities of being LINEAR for the w-form. However, a minority of individuals (36) have more than 50% posterior probability of being LINEAR. So again, there is evidence that for a minority of individuals that the LINEAR w-form remains a good candidate model, but this is certainly not true for the majority.

4.2. Overall Model Comparison

The first finding to note is that there has been a large fall in the LML values for all aspect forms as a result of imposing the representative agent restriction (Level 0) relative to either Levels 1 or 2

(that is comparing $l_0(N, \mathcal{R})$, $l_1(N, \mathcal{R})$ and $l_2(N, \mathcal{R})$). If we were to treat the representative agent model as a hypothesis, we would reject this restriction in complete confidence, in favour of agents having different parameters, even if they have the same models (combination of aspect forms) imposed upon them. Both $l_1(N, \mathcal{R})$ and $l_2(N, \mathcal{R})$ exceed -4199.74, which is the top performing Level 0 model. Thus, while the notion of a representative agent may be an attractive assumption, such a construction disguises the true heterogeneous nature of risk attitudes across individuals, at least for this data set.

Another important observation is that, as $l_1(N, \mathcal{R})$ exceeds $l_2(N, \mathcal{R})$, this in a sense supports the common model restriction ($\mathcal{R}$) as specified by the set of aspect forms above. However, by narrowing the model space to a subset of well performing aspect forms one can achieve LML values that exceed the top performing Level 1 model. In order to explore this further, we conducted a search[13] over model spaces. Our results indicate that the top model space (at Level 2) contained the POWER-I, + (PRELEC-II and G&E) + (CONSTANT and BETA I and BETA-II) + (C-UTILITY) aspect forms, with a Level 2 LML equal to -3518.49. This exceeds the top Level 1 model LML (-3524.84). The subtraction or replacement of any aspect form (or PH model) reduces the LML at Level 2. Therefore, although at Level 2 in the full model space there is evidence that removal of the POWER w-form, and the UTILITY Inner Link reduced performance, when seeking an optimal combination of forms they played no part. Thus, this result further supports the C-UTILITY $\tilde{\lambda}$-form as the optimal choice even at Level 2.

Overall, these results support the contention that no single w-form or $\bar{\lambda}$-form was sufficiently flexible to adequately model all individual behaviour. As discussed above, overall non-linear w-forms, other than the TAX model, do better at explaining the majority of individual behaviour. However, the nature of the non-linearity has not been broadly discussed in the literature. This issue is explored in Figure 3.

{Approximate Position of Figure 3}

Figure 3 gives the estimates by individual for the five non-linear PT w-forms, along with the model averaged estimates over all of the six in the top left hand corner (averaging uses [20]). Each plot has curves for all individuals though it may appear as shading. As can be seen the model

---

[13]Our search was not over the entire model space. We started by including all aspect forms for which elimination was not supported in Table 2. The search was then over all model spaces in which there was an elimination of one or more of these aspect forms.

averaged version take some of the attributes of each of the components. What is also clear from both the averaged and two parameter w-forms is that there are individuals which appear to be mainly concave, mainly convex, IS or S-shaped.

Individuals were also grouped into whether they had an S or IS shape or were (almost) purely concave or convex. What we can see is that there is a mix of individuals. Five individuals are concave (so purely "optimistic"), but with a larger number of individuals being convex (so purely "pessimistic), but with the remainder being fairy evenly split between being S or IS shaped. This paints quite a different picture from that of the representative agent model (Level 0). Therefore, the combined choices of individuals are best modelled by a primarily concave w-form, but an examination of all individual level results in no way supports the contention that most people behave in this way.

Finally, it is worth noting that we have found this degree of heterogeneity in spite of the limited prospect employed to generate the data examined. This of course does not imply that a different prospect would yield similar results, but it does suggest that more attention needs to be given to potential heterogeneity present in such data.

5. Conclusions

This paper has reexamined models of choice under risk using a Bayesian approach to estimation and model selection. We compared a large range of model specifications including PT models, the TAX model of Birnbaum and Chavez (1997) and a generalisation of the PH of Brandstätter et al. (2006) for which the thresholds were estimated. In addition, all models have been examined at different levels of heterogeneity so that model performance can be assessed in relation to aggregate as well as individual behaviour.

In terms of the v-form aspect (value functions), our results are in general accordance with the findings of Stott (2006). The one parameter POWER-I was far superior to the other forms considered, whether it was applied at the representative agent level (Level 0) or at the individual level (Level 2). In addition, our results support the use of non-linear w-form as suggested by PT, but this conclusion comes with some caveats. Whereas Stott (2006) preferred the one parameter PRELEC-I specification, we found that the two parameter w-forms were superior, and our findings were different depending on the level of heterogeneity that was permitted. For the representative agent (Level 0) the two parameter PRELEC-II was preferred, whereas with heterogeneity in parameter values (Level 1) the G&E specification was generally preferred to the PRELEC-II. However, where

heterogeneity in both forms and parameters was permitted (Level 2), neither of the generalised forms alone seemed sufficient to explain the behaviour of all individuals. This was also reflected in the fact that individuals seemed to have a great degree of heterogeneity with respect to w-form (i.e., probability weightings).

Overall, at the representative agent level (Level 0), there appeared to the familiar overweighting of small probability high payoffs, but of a more concave form than the IS form than commonly assumed within the literature. While all or nearly all individuals appeared to have concave v-form (value functions), the individual w-form (probability weightings) were commonly of IS, S, concave or convex functions, consistent with the observation of Wakker (2010, p.228) that "In general, probability weighting is a less stable component than outcome utility". In behavioural terms what this means is that there are individuals that behave in a pure pessimistic way, pure optimistic way, as well as having the kind of reversal in probability weightings dictated by the S or IS forms. This also means that researchers should be careful in the implementation of the IS approach, as recommended by Kahnmann and Tversky (1992). Researchers should not automatically jump to the conclusion that a form that ostensibly facilitates IS behaviour should be imposed on all individuals.

Across the different levels of heterogeneity, the contextual utility approach introduced by Wilcox (2011) was found to have the most support relative to the utility difference or the certainty equivalent difference approaches. The certainty equivalent approach was significantly inferior to the other two. While the contextual utility approach was supported empirically, some of the theoretical motivation for the contextual utility approach is weakened by the fact that individuals have a wide range of probability weightings meaning that the categorisation of somebody being more or less stochastically risk averse relative to others will prove impossible.

More generally, the results herein also remind us that for all the classes of models investigated here, no one model could adequately predict everybody, and the collective set of models failed to predict the behaviour of all individuals. It is, of course, possible that such a framework that can explain all behaviour simply does not exist and individuals employ different strategies when making choices under risk. Indeed, we found little support for either the TAX or PH model being applied to all individuals, though these models outperformed others for a small number of individuals. Furthermore, our generalisations of the PH approach improved its performance, but not sufficiently for it to outperform compensatory approaches.

This paper also introduced a BETA Outer Link which was found to outperform those commonly employed in the literature such as the LOGIT, PROBIT or CONSTANT probability link when applied at the individual level, though the LOGIT was preferred at the representative agent level. Notably, the CONSTANT probability Outer Link performed poorly relative to a others if applied to everybody, but seemed a good descriptor for some individuals. Therefore, in line with our expectations, deterministic compensatory models were more likely to predict choices if there were large differences in utility (contextual utility or certainty equivalents) rather than small ones. However, some individuals seemed to be performing in a way that was more consistent with the 'trembles' characterisation.

Looking to the future, we would contend that there is room for further empirical studies aimed specifically at examining the nature of risk functionals in the loss and mixed domains, taking a further look at PT propositions such as convexity of the Value function within the loss domain and loss aversion. These propositions could be usefully examined under a wide range of specifications using the model averaging approaches employed in this paper, or perhaps employing a reversible jump approach (e.g. Green, 1995) so that computational burdens of computing thousands of models can be reduced. Some may take the view that since there are now a number of papers which estimate preference parameters, this literature is already exhibiting decreasing returns. We take a different view. On such a fundamental issue there is a significant need for further work to be done.

Indeed, there are a significant range of estimates in the literature for key preference parameters that suggest that perhaps behavioural parameters such as those governing probability weightings may be heavily dependent on the experimental design, or more generally the context in which decisions are made. If, for example, further studies find quite different probability weighting patterns we would question whether the conditions and environment within which the experiment are having a significant role in peoples attitudes towards risk and use of probabilities, which PT and RDU theories to do not permit.

Finally, we believe that there is benefits in taking on board some of the "process based" approaches used in psychology (e.g. Fiedler and Glockner, 2012) to give a further insight into the behaviour of individuals, but combining them with econometric analyses of the sort conducted here.

References

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. (2008). Eliciting Risk and Time

Preferences. **Econometrica**, 76(3): 583–618.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. (2010). Behavioral Econometrics for Psychologists, **Journal of Economic Psychology,** 31: 553–576.

Andrews D.K. (1998). Hypothesis Testing with a restricted parameter space. **Journal of Econometrics**, 84:155-199.

Andrieu C. and Thoms J. (2008). A tutorial on adaptive MCMC. **Statistical Computation**,18: 343–373

Barberis, N.C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment, **Journal of Economic Perspectives**, 27(1): 173-196.

Birnbaum, M. H. (2006). Evidence against prospect theories in gambles with positive, negative, and mixed consequences. **Journal of Economic Psychology**, 27: 737–761.

Birnbaum, M. H. (2008). Evaluation of the priority heuristic as a descriptive model of risky decision making: Comment on Brandstätter, Gigerenzer, Hertwig (2006). **Psychological Review**, 115(1): 253–260.

Birnbaum, M. H. and Chavez, A. (1997). Tests of theories of decision-making: Violations of branch independence and distribution independence. **Organizational Behavior and Human Decision Processes**, 71(2): 161–194.

Booij, A.S., van Praag, B.M.S. and van de Kuilen, G. (2010). A Parametric Analysis of Prospect Theory's Functionals for the General Population, **Theory and Decision**, 68: 115-148.

Brandstätter, E., Gigerenzer, G., and Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. **Psychological Review**, 113(2): 409–432.

Brandstätter, E., Gigerenzer, G. and Hertwig, R. (2008). Risky Choice with Heuristics Reply to Birnbaum (2008) Johnson Schulte-Mecklenbeck and Willemsen (2008) and Riger and Wang (2008). **Psychological Review**, 115(1): 281-290.

Bruhin, A., Fehr-Duda, H. and Epper, T. (2010). Risk and Rationality: Uncovering Heterogeneity in Probability Distortion, **Econometrica**, 78(4): 1375-1412.

Camerer, C.F. and Ho, T-H. (1994). Violations of the betweenness axiom and nonlinearity in probability. **Journal of Risk and Uncertainty**, 8: 167–196.

Carbone, E. and Hey, J.D. (2000). Which Error Story is Best, **Journal of Risk and Uncertainty** 20: 161–176.

Conte, A., Hey, J.D. and Moffat, J. (2011). Mixture models of choice under risk **Journal of**

Econometrics 162: 79–88

Cox, J.C. and Sadiraj, V. (2006). Small- and Large-Stakes Risk Aversion: Implications of Concavity Calibration for Decision Theory, **Games and Economic Behaviour**, 56: 45-60.

Cox, J.C., Sadiraj, V., Vogt, B. and Dasgupta, U. (2012). Is There a Plausible Theory for Decision Under Risk? A Dual Calibration Critique, **Economic Theory**, (Forthcoming).

Fiedler, S. and Glöckner, A. (2012) The dynamics of decision making in risky choice: an eye-tracking analysis. **Frontiers in Psychology**, 3: Article 335.

Fehr-Duda, H. and Epper, T. (2012). Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences, **Annual Review of Economics**, 4(1): 567-593.

Fernandez, C., Ley, E. and Steel, M.F. (2001). Benchmark priors for Bayesian model averaging. **Journal of Econometrics** 100: 381-427.

Findley D.F. (1990) Making difficult model comparisons, Bureau of the Census Statistical Research Division Report Series, SRD Research Report Number: CENSUS/SRD/RR-90/11.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. **Journal of the Royal Statistical Society (Series B)** 56: 501–504.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, **Biometrika**, 82(4): 711-32.

Goldstein, W. M. and Einhom, H. J. (1987). Expression theory and the preference reversal phenomena. **Psychological Review**, 94(2): 236-254.

Harless, D.W. and Camerer, C.F. (1994). The predictive utility of generalized expected utility theories. **Econometrica**, 62(6): 1251–1289.

Harrison, G.W., Humphrey, S.J. and Verschoor, A. (2010). Choice Under Uncertainty: Evidence from Ethiopia, India and Uganda, **Economic Journal**, 120: 80-104.

Harrison, G.W. and Rutström, E. (2009). Expected utility and prospect theory: One wedding and a decent funeral. **Experimental Economics**, 12(2): 133–158.

Hey, J. D., and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. **Econometrica**, 62(6): 1291–1326.

Ingersoll, J. (2008), Non-Monotonicity of the Tversky-Kahneman Probability-Weighting Function: A Cautionary Note, **European Financial Management**, 14(3): 385–390

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

Koop G. (2003). **Bayesian Econometrics**. Wiley: Chichester

Nilsson H., Rieskampa J. and Wagenmakers, E.J.(2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. **Journal of Mathematical Psychology** 55: 84–93.

Pesaran, M. H. and Weeks, M. (2007) Nonnested Hypothesis Testing: An Overview, in **A Companion to Theoretical Econometrics** (ed B. H. Baltagi), Blackwell Publishing Ltd, Malden, MA, USA.

Prelec, D. (1998). The probability weighting function. **Econometrica**, 66: 497–527.

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

Rabin, M. (2000). Risk Aversion and Expected-Utility Theory: A calibration Theorem, **Econometrica**, 68: 1281-1292.

Rabin, M. and Thaler, R. H. (2001). Anomalies Risk Aversion **Journal of Economic Perspectives**, 15(1): 219-232.

Sala-i-Martin, X., Doppelhofer, G. and Miller, R. (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach, **American Economic Review**, 94: 813–835.

Shleifer, A. (2012). Psychologists at the Gate: A Review of Daniel Kahneman's *Thinking, Fast and Slow*, **Journal of Economic Literature**, 50(4): 1-12.

Stott , H.P. (2006). Cumulative prospect theory's functional menagerie. **Journal of Risk and Uncertainty** 32: 101–130

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

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. **Econometrica**, 57(2): 307–333.

Wakker, P.P. (2010). **Prospect Theory: For Risk and Ambiguity**, Cambridge University Press.

Wilcox, N.T. (2011) Stochastically more risk averse:' A contextual theory of stochastic discrete choice under risk, **Journal of Econometrics,** 162: 89–104.

Table 1: Log Bayes Ratios by Sole Inclusion of Aspect Forms

(Elimination of alternative aspect forms other than that listed)

Pr Heuristic	Level 0	Level 1+2
PH-0	-976.49	-1559.71
PH-I	-538.14	-882.41
PH-II	-502.89	-856.51

Value (v-form)	Level 0	Level 1	Level 2	P-weight (w-form)	Level 0	Level 1	Level 2
POWER-I	1.71	1.80	66.04	TAX	-156.27	-399.31	-289.05
EXPO-I	-106.33	-142.90	-81.54	LINEAR	-11.33	-199.05	-140.65
LOG	-0.65	-82.60	5.31	PRELEC-I	-7.94	-30.71	-22.34
QUAD	-375.02	-445.26	-500.24	K&T	-8.34	-162.74	-110.31
POWER-II	-22.33	-395.41	-390.41	POWER	-5.29	-130.99	-54.15
EXPO-II	-91.07	-128.07	-61.60	PRELEC-II	1.95	-5.75	13.61
				G&E	-9.31	1.95	9.04

Outer Link ($\bar{\lambda}$-form)	Level 0	Level 1	Level 2	Inner Link ($\tilde{\lambda}$-form)	Level 0	Level 1	Level 2
LOGIT	1.60	-74.30	-90.65	UTILITY	-5.85	-2.93	10.88
PROBIT	-8.73	-23.38	-47.49	C-UTILITY	1.10	1.09	13.57
CONSTANT	-334.48	-208.73	-144.15	E-EQUIV	-77.76	-209.44	-145.74
BETA-I	-10.22	-1.73	-2.81				
BETA-II	-2.78	1.58	-7.86				

Notes: Values are differences of the form $l_j(N, R) - l_j(N, \mathcal{R})$ where j denotes level

$l_0(N, \mathcal{R}) = -4206.09$, $l_1(N, \mathcal{R}) = -3531.26$, $l_2(N, \mathcal{R}) = -3633.92$.

Top Model (Level 0) POWER-I, PRELEC-II, LOGIT,C-UTILITY: -4199.75

Top Model (Level 1): POWER-I,G&E, BETA-II, C-UTILITY: -3524.85

Top Model Space (Level 2) POWER-I, G&E+PRELEC-II,

BETA-I+BETA-II+CONST, C-UTILITY: -3518.49

A Positive LBR supports the inclusion of an aspect form

Table 2: Log Bayes Ratios Under Exclusion of Aspect Forms

(Elimination of the listed aspect form)

PH	Level 0	Level 1	Level 2
PH-0	0.0015	0.0027	0.14
PH-I	0.0015	0.0027	-0.04
PH-II	0.0015	0.0027	-0.23

Value (v-form)	Level 0	Level 1	Level 2	P-weight (w-form)	Level 0	Level 1	Level2
POWER-I	-2.27	-84.21	-49.71	TAX	0.15	0.16	5.96
EXPO-I	0.18	0.18	5.62	LINEAR	0.15	0.16	.96
LOG	0.09	0.18	-11.50	PRELEC-I	0.15	0.16	-.98
QUAD	0.18	0.18	13.00	K&T	0.15	0.16	2.56
POWER-II	0.18	0.18	12.05	POWER	0.15	0.16	-2.71
EXPO-II	0.18	0.18	4.33	PRELEC-II	-6.96	0.16	-7.19
				G&E	0.15	-7.55	-6.79

Outer Link ($\bar{\lambda}$-form)	Level 0	Level 1	Level 2	Inner Link ($\tilde{\lambda}$-form)	Level 0	Level 1	Level 2
LOGIT	-4.17	0.22	10.48	UTILITY	0.40	0.39	-13.22
PROBIT	0.22	0.22	2.22	C-UTILITY	-6.55	-3.62	-11.88
CONSTANT	0.22	0.22	-22.10	E-EQUIV	0.41	0.41	14.26
BETA-I	0.22	0.19	-10.19				
BETA-II	0.21	-3.12	-5.62				

Values are differences of the form $l_j(N, R) - l_j(N, \mathcal{R})$ where j denotes level

$l_0(N, \mathcal{R}) = -4206.09$, $l_1(N, \mathcal{R}) = -3531.26$, $l_2(N, \mathcal{R}) = -3633.92$.

A positive LBR supports the exclusion of an aspect form

Table 3: Occurrences of Best-Worst Performing

Forms over Individual at Level 2

PH	Best	Worst
PH-0	0	12
PH-I	0	0
PH-II	5	0

Value (v-form)	Best	Worst	P-weight (w-form)	Best	Worst
POWER-I	44	0	TAX	5	0
EXPO-I	1	1	LINEAR	20	68
LOG	31	1	PRELEC-I	8	3
QUAD	5	76	K&T	10	7
POWER-II	3	0	POWER	17	0
EXPO-II	1	0	PRELEC-II	15	0
			G&E	10	0

Outer Link ($\bar{\lambda}$-form)	Best	Worst	Inner Link ($\tilde{\lambda}$-form)	Best	Worst
LOGIT	2	0	UTILITY	42	7
PROBIT	21	0	C-UTILITY	23	8
CONSTANT	19	21	E-EQUIV	20	63
BETA-I	29	1			
BETA-II	14	56			

Note: There are 90 individuals in the sample. 85 out 90 are best described by a compensatory model. Only 5 are best described by the PH.

Representive Agent Probability Weighting Function
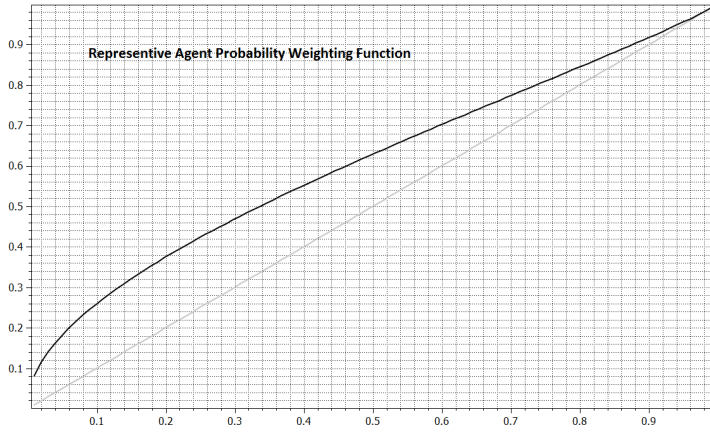
Figure 1: Representative Agent w-form
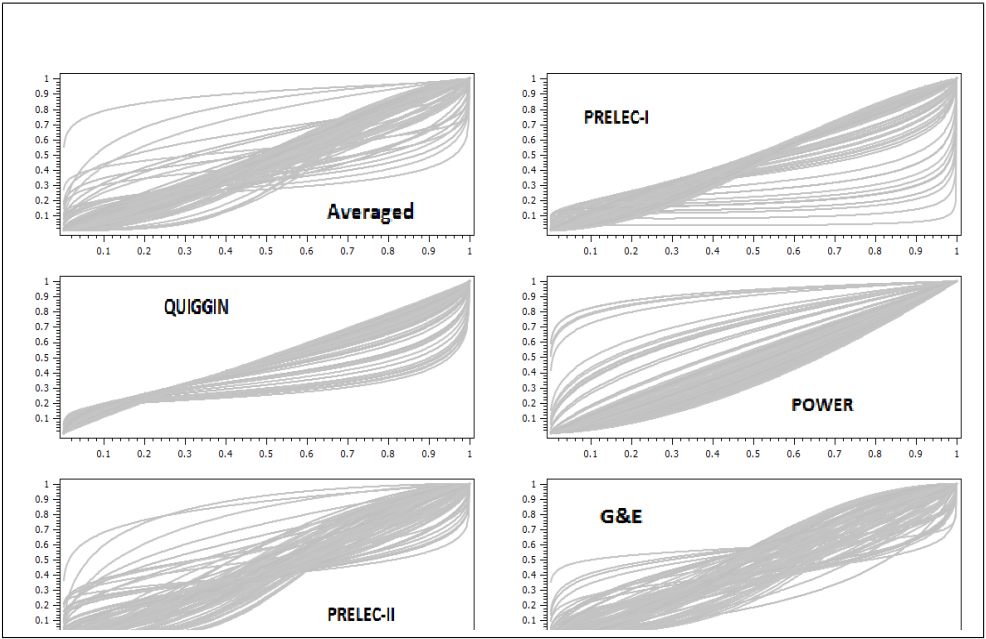
Function

Figure 2: Model Probabilities by Main Model Types

Figure 3: Probability Weighting Function Plots by Individual

Appendix A1: Transformations

The parameters of interest in the models take θ in only one of two forms. That is, we parameterise our model by using $\theta = t_1(\vartheta; \delta_l, \delta_u) = \delta_l + (\delta_u - \delta_l)\frac{e^\vartheta}{1+e^\vartheta}$ or $\theta = t_2(\vartheta) = \exp(\vartheta)$ where $\vartheta \in R$. In the case of $t_1(\vartheta; \delta_l, \delta_u)$ the transformed parameter lies within the interval (δ_l, δ_u). We set the values for $\{\delta_i\}$ *a priori* in accordance with the inequality constraints. The priors for parameters of the form $t_1(\vartheta; \delta_l, \delta_u)$ are $(\vartheta \sim N(0, \zeta))$ where they are assigned a variance ζ equal to $\frac{9}{4}$, yielding an approximately uniform prior within the specified interval, although there is less mass at the very extremes. Thus, in a sense we are being 'non-informative' about the values except that we have specified the interval over which the parameters lie. For parameters of the form $t_2(\vartheta)$ we assume that ϑ is normally distributed so that the implied prior distribution for the transformed parameter is log-normal.

Appendix A2: Pratt Coefficients

For the v-forms in the text are as follows:

$$\text{POWER-I} \quad : \quad pc = \frac{(1-\alpha_1)}{x} : \alpha_1 > 0$$

$$\text{EXPO-I} \quad : \quad pc = \alpha_2 : \alpha_2 > 0$$

$$\text{LOG} \quad : \quad pc = \frac{\alpha_3}{(1+\alpha_3 x)} : \alpha_3 > 0$$

$$\text{QUAD} \quad : \quad pc = \frac{\alpha_4}{1-\alpha_4 x} : \quad \alpha_4 > 0, \ \alpha_4 < \frac{2}{x_{\max}}$$

$$\text{POWER-II} \quad : \quad pc = \frac{(1-\alpha_5)}{\alpha_6 + x} : \ \alpha_5 > 0, \ \alpha_6 > 0$$

$$\text{EXPO-II} \quad : \quad pc = \left((\alpha_7 - 1)x^{-1} + \alpha_7 \alpha_8 x^{\alpha_8 - 1}\right) : \alpha_7 > 0, \ 0.5 < \alpha_8 < 1.5$$

Appendix A3: Marginal Likelihoods and Model Probabilities

Individuals are indexed by n and the collection of all individuals as the set $\mathcal{N} = (1, ...N)$. Models are indexed by r with the set of all models being contained in the set $\mathcal{R} = (1, ..., \#\mathcal{R})$. However, the model spaces may be limited to subsets of $\mathcal{R}$ which we call R, which contain $\#R$ elements. For the representative agent model (Level 0), and collective choices of all individuals as $Y = (y_1, ...y_N)$ where individual n makes choices y_n, the marginal likelihood for model m_r is the

prior predictive density of the observed data:

$$f\left(Y|m_r\right) = \int_{\theta_r \in \Theta_r} f\left(Y|m_r, \theta_r\right) f\left(\theta_r|m_r\right) d\theta_r \tag{7}$$

where $f\left(\theta_r|m_r\right)$ prior distributions of the parameters θ_r and $f\left(Y|m_r, \theta_r\right)$ is the likelihood of the data. Commonly, it is the log marginal likelihood (LML) that is the reported quantity which we define as $l_0\left(\mathcal{N}, r\right) = \ln f\left(Y|m_r\right)$ and can be 'averaged' over models in the sense that if a class of models is defined as $\mathrm{M}_R = \{m_r\}_{r \epsilon R}$ where $R \subseteq \mathcal{R}$ ($\mathcal{R}$ being the index set of all models) then:

$$l_0\left(\mathcal{N}, R\right) = \ln f\left(Y|\mathcal{M}_R\right) = \ln\left(\sum_{r \epsilon R} e^{l(\mathcal{N},r)} f\left(m_r|\mathcal{M}_R\right)\right) \tag{8}$$

where $f\left(m_r|\mathcal{M}_R\right)$ is the prior distribution for a given model conditional on $\mathcal{M}_R$. If all models within that class are equally likely, then $f\left(m_r|\mathcal{M}_R\right) = \frac{1}{\#R}$. (with $\#R$ denoting the number of elements within R). Therefore:

$$l_0\left(\mathcal{N}, R\right) = \ln\left(\sum_{r \in R} \frac{e^{l(\mathcal{N},r)}}{\#R}\right) \tag{9}$$

At the individual level, the marginal likelihood of a model with respect to an individual is,

$$f\left(y_n|m_r\right) = \int_{\theta_{n,r} \in \Theta_r} f\left(y_n|m_r, \theta_{n,r}\right) f\left(\theta_{n,r}|m_r\right) d\theta_{n,r} \tag{10}$$

Under the assumption that the same parameter priors are assigned to every individual $f\left(\theta_{n,r}|m_r\right) = f\left(\theta_r|m_r\right)$ for all n. The LML is $l\left(n, r\right) = \ln f\left(y_n|m_r\right)$ as at Level 0, the LML at the individual level can be 'averaged' over models in the sense that if a class of models is defined as $\mathrm{M}_R = \{m_r\}_{r \epsilon R}$ where $R \subseteq \mathcal{R}$ ($\mathcal{R}$ being the index set of all models) then:

$$l\left(n, R\right) = \ln f\left(y_n|\mathcal{M}_R\right) = \ln\left(\sum_{r \epsilon R} e^{l(n,r)} f\left(m_r|\mathcal{M}_R\right)\right) \tag{11}$$

and if all models are equally likely within the class of $\mathcal{M}_R$:

$$l\left(n, R\right) = \ln\left(\sum_{r \in R} \frac{e^{l(n,r)}}{\#R}\right) \tag{12}$$

Where there are no parameter restrictions across individuals, but where individuals share models

(r), the LML is additive over individuals in the sense that for the set of individuals $\mathcal{N}$

$$l_1(\mathcal{N}, r) = \sum_{n \in \mathcal{N}} l(n, r) \tag{13}$$

and for a class of models we can define:

$$l_1(\mathcal{N}, R) = \ln\left(\sum_{r \in R} \frac{e^{l_1(\mathcal{N}, r)}}{\#R}\right) \tag{14}$$

$l_0(\mathcal{N}, r)$ and $l_0(\mathcal{N}, R)$ we refer to as Level 0 log marginal likelihoods and and $l_1(\mathcal{N}, r)$ and $l_1(\mathcal{N}, R)$ as Level 1 LML.

Finally, we define the marginal likelihoods which model over the model class where common parameters nor common model are imposed on individuals. Define an index of models for all individuals $\tau = (r_1, ...r_N)$ with a restriction that each $r_n \in R$ for all $n = 1, ...N$ (but with no requirement that $r_n = r_{n^*}$ unless $n = n^*$) and the set of all such model combinations as $\Omega(R)$. We assume that all elements of $\Omega(R)$ are equally likely *a priori*, and equally likely to apply to all individuals, then the prior probability for τ is $(\#R)^N$ and the LML is

$$l_2(\mathcal{N}, R) = \ln \sum_{\tau \in \Omega(R)} e^{\frac{\left(\sum_{n \in \mathcal{N}} l(n, r_n)\right)}{\#\Omega(R)}} \tag{15}$$

We refer to this quantity as the Level 2 LML. Note, that under the restriction that $r_1 = r_2 = ... = r_N$, $l_2(\mathcal{N}, R)$ collapses to $l_1(\mathcal{N}, R)$.

Model Probabilities

Individual model probabilities can be constructed using:

$$f(m_{n,r}|y_n) = \frac{f(y_n|m_{n,r}) f(m_r)}{f(y_n)} = \frac{e^{l(n,r)} f(m_{n,r})}{f(y_n)} \tag{16}$$

where under equal prior odds for all models applying to all individuals $f(m_{n,r}) = \frac{1}{\#\mathcal{R}}$, then using $\sum_{r \in \mathcal{R}} f(m_{n,r}|y_n) = 1$, the probability for individual n having model m_r is:

$$\pi_{n,r} = \frac{e^{l(n,r)}}{\sum_{r \in \mathcal{R}} e^{l(n,r)}} \tag{17}$$

$\pi_{\mathcal{N},r}$ is the probability that model r can be applied to all individuals, and is defined as

$$\pi_{\mathcal{N},r} = \frac{e^{l(\mathcal{N},r)}}{\sum_{r\in\mathcal{R}} e^{l(\mathcal{N},r)}} \tag{18}$$

and likewise, the probability that a class of models (R) pertain to an individual

$$\pi_{n,R} = \frac{e^{l(n,R)}}{\sum_{R\in\mathcal{R}} e^{l(n,R)}} \tag{19}$$

The LMLs form the framework of comparisons for models, and penalise those models with more parameters. The values of $l(n,r)$ or $l_0(\mathcal{N},r)$ are estimated after simulation of the posterior distribution of the models, using the method of Gelfand and Dey (1994), which is also outlined in Koop 2003 the associated estimates for , $l_1(\mathcal{N},R)$, $l_2(\mathcal{N},R)$ $\pi_{n,r}$ etc., can then be calculated using the relationships above. For set where $(\#R)^N$ is large $l_2(\mathcal{N},R)$ cannot be feasibly calculated as above. However, using a proposal density for τ based on the posterior probabilities, $l_2(\mathcal{N},R)$ can be estimated to a high degree of accuracy by simulation.

Within the paper the set R will usually be a model class defined by a particular aspect form. For example R_A could refer to all the models with the POWER-I value form and R_B the set of models with the QUADRATIC value form. Alternatively, R_A and R_B can be defined by the absence of these forms. . Thus, for example a Bayes Ratio of 10 would, under uniform model, priors indicate that the model space with higher LML was 10 times as likely compared to the lesser one. In our empirical section we report logged Bayes Ratios since the raw ones can be very large. We also calculate and present the individual model probabilities $\pi_{n,R}$ as histograms in the empirical section. These probabilities are also used to produce weighted estimates of "quantities of interests" Δ. For example, if an individual has $\Delta_{n,r}$ under model r, then a model averaged estimate of this quantity can be obtained using

$$\bar{\Delta}_{n,R} = \sum_{r\in R} \Delta_{n,r}\pi_{n,r} \tag{20}$$

While we have calculated these for a wide range of parameters, we present only the model averaged estimates of the probability weighting functions due to space constraints.

Appendix A4: The Adaptive MCMC sampler.

The following sequence was used:

Step 1: There was a initial random walk Metropolis-Hastings phase of 5,000 draw in order to

reach a starting point for phase 2

Step 2: A 'heated' random walk phase of 50,000 draws from which a multivariate normal proposal density was constructed. This proposal was a mixture of

$$f\left(x\right) = \frac{1}{2}N\left(\hat{\mu}, \hat{\Sigma}\right) + \frac{1}{2}N\left(\hat{\mu}, 3\tilde{\Sigma}\right)$$

where $\hat{\mu}$ and $\hat{\Sigma}$ were the mean and variance of the sample from the heated random walk phase, and $\tilde{\Sigma}$ was a diagonal matrix with diagonal elements $\left\{\hat{\Sigma}_{ii}\right\}$. The draws for this phase were not used as posterior samples, but only in the construction of the initial proposal densities.

Step 3: 100,000 draws were then taken using the $f\left(x\right)$ as the proposal, with the proposal being updated every 100 iterations. At the end of this phase convergence was tested, and if it passed then moved to step 4 and if it did not pass Step 5.

Step 4: Estimates and the marginal likelihood was calculated using the method of Gelfand and Dey (1994) with a multivariate normal tuning density (see Koop, 2003).

Step 5: If convergence had not been achieved then the number of draws from the sampler was doubled and convergence retested. Sample size were doubled continuously until convergence of the model was achieved.

We employed two tests. First, the convergence diagnostic outlined in Koop (2003), and second, a requirement that the serial correlation of the thinned sequence of 10,000 draws constructed from the 100,000 had a serial correlation of less than 0.5. The entire estimation process was run twice so to check if there were any substantial differences. There were no substantive differences in the two runs. Finally, where results were obtained that seemed 'unusual', these models were investigated more closely to uncover any problems. However, while there were a few of these cases, there was no evidence that they are due to estimation problems. Note that updating of the proposal density violates as in step 3 reversibility of the chain. However, as outlined in Andrieu and Johannes (2008), this updating is allowed providing it obeys the principle of vanishing adaptation.