Information customization and food choice

Article

Accepted Version

Balcombe, K., Fraser, I., Lowe, B. and Souza Monteiro, D. (2016) Information customization and food choice. American Journal of Agricultural Economics, 98 (1). pp. 54-73. ISSN 0002-9092 doi: https://doi.org/10.1093/ajae/aav033 Available at http://centaur.reading.ac.uk/40135/

To link to this article DOI: http://dx.doi.org/10.1093/ajae/aav033

Publisher: Oxford University Press

Information Customization and Food Choice.

Abstract

In this article we employ a hypothetical discrete choice experiment (DCE) to examine how much consumers are willing to pay to use technology to customize their food shopping. We conjecture that customized information provision can aid in the composition of a healthier shop. Our results reveal that consumers are prepared to pay relatively more for individual specific information as opposed to generic nutritional information that is typically provided on food labels. In arriving at these results we have examined various model specifications including those that make use of ex-post de-briefing questions on attribute non-attendance and attribute ranking information and those that consider the time taken to complete the survey. Our main results are robust to the various model specifications we examine.

Key Words: Discrete Choice Experiment, Food Labels, Information Customization.

There is an ever expanding literature that examines the use of food labels as a means to provide information for consumers so that they can make informed and healthy food choice resulting in, it is hoped, improvements in public health. This provision of information is part of a wider effort to deal with the rapid rise of public health costs associated with food related diseases (Cowburn and Stockley, 2005; Mazzocchi, Traill and Shogren, 2009; Wansink, 2015; Lowe, Fraser and Souza Monteiro, 2015a). The expectation, on the part of policy makers, is that consumers will use the information provided on labels to make more informed (i.e., healthier) choices about the food they purchase (Grunert and Wills 2007).

To date, there has been rapid development and adoption by food retailers of various forms of health and nutritional information on food packaging (both front and back) (Hodgkins et al., 2012; Van Camp, Souza Monteiro and Hooker, 2012). On a positive note existing research tells us that consumers generally understand nutritional labels (e.g., Aschemann-Witzel, et al., 2013). However, on a more pessimistic note the degree to which health and nutritional information is actually used by consumers is significantly less than might be expected (Grunert et al., 2010). In addition, it has been noted by Balcombe, Fraser and Di Falco (2010) and Lowe, Souza Monteiro and Fraser (2013) that the vast majority of studies (hypothetical or real) have analyzed consumers' understanding and use of different food label formats for a single product. This is in contrast to many grocery shopping experiences where consumers are exposed to multiple marketing stimuli, purchasing a range of products as a multi-dimensional decision problem. In these circumstances, even for highly motivated consumers who are knowledgeable and aware of the importance of making healthier food choices, the evaluation of the overall nutritional value of a grocery shop can be a daunting task.

In reality, for most shoppers most of the time, a food shopping trip, planned or unplanned (Walters and Jamil, 2003; Nordfalt, 2009), will result in a range of products being bought at any given time. In this more complex purchasing environment the ability of the consumer to keep track of the nutritional composition and quality of their food shopping can be questioned no matter how the information is presented. Indeed, there is related evidence to suggest that when consumers are confronted by relatively simple tasks, that they can struggle. For example, it has been noted that shoppers struggle to accurately assess the monetary value of a shopping trip (see van Ittersum, Pennings and Wansink, 2010). What appears to happen is that consumers rely on heuristics which can, and frequently do, yield an incorrect answer. This applies both to consumers trying to keep expenses within a budget as well as those trying to achieve a balanced and healthy diet. Also, Wansink, Just and Payne (2009) note that consumers struggle to assess calorie intake with a strong tendency to under-estimate. Therefore, we contend that assessing the overall nutrition value of a basket of goods for specific dietary requirements is a complex task. So much so that even if consumers do consider health and nutritional labels on the food items they purchase, they are likely to struggle to determine the actual aggregate nutritional value of a multiple item shop. Thus, a tool that facilitates the customization of the nutrition and/or healthiness of a shopping basket may appeal to consumers. Such tools are becoming increasingly widespread, although little is known about their impact on consumer behavior (Lowe, Fraser and Souza Monteiro, 2015b).

In this article, we examine if customization of information provision might be valued as a means to reduce cognitive complexity and to improve the health and nutritional quality of the goods being purchased. To do this we employ a discrete choice experiment (DCE) that is designed to examine consumer interest in customization of the grocery shopping experience with respect to health and nutritional information.

The opportunities for consumers to customize goods and services they purchase is rapidly growing in all areas of retail (see Coker and Nagpal, 2013). For example, within the area of food choice there is growing interest in the development and delivery of individually personalized nutrition. Personalized nutrition can in principle yield the development of a set of individual-specific dietary choices with the aim of increasing well being as well as reducing the incidence of disease (Stewart-Knox et al., 2013; Fallaize et al., 2013).

In this research we consider different means by which personalized nutrition information can be conveyed. We contend that smart technology may be able to assist consumers at the moment of food choice by providing contextually valid information that is less likely to be biased by consumers' heuristics that might be used to overcome complexity. As such this study is contributing to the literature on how to improve the delivery and use of information by consumers when grocery shopping (e.g., Salaün and Flores, 2001; Lowe, Souza Monteiro and Fraser, 2013; Lowe, Fraser and Souza Monteiro, 2015b). We argue that the use of such technologies may contribute to social welfare by facilitating the way in which consumers gather, process and use information.

When implementing a hypothetical DCE it is generally assumed that respondents use all of the attributes presented to them in reaching their choices. Growing evidence suggests this is not always the case (Hensher, Rose and Greene, 2005), and this type of behavior is known as attribute non-attendance (ANA) (see Scarpa et al., 2010, 2013; Thiene, Scarpa and Louviere, 2015; Balcombe, Fraser and McSorley, 2015). What is also apparent from the literature is that many respondents do not necessarily ignore one or more attributes fully, or all of the time (Hess and Hensher, 2010). Therefore, the dichotomous yes/no ANA debriefing question may be too coarse/simple to reveal true ANA and may only be indicative of attributes of lesser importance. Consequently, an ANA response may in fact not imply a zero value for that attribute, so setting the marginal utility of a specific attribute to zero may bias model estimates. We explore an alternative approach by asking respondents to rank the attributes used within the DCE in order of importance to them. We refer to this as the Attribute Importance Ranking (AIR) approach. By employing an AIR debriefing question, we allow survey respondents to indicate a lower value for particular attributes without implying that some have no value at all.

We examine both approaches to dealing with ANA by estimating three Mixed Logit specifications: the standard Mixed Logit; the Mixed Logit modified to include ANA data; and the Mixed Logit modified to include AIR data. The debriefing information is employed within a modified Mixed Logit that is closely related to the Generalized Multinomial Logit introduced by Fiebig et al. (2010). As part of this analysis we also include a discussion of the potential biases that might arise from the use of debriefing data.

In addition, we examine scale heterogeneity following Savage and Waldman (2008) and Keane and Wasi (2012) who suggest that there may be learning by respondents through the course of completing a survey instrument, and that this can be captured by scale heterogeneity in the Gumbel error. We also assess the quality of survey responses by employing a measure of time (i.e., how long it took them to complete the entire survey). Within the literature the time taken to complete a survey is considered a source of information about quality of responses provided. For example, Cook et al. (2012) and Snowball and Willis (2011) both suggest that online surveys give respondents "time-to-think" and as a result provide more reliable results. This idea is in part being tested when time is being constrained as part of a choice task that sets out to assess search theory (Caplin, Dean and Martin, 2011). Thus, our most general models allow for random parameters, scale heterogeneity (i.e., time) and the inclusion of ANA and AIR data. Within the literature to date, our approach to modelling is similar to that of Thiene, Scarpa and Hensher (2015), albeit they employ a finite mixture approach.

Finally, we employ Bayesian methods to estimate our models which in turn allows us to undertake model comparison using log marginal likelihoods following Balcombe, Fraser and Chalak (2009). Also the use of Bayesian methods overcomes problems of empirical identification associated with classical approaches to simulation noted by Greene and Hensher (2010).

In general, we find interesting results in relation to the value attached to specific types of information as well as model performance. First, we find that there is latent demand for the customization of information, although the specific type preferred by respondents is not of the type which aligns with current public policy approaches. In particular, respondents are prepared to pay for information that relates to a specific dietary requirement, whereas more general (non-specific) information about nutrient content is valued far less, regardless of how it is provided. The implication of this finding is that the current dominant emphasis in public policy on generic nutrition labels might well be unwarranted. Second, in terms of model performance a model specification that employs attribute ranking information outperforms all other specifications. However, the message that emerges in relation to food label information provision and preferences for customization remains consistent and robust irrespective of which econometric model specification we report.

Survey Instrument Design and Basic Data

The survey instrument examines consumer willingness to pay for information to customize a grocery shopping experience. The design of the survey instrument began with the construction of a concept statement. The development of the concept statement was based on the literature and recent technological developments such as hand held scanners that are able to read bar codes, retrieve information and display it on a screen from which consumers can read it. However, these scanners could just as well read quick response (QR) or bar matrix codes, which are becoming standard and allow for more information to be recorded. Inspired by these developments, we propose a hypothetical service which would read nutrition information recorded on a QR label placed on food packages.

In its final form the concept statement proposed a customization service which would read health and nutrition information on food labels. The service enables shoppers to keep a tally of the overall nutrition value of their grocery shop as if they were shopping in a supermarket as well as a number of other food choice related features.

The concept statement states,"*This research is about how you perceive a new service enabling instantaneous access to nutrition information on the food you buy.*"

Then to make the situation more realistic we included a shopping list comprising both raw and pre-prepared foods. This list was based on the one used in Jetter and Cassidy (2006), but was adjusted for British shoppers and the need to have products with nutritional variation. This information is presented in figure 1:

{Approximate Position of figure 1}

Next we described the attributes used in the DCE. As a result of extensive focus group work and pilot research we settled on five attributes: appearance, nutrition label format, allergy alert, diet alert and price to be paid.

Appearance: this attribute relates to how the nutritional information is presented. By summarizing (i.e., aggregating) nutritional information to consumers for their entire shopping basket, rather than for each individual product, this could potentially reduce consumer cognitive burden, identified by Grunert and Wills (2007, p. 391), as a key concern regarding the use of nutritional food labels. Therefore, the information presented can either be product by product, or for all the products purchased in an aggregate form.

Nutrition label: with this attribute we provide alternative forms of the nutrition labels for salt, sugar, saturated fat and fat. In keeping with the U.K. government policy, we offered a hybrid label that includes a color scheme (known as the traffic light), plus a reference to a guideline daily amount per nutrient and the words "high", "medium" or "low" relating to the level of content of a nutrient. As an alternative we offered the basic traffic light approach which simply color codes the nutrients as Green (low), Amber (medium) and Red (high).

Allergy alert: this attribute allows consumers who may be subject to food allergies an ability to check for potential issues. Thus, consumers with health conditions such as a nut allergy or gluten intolerance can undertake food choice effectively, reducing the possibility that they have mistakenly overlooked the nutritional content of some products that have associated health issues. This attribute has two levels: either the allergy alert is available or it is not.

Diet alert: this attribute is offered for consumers who might have health conditions and lifestyle options that mean they should follow a certain diet. For example, some people may need to follow a low sugar or gluten free diet to mitigate type 2 Diabetes or Coeliac disease respectively. A device that can quickly alert shoppers to products containing nutrients or ingredients that need to be avoided can considerably reduce search costs. Finding a set of foods that align with a certain type of diet in a supermarket can be a very time consuming activity.

Price: we assume that the customization service will incur significant transaction costs to develop, implement and maintain. Based on a pilot study (i.e., n=32), using a price sensitivity meter for estimating thresholds of consumer price acceptability, we derived a range of acceptable prices asking for the maximum and minimum price respondents would be willing to pay each time they used this service. The mode of maximum price was £5 (mean of £3.44), while the mode of the lowest price was £1 (£1.55). Based on these results we decided to use five levels of price, that is: £0.50, £1.00, £1.50, £2.50 and £5.00. Moreover, although the price could have been set based on a monthly subscription, because of the nature of the product and because of the way consumers responded to the pilot study a pay per use price was implemented.

So, in summary, we had two levels for the first four attributes and five levels for Price. For Appearance we code the option of an individual product as zero and the aggregate nutritional information as one. For the Nutrition Label we code the hybrid label as zero and the basic traffic light as one. For both Allergy and Diet we code absence as a zero and inclusion as a one. Finally, Price is coded in the levels employed in the DCE. Having finished reading the concept statement respondents were then exposed to the hypothetical DCE comprising 12 choice cards each asking for a choice from three options.

DCE Design

Based on the five attributes described and levels employed (i.e., 2^4x5) we generated a basic D-optimal efficient design (Scarpa and Rose, 2008) assuming a Multinomial Logit functional form. Employing NGENE (Version 1.1) we generated 48 choice cards, each with two choices and an opt-out "no buy" option (i.e., a status quo). The "no buy" option is not given a specific set of attribute levels as there are a number of ways in which a consumer might undertake their shopping that means they are not necessarily interested in employing the new technology. To avoid respondent fatigue, these cards where blocked into four groups of twelve cards (Louviere, Hensher and Swait, 2001). An example choice card is shown in figure 2:

{Approximate Position of figure 2}

Once all the choice cards had been completed we presented our two de-briefing questions. The order of the two de-briefing questions was randomized as was the order of the attributes. The ANA questions took the following form:

"*Which of the following attributes (if any) have you IGNORED when making your choices? (Please tick all that you IGNORED)*

- ■ *Appearance (1)*
- ■ *Diet Alert (2)*

- *Allergy Alert (3)*
- *Nutrition label format (4)*
- *Price (5)"*

The AIR question was presented as follows:

"*Please rank which of the attributes you MOST CONSIDERED when making your choices? (please click and drag the options into the correct order such that 1=most considered attribute to 5=least considered attribute)*

1______ Appearance

2______ Diet Alert

3______ Allergy Alert

4______ Nutrition label format

5______ Price"

As already explained, we requested respondents to rank the importance of the attributes (no ties allowed) as opposed to simply indicating which of them they used. This should in principle yield more precise information about the value consumers give to each attribute.

DCE Implementation and Responses

The survey was implemented in Qualtrics (http://www.qualtrics.com/) and then administered to an online panel of UK citizens by the Toluna Group Limited, an online based pollster (https://uk.toluna.com/). In total we obtained 791 completed surveys. The sample is almost an even split between males (48%) and females (52%). About one half of the respondents had a college education and two thirds had a gross monthly income between £2,500 and £5,000. The majority of respondents (91.2%) did not report having any food related health conditions. About a third of respondents in our sample expressed they might use this type of service should it become available in supermarkets.

As noted, an important aspect of the survey implementation is that it yielded a measure of the time taken to complete the survey. The average length of time taken to complete the survey was 11.4 minutes with a median of 8.0 minutes and a standard deviation of 10.9 minutes. As we employ a panel of respondents it is feasible that they are experienced at completing an online survey and as such are able to complete the survey very rapidly. But, rapid survey completion could also capture responses that are simply random and conducted in such a way so as to minimize the effort required to earn the payment obtained from participation (Olson, 2009). Regardless of the competing reasons that might help to explain survey engagement the collection of time taken to complete the survey is a useful piece of information to include within our analysis.

Model Specification and Estimation

We employ three different model specifications: no de-briefing data; ANA data; and AIR data. For each specification we allow for scale heterogeneity in the manner described by Fiebig et al. (2010). We also assume the random parameter distribution for the Price attribute in the DCE to be normal and log-normal, which are both popular choices within the literature. Taken together this means that we have four models per specification giving a total of 12 models to be estimated.

Model 1: Mixed Logit with Heteroscedastic Scale Variance

We begin by describing the Mixed Logit specification which is the base model in our analysis. Assume that individual j $(j = 1, ..., J)$ obtains utility (U) by making choice i $(i = 1, ..., I)$ from a choice set s $(s = 1, ..., S)$. We then assume that U takes the following form:

$$\dot{U}_{ijs} = \dot{x}'_{ijs}\dot{\Lambda}\dot{g}\left(\beta_j\right) + \sigma_j e_{ijs} \tag{1}$$

where $\dot{x}'_{ijs}$ is a $k \times 1$ vector of attributes employed in the DCE. We assume that e_{ijs} is an 'extreme value' (Gumbel) distributed error, that is independent of $\dot{x}'_{ijs}$, and uncorrelated across individuals or across choices. Finally, β_j is a $K \times 1$ vector describing the preferences of individual j such that

$$\beta_j = \alpha + u_j \tag{2}$$

where α is the mean and u_j is an independently and identically normally distributed vector of errors with variance covariance matrix Ω. The errors are assumed to be uncorrelated across individuals. For the standard Mixed Logit, the matrix $\dot{\Lambda}$ is defined as $\dot{\Lambda} = I_K$ where the function $\dot{g}\left(\beta_j\right) = \left(\dot{g}_1\left(\beta_{1j}\right), ..., \dot{g}_K\left(\beta_{Kj}\right)\right)$ is a dimension preserving transformation of the vector β_j. This allows us to use an exponential transformation for any given attribute coefficient, such that the marginal utility for that attribute will be log-normal. In our analysis we only apply this transformation to the Price attribute.

Scale Heterogeneity

To accommodate scale heterogeneity, the model above is generalized so that the variance of the Gumbel error $\{\sigma_j\}$ is specified as dependent on j. We specify the following functional form for the scale variance

$$\sigma_j = e^{(-\phi t_j)} \tag{3}$$

where the parameter ϕ is to be estimated. In equation (3) term t_j is the log of the time taken to complete the survey. This generalization holds for all model specifications employed in the paper.

As previously noted, we use time in our specification because of the method (i.e., online) used to distribute the DCE survey instrument allowed us to collect this information. Based on the model specification, we can in principle assess if the views of Snowball and Willis (2011) and Cook et al. (2012) are as we would expect, such that those individuals who dwelt on the survey for longer will tend to have a lower variance attached to the Gumbel error, reflecting greater certainty about their choice. Equally, if ϕ is negative then it might be that the experience of some of the panel members is such that it allows them to complete the survey more rapidly. However, a priori, we assume that ϕ will be positive which implies rapid responses potentially signal low involvement with the survey instrument.

Finally, within the literature, there has been extensive discussion of the potential confounding of heterogeneity in scale and taste (e.g., Greene and Hensher, 2010; Hess and Rose, 2012; Thiene, Scarpa and Louviere, 2015). On one level we cannot separate scale heterogeneity from heterogeneity in tastes. They are confounded since dividing all terms in the utility function by the scale standard deviation yields a model with no scale heterogeneity. But, that means heterogeneity is then embodied in the marginal utilities. This means that the difference between scale and taste heterogeneity is that scale heterogeneity leaves the marginal rates of substitution between the attributes unchanged. Of course, in practice, it may become difficult to distinguish between the types of heterogeneity. Thus, when estimating the parameters in equation [3], our findings may be dependent on the assumptions about the distribution of β_j.

Model 2: Attribute Non-Attendance Approach

As with Model 1 we allow for the transformation of the vector β_j as well as scale heterogeneity. However, unlike Model 1 we assume that an individual is either a serial attender or nonattender throughout the DCE given their response to the attribute non-attendance de-briefing question.

We use the data collected by the debriefing question in the following way. We begin by modifying the distribution of the original marginal utilities that are defined by $\{\beta_j\}$. We do this by replacing $\dot{\Lambda}_j = I$ with $\bar{\Lambda}_j = diag(\bar{\lambda}_{j1}, ...\bar{\lambda}_{jK})$ whose elements are

$$(4) \qquad \bar{\lambda}_{jk} = (1 - \delta_{jk} + \bar{\rho}\delta_{jk})$$

where δ_{jk} is an indicator variable that equals one if individual j is classified as a nonattender of attribute k. With this specification we assume that $\bar{\rho}$ is bounded within the unit interval

[0,1], such that $\bar{\rho} = 0$ implies that a nonattender ($\delta_{jk} = 1$) has zero marginal utility for an attribute that they do not attend, and $\bar{\rho} = 1$ implies no difference between the distributions of the marginal utility of the attender and nonattender. Thus, the smaller the value of $\bar{\rho}$ the larger the reduction of marginal utility towards zero. We note that this approach has similarities to that taken by Scarpa, Thiene and Hensher (2010), though here we allow for any value of $\bar{\rho}$ between 0 and 1.

We then use $\bar{\Lambda}_j$ to modify the utility function so that the specification becomes:

$$\dot{U}_{ijs} = \dot{x}'_{ijs}\bar{g}(\beta_j) + \sigma_j e_{ijs} \tag{5}$$

where

$$\bar{g}\left(\beta_j\right) = \bar{\Lambda}_j \dot{g}\left(\beta_j\right). \tag{6}$$

Model 3: Attribute Importance Ranking Approach

The approach that we employ for our AIR data is to take the approach above but instead define $\tilde{\Lambda}_j = diag(\tilde{\lambda}_{j1}, ...\tilde{\lambda}_{jK})$ which has the elements

$$\tilde{\lambda}_{jk} = (1 - \tilde{\rho}) + \tilde{\rho}\frac{(R - z_{jk})}{R - 1} \tag{7}$$

where the parameter $\tilde{\rho}$ is estimated and can take values between zero and one, R is the number of attributes in the DCE, and z_{jk} the rank score given to attribute k by individual j. Given this specification as $\tilde{\rho} \to 0$ it follows that the ranking data is not important in determining the mean and variance of the coefficients. In contrast, when $\tilde{\rho} = 1$ this means that the lowest ranked attribute will have zero marginal utility. Thus, an estimate of $\tilde{\rho}$ closer to one implies that the AIR data is providing important information in terms of model performance. How this works within (7) is such that the higher the (mean) rank of an attribute the bigger the relative estimate of $\tilde{\lambda}_{jk}$ and the lower the impact on the resulting estimate of α. We also note that for the highest ranked attribute $\tilde{\lambda}_{jk} = 1$ irrespective of the value of $\tilde{\rho}$. This condition is required so that the model is identified. We can write this in vector form as an alternative to (6) as follows: $\tilde{g}\left(\beta_j\right) = \tilde{\Lambda}_j \dot{g}\left(\beta_j\right)$.

Model Estimation

Model estimation employs Bayesian methods closely related to that in Balcombe, Fraser and McSorley (2015). Model 1, the standard Mixed Logit, is estimated in a standard manner. Both Model 2 and Model 3 require us to modify estimation of the standard Mixed Logit. The model is relatively straightforward to estimate since it can be specified in a manner similar to the standard Mixed Logit. The main difference is that the normal latent variables are multiplied by the terms that capture the impact of the ANA/AIR data. By defining:

$$x'_{ijs} = \sigma_j^{-1} \dot{x}'_{ijs} \tag{8}$$

the (rescaled) utility function can be expressed as

$$U_{ijs} = x'_{ijs} g(\beta_j) + e_{ijs} \tag{9}$$

where $g(\beta_j)$ can take the forms $\dot{g}(\beta_j)$, $\bar{g}(\beta_j)$ or $\tilde{g}(\beta_j)$ where trivially we can define $\dot{g}(\beta_j) = \dot{\Lambda}_j \dot{g}(\beta_j)$ where $\dot{\Lambda}_j = I$. The non-stochastic component of utility is defined as

$$V_{ijs} = x'_{ijs} g(\beta_j) \tag{10}$$

and the posterior densities for the parameters $\left\{\beta_j\right\}, \alpha, \Omega$, and ρ $(= \tilde{\rho}$ or $\bar{\rho})$ are obtained by observing the probability of i being chosen in the circumstance js is the standard logit probability

$$p_{ijs} = \frac{e^{V_{ijs}}}{\sum_{i} e^{V_{ijs}}}. \tag{11}$$

If the observed choices are defined by $y_{ijs} = 1$ where the *ith* option is chosen in circumstance js and $y_{ijs} = 0$ otherwise, then the likelihood function for all the observed choices (Y) is

$$f(Y|\rho, \alpha, \Omega) = \prod_{i} \prod_{j} \prod_{s} p_{ijs}^{y_{ijs}}. \tag{12}$$

Conditional on ϕ, and $\Lambda_j = \dot{\Lambda}_j$, $\Lambda_j = \bar{\Lambda}_j$, or $\Lambda_j = \tilde{\Lambda}_j$ the steps for generating latent variables $\{\beta_j\}$ along with α and Ω can be estimated using Markov Chain Monte Carlo (MCMC) steps employed for the standard Mixed Logit (e.g., Train and Sonnier, 2005). Thus, having normalised the attributes $\left(x'_{ijs} = \dot{x}'_{ijs}\Lambda_j\right)$ the conditional distributions for β_j along with α and Ω are defined in the usual way (in terms of x_{ijs}). However, since ϕ along with ρ $(= \tilde{\rho}$ or $\bar{\rho})$ are estimated, the normalised attributes need to be updated at each iteration. In this case the posterior distribution for ϕ and ρ $(= \tilde{\rho}$ or $\bar{\rho})$ are needed. The precise priors that we use are a mean of zero for α and a diagonal covariance matrix for α with a variance of 9 for each of the elements. The precision matrix has a Wishart prior $W(I, K + 4)$ where K is the dimension of the covariance matrix. The prior variance for α was set so as to be relatively uninformative for the estimates, but small enough so that the penalty for additional parameters in the model would not be overly restrictive. The posterior distributions for ρ $(= \tilde{\rho}$ or $\bar{\rho})$ conform to the following

$$f(\rho|Y, \alpha, \Omega, \phi) \propto f(Y|\rho, \alpha, \Omega, \phi) \tag{13}$$

and

$$f(\phi|Y,\alpha,\Omega,\rho) \propto f(Y|\rho,\alpha,\Omega,\phi)\, f(\phi) \tag{14}$$

where $f(\rho)$ and $f(\phi)$ are the prior distributions. In model specifications 2 and 3 we specify $f(\rho) = I(\rho \in [0,1])$, where $I(.)$ denotes an indicator function which is one where the internal condition is obeyed and zero otherwise, and ϕ is standard normal. Estimation proceeds by iterating through the sequence of conditional draws:

i) $\{\beta_j\}\,|\alpha,\Omega,\rho,\phi,Y$;

(*ii*) $\alpha|\{\beta_j\},\Omega,\phi,\rho,Y$;

(*iii*) $\Omega|\{\beta_j\},\phi,\alpha,\rho,Y$;

(*iv*) $\rho|\alpha,\Omega,\phi,\{\beta_j\},Y$; and

(*v*) $\phi|\alpha,\Omega,\{\beta_j\},\rho,Y$.

The conditional posterior distributions for the first three components (i.e., i, ii, iii) are the same as in Train and Sonnier (2005). The conditional posterior distribution for the last two are obtained from (13). Estimation proceeds by iterating through the sequence of conditional draws as is standard for the Mixed Logit. Train (2009) provides an informative description and explanation of Bayesian computation and estimation for the Mixed Logit. The main difference introduced here is that the conditional posterior distribution for ρ ($= \tilde{\rho}$ or $\bar{\rho}$) and ϕ are obtained from (13), sampled using Metropolis Hastings steps with a random walk proposal density. That is, in *iv*) and *v*) we propose the parameter θ^* (which could be $\tilde{\rho}$ or $\bar{\rho}$ or ϕ), as $\theta^* = \theta +$ random normal, the variance of which is chosen endogeneously during the burn in phase so as to have an acceptance rate of around 40%, then choose to either stick with the old value θ or accept the proposed value θ^* with probability

$$\max\left(1, \frac{f(\theta^*|Y,\text{other parameters})}{f(\theta|Y,\text{other parameters})}\right). \tag{15}$$

Model Comparison

The support for each model ($\mathcal{M}$) was evaluated by calculating the marginal likelihood ($f(Y|\mathcal{M})$) as outlined in Balcombe, Fraser and Chalak (2009). The marginal likelihood for any particular model is defined as:

$$f(Y|\mathcal{M}) = \int f(Y|\Theta, \mathcal{M}) f(\Theta|\mathcal{M}) \, d\Theta \tag{16}$$

where $\mathcal{M}$ represents the model; Y is the observed data; and, $f(\Theta|\mathcal{M})$ is the prior distribution for the parameters Θ. Each model has its own marginal likelihood for the observed data, and we calculate this for each model that we estimate.

Endogeneity Bias in Choice Experiments

The use of ANA and AIR data has generated some concern in the literature (e.g., Hess and Hensher, 2013) about introducing possible forms of bias into CE data analysis. The aim when conducting a CE is to uncover how preferences are shaped by stimuli such as the attributes of a product and its price. However, in circumstances where there are hidden or unaccounted influences governing choices, we may falsely attribute the impact of one stimulus to the effect of another. These problems are arguably most acute in non-experimental circumstances where price changes may be confounded, because consumers may be simultaneously responding to price changes and other forms of product promotion (e.g., Petrin and Train, 2010). If we cannot control for this, we observe the joint impact of promotions and prices, not the impact of price alone. The failure to recognize this confounding effect will result in "bias".

The attraction of a CE is that we can maintain control, but effects can be introduced into the CE in a way that may undermine the internal and/or external validity of the findings. Small changes in presentation/framing in a CE can sometimes play a role in determining choices (e.g., Hensher, 2006). The CE literature contains a long debate about how to keep these to a minimum. However, concerns about "endogeneity" and bias have also been expressed (Hess and Hensher, 2013):

1. It is suggested that unobserved respondent characteristics can influence choices, thus, potentially imparting a form of bias.

2. The use of 'auxiliary data' (AD) can lead to potential bias. The information used to construct estimates of respondent preferences can come from both the choices that they make when answering choice tasks as well as from AD. AD can be answers to specific questions about attitudes and beliefs of respondents, or how they engaged with the CE, and also in the form of observational data such as eyetracking.

Let us deal with each concern in turn.

Unobserved Respondent Characteristics

Estimation of models using CE data allows for individual specific parameters. Each respondent is asked to make multiple choices, such that we are able to estimate parameters that characterize individual preferences. However, rather than treat each individual as an island, latent class models or random parameter models 'borrow' information about one individual to help improve the estimates of others. In the case of the 'Hierarchical Bayes Mixed Logit' (the structure we use in the paper, and therefore, the one mainly we allude to here) assumptions are made about the distributions of preference parameters, which are assumed to be transformations of multivariate normal distributions. These can be conditioned on individuals characteristics, but many of these will remain unobserved.

Not to observe important factors determining preferences does not mean, in general, that conditional estimates are biased. For unobserved heterogeneity to be a problem it has to be either incompatible with the types of distributional assumptions that we are making, or lead to some more fundamental undermining of the CE. Nonetheless, it is possible to see how either of these might happen. For example, in the context of our CE some respondents may have severe food allergies. This could induce a form of non-compensatory behavior that is incompatible with the underlying assumptions of our model. Alternatively, a small but substantial group of people might have distinct preferences because they have food allergies.

Consequently, if we model the distribution using a single multi-model normal latent variable, this may fail to reflect the actual distribution of underlying preferences.

The Use of AD

It has been proposed that the use of AD can actually induce bias, particularly when AD is of the ANA or AIR type. Here the concern has been raised that because there are common factors driving both the AD responses and the choices made during the CE that there is an endogeneity issue (bias). However, it is not, in general, a problem that common factors drive both the systematic response of the individual across all choices and AD. In order to understand this, let us first reflect on the nature of the utility function (U_{ijs}):

$$
\begin{aligned}
(17) \qquad U_{ijs} &= V_{ijs} + e_{ijs} \\
(18) \qquad \text{where } V_{ijs} &= V\left(\beta_j, x_{ijs}\right).
\end{aligned}
$$

The V_{ijs} term represents systematic preferences of the individual and e_{ijs} is the random error (Gumbel distributed) assumed to be independent of V_{ijs}. The e_{ijs} reflect choices, but not at a deeper level 'preferences' in the sense that, presented with exactly the same attributes in a different circumstance V_{ijs} remains unchanged, whereas e_{ijs} may differ.

Let us first highlight circumstances where there would be a definite problem - the case where the AD is related to the Gumbel error e_{ijs}. Imagine, we asked somebody whether they liked the status-quo option in the CE they have just completed. Perhaps in order to answer the question, an individual physically examined the number of times they chose the status-quo, and responded with this in mind. In doing so they would have used knowledge not only of their systematic utility component, but of the random error (e_{ijs}). Subsequent use of this data to aid estimation of the preference parameters β_j would, therefore, be using a variable that is associated (i.e., correlated) with e_{ijs}. This has parallels to "endogeneity" in the standard linear regression model, since the conditioning of β_j on information dependent on e_{ijs} means that V_{ijs} and e_{ijs} are also dependent.

Furthermore, if the stimuli x_{ijs} are not constructed independently of the characteristics of the individuals (where z_j is a form of AD) and/or more directly preferences reflected in β_j, bias is likely to occur. Examples of where this may occur, are where an individual's perception of risks or attitudes are elicited and used to form or modify x_{ijs}. In such circumstances steps must be taken to try and account for this 'endogeneity' (e.g., Teisl and Roe, 2010) or tests conducted for its potential existence (e.g., Lusk, Schroeder and Tonsor, 2014).

Now imagine instead that we have a standard response to an ANA question; data is collected after all the choice sets have been completed. This type of question can be addressed by the respondent without reference to the choices they made (U_{ijs}), but simply about whether attributes were used (or not) when making choices, implying that the impact will be through β_j. In this case there is no reason to suppose an association with e_{ijs}. Therefore, conditioning of β_j on ANA or AIR data need not necessarily induce a bias in this sense. This does not mean there is no possibility of endogeneity bias. If the $\beta_j s$ are conditioned on AD z_j (but x'_{ijs} is set exogenously) "bias" can occur when our way of conditioning β_j on z_j is either flawed, or when we fail to model parameters that determine conditioning correctly.

Consider the following example. Assume, we have $\beta_{jk} = z'_j\alpha_k + v_{jk}$, with v_{jk} normally distributed, there is always a construction of α_k that makes v_{jk} independent of z'_j regardless of whether there are latent variables driving both z'_j and β_{jk}. If, however, we endow α_k with a structural interpretation that is incompatible with the independence between v_{ik} we then have "bias" with regard to α_k (but not with regard to β_{jk}). Thus, we observe that (un)biasedness is not a property that exists separately from a claim about what the parameter of interest represents. For example, some people may refuse to state their income (a type of AD). When estimating MUs conditioned on income data we need to recognize that these utilities are being conditioned on a persons' income plus their willingness to divulge this information. If we have significant non declaration of income, it is material that the class of people who earn between 50 and 100 thousand dollars, is not the same class of people that are prepared to **say** they earn this amount. The two groups may differ systematically in their responses and, therefore, confusing these two classes leads to what might be called 'bias'. Likewise, if we use ANA or AIR data we must be cognizant of the difference between the class of people that say they ignore an attribute, and those that actually did. The interpretation of any parameter that is derived must be interpreted accordingly.

Finally, bias may also result from the incorrect way in which β_j is conditioned on AD. For example, Hess and Hensher (2013) posit a structure whereby ANA and AIR data is driven by a latent variable that also drives preferences and, hence, choices. If direct attendance or ranking data is simply inserted in place of a latent variable within this framework, then there will be misspecification bias. However, the direct use of ANA or AIR data need not be interpreted as the direct replacement of a latent variable, as any model has a likelihood with the data remaining but the latent structures integrated out (numerically or in closed form). Bias will occur to the extent which the employed model differs from the true structure (assuming of course that one posits a true structure to exist). The observation that model specification matters is important, but does not imply that the conditioning of preferences on ANA or AIR data leads to bias or problems associated with endogeneity.

Empirical Specification

Given our DCE and model specification, the empirical utility structure estimate is:

$$U_{ijs} = V_{ijs} + e_{ijs} \tag{19}$$

$$\begin{aligned} V_{ijs} = {} & (ASCNB_{ijs} + \lambda_j g(\beta)\, App_{ijs} + \lambda_j g(\beta)\, NutLabel_{ijs} \\ & + \lambda_j g(\beta)\, Allegy_{ijs} + \lambda_j g(\beta)\, Diet_{ijs} + \lambda_j g(\beta)\, Cost_{ijs}) \\ & \times \exp(-\phi t_j) \end{aligned} \tag{20}$$

where the $ASCNB_{ijs}$ captures the no buy option that results from not selecting to use the technology. Importantly, as we employ a no buy option, that is, given no specific form, it then follows that our parameter estimates for $ASCNB_{ijs}$ have no specific model interpretation. λ_j is equal to one if estimating model 1, $\bar{\lambda}_{jk}$ if model 2 and $\tilde{\lambda}_{jk}$ if model 3. The final term in the specification, $\exp(-\phi t_j)$, captures the scale heterogeneity and is assumed a function of time (t). Finally, the transformation $g(.)$ can take one of two forms, $g(\beta) = \beta$ or $g(\beta) = \exp(\beta)$. In the models presented we only allow the Price attribute to be both a normal and log-normal random parameter.

Results

Given the models described above we estimate 12 different model specifications. We estimated all models using GAUSS 11.0 employing a burn in of 1,000 iterations followed by every 100th draw being kept to yield 10,000 in total from 1,000,000 iterations. To ensure that we had achieved model convergence we tested all specifications using standard diagnostics (Koop, 2003) (i.e., visually, and modified t-tests for the difference between the first and second halfs of the iterations).

In terms of the specific models estimated we have the standard Mixed Logit (Model 1), the Mixed Logit plus the ANA data (Model 2) and the Mixed Logit plus the AIR data (Model 3). For each specification, we also estimate results assuming that the Price attribute is both a normal and log-normal random parameter (denoted by N for normal and L for log-normal). We also assess scale heterogeneity and those model specifications are labeled using T.

We begin by reporting some descriptive analysis of the survey data. This is then followed by an examination of relative model performance. Next, we report model specific results for the best performing specification which in turn yield results that allow us to compare model specification impacts on WTP estimates. Finally, we investigate the effect of time taken to complete the survey instrument on model selection and WTP by re-examining a subset of the data for responses longer than seven minutes. We selected seven minutes, because this was the length of time we deemed necessary to complete and engage with the survey as intended. This estimate was based on discussions undertaken during focus group activity and examination of the average time to complete the survey during the piloting exercise. For our full sample of data, 360 out of 791 respondents completed the survey in less than seven minutes.

Descriptive Data Analysis

Some preliminary remarks on the behavior of the data may be helpful to the reader. First, the AIR and ANA data were broadly consistent in the sense that the ordering by mean attribute rankings were consistent with ordering of importance by attendance (i.e., higher mean attendance meant better mean rankings for a given attribute). Price was the highest attended and highest ranked attribute, though about 20% of individuals still indicated ANA for this attribute with an average rank of about 1.9 (one being the highest). As expected, the two measures (ANA and AIR) were strongly correlated over individuals. However, there were also "inconsistencies" in the sense that a substantive number (just over 30%) of individuals indicated ANA for attributes that had higher rankings than those they indicated attendance for. We note that Cameron and DeShazo (2010) might alternatively consider these choices as being "counterintuitive" where this outcome can be explained by the fact that in specific contexts an important attribute might not be "pivotal" in terms of the choice that is made.

A natural question that follows from the basic data is whether ANA and AIR data could be associated with time taken to complete the survey. We investigated this question by a series of regressions we report in the supplemental appendix online. Our overall conclusion, is that time taken to complete the survey was not a likely candidate for explaining ANA or AIR. However, time taken to complete the survey may still explain the level of noise in decision making and we investigate this further below.

Model Comparisons

We now examine the relative performance and results of the various models estimated. All the logged marginal likelihoods (MargLL) for the models estimated are presented in table

1.

[Approximate Position of table 1]

The results in table 1 clearly indicate that the inclusion of the de-briefing information significantly improves model performance. In terms of the ANA and AIR specifications it appears that the AIR specification assuming a log-normal distribution on Price and scale heterogeneity, is the top performing model (Model 3LT). Indeed, for all three model variants the use of the log-normal distribution and the inclusion of scale heterogeneity improve model performance significantly.

From a Bayesian perspective the MargLLs are sufficient for us to make model comparisons (Balcombe, Fraser and Chalak, 2009). To understand the extent of model improvement from the introduction of the de-briefing data the exponential of the difference between the MargLL for two models gives the 'Bayes Factor' between the models when each is considered equally plausible *a priori*. The MargLL also implicitly takes into account whether one model has more parameters than another, so no adjustment needs to be made to the MargLL in order to make model comparisons.

Model Results

We now report results for our preferred model specifications. Given the clear support for the use of the log-normal distribution for Price and the inclusion of scale heterogeneity we report the best model for each of the three specifications (i.e., Models 1LT, 2LT and 3LT). These results are all reported in table 2 (see the supplementary online appendix for the full set of model results).

[Approximate Position of table 2]

The first thing that we can see from table 2 is that the parameter estimates on all attributes for each of the specifications are consistent. That is, the main changes to the model specification do not, in general, impact the magnitude or sign of the parameter estimates.

Indeed, if we consider the posterior distributions for the coefficients, we find a strong degree of consistency across all three models in that the same coefficients (mean of α) generally have standard deviations that are less than half the level of the estimates. Classically speaking this would be taken as evidence of statistical signifcance.

Turning to the specific attributes, Appearance is negative for each specification indicating that respondents have a clear preference for information on individual products as opposed to that which is summarized at the aggregate level. The sign for Nutrition Label is positive and strongly significant, which indicates a strong preference for a simple traffic light rather than a hybrid label. The Allergy alert attribute is strongly significant with a positive sign indicating that respondents prefer to have this option available. Similarly, the Diet alert attribute is positive and significant indicating that respondents would like to know about food products that they may have to include, or avoid, as part of a diet. The Price attribute is negative as we would expect. Finally, for the three specifications reported in table 2 the ASC NB is negative which as previously noted, we cannot attribute a specific meaning to because we have employed an unspecified no buy option.

Next we consider the ANA coefficient ($\bar{\rho}$) for Model 2LT. We can see that the coefficient is statistically robust and its relative value (i.e., $\bar{\rho} \to 0$) indicates that the ANA data has had a significant impact on model performance. Essentially, the smaller the value of $\bar{\rho}$ (i.e., closer to zero) the greater the reduction in marginal utility such that it becomes closer to zero. Also, the magnitude of $\bar{\rho}$ is similar for all specifications, not just this preferred model, which suggests that the impact of ANA responses is reasonably robust to other aspects of model specification such as the choice of random parameter distribution and the inclusion or exclusion of scale heterogeneity.

We now consider in table 2 the AIR coefficient ($\tilde{\rho}$) for Model 3LT. We can see that it is statistically robust and like the ANA coefficient its magnitude indicates that the AIR data has had a significant impact on the model results. As already explained, when $\tilde{\rho} = 1$ it follows that the lowest ranked attribute will have zero marginal utility. Therefore, the fact that $\tilde{\rho} \rightarrow 1$ (0.884) and its associated standard deviation indicate that the AIR data is statistically significant. The associated improvement in model performance is reflected in the results presented in table 1 which indicates that by taking account of this information in model estimation we improve model performance.

Finally, we can consider the scale heterogeneity (ϕ) estimate for each model. The parameter estimates for all the models are positive and very similar in magnitude. The positive sign indicates increased precision in responses for those respondents who took longer as time is assumed inversely related to variance. Or to put it another way, people who took longer to complete the survey, have yielded more precise (i.e., greater choice determinacy with lower variance) responses. This finding is consistent with the view that the longer the time taken to consider and think about responses the more precise is the information revealed.

WTP Results

To be able to compare the impact of the ANA and AIR data on the model output we now consider WTP estimates. Our WTP estimates have been computed by first taking the estimates for α and Ω (i.e., the distribution of the latents), and taking multiple draws (e.g., 50,000) of these latents. Then for each draw we calculate the ratio of the marginal utilities (with Price as the denominator) after having made any distributional transformations. This gives the distributions for the WTPs. We then use these distributions to generate the median plus lower and upper and quartiles. The reason we report the median is that previous research has shown that mean WTP estimates generated in preference space as opposed to WTP space can be very unstable. Furthermore, median estimates generated in preference space closely match those derived in WTP space, as demonstrated by Balcombe, Chalak and Fraser (2009).

We begin by examining median WTP estimates for all the models estimated so that the impact of the difference in specification is clear. These results are presented in table 3.

[Approximate Position of table 3]

As can be observed in table 3 the WTP estimates yield relatively consistent evidence about the specific importance of each attribute, and the results provide several interesting insights into how respondents value customized information. First, we can see that for each model the rank order by magnitude of WTP remains the same. In each case it is the inclusion of the Allergy attribute that yields the highest WTP followed by Diet. Second, the form of the customized information (i.e., Appearance) matters very little, and the WTP for type of Nutrition Label is also of only marginal importance. Third, there is some variation in value estimates that is, as we might expect, related to choice of distributional assumption.

Next, we can consider the results, in this case median plus lower and upper quartile, for the best performing specification for each model type. These results are presented in table 4.

[Approximate Position of table 4]

The results in table 4 reveal that, statistically, the key attributes are Allergy and Diet. Clearly, the estimates for Appearance and Nutrition Label are far less significant as the range between the quartiles includes zero. What is also apparent is that the upper quartile values are significantly larger than the median confirming the observations of Balcombe, Chalak and Fraser (2009) that WTP estimates generated in preference space can be unstable. These results suggest that information that has specific relevance to the individual is more highly valued compared to generic public health information of the type provide by nutrition labels regardless of their format.

Survey Response Time

The final piece of analysis we present relates to the time taken to complete the survey. As noted above, longer survey completion time correlated positively with choice determinacy across sequences. This result raises an interesting dilemma for the use of data collected in which the speed of response cannot necessarily be controlled. Although the analysis undertaken here "controls" for the speed of survey completion by the choice of model specification an alternative approach is that very rapid responses might be dropped from the sample data. This raises an interesting question: should we use all sample data and control for behavior or should we remove "outliers" prior to estimation? In an effort to address this question, we have re-examined our data and removed the 46% of responses that took less than the previously discussed seven minutes to complete the survey instrument.

To see if the exclusion of this data impacts our results we have re-estimated all 12 models (we report marginal likelihoods as well as specific model results in the supplementary online appendix). Based on these results, we find that the best performing model is still Model 3LT although the margin by which it is preferred is reduced. What we observe with respect to the model estimates is that they are almost identical in sign and very similar in magnitude to the earlier results. However, we find that the sign on our scale heterogeneity estimate is now negative (the only sign change) and in most specifications provides little explanatory power. The negative sign suggests that respondents who took a long time to complete the survey yielded less precise (i.e., reduced choice determinacy with higher variance) responses. Lastly, the AIR coefficient ($\tilde{\rho}$) is almost identical to that of the full sample (mean = 0.882 and standard deviation = 0.031).

Turning to the WTP estimates for this specification, we find that these are almost identical (i.e., magnitude and explanatory power) to the full sample. Our new results are reported in table 5.

[**Approximate Position of table 5**]

Thus, the WTP estimates demonstrate that the reduction in sample data to account for response time, based on our prior views about quality of responses, has had minimal impact. Therefore, although it makes statistical sense to take account of response times (i.e., controlling for them) via the inclusion of time in the scale heterogeneity part of the model, at least for this data set excluding these data has not impacted WTP estimates.

Summary and Conclusions

In this paper we have investigated how the customization of information about diet and health is valued by consumers. Using a DCE we have examined the extent to which consumers are willing to pay to adopt and use some form of "technology" to provide this information. Our model results indicate that respondents appear willing to pay for the customization of information to help inform their grocery shopping. However, the nature of the information being sought is not necessarily of the form currently being provided on the front and back of food packaging by food manufacturers. This is an important result as it implies a need to better understand actual consumer requirements as opposed to assuming a one model fits all approach. By specifically identifying this behavior, we can start to understand why all too frequently consumers appear to be knowledgeable about nutritional labels but they do not respond to them in the manner required. These results fit into the wider literature on personalized nutrition which is an area of growing research interest in relation to public health and wellbeing (Stewart-Knox et al., 2013).

Clearly, an important dimension of the finding is that our results indicate that the latent demand for customization is much stronger for aspects of food consumption that are specific to the individual as opposed to undifferentiated mass communications which are currently used by retailers and policy makers to provide nutritional information. This result has interesting implications for food label design and more importantly for how to influence consumers to improve their diet. The evidence presented here suggests that simply reformulating the existing form of nutritional information is of little interest to respondents. If nutritional information is going to be heeded by food customers then there needs to be more thought given on how to provide this type of information in a manner which is consumer-specific. What is really required is for such information to be relevant at the point of purchase and/or during consumption (Lowe, Souza Monteiro and Fraser, 2013; Lowe, Fraser and Souza Monteiro, 2015b). Thus, it is likely, that if the provision of nutritional information continues to ignore specific requirements of the individual consumer, then the desired public benefits of nutritional labels will remain unachieved.

Turning to model performance for the specific DCE reported in this research, we find that the AIR data in combination with scale heterogeneity and a log-normal specification for the Price yields the best performing model. Indeed, our model comparison results suggest that there is a positive impact of employing debriefing questions on DCE model performance. More specifically, we have demonstrated that the use of an AIR question as opposed to the more conventional ANA question improves model performance for this specific DCE. The conventional ANA question is coarse in terms of the information which is revealed in relation to respondent use of attributes. In our opinion the ranking question better allows respondents to express their views in relation to how they have actually interpreted and used the attributes within the DCE. However, what our results also reveal, at least for the DCE examined here, is that the inclusion of this information does not seriously impact the resulting WTP estimates reported.

Another interesting aspect of our results relates to inclusion of time as a variable with which to model scale heterogeneity. In the literature, as noted earlier, time has been used to assess various hypotheses that might explain the quality, or lack of, in resulting DCE responses. In the results reported here, we find very little evidence that time has had a significant impact on model outcomes. We do find that the inclusion of time improves model fit but the overall impact on WTP is marginal. As such the impact of time is clear in terms of model performance but overall it has little qualitative impact.

At this point it is important to view our results in relation to potential forms of bias that might be present. As we have discussed a facet of the research and results presented is the possibility that the use of auxiliary data (AD), such as ANA and AIR, might introduce bias into model results. We have discussed this controversial topic attempting to shed light on circumstances when bias may well be a serious problem. As we note, care needs to be taken when considering this issue as the types of bias being identified are in many cases more likely to be examples of model misspecification.

More generally, our results present a challenge for the food industry and public policy makers alike on two levels. First, if the focus on nutrition information continues to prove largely ineffective, in terms of use as opposed to understanding, then an alternative public marketing policy needs to be formulated. Here we align with Andrews, Netemeyer and Burton (2009), who challenge the viability of increasing levels of nutrition literacy for the majority of the population. Consistent with Lowe, Souza Monteiro and Fraser (2015b), we suggest that policies targeting consumers at risk and helping them make choices consistent with a healthy diet hold more promise. Thus, we can see merit in the development of personalized nutrition, although we remain somewhat skeptical about the more extreme methods of implementation being discussed in the literature (e.g., nutrigenomics). Second, existing information policies may be more effective if they make use of newly developed technological platforms, such as apps or hand held scanners increasingly available in retail environments, which allow consumers to tailor shopping to their specific requirements. This degree of intervention is far less than that required by nutrigenomics and it raises far fewer ethical questions about the use of personalized information that might be deemed invasive/inappropriate for the individual consumer.

References

Andrews, C.J., R.G. Netemeyer, and S. Burton. 2009. "The Nutrition Elite: Do Only the Highest Levels of Caloric Knowledge, Obesity Knowledge, and Motivation Matter in Processing Nutrition Ad Claims and Disclosures?" *Journal of Public Policy and Marketing* 28(1):41-55.

Aschemann-Witzel, J., K.G. Grunert, H.C.M. van Trijp, S. Bialkova, M.M. Raats, C. Hodgkins, G. Wasowicz-Kirylo, and J. Koenigstorfer. 2013. "Effects of Nutrition Label Format and Product Assortment on the Healthfulness of Food Choice." *Appetite* 71:63-74.

Balcombe, K.G., I.M. Fraser, and A. Chalak. 2009. "Model Selection in the Bayesian Mixed Logit." *Journal of Environmental Economics and Management* 57(2):226-237.

Balcombe, K.G., I.M. Fraser, and S. Di Falco. 2010. "Traffic Lights and Food Choice: A Choice Experiment Examining the Relationship between Food Labels and Price." *Food Policy* 35:211-220.

Balcombe, K.G., I.M. Fraser, and E. McSorley. 2015. "Visual Attention and Attribute Attendance in Multi-Attribute Choice Experiments." *Journal of Applied Econometrics.* 30(3): 447-467.

Cameron, T.A. and J.R. DeShazo. 2010. "Differential Attention to Attributes in Utility-Theoretic Choice Models." *Journal of Choice Modelling* 3(3):73-115.

Caplin, A., M. Dean, and D. Martin. 2011. "Search and Satisficing." *American Economic Review* 101(7):2899-2922.

Coker, B. and A. Nagpal. 2013. "Building-Up Versus Paring-Down: Consumer Responses to Recommendations When Customizing." *Journal of Retailing* 89(2):190-206.

Cook, J., M. Jeuland, B. Maskery, and D. Whittington. 2012. "Giving Stated Preference Respondents "Time to Think": Results From Four Countries." *Environmental and Resource Economics* 51:473-496.

Cowburn, G. and L. Stockley. 2005. "Consumer Understanding and Use of Nutrition Labelling: A Systematic Review." *Public Health Nutrition* 8 (1):21-28.

Greene W.H. and D.A. Hensher. 2010. "Does Scale Heterogeneity Across Individuals Matter? An Empirical Assessment of Alternative Logit Models." *Transportation* 37(3): 413-428.

Fallaize, R., A.L. Macready, L.T. Butler, J.A. Ellis, and J.A. Lovegrove. 2013. "An Insight into the Public Acceptance of Nutrigenomic-Based Personalised Nutrition." *Nutrition Research Reviews* 26:39-48.

Fiebig, D.G., M.P. Keane, J. Louviere, and N. Wasi. 2010. "The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity." *Marketing Science* 29(3):393-421.

Food Standards Agency. 2007. Front of Pack Nutritional Signpost Labelling Technical Guidance, [available http://www.foodstandards.gov.uk/multimedia/pdfs/frontofpackguidance.pdf].

Grunert, K.G. and J.M. Wills. 2007. "A Review of European Research on Consumer Response to Nutrition Information on Food Labels." *Journal of Public Health* 15 (October):385-399.

Grunert, K.G., J.M. Wills, and L. Fernandez-Celemin. 2010. "Nutrition Knowledge, and Use and Understanding of Nutrition Information on Food Labels Among Consumers in the UK." *Appetite* 55:177-189.

Hensher, D.A. 2006. "How Do Respondents Process Stated Choice Experiments? Attribute Consideration Under Varying Information Load." *Journal of Applied Econometrics* 21:861-878.

Hensher, D.A., J.M. Rose, and W.H. Greene. 2005. "The Implications on Willingness to Pay of Respondents Ignoring Specific Attributes." *Transportation* 32:203-222.

Hensher, D.A., J.M. Rose, and W.H. Greene. 2012. "Inferring Attribute Non-Attendance from Stated Choice Data: Implications for Willingness to Pay Estimates and a Warning for Stated Choice Experiment Design." *Transportation* 39:235-245.

Hess, S. and D.A. Hensher. 2010. "Using Conditioning on Observed Choices to Retrieve Individual-Specific Attribute Processing Strategies." *Transportation Research B* 44(6):781-

790.

Hess S. and J. Rose. 2012. "Can Scale and Coefficient Heterogeneity be Separated in Random Coefficients Models?" *Transportation* 39(6):1225-1239.

Hess, S. and D.A. Hensher. 2013. "Making Use of Respondent Reported Processing Information to Understand Attribute Importance: A Latent Variable Scaling Approach." *Transportation* 40(2):397-412.

Hodgkins, C., J. Barnett, G. Wasowicz-Kirylo, M. Stysko-Kunkowska, Y. Gulcan, Y. Kustepeli, S. Akgungor, G. Chryssochoidis, L. Fernandez-Celemin, S. Strocksdieck, M. Gibbs, and M. Raats. 2012. "Understanding How Consumers Categorise Nutritional Labels: A Consumer Derived Typology for Front-of-Pack Nutrition Labelling." *Appetite* 59:806-817.

Jetter, K.M. and D.L. Cassady. 2006. "The Availability and Cost of Healthier Food Alternatives." *American Journal of Preventative Medicine*, 30(1):38-44.

Keane, M. and N. Wasi. 2012. "Comparing Alternative Models of Heterogeneity in Consumer Choice Behavior." *Journal of Applied Econometrics* 28(6):1018-1045.

Koop, G. 2003. *Bayesian Econometrics*. Chichester: John Wiley & Sons Ltd.

Louviere, J.L., D.A. Hensher, and J.D. Swait. 2001. *Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation.* Cambridge: Cambridge University Press.

Lowe, B., D.M. Souza Monteiro, and I.M. Fraser. 2013. "Nutritional Labelling Information: Utilisation of New Technologies." *Journal of Marketing Management* 29(11-12):1337-1366.

Lowe, B., I.M. Fraser, and D.M. Souza Monteiro. 2015a. "Changing Food Consumption Behaviors." *Psychology & Marketing*, 32(5): 481-485.

Lowe, B., I.M. Fraser, and D.M. Souza Monteiro. 2015b. "A Change for the Better? Digital Health Technologies and Changing Food Consumption Behaviors." *Psychology & Marketing*, 32(5): 585-600.

Lusk, J.L., T.C. Schroder, and G.T. Tonsor. 2014. "Distinguishing Beliefs from Preferences in Food Choice." *European Review of Agricultural Economics* 41(4):627-655.

Mazzocchi, M., W.B. Traill, and J.F. Shogren. 2009. *Fat Economics. Nutrition, Health, and Economic Policy.* Oxford: Oxford University Press.

Nordfalt, J. 2009. "Unplanned Grocery Purchases: The Influence of the Shopping-Trip Type Revisited." *Journal of Consumer Behaviour* 8:1-13.

Olsen, S.B. 2009. "Choosing Between Internet and Mail Survey Modes for Choice Experiment Surveys Considering Non-Market Goods." *Environmental and Resource Economics* 44:591-610.

Petrin, A. and K. Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research* 47:3-13.

Salaün, Y. and K. Flores. 2001. "Information Quality: Meeting the Needs of the Consumer." *International Journal of Information Management* 21:21-37.

Savage, S.J. and D.M. Waldman. 2008. "Learning and Fatigue During Choice Experiments: A Comparison of Online and Mail Survey Modes." *Journal of Applied Econometrics* 23:351-371.

Scarpa, R. and J.M. Rose. 2008. "Design Efficiency for Non-Market Valuation with Choice Modelling: How to Measure it, What to Report and Why." *Australian Journal of Agricultural and Resource Economics* 52:253-282.

Scarpa, R., M. Thiene, and D.A. Hensher. 2010. "Monitoring Choice Task Attribute Attendance in Nonmarket Valuation of Multiple Park Management Services: Does it Matter?" *Land Economics* 86(4):817-839.

Scarpa, R., S. Notaro, J. Louviere, and R. Raffaelli. 2011. "Exploring Scale Effects of Best/Worst Rank Ordered Choice Data to Estimate Benefits of Tourism in Alpine Grazing Commons." *American Journal of Agricultural Economics* 93(3):813-828.

Scarpa, R., Z. Raffaele, V. Bruschi, and S. Naspetti. 2013. "Inferred and Stated Attribute Non-Attendance in Food Choice Experiments." *American Journal of Agricultural Economics* 95(1):165-180.

Snowball, J.D. and K.G. Willis. 2011. "Interview Versus Self-Competion Questionnaires in Discrete Choice Experiments." *Applied Economic Letters* 18:1521-1525.

Stewart-Knox, B., S. Kuznesof, J. Robinson, A. Rankin, K. Orr, M. Duffy, R. Poinhos, M.D. Vaz de Almeida, A. Macready, C. Gallagher, A. Berezowska, A.R.H. Fischer, S. Navas-Carretero, M. Riemer, I. Traczyk, I.M.F. Gjelstad, C. Mavrogianni, and L.J. Frewer. 2013. "Factors Influencing European Consumer Uptake of Personalised Nutrition. Results of a Qualitative Analysis." *Appetite* 66:67-74.

Teisl, M.F. and B.E. Roe. 2010. "Consumer Willingness-to-Pay to Reduce the Probability of Retail Foodborne Pathogen Contamination." *Food Policy* 35:521-530.

Thiene, M., R. Scarpa, and J.J. Louviere. 2015. "Addressing Preference Heterogeneity, Multiple Scales and Attribute Attendance with a Correlated Finite Mixing Model of Tap Water Choice." *Enviromental and Resource Economics*, in press.

Train, K. 2009. *Discrete Choice Methods with Simulation,* 2nd ed. Cambridge: Cambridge University Press.

Train, K. and G. Sonnier. 2005. "Mixed Logit with Bounded Distributions of Part-worths." In R. Scarpa and A. Alberini, eds. *Applications of Simulation Methods in Environmental Resource Economics.* Dordrecht: Kluwer Academic Publishing, pp. 117-134.

Van Camp, D., D.M. Souza Monteiro, and N.H. Hooker. 2012. "Stop or Go? How is the UK Food Industry Responding to Front-of-Pack Nutrition Labels?" *European Review of Agricultural Economics* 39(5):821-842.

van Ittersun, K., J.M.E. Pennings, and B. Wansink. 2010. "Trying Harder and Doing Worse: How Grocery Shoppers Track In-Store Spending." *Journal of Marketing* 74(March):90-104.

Walters, R.G. and M. Jamil. 2003. "Exploring the Relationships Between Shopping Trip Type, Purchases of Products on Promotion, and Shopping Basket Profit." *Journal of Business Research* 56:17-29.

Wansink, B. 2003. "How Do Front and Back Package Labels Influence Beliefs About

Health Claims." *Journal of Consumer Affairs* 37(2):305-316.

Wansink, B. 2015. "Change Their Choice! Changing Behavior Using the CAN Approach and Activism Research." *Psychology & Marketing*, 32(5): 486-500.

Wansink, B., D.R. Just, and C.R. Payne. 2009. "Mindless Eating and Healthy Heuristics for the Irrational." *American Economic Review: Papers and Proceedings* 99(2):165-169.

Table 1: Marginal Log Likelihoods

Model Specifications	**MargLL**
Model 1N: Mixed Logit (Normal)	-5783.97
Model 1L: Mixed Logit (Log-normal)	-5628.75
Model 1NT: Mixed Logit (Normal) + Time	-5730.72
Model 1LT: Mixed Logit (Log-Normal) + Time	**-5611.44**
Model 2N: Mixed Logit and ANA (Normal)	-5629.02
Model 2L: Mixed Logit and ANA (Log-normal)	-5526.92
Model 2NT: Mixed Logit and ANA (Normal) + Time	-5600.96
Model 2LT: Mixed Logit and ANA (Log-normal) + Time	**-5505.01**
Model 3N: Mixed Logit and AIR (Normal)	-5570.70
Model 3L: Mixed Logit and AIR (Log-Normal)	-5466.79
Model 3NT: Mixed Logit and AIR (Normal) + Time	-5549.55
Model 3LT: Mixed Logit and AIR (Log-Normal) + Time	**-5449.29**

Note: Models in bold are best by model type (1,2 and 3)

Table 2: Best Model Results

Model 1LT	**Mean α**	**St Dev α**	**Mean Var**	**St Dev Var**
Appearance	-0.158	0.078	0.185	0.055
Nutrition Label	0.372	0.087	0.641	0.131
Allergy	0.993	0.086	0.977	0.172
Diet	0.841	0.110	1.072	0.204
Price	-0.813	0.472	11.03	3.052
ASC NB	-1.849	0.986	46.79	11.71
Hetero (ϕ)	0.383	0.052		
Model 2LT	**Mean α**	**St Dev α**	**Mean Var**	**St Dev Var**
Appearance	-0.148	0.087	0.178	0.068
Nutrition Label	0.437	0.094	0.812	0.165
Allergy	1.5589	0.123	1.676	0.313
Diet	1.144	0.106	1.335	0.233
Price	-0.978	0.181	4.884	0.892
ASC NB	-0.037	0.369	56.86	6.195
Hetero (ϕ)	0.398	0.053		
ANA Coeff ($\tilde{\rho}$)	0.269	0.036		
Model 3LT	**Mean α**	**St Dev α**	**Mean Var**	**St Dev Var**
Appearance	-0.244	0.131	0.391	0.140
Nutrition Label	0.621	0.120	1.264	0.280
Allergy	2.033	0.156	2.559	0.494
Diet	1.669	0.165	3.174	0.631
Price	-0.745	0.221	6.052	1.259
ASC NB	-0.597	0.499	51.23	6.853
Hetero (ϕ)	0.352	0.053		
AIR Coeff ($\tilde{\rho}$)	0.884	0.030		

Note: n=791 respondents, 12 cards and 3 options yielding 28,476 observations.

Table 3: Median WTP Estimates

	Normal			Log-Normal		
Model	1N	2N	3N	1L	2L	3L
Appearance	-0.051	-0.037	-0.027	-0.033	-0.077	-0.073
Nutrition Label	0.133	0.146	0.139	0.054	0.308	0.215
Allergy	0.524	0.425	0.344	0.330	1.469	0.958
Diet	0.438	0.361	0.341	0.199	0.978	0.546
	Normal + Time			**Log-Normal + Time**		
Model	1NT	2NT	3NT	1LT	2LT	3LT
Appearance	-0.049	-0.035	-0.040	-0.036	-0.055	-0.063
Nutrition Label	0.134	0.147	0.146	0.110	0.304	0.276
Allergy	0.531	0.431	0.341	0.764	1.520	1.026
Diet	0.438	0.370	0.354	0.408	1.009	0.675

Note: n=791 respondents, 12 cards and 3 options yielding 28,476 observations.

Table 4: **WTP Estimates (Full Sample)**

Model 1LT	**Median**	**Lower Quartile**	**Upper Quartile**
Appearance	-0.036	-1.323	0.103
Nutrition Label	0.111	-0.099	2.725
Allergy	0.764	0.031	9.164
Diet	0.409	0.008	4.528
Model 2LT	**Median**	**Lower Quartile**	**Upper Quartile**
Appearance	-0.055	-0.601	0.260
Nutrition Label	0.304	-0.143	3.192
Allergy	1.520	0.197	8.724
Diet	1.009	0.102	5.847
Model 3LT	**Median**	**Lower Quartile**	**Upper Quartile**
Appearance	-0.063	-1.288	0.048
Nutrition Label	0.276	-0.042	3.389
Allergy	1.026	0.116	7.111
Diet	0.675	0.056	4.128

Note: n=791 respondents, 12 cards and 3 options yielding 28,476 observations.

Table 5: WTP Estimates (More than 7 Minutes Sample)

Model 3LT	Median	Lower Quartile	Upper Quartile
Appearance	-0.018	-1.544	0.077
Nutrition Label	0.327	-0.010	3.931
Allergy	1.090	0.175	6.156
Diet	0.546	0.047	3.017

Note: n=431 respondents,12 cards and 3 options yielding 15,516 observations.

"*New Product Description Please carefully read the following new product description. Imagine you are considering purchasing thisproduct, we would like your thoughts on how this service might be developed. Please note this is a hypothetical product and we have no commercial interest in this new service. Consider you are doing your weekly food shopping in a supermarket and you are purchasing the following list of food products*:

12 inch pepperoni pizza	*Chicken curry & rice (400g pack)*
Froze battered cod fish (400g pack)	*Coca-cola pack 8x250 ml*
Butchers lamb chops (450 g)	*Corn flakes (500 g)*
Bottle of extra virgin olive oil (500 ml)	*Tomato soup (2 pints)*
2 cans of tuna in water	*Pasta (2 x 500 g packages)*
6 Pink Lady apples	*Fruity Bars (pack 5 bars 25 g each)*
Multipack of crisps (12 bags of 120 g each)	*Ginger nut biscuits (200 gr)*
Cheddar cheese (400 g)	

Currently, if you wanted to find out the nutritional value of these products as you add them to your basket or trolley, you would need to read the nutrition labels on the front or the back of packages. You then would need to keep a mental tally of the nutritional content for all the foods you have purchased. Suppose there is an alternative way to access nutrition information from packages, based on widely available technologies. The technological device can read QR labels and display information on a portable device as shown in figure 1(eg, mobile phone, hand held scanner). The advantage of this device is that it allows you to instantaneously access nutrition information for the foods you are purchasing as you are adding them to your trolley. Furthermore, the device can be developed such that you may add features that you would like to have."

Figure 1: Concept statement for shopping list

Attributes	Option A	Option B	Option C
Information display	Itemized	Whole foods	I would not buy either of option A or B
Nutrition label format	Traffic lights	Hybrid	
Allergy alert	Available	Available	
Diet alert	Not available	Available	
Price per use	£2.5	£5	

Figure 2: Example choice card