JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **28**: 1018–1045 (2013)
Published online 24 September 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2304

COMPARING ALTERNATIVE MODELS OF HETEROGENEITY IN CONSUMER CHOICE BEHAVIOR

MICHAEL KEANE* AND NADA WASI
School of Economics, University of New South Wales, Sydney, Australia

SUMMARY

When modeling demand for differentiated products, it is vital to adequately capture consumer taste heterogeneity, But there is no clearly preferred approach. Here, we compare the performance of six alternative models. Currently, the most popular are mixed logit (MIXL), particularly the version with normal mixing (N-MIXL), and latent class (LC), which assumes discrete consumer types. Recently, several alternative models have been developed. The ’generalized multinomial logit’ (G-MNL) extends N-MIXL by allowing for heterogeneity in the logit scale coefficient. Scale heterogeneity logit (S-MNL) is a special case of G-MNL with scale heterogeneity only. The ’mixed-mixed’ logit (MM-MNL) assumes a discrete mixture-of-normals heterogeneity distribution. Finally, one can modify N-MIXL by imposing theoretical sign constraints on vertical attributes. We call this ’T-MIXL’. We find that none of these models dominates the others, but G-MNL, MM-MNL and T-MIXL typically outperform the popular N-MIXL and LC models. Copyright © 2012 John Wiley & Sons, Ltd.

Received 27 July 2010; Revised 24 May 2012

Supporting information may be found in the online version of this article.

1. INTRODUCTION

For at least 25 years, a large research program in economics and marketing has focused on modeling demand for differentiated products. Within economics, this program is especially active in the industrial organization (IO), transport, environmental and health fields. From the outset, this literature has emphasized the great extent of heterogeneity in brand preferences across consumers, and the resultant ’loyalty’ that consumers exhibit towards particular brands.

Thus a large literature has emerged on modeling consumer taste heterogeneity. The treatment of heterogeneity in demand models is important for many reasons. Most obviously, estimates of price elasticities may be severely biased if one fails to account for heterogeneity. But taste heterogeneity is critical for a host of other problems, such as new product development, product positioning and advertising, price discrimination, development of menus of product offerings, product image and/or brand equity considerations, and consumer welfare calculations.

For example, consider Small *et al.* (2005, 2006). Using revealed and stated preference data on express vs. regular lane choices of commuters on California State Route 91, they find substantial heterogeneity in how motorists value speedy and reliable travel. Given the estimated distribution of tastes, they show how to design a more efficient fee structure. As in most studies, they find most heterogeneity is ’unobserved’ (not explained by observed consumer attributes).

The multinomial logit (MNL) model of McFadden (1974), long the ’workhorse’ model of multinomial choice, assumes homogeneous tastes for observed product attributes. But given the evidence for the importance of *unobserved* taste heterogeneity built up over the past 25 years, most researchers would now agree that simple MNL is inadequate to capture choice behavior in many contexts.

* Correspondence to: Michael P. Keane, School of Economics, University of New South Wales, Sydney, NSW 2052, Australia. E-mail: m.keane@unsw.edu.au

Consequently, many models have been developed that extend MNL to allow for unobserved taste heterogeneity. However, there is no consensus on a preferred approach.

Our aim is to compare the performance of several alternative models of taste heterogeneity, using a wide range of datasets. Two of the most popular alternatives are (i) the latent class model (LC) (see, for example, Kamakura and Russell, 1989; Arcidiacono and Jones, 2003) and (ii) the mixed logit model (MIXL) (see, for example, McFadden and Train, 2000; Hensher and Greene, 2003; Bajari *et al*., 2007). These models extend MNL to allow for unobserved taste heterogeneity. In LC, consumers come from a finite set of types, each with a different vector of preference weights. In contrast, an infinity of MIXL models may be obtained by choosing different mixing distributions, either continuous or discrete (so that LC is actually a special case of MIXL).

Most applications of MIXL, however, assume the vector of preference weights is multivariate normal (MVN) in the population. We refer to this model as N-MIXL. A few papers use one-sided distributions, especially log-normals, to theoretically constrain the sign on price and other vertical attributes. We call this version T-MIXL (where T stands for theory-constrained).

However, the fact that MIXL nests an infinite number of models raises an interesting methodological point: it is obviously cumbersome for a researcher to seek the ’best’ marginal distribution for each attribute in a choice model (and then combine these into the ’best’ joint distribution); and any such search is necessarily ad hoc, as not every possible distribution can be considered. This raises the question of whether the specification search can be *automated*.

One option is to adopt a MIXL model where the mixing distribution is a discrete mixture-of-normals. We call this a mixture-of-normals logit, or ’mixed-mixed’ logit (MM-MNL). The appeal of MM-MNL is that a mixture-of-normals can approximate any distribution arbitrarily well. Indeed, it has been shown to fit choice behavior better than N-MIXL in some recent studies (Rossi *et al*., 2005; Burda *et al*., 2008). Thus it may be both easier and more systematic for a researcher to use established methods (information criteria) to choose the number of elements of a normal mixture, rather than searching over many possible parametric mixing distributions.

Alternatively, Fiebig *et al*. (2010) have recently proposed two new models: scale heterogeneity logit (S-MNL) and generalized multinomial logit (G-MNL). S-MNL extends MNL by letting the scale of the errors vary across consumers. Fiebig *et al*. (2010) find that this simple structure captures much of the taste heterogeneity in many datasets. The G-MNL model nests S-MNL and N-MIXL. In G-MNL the heterogeneity distribution is a *continuous* mixture of scaled normals. In contrast to MM-MNL, the researcher does not choose the number of discrete mixture elements, but instead estimates a single scaling (or dispersion) parameter.

Fiebig *et al*. (2010) find that G-MNL fits choice behavior much better than N-MIXL across a wide range of datasets. As we noted, some recent papers also find that MM-MNL fits better than N-MIXL. However, there is no study comparing the new G-MNL and MM-MNL models, nor is there any study that systematically compares them to the T-MIXL or LC approaches.

Thus our plan is to run a ’horse race’ between the six alternative models of heterogeneity (G-MNL, S-MNL, MM-MNL, T-MIXL, N-MIXL and LC), by comparing their performance across 10 datasets that cover a wide range of products. One way to summarize our results is how often each model is preferred by the Bayes information criterion (BIC). We find MM-MNL is preferred in four datasets, the S-MNL special case of G-MNL is preferred in three, the full G-MNL model is preferred in one and T-MIXL is preferred in two. These diverse results suggest researchers should investigate multiple approaches to modeling heterogeneity in any given application.

Strikingly, N-MIXL and LC, the most popular models of heterogeneity in the literature, are never preferred; their BIC values rarely even come close to the best models. Nevertheless, we show that LC is still very useful for understanding the structure of heterogeneity in each dataset.

Of course, knowing one model fits better than another is not all that matters. It is also important to understand what *behavioral patterns* each captures better. Specifically, we find that G-MNL,

DOI: 10.1002/jae

MM-MNL and T-MIXL all do a much better job than N-MIXL at capturing consumer behavior that is ’extreme’ or lexicographic, meaning choice is based largely on a single attribute (e.g. lowest price, highest quality). *At the same time*, these three models are also better at capturing ’random’ behavior (i.e. choice is only influenced slightly by observed attributes).

Which is preferred among G-MNL, T-MIXL and MM-MNL is subtler, as they tend to give a similar overall fit. MM-MNL is preferred in datasets where the structure of heterogeneity is ’complex’ (in a sense made more precise below). This complexity justifies the large number of parameters that typically arise in discrete mixture models. T-MIXL or G-MNL (or its S-MNL special case) is preferred over MM-MNL when heterogeneity is ’simpler’. We find a key feature of ’complex’ heterogeneity structures is existence of small segments of consumers with a strong preference for attributes not viewed as important by the majority. For example, in pizza, many consumers place great weight on ’major’ attributes like price and/or quality. But small segments of consumers also care greatly about ’minor’ attributes, like crust style, wood-fire cooking, etc.. MM-MNL provides a much better fit to consumers in these small segments.

We also conduct counterfactual simulations to examine differences in demand predictions of G-MNL, T-MIXL and MM-MNL. In most cases the three models predict similar effects of changing price or attributes. Predicted effects of changing ’major’ attributes are always similar, but in some cases the predicted effects of changing ’minor’ attributes are quite different.

2. ALTERNATIVE MODELS OF CONSUMER CHOICE BEHAVIOUR

In the traditional multinomial logit (MNL) model (McFadden, 1974), the utility to person n from choosing alternative j on purchase occasion (or in choice scenario) t is given by

$$U_{njt} = \beta x_{njt} + \varepsilon_{njt} \quad n = 1, \ldots, N; \quad j = 1, \ldots, J; \quad t = 1, \ldots, T \tag{1}$$

where x_{njt} is a K-vector of observed attributes of alternative j, β is a vector of utility weights (homogeneous across consumers) and the error ε_{njt} ~ i.i.d. extreme value. The x_{njt} for $j = 1, \ldots, J$ may include alternative specific constants (ASCs) to capture unobserved attributes of each option j. The i.i.d. extreme value assumption leads to a closed-form expression for the choice probabilities:

$$P(j|X_{nt}) = \exp\left(\beta x_{njt}\right) \Big/ \sum_{k=1}^{J} \exp(\beta x_{nkt})$$

where X_{nt} is the vector of attributes of all alternatives $j = 1, \ldots, J$. Due to the restrictive assumptions that (i) the ε_{njt} are i.i.d. and (ii) tastes for observed attributes are homogeneous, MNL imposes the special independence of irrelevant alternatives (IIA) structure on how changes in the x_{njt} can affect choice probabilities.

2.1. **Models with Unobserved Taste Heterogeneity: Mixed Logit and Latent Class models**

The MIXL model extends MNL to allow for random coefficients on observed attributes, but continues to assume that the ’idiosyncratic’ error component ε_{njt} is i.i.d. extreme value:[1]

[1] Other models that avoid IIA and/or allow for unobserved heterogeneous tastes over the observed product attributes have been proposed. These included the nested logit model (McFadden, 1978), the generalized extreme value (GEV) model (McFadden, 1978) and the multinomial probit (MNP) model (Thurstone, 1927).

$$U_{njt} = \beta_n x_{njt} + \varepsilon_{njt} \; n = 1, \ldots, N; \;\; j \; = \; 1, \ldots, J; \;\; t = 1, \ldots, T \tag{2}$$

The MIXL model is often written as

$$U_{njt} = (\beta + \eta_n)x_{njt} + \varepsilon_{njt} \quad n = 1, \ldots, N; \quad j = 1, \ldots, J; \quad t = 1, \ldots, T \tag{3}$$

Here, β is the vector of *mean* attribute utility weights in the population, while η_n is the person n specific deviation from the mean.

The mixing distribution (i.e. the distribution for β_n) in the MIXL model can in principle be anything. For example, by choosing $\eta_n \sim \mathrm{MVN}(0,\Sigma)$ or, equivalently,$\beta_n \sim \mathrm{MVN}(\beta,\Sigma)$, one obtains the MIXL model with normal mixing. We call this N-MIXL. Most MIXL applications in the literature do assume normal mixing.

There are exceptions, however. Some studies assume the price coefficient is (minus) log-normal to force it to have the theoretically correct negative sign. This also assures that the distributions of willingness-to-pay for product attributes have finite moments. In general, it may be desirable to adopt one-sided distributions for all vertical attributes. We will call MIXL models with theoretically restricted one-sided distributions on price and vertical attributes 'T-MIXL' models.

As we will see in Section 4, it is not always clear which attributes are vertical, and sign constraints can cause fit to deteriorate. Therefore finding the best-fitting T-MIXL model may require one to estimate many models with one-sided distributions on different attributes. Obviously this is much more difficult if one considers distributions besides the log-normal, so we only consider T-MIXL models where the one-sided distributions are log-normal (while others are normal).

In the special case that the mixing distribution is discrete we obtain the latent class (LC) model. In the LC model consumers belong to one of S latent classes (also called 'segments' or 'types'). The β_n differ across classes but are identical within each class. That is:

$$\beta_n = \beta_s \text{ with probability } w_{n,s} \text{ for} s \; = 1, \; \ldots, S, \text{ where} \sum_s w_{n,s} = 1 \text{ and } w_{n,s} > 0 \; \forall \; s \tag{4}$$

Here $w_{n,s}$ is the probability person n is a member of class s. $w_{n,s}$ may depend on characteristics of person n or may be assumed identical across consumers, in which case $w_{n,s}$=w_s. Of course, the number of classes is unknown a priori. Typically, the researcher estimates models with different numbers of classes, and the best model is chosen using BIC or AIC.

MIXL choice probabilities *conditional* on β_n have the logit form:

$$P(j|X_{nt}, \beta_n) = \exp\left(\beta_n x_{njt}\right) \Bigg/ \sum_{i=1}^{J} \exp(\beta_n x_{nit})$$

Let $y_{njt} = 1$ if person n chooses j on occasion t, and 0 otherwise. The probability of a sequence of choices $\left\{y_{njt}\right\}_{t=1}^{T}$ is the product of the period-by-period logit expressions. Since β_n is unobserved, the *unconditional* choice probabilities are obtained by integrating over all possible values of β_n. The choice probabilities for MIXL (with a continuous mixing distribution) and LC are given by

$$\mathrm{prob}\left(\left\{y_{njt}\right\}_{t=1}^{T}\right) = \int \left[\prod_t \prod_j \left(e^{\beta_n x_{njt}} \Big/ \sum\nolimits_i e^{\beta_n x_{nit}}\right)^{y_{njt}}\right] f(\beta_n) d\beta_n \tag{5}$$

$$\mathrm{prob}\left(\left\{y_{njt}\right\}_{t=1}^{T}\right) = \sum_{s=1}^{S} w_{n,s} \left[\prod_t \prod_j \left(e^{\beta_s x_{njt}} \Big/ \sum\nolimits_i e^{\beta_s x_{nit}}\right)^{y_{njt}}\right] \tag{6}$$

DOI: 10.1002/jae

In (5) β_n has the continuous distribution $f(\beta_n)$. If $f(\beta_n)$ is MVN we get N-MIXL. If it combines normal and log-normal elements we get T-MIXL. In (6) β_n has the discrete distribution defined in (4), giving the LC model. Equation (5) does not have a closed form but it can be simulated using

$$\hat{\text{prob}}\left(\left\{y_{njt}\right\}_{t=1}^{T}\right) = \frac{1}{D}\sum_{d=1}^{D}\prod_{t}\prod_{j}\left(e^{\beta_d x_{njt}}\Big/\sum\nolimits_{i} e^{\beta_d x_{nit}}\right)^{y_{njt}}$$

The simulator has a form similar to (6), except that the β_s, which are estimated parameters in (6), are replaced by draws from $f(\beta_n)$, denoted $\{\beta_d\}_{d=1}^{D}$. Also, the weight on each draw equals one.

Notably, recent work has raised concerns with these models. By using a finite number of classes, LC may understate the extent of heterogeneity in the data (see Elrod and Keane, 1995; Allenby and Rossi, 1998).

Furthermore, Louviere *et al.* (1999, 2002) argue that a major source of heterogeneity is 'scale heterogeneity'—i.e. a general scaling up or down of the entire vector of attribute weights—that is not captured explicitly by N-MIXL or T-MIXL.

2.2. Scale Heterogeneity and Generalized Multinomial Logit (G-MNL) Models

Recently, Fiebig *et al.* (2010) developed the G-MNL model, which extends N-MIXL by incorporating both scale heterogeneity and a random coefficient vector. To understand what scale heterogeneity means, note that the variance of the extreme value error in the MNL model is $\sigma^2\pi^2/6$, where σ, the scale parameter, is usually normalized to one to achieve identification. The simple MNL model can be written with the scale of the error made explicit:

$$U_{njt} = \beta x_{njt} + \varepsilon_{njt}/\sigma \quad n = 1, \dots, N; \quad j = 1, \dots, J; \quad t = 1, \dots, T$$

The scale heterogeneity logit (S-MNL) model assumes that σ is heterogeneous in the population, and hence denotes its value for person n by the scalar random variable σ_n:

$$U_{njt} = \beta x_{njt} + \varepsilon_{njt}/\sigma_n \; n = 1, \dots, N; \;\; j = 1, .., J; \;\; t = 1, \dots, T \tag{7}$$

In (7), all heterogeneity is in the variance of the error term, while the β vector is homogeneous. However, heterogeneity in scale is observationally equivalent to a certain type of heterogeneity in the utility weights. Multiplying (7) through by the scalar random variable σ_n, we obtain

$$U_{njt} = (\sigma_n\beta)x_{njt} + \varepsilon_{njt} \; n = 1, \dots, N; \;\; j = 1, .., J; \;\; t = 1, \dots, T \tag{8}$$

Note that equation (8) can be interpreted as an MIXL model with$\beta_n = \sigma_n\beta$. Thus, in S-MNL, the vector of utility weights β is scaled up or down proportionately across consumers by the scaling factor σ_n. (We discuss the separate identification of β and the σ_n distribution below.)

The G-MNL model nests S-MNL and N-MIXL. This can be done in two ways. The first approach, which we call G-MNL-I, combines (3) with (8):

$$U_{njt} = (\sigma_n\beta + \eta_n)x_{njt} + \varepsilon_{njt} \tag{9}$$

The other approach, called G-MNL-II, starts with N-MIXL and multiplies through by σ_n:

$$U_{njt} = \sigma_n(\beta + \eta_n)x_{njt} + \varepsilon_{njt} \tag{10}$$

DOI: 10.1002/jae

Both (9) and (10) include N-MIXL and S-MNL as special cases. In (9), the random part of the attribute coefficients maintains a constant variance as the mean attribute weights are scaled. In (10), the standard deviations of the random coefficients are scaled proportionately to their means.

G-MNL nests G-MNL-I and II by adding a parameter γ that determines how the standard deviation of the random coefficients is scaled. Specifically, the G-MNL model is given by

$$U_{njt} = [\sigma_n \beta + \gamma \eta_n + (1-\gamma)\sigma_n \eta_n] x_{njt} + \varepsilon_{njt} \tag{11}$$

If $\gamma = 1$ we get G-MNL-I (equation (9)), while if $\gamma = 0$ we get G-MNL-II (equation (10)). Of course, γ can take on a continuum of values indicating different ways of scaling the variance of β_n. Here we list key special cases of G-MNL[2]:

$\sigma_n = \sigma = 1$	$\mathrm{var}(\eta_n) = 0$		$\beta_n = \beta$	MNL
$\sigma_n \neq \sigma$	$\mathrm{var}(\eta_n) = 0$		$\beta_n = \sigma_n \beta$	S-MNL
$\sigma_n = \sigma = 1$	$\mathrm{var}(\eta_n) \neq 0$		$\beta_n = \beta + \eta_n$	N-MIXL
$\sigma_n \neq \sigma$	$\mathrm{var}(\eta_n) \neq 0$	$\gamma = 1$	$\beta_n = \sigma_n \beta + \eta_n$	G-MNL-I
$\sigma_n \neq \sigma$	$\mathrm{var}(\eta_n) \neq 0$	$\gamma = 0$	$\beta_n = \sigma_n(\beta + \eta_n)$	G-MNL-II

Fiebig *et al.* (2010) imposed $0 \le \gamma \le 1$ with the idea that G-MNL-I and II were extremal cases. Thus they used a logistic transformation $\gamma = \exp(\gamma^*)/(1+\exp(\gamma^*))$ and estimated γ^* instead of γ. This caused numerical problems because in many datasets γ^* ran off to $\pm\infty$ (sending γ to 1 or 0).[3]

In fact, there is no reason to impose that $0 \le \gamma \le 1$. If $\gamma < 0$ the standard deviation of β_n increases more than proportionately as β is scaled up by σ_n. If $\gamma > 1$ the standard deviation of β_n falls as β increases. Thus we estimate γ directly. This resolves the numerical problem; and often we estimate $\gamma < 0$, giving significant likelihood improvements over the constrained model.[4]

To complete the G-MNL model the distribution of σ_n must be specified. As it is the 'scale' parameter, it should have positive support. We have used the log-normal distribution, $\ln(\sigma_n) \sim N(\bar{\sigma}, \tau^2)$. Note that the parameters $\bar{\sigma}, \tau$ and β are not separately identified. To achieve identification, we estimate only β and τ and then calibrate $\bar{\sigma}$ so as to normalize $E[\sigma_n]$ to one. Thus the estimated β is interpretable as the mean vector of the random preference weights.

2.3. The Mixed-Mixed Multinomial Logit Model (MM-MNL)

A few discrete-choice papers use a mixture-of-normals for the heterogeneity distribution, but it is not common. In the Bayesian approach, some authors use mixture-of-normal priors for individual level parameters. For probit see Geweke and Keane (1999, 2001, 2007), and for logit see Rossi *et al.* (2005) or Burda *et al.* (2008).[5] The appeal of the discrete mixture-of-normals is that it can approximate any distribution arbitrarily well (Ferguson, 1973). Indeed, Figure 5.7 in Rossi *et al.* (2005) provides a nice illustration of the flexibility of the posterior distribution of household-level parameters in a mixture-of-normals model vs. an N-MIXL model.

[2] Note that so long as $\mathrm{var}(\eta_{nk}) \neq 0$, where k is the price coefficient, G-MNL will necessarily put some positive mass on price coefficients with the 'wrong' sign. So, as with N-MIXL, the moments of WTP are not defined.

[3] As $\gamma^* \to \pm\infty$ the derivative of the likelihood with respect to γ^* approaches 0, and the Hessian becomes singular, requiring that γ be pegged at 1 or 0.

[4] We thank seminar participants in the econometrics workshop at Princeton, particularly Bo Honore, for pointing out to us that there is no logical reason to constrain γ to the [0,1] interval.

[5] Burda *et al.* (2008) allow a *subset* of coefficients in the MIXL model to have mixture-of-normal distributions, while others have normal distributions. Train (2008) and Bajari *et al.* (2007) consider MIXL with a mixture-of-normals, but they focus on alternative estimation algorithms, not the fit of that model compared to alternative models.

DOI: 10.1002/jae

In a classical framework, we consider a model where the mixing distribution in N-MIXL is generalized to a discrete mixture-of-multivariate normals. That is, we replace (4) with

$$\beta_n \sim \mathrm{MVN}(\beta_s, , \Sigma_s) \text{ with probability } w_{n,s} \text{ for s } = 1, \ \ldots, S \tag{12}$$

We call this as a 'mixed-mixed' logit (MM-MNL). Note that if $w_{n,s} \to 0$ for all but one class, (12) becomes the N-MIXL model in (3). If $\Sigma_s \to 0 \ \forall \ s$, (12) becomes the LC model in (4). Thus MM-MNL nests N-MIXL and LC. The choice probabilities for MM-MNL are given by

$$\mathrm{prob}\left(\left\{y_{njt}\right\}_{t=1}^{T}\right) = \sum_{s=1}^{S} w_{n,s}\left\{\int\left[\prod_t \prod_j \left(e^{\beta_{n|s}x_{njt}} \Big/ \sum\nolimits_k e^{\beta_{n|s}x_{nkt}}\right)^{y_{njt}}\right] f\left(\beta_{n|s}\right) d\beta_{n|s}\right\} \tag{13}$$

where $f\,(\beta_{n|s})$ refers to MVN(β_s,Σ_s).

3. SOME NOTES ON THE ESTIMATION PROCEDURES

For MM-MNL, the probabilities in (13) can be simulated as follows. First, *conditional* on being in class s, the simulated probability of observing person n choose a sequence$\left\{y_{njt}\right\}_{t=1}^{T}$is

$$\hat{P}_n(s) = \frac{1}{D}\sum_{d=1}^{D}\prod_t\prod_j \left(P\left(j|X_{nt}, \eta^{s,d}, s\right)\right)^{y_{njt}} = \frac{1}{D}\sum_{d=1}^{D}\prod_t\prod_j \left[\frac{\exp\left(\beta^s + \eta^{s,d}\right)x_{njt}}{\sum_{k=1}^{J}\exp\left(\beta^s + \eta^{s,d}\right)x_{nkt}}\right]^{y_{njt}}$$

where $\eta^{s,d}$ is a K-vector distributed MVN(0,Σ_s). We draw $\{\eta^{s,d}\}$ for $d = 1, \ldots, D$; and $s = 1, \ldots, S$.

To obtain the simulated *unconditional* probability, we take a weighted average of these conditional probabilities: $\hat{P}_n = \sum_s w_{n,s}\hat{P}_n(s)$. The simulated log-likelihood for the sample is the sum of the simulated log-likelihood contributions for all individuals: $\ln\hat{L} = \sum_n \ln\hat{P}_n$.

Few personal characteristics are available in our datasets, so in the LC and MM-MNL models we set $w_{n,s} = w_s$. To impose $\sum_s w_s = 1$ we use a logistic transformation, $w_s = \exp\left(w_s^*\right) \div \left[1 + \sum_{s=1}^{S-1}\exp\left(w_s^*\right)\right]$, and set $w_s^* = 0$. To avoid cases where w_s^* may run off to infinity as we iterate, we set upper and lower bounds of 5 and −5, so the frequency of each class is at least 0.01. Also, when a large fraction of respondents choose based on one or two attributes, LC and MM-MNL tend to generate one class to capture their behavior (with the utility weights on those attributes running off to infinity). We impose upper and lower bounds on taste parameters to prevent this.

In MM-MNL parameters proliferate rapidly with the number of classes if Σ_s is a full covariance matrix. Thus we consider two alternative restrictions. First, we assume Σ_s is diagonal for all s, except for correlation among alternative specific constants. Second, we restrict covariance matrices for all classes to be proportional, $\Sigma_s = k_s\Sigma$, where Σ is a full covariance matrix.

The LC and MM-MNL models are estimated using several alternative values for S, the number of classes. We report results for the S values preferred by BIC. Estimation results for LC models can be sensitive to starting values. Thus, for each value of S, we use the solution to the model with one fewer class as starting values (and we also try 50 other randomly chosen starting points). We increase S until the model yields a smaller BIC than the model with one fewer class.

For N-MIXL, T-MIXL and G-MNL, we consider both the case where the covariance matrix of η, denoted Σ, is a full matrix and the case where Σ is diagonal (except that intercepts are still correlated). We again report results for the version of each model preferred by BIC.

For datasets where choices are labeled (e.g. buy or don't buy), all our models include alternative specific constants, or 'ASCs'. Of course, ASCs are not needed if choices are generic (e.g. Pizza A or Pizza B). Fiebig *et al*. (2010) found that scaling the ASCs in the S-MNL model leads to a poor fit.[6] We therefore assume instead that the ASCs in S-MNL are normally distributed random coefficients (β_{0n}) that are not scaled, giving the model

$$U_{njt} = \beta_{0n} + (\sigma_n\beta)x_{njt} + \varepsilon_{njt}\; n = 1, \ldots, N; \quad j = 1, .., J; \quad t = 1, \ldots, T \tag{14}$$

As neither β_{0n} nor ε_{njt} is scaled, this model implies that observed attributes x_{njt} are more important for choice (relative to *all* unobserved factors) for some consumers than others. This is in keeping with the scale heterogeneity idea.

4. EMPIRICAL RESULTS

We evaluate our six models of heterogeneity (LC, N-MIXL, T-MIXL, G-MNL, S-MNL, MM-MNL) using data from 10 stated preference (SP) discrete-choice experiments (DCEs). In recent years there has been a rapid increase in the use of DCEs in both marketing and economics to predict demand for products—especially new products and public goods, neither of which is traded *explicitly* in a market.[7] Economists have traditionally been skeptical of SP data, but many recent studies show that well-executed DCEs can give reliable predictions of demand (Carson *et al*., 1994; Louviere *et al*., 2000; Kanninen, 2007). Indeed, all of our SP datasets were collected for actual demand-forecasting exercises (i.e. they were not collected specifically for this paper).

Our datasets cover a wide range of settings. Three involve medical decision making (i.e. decisions about genetic and cervical cancer tests). Seven involve choice of various consumer products, ranging from pizza delivery services to holiday packages, mobile phones to charge cards. Supporting information Appendix A, Table A1, describes the general characteristics of each dataset (i.e. number of attributes, number of choices, number of choice occasions), while Table A2 gives details about the design of each experiment (i.e. the attributes of the choice alternatives).

In Tables I–V we report results for five datasets that illustrate well the contexts where alternative models of heterogeneity perform best. We report results for the other five datasets in the Appendix (supporting information), and give brief descriptions in the text. To save space, we only present a subset of the parameter estimates for each model. Also, we have generally estimated several different versions of each model (with different numbers of latent classes, correlated or uncorrelated errors, etc.). We only report results for the version preferred by BIC.[8]

4.1. Estimation Results for the 10 Datasets

Table I presents the results of the first dataset. Here, subjects were asked whether they would choose to receive diagnostic tests for Tay–Sachs disease, cystic fibrosis, both or neither, giving four alternatives. Covariates include cost of the tests, whether the person's doctor recommends it, risk factors and

[6] This is not surprising. Given that some consumers are loyal to particular brands while other consumers are loyal to other brands, one clearly must allow for heterogeneity in the *sign* of the various brand intercepts, not just the scale.

[7] See Small *et al*. (2005), Ortúzar and Rizzi (2003), Louviere and Street (2000) and Hensher (1994) for transport economics; Carson and Hanemann (2005), Adamowicz (2004), Bennett and Blamey (2001) for environmental economics; Whitehead *et al*. (2008) for agricultural economics; Ryan *et al*. (2007) and Guttmann *et al*. (2009) for health economics; and Louviere (1994) and Rao (2008) for surveys of the many papers in marketing.

[8] Given N people and T choices per person we have that BIC = −2LL + (#parameters) × log(NT). Monte Carlo work in Fiebig *et al*. (2010) found that BIC was the most reliable criterion for choosing the correct model in this type of data.

DOI: 10.1002/jae

alternative specific constants. The sample members are all Ashkenazi Jews, who have a relatively high probability of carrying Tay–Sachs.

Column 3 reports estimates of N-MIXL, which has a BIC of 5626. Mean preference weights have the expected signs and most are statistically significant. The mean ASCs are not significantly different from zero, but their estimated variances (not reported) are large and significant, implying substantial unobserved heterogeneity in how people value the test options.

Many T-MIXL models are possible here, as one could plausibly constrain the sign of all eight variables in this model (price, risk, accuracy, etc.). However, after considerable experimentation we found the best-fitting model only constrains the three price coefficients. This T-MIXL model with log-normal price coefficients is reported in column 4. It has a BIC value of 5563, which is a substantial 63-point improvement over N-MIXL. The experimentation required to arrive at this specification, however, illustrates that use of T-MIXL may require considerable care.

Column 5 reports estimates of the G-MNL model, which achieves a BIC of 5600. Thus G-MNL is preferred to N-MIXL but not to T-MIXL. The estimate of τ is 0.45, which implies substantial scale heterogeneity. As $\sigma_n = \exp(-\tau^2/2 + \tau\varepsilon_{0n})$, where $\varepsilon_{0n} \sim N(0,1)$, this value implies that a person at the 90th percentile of τ has his/her vector of utility weights scaled up by 57%, while a person at the 10th percentile has his/her utility weights scaled down by 46%. The estimate of γ is 0.07, so the data are much closer to the G-MNL-II model, where the variance of residual taste heterogeneity increases with scale, than to the G-MNL-I model, where it is invariant to scale.

Column 6 reports the LC estimates. BIC prefers a model with five classes. Given the large number of parameters, we only report β vectors for the three largest, which make up 76% of the population. The largest class (#1) places great weight on risk factors. Class 2 places great weight on cost. Class 3 has a high intercept for the ’both’ option, so they tend to get both tests regardless of attribute settings. Class 4 (not reported) cares *extremely* much about cost, and class 5 (not reported) behaves fairly ’randomly,’ with little effect of observed attributes on choice.

Column 7 presents estimates of MM-MNL. We tried several versions, with different numbers of classes, and independent or correlated random coefficient vectors. BIC preferred a mixture of two independent normal vectors, but with correlated intercepts. This model achieves a BIC of 5560. The larger class (class 1) places much more weight on risk factors, while the smaller one (class 2) places much more weight on costs.

In summary, BIC prefers MM-MNL over all alternative models, with T-MIXL a close second (5560 vs. 5563). Next are G-MNL (5600) and N-MIXL (5626). There is then a rather wide gap before we get to S-MNL (5777) and another wide gap before we get to LC (5882).

As we will see below, this pattern of LC performing poorly relative to other models holds across all 10 datasets. Nevertheless, we will also see that LC estimates are still useful for gaining an intuitive understanding of the nature of heterogeneity in each category. For instance, take the five classes identified by the LC model (described above), and compare them with posteriors of the attribute weights derived from the better-fitting MM-MNL, T-MIXL and G-MNL models. Most consumers have posteriors that imply behavior similar to one of the five LC model types. Also, the two classes identified by the preferred MM-MNL model are similar to the two *largest* classes identified by the LC model (i.e. class 1 cares a lot about risk; class 2 cares a lot about costs).[9]

Supporting information Table A3 reports results of an identical Tay–Sachs/cystic fibrosis diagnostic test choice experiment, except using a sample of the *general* population. Here, the MM-MNL model

[9] MM-MNL captures the behavior of smaller segments by relying on the randomness of its coefficient vectors. In contrast, G-MNL has no explicit segments (it has a single mean β vector). It captures behavior of various segments via the interaction of the random coefficients with the scaling parameter. For example, there are some cases where the random draws for the price coefficients are large, and the random draw for the scale parameter is also large. This generates behavior where consumers care very much about price.

DOI: 10.1002/jae

Table I. Tay–Sachs disease (TS) and cystic fibrosis (CF) test: Jewish sample (3 ASCs)

											Latent class[c]						MM-MNL[d]			
	MNL		S-MNL (with RE)		N-MIXL[a]		T-MIXL[a,b]		G-MNL[a]		Class 1		Class 2		Class 3		Class 1		Class 2	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
ASC for TS test	**−0.57**	0.14	**−0.57**	0.20	−0.67	0.47	−0.31	0.46	−0.21	0.41	−0.33	0.36	**−0.64**	0.28	2.64	2.16	0.20	0.48	−0.77	0.42
ASC for CF test	**−0.82**	0.15	**−0.88**	0.22	−0.74	0.42	−0.59	0.41	−0.37	0.36	−0.39	0.30	**−0.92**	0.35	3.18	2.35	0.44	0.50	**−0.89**	0.44
ASC for both tests	−0.08	0.15	0.01	0.27	−0.38	0.52	0.31	0.59	−0.09	0.45	**−0.98**	0.31	−0.42	0.29	**6.26**	2.63	0.18	0.59	**0.94**	0.43
TS cost	**−2.51**	0.24	**−3.45**	0.34	**−4.75**	0.63	**−7.14**	1.01	**−5.63**	0.73	**−2.03**	0.40	**−5.20**	0.68	−1.78	1.21	**−2.81**	0.43	**−9.53**	1.50
CF cost	**−1.43**	0.13	**−1.96**	0.20	**−3.24**	0.38	**−4.48**	0.73	**−3.55**	0.43	**−1.46**	0.22	**−3.03**	0.30	−2.50	1.29	**−2.32**	0.26	**−6.23**	1.08
Both cost	**−1.20**	0.07	**−2.70**	0.17	**−3.65**	0.26	**−5.45**	0.59	**−4.23**	0.34	**−1.88**	0.14	**−4.77**	0.32	**−1.94**	0.44	**−2.47**	0.18	**−5.15**	0.48
Recommend	**0.33**	0.04	**0.56**	0.06	**0.95**	0.13	**0.77**	0.13	**1.01**	0.19	**0.43**	0.10	**0.64**	0.08	0.72	0.89	**0.67**	0.14	**1.08**	0.18
Inaccuracy	**−0.12**	0.02	**−0.15**	0.03	−0.14	0.09	**−0.33**	0.09	**−0.35**	0.10	**−0.29**	0.06	**−0.12**	0.04	**−0.55**	0.26	**−0.40**	0.10	−0.06	0.07
Form	0.07	0.04	**0.12**	0.05	0.28	0.16	0.13	0.14	0.15	0.19	**−0.21**	0.10	**0.19**	0.08	−0.43	0.41	0.01	0.14	**0.52**	0.19
Own risk of TS	**0.50**	0.03	**1.05**	0.08	**1.39**	0.12	**1.41**	0.15	**1.66**	0.17	**1.43**	0.12	**0.81**	0.08	**0.96**	0.45	**1.81**	0.14	**0.46**	0.10
Own risk of CF	**0.47**	0.04	**1.02**	0.07	**1.26**	0.12	**1.23**	0.12	**1.51**	0.18	**1.30**	0.09	**0.77**	0.12	**0.66**	0.22	**1.61**	0.12	**0.38**	0.11
τ			**0.64**	0.06					**0.45**	0.08										
γ									0.07	0.15										
Class probability											**0.29**	0.03	**0.27**	0.03	**0.20**	0.03	0.62	0.04	0.38	0.04
No. of parameters	11		18		77		77		79		59						51			
LL	−3717		−2815		−2500		−2469		−2479		−2701						−2573			
BIC	7523		5777		5626		5563		5600		5882						**5560**			

[a] Estimates from correlated coefficient specification.

[b] Coefficients of TS cost, CF cost, and Both cost are assumed to be distributed as log-normal, while other coefficients are assumed to be normally distributed.

[c] Estimates from LC with five classes.

[d] Estimates from MM-MNL with two independent normals but the random intercepts in each class are allowed to be correlated.

Bold estimates are statistically significant at 5%.

Table II. Mobile phones (1 ASC)

											Latent class[c]						MM-MNL[d]			
	MNL		S-MNL (with RE)		N-MIXL[a]		T-MIXL[a,b]		G-MNL[a]		Class 1		Class 2		Class 3		Class 1		Class 2	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
ASC for purchase	**−0.80**	0.05	**−0.35**	0.12	**−0.50**	0.11	**−0.29**	0.13	**−0.45**	0.12	**−1.15**	0.14	**−2.96**	0.27	**1.49**	0.26	0.22	0.25	**−1.31**	0.20
No voice comm.	0.04	0.04	0.06	0.05	0.04	0.05	0.04	0.05	0.04	0.06	0.02	0.08	0.03	0.31	0.07	0.11	0.07	0.08	−0.02	0.12
Voice dialing	**0.08**	0.04	0.05	0.06	**0.10**	0.05	**0.11**	0.05	0.08	0.06	0.08	0.09	0.23	0.26	−0.12	0.12	0.12	0.08	0.10	0.12
Voice operation	**−0.12**	0.04	−0.11	0.06	**−0.13**	0.05	**−0.13**	0.05	**−0.12**	0.06	**−0.21**	0.10	−0.37	0.39	0.07	0.11	−0.08	0.08	−0.21	0.14
No push to com.	0.06	0.04	**0.12**	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.10	−0.22	0.32	0.18	0.12	0.07	0.08	0.05	0.12
Push to talk	0.03	0.04	0.03	0.07	0.05	0.05	0.06	0.05	0.07	0.06	0.17	0.09	−0.21	0.39	0.05	0.14	0.00	0.09	0.12	0.11
Push to share pics/video	−0.02	0.04	−0.08	0.07	−0.02	0.05	−0.03	0.05	−0.04	0.06	**−0.23**	0.11	0.51	0.28	−0.06	0.13	0.05	0.09	−0.18	0.13
Personal e-mail	−0.07	0.04	−0.04	0.06	−0.08	0.05	−0.08	0.05	−0.07	0.06	−0.15	0.10	0.32	0.27	0.03	0.13	−0.03	0.09	−0.13	0.13
Corporate e-mail	**0.09**	0.04	0.08	0.07	0.08	0.05	0.08	0.05	0.08	0.06	0.10	0.08	0.00	0.31	−0.01	0.14	0.06	0.09	0.09	0.11
both e-mails	−0.05	0.04	−0.08	0.06	−0.03	0.05	−0.03	0.05	−0.04	0.06	0.08	0.09	−0.41	0.39	−0.05	0.13	−0.11	0.09	0.08	0.12
WiFi	0.001	0.02	−0.02	0.03	−0.002	0.03	0.00	0.03	−0.01	0.03	0.08	0.06	0.05	0.17	−0.08	0.07	−0.07	0.05	0.09	0.07
USB cable/cradle	**0.06**	0.03	**0.08**	0.04	**0.07**	0.03	**0.07**	0.03	**0.08**	0.04	0.05	0.06	−0.01	0.18	**0.20**	0.08	0.08	0.05	0.07	0.07
Thermometer	**0.07**	0.03	0.05	0.03	**0.07**	0.03	**0.07**	0.03	**0.08**	0.03	0.05	0.05	0.00	0.18	0.10	0.06	**0.11**	0.05	0.02	0.07
Flashlight	0.05	0.03	0.01	0.03	0.05	0.03	0.05	0.03	0.04	0.03	**0.16**	0.06	−0.10	0.17	−0.03	0.08	−0.02	0.05	**0.18**	0.07
Price/100	**−0.32**	0.02	**−1.02**	0.16	**−0.76**	0.06	**−1.47**	0.36	**−0.88**	0.10	−0.04	0.05	**−0.64**	0.20	**−2.06**	0.21	**−1.57**	0.20	−0.05	0.08
τ			**1.45**	0.15					**0.69**	0.19										
γ									0.08	0.24										
Class probability											**0.32**	0.03	**0.28**	0.03	**0.22**	0.03	**0.67**	0.05	**0.33**	0.05
No. of parameters	15		17		30		30		32		63						61			
LL	−4475		−3990		−3971		−3978		−3966		−3952						−3927			
BIC	9074		**8121**		8190		8204		8197		8426						8359			

[a] Estimates from uncorrelated coefficient specification.

[b] Coefficient of price is assumed to be distributed as log-normal, while other coefficients are assumed to be normally distributed.

[c] Estimates from LC with four classes.

[d] Estimates from MM-MNL with two independent normals.

Bold estimates are statistically significant at 5%.

Table III. Pizza A (no ASC)

											Latent class[c]						MM-MNL[d]			
	MNL		S-MNL		N-MIXL[a]		T-MIXL[a,b]		G-MNL[a]		Class 1		Class 2		Class 3		Class 1		Class 2	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Gourmet	0.02	0.02	0.03	0.04	0.03	0.05	0.01	0.05	**0.49**	0.24	−0.01	0.05	0.02	0.02	0.08	0.10	0.02	0.07	0.14	0.47
Price	**−0.16**	0.02	**−0.19**	0.05	**−0.35**	0.06	**−1.17**	0.45	**−1.82**	0.72	**−0.20**	0.06	**−0.16**	0.03	**−0.39**	0.09	**−0.18**	0.06	−4.63	2.71
Ingredient freshness	**0.48**	0.03	**1.45**	0.29	**0.96**	0.08	**4.09**	1.71	**5.06**	1.91	**1.57**	0.09	**0.12**	0.06	0.30	0.16	**0.59**	0.08	13.47	7.73
Delivery time	**0.09**	0.03	**0.16**	0.08	**0.16**	0.05	**0.28**	0.12	**0.81**	0.39	0.10	0.09	**0.10**	0.04	**0.32**	0.09	0.06	0.05	3.95	2.36
Crust	0.02	0.03	0.01	0.04	0.02	0.06	0.01	0.07	0.48	0.29	**−0.12**	0.06	0.01	0.05	**−0.30**	0.09	−0.06	0.08	1.18	1.05
Sizes	**0.09**	0.03	**0.12**	0.06	**0.20**	0.05	**0.18**	0.05	**0.88**	0.40	**0.15**	0.07	0.06	0.04	**0.23**	0.11	**0.23**	0.07	0.92	0.81
Steaming hot	**0.38**	0.03	**1.02**	0.24	**0.87**	0.08	**0.88**	0.08	**4.86**	1.84	**0.50**	0.08	**0.12**	0.06	**1.60**	0.18	**0.50**	0.08	9.85	5.76
Late open hours	**0.04**	0.02	0.08	0.06	0.07	0.05	0.08	0.05	0.30	0.17	0.09	0.08	**0.06**	0.03	0.02	0.07	**0.12**	0.06	−0.97	0.72
τ	—		**1.69**	0.18	—				**1.80**	0.24										
γ									−0.02	0.02										
Class probability											**0.36**	0.04	**0.32**	0.04	**0.23**	0.04	**0.57**	0.04	**0.43**	0.04
No. of parameters	8		9		16		16		18		35						33			
LL	−1657		−1581		−1403		−1342		−1372		−1418						−1328			
BIC	3378		3233		2933		2812		2886		3115						2919			

[a] Estimates from uncorrelated coefficient specification.

[b] Coefficients of price, ingredient freshness and delivery time are assumed to be distributed as log-normal, while other coefficients are assumed to be normally distributed. If only price coefficient is assumed to be log-normal distributed, BIC of T-MIXL is 2907.

[c] Estimates from LC with four classes.

[d] Estimates from MM-MNL with two independent normals.

Bold estimates are statistically significant at 5%.

Table IV. Pizza B (no ASC)

											Latent class[c]						MM-MNL[d]					
	MNL		S-MNL		N-MIXL[a]		T-MIXL[a,b]		G-MNL[a]		Class 1		Class 2		Class 3		Class 1		Class 2		Class 3	
	Est	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Gourmet	0.01	0.01	**0.05**	0.01	0.01	0.02	0.02	0.02	0.02	0.04	0.01	0.02	0.02	0.07	0.09	0.05	−0.03	0.04	−0.12	0.07	**0.37**	0.08
Price	**−0.17**	0.01	**−0.25**	0.02	**−0.30**	0.03	**−0.71**	0.16	**−0.94**	0.10	−0.04	0.02	**−1.71**	0.09	0.24	0.11	−0.10	0.04	**−0.86**	0.10	−0.17	0.13
Ingredient freshness	**0.21**	0.01	**0.36**	0.03	**0.34**	0.03	**1.65**	0.50	**1.22**	0.11	**0.10**	0.02	**0.46**	0.06	**2.17**	0.19	**0.12**	0.03	**0.29**	0.07	**1.02**	0.13
Delivery time	**0.03**	0.01	0.04	0.02	**0.05**	0.02	**0.06**	0.02	**0.21**	0.06	0.02	0.02	0.14	0.10	−0.03	0.16	0.02	0.03	0.19	0.07	0.14	0.08
Crust	**0.08**	0.01	**0.09**	0.01	**0.08**	0.03	**0.12**	0.03	**0.64**	0.07	**−0.04**	0.01	−0.05	0.04	**0.31**	0.08	−0.03	0.03	**0.62**	0.09	0.15	0.07
Sizes	**0.07**	0.01	**0.08**	0.02	**0.11**	0.02	**0.10**	0.02	**0.21**	0.04	**0.05**	0.02	**0.19**	0.07	**0.28**	0.07	0.06	0.03	**0.31**	0.07	**0.26**	0.09
Steaming hot	**0.20**	0.01	**0.35**	0.03	**0.34**	0.02	**0.38**	0.03	**1.42**	0.14	**0.10**	0.02	**0.22**	0.07	**0.67**	0.07	**0.11**	0.03	**0.37**	0.06	**1.43**	0.17
Late open hours	**0.04**	0.01	0.02	0.02	**0.08**	0.02	**0.07**	0.02	0.10	0.04	**0.04**	0.01	0.06	0.06	0.07	0.10	0.01	0.02	**0.29**	0.07	**0.19**	0.06
Free delivery charge	**0.12**	0.01	**0.15**	0.02	**0.20**	0.02	**0.24**	0.02	**0.69**	0.08	**0.11**	0.01	**0.56**	0.04	0.15	0.08	**0.22**	0.05	**0.26**	0.06	**0.28**	0.07
Local store	**0.08**	0.01	**0.06**	0.02	**0.15**	0.02	**0.15**	0.02	**0.60**	0.07	**0.14**	0.01	−0.01	0.07	0.10	0.12	**0.09**	0.03	**0.43**	0.07	0.08	0.08
Baking method	**0.07**	0.01	**0.07**	0.02	**0.11**	0.02	**0.11**	0.02	**0.27**	0.05	**0.06**	0.01	0.16	0.07	**0.29**	0.07	0.01	0.03	**0.32**	0.06	**0.35**	0.11
Manners	0.01	0.01	−0.004	0.02	0.02	0.02	0.02	0.02	0.02	0.06	0.03	0.02	0.03	0.08	−0.06	0.11	0.03	0.03	−0.06	0.08	0.11	0.11
Vegetarian availability	**0.09**	0.01	**0.06**	0.01	**0.13**	0.03	**0.14**	0.03	**0.42**	0.08	0.02	0.02	**0.15**	0.04	0.04	0.11	0.04	0.03	**0.35**	0.09	0.04	0.07
Delivery time guaranteed	**0.07**	0.01	**0.07**	0.02	**0.11**	0.02	**0.11**	0.02	**0.15**	0.04	**0.08**	0.02	**0.17**	0.05	0.12	0.12	**0.14**	0.04	0.07	0.08	**0.19**	0.07
Distance to the outlet	**0.06**	0.01	0.04	0.02	**0.09**	0.02	**0.08**	0.02	0.08	0.05	**0.09**	0.02	0.11	0.07	−0.12	0.10	**0.11**	0.04	0.09	0.07	0.06	0.07
Range/variety availability	**0.06**	0.02	0.04	0.02	**0.09**	0.02	**0.09**	0.02	0.15	0.06	**0.07**	0.03	0.03	0.07	0.07	0.10	**0.10**	0.03	0.03	0.07	0.19	0.08
τ	—		**1.22**	0.08	—				**1.50**	0.08												
γ									−0.05	0.02												
Class probability											**0.51**	0.03	**0.14**	0.02	**0.12**	0.02	**0.41**	0.03	**0.31**	0.03	**0.28**	0.03
No. of parameters	16		17		32		32		34		101						98					
LL	−6747		−6607		−5892		−5652		−5662		−5591						−5310					
BIC	13,641		13,372		12,081		11,600		11,639		12,118						**11,527**					

[a] Estimates from uncorrelated coefficient specification.

[b] Coefficients of price, ingredient freshness, delivery time and delivery charge are assumed to be distributed as log-normal, while other coefficients are assumed to be normally distributed. If only price coefficient is assumed to be log-normal distributed, BIC of T-MIXL is 11,868.

[c] Estimates from LC with six classes.

[d] Estimates from MM-MNL with three independent normals.

Bold estimates are statistically significant at 1%.

Table V. Papsmear test (1 ASC)

											Latent class[c]						MM-MNL[d]			
	MNL		S-MNL		N-MIXL[a]		T-MIXL[a,b]		G-MNL[a]		Class 1		Class 2		Class 3		Class 1		Class 2	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
ASC for test	**−0.40**	0.14	−0.60	0.37	**−1.26**	0.30	**−1.49**	0.34	**−1.01**	0.34	**−1.59**	0.22	4.31	9.57	**−1.37**	0.34	−0.16	0.43	−1.35	1.10
If know doctor	**0.32**	0.09	**0.63**	0.14	**0.78**	0.18	**−0.69**	0.14	**0.72**	0.22	0.02	0.27	−1.34	9.51	**1.27**	0.13	0.20	0.28	2.15	1.21
If doctor is male	**−0.70**	0.09	**−1.24**	0.16	**−1.39**	0.30	**−1.72**	0.38	**−1.92**	0.33	−0.18	0.25	0.90	4.55	**−0.75**	0.27	−0.40	0.23	**−6.14**	1.46
If test is due	**1.23**	0.10	**2.74**	0.29	**3.26**	0.31	**−3.44**	0.43	**3.44**	0.39	**3.15**	0.16	2.67	12.80	**0.88**	0.22	**3.20**	0.41	**3.82**	0.65
If doctor recommends	**0.51**	0.10	**0.74**	0.17	**1.33**	0.23	**−1.31**	0.21	**1.65**	0.34	**1.57**	0.18	0.62	15.60	**0.52**	0.27	**1.31**	0.38	**1.53**	0.69
Test cost	**−0.08**	0.04	**−0.17**	0.07	**−0.22**	0.09	**−0.23**	0.08	**−0.27**	0.09	**−0.18**	0.09	−0.50	1.85	−0.23	0.14	−0.16	0.12	−0.45	0.34
τ			**0.81**	0.11					**0.92**	0.13										
γ									−0.09	0.16										
Class probability											**0.37**	0.05	**0.20**	0.04	**0.19**	0.04	**0.70**	0.07	**0.30**	0.07
No. of parameters	6		8		12		12		14		34						25			
LL	−1528		−1063		−945		−943		−934		−958						−923			
BIC	3104		2189		1984		1980		**1978**		2183						2042			

[a] Estimates from uncorrelated coefficient specification.

[b] Coefficients of ’if know doctor’, ’if test is due’, ’if doctor recommend’ and cost are assumed to be distributed as log-normal, while other coefficients are assumed to be normally distributed. If only cost coefficient is assumed to be log-normal distributed, BIC of T-MIXL is 1984.

[c] Estimates from LC with five classes.

[d] Estimates from MM-MNL with two independent normals.

Bold estimates are statistically significant at 5%.

(with two independent normal vectors of attribute weights but correlated intercepts) is again preferred by BIC (6403). Second is T-MIXL at 6404. Now the preferred T-MIXL model has log-normal coefficients on cost variables *and* doctor recommendation. The order of the other models is the same as before: G-MNL (6465), N-MIXL (6535), S-MNL (6601) and LC (6723).

Interestingly, the structure of heterogeneity is more complex for the general population than the Ashkenazis. The LC model identifies seven segments, compared to only five in the Ashkenazi data.[10] The segments are also very different. In the general population, the largest (22%) rarely chooses to get tested. This replaces segment 3 in the Ashkenazi data (20%), which almost always gets tested. This shift is not surprising *unconditionally*, as we would expect Ashkenazis to care more about getting tested as they are at higher risk. But here it occurs *conditional* on risk. This is consistent with a view that the experimental subjects are behaving as Bayesians—i.e. updating priors on risk with the information given in the experiment. So, even in an experiment, it is not possible to fully control subjects' perceptions of attribute levels.

Table II reports estimates from the mobile phone DCE. The choice is simply whether or not to buy a phone with certain attributes. The structure of heterogeneity here is very simple: the LC model identifies only four segments. The largest (32%) is not very sensitive to any particular attribute (i.e. they exhibit fairly 'random' behavior). Segment 2 (28%) is sensitive to price but not other attributes. Segment 3 (22%) is very sensitive to price. Segment 4 (not reported, 18%) is modestly sensitive to price. Thus consumers are grouped into four levels of price sensitivity, while other attributes are fairly unimportant. This lack of sensitivity to extra cell phone features (beyond the basics all phones have) is consistent with views expressed by industry executives.

Given the simple structure of heterogeneity, it is not surprising that the very parsimonious S-MNL model is preferred here, with a BIC of 8121. There is little to choose between N-MIXL, G-MNL and T-MIXL, which have BIC values of 8190, 8197 and 8204, respectively. MM-MNL and LC lag far behind, with BIC values of 8359 and 8426. Given the simplicity of the data, they are heavily penalized for lack of parsimony.

Recall that both S-MNL and N-MIXL are nested in G-MNL. They are (slightly) preferred by BIC in Table II because the likelihood improvement achieved by G-MNL is too small to justify the extra parameters. Indeed, a researcher using G-MNL should test down to the S-MNL nested case here. So, technically, the preferred G-MNL model is simply S-MNL. This occurs in three datasets. In these cases, we continue (for completeness) to report results of the full G-MNL model (which includes the insignificant variance–covariance matrix parameters Σ).

Notably, we were unable to find a T-MIXL model that improves upon N-MIXL in this dataset. Thus Table II reports the T-MIXL model that assumes only a log-normal price coefficient. Imposing this assumption causes BIC to *deteriorate* from 8190 to 8204. There are three datasets where T-MIXL does not improve upon N-MIXL, and in these cases we report the T-MIXL model that puts a log-normal coefficient only on price. In these cases, the researcher has to decide if it is more important to impose the theoretical sign restriction or to use a better-fitting model.

Table III reports results from pizza delivery service choice experiment A. The two services are generic (labeled A and B), so the model does not contain ASCs. Again the LC model identifies four segments, but the structure of heterogeneity is much more complex than in the mobile phone data, because now the four segments are very different: type 1 cares greatly about ingredient freshness; type 2 exhibits 'random' behavior (i.e. insensitive to observed attributes); type 3 cares greatly about the pizza being hot on delivery; and type 4 (not reported) cares a lot about price.

After some experimentation, we found that the preferred T-MIXL model for the Pizza A dataset has log-normal coefficients on price, freshness and delivery time. This model also has the best BIC

[10] The two extra segments that appear in the general population are #6, who rely heavily on doctor recommendation (7%), and #7, who care a lot about *both* risk and price (7%).

DOI: 10.1002/jae

of 2812. Second best is G-MNL, with a BIC of 2886. Interestingly, if we only put a log-normal coefficient on price, the BIC for T-MIXL is 2907, so G-MNL is preferred. This again illustrates the need for care in specifying the T-MIXL model. The third-best model is MM-MNL at 2919, followed by N-MIXL at 2933. The LC and S-MNL models trail by a wide margin.

Table IV reports results from pizza choice experiment B. It differs from experiment A in that the number of attributes of the pizza (and the delivery service) is increased from eight to 16. Not surprisingly, this increases the number of classes identified by the LC model from four to six. Given this more complex heterogeneity structure, MM-MNL is the preferred model (BIC = 11,527). The next-best model is T-MIXL with log-normal coefficients on price, ingredient freshness, delivery time and delivery charge (BIC = 11,600).[11] Third is G-MNL (11,639). There is then quite a large gap before we come to N-MIXL (12,081), followed by LC (12,118) and then S-MNL (13,372).

The structure of heterogeneity in the Pizza B dataset is quite interesting. Segment 1 identified by the LC model makes up 51% of the population, and shows very modest sensitivity to attributes (i.e. close to random choice behavior). The second segment (14%) cares greatly about price, the third (12%) cares greatly about fresh ingredients, the fourth (10%) cares greatly about crust type, the fifth (9%) wants hot delivery and the sixth (4%) wants vegetarian. Thus we have several small segments that care about different attributes. This is the first dataset we have seen where the structure of heterogeneity is complex enough that MM-MNL supports a three-class model.

Table V reports results from the Papsmear test experiment. Here the LC model identifies five segments. Substantively, it is interesting that type 1s get tested as needed (i.e. test is due and the doctor recommends it), type 2s almost always get tested, and type 3s care about a range of factors (if test is due, doctor characteristics, doctor recommends). Types 4 and 5 (not reported) make up 24% of the population. They are, respectively, either *very* averse or *extremely* averse to male doctors. No type is very concerned about price.

In the Papsmear dataset the G-MNL model is preferred, with a BIC of 1978. The preferred T-MIXL model, which has log-normal coefficients on price, doctor recommends, test due, and know doctor, is second, with a BIC of 1980. Interestingly, however, this model barely improves over N-MIXL, which has a BIC of 1984. Finally, MM-MNL, LC and S-MNL trail considerably.

Supporting information Table A4 reports results for the Holiday A dataset. Participants choose between two generic holiday packages (labeled A and B), so again there is no ASC. Here, the LC model identifies five segments. The first cares a lot about price, and the second about quality accommodation. The third cares modestly about both. The fourth and fifth either like or do not like overseas destinations, respectively. The preferred model is T-MIXL with log-normal coefficients on price and four-star accommodation. G-MNL is a close second (BIC of 5165 vs. 5178). Notably, however, if we assume log-normal only for price, then G-MNL is preferred.

Supporting information Table A5 reports results for the Holiday B data. It differs from A in that the number of attributes is increased from eight to 16. Here, the LC model identifies nine segments—more than in any other dataset.[12] Given this complex heterogeneity structure, MM-MNL is again the preferred model (BIC = 23002) and, as in Pizza B, a mixture of three types is preferred. Second is T-MIXL

[11] By putting log-normal distributions on price, freshness, delivery time and charges, T-MIXL obtains a 481-point BIC improvement over N-MIXL. Note, however, that it required considerable effort to determine that an assumption of log-normality on these four particular coefficients gave a much better fit. In contrast, MM-MNL and G-MNL capture non-normality 'automatically' using normal mixtures. They achieve 554- and 442-point improvements in fit, respectively.

[12] Segment 1 (27%) cares modestly about price, but puts little weight on other attributes. Segment 2 (15%) cares modestly about four-star hotel, price, meals and length of stay. Segment 3 (14%) cares intensely about four-star hotel. Segment 4 (14%) likes overseas travel. Segment 5 cares about price and meals. Segment 6 (7%) wants a beach or swimming pool. Segment 7 (6%) hates overseas travel. Segment 8 (5%) likes personal tours. Segment 9 (4%) cares about price, four-star hotel and length of stay. Too few respondents value cultural activities to form a segment.

DOI: 10.1002/jae

with log-normal coefficients on price, four-star hotel and meal inclusion (23,021).[13] Third is G-MNL (23,264). Next comes N-MIXL (23,519), followed by LC (23,981) and S-MNL (26,224).

Supporting information Table A6 reports results from an experiment where a bank offers a credit and a debit card. Along with 'neither' this gives three alternatives. The structure of heterogeneity is quite simple here. According to the LC model, there are only four segments. The largest (48%) does not like either card (large negative intercepts). Segment 2 (27%) prefers a debit while segment 3 (19%) prefers a credit card. Segment 4 (not reported, 7%) is fairly indifferent between the two. All types dislike interest and fees. Given this simple structure, it is not surprising that S-MNL is preferred by BIC (5707). It is followed by N-MIXL, T-MIXL and G-MNL, with very similar BIC values of 5770, 5776 and 5786, respectively. Then comes MM-MNL (5988) and, finally, LC (6039). Note that here we are unable to find a T-MIXL model that improves on N-MIXL.

Supporting information Table A7 reports results from the Charge Card B experiment. Here a fourth option of a transaction card is added. LC again identifies four segments. The largest (44%) has little desire for any card. Segment 2 (21%) is indifferent among types of card, but cares about interest, fees and access. Segment 3 (19%) prefers a transaction card and cares about interest rates. Segment 4 (16%) is indifferent among types of card, but cares about interest, fees and access. All types dislike interest. Given this simple structure, the S-MNL model is again preferred by BIC (7007). It is followed by N-MIXL, T-MIXL and G-MNL, whose BIC values are again very close (7060, 7074 and 7076, respectively). Then comes MM-MNL (7258) and, finally, LC (7391). Here again we are unable to find a T-MIXL model that improves on N-MIXL.

4.2. Comparing Model Fit across the 10 Datasets

In Table VI we summarize results from the 10 datasets. According to BIC, the MM-MNL model is preferred in four datasets, G-MNL is preferred in one, S-MNL is preferred in three and T-MIXL is preferred in two. It is worth recalling that S-MNL is nested within G-MNL—so a researcher using G-MNL would test down to S-MNL in the three cases where the latter is preferred.

Strikingly, N-MIXL and LC—currently the most popular models of heterogeneity—are never preferred by BIC. The performance of LC is particularly weak as it ranks last or next to last in every dataset. Nevertheless, we found that LC is very useful for gaining an intuition for the structure of heterogeneity. Thus we would advise using LC to gain intuition about consumer types, and G-MNL, MM-MNL or T-MIXL for actual demand prediction.

One clear pattern is that the four datasets where MM-MNL is preferred are those with the most complex heterogeneity structures. That is, these are the datasets where the LC model finds five to nine segments, and these segments care about rather different attributes. In contrast, S-MNL is preferred in three simple datasets (mobile phones, charge card A and B) where there are only four segments that care about similar things (e.g., price, interest)—only to different degrees.

A careful inspection of Table VI reveals that, regardless of which is preferred, the fit of MM-MNL, T-MIXL and G-MNL (or its S-MNL special case) is quite comparable across all 10 datasets. This is a bit hidden, as in four datasets (#3, 6, 9, 10) MM-MNL seems to lose by a wide margin. But N-MIXL, which is a special case of MM-MNL, performs rather well in those datasets. A researcher using the MM-MNL framework would test down to N-MIXL in those cases.

Given that G-MNL, MM-MNL and T-MIXL provide similar fits, the choice among them may be based on other considerations, like ease of use. In the case of T-MIXL, it can be difficult to decide which coefficients are log-normal. For instance, in the Tay–Sachs data eight variables may plausibly

[13] If one only adopts a log-normal price coefficient the improvement over N-MIXL is minor. It is the assumption of a log-normal coefficient on four-star status and meal inclusion that really drives the likelihood improvement. We found a similar situation with Pizza B (where it was ingredient quality and not price where log-normality really matters).

DOI: 10.1002/jae

Table VI. Comparing model fit across datasets

	MNL	S-MNL	N-MIXL	T-MIXL	G-MNL	LC	MM-MNL
Tay–Sachs disease and cystic fibrosis test, Jewish sample (3 ASCs), $T = 16$; $N = 210$							
# parameters	11	18	77	77	79	59	51
LL	−3717	−2815	−2500	−2469	−2479	−2701	−2573
BIC	7523	5777	5626	5563	5601	5882	**5560**
Tay–Sachs disease and cystic fibrosis test, general population (3 ASCs) $T = 16$, $N = 261$							
# parameters	11	18	77	77	27	83	51
LL	−4649	−3226	−2946	−2881	−3120	−3016	−2989
BIC	9390	6601	6535	6404	6465	6723	**6403**
Mobile phone (1 ASC), $T = 8$; $N = 493$							
# parameters	15	17	30	30	32	63	61
LL	−4475	−3990	−3971	−3978	−3966	−3952	−3927
BIC	9074	**8121**	8190	8204	8197	8426	8359
Pizza A (no ASC), $T = 16$; $N = 178$							
# parameters	8	9	16	16	18	35	33
LL	−1657	−1581	−1403	−1342	−1372	−1418	−1328
BIC	3378	3233	2933	**2812**	2886	3115	2919
Holiday A (no ASC), $T = 16$; $N = 331$							
# parameters	8	9	16	16	18	44	33
LL	−3066	−2967	−2553	−2514	−2512	−2502	−2464
BIC	6201	6011	5244	**5165**	5178	5381	5211
Papsmear test (1 ASC), $T = 32$; $N = 79$							
# parameters	6	8	12	12	14	34	25
LL	−1528	−1063	−945	−943	−934	−958	−923
BIC	3104	2189	1984	1980	**1978**	2183	2042
Pizza B (no ASC), $T = 32$; $N = 328$							
# parameters	16	17	32	32	34	101	98
LL	−6747	−6607	−5892	−5652	−5662	−5591	−5310
BIC	13,642	13,372	12,081	11,600	11,639	12,118	**11,527**
Holiday B (no ASC), $T = 32$; $N = 683$							
# parameters	16	17	32	32	34	152	98
LL	−13,478	−13,027	−11,600	−11,351	−11,462	−11,231	−11,012
BIC	27,116	26,224	23,519	23,021	23,264	23,981	**23,002**
Credit card A (2 ASCs), $T = 4$; $N = 827$							
# parameters	17	21	35	35	37	71	69
LL	−3354	−2768	−2743	−2746	−2743	−2732	−2714
BIC	6846	**5706**	5770	5776	5786	6039	5988
Credit card B (3 ASCs), $T = 4$; $N = 827$							
# parameters	18	25	39	39	41	75	74
LL	−4100	−3402	−3372	−3379	−3372	−3391	−3329
BIC	8346	**7007**	7060	7074	7076	7391	7258

have log-normal coefficients, giving $2^8 = 256$ potential models. A researcher must use judgment to narrow this down. Also, we found that imposing log-normal price coefficients caused the fit to deteriorate in three datasets (#3, 9, 10). In five other cases (#2, 4, 5, 6, 7), the best model had constraints on non-obvious sets of coefficients (e.g. Holiday B had a log-normal on meals, while Holiday A did not). This suggests one should use care in choosing a T-MIXL specification.

To use the MM-MNL framework one must estimate multiple models with (i) different numbers of mixture elements and (ii) different assumptions about the covariance parameters. One must then determine which model is preferred. In G-MNL, a researcher must (i) decide whether or not to allow error correlations, and (ii) test against the nested S-MNL and N-MIXL models. But, in contrast to MM-MNL, one need not determine the number of mixture elements.

DOI: 10.1002/jae

Overall, the advantage of G-MNL over MM-MNL and T-MIXL is that one has to estimate fewer models, but the extra complexity of MM-MNL is justified if the heterogeneity structure is very complex. T-MIXL may be preferred if the researcher has strong priors on coefficient signs, so that only a few specifications need be tried. This is because *individual* T-MIXL models can be estimated faster than G-MNL or MM-MNL. Also, T-MIXL allows one to impose the correct sign on the price coefficient, so that WTP distributions will have finite moments.

4.3. Understanding the Behavioral Differences between Models

Knowing a model fits better than others is not all that matters. It is also important to understand why. What aspects of behavior does it capture better? Here we examine behavioral differences among our six models, by looking at how well each model fits key patterns in the data.

First, however, it is useful to examine what each model implies about the distribution of consumer taste heterogeneity. We adopt what Allenby and Rossi (1998) call an ’approximate Bayesian’ approach: a model’s estimated heterogeneity distribution is taken as the prior. Then, posterior means of the person-specific vectors of preference weights are calculated conditional on each person’s observed choices (see Train, 2003, ch. 11, for details).

For the Pizza B dataset, Figure 1 plots the posterior distributions of the person-level price coefficients. Note how the N-MIXL posterior has a distinctly normal shape. As Allenby and Rossi (1998) point out, the normal prior of N-MIXL has a strong tendency to draw in outliers, so N-MIXL has difficulty capturing ’extreme’ consumers who place great weight on price.

In contrast, the posteriors of T-MIXL, G-MNL, MM-MNL and LC depart substantially from normality. They all generate a mass of consumers in the left tail who care intensely about price. They also generate excess kurtosis—i.e. a mass of consumers with price coefficients near zero. But, as noted

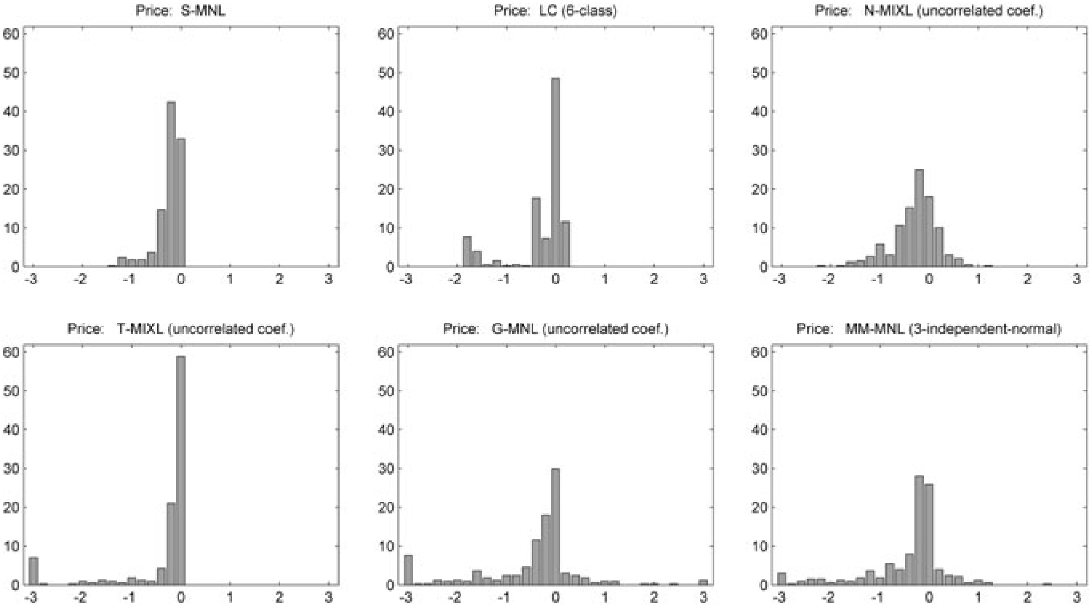

Figure 1. Posterior distribution of individual-level PRICE coefficient from Pizza B dataset.
Note: The first bin includes data between – infinity and −2.9 and the last bin includes data between 2.9 and infinity. For T-MIXL, G-MNL and MM-MNL, their left tails span to −26.8, −15.6 and −4.4, respectively

DOI: 10.1002/jae

by Elrod and Keane (1995), the LC posterior understates heterogeneity, as it is constrained to lie within the convex hull of the $\{\beta_s\}$.

The G-MNL and MM-MNL posteriors for price are quite similar. This is not surprising as the difference between these models is the use of continuous vs. discrete normal mixtures. Both priors are flexible, letting the data have more impact on the shape of the posterior. T-MIXL also captures *both* outliers and a large mass near zero, because this is a feature of the log-normal. The main difference among the three models is that kurtosis is much greater for T-MIXL.

Figure 2 plots the posterior distributions for the ingredient quality coefficient. The story is very similar to Figure 1. G-MNL, MM-MNL and T-MIXL are all able to generate a segment of consumers that puts great positive weight on fresh ingredients. N-MIXL is again unable to capture this, as these outliers are pulled in by the normal prior. LC captures the segment that cares a lot about freshness, but it understates the extent of heterogeneity in the data.

Next, Figure 3 reports selected model predictions of how changes in product attributes affect consumer demand. We start from a baseline where pizza delivery services A and B have identical attributes. In that case people are indifferent between the two, and all models predict that 100% of consumers choose service A exactly 50% of the time. In the experiment, service A improves ingredient quality (i.e. fresh ingredients) while also increasing price by $4.

After the policy change G-MNL predicts that 23% of consumers still have about a 50% chance of choosing each option. Strikingly, 9% of consumers shift to a near 100% chance of choosing the high-quality option A (they put great weight on freshness), while 5% shift to a near 100% chance of choosing the low-price option B (they put great weight on price). (We define ’about 50%’ as 0.475 to 0.525, ’near 100%’ as greater than 0.95 and ’near 0%’ as less than 0.05.)

The predictions of MM-MNL are similar. It predicts that 16% of consumers remain near 50%, while 9% have a near 100% chance of choosing A and 7% have a near 100% chance of choosing B. The T-MIXL predictions are a bit different, in that more consumers remain close to indifferent

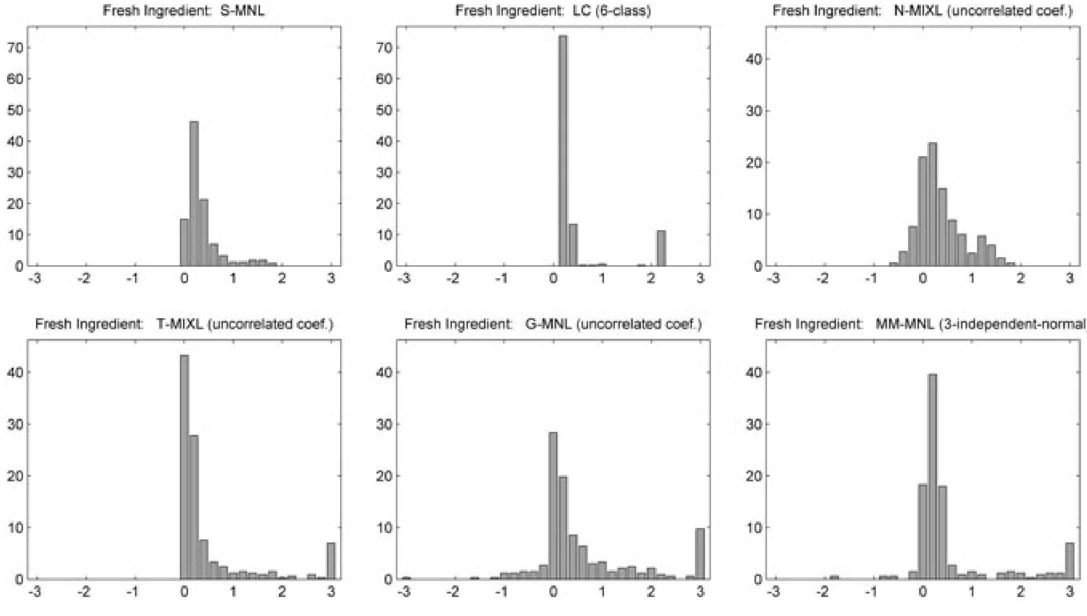

Figure 2. Posterior distribution of individual-level FRESH INGREDIENT coefficient from Pizza B dataset. Note: The first bin includes data between – infinity and −2.9 and the last bin includes data between 2.9 and infinity. The maximum values of the right tails of T-MIXL, G-MNL and MM-MNL are 43.8, 21.5 and 5.2, respectively. MM-MNL also has a small mode at 3

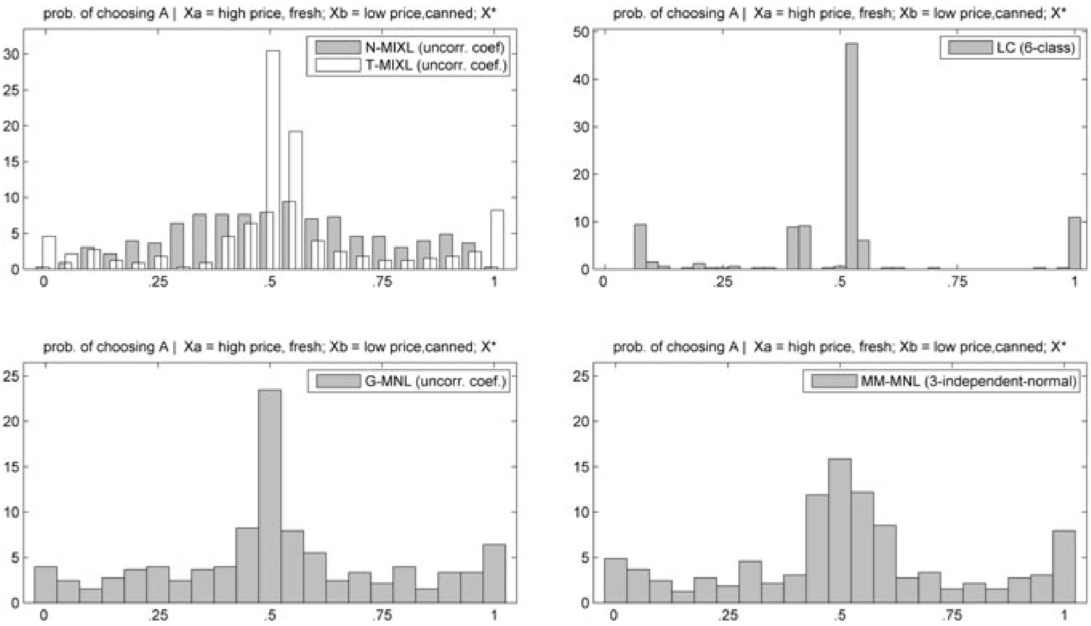

Figure 3. Predicted distribution of probability of choosing firm A from MIXL, LC, G-MNL and MM-MNL models when firm A improves ingredient quality and increases price $4

(30%). This is not surprising, given the frequency of price and ingredient coefficients near zero in this model (see Figures 1 and 2). But T-MIXL gives similar predictions for the fractions of consumers who shift to a 100% chance of choosing either A or B (9.5% and 6%, respectively).

As we would expect, given the coefficient distributions in Figures 1 and 2, N-MIXL predicts fewer people stay indifferent, but also that fewer people have extreme reactions. Specifically, it predicts that only 8% of consumers stay at roughly a 50% chance of choosing A, while almost no consumers have their choice probabilities move close to 100% or 0%.

To summarize, we see that G-MNL, MM-MNL and T-MIXL can generate subjects who exhibit either (i) 'extreme' reactions to changes in price or quality or (ii) near indifference to such changes. But N-MIXL seems unable to generate such patterns. This gives us a clue as to why these models may fit better than N-MIXL. Given this, we turn to look at patterns in the data.

In the Pizza B data, 24 of the 328 subjects choose the fresh ingredient pizza on all 32 choice occasions regardless of other attribute settings, while 27 always choose the cheaper pizza. Thus 51 subjects exhibit extreme (or lexicographic) preferences for price or quality. For these 51 subjects, T-MIXL, G-MNL and MM-MNL have BIC advantages over N-MIXL of 206, 135 and 158 points, respectively. There are 62 additional subjects who are lexicographic regarding some other attribute (e.g. hot delivery, vegetarian, crust type). For these subjects, T-MIXL, G-MNL and MM-MNL have BIC advantages over N-MIXL of 160, 161 and 371 points, respectively.

Thus, among the 113 subjects who exhibit lexicographic preferences for some attribute, T-MIXL, G-MNL and MM-MNL have large BIC advantages over N-MIXL of 366, 296 and 529 points. Recall (Table IV) that the overall BIC advantages of these models over N-MIXL are 481, 442 and 554 points, respectively. Thus the lexicographic subjects account for 76%, 67% and 95% of these overall BIC gains. This shows that better ability to fit 'extreme' or lexicographic behavior is a key reason that T-MIXL, G-MNL and MM-MNL fit better than N-MIXL.

Next, if a consumer chooses randomly between options A and B, the mean attribute differences between their chosen and non-chosen options will be ’close’ to zero (except for variation due to sampling). We define precisely how we measure ’close’ in Appendix B (supporting information). Given this definition, we classify 31 consumers as exhibiting random behavior. For these subjects, T-MIXL, G-MNL and MM-MNL have BIC advantages over N-MIXL of 52, 107 and 58 points.

Combining results for the 113 lexicographic and 31 random subjects, the BIC advantages of T-MIXL, G-MNL and MM-MNL over N-MIXL are 418, 403 and 587 points. Thus the lexicographic and random consumers together account for 87%, 91% and 106% of the overall BIC gains of these three models over N-MIXL. Note that these consumers account for 144/328 = 44% of the subjects. For the remaining 56% of subjects the fit of all four models is quite similar.

In summary, T-MIXL, G-MNL and MM-MNL fit better than N-MIXL because: (i) they are better at capturing ’extreme’ or lexicographic behavior; and (ii) they are better at capturing ’random’ choice behavior that is only slightly influenced by observed attributes. These advantages arise from the more flexible heterogeneity distributions of these models, which allow them to generate attribute weight distributions with both more outliers and more mass near zero.

Understanding differences among G-MNL, MM-MNL and T-MIXL is more difficult, as these models predict rather similar behavior. As we have noted, they can all capture behavior that is nearly lexicographic or nearly random, so we sought more subtle patterns to distinguish them.

A clue is provided by the fact that MM-MNL is preferred in four datasets with very complex heterogeneity. For example, in Pizza B there are five ’major’ attributes important enough for LC to devote segments to people who value them highly (i.e. price, freshness, crust, hot delivery, vegetarian). However, examination of the posterior distribution of attribute weights revealed additional ’minor’ attributes that small but non-trivial segments of consumers also value highly.

For instance, the top panel of Figure 4 shows posteriors for the coefficient on wood-fire cooking. Those for G-MNL and T-MIXL are concentrated near zero, but for MM-MNL 29% of the mass of the cooking method coefficient is in the 0.30–0.50 range.[14] Thus MM-MNL implies that a non-negligible segment of consumers has a modest preference for wood-fire cooking.

The bottom panel of Figure 4 examines how demand predictions differ for the three models. In the experiment firms A and B are identical, *except* that A offers wood-fire cooked pizza. Both G-MNL and T-MIXL predict that the large majority of consumers are nearly indifferent between A and B. But MM-MNL predicts that for about 30% of consumers the probability of choosing A is about 65%.

Thus G-MNL, MM-MNL and T-MIXL predict similar responses to ’major’ attributes, but rather different responses to ’minor’ attributes. This may explain why/when the models fit differently. To further explore this issue, we turn to the actual data, and classify consumers into types based on how strongly they prefer certain attributes. Details of the classification procedure are provided in Appendix B (supporting information). Here we give an overview.

Consider again the Pizza B dataset. Some consumers exhibit a strong preference for *one* attribute. We subdivide them based on whether this is a ’major’ attribute (price, quality, crust, hot, vegetarian) or a ’minor’ attribute. These groups are reported in the first two rows of Table VII.

We subdivide these two groups into those with (i) little interest in all other attributes, (ii) modest preference for a few other attributes, or (iii) modest preference for many other attributes. These subgroups are given in the three panels of Table VII. For example, in row 1, panel 1, we see that 39 people have a strong taste for one major attribute and care little about other attributes.

Similarly, some people have strong tastes for *two* attributes. We subdivide them based on whether these are major, minor or both. These groups are reported in rows 3–5. Next, in rows 6–8, we have

[14] In contrast, for G-MNL, most of the mass is near zero, and only 8% is in the 0.30–0.50 range.

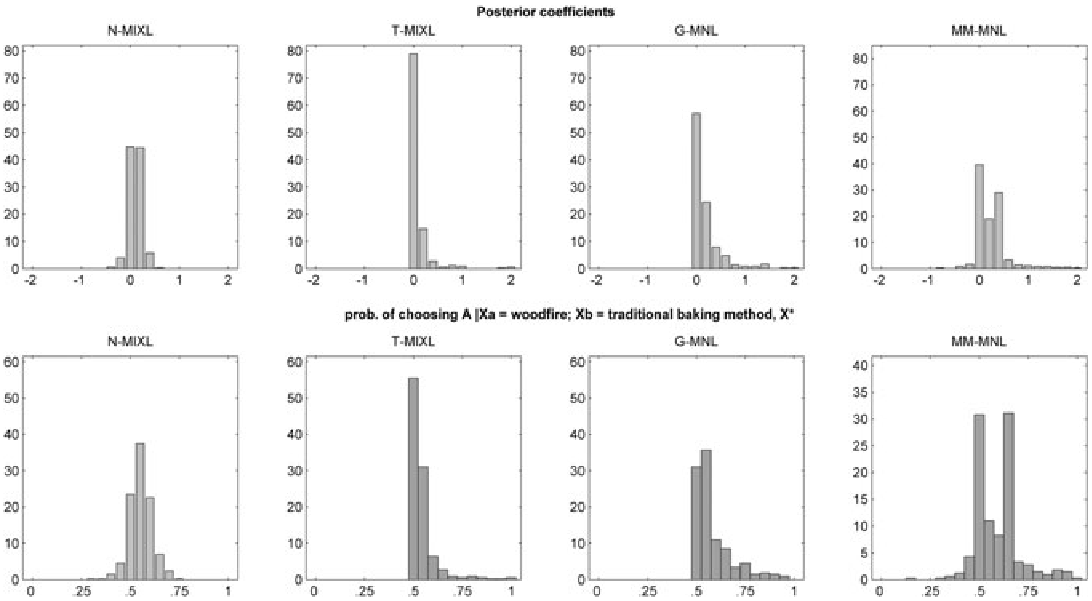

Figure 4. Posterior distribution of individual-level BAKING METHOD coefficient and predicted probability of choosing firm A from MIXL, G-MNL and MM-MNL models when firm A uses wood-fire baking method but does not increase price. The coefficient of baking method in T-MIXL is assumed to be normally distributed

consumers with strong tastes for *three* or more attributes. Finally, in rows 9–10, we have ’non-extreme’ people who do not exhibit a strong taste for any attribute. As above, these groups are subdivided into three subgroups, based on their (modest) tastes for other attributes.

Table VII lists the number of people in each group, and the BIC gain for MM-MNL over G-MNL and T-MIXL in each case. We highlight in bold cases where MM-MNL has a large advantage and italic cases where G-MNL or T-MIXL has a large advantage. For the dataset as a whole, MM-MNL has a BIC advantage of 112 points over G-MNL and 73 points over T-MIXL.

Strikingly, as we see in row 2, panel 1, MM-MNL achieves an advantage of 186 points over G-MNL on just 17 consumers with an extreme taste for a *minor* attribute. In panel 1, rows 4 and 5, we see that MM-MNL has a BIC advantage of 34 + 20 =54 points on just 11 consumers with a strong taste for one or two minor attributes. Finally, in row 5, panel 2, MM-MNL achieves a BIC advantage of 31 points on just three consumers with a strong taste for two minor attributes.

There are also groups where G-MNL is favored over MM-MNL. In particular, in rows 9 and 10, we see G-MNL has a BIC advantage over MM-MNL of 53 + 75 = 128 points in fitting behavior of 61 ’non-extreme’ consumers with modest preference weights on multiple attributes.

As is clear from Table VII, the patterns for T-MIXL are rather similar; that is, it tends to be superior or inferior to MM-MNL for the same subgroups. The main difference is that gaps between T-MIXL and MM-MNL are usually smaller (in both positive and negative directions).

We performed the same type of analysis on other datasets and came to the same general conclusion: MM-MNL fits better than G-MNL for consumers who have strong tastes for ’minor’ attributes not valued highly by the majority. On the other hand, G-MNL outperforms MM-MNL for consumers with (i) moderate tastes for multiple attributes and (ii) fairly random-choice behavior. T-MIXL shares the advantages/disadvantages of G-MNL. Indeed, across all 10 datasets, if G-MNL is preferred to MM-MNL then so is T-MIXL, and vice versa.

This analysis is consistent with the results in Section 4.1. There we found MM-MNL was preferred to G-MNL and T-MIXL in the four datasets with the most complex heterogeneity patterns. In those

Table VII. BIC gain of MM-MNL over G-MNL and T-MIXL from different types of observed choice pattern from Pizza B dataset

	Attribute preferences		Totally indifferent to ’other’ attributes			Also like some ’other’ attributes			Also like many ’other’ attributes	
			BIC gain of MM-MNL			BIC gain of MM-MNL			BIC gain of MM-MNL	
			Over	Over	Freq.	Over	Over	Freq.	Over	Over
		Freq.	G-MNL	T-MIXL		G-MNL	T-MIXL		G-MNL	T-MIXL
Extremely prefer one attribute										
(1)	One of major attributes (price, fresh ingredient, crust, hot or vegetarian)	39	43	73	42	−5	−67	17	−19	−30
(2)	One of minor attributes (other attributes)	17	**186**	**111**	12	32	17	16	10	−4
Extremely prefer 2 attributes										
(3)	both in major attributes	9	*−36*	8	14	−14	14	5	*−21*	−9
(4)	1 in major and 1 in minor attributes	5	**34**	−7	21	24	−2	10	−6	−11
(5)	both in minor attributes	6	**20**	**37**	3	**31**	**22**	7	17	−12
Extremely prefer 3 or more attributes										
(6)	2 in major attributes				12	−25	7	2	0	1
(7)	1 in major and 1 in minor attributes				7	8	4	3	−5	2
(8)	2 in minor attributes				1	**13**	**3**			
Non-extreme										
(9)	like at least 3 of major attributes	2	*−12*	*−8*	17	*−53*	−27	6	*−27*	*−20*
(10)	like 2 of major attributes	1	*−5*	−1	44	−75	−13	10	−7	−15

Note: The bold cells are cases where MM-MNL has a BIC advantage over G-MNL or T-MIXL for at least 3 points per person on average. The italic cells are cases where G-MNL or T-MIXL has a BIC advantage over MM-MNL for at least 3 points per person on average.

DOI: 10.1002/jae

datasets there exist fairly small subsets of consumers with strong preferences for several ’minor’ attributes that the majority of consumers are relatively uninterested in.

5. CONCLUSION

Here we have compared the empirical performance of six alternative models of consumer taste heterogeneity. Arguably, the most popular model of taste heterogeneity in use today is the mixed logit or MIXL model, where the coefficient vector of the basic logit model is allowed to be heterogeneous in the population. Many applications of MIXL posit a multivariate normal coefficient vector, leading to what we call ’N-MIXL’ (i.e. logit with normal mixing).

In a recent paper, Fiebig *et al*. (2010) introduced a new choice model called ’generalized multinomial logit’ or G-MNL. It generalizes N-MIXL by allowing for heterogeneity in the scale of the logit error terms. A key special case of G-MNL is the scale heterogeneity model (S-MNL), where *only* scale heterogeneity is present. An equivalent representation of G-MNL is that the scale of the logit errors is fixed, but the multivariate normal coefficient vector of N-MIXL is scaled by a continuously distributed scalar. This gives a *continuous* mixture of scaled normals.

There is also a rapidly growing literature in statistics and econometrics using *discrete* mixtures-of-normals models. This is appealing because it can approximate any distribution arbitrarily well (see Ferguson, 1973). We refer to a MIXL model with a discrete mixture-of-normals heterogeneity distribution as the ’mixed-mixed logit’ or ’MM-MNL’ model.

A possible limitation of N-MIXL, G-MNL and MM-MNL is that they do not impose theoretical restrictions on coefficients. For example, they are all based on normals, so they must put some mass on price coefficients greater than or equal to zero. Of course, depending on the parameters of the normal, this mass may be quite small, but it still creates a problem if moments of WTP are of interest.[15] We call models that place theoretical restrictions on price and other vertical attributes ’T-MIXL’ models. Of course, this can be done in an infinite number of ways, so for tractability we focus on log-normal distributions (the most common approach in the literature).

Finally, we also consider latent class (LC) models. These assume that consumers fall into a discrete set of types, each with its own coefficient vector, and they have also been very popular.

We compare the performance of these six models of heterogeneity (MM-MNL, G-MNL, S-MNL, T-MIXL, N-MIXL, LC) on data from 10 stated preference discrete choice experiments previously analyzed in Fiebig *et al*. (2010). Our main results can be summarized as follows:

1. MM-MNL, T-MIXL and G-MNL (or its S-MNL special case) usually dominate N-MIXL and LC in terms of fit (as measured by BIC). The poor performance of N-MIXL and LC is notable, as these are the most popular models in use today.
2. N-MIXL and S-MNL are only preferred to MM-MNL, T-MIXL and G-MNL in three datasets with very simple heterogeneity patterns.
3. The superior fit of MM-MNL, G-MNL and T-MIXL arises because all three can capture *simultaneously* the existence of (i) segments of consumers who exhibit approximately lexicographic behavior with respect to one or two attributes, and (ii) a segment that exhibits ’random’ behavior in the sense that choice is little influenced by observed attributes. N-MIXL has difficulty capturing both patterns of behavior.

[15] Of course, if the price coefficient is zero or the wrong sign it implies infinite WTP for a product (or an attribute). But existence of a subset of consumers with zero or positive price coefficients may not invalidate a demand model. For the inexpensive products considered here, and the modest price variation in the data, it is not surprising if some consumers behave as if indifferent to price (or use price as a quality signal). But it would be a mistake to conclude these same consumers would be willing to pay far higher prices. Thus models that do not impose negative price coefficients may be useful for predicting demand, *provided* one does not extrapolate to very high prices that are out of the range of the data. The ’infinite WTP’ problem arises from extrapolating to behavior at very high prices.

4. MM-MNL, T-MIXL and G-MNL often give similar fits, so it can be hard to choose among them: the preferred model depends on context, and researchers should experiment with multiple models. Ease of use may also be an important consideration.
5. A drawback of T-MIXL is there are many possible models, and it can be laborious to determine which combination of attribute coefficients should be constrained (so as to maximize fit). On the other hand, each individual T-MIXL model is fast to estimate. Thus T-MIXL may be preferred if one has strong priors on the sign constraints.
6. MM-MNL can also be hard to use, because one must test for the number of types, and parameters proliferate quickly as types are added. But it allows for the most flexible heterogeneity distribution of the models we considered, so it may be preferred in datasets where the structure of heterogeneity is very complex (to justify the extra parameters).
7. G-MNL has the advantage of being more ’automated’ than MM-MNL or T-MIXL. That is, a single estimated scale parameter determines how the coefficients depart from normality. One does not need to choose a number of types as in MM-MNL, or try alternative distributional assumptions on different coefficients as in T-MIXL. However, G-MNL’s heterogeneity distribution is not as flexible as that of MM-MNL.
8. The number of segments identified by LC, and how sharply they differ, is a good preliminary diagnostic to use to determine the complexity of heterogeneity in a dataset.

While G-MNL, MM-MNL and T-MIXL provide similar fits, there are subtle differences among them. In particular, MM-MNL outperforms G-MNL and T-MIXL when there are *several* attributes that different segments of consumers value highly. In contrast, G-MNL and T-MIXL give a better fit to consumers who put modest weight on several attributes. Aside from these modest differences, these three models predict similar behavioral patterns for most consumers.

We found that G-MNL, MM-MNL and T-MIXL outperform N-MIXL because they can capture, to a good *approximation*, a wide range of behavioral types, from lexicographic/non-compensatory at one extreme to ’random’ choice at the other. Recently, a number of authors have developed models where consumer types *explicitly* follow a variety of compensatory, non-compensatory or random choice rules. For instance, Swait (2001) extends MIXL to include (fuzzy) attribute cut-offs that options must satisfy to be considered, and shows that it provides a much better fit to rental car choices than MIXL.[16] Gilbride and Allenby (2004, 2006) assume that consumers may use a variety of conjunctive or compensatory rules. They find that this fits data on camera and movie rental choices better than MIXL.[17] These results suggest that a useful avenue for future research is to compare the performance of G-MNL, T-MIXL and MM-MNL to such mixed compensatory/non-compensatory models. Note, however, that these models are relatively difficult to estimate.[18] Furthermore, as we have seen, G-MNL, MM-MNL and T-MIXL can generate behavior that looks close to lexicographic. Thus it is not clear that more complex models are needed to capture non-compensatory behavior.

[16] The comparison is not ’fair’ because Swait collected data on attribute cut-offs and used it to help fit the new model. So it uses more information than MIXL. Swait (2001) advocated the collection of such data. Thus his result has a similar flavour to Harris and Keane (1999), who developed an ’extended MIXL’ model that uses data on consumers’ stated preference weights. They found that this model fit health plan choice data vastly better than MIXL.

[17] See also Swait (2009), Arana *et al*. (2008) and von Haefen *et al*. (2005).

[18] As Gilbride and Allenby (2004) note, non-compensatory rules lead to non-differentiable likelihoods; i.e. changing an attribute cut-off can cause the probability of an observed choice to jump discretely from zero to a positive value. They develop an MCMC algorithm that handles this problem by treating cut-offs as latent variables. Drawing them, however, requires a complex Metropolis–Hastings step. And in many models the behavior of an agent following a non-compensatory rule is deterministic. Thus fitting non-compensatory models can require sophisticated non-statistical methods, such as dynamic programming algorithms based on Greedoid languages. See Yee *et al*. (2007) or Kohli and Jedidi (2007).

DOI: 10.1002/jae

ACKNOWLEDGEMENTS

This project was supported by Australian Research Council grants FF0561843 and FL1100100247. Keane was also a visiting research professsor at Arizona State University during much of this project. We thank the editor and three referees for helpful comments.

REFERENCES

Adamowicz W. 2004. What's it worth? An examination of historical trends and future directions in environmental valuation. *Australian Journal of Agricultural and Resource Economics* **48**: 419–443.

Allenby G, Rossi P. 1998. Marketing models of consumer heterogeneity. *Journal of Econometrics* **89**: 57–78.

Arana JE, Leon CL, Hanemann WM. 2008. Emotions and decision rules in discrete choice experiments for valuing health care programmes for the elderly. *Journal of Health Economics* **27**: 753–769.

Arcidiacono P, Jones JB. 2003. Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica* **71**: 933–946.

Bajari P, Fox JT, Ryan S. 2007. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *American Economic Review* **97**: 459–463.

Bennett J, Blamey R. 2001. *The Choice Modelling Approach to Environmental Valuation*. Edward Elgar: Cheltenham, UK.

Burda M, Harding M, Hausman J. 2008. A Bayesian mixed logit–probit model for multinomial choice. *Journal of Econometrics* **147**: 232–246.

Carson RT, Hanemann WM. 2005. Contingent valuation. In *Handbook of Environmental Economics*, Vol. **2**, Mäler K-G, Vincent JR (eds). Elsevier Science: Amsterdam; 821–936.

Carson RT, Louviere JJ, Anderson DA, Arabie P, Bunch DS, Hencher DA, Johnson RM, Kuhfeld WF, Steinberg D, Swait J, Timmermans H, Wiley JB. 1994. Experimental analysis of choice. *Marketing Letters* **5**: 351–361.

Elrod, T, Keane MP. 1995. A factor analytic probit model for representing the market structure in panel data. *Journal of Marketing Research* **32**: 1–16.

Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**: 209–230.

Fiebig D, Keane M, Louviere J, Wasi N. 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science* **29**: 393–421.

Geweke J, Keane M. 1999. Mixture of normals probit models. In *Analysis of Panels and Limited Dependent Variable Models*, Hsiao C, Lahiri K, Lee L-F, Pesaran H (eds). Cambridge University Press: Cambridge, UK; 49–78.

Geweke J, Keane M. 2001. Computationally intensive methods for integration in econometrics. In *Handbook of Econometrics*, Vol. **5**, Heckman JJ, Leamer EE (eds). Elsevier Science: Amsterdam; 3463–3568.

Geweke J, Keane M. 2007. Smoothly mixing regressions. *Journal of Econometrics* **138**: 291–311.

Gilbride TJ, Allenby GM. 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science* **23**: 391–406.

Gilbride TJ, Allenby GM. 2006. Estimating heterogeneous EBA and economic screening rule choice models. *Marketing Science* **25**: 494–509.

Guttmann R, Castle R, Fiebig DG. 2009. Use of discrete choice experiments in health economics: an update of the literature. CHERE working paper 2009/2, Centre for Health Economics Research and Evaluation, Sydney.

Harris K, Keane M. 1999. A model of health plan choice: inferring preferences and perceptions from a combination of revealed preference and attitudinal data. *Journal of Econometrics* **89**: 131–157.

Hensher DA. 1994. Stated preference analysis of travel choices: the state of practice. *Transportation* **21**: 107–133.

Hensher DA, Greene WH. 2003. The mixed logit model: the state of practice. *Transportation* **30**: 133–176.

Kamakura W, Russell G. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **25**: 379–390.

Kanninen BJ. 2007. *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Approach to Theory and Practice*. Springer: Berlin.

Kohli R, Jedidi K. 2007. Representation and inference of lexicographic preference models and their variants. *Marketing Science* **26**: 380–399.

Louviere J. 1994. Conjoint analysis. In *Handbook of Marketing Research*, Bagozzi R (ed.). Blackwell: Oxford; 223–259.

Louviere J, Street D. 2000. Stated-preference methods. In *Handbook of Transport Modeling*, Hensher DA, Button KJ (eds). Elsevier Science: Amsterdam; 131–143.

DOI: 10.1002/jae

Louviere J, Meyer R, Bunch D, Carson R, Dellaert B, Hanemann WM, Hensher D, Irwin J. 1999. Combining sources of preference data for modelling complex decision processes. *Marketing Letters* **10**: 205–217.

Louviere J, Hensher D, Swait J. 2000. *Stated Choice Methods: Analysis and Application*. Cambridge University Press: New York.

Louviere J, Carson R, Ainslie A, Cameron T, DeShazo JR, Hensher D, Kohn R, Marley T, Street D. 2002. Dissecting the random component of utility. *Marketing Letters* **13**: 177–193.

McFadden D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, Zarembka P (ed.). Academic Press: New York; 105–142.

McFadden D. 1978. Modeling the choice of residential location. In *Spatial Interaction Theory and Planning Models*, Karlqvist A, Lundqvist L, Snickars F, Weibull J (eds). North-Holland: Amsterdam; 75–96.

McFadden D, Train K. 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* **15**: 447–470.

Ortúzar JD, Rizzi LI. 2003. Valuation case studies. In *Handbook of Transport and the Environment*, Hensher DA, Button KJ (eds). Pergamon: Amsterdam; 391–409.

Rao VR. 2008. Developments in conjoint analysis. In *Handbook of Marketing Decision Models*, Weirenga B (ed.). Springer: Berlin; 28–53.

Rossi P, Allenby G, McCulloch R. 2005. *Bayesian Statistics and Marketing*. Wiley: Hoboken, NJ.

Ryan M, Gerard K, Amaya-Amaya M. 2007. *Using Discrete Choice Experiments to Value Health and Health Care*. Springer: Dordrecht.

Small K, Winston C, Yan J. 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* **73**: 1367–1382,

Small K, Winston C, Yan J. 2006. Differentiated road pricing, express lanes, and carpools: exploiting heterogeneous preferences in policy design. *Brookings-Wharton Papers on Urban Affairs* 53–96.

Swait J. 2001. A non compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B* **35**: 903–928.

Swait J. 2009. Choice models based on mixed discrete/continuous PDFs. *Transportation Research Part B* **43**: 766–783.

Thurstone L. 1927. A law of comparative judgment. *Psychological Review* **34**: 273–286.

Train K. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press: Cambridge, UK.

Train K. 2008. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modeling* **1**: 40–69.

Von Haefen RH, Massey DM, Adamowicz WL. 2005. Serial nonparticipation in repeated discrete choice models. *American Journal of Agricultural Economics* **87**: 1061–1076.

Whitehead JC, Pattanayak SK, Van Houtven GL, Gelso BR. 2008. Combining revealed and stated preference data to estimate the nonmarket value of ecological services: an assessment of the state of the science. *Journal of Economic Surveys* **22**: 872–908.

Yee M, Dahan E, Hauser J, Orlin J. 2007. Greedoid-based non-compensatory two-stage consideration-then-choice inference. *Marketing Science* **26**: 532–549.

DOI: 10.1002/jae