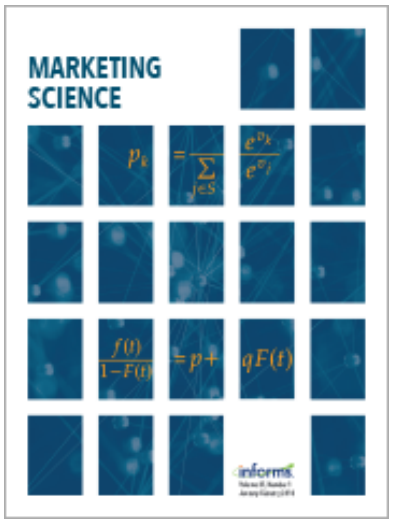


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity

Denzil G. Fiebig, Michael P. Keane, Jordan Louviere, Nada Wasi,

To cite this article:
Denzil G. Fiebig, Michael P. Keane, Jordan Louviere, Nada Wasi, (2010) The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. Marketing Science 29(3):393-421. https://doi.org/10.1287/mksc.1090.0508

MARKETING SCIENCE
Vol. 29, No. 3, May–June 2010, pp. 393–421
ISSN 0732-2399 | EISSN 1526-548X | 10 | 2903 | 0393

informs

DOI 10.1287/mksc.1090.0508

The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity

Denzil G. Fiebig
School of Economics, University of New South Wales, Sydney, New South Wales 2052, Australia, d.fiebig@unsw.edu.au

Michael P. Keane
University of Technology Sydney, Sydney, New South Wales 2007; and Arizona State University, Tempe, Arizona 85287, michael.keane@uts.edu.au

Jordan Louviere
School of Marketing, Centre for the Study of Choice, University of Technology Sydney, Sydney, New South Wales 2007, Australia, jordan.louviere@uts.edu.au

Nada Wasi
School of Finance and Economics, Centre for the Study of Choice, University of Technology Sydney, Sydney, New South Wales 2007, Australia, nada.wasi@uts.edu.au

The mixed or heterogeneous multinomial logit (MIXL) model has become popular in a number of fields, especially marketing, health economics, and industrial organization. In most applications of the model, the vector of consumer utility weights on product attributes is assumed to have a multivariate normal (MVN) distribution in the population. Thus, some consumers care more about some attributes than others, and the IIA property of multinomial logit (MNL) is avoided (i.e., segments of consumers will tend to switch among the subset of brands that possess their most valued attributes). The MIXL model is also appealing because it is relatively easy to estimate. Recently, however, some researchers have argued that the MVN is a poor choice for modelling taste heterogeneity. They argue that much of the heterogeneity in attribute weights is accounted for by a pure scale effect (i.e., across consumers, all attribute weights are scaled up or down in tandem). This implies that choice behaviour is simply more random for some consumers than others (i.e., holding attribute coefficients fixed, the scale of their error term is greater). This leads to a "scale heterogeneity" MNL model (S-MNL). Here, we develop a generalized multinomial logit model (G-MNL) that nests S-MNL and MIXL. By estimating the S-MNL, MIXL, and G-MNL models on 10 data sets, we provide evidence on their relative performance. We find that models that account for scale heterogeneity (i.e., G-MNL or S-MNL) are preferred to MIXL by the Bayes and consistent Akaike information criteria in all 10 data sets. Accounting for scale heterogeneity enables one to account for "extreme" consumers who exhibit nearly lexicographic preferences, as well as consumers who exhibit very "random" behaviour (in a sense we formalize below).

Key words: choice models; mixture models; consumer heterogeneity; choice experiments
History: Received: December 5, 2007; accepted: March 27, 2009; processed by Peter Fader. Published online in *Articles in Advance* July 23, 2009.

1. Introduction

It is well known that consumer choice behaviour exhibits substantial heterogeneity. In choice modelling, adequate modelling of heterogeneity is important for many reasons. Most obviously, estimates of own- and cross-price elasticities of demand may be severely biased if one does not properly account for taste heterogeneity. More subtle and interesting, perhaps, are issues that arise with respect to new product development (NPD), product positioning and advertising, optimal price discrimination strategies, the development of menus of product offerings, and considerations of product image or brand equity, or both.

For example, in NPD, the estimation of only average preferences, as in a simple multinomial logit model—or misspecification of the taste distribution more generally—may lead one to miss that a product with particular attributes would have great appeal for a subset of the population. Similarly, welfare analysis requires correct modelling of taste distributions. Failure to understand the nature of taste heterogeneity may result in failure to optimally target advertising that stresses certain product features to groups that favour those features. Also, in many instances, one cares at least as much about the composition of buyers by type as about market share (e.g., any insurance or usage fee-based product where profits/revenues

depend on subsequent usage, not just on purchase). Many more examples could be provided.

For over 25 years there has been a major research program in marketing on alternative ways to model consumer heterogeneity. As Keane (1997a, b) discusses, the traditional multinomial logit (MNL) of McFadden (1974) and multinomial probit (MNP) of Thurstone (1927) have an asymmetric heterogeneity structure, because they can be motivated by assuming consumers have heterogeneous tastes for unobserved (or unmeasured/intangible) attributes of products but common tastes for observed attributes. Much recent work focuses on extending these models to also allow for heterogeneous tastes over observed attributes as well.

For example, the mixed or heterogeneous logit (MIXL) model is currently quite popular (see, e.g., Ben-Akiva et al. 1997, McFadden and Train 2000, Dubé et al. 2002). MIXL extends MNL to allow random coefficients on observed attributes while continuing to assume the idiosyncratic error is independent and identically distributed (i.i.d.) extreme value. A researcher has great latitude in specifying distributions for the attribute coefficients, but the multivariate normal is used in most applications. (An exception is the price coefficient, which is often modelled as log-normal to impose the constraint that it be negative.) Indeed, it is common to call MNL with normal heterogeneity the mixed logit (see, e.g., Dubé et al. 2002, p. 210).[1]

Of course, one can also estimate multinomial probit (MNP) models with normally distributed attribute weights using the Gewecke Hajivassiliou Keane (GHK) simulator to evaluate the choice probabilities (see Keane 1994, 1997b). However, the popularity of the MIXL stems from its greater ease of use (i.e., GHK is harder to program, and MIXL procedures are now available in standard estimation software packages). Thus, the use of MNP has been mostly limited to more sophisticated academic users, whereas various logit models are widely used by practitioners.

One may also specify a discrete distribution for heterogeneity in either the MNL or MNP. This leads to what is known as the "latent class" (LC) model (see, e.g., Kamakura and Russell 1989). Most applications of LC have used MNL as the base model, again based on ease of use. LC models typically generate a few discrete types of consumers. Part of the appeal of this approach is that one can "name" the types (e.g., couch potatoes, trend setters) leading to easier interpretation of market segments (see, e.g., Wedel and Kamakura 1998). On the other hand, work by Elrod and Keane (1995) and Allenby and Rossi (1998) suggests that latent class models understate the extent of heterogeneity in choice data.

Other ways of capturing heterogeneity have been proposed, such as Harris and Keane (1999), who extended MIXL to allow the means of the random coefficients to depend on observed characteristics of consumers—in particular, attitudinal questions about how much they value each attribute. This led to dramatic improvements in model fit (i.e., doubling of pseudo R-squared). In general, however, choice modellers have favored models that rely largely or exclusively on unobserved heterogeneity, largely abandoning attempts to explain heterogeneous tastes using observables. To be fair, this is partly due to the rather limited set of consumer characteristics recorded in most data sets used in choice modelling.

Recently, Louviere et al. (1999, 2002, 2008b), Louviere and Eagle (2006), and Louviere and Meyer (2007) have argued that the normal mixing distribution commonly used in the MIXL model is seriously misspecified. They argue that much of the taste heterogeneity in most choice contexts can be better described as "scale" heterogeneity—meaning that for some consumers, the scale of the idiosyncratic error term is greater than for others. However, the scale or standard deviation of the idiosyncratic error is not identified in discrete choice data—a problem that is typically resolved by normalizing it to a constant. Thus, the statement that all heterogeneity is in the scale of the error term is observationally equivalent to the statement that heterogeneity takes the form of the vector of utility weights being scaled up or down proportionately as one "looks" across consumers.

It is important to note that the scale heterogeneity model is not nested within the heterogeneous logit with normal mixing. It might appear that the scale model is a limiting case of MIXL in which the attribute weights are perfectly correlated. However, the scale parameter must be positive for all consumers, so although attribute weights may vary in the population, for all consumers they must have the same sign. A normal mixture model with perfectly correlated errors does not impose this constraint. What is clear is that MIXL with independent normal mixing is likely to be a poor approximation to the data-generating process if scale heterogeneity is important. MIXL with correlated random coefficients may or may not provide a better approximation—an empirical question we address below.

Consider then the more general case with both scale heterogeneity and normally distributed taste heterogeneity (independent of the variation induced by scale). In that case, MIXL with normal mixing is clearly misspecified. For example, suppose the utility weight on attribute k, $k = 1, \ldots, K$, for person n, $n = 1, \ldots, N$,

[1] Even when other distributions have been considered, computational problems have often led researchers to revert back to a normality assumption (see, e.g., Bartels et al. 2006, Small et al. 2005).

is given by $\beta_{nk} = \sigma_n \beta_k + \varepsilon_{nk}$, where β_k is the population mean utility weight on attribute k; σ_n is the person n-specific scaling parameter, which for illustration we assume is log-normal; and ε_{nk} is what we will call "residual" heterogeneity not explained by scale heterogeneity. Assume ε_{nk} is distributed normally. Then, writing $\beta_{nk} = \beta_k + \upsilon_{nk}$ as in the conventional MIXL model, the error term $\upsilon_{nk} = \beta_k(\sigma_n - 1) + \varepsilon_{nk}$ is a complex mixture of normal and log-normal errors—the nature of the mixing depending on the unknown parameter vector β_k. Thus, the normal mixing model is misspecified.[2]

There are a number of possible responses to this problem. Some argue for estimation of individual-level models, which circumvent the need to specify a heterogeneity distribution (see Louviere et al. 2008b, Louviere and Eagle 2006, Louviere and Meyer 2007). However, revealed preference data rarely provide enough observations per individual to make this approach feasible. However, as shown by Louviere et al. (2008b), it is possible to estimate individual-level models from stated preference data obtained using efficient experimental designs. Their results suggest that distributions of preference weights generally depart substantially from normality. Nevertheless, models are by definition only approximations to "reality," so if the estimation of individual models is not feasible, it remains an empirical matter whether assuming normal preference weights is a good modelling choice.

The hierarchical Bayes (HB) approach has also become popular recently, in part because advances in simulation methods (Markov chain Monte Carlo) made it computationally practical (see Allenby and Rossi 1998, Geweke and Keane 2001). In HB, the ease of use advantage of MIXL vanishes, so both MNP and MIXL are widely used.[3] An appeal of HB is that by specifying weak priors that individual level parameters are normally distributed, one can allow considerable flexibility in their posterior distribution. However, as Allenby and Rossi (1998) note, HB procedures shrink individual level estimates toward the prior. Furthermore, as Rossi et al. (2005, p. 142) note, "the thin tails of the normal model tend to shrink outlying units greatly toward the center of the data." Thus, if one shrinks toward a normal prior, when the true heterogeneity distribution is highly nonnormal, it may result in unreliable inferences. A large amount of data per person may be necessary before the data "overwhelms" the prior.

In response to this problem, a literature has emerged on using mixtures of normal distributions to generate flexible priors that can easily accommodate a wide range of nonnormal posteriors. This approach, known as "Bayesian nonparametrics," originated with Ferguson (1973) for density estimation. It has been extended to probit by Geweke and Keane (1999, 2001), and to MIXL by Rossi et al. (2005) and Burda et al. (2008).[4] Figure 5.7 in Rossi et al. (2005) provides a nice illustration of how flexible the distribution of household posterior means can be in a mixture of normals model.[5] Of course, one can also adopt a mixture-of-normals specification for the heterogeneity distribution within the classical framework.

Here, we propose an alternative approach to modelling heterogeneity that stays within the classical framework and retains the simplicity of use of MIXL while extending it to accommodate both scale and residual taste heterogeneity. We show how to nest the MIXL model and the MNL model with scale heterogeneity (S-MNL) within a single framework. We refer to this new model as the "generalized multinomial logit model," or G-MNL. Estimating the G-MNL model allows one to assess whether including scale heterogeneity leads to a significant improvement in fit over the conventional MIXL model in any given data set.

[2] McFadden and Train (2000) show MIXL can approximate any random utility model arbitrarily well. However, this result relies on the investigator using the correct mixing distribution—which must be specified a priori. Unfortunately, their result seems to be widely misinterpreted among practitioners to mean MIXL with normal mixing can approximate any random utility model, which is certainly not true. Indeed, the correct mixing distribution is normal only in the case of the MNP model.

[3] Indeed, as noted in Train (2003, p. 316), MNP is actually somewhat computationally easier, because it has the same distribution (normal) for both the attribute weights and the idiosyncratic errors.

[4] In mixture of normals models, there is probability of drawing from each class in the mixture, and one must put a prior on that probability vector. Some authors adopt the Dirichlet process prior (DPP), which says there may be a countably infinite number of classes, but that is typically specified to put more prior mass on models with fewer classes. This is the approach of Ferguson (1973) recently extended to MIXL models by Burda et al. (2008). A second approach is to assume a fixed number of classes and put a Dirichlet (i.e., multivariate Beta) distribution on the vector of type probabilities. This approach is adopted in Geweke and Keane (1999, 2001) and Rossi et al. (2005). In practice there is no fundamental difference between the approaches. This is because (1) in the second method one can consider models with different numbers of classes and compare them based on the marginal likelihood, and (2) in the first approach, inference invariably puts essentially all mass on a fairly small number of types anyway. Thus, in practice, the two kinds of prior on the hyperparameters of the Dirichlet distribution produce essentially identical results, and the real point of this literature is the mixture of normals specification.

[5] Recently, Geweke and Keane (2007) introduced the "smoothly mixing regression" (SMR) model in which the class probabilities in a mixture of normals model are determined by a multinomial probit. The key advantage of SMR is that it allows class probabilities to depend on covariates, which is critical for modelling nonstationary processes. SMR is closely related to what are known as "mixture of experts" models in statistics (see Jiang and Tanner 1999, Villani et al. 2007).

Although it is not immediately obvious, G-MNL is closely related to mixture of normals models. The relation becomes clear if we adopt an approximate Bayesian perspective and use our estimated model to calculate person-specific parameters a posteriori (see Train 2003, Chapter 11). Then the estimated heterogeneity distribution plays the same role as the prior in the Bayesian framework. One can interpret our model as allowing a more flexible prior on the distribution of individual-level parameters than does a normal model but via a different means than the standard discrete mixtures of normals approach. Specifically, G-MNL implies the attribute coefficients are a continuous mixture of scaled normals.

We apply G-MNL to data from 10 stated preference choice experiments that cover several different types of choices: choices about medical procedures, mobile phones, food delivery services, holiday packages, and charge cards. We also estimate MIXL and S-MNL on each data set and compare the performance of the models using three information criteria: the Akaike (AIC), Bayes (BIC), and consistent Akaike (CAIC) criteria.

Our main finding is that models that include scale heterogeneity are preferred over MIXL by both BIC and CAIC in all 10 data sets: G-MNL in 7 and S-MNL in 3. The MIXL model is only (very slightly) preferred by the AIC for the two charge card data sets. However, our Monte Carlo results indicate that BIC and CAIC are more reliable measures for determining if scale heterogeneity is present. Interestingly, among practitioners the MIXL model with uncorrelated errors is very widely used (see Train 2007).[6] However, we find this model is dominated by either G-MNL or S-MNL in all 10 data sets, and in some cases by both.

Of course, it is also important to assess why the scale heterogeneity models fit better, in terms of what behavioural patterns they capture better than MIXL. We show that G-MNL can account for "extreme" consumers who exhibit nearly lexicographic preferences, whereas MIXL does not. We also show that G-MNL is better able to explain consumers who exhibit very "random" behaviour (in a sense we formalize below). Both of these advantages follow directly from the fact that the G-MNL model allows for much greater flexibility in the shape of posterior distribution of person-specific parameters than does MIXL.

A comparison of our results across data sets revealed two other interesting patterns. First, we can assess the importance of heterogeneity in general (both scale and residual) by looking at the percentage log-likelihood improvement in going from the simple MNL to the G-MNL model. By this metric, heterogeneity is roughly twice as important in the data sets that involve medical decisions as in those that involve product choices. We speculate that this may be because medical decisions involve more complex emotions, greater involvement, higher brain functions, or all three, than do consumer purchase decisions. But regardless of the reason, the result has important implications for the study of medical decision making.

Second, we look at the fraction of the overall likelihood improvement (from including all forms of heterogeneity) that is attained by including scale heterogeneity alone. This fraction is far greater in the four data sets involving medical decisions or cell phones than in the six involving food delivery, holiday packages, or charge cards. Similarly, the pseudo-R^2 improvements from including scale heterogeneity are greatest in the medical and cell phone data sets. These findings are consistent with a hypothesis that scale heterogeneity is more important in contexts involving more complex choice objects (medical tests or high-tech goods versus consumer goods). However, research on sources of heterogeneity is in its infancy (see, e.g., Louviere et al. 2002, Cameron et al. 2002),[7] so this hypothesis is only preliminary.

2. The Generalized Multinomial Logit Model

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

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

where x_{njt} is a K-vector of observed attributes of alternative j, β is a vector of utility weights (homogenous across consumers), and $\varepsilon_{njt} \sim$ i.i.d. extreme value is the "idiosyncratic" error. As emphasized by Keane (1997b), this error can be motivated as consumer heterogeneity in tastes for unobserved (or intangible or

[6] Presumably, a key reason is the ready availability of Ken Train's program for MIXL. The classical version of his program imposes uncorrelated errors (although the Bayesian version has an option to allow correlation).

[7] This previous work has examined complexity as a source of scale heterogeneity, defining complexity to be the amount of information subjects must process to make choices. Factors examined as contributors to complexity include the number of attributes, number of alternatives, number of attributes that differ among alternatives, and number of scenarios. However, complexity may also derive from the nature of attributes themselves (i.e., attributes of high-tech goods may be intrinsically harder to evaluate than those of simple consumer goods). There may also be individual differences in ability to deal with complexity, arising because of literacy differences, age differences, etc. For example, Fang et al. (2006) find that the ability to choose among insurance plans differs by level of cognitive ability.

latent) product attributes. The x_{njt} for $j = 1, \ldots, J$ may include alternative specific constants (ASCs), which capture persistence in the unobserved attributes (for each option j) over choice occasions. If the average consumer views option j as having desirable unmeasured attributes, it will have a positive ASC.

Of course, the great popularity of MNL stems from the fact that it generates simple closed-form expressions for the choice probabilities

$$P(j \mid X_{nt}) = \exp(\beta x_{njt}) \Big/ \sum_{k=1}^{J} \exp(\beta x_{nkt}), \tag{2}$$

where X_{nt} is the vector of attributes of all alternatives $j = 1, \ldots, J$. However, because of the restrictive assumptions that (1) the ε_{njt} are i.i.d. extreme value and (2) tastes for observed attributes are homogenous, MNL imposes a very special structure on how changes in elements of x_{njt} can affect choice probabilities. For instance, from (2) we see the restrictive IIA property

$$P(j \mid X_{nt})/P(k \mid X_{nt}) = \exp(\beta x_{njt} - \beta x_{nkt}),$$

which says the ratio of choice probabilities for alternatives j and k depends only on the attributes of j and k. Thus, changes in the attributes of any product l, or the introduction of a new product into the choice set, cannot alter the relative probabilities of j and k. This is obviously unrealistic in cases where product l is much more similar to j than to k.

One model that avoids IIA is the MIXL model. In MIXL the utility to person n from choosing alternative j on purchase occasion (or in choice scenario) t is given by

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

Here, β is the vector of mean attribute utility weights in the population, whereas η_n is the vector of person n-specific deviations from the mean. The idiosyncratic error ε_{njt} is still assumed to be i.i.d. extreme value. The investigator may specify any distribution for the η_n vector, but in most applications it is assumed to be multivariate normal, MVN$(0, \Sigma)$. However, the price coefficient is sometimes assumed to be log-normal to impose the proper sign restriction.

Many MIXL applications have assumed Σ is diagonal. This rules out that consumers who like a certain attribute will also tend to like (dislike) some other attribute. That is, it rules out correlation in tastes across attributes but not correlation in tastes across alternatives.[8]

A major appeal of MIXL is ease of use. It relaxes IIA yet it is still simple to program.[9] This is clear from the expression for how choice probabilities are simulated

$$P(j \mid X_{nt}) = \frac{1}{D} \sum_{d=1}^{D} \frac{\exp[(\beta + \eta^d) x_{njt}]}{\sum_{k=1}^{J} \exp[(\beta + \eta^d) x_{nkt}]}. \tag{4}$$

Thus, given D draws $\{\eta^d\}_{d=1,\ldots,D}$ from the multivariate normal MVN$(0, \Sigma)$, one obtains simulated choice probabilities just by averaging simple logit expressions over these draws.[10]

The scale heterogeneity model (S-MNL) can be understood by recognizing that the idiosyncratic error in both (1) and (3) has a scale or variance that has been implicitly normalized (to that of the standard extreme value distribution) to achieve identification. To proceed, let us write out the simple logit model with the scale of the error made explicit:

$$U_{njt} = \beta x_{njt} + \varepsilon_{njt}/\sigma$$
$$n = 1, \ldots, N;\ j = 1, \ldots, J;\ t = 1, \ldots, T. \tag{5}$$

Here, σ is the scale of the error term. Obviously, it is not possible to identify both β and σ, so it is standard practice to normalize σ to 1, which is equivalent to multiplying (5) through by σ. Now, suppose that σ is heterogeneous in the population, and denote its value

[8] It is important to emphasize that the assumption that Σ is diagonal does not rule out correlation across alternatives or within alternatives over time. Note that (3) can be rewritten as

$$U_{njt} = \beta x_{njt} + (\eta_n x_{njt} + \varepsilon_{njt}) = \beta x_{njt} + v_{njt}.$$

The composite error term $v_{njt} = (\eta_n x_{njt} + \varepsilon_{njt})$ will be positively (negatively) correlated across alternatives j that have similar (dissimilar) attributes, which is indeed the essential idea of the MIXL model. Thus, MIXL with diagonal Σ avoids IIA. It also allows for correlation over time, as a person who places high utility weights on certain attributes will persist in preferring brands with high levels of those attributes over time.

[9] The MNP model, which assumes the idiosyncratic errors have a multivariate normal distribution, also avoids IIA. It can be extended to allow the whole β vector to be normally distributed. MNP generates choice probabilities that are $J-1$ dimensional integrals with no closed form. Thus, estimation beyond the $J = 2$ case was precluded for many years by computational limits. The development of the GHK algorithm in the late 1980s (see Keane 1994, 1997b) made MNP estimation feasible. MNP algorithms are now available in popular packages like SAS and STATA, but these packaged programs only allow correlation across alternatives, not choice occasions. Geweke et al. (1997) discuss the MNP with correlation across alternatives and over choice occasions. As they make clear, the programming required in this general case is much more involved.

[10] Furthermore, with panel data or multiple-choice occasions per subject, the simulated choice probabilities are obtained simply by taking the products of period-by-period logit expressions and averaging them over draws d.

for person n by σ_n. Then, multiplying (5) through by σ_n we obtain the S-MNL model:

$$U_{njt} = (\beta\sigma_n)x_{njt} + \varepsilon_{njt}$$
$$n = 1, \ldots, N;\ j = 1, \ldots, J;\ t = 1, \ldots, T. \quad (6)$$

Notice that heterogeneity in scale is observationally equivalent to a particular type of heterogeneity in the utility weights. That is, Equation (6) implies that the vector of utility weights β is scaled up or down proportionately across consumers n by the scaling factor σ_n.[11]

If valid, S-MNL provides a much far parsimonious description of the data than MIXL, because $\beta\sigma_n$ is a much simpler object than $(\beta + \eta_n)$. For example, say there are 10 attributes. Then η_n is a 10-vector of normals, with a 10×10 covariance matrix containing 55 unique elements to be estimated. In contrast, σ_n is a scalar random variable, so its distribution will typically have far fewer parameters. For example, if σ_n is log-normal, we need only estimate its variance—a single parameter—because the mean must be constrained for identification (see below).

Recently, Louviere et al. (2008b) have criticized the MIXL model in (3). Based on the distributions of utility weights obtained from individual-level estimations, they have argued that (1) distributions do not appear very close to being normal, as assumed in most MIXL applications; and (2) when comparing coefficient vectors across consumers, something close to the scaling property implied by (6) seems to hold. Thus, they have argued that much of the heterogeneity in discrete models would be better captured by S-MNL than by MIXL.

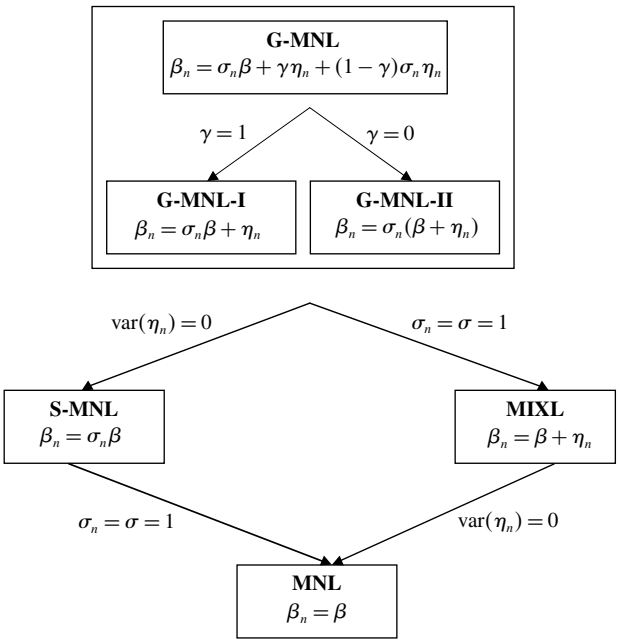
In an attempt to shed light on this issue, Keane (2006) noted that MIXL and S-MNL could be nested, to obtain a "generalized multinomial logit" model (G-MNL). Estimation of G-MNL would shed light on whether heterogeneity is better described by scale heterogeneity, normal mixing, or some combination of the two. In the G-MNL model, the utility to person n from choosing alternative j on purchase occasion (or in choice scenario) t is given by

$$U_{njt} = [\sigma_n\beta + \gamma\eta_n + (1-\gamma)\sigma_n\eta_n]x_{njt} + \varepsilon_{njt}, \quad (7)$$

where γ is a parameter between 0 and 1. Figure 1 describes how G-MNL nests MIXL, S-MNL, and MNL, as well as two other models we call G-MNL-I and G-MNL-II. To obtain MIXL one sets the scale parameter $\sigma_n = \sigma = 1$. To obtain the S-MNL model one sets $\mathrm{Var}(\eta_n) = 0$, meaning the variance-covariance matrix of η_n, denoted Σ, is degenerate.

[11] A common misconception is that random coefficient and scale heterogeneity models are fundamentally different. In fact, they are just different ways of specifying the distribution of coefficient heterogeneity.

Figure 1 The G-MNL Model and Its Special Cases



The parameter γ does not arise in either the MIXL or S-MNL special cases. It is only present in the G-MNL model, and its interpretation is more subtle than either σ_n or Σ. The parameter γ governs how the variance of residual taste heterogeneity varies with scale in a model that includes both. To see this, note that there are two equally sensible ways to nest MIXL and S-MNL. One might simply combine (3) and (6) to obtain what we call G-MNL-I:

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

Alternatively, one might start with (3) and make the scale parameter explicit:

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

Then, multiplying through by σ_n we obtain G-MNL-II:

$$U_{njt} = \sigma_n(\beta + \eta_n)x_{njt} + \varepsilon_{njt}$$
$$n = 1, \ldots, N;\ j = 1, \ldots, J;\ t = 1, \ldots, T. \quad (9)$$

Note that, in either model (8) or (9), we can write the vector of utility weights as

$$\beta_n = \sigma_n\beta + \eta_n^*.$$

In our terminology, the random variable σ_n captures scale heterogeneity and η_n^* captures residual taste heterogeneity. The difference between G-MNL-I and G-MNL-II is that in G-MNL-I, the standard deviation

of η_n^* is independent of the scaling of β, whereas in G-MNL-II, it is proportional to σ_n. As noted in Figure 1, G-MNL approaches G-MNL-I as $\gamma \to 1$, and it approaches G-MNL-II as $\gamma \to 0$. In the full G-MNL model, $\gamma \in [0, 1]$.

To enhance intuition, it is useful to consider using an estimated G-MNL model to calculate the posterior means of individual-level parameters. G-MNL-I adopts the prior they are a mixture of normals with different means but equal variances. G-MNL-II adopts the prior they are a mixture of normals with proportionally different means and standard deviations. The full G-MNL model allows for differential scaling of β and η_n^*. To impose that $\gamma \in [0, 1]$ in estimation, we use a logistic transform $\gamma = \exp(\gamma^*)/[1+\exp(\gamma^*)]$ and estimate the parameter γ^*. Thus, G-MNL approaches G-MNL-I as $\gamma^* \to \infty$ and approaches G-MNL-II as $\gamma^* \to -\infty$.

Finally, we must choose a distribution for σ_n. Because σ_n represents the (person-specific) scale of the idiosyncratic error, it should be positive. Thus, we assume it is log-normal with mean 1 and standard deviation τ, or LN(1, τ^2). Thus, τ is the key parameter that captures scale heterogeneity. As $\tau \to 0$, G-MNL approaches MIXL. If $\tau > 0$, then G-MNL approaches S-MNL as the diagonal elements of Σ approach 0. As both τ and Σ go to 0 we approach MNL.

3. Computation and Estimation

Here, we discuss the details of computation and estimation for the G-MNL model. To constrain the scale parameter σ_n to be positive, we use an exponential transformation

$$\sigma_n = \exp(\bar{\sigma} + \tau\varepsilon_{0n}), \quad \text{where } \varepsilon_{0n} \sim N(0, 1).$$

Thus, as the parameter τ increases, the degree of scale heterogeneity increases.

Obviously, as σ_n and β only enter the model as a product $\sigma_n\beta$, some normalization on σ_n is necessary to identify β. The natural normalization is to set the mean of σ_n equal to 1 so that β is the mean vector of utility weights. To achieve this $\bar{\sigma}$ must be a decreasing function of τ. Note that $E\sigma_n = \exp(\bar{\sigma} + \tau^2/2)$. Thus, to set $E\sigma_n = 1$ we should set $\bar{\sigma} = -\tau^2/2$.

Then, the simulated choice probabilities in the G-MNL model take the form

$$P(j \mid X_{nt}) = \frac{1}{D}\sum_{d=1}^{D} \frac{\exp(\sigma^d\beta + \gamma\eta^d + (1-\gamma)\sigma^d\eta^d)X_{njt}}{\sum_{k=1}^{J}\exp(\sigma^d\beta + \gamma\eta^d + (1-\gamma)\sigma^d\eta^d)X_{nkt}}, \quad (10)$$

where $\sigma^d = \exp(\bar{\sigma} + \tau\varepsilon_0^d)$. Note that η^d is a K-vector distributed MVN(0, Σ), whereas ε_0^d is a $N(0, 1)$ scalar. The simulation involves drawing the $\{\eta^d\}$ and $\{\varepsilon_0^d\}$ for $d = 1, \ldots, D$. Note that the computation in (10) is no more difficult than for the MIXL model in (4).

Now suppose we have either panel data or multiple-choice tasks per subject.[12] Let $y_{njt} = 1$ if person n chooses option j at time t, and 0 otherwise. Then, the simulated probability of observing person n choosing a sequence of choices $\{y_{njt}\}_{t=1}^{T}$ is given by

$$\hat{P}_n = \frac{1}{D}\sum_{d=1}^{D}\prod_t\prod_j (P(j \mid X_{nt}, \sigma^d, \eta^d))^{y_{njt}} = \frac{1}{D}\sum_{d=1}^{D}\prod_t\prod_j \left(\frac{\exp(\sigma^d\beta + \gamma\eta^d + (1-\gamma)\sigma^d\eta^d)X_{njt}}{\sum_{k=1}^{J}\exp(\sigma^d\beta + \gamma\eta^d + (1-\gamma)\sigma^d\eta^d)X_{nkt}}\right)^{y_{njt}},$$

which strings together period-specific probabilities like that inside the summation in (10).

In practice, we found the numerical performance of the algorithm is substantially improved by adopting two slight modifications to the above specification. First, we set the mean of σ_n equal to 1 in the simulated data, not merely in expectation. This means setting

$$\bar{\sigma} = -\ln\left[\frac{1}{N}\sum_{n=1}^{N}\exp\left(\tau\varepsilon_0^{d(n)}\right)\right],$$

where $\varepsilon_0^{d(n)}$ denotes the dth draw for the nth person. Second, if τ is too large, it causes numerical problems (i.e., overflows and underflows in exponentiation) for extreme draws of ε_0. To avoid this, we draw ε_0 from a truncated normal with truncation at ± 2.

We also discovered that the S-MNL model performs poorly when ASCs are scaled. That is, (1) the estimates often "blow up," with τ taking on very large values and the standard errors of the elements of β becoming very large; and (2) the model produces a substantially worse fit than one where only the coefficients on observed attributes are scaled, whereas ASCs are assumed homogenous in the population. Mechanically, these problems arise because data sets typically contain a set of individuals who always (or almost always) chose the same option, regardless of the elements of X_{nt}. If ASCs are scaled, the model can explain this phenomenon by making τ very large so that ASCs can vary substantially across individuals. The estimation algorithm usually decides to take this route.

[12] Applications of MIXL, S-MNL, and G-MNL typically involve multiple observations per subject. Although the MIXL model is formally identified given only one observation per subject (as a mixture of normal and extreme value errors may provide a slightly better fit to the data than extreme value alone), Harris and Keane (1999) showed the likelihood surface is extremely flat without repeated observations. We cannot expect to identify parameters that characterize heterogeneity in choice behaviour if we only see each person once.

On a more conceptual level, ASCs are fundamentally different from most observed attributes. For instance, for attributes like price or quality, it makes sense that all consumers have utility weights of the same sign, but these weights are scaled up or down across consumers (e.g., all consumers value quality, but some value it more than others). In contrast, ASCs tend to measure intangible aspects of products, which some consumers may like and others dislike. Thus, a model where all consumers have ASCs of the same sign, and where these are merely scaled, is unlikely to explain key patterns in choice behaviour, such as the strong persistence in choices (i.e., the loyalty consumers often exhibit for specific brands).

Thus, in models with ASCs, we specify S-MNL in one of two ways: In version one we assume the ASCs are homogenous in the population and do not scale them. In version two we treat the ASCs as random effects with a MVN distribution. We treat the ASCs in the G-MNL model in exactly the same way. Thus, we can rewrite (7) as

$$U_{njt} = (\beta_{0j} + \eta_{0nj}) + [\sigma_n \beta + \gamma \eta_n + (1-\gamma)\sigma_n \eta_n] x_{njt} + \varepsilon_{njt}, \quad (11)$$

where x_{njt} is now interpreted to include only observed attributes and not ASCs, and $\beta_{0j} + \eta_{0nj}$ is the ASC for alternative j. This consists of the component β_{0j}, which is constant across people, and the component η_{0nj}, which is heterogeneous across people. Thus, we have that $(\beta_0 + \eta_{0n})$ is a vector of ASCs, with β_0 being the mean vector and η_{0n} being the stochastic component. We assume that the entire vector $\{\eta_{0n}, \eta_n\}$ has a MVN distribution.

Finally, to explain why scale differs across people, or across choice occasions, we can let σ_n be a function of characteristics of people or choice occasions. That is, we could write

$$\sigma_{nt} = \exp(\bar{\sigma} + \theta z_{nt} + \tau \varepsilon_0), \quad (12)$$

where z_{nt} is a vector of attributes of person n and choice occasion t. One might let z_{nt} contain demographics, or some measure of the entropy of the choice occasion (e.g., how similar or dissimilar the choices are; see Swait and Adamowicz 2001, DeShazo and Fermo 2002).

4. Monte Carlo Results

In this section we present two Monte Carlo experiments to evaluate the properties of G-MNL model estimates. Of particular interest is whether the model can accurately assess the extent of scale heterogeneity (captured by τ) versus residual taste heterogeneity captured by Σ. To make the Monte Carlo experiments realistic, we constructed simulated data sets based on two of the empirical data sets that we will analyse in §5. The first is a data set where women choose whether to have a Pap smear exam, and the second is a data set where people choose between holiday locations (the "Holiday A" data set).

In each experiment, we use the actual Xs from the empirical data sets. The "true" parameters are obtained by estimating the G-MNL model on the empirical data sets. We generate 20 artificial data sets based on each empirical data set. We then estimate the G-MNL model on these 20 Monte Carlo data sets. In these estimations, as in our empirical work in §5, we use $D = 500$ draws to simulate the likelihood.

Table 1 presents results of estimating G-MNL on 20 data sets based on the Pap smear data. There are five attributes, including cost, doctor attributes, and contextual variables (i.e., if the test is recommended), and an ASC for the "Yes" option. There are 79 hypothetical respondents and 32 choice occasions per respondent. Table 1 reports the true parameter values, the mean estimates across the 20 replications, the empirical standard deviation of the estimates across data sets, and the mean of the asymptotic standard errors. An asterisk indicates that the bias in an estimated parameter is significant at the 5% level.

The results in Table 1 show evidence of significant bias for only a handful of parameters. Of the six elements of the β vector, only β_5 exhibits significant bias. However, the magnitude of the bias is less than 2/3 of an empirical standard deviation.[13] The standard deviations of residual taste heterogeneity are also rather precisely estimated, except for that on attribute 6, where it is upward biased.[14] Most importantly, the scale heterogeneity parameter τ is estimated quite precisely: its true value is 0.890 and the mean estimate is 0.891.

In these data sets the true value of the parameter γ is almost 0, but our mean estimate is 0.156. This is significantly greater than zero but still quantitatively small. In addition, the median estimate of γ is 0.08 (it exceeds 0.50 in only one out of 20 data sets). Thus, the model does a reasonable job of uncovering the fact that the true γ is small. For the most part, the empirical standard errors and mean asymptotic standard errors are close, suggesting the asymptotic theory is a good guide to the variability of the estimates.

[13] Bear in mind that ML estimates are only consistent—not unbiased in finite samples. This is why in Monte Carlo work it is generally argued that modest biases are to be expected and are not a major concern. Only quantitatively large biases that would substantially alter the interpretation of results would be a major concern.

[14] The estimates of the correlations among the errors (i.e., the residual taste heterogeneity) fall reasonably well in line with the true values, although significant biases show up in 6 out of 15 cases.

Table 1 Monte Carlo Simulation Results—Pap Smear Test Configuration

	True	$\bar{\theta}$	Std. dev.	$\overline{\text{ASE}}$		True	$\bar{\theta}$	Std. dev.	$\overline{\text{ASE}}$
β_1	−1.201	−1.192	0.742	0.542	ρ_{12}	−0.392	−0.271	0.377	0.308
β_2	0.466	0.525	0.405	0.444	ρ_{23}	−0.116	−0.031	0.270	0.281
β_3	−1.475	−1.458	0.818	0.601	ρ_{34}	−0.182	−0.186	0.170	0.216
β_4	3.563	3.983	1.045	0.908	ρ_{45}	−0.120	−0.016	0.265	0.406
β_5	1.657	1.928	0.408*	0.551	ρ_{56}	0.398	0.175	0.346*	0.565
β_6	−0.215	−0.241	0.123	0.192	ρ_{13}	0.075	0.211	0.287*	0.245
					ρ_{24}	0.073	0.030	0.319	0.233
σ_1	4.036	4.107	0.704	0.786	ρ_{35}	0.280	0.249	0.237	0.386
σ_2	1.631	1.768	0.469	0.555	ρ_{46}	−0.115	−0.169	0.355	0.497
σ_3	2.454	2.685	0.781	0.724	ρ_{14}	−0.385	−0.231	0.275*	0.233
σ_4	2.992	3.436	0.959	0.782	ρ_{25}	−0.418	−0.494	0.257	0.322
σ_5	1.506	1.787	0.674	0.640	ρ_{36}	−0.483	−0.219	0.329*	0.474
σ_6	0.226	0.490	0.175*	0.254	ρ_{15}	0.132	0.262	0.262*	0.424
					ρ_{26}	0.110	−0.077	0.370*	0.478
τ	0.890	0.891	0.328	0.214	ρ_{16}	0.239	0.048	0.415	0.522
γ^*	−5.000	−2.699	1.978*	16.215					
γ	0.007	0.156	0.198*	0.232					

Notes. The attributes and true values are constructed from the Pap smear data set (X_1 is ASC). The number of draws used in simulated maximum likelihood estimation is 500. We construct 20 artificial data sets (indexed by $m = 1, \ldots, 20$) and compare the estimates to the true values. $\bar{\theta} = (1/20)\sum_{m=1}^{20} \theta_m$; Std. dev. $= \sqrt{(1/19)\sum_{m=1}^{20}(\theta_m - \bar{\theta})^2}$; $\overline{\text{ASE}} = (1/20)\sum_{m=1}^{20} \text{ASE}_m$, where θ_m and ASE_m denote parameter estimates and asymptotic standard errors, respectively, from each data set. An asterisk indicates the t-statistic for the estimated bias greater than the critical value at the 5% level; i.e., $|t| > 2.09$ where $t = \sqrt{20}(\bar{\theta} - \theta_{\text{true}})\text{Std. dev.}^{-1}$. The true values are from Table 11.

Table 2 Monte Carlo Simulation Results—Holiday A Configuration

	True	$\bar{\theta}$	Std. dev.	$\overline{\text{ASE}}$		True	$\bar{\theta}$	Std. dev.	$\overline{\text{ASE}}$		True	$\bar{\theta}$	Std. dev.	$\overline{\text{ASE}}$
β_1	−0.905	−1.134	0.415*	0.242	ρ_{12}	0.216	0.092	0.219*	0.033	ρ_{47}	0.194	0.105	0.125*	0.119
β_2	1.012	1.214	0.619	0.277	ρ_{23}	0.012	0.045	0.303	0.180	ρ_{58}	0.113	0.193	0.156*	0.050
β_3	−0.189	−0.243	0.174	0.111	ρ_{34}	−0.092	−0.129	0.288	0.122	ρ_{15}	−0.065	−0.148	0.227	0.066
β_4	1.924	2.223	0.814	0.449	ρ_{45}	0.243	0.304	0.127*	0.067	ρ_{26}	−0.165	−0.157	0.197	0.113
β_5	1.771	2.032	0.733	0.413	ρ_{56}	0.225	0.255	0.185	0.102	ρ_{37}	−0.070	−0.011	0.294	0.164
β_6	0.860	0.885	0.299	0.209	ρ_{67}	0.094	0.173	0.186	0.143	ρ_{48}	−0.060	0.039	0.167*	0.050
β_7	0.262	0.232	0.180	0.120	ρ_{78}	−0.350	−0.201	0.196*	0.083	ρ_{16}	0.244	0.212	0.247	0.105
β_8	3.200	3.803	1.296	0.745	ρ_{13}	−0.043	−0.081	0.410	0.184	ρ_{27}	0.194	0.217	0.182	0.112
					ρ_{24}	0.446	0.453	0.129	0.050	ρ_{38}	0.106	0.018	0.254	0.076
σ_1	0.982	1.157	0.403	0.241	ρ_{35}	0.056	0.064	0.221	0.098	ρ_{17}	0.620	0.489	0.204*	0.106
σ_2	3.590	4.503	1.503*	0.854	ρ_{46}	0.182	0.162	0.198	0.104	ρ_{28}	0.015	0.035	0.139	0.048
σ_3	0.616	0.598	0.245	0.162	ρ_{57}	−0.358	−0.308	0.199	0.117	ρ_{18}	0.129	0.031	0.230	0.042
σ_4	1.891	2.451	0.785*	0.473	ρ_{68}	0.181	0.154	0.198	0.067					
σ_5	1.693	2.127	0.686*	0.414	ρ_{14}	0.007	−0.057	0.217	0.055					
σ_6	1.006	1.305	0.445*	0.283	ρ_{25}	0.072	0.087	0.126	0.070					
σ_7	0.877	1.119	0.407*	0.247	ρ_{36}	0.403	0.283	0.351	0.132					
σ_8	2.351	3.323	1.212*	0.638										
τ	1.000	0.968	0.344	0.138										
γ^*	−1.380	−2.633	2.260*	10.379										
γ	0.200	0.137	0.254*	0.118										

Notes. The attributes and true values are constructed from the Holiday A data set. The number of draws used in simulated maximum likelihood estimation is 500. We construct 20 artificial data sets (indexed by $m = 1, \ldots, 20$) and compare the estimates to the true values. $\bar{\theta} = (1/20)\sum_{m=1}^{20} \theta_m$; Std. dev. $= \sqrt{(1/19)\sum_{m=1}^{20}(\theta_m - \bar{\theta})^2}$; $\overline{\text{ASE}} = (1/20)\sum_{m=1}^{20} \text{ASE}_m$, where θ_m and ASE_m denote parameter estimates and asymptotic standard errors, respectively, from each data set. An asterisk indicates the t-statistic for the estimated bias greater than the critical value at the 5% level; i.e., $|t| > 2.09$, where $t = \sqrt{20}(\bar{\theta} - \theta_{\text{true}})\text{Std. dev.}^{-1}$. The true values are from Table 10, except τ is reduced from 1.51 to 1.00 to make detection of scale heterogeneity more challenging. In addition, γ is increased from 0 to 0.20 so to contrast with Table 1, where $\gamma = 0$.

The results of estimating the G-MNL model on the 20 artificial data sets based on the Holiday A data set are reported in Table 2. These data sets contain eight attributes, as described later. There are 331 hypothetical respondents and 16 choice occasions per respondent.

Of the eight elements of the β vector, only β_1 exhibits significant bias. However, the magnitude of the bias is only 1/2 of an empirical standard deviation. The scale heterogeneity parameter τ is again estimated quite precisely: its true value is 1.0 and the mean estimate is 0.968. However, the standard deviations of residual taste heterogeneity show a tendency to be *upward* biased, and this bias is significant for six out of eight parameters.

We would argue that this upward bias in the error variances is not a great cause for concern. The largest bias, which is for σ_8, is only 80% of an empirical standard deviation, and other significant biases are about 2/3 of a standard deviation. Biases of this magnitude are not surprising, in light of prior work showing it is often difficult to pin down error variance-covariance parameters in discrete-choice models (e.g., Geweke et al. 1994).

The true γ is 0.20 and the mean estimate is 0.137. This downward bias is significant, but the model does a reasonable job of uncovering the fact that γ is small. The greatest cause for concern in Table 2 is that asymptotic standard errors are systematically smaller than empirical standard errors. This was not the case in Table 1. The difference may arise because here we attempt to estimate a larger number of variance-covariance parameters (36 versus 21).

In §5 we use AIC, BIC, and CAIC to choose between the G-MNL, MIXL, and S-MNL models; it is important to consider if these criteria are reliable. To address this issue, we perform a 3×6 factorial experiment where we (1) simulate data where the true model is S-MNL, MIXL, or G-MNL (both with correlated errors); and (2) estimate the MNL, S-MNL, MIXL, and G-MNL models (both with and without correlated errors) on those data sets. As in Tables 1 and 2, this was done using data sets constructed to look like the Pap smear and Holiday A data.[15] We then counted the number of times that AIC, BIC, and CAIC preferred each model in each case. The results are reported in Table 3.

Consider first the case where G-MNL with correlated errors is the true model. In the Pap smear data sets, AIC correctly picks it in 9/20 cases. However, in 11/20 cases AIC chooses instead the more parsimonious G-MNL with uncorrelated errors. In contrast, for the Holiday A data sets, AIC correctly picks G-MNL with correlated errors in all 20 cases. Now consider BIC and CAIC (which have larger penalties for adding parameters). In both data sets, these criteria tend to pick the more parsimonious G-MNL with uncorrelated errors, even though errors are correlated. Indeed, they occasionally even pick MIXL with uncorrelated errors.

Next consider the case where MIXL with correlated errors is the true model. In this case BIC and CAIC correctly pick MIXL as the true model in the large majority of cases, but they always choose the more parsimonious version with uncorrelated errors. The performance of AIC in this case is poor, because it incorrectly chooses G-MNL in 12/20 cases in the Pap smear data sets, and 7/20 cases in the Holiday A data sets.

Table 3 Monte Carlo Simulation Results

			Correlated error		Uncorrelated error	
	MNL	S-MNL	MIXL	G-MNL	MIXL	G-MNL
(a) Pap smear test data configuration						
True DGP is G-MNL						
AIC	0	0	0	9	0	11
BIC	0	0	0	0	2	18
CAIC	0	0	0	0	4	16
True DGP is MIXL						
AIC	0	0	8	10	0	2
BIC	0	0	0	0	19	1
CAIC	0	0	0	0	19	1
True DGP is S-MNL						
AIC	0	20	0	0	0	0
BIC	0	20	0	0	0	0
CAIC	0	20	0	0	0	0
(b) Holiday A data configuration						
True DGP is G-MNL						
AIC	0	0	0	20	0	0
BIC	0	0	0	1	5	14
CAIC	0	0	0	0	6	14
True DGP is MIXL						
AIC	0	0	13	7	0	0
BIC	0	0	0	0	16	4
CAIC	0	0	0	0	17	3
True DGP is S-MNL						
AIC	0	20	0	0	0	0
BIC	0	20	0	0	0	0
CAIC	0	20	0	0	0	0

G-MNL or MIXL

		Wrong	
	Right	MIXL	G-MNL
BIC	68	7	5
CAIC	66	10	4
AIC	61	0	19

[15] For example, we fit the S-MNL model to the Pap smear data set, and use those estimates to generate the data where S-MNL is the true model. We fit MIXL with correlated errors to the Pap smear data set, and use those estimates to generate the data where MIXL is the true model. Finally, we fit G-MNL with correlated errors to the Pap smear data set and use those estimates to generate the data where G-MNL is the true model.

Table 4 Empirical Data Sets

	No. of choices	No. of choice occasions	No. of respondents	No. of observations	No. of attributes	Products	Meaningful ASC	Complicated attributes	Variation in attributes	All consumers are likely to have same signs
1 Tay Sachs disease and cystic fibrosis test—Jewish sample (3 ASCs)	4	16	210	3,360	11	Medical	Yes	Yes	High	Yes
2 Tay Sachs disease and cystic fibrosis test—general population sample (3 ASCs)	4	16	261	4,176	11	Medical	Yes	Yes	High	Yes
3 Mobile phone (1 ASC)	4	8	493	3,944	15	Consumption	No	Yes	High	Yes
4 Pizza A (no ASC)	2	16	178	2,848	8	Consumption	No	No	Low	No
5 Holiday A (no ASC)	2	16	331	5,296	8	Consumption	No	No	Low	No
6 Pap smear test (1 ASC)	2	32	79	2,528	6	Medical	Yes	No	Medium	Yes
7 Pizza B (no ASC)	2	32	328	10,496	16	Consumption	No	No	Low	No
8 Holiday B (no ASC)	2	32	683	21,856	16	Consumption	No	No	Low	No
9 Charge card A (2 ASCs)	3	4	827[a]	3,308	17	Consumption	Yes	No	High	Yes
Charge card B (3 ASCs)	4	4	827[a]	3,308	18	Consumption	Yes	No	High	Yes

[a]The respondents in the two credit card data sets are the same. They first complete four tasks with three options and then answer four tasks with four options. Some data sets were used in previous research (see Hall et al. 2006 for data sets 1 and 2; Fiebig and Hall 2005 for data set 6; Louviere et al. 2008a for data sets 4, 5, 7, and 8).

Finally, when S-MNL is the true model it is correctly identified by all three information criteria in all 40 cases. The reason for this success is that MIXL and G-MNL both involve a large increase in the number of parameters over S-MNL. In summary, although the results for the case when S-MNL is the true model are clear-cut, those for the cases where MIXL or G-MNL is the true model appear more ambiguous.

How can we make sense of these results? The bottom panel of Table 3 provides a useful summary. Here, we look only at the cases where MIXL or G-MNL is the true model, ignore the distinction between correlated and uncorrelated errors, and combine the results from the two data generating processes. We simply ask how reliably the three information criteria determine if the true model contains scale heterogeneity. Note that BIC makes the correct determination in 68/80 cases. It wrongly concludes that the true model is MIXL in 7/80 cases, and it only gives a false positive for scale heterogeneity in 5/80 cases. The results for CAIC are similar. In contrast, AIC has a bias towards accepting scale heterogeneity when it is not present (19/80 cases). This is not surprising, because the AIC has a smaller penalty for adding parameters, and G-MNL has only two more parameters than MIXL.

Given these results, we conclude that both BIC and CAIC provide accurate guides for whether scale heterogeneity is present—i.e., for distinguishing between MIXL and G-MNL. However, they are biased toward rejecting the presence of error correlations. This is not surprising as error correlations add many parameters, which these criteria penalize heavily. On the other hand, AIC correctly picks models where errors are correlated in 69/80 cases. Thus, we would recommend using the information criteria in conjunction: using BIC and/or CAIC as reliable measures of whether scale heterogeneity is present (i.e., MIXL versus G-MNL or S-MNL) and then using AIC to evaluate whether error correlations are important.

5. Empirical Results

5.1. Estimation Results

Our empirical results are based on data from 10 stated preference choice experiments described in Table 4. The data sets differ widely along several dimensions, including the object of choice (i.e., medical tests, mobile phones, pizza delivery services, holiday packages, and charge cards), the number of attributes (6 to 18), the number of choices (2 to 4), and the number of choice occasions (or choice sets) that each person faced in the experiment (4 to 32). All data sets are fairly large, but the number of observations also varies substantially (from 2,528 to 21,856). Table 5 lists all attributes and how they are coded in each data set.

Tables 6–15 present estimation results for the 10 data sets. We only discuss the results for data set 1 in detail, giving an overview of results for the other data sets in §§5.2 and 5.4. In data set 1, participants were asked whether they would chose to receive diagnostic tests for Tay Sachs disease, cystic fibrosis, both, or neither, giving four alternatives. The attributes that vary across choice scenarios are the cost of the tests, whether the person's doctor recommends the tests, the chance that the test is inaccurate, how the results of the tests will be communicated, and what the person is told about the probability that they are a carrier

Table 5 Attributes and Levels

Attributes	Levels
Tay Sachs disease (TS) and cystic fibrosis (CF) test: Jewish and general population	
1 ASC for TS test	0, 1
2 ASC for CF test	0, 1
3 ASC for both tests	0, 1
4 Your cost of being tested for TS	(0, 150, 300, 600)/1,000
5 Your cost of being tested for CF	(0, 375, 750, 1,500)/1,000
6 Your cost of being tested both TS and CF	(0, 150, . . . , 1,800, 2,100)/1,000
7 Whether your doctor recommends you have a test	−1 (no), 1 (yes)
8 The chance that you are a carrier even if the test is negative	(15, 30, 45, 60)/10
9 Whether you are told your carrier status as an individual or as a couple	−1 (individual), 1 (couple)
10 Risk of being a carrier for TS	log base 10 of (0.004, 0.04, 0.4, 4) × 10^3
11 Risk of being a carrier for CF	log base 10 of (0.004, 0.04, 0.4, 4) × 10^3
Mobile phone	
1 ASC for purchase (phone 1, phone 2, or phone 3)	0, 1
Voice commands (omitted text to voice or voice to text converter)	
2 (1) No	−1, 0, 1
3 (2) Voice dialling by number or name	−1, 0, 1
4 (3) Voice operating commands	−1, 0, 1
Push to communicate (omitted to share video)	
5 (1) No	−1, 0, 1
6 (2) To talk	−1, 0, 1
7 (3) To share pictures or video	−1, 0, 1
E-mail access (omitted e-mail with attachments)	
8 (1) Personal e-mails	−1, 0, 1
9 (2) Corporate e-mails (VPN, RIM)	−1, 0, 1
10 (3) Both personal and corporate e-mails on multiple accounts	−1, 0, 1
11 WiFi	−1 (no), 1 (yes)
12 USB cable or cradle connection	−1 (no), 1 (yes)
13 Thermometer	−1 (no), 1 (yes)
14 Flashlight	−1 (no), 1 (yes)
15 Price	(0, 11.7, 19.5, . . . , 497.25, 563.55)/100 (36 unique values)
Pap smear test	
1 ASC for test	0 (no), 1 (yes)
2 Whether you know doctor	0 (no), 1 (yes)
3 Whether doctor is male	0 (no), 1 (yes)
4 Whether test is due	0 (no), 1 (yes)
5 Whether doctor recommends	0 (no), 1 (yes)
6 Test cost	{0, 10, 20, 30}/10
Pizza A: Attributes 1–8; Pizza B: Attributes 1–16 (no ASC)	
1 Gourmet	−1 (traditional), 1 (gourmet)
2 Price	−1 ($13), 1 ($17)
3 Ingredient freshness	−1 (some canned), 1 (all fresh ingredients)
4 Delivery time	−1 (45 mins), 1 (30 min)
5 Crust	−1 (thin), 1 (thick)
6 Sizes	−1 (single size), 1 (3 sizes)
7 Steaming hot	−1 (warm), 1 (steaming hot)
8 Late open hours	−1 (till 10 P.M.), 1 (till 1 A.M.)
9 Free delivery charge	−1 ($2), 1 (free)
10 Local store	−1 (chain), 1 (local)
11 Baking method	−1 (traditional), 1 (woodfire)
12 Manners	−1 (friendly), 1 (polite and friendly)
13 Vegetarian availability	−1 (no), 1 (yes)
14 Delivery time guaranteed	−1 (no), 1 (yes)
15 Distance to the outlet	−1 (in other suburb), 1 (in own suburb)
16 Range/variety availability	−1 (restricted menu), 1 (large menu)
Holiday A: Attributes 1–8; Holiday B: Attributes 1–16 (no ASC)	
1 Price	−1 ($999), 1 ($1,200)
2 Overseas destination	−1 (Australia), 1 (overseas)
3 Airline	−1 (Qantas), 1 (Virgin)

Table 5 (Cont'd.)

Attributes	Levels
Holiday A: Attributes 1–8; Holiday B: Attributes 1–16 (no ASC)	
4 Length of stay	−1 (7), 1 (12)
5 Meal inclusion	−1 (no), 1 (yes)
6 Local tours availability	−1 (no), 1 (yes)
7 Peak season	−1 (off-peak), 1 (peak)
8 Four-star accommodation	−1 (2-star), 1 (4-star)
9 Length of trip	−1 (3 hours), 1 (5 hours)
10 Cultural activities	−1 (historical sites), 1 (museum)
11 Distance from hotel to attractions	−1 (200 m), 1 (5 km)
12 Swimming pool availability	−1 (no), 1 (yes)
13 Helpfulness	−1 (helpful), 1 (very helpful)
14 Individual tour	−1 (organized tour), 1 (individual)
15 Beach availability	−1 (no), 1 (yes)
16 Brand	−1 (jetset), 1 (creative holidays)
Charge card A and B (no transaction option for card A)	
1 ASC for credit card	0, 1
2 ASC for debit card	0, 1
3 ASC for transaction card	0, 1
4 Annual fee	$(-70, -30, 10, 70)/10$
5 Transaction fee	$(-0.5, -0.3, 0.1, 0.5) * 10$
6 Permanent overdraft facility	
Credit:	0 (N/A)
Debit/trans:	−1 (available), 1 (not available)
7 Overdraft interest free days (up to)	$(-30, 5, 15, 30)/10$
8 Interest charged on outstanding credit/overdraft	$(-0.075, -0.035, 0.015, 0.075) * 100$
9 Interest earned on positive balance	
Credit:	$(-0.025, 0.025) * 100$
Debit/trans:	$0.015 * 100$
10 Cash advance interest rate	
Credit:	$(-0.035, -0.005, 0.015, 0.035) * 100$
Debit/trans:	$0.015 * 100$
Location and shop access (omitted EFTPOS + telephone + Internet + mail, use worldwide)	
11 (1) Nowhere else, use Australia-wide	−1, 0, 1
12 (2) EFTPOS + telephone + Internet + mail, use Australia-wide	−1, 0, 1
13 (3) Nowhere else, use worldwide	−1, 0, 1
14 Loyalty scheme	0 (none), 1 (frequent flyer/fly buys and other rewards)
15 Loyalty scheme annual fees	$(-40, 40)/10$ if loyalty scheme = 1; 0 if loyalty scheme = 0
16 Loyalty scheme points earning	−1 (points on outstanding balance interest paid on), 1 (points on purchases only)
17 Merchant surcharge for using card	$(-0.03, -0.01, 0.01, 0.03) * 100$
18 Surcharge for transactions at other banks ATM	
Credit:	−1.5
Debit/trans:	(−1.5, −0.5, 0.5, 1.5)

for each disease.[16] The members of the sample in data set 1 are Ashkenazi Jews, who are a population of interest as they have a relatively high probability of carrying Tay Sachs.

The estimation results are presented in Table 6. The first column presents results for a simple MNL model. All attribute coefficients are significant with expected signs (except for how the result is communicated, which is not significant). Cost has a negative effect, doctor recommendation and risk factors have positive effects, and inaccuracy has a negative effect.

The second column contains results for the S-MNL model with homogeneous ASCs. The scale parameter τ is 1.14 with a standard error of 0.09, implying

[16] Note that the four alternative choice set in this experiment is of the form {A, B, AB, 0}, where A and B are the options to get one or the other medical test, AB is the option to get both tests, and 0 is the option to get neither test. In applying the multinomial choice framework in such a context, it is important to be careful in specifying the utility of the AB option. An early reference in this regard is Keane and Moffitt (1998), who modelled the decision to participate in combinations of available public welfare programs. They noted that the costs and benefits of jointly participating in two programs may be a complex function of the costs and benefits of participating in each program individually (e.g., one program may tax part of another program's benefits). In the present case, the interaction between the two tests is rather simple. First, the cost of getting both tests may be less than the sum of the costs getting each test separately. Thus the cost variable in the AB utility function is the joint cost of getting both tests together. Second, there may be savings on time, discomfort, or both in getting the two tests jointly rather than separately (e.g., it may require just one office visit, one blood sample, etc.). Because we do not measure such attributes directly, they are captured by the AB option having its own intercept. As Keane and Moffitt (1998) noted, estimation of separate models for options A and B cannot adequately capture such synergies between the two alternatives (even if the errors in the separate models are allowed to be correlated).

Table 6 Tay Sachs Disease (TS) and Cystic Fibrosis (CF) Test: Jewish Sample (Three ASCs)

							Correlated errors				Uncorrelated errors			
	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Mixed logit MIXL		G-MNL		Mixed logit MIXL		G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for TS test	**−0.57**	0.14	**−2.24**	0.11	**−0.57**	0.20	−0.67	0.47	−0.17	0.41	**−1.07**	0.18	**−0.95**	0.17
ASC for CF test	**−0.82**	0.15	**−2.39**	0.13	**−0.88**	0.22	−0.74	0.42	−0.27	0.36	**−1.14**	0.20	**−1.15**	0.19
ASC for both tests	−0.08	0.15	**−3.01**	0.12	0.01	0.27	−0.38	0.52	0.01	0.45	−0.43	0.20	−0.32	0.18
TS cost	**−2.51**	0.24	**−2.87**	0.40	**−3.45**	0.34	**−4.75**	0.63	**−5.62**	0.78	**−4.24**	0.34	**−5.41**	0.49
CF cost	**−1.43**	0.13	**−1.38**	0.20	**−1.96**	0.20	**−3.24**	0.38	**−3.57**	0.42	**−3.07**	0.25	**−4.11**	0.35
Both cost	**−1.20**	0.07	**−0.95**	0.13	**−2.70**	0.17	**−3.65**	0.26	**−4.25**	0.37	**−3.13**	0.20	**−4.23**	0.24
Recommend	**0.33**	0.04	**0.66**	0.11	**0.56**	0.06	**0.95**	0.13	**1.00**	0.19	**0.64**	0.08	**0.81**	0.10
Inaccuracy	**−0.12**	0.02	**0.22**	0.04	**−0.15**	0.03	−0.14	0.09	**−0.36**	0.10	**−0.12**	0.04	**−0.19**	0.05
Form	0.07	0.04	0.13	0.08	0.12	0.05	0.28	0.16	0.15	0.19	**0.24**	0.08	0.24	0.10
Own risk of TS	**0.50**	0.03	**1.62**	0.17	**1.05**	0.08	**1.39**	0.12	**1.67**	0.18	**1.10**	0.07	**1.20**	0.09
Own risk of CF	**0.47**	0.04	**1.27**	0.14	**1.02**	0.07	**1.26**	0.12	**1.50**	0.18	**1.02**	0.07	**1.37**	0.10
τ	—		**1.14**	0.09	**0.64**	0.06	—		**0.45**	0.08	—		**0.52**	0.05
γ									0.11	0.15			0.01	0.02
No. of parameters	11		12		18		77		79		22		24	
LL	−3,717		−3,223		−2,815		−2,500		−2,480		−2,753		−2,744	
AIC	7,455		6,469		5,666		5,154		**5,118**		5,550		5,535	
BIC	7,523		6,543		5,777		5,626		**5,601**		5,684		5,682	
CAIC	7,534		6,555		5,795		5,703		**5,680**		5,706		5,706	

Note. Bold estimates are statistically significant at the 1% level.

substantial scale heterogeneity in the data. Including scale heterogeneity leads to a dramatic improvement in the likelihood over MNL, from −3,717 to −3,223 (i.e., 494 points, or 13%). S-MNL adds only one parameter, so it leads to substantial improvements in all three information criteria (AIC, BIC, and CAIC).

In the third column we report results of the S-MNL model with heterogeneity in the ASCs. Allowing for such heterogeneity leads to a further substantial improvement in fit (e.g., 408 points in the likelihood, or 11%). Note that the scale heterogeneity parameter τ falls from 1.22 to 0.64 but remains highly significant with a standard error of 0.06.

The next two columns of Table 6 present results from MIXL and the G-MNL model that nests MIXL and S-MNL. Two aspects of the results are notable. First, although the S-MNL model provides a great improvement in fit compared to simple MNL, the improvement achieved by MIXL is, at least in this data set, considerably greater. MIXL achieves a log-likelihood of −2,500 versus −3,717 for MNL. This is a 33% improvement compared to the 24% improvement achieved by S-MNL (with random ASCs). Of course, this is not too surprising, because MIXL adds 66 parameters, whereas S-MNL (with random ASCs) adds only 7.

Second, G-MNL provides a better fit than either MIXL or S-MNL alone. By adding two parameters, it achieves a log-likelihood improvement of 20 points over MIXL, and it beats MIXL on all three information criteria (AIC, BIC, and CAIC).

Note that the G-MNL estimate of the scale parameter τ is 0.45 with a standard error of 0.08. Thus, the estimates imply a substantial degree of scale heterogeneity in the data, even after allowing for correlated normal random coefficients. As $\sigma_n = \exp(-\tau^2/2 + \tau\varepsilon_{0n})$, the estimates imply a person at the 90th percentile of the scale parameter would have his or her vector of utility weights scaled up by 57%, whereas a person at the 10th percentile would have his or her vector of utility weights scaled down by 46%.

The estimate of γ is 0.11, which implies the data is closer to the G-MNL-II model (see Equation (9)), where the variance of residual taste heterogeneity increases with scale, than the G-MNL-I model (see Equation (8)), where it is invariant to scale.

Finally, the last two columns of Table 6 report estimates of MIXL and G-MNL with uncorrelated residual taste heterogeneity. These are of interest in part because MIXL with uncorrelated coefficients is popular among practitioners. Restricting residual heterogeneity to be independent across attributes leads to a sharp deterioration of the log-likelihood—by over 250 points for both MIXL and G-MNL.[17] AIC, BIC, and

[17] A priori, one might have expected the deterioration in the likelihood to be less in the model with scale heterogeneity, because scale heterogeneity could "sop up" much of the positive correlation among the attribute weights. However, when we look at the estimated correlation matrix (not reported but available on request), we find that at least half of the correlations among attribute weights in this data set are negative.

Table 7 Tay Sachs Disease (TS) and Cystic Fibrosis (CF) Test: General Population Sample (Three ASCs)

	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Correlated errors: Mixed logit MIXL		Correlated errors: G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for TS test	**−3.18**	0.13	**−3.43**	0.11	**−3.17**	0.20	**−3.24**	0.32	**−3.29**	0.31	**−3.20**	0.18	**−3.36**	0.17
ASC for CF test	**−1.92**	0.12	**−3.18**	0.11	**−2.92**	0.21	**−2.61**	0.33	**−2.64**	0.29	**−2.61**	0.17	**−2.92**	0.15
ASC for both tests	**−1.49**	0.13	**−4.11**	0.11	**−3.63**	0.27	**−3.13**	0.44	**−3.73**	0.40	**−2.75**	0.20	**−3.03**	0.19
TS cost	**−1.12**	0.25	**−1.60**	0.40	**−1.39**	0.25	**−2.71**	0.50	**−2.99**	0.52	**−2.04**	0.32	**−1.62**	0.27
CF cost	**−0.73**	0.10	**−0.82**	0.17	**−0.87**	0.11	**−2.17**	0.30	**−2.56**	0.32	**−1.60**	0.15	**−1.24**	0.14
Both cost	**−0.51**	0.06	**−0.51**	0.12	**−1.11**	0.10	**−2.13**	0.23	**−2.27**	0.22	**−1.49**	0.14	**−1.37**	0.11
Recommend	**0.35**	0.03	**1.11**	0.15	**0.61**	0.05	**0.95**	0.12	**0.94**	0.12	**0.69**	0.07	**0.70**	0.08
Inaccuracy	0.02	0.02	**0.45**	0.06	**0.05**	0.02	0.10	0.07	0.02	0.06	−0.01	0.05	0.10	0.04
Form	0.06	0.03	**0.23**	0.07	0.08	0.04	0.25	0.10	0.21	0.13	**0.17**	0.06	**0.18**	0.06
Own risk of TS	**0.39**	0.03	**1.54**	0.19	**0.91**	0.07	**1.06**	0.11	**1.26**	0.13	**0.86**	0.06	**0.85**	0.06
Own risk of CF	**0.37**	0.03	**1.43**	0.18	**0.88**	0.06	**0.99**	0.10	**1.16**	0.10	**0.87**	0.06	**0.87**	0.06
τ	—		**1.53**	0.11	**0.89**	0.07	—		**0.56**	0.07	—		**0.64**	0.06
γ									**0.64**	0.08			**0.99**	0.02
No. of parameters	11		12		18		77		79		22		24	
LL	−4,649		−3,567		−3,221		−2,946		−2,914		−3,232		−3,199	
AIC	9,320		7,158		6,477		6,047		**5,986**		6,507		6,446	
BIC	9,390		7,234		6,591		6,535		**6,487**		6,646		6,598	
CAIC	9,401		7,246		6,610		6,612		**6,566**		6,668		6,622	

Note. Bold estimates are statistically significant at the 1% level.

CAIC all prefer the G-MNL model with *correlated* taste heterogeneity. This is a bit surprising, in light of our Monte Carlo result that BIC and CAIC tend to prefer the uncorrelated model even if correlation is present.

With the above discussion as a guide, the interested reader should be able to follow the empirical results in Tables 7–15. Rather than describe each of these in detail, we turn to a discussion of general patterns that emerge across data sets.

5.2. Comparing Model Fit Across Data Sets

Table 16 compares the fit of our seven alternative models (MNL, S-MNL, MIXL, G-MNL, and the latter two with uncorrelated taste heterogeneity) across the 10 data sets. Recall that our Monte Carlo results in §4

Table 8 Mobile Phones (One ASC)

	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Correlated errors: One-factor MIXL		Correlated errors: One-factor G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for purchase	**−0.80**	0.05	0.00	0.04	**−0.35**	0.12	**−0.54**	0.11	**−0.51**	0.13	**−0.50**	0.11	**−0.46**	0.12
No voice comm.	0.04	0.04	0.03	0.13	0.06	0.05	0.03	0.05	0.04	0.06	0.04	0.05	0.04	0.06
Voice dialing	**0.08**	0.04	0.20	0.14	0.05	0.06	0.03	0.06	0.07	0.06	**0.10**	0.05	0.09	0.06
Voice operation	**−0.12**	0.04	−0.22	0.17	−0.11	0.06	−0.10	0.06	−0.12	0.07	**−0.13**	0.05	**−0.12**	0.06
No push to com.	0.06	0.04	0.15	0.16	**0.12**	0.06	0.06	0.05	0.09	0.06	0.05	0.05	0.06	0.06
Push to talk	0.03	0.04	0.07	0.18	0.03	0.07	0.04	0.06	0.03	0.07	0.05	0.05	0.07	0.06
Push to share pics/video	−0.02	0.04	−0.14	0.19	−0.08	0.07	−0.04	0.06	−0.03	0.07	−0.02	0.05	−0.04	0.06
Personal e-mail	−0.07	0.04	−0.09	0.16	−0.04	0.06	−0.09	0.06	−0.11	0.07	−0.08	0.05	−0.07	0.06
Corporate e-mail	**0.09**	0.04	0.24	0.19	0.08	0.07	0.09	0.05	0.10	0.06	0.08	0.05	0.08	0.06
Both e-mails	−0.05	0.04	−0.16	0.17	−0.08	0.06	−0.01	0.06	−0.003	0.06	−0.03	0.05	−0.04	0.06
WiFi	0.001	0.02	−0.03	0.09	−0.02	0.03	0.02	0.03	0.02	0.04	−0.002	0.03	−0.01	0.03
USB cable/cradle	**0.06**	0.03	0.02	0.09	**0.08**	0.04	**0.07**	0.03	**0.08**	0.04	**0.07**	0.03	**0.08**	0.03
Thermometer	**0.07**	0.03	0.02	0.08	0.05	0.03	**0.06**	0.03	0.06	0.04	**0.07**	0.03	**0.08**	0.03
Flashlight	0.05	0.03	0.07	0.08	0.01	0.03	0.05	0.03	**0.08**	0.04	0.05	0.03	0.04	0.03
Price/100	**−0.32**	0.02	**−3.07**	0.47	**−1.02**	0.16	**−0.76**	0.06	**−0.84**	0.10	**−0.76**	0.06	**−0.88**	0.10
τ	—		**2.14**	0.13	**1.45**	0.15	—		**0.77**	0.19	—		**0.66**	0.18
γ									0.28	0.24			0.01	0.49
No. of parameters	15		16		17		45		47		30		32	
LL	−4,475		−4,102		−3,990		−3,962		−3,949		−3,971		−3,966	
AIC	8,980		8,236		8,014		8,014		**7,986**		8,002		7,996	
BIC	9,074		8,336		**8,121**		8,297		8,281		8,190		8,197	
CAIC	9,089		8,352		**8,138**		8,342		8,328		8,220		8,229	

Note. Bold estimates are statistically significant at the 5% level.

Table 9 Pizza A (No ASC)

	MNL		Scale heterogeneity S-MNL		Correlated errors: Mixed logit MIXL		Correlated errors: G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Gourmet	0.02	0.02	0.03	0.04	−0.01	0.06	0.16	0.45	0.03	0.05	**0.45**	0.22
Price	**−0.16**	0.02	**−0.19**	0.05	**−0.38**	0.06	−3.44	1.81	**−0.35**	0.06	**−1.67**	0.65
Ingredient freshness	**0.48**	0.03	**1.45**	0.29	**1.06**	0.10	**11.10**	5.46	**0.96**	0.08	**4.65**	1.69
Delivery time	**0.09**	0.03	**0.16**	0.08	**0.17**	0.07	1.24	0.72	**0.16**	0.05	**0.74**	0.35
Crust	0.02	0.03	0.01	0.04	0.08	0.08	0.70	0.70	0.02	0.06	0.42	0.26
Sizes	**0.09**	0.03	**0.12**	0.06	**0.17**	0.07	1.21	0.91	**0.20**	0.05	**0.81**	0.37
Steaming hot	**0.38**	0.03	**1.02**	0.24	**0.86**	0.11	**8.93**	4.32	**0.87**	0.08	**4.46**	1.64
Late open hours	**0.04**	0.02	0.08	0.06	0.07	0.06	0.39	0.55	0.07	0.05	0.29	0.17
τ	—		**1.69**	0.18	—		**2.00**	0.26	—		**1.79**	0.24
γ							0.02	0.01			0.01	0.01
No. of parameters	8		9		44		46		16		18	
LL	−1,657		−1,581		−1,379		−1,324		−1,403		−1,373	
AIC	3,330		3,179		2,847		**2,741**		2,838		2,782	
BIC	3,378		3,233		3,109		3,015		2,933		**2,889**	
CAIC	3,386		3,242		3,153		3,061		2,949		**2,907**	

Note. Bold estimates are statistically significant at the 5% level.

indicated that BIC and CAIC were the most reliable criteria for determining whether scale heterogeneity is present. According to BIC and CAIC, G-MNL is the preferred model in 7 out of 10 data sets. S-MNL (with random intercepts) is preferred in the remaining three data sets (mobile phones and charge cards A and B). Thus, models that include scale heterogeneity are preferred over MIXL in all cases.

That S-MNL is preferred in three cases is striking, given the great simplicity of this model relative to its competitors. For example, in the mobile phone data set, S-MNL (with a random intercept) beats MIXL by 176 points on BIC and beats G-MNL by 160 points. However, it has only 17 parameters, compared to 45 for MIXL and 47 for G-MNL. Similarly, in the charge card A and B data sets, S-MNL beats MIXL by 176 and 159 points on BIC, respectively.

Among the seven data sets where G-MNL is preferred by BIC and CAIC, the G-MNL model with correlated errors is preferred only in the two Tay Sachs data sets. The G-MNL model with uncorrelated residual taste heterogeneity is preferred in five data sets (Pap smear, Pizzas A and B, and Holidays A and B), but this result should be interpreted with caution in light of our Monte Carlo results in §4, showing that BIC and CAIC tend to prefer simpler models without correlation even when error correlations are present.

In the seven cases where BIC and AIC prefer G-MNL, the AIC, which imposes a smaller penalty for additional parameters, always prefers the full version of G-MNL with correlated errors. This is not too surprising, because our Monte Carlo results suggest that AIC is more likely to prefer models with correlated errors when correlation is in fact present.

The AIC results regarding the preferred model contradict BIC and CAIC in only three data sets. For mobile phones, AIC prefers G-MNL with correlated errors, whereas BIC and CAIC both prefer S-MNL. Also, in the two credit card data sets, AIC slightly prefers MIXL with correlated errors, although the advantage over G-MNL and S-MNL is very small.

In summary, models with scale heterogeneity (G-MNL or S-MNL) are preferred by all three information criteria in 8 out of 10 cases. In the other two cases, AIC picks MIXL, whereas BIC and CAIC pick S-MNL. Thus, there is clear evidence that scale heterogeneity is important in eight data sets and substantial evidence it is important in the other two.[18]

A final notable result is that MIXL with uncorrelated errors, which is very widely used (see Train 2007), is never preferred. According to BIC and CAIC, it is beaten by G-MNL with uncorrelated errors in every data set except mobile phones. It is beaten by S-MNL in mobile phones, as well as the Tay Sachs general population data, and the two charge card data sets. It is beaten by MIXL with correlated errors in the Tay Sachs and charge card data. According to BIC and CAIC, it is beaten by G-MNL with correlated errors in those four data sets plus Pizza B and Holiday B,

[18] Among the 10 data sets, γ ran off to 0—and had to be pegged near 0—in five cases, and it ran off to one in another. Only in four cases (the Tay Sachs data sets, mobile phones, and charge card A) do we find intermediate values of γ. Usually γ was small, implying the G-MNL-II model is often a reasonable description of the data. In practice, we would advise users of G-MNL who experience the problem of γ running off to 0 or 1 to simply peg the parameter (at 0 or 1) and estimate either G-MNL-I or G-MNL-II, whichever is appropriate.

Table 10 Holiday A (No ASC)

					Correlated errors				Uncorrelated errors			
	MNL		Scale heterogeneity S-MNL		Mixed logit MIXL		G-MNL		Mixed logit MIXL		G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Price	−**0.16**	0.02	−**0.17**	0.03	−**0.36**	0.04	−**0.91**	0.22	−**0.33**	0.04	−**0.74**	0.12
Overseas destination	**0.09**	0.02	**0.17**	0.02	**0.19**	0.07	**1.01**	0.26	**0.23**	0.06	**0.32**	0.11
Airline	−0.01	0.02	−0.05	0.02	−0.05	0.03	−0.19	0.11	−0.02	0.03	−0.1	0.06
Length of stay	**0.26**	0.02	**0.35**	0.04	**0.55**	0.05	**1.92**	0.42	**0.52**	0.04	**1.24**	0.19
Meal inclusion	**0.27**	0.02	**0.31**	0.03	**0.61**	0.05	**1.77**	0.39	**0.56**	0.04	**1.29**	0.2
Local tours availability	**0.09**	0.02	**0.09**	0.03	**0.23**	0.05	**0.86**	0.21	**0.19**	0.03	**0.45**	0.09
Peak season	0.03	0.02	0	0.03	0.08	0.05	0.26	0.12	0.06	0.03	0.14	0.07
Four-star accommodation	**0.44**	0.02	**0.65**	0.05	**0.92**	0.06	**3.2**	0.68	**0.86**	0.06	**1.99**	0.29
τ	—		**0.97**	0.08	—		**1.51**	0.14	—		**1.19**	0.10
γ							0.00	0.14			0.00	0.18
No. of parameters	8		9		44		46		16		18	
LL	−3,066		−2,967		−2,504		−2,469		−2,553		−2,519	
AIC	6,149		5,952		5,097		**5,031**		5,139		5,074	
BIC	6,201		6,011		5,386		5,333		5,244		**5,192**	
CAIC	6,209		6,020		5,430		5,379		5,260		**5,210**	

Note. Bold estimates are statistically significant at the 1% level.

and under AIC it is beaten by G-MNL with correlated errors in every data set. The only case where it comes even close to being the preferred model is the Pap smear data. Thus, the data offer no empirical support for MIXL with uncorrelated errors.

5.3. Why Do Models with Scale Heterogeneity Fit Better than MIXL?

We have shown that models with scale heterogeneity (either G-MNL or S-MNL) are preferred by BIC and CAIC in all 10 data sets, and preferred by AIC in 8 out of 10. Thus, we have strong evidence that models with scale heterogeneity provide a better fit to a wide range of data sets than do models like MIXL that rely on residual taste heterogeneity alone. In this section we ask, why do models with scale heterogeneity fit better? That is, what behavioural patterns can they explain better than the MIXL model? Moreover, what substantive behavioural predictions differ between the G-MNL model and simpler, nested models like MIXL?

These key questions are addressed in Figures 2 to 5 and looking specifically at the Pizza B data, although we could have shown similar figures for other data sets. In Figure 2, we order the 328 individuals from the one with the least negative log-likelihood contribution in the MIXL model (i.e., the person the model fits best) to the one with the most negative contribution (i.e., the person the model fits worst). We then plot these people from left to right (the dark circles). We also plot each person's log-likelihood contribution according to the G-MNL model (the light crosses). The horizontal line is the log-likelihood of the naïve model that assumes equal choice probabilities for both alternatives. We also divide the sample into thirds: the type I people on the left that MIXL fits best, the type II people in the middle, and the type III people on the right (for whom the fit is often worse than the naïve model).

The key result of Figure 2 is that G-MNL generally fits types I and type III better than MIXL, whereas the fit for type II is about the same. What does this mean? It turns out the type I people are "extreme" and have preferences close to lexicographic. For instance, of these 109 people, 22 always choose the pizza with fresher ingredients on all choice occasions, regardless of other attributes, 18 always choose the pizza with the lower price, etc.[19] The G-MNL model is better able to explain such extreme behaviour by saying that (1) some people have a very small scale for the error term (or, conversely, very large attribute weights), so there is little randomness in their behaviour; and (2) because attribute weights are random, for some people one or a few attributes are much more important than others, so that one or few attributes almost entirely drive choices.

Type III people present behaviour that is highly random. That is, their behaviour is largely driven by the idiosyncratic error term ε_{njt} and is little affected by attributes. Indeed, the naïve model that assumes equal choice probabilities regardless of attribute settings generally fits their behaviour better than MIXL. G-MNL still has trouble fitting the behaviour of such

[19] There are 32 choice occasions, but each attribute only differs between the two options on 16 occasions.

Table 11 Pap Smear Test (One ASC)

							Correlated errors				Uncorrelated errors			
	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Mixed logit MIXL		G-MNL		Mixed logit MIXL		G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for test	**−0.40**	0.14	**−1.93**	0.11	−0.60	0.37	**−1.93**	0.57	**−1.20**	0.45	**−1.26**	0.30	**−0.80**	0.31
If know doctor	**0.32**	0.09	**1.83**	0.45	**0.63**	0.14	**0.97**	0.29	0.47	0.30	**0.78**	0.18	**0.68**	0.21
If doctor is male	**−0.70**	0.09	**−0.97**	0.34	**−1.24**	0.16	−1.07	0.46	**−1.48**	0.53	**−1.39**	0.30	**−1.99**	0.32
If test is due	**1.23**	0.10	**5.35**	1.38	**2.74**	0.29	**3.33**	0.48	**3.56**	0.58	**3.26**	0.31	**3.35**	0.42
If doctor recommends	**0.51**	0.10	**2.68**	0.77	**0.74**	0.17	**1.31**	0.30	**1.66**	0.46	**1.33**	0.23	**1.65**	0.31
Test cost	**−0.08**	0.04	0.00	0.13	−0.17	0.07	−0.18	0.12	−0.22	0.16	−0.22	0.09	**−0.28**	0.09
τ	—		**1.45**	0.18	**0.81**	0.11	—		**0.89**	0.18	—		**1.00**	0.11
γ									0.00	0.42			0.01	0.38
No. of parameters	6		7		8		27		29		12		14	
LL	−1,528		−1,124		−1,063		−923		−914		−945		−935	
AIC	3,069		2,262		2,143		1,899		**1,887**		1,914		1,897	
BIC	3,104		2,303		2,189		2,057		2,056		1,984		**1,979**	
CAIC	3,110		2,310		2,197		2,084		2,085		1,996		**1,993**	

Note. Bold estimates are statistically significant at the 1% level.

Table 12 Pizza B (No ASC)

					Correlated errors				Uncorrelated errors			
	MNL		Scale heterogeneity S-MNL		One-factor MIXL		One-factor G-MNL		Mixed logit MIXL		G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Gourmet	0.01	0.01	**0.05**	0.01	0.01	0.02	0.06	0.03	0.01	0.02	0.03	0.03
Price	**−0.17**	0.01	**−0.25**	0.02	**−0.32**	0.03	**−0.54**	0.05	**−0.30**	0.03	**−0.79**	0.07
Ingredient freshness	**0.21**	0.01	**0.36**	0.03	**0.39**	0.03	**0.74**	0.06	**0.34**	0.03	**1.05**	0.08
Delivery time	**0.03**	0.01	0.04	0.02	0.05	0.02	**0.15**	0.04	0.05	0.02	**0.15**	0.04
Crust	**0.08**	0.01	**0.09**	0.01	**0.16**	0.03	**0.33**	0.05	**0.08**	0.03	**0.59**	0.06
Sizes	**0.07**	0.01	**0.08**	0.02	**0.10**	0.02	**0.17**	0.03	**0.11**	0.02	**0.23**	0.03
Steaming hot	**0.20**	0.01	**0.35**	0.03	**0.34**	0.03	**0.76**	0.05	**0.34**	0.02	**1.15**	0.09
Late open hours	**0.04**	0.01	0.02	0.02	**0.07**	0.02	**0.12**	0.03	**0.08**	0.02	0.08	0.04
Free delivery charge	**0.12**	0.01	**0.15**	0.02	**0.21**	0.02	**0.41**	0.04	**0.20**	0.02	**0.56**	0.06
Local store	**0.08**	0.01	**0.06**	0.02	**0.13**	0.02	**0.24**	0.04	**0.15**	0.02	**0.42**	0.05
Baking method	**0.07**	0.01	**0.07**	0.02	**0.10**	0.02	**0.22**	0.03	**0.11**	0.02	**0.25**	0.04
Manners	0.01	0.01	−0.004	0.02	0.02	0.02	−0.07	0.04	0.02	0.02	0.01	0.04
Vegetarian availability	**0.09**	0.01	**0.06**	0.01	**0.11**	0.03	**0.21**	0.05	**0.13**	0.03	**0.34**	0.06
Delivery time guaranteed	**0.07**	0.01	**0.07**	0.02	**0.11**	0.02	**0.16**	0.03	**0.11**	0.02	**0.15**	0.04
Distance to the outlet	**0.06**	0.01	0.04	0.02	**0.09**	0.02	**0.12**	0.03	**0.09**	0.02	0.10	0.04
Range/variety availability	**0.06**	0.02	0.04	0.02	**0.09**	0.02	**0.12**	0.04	**0.09**	0.02	**0.14**	0.05
τ	—		**1.22**	0.08	—		**1.12**	0.06	—		**1.26**	0.06
γ							0.01	0.01			0.01	0.01
No. of parameters	16		17		48		50		32		34	
LL	−6,747		−6,607		−5,857		−5,668		−5,892		−5,689	
AIC	13,525		13,249		11,810		**11,436**		11,849		11,446	
BIC	13,641		13,372		12,159		11,799		12,081		**11,693**	
CAIC	13,657		13,389		12,207		11,849		12,113		**11,727**	

Note. Bold estimates are statistically significant different at the 1% level.

Table 13 Holiday B (No ASC)

	MNL		Scale heterogeneity S-MNL		Correlated errors: One-factor MIXL		Correlated errors: One-factor G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Price	−**0.16**	0.01	−**0.16**	0.01	−**0.25**	0.02	−**0.31**	0.02	−**0.25**	0.02	−**0.34**	0.02
Overseas destination	**0.08**	0.01	**0.12**	0.01	**0.17**	0.02	**0.21**	0.03	**0.12**	0.02	**0.24**	0.03
Airline	−0.02	0.01	−0.02	0.01	−**0.03**	0.01	−0.03	0.02	−**0.03**	0.01	−0.03	0.02
Length of stay	**0.18**	0.01	**0.19**	0.01	**0.30**	0.02	**0.40**	0.02	**0.29**	0.02	**0.40**	0.02
Meal inclusion	**0.20**	0.01	**0.24**	0.02	**0.33**	0.02	**0.45**	0.02	**0.34**	0.02	**0.46**	0.03
Local tours availability	**0.07**	0.01	**0.08**	0.01	**0.11**	0.01	**0.15**	0.02	**0.11**	0.01	**0.17**	0.02
Peak season	0.003	0.01	0.02	0.01	0.003	0.01	0.005	0.01	0.001	0.01	−0.01	0.02
Four-star accommodation	**0.34**	0.01	**0.54**	0.03	**0.51**	0.02	**0.65**	0.03	**0.50**	0.02	**0.69**	0.03
Length of trip	−0.02	0.01	−**0.03**	0.01	−**0.04**	0.01	−**0.03**	0.01	−**0.03**	0.01	−0.03	0.02
Cultural activities	−**0.05**	0.01	−**0.05**	0.01	−**0.09**	0.01	−**0.11**	0.02	−**0.09**	0.01	−**0.12**	0.01
Distance from hotel to attractions	−**0.08**	0.01	−**0.07**	0.01	−**0.13**	0.01	−**0.17**	0.02	−**0.12**	0.01	−**0.17**	0.02
Swimming pool availability	**0.09**	0.01	**0.09**	0.01	**0.15**	0.01	**0.19**	0.02	**0.15**	0.01	**0.23**	0.02
Helpfulness	**0.04**	0.01	**0.03**	0.01	**0.06**	0.01	**0.08**	0.02	**0.06**	0.01	**0.07**	0.02
Individual tour	**0.07**	0.01	**0.07**	0.01	**0.11**	0.02	**0.19**	0.02	**0.13**	0.02	**0.20**	0.02
Beach availability	**0.11**	0.01	**0.10**	0.01	**0.19**	0.01	**0.23**	0.02	**0.18**	0.01	**0.22**	0.02
Brand	0.001	0.01	−0.01	0.02	−0.004	0.02	0.01	0.02	0.003	0.02	0.004	0.02
τ	—		**1.13**	0.05	—		**0.67**	0.04	—		**0.72**	0.04
γ							0.01	0.02			0.01	0.02
No. of parameters	16		17		48		50		32		34	
LL	−13,478		−13,027		−11,570		−11,446		−11,600		−11,476	
AIC	26,988		26,088		23,236		**22,992**		23,263		23,019	
BIC	27,116		26,224		23,619		23,391		23,519		**23,291**	
CAIC	27,132		26,241		23,667		23,441		23,551		**23,325**	

Note. Bold estimates are statistically significant at the 1% level.

Table 14 Charge Card A (Two ASCs)

	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Correlated errors: One-factor MIXL		Correlated errors: One-factor G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est	Std. err.	Est.	Std. err.	Est	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for credit	−**0.85**	0.08	−**1.00**	0.05	−**0.90**	0.18	−**1.31**	0.27	−**1.31**	0.27	−**2.51**	0.36	−**3.15**	0.34
ASC for debit	−**0.99**	0.08	−**1.35**	0.05	−**1.22**	0.18	−**2.07**	0.31	−**2.05**	0.32	−**3.34**	0.44	−**4.16**	0.46
Annual fee	−**0.08**	0.01	−**0.14**	0.02	−**0.13**	0.01	−**0.18**	0.02	−**0.19**	0.02	−**0.27**	0.03	−**1.14**	0.23
Trans fee	−**0.53**	0.07	−**0.80**	0.11	−**0.82**	0.11	−**1.34**	0.20	−**1.37**	0.21	−**1.53**	0.27	−**7.90**	1.77
Overdraft facility	**0.28**	0.06	**0.58**	0.09	**0.43**	0.09	**0.70**	0.15	**0.75**	0.16	**0.80**	0.20	**4.05**	0.93
Overdraft free days	0.04	0.02	**0.06**	0.02	**0.06**	0.02	0.07	0.03	0.07	0.03	0.08	0.04	**0.35**	0.13
Interest charged	−**0.43**	0.06	−**1.26**	0.15	−**0.67**	0.09	−**1.00**	0.15	−**1.01**	0.16	−**1.29**	0.21	−**6.45**	1.26
Interest earned	**0.04**	0.01	0.02	0.01	0.04	0.02	0.06	0.03	0.06	0.03	0.08	0.04	0.31	0.13
Access_1	−0.05	0.02	−**0.06**	0.02	−**0.08**	0.02	−0.05	0.03	−0.06	0.03	−**0.14**	0.05	−0.30	0.16
Access_2	−**0.21**	0.05	−**0.35**	0.07	−**0.31**	0.08	−**0.42**	0.13	−**0.39**	0.12	−**0.54**	0.16	−**1.77**	0.61
Access_3	0.06	0.05	**0.26**	0.07	0.11	0.07	0.22	0.11	0.23	0.11	0.32	0.14	0.81	0.50
Cash advance interest	−0.06	0.05	−**0.39**	0.07	−0.12	0.08	−0.29	0.13	−0.34	0.14	−0.33	0.15	−**1.91**	0.63
Loyal scheme	**0.26**	0.06	**0.56**	0.08	**0.33**	0.08	**0.44**	0.14	**0.47**	0.15	0.37	0.20	**3.18**	0.83
Loyal fee	−**0.03**	0.01	−**0.04**	0.01	−**0.05**	0.01	−**0.06**	0.02	−**0.06**	0.02	−**0.08**	0.03	−0.15	0.11
Loyal point	−0.04	0.04	0.04	0.04	0.04	0.06	0.07	0.09	0.07	0.09	0.13	0.13	0.73	0.49
Merchant surcharge	−0.02	0.01	−**0.09**	0.02	−**0.07**	0.02	−**0.08**	0.03	−**0.08**	0.03	−0.10	0.04	−**0.57**	0.16
Surcharge at other ATM	−0.10	0.04	−**0.22**	0.04	−**0.17**	0.06	−0.20	0.11	−0.19	0.11	−0.20	0.12	−**1.99**	0.44
τ	—		**1.86**	0.17	**0.40**	0.17	—		0.21	0.24	—		**2.17**	0.20
γ									0.50	0.56			0.00	0.18
No. of parameters	17		18		21		51		53		35		37	
LL	−3,354		−3,217		−2,768		−2,735		−2,734		−2,868		−2,820	
AIC	6,742		6,470		5,579		**5,572**		5,574		5,806		5,714	
BIC	6,846		6,580		**5,707**		5,883		5,898		6,020		5,940	
CAIC	6,863		6,598		**5,728**		5,934		5,951		6,055		5,977	

Note. Bold estimates are statistically significant at the 1% level.

Table 15 Charge Card B (Three ASCs)

	MNL		Scale heterogeneity S-MNL		Random effects S-MNL		Correlated errors: One-factor MIXL		Correlated errors: One-factor G-MNL		Uncorrelated errors: Mixed logit MIXL		Uncorrelated errors: G-MNL	
	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
ASC for credit	**−0.97**	0.07	**−1.02**	0.05	**−0.83**	0.18	**−1.29**	0.24	**−1.29**	0.24	**−3.06**	0.29	**−2.72**	0.25
ASC for debit	**−1.29**	0.08	**−1.78**	0.07	**−1.47**	0.20	**−1.99**	0.27	**−1.99**	0.27	**−4.18**	0.37	**−4.54**	0.34
ASC for transaction	**−1.32**	0.08	**−1.72**	0.06	**−1.59**	0.21	**−2.12**	0.29	**−2.12**	0.29	**−4.30**	0.38	**−4.76**	0.36
Annual fee	**−0.10**	0.01	**−0.13**	0.01	**−0.16**	0.01	**−0.22**	0.02	**−0.22**	0.02	**−0.28**	0.02	**−0.49**	0.05
Trans fee	**−0.61**	0.07	**−0.71**	0.07	**−0.94**	0.10	**−1.32**	0.17	**−1.32**	0.17	**−1.72**	0.23	**−3.48**	0.45
Overdraft facility	**0.30**	0.06	**0.54**	0.06	**0.42**	0.08	**0.48**	0.11	**0.48**	0.11	**0.98**	0.15	**1.77**	0.22
Overdraft free days	**0.06**	0.02	**0.08**	0.01	**0.09**	0.02	**0.10**	0.03	**0.10**	0.03	**0.15**	0.04	0.18	0.07
Interest charged	**−0.56**	0.06	**−0.96**	0.08	**−0.80**	0.08	**−0.90**	0.12	**−0.90**	0.13	**−1.65**	0.19	**−3.21**	0.34
Interest earned	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.00	0.04
Access_1	−0.01	0.02	**0.11**	0.02	0.00	0.02	−0.01	0.03	−0.01	0.03	−0.01	0.05	**0.19**	0.07
Access_2	**−0.21**	0.05	**−0.22**	0.05	**−0.35**	0.07	**−0.44**	0.10	**−0.44**	0.10	**−0.58**	0.13	**−1.08**	0.22
Access_3	**0.13**	0.05	**0.16**	0.04	**0.19**	0.06	**0.32**	0.09	**0.32**	0.09	0.28	0.11	**0.48**	0.17
Cash advance interest	**−0.19**	0.05	**−0.27**	0.05	**−0.32**	0.06	**−0.45**	0.11	**−0.45**	0.11	**−0.50**	0.13	**−0.89**	0.20
Loyal scheme	**0.24**	0.05	**0.50**	0.05	**0.37**	0.07	**0.46**	0.11	**0.46**	0.11	**0.44**	0.17	**1.56**	0.23
Loyal fee	−0.02	0.01	**−0.03**	0.01	**−0.04**	0.01	−0.04	0.02	−0.04	0.02	−0.07	0.03	−0.07	0.04
Loyal point	−0.03	0.04	−0.09	0.04	−0.06	0.06	−0.06	0.08	−0.06	0.08	−0.22	0.11	−0.41	0.16
Merchant surcharge	**−0.06**	0.01	**−0.04**	0.01	**−0.08**	0.02	**−0.13**	0.03	**−0.13**	0.03	**−0.13**	0.03	**−0.20**	0.05
Surcharge at other ATM	−0.07	0.03	−0.05	0.03	−0.11	0.04	**−0.19**	0.07	**−0.19**	0.07	−0.18	0.08	−0.17	0.11
τ	—		**1.18**	0.09	**0.38**	0.12	—		0.00	0.19	—		**1.52**	0.11
γ									0.99	171			0.01	0.25
No. of parameters	18		19		25		54		56		36		38	
LL	−4,100		−3,947		−3,402		−3,364		−3,364		−3,528		−3,447	
AIC	8,236		7,932		6,854		**6,836**		6,840		7,128		6,970	
BIC	8,346		8,048		**7,007**		7,166		7,182		7,348		7,202	
CAIC	8,364		8,067		**7,032**		7,220		7,238		7,384		7,240	

Note. Bold estimates are statistically significant at the 1% level.

Table 16 Comparing Model Fit Across Data Sets

	Criteria	MNL	Scale heterogeneity S-MNL	Random effects S-MNL	Correlated errors: MIXL	Correlated errors: G-MNL	Uncorrelated errors: MIXL	Uncorrelated errors: G-MNL
Tay Sachs disease and cystic fibrosis test—Jewish sample (3 ASCs)	AIC	7,455	6,469	5,666	5,154	**5,118**	5,550	5,535
	BIC	7,523	6,543	5,777	5,626	**5,601**	5,684	5,682
	CAIC	7,534	6,555	5,795	5,703	**5,680**	5,706	5,706
Tay Sachs disease and cystic fibrosis test—general population (3 ASCs)	AIC	9,320	7,158	6,477	6,047	**5,986**	6,507	6,446
	BIC	9,390	7,234	6,591	6,535	**6,487**	6,646	6,598
	CAIC	9,401	7,246	6,610	6,612	**6,566**	6,668	6,622
Mobile phone (1 ASC)	AIC	8,980	8,236	8,014	8,014	**7,986**	8,002	7,996
	BIC	9,074	8,336	**8,121**	8,297	8,281	8,190	8,197
	CAIC	9,089	8,352	**8,138**	8,342	8,328	8,220	8,229
Pizza A (no ASC)	AIC	3,330	3,179		2,847	**2,741**	2,838	2,782
	BIC	3,378	3,233		3,109	3,015	2,933	**2,889**
	CAIC	3,386	3,242		3,153	3,061	2,949	**2,907**
Holiday A (no ASC)	AIC	6,149	5,952		5,097	**5,031**	5,139	5,074
	BIC	6,201	6,011		5,386	5,333	5,244	**5,192**
	CAIC	6,209	6,020		5,430	5,379	5,260	**5,210**
Pap smear test (1 ASC)	AIC	3,069	2,262	2,143	1,899	**1,887**	1,914	1,897
	BIC	3,104	2,303	2,189	2,057	2,056	1,984	**1,979**
	CAIC	3,110	2,310	2,197	2,084	2,085	1,996	**1,993**
Pizza B (no ASC)	AIC	13,525	13,249		11,810	**11,436**	11,849	11,446
	BIC	13,641	13,372		12,159	11,799	12,081	**11,693**
	CAIC	13,657	13,389		12,207	11,849	12,113	**11,727**
Holiday B (no ASC)	AIC	26,988	26,088		23,236	**22,992**	23,263	23,019
	BIC	27,116	26,224		23,619	23,391	23,519	**23,291**
	CAIC	27,132	26,241		23,667	23,441	23,551	**23,325**
Charge card A (2 ASCs)	AIC	6,742	6,470	5,579	**5,572**	5,574	5,806	5,714
	BIC	6,846	6,580	**5,707**	5,883	5,898	6,020	5,940
	CAIC	6,863	6,598	**5,728**	5,934	5,951	6,055	5,977
Charge card B (3 ASCs)	AIC	8,236	7,932	6,854	**6,836**	6,840	7,128	6,970
	BIC	8,346	8,048	**7,007**	7,166	7,182	7,348	7,202
	CAIC	8,364	8,067	**7,032**	7,220	7,238	7,384	7,240

Note. Bold estimates indicate the preferred model in each row.

people, but it gives a clear improvement over MIXL. G-MNL is better able to explain "random" people because it can say that some people have a very large scale of the error term (or, conversely, very small attribute weights).[20]

Some further insight is gained by looking at the bottom panel of Figure 2. Here, we fit simple MNL models to the type I, II, and III groups separately. Note that for the type III people we obtain very small attribute weights. Thus, choices made by type III people are largely driven by the error terms. In this sense they are highly random. In contrast, for type I people we estimate very large attribute weights. This makes their choices very sensitive to attribute settings. The type II people are in the middle. One can clearly see a general scaling up of the attribute weights as one moves from type III to type II to type I.

Having isolated why G-MNL fits better than MIXL, we turn to the question of how its substantive predictions differ. In Figure 3, the four graphs correspond to MNL, S-MNL, MIXL, and G-MNL, respectively. Each graph shows the distribution of people in terms of their probability of choosing between the two pizza delivery services.[21] The distribution is shown under two scenarios: a baseline where services A and B have identical attributes, and a scenario where service A improves ingredient quality (to all fresh) while also raising the price by \$4.

Of course, under the baseline, each model says that 100% of the people have a 50% probability of choosing A. After the policy change, MNL (which assumes homogeneous preferences) predicts that *all* people have a 52% chance of choosing A. In contrast, S-MNL predicts heterogeneity in consumer responses. Forty-one percent of consumers continue to have a roughly 50% chance of choosing A, whereas for 43% the probability of choosing A increases to about 55%, and for 17% of the probability of choosing A increases into the 60%–75% range.

The more interesting comparison is between MIXL and G-MNL. G-MNL predicts that after the policy change 14% of consumers still have a 50% chance of choosing A. Strikingly, 8% of consumers would have essentially a 100% chance of choosing A (these are the types who put great weight on fresh ingredients), whereas 5% would have essentially a 0% chance of choosing A (these are the types who care primarily about price). As we would expect based on the Figure 2 results, MIXL predicts that fewer people stay indifferent and also that fewer people have extreme reactions. Specifically, MIXL predicts that only 8% of consumers stay at roughly a 50% chance of choosing A, whereas essentially no consumers have their choice probabilities move all the way to 100% or 0%.

In the actual Pizza B data, there are $24/328 = 7.3\%$ of subjects who choose the fresh ingredient pizza on all choice occasions regardless of other attribute settings, and there are $27/328 = 8.2\%$ who always choose the less expensive pizza. The Figure 3 results show that G-MNL can generate such extreme (or lexicographic) behaviour, whereas MIXL cannot.

To gain additional insight into why the behavioural predictions of G-MNL and MIXL differ, we report for each model the posterior means of the person-level coefficients on fresh ingredients and price. To do this we condition on the estimated model parameters and the 32 observed choices of each person, using the algorithm in Train (2003, p. 266). Distributions of the posterior means of the person-specific parameters are reported in Figure 4.[22]

The difference in posteriors generated from MIXL and G-MNL is striking. Although the MIXL posteriors depart modestly from normality, the strong influence of the normal prior, and its tendency to pull in tail observations, is evident. In contrast, the G-MNL posteriors are multimodal, with considerable tail mass. For instance, the G-MNL posterior for the price coefficient has a substantial mass of people in the left tail who care greatly about price. In addition, the G-MNL posterior for the coefficient on fresh ingredients has a substantial mass of people in the right tail who greatly value freshness. This illustrates the flexibility of the continuous mixture of scaled normals prior for individual-level coefficients in the G-MNL model.

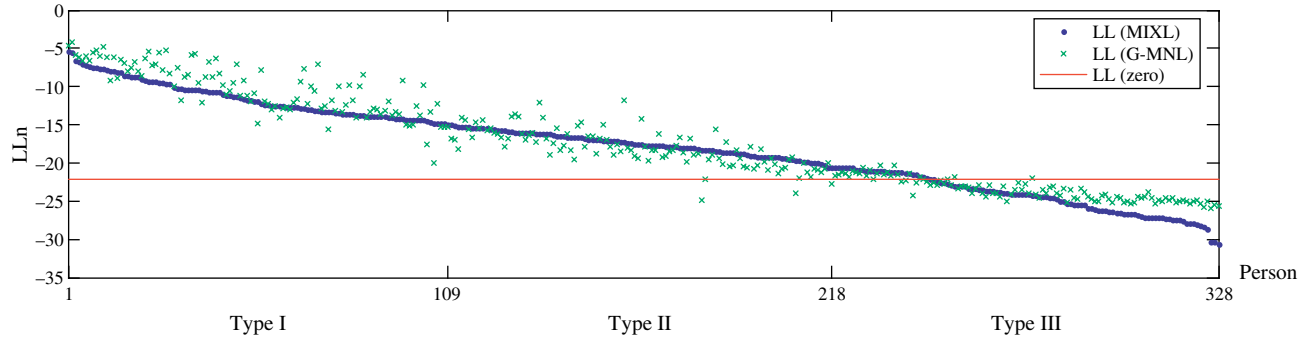
Finally, Figure 5 shows that the pattern shown in Figure 3 emerges not just for fresh ingredients but

[20] Note that, in principle, the MIXL model can also generate some people for whom all attribute weights are small. However, in a model with several attributes, this would be a very unlikely event. This would still be a problem if MIXL used an alternative heterogeneity distribution, such as multivariate t.

[21] Note that here we adopt the mathematical psychology view that choice is random for an individual across choice occasions. In the economist's view, the randomness in choice exists solely from the point of view of the analyst, who does not observe a consumer's preference type or all the relevant product attributes. In this view, a consumer faced with the same choice situation on two occasions should make the same choice. However, this view is hard to reconcile with behaviour in choice experiments where consumers make repeated choices (e.g., 32 in the Pizza B data set). Inevitably one sees cases where, when presented with A versus B, a person chooses A, and then later, in a situation that is identical except that an attribute of A is improved, the person chooses B. Randomness across choice occasions at the individual level is necessary to explain this. Such randomness is present in choice models applied to experimental data whenever one lets the stochastic terms differ across choice occasions.

[22] Allenby and Rossi (1998) call this an "approximate Bayesian" approach. We tried integrating over uncertainty in the estimated model parameters, as well as uncertainty in calculating the posterior means. This made little difference, presumably because (1) the model coefficients are estimated quite precisely, and (2) with 32 observations per person, the posterior means are also estimated rather precisely.

Figure 2 **Individuals' Log-Likelihood from Pizza B Data Set ($N = 328$; $T = 32$; Binary Choice)**



"Extremely persistent" type
22 people—Fresh ingredients only
18 people—Low price only
14 people—Steaming hot only
8 people—Vegetarian only (small overlap)

"More random" type
Utility weights are close to zero

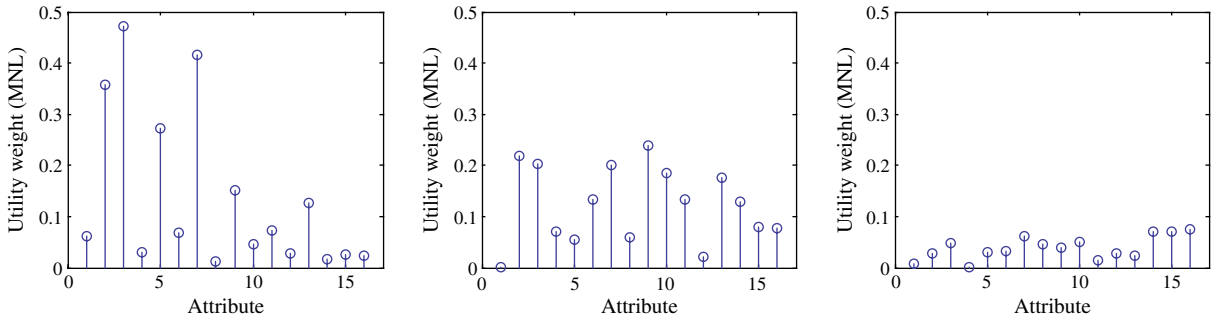


Figure 3 **Predicted Distribution of Probability of Choosing Firm A from MNL, S-MNL, MIXL, and G-MNL Models When Firm A Improves Ingredient Quality and Increases Price by $4**

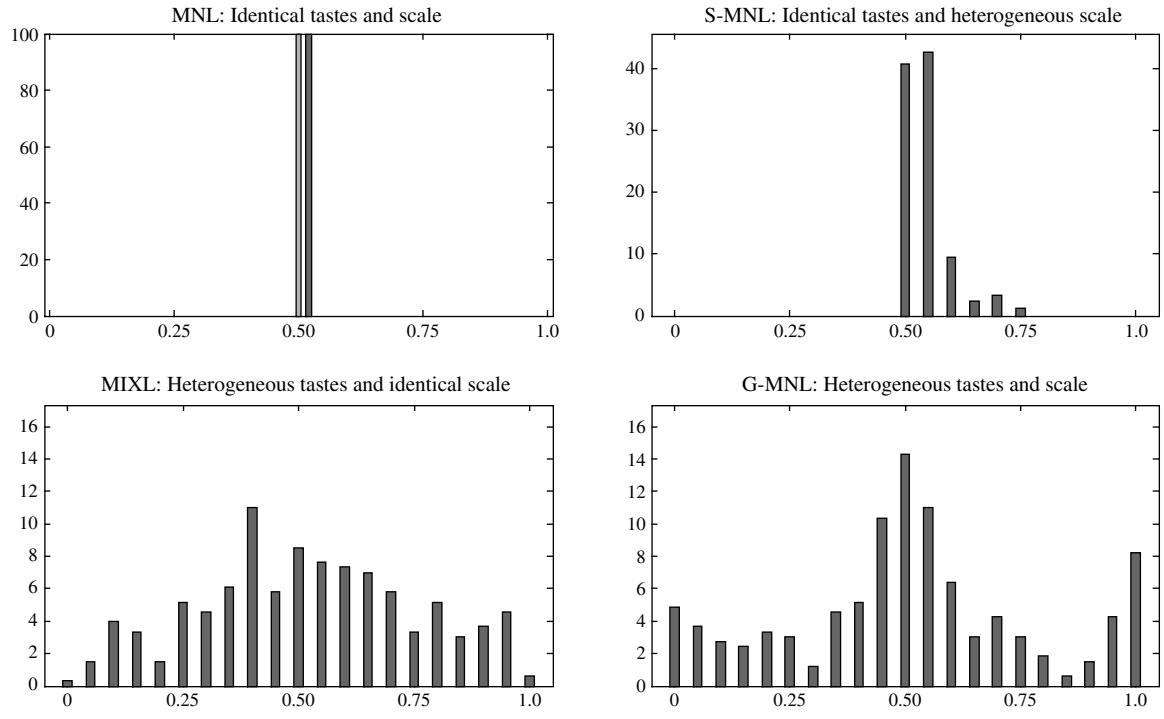


Figure 4 Posterior Distribution of Individual-Level Parameters: MIXL vs. G-MNL

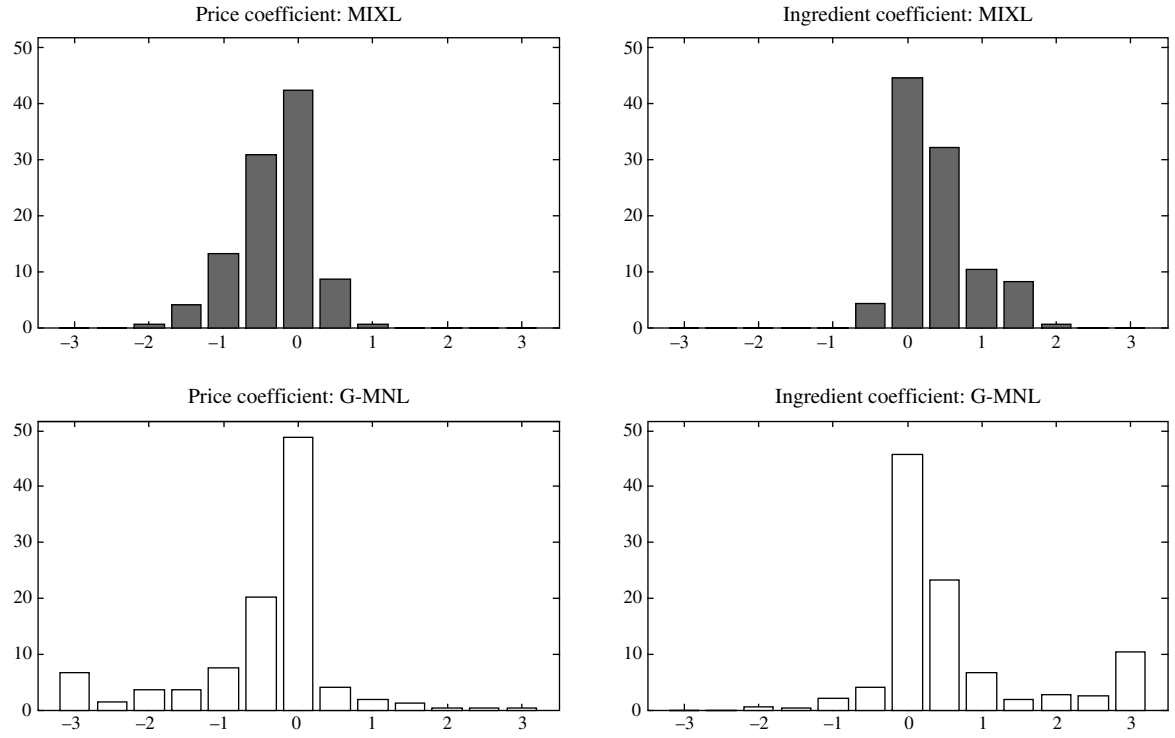
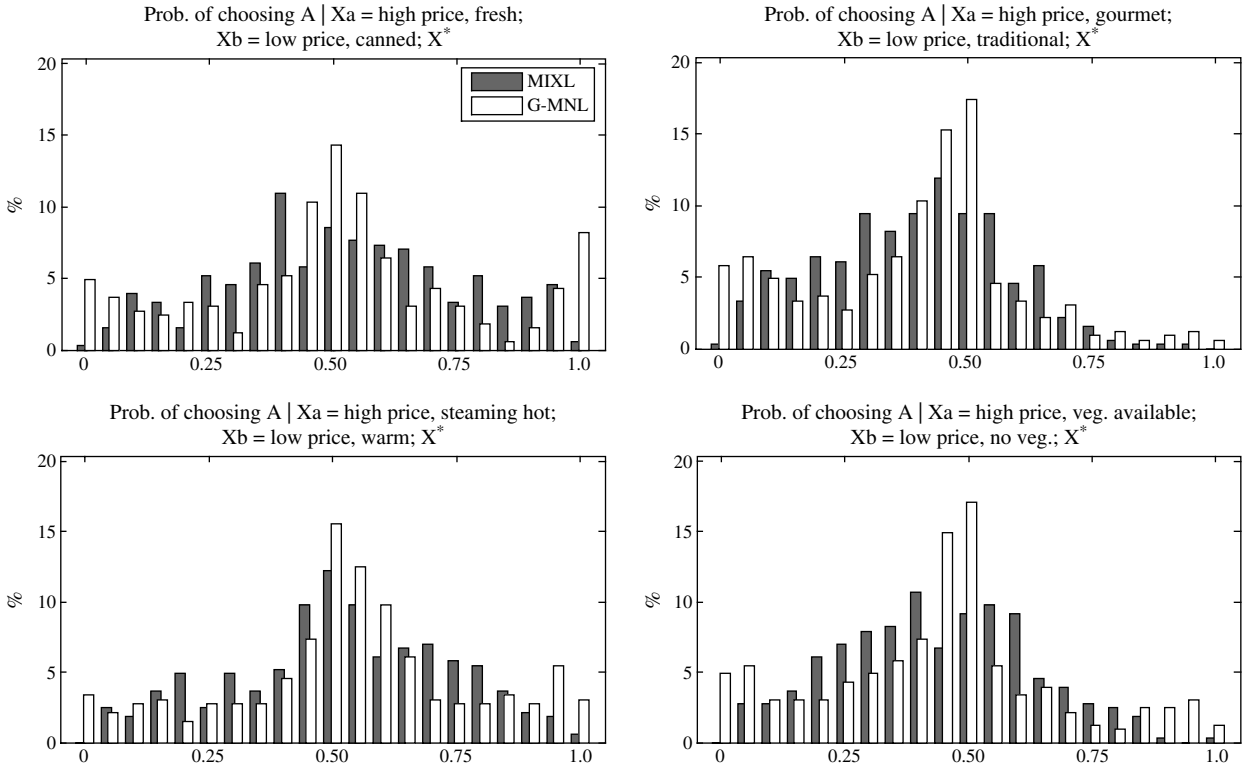


Figure 5 Predicted Distribution of Probability of Choosing Firm A from MIXL and G-MNL Models When Firm A Improves One Attribute and Increases Price by $4



Note. Attribute scenarios clockwise are fresh or canned ingredients, gourmet or traditional, hot or warm, and vegetarian availability.

Table 17 Comparing the Importance of Heterogeneity Across Data Sets

	No. of choices	No. of attributes	No. of occasions	No. of people	MNL	S-MNL	MIXL	G-MNL	Percent improvement: MNL to S-MNL/ MNL to G-MNL	Percent improvement: MNL to G-MNL
1 Tay Sachs disease and cystic fibrosis test—Jewish sample (3 ASCs)	4	11	16	210	−3,717	−3,223[a] −2,815[b]	−2,500	−2,480	40 73	33
2 Tay Sachs disease and cystic fibrosis test—general population sample (3 ASCs)	4	11	16	261	−4,649	−3,567[a] −3,221[b]	−2,946	−2,914	62 82	37
3 Mobile phone (1 ASC)	4	15	8	493	−4,475	−4,102[a] −3,990[b]	−3,962[c]	−3,949[c]	71 92	12
4 Pizza A (no ASC)	2	8	16	178	−1,657	−1,581	−1,379	−1,324	23	20
5 Holiday A (no ASC)	2	8	16	331	−3,066	−2,967	−2,504	−2,469	17	19
6 Pap smear test (1 ASC)	2	6	32	79	−1,528	−1,124[a] −1,063[b]	−923	−914	66 76	40
7 Pizza B (no ASC)	2	16	32	328	−6,747	−6,607	−5,857[c]	−5,668[c]	13	16
8 Holiday B (no ASC)	2	16	32	683	−13,478	−13,027	−11,570[c]	−11,446[c]	22	15
9 Charge card A (2 ASCs)	3	17	4	827	−3,354	−3,217[a] −2,768[b]	−2,735[c]	−2,734[c]	22 95	18
10 Charge card B (3 ASCs)	4	18	4	827	−4,100	−3,947[a] −3,402[b]	−3,364[c]	−3,364[c]	21 95	18

[a]S-MNL with fixed ASCs.
[b]S-MNL with random ASCs.
[c]Imposing one-factor model restriction on variance-covariance matrix.

for other attributes as well. Each panel plots the distribution of consumer choice probabilities under an experiment where one attribute of pizza delivery service A is improved, and price is also increased by \$4. The first panel repeats the fresh ingredients experiment from Figure 3. However, now the probability distributions of the G-MNL and MIXL models are plotted side by side, making the differences easier to see. Clearly, the G-MNL model puts more mass near the center of the distribution of choice probabilities (i.e., close to 50%) and more mass in the tails (close to 0% or 100%). The same basic pattern holds in experiments where firm A offers gourmet pizza, steaming hot pizza, or a vegetarian option.

What are the managerial implications of these results? Unlike the situation in the Pizza B choice experiment, in the real world, pizza delivery firms do not offer a single type of pizza (or a small range of options) at a single price. They offer a wide range of pizzas with different attributes at different prices. To determine the optimal menu of offerings, a firm needs to know the entire distribution of demand. It is beyond the scope of this paper to design optimal menus. However, it is clear that optimal price discrimination strategies would differ between a market where a significant fraction of consumers have essentially lexicographic preferences versus a market where attribute weights differs less markedly across consumers.

In summary, these results make clear why models with scale heterogeneity (G-MNL or S-MNL) fit better than MIXL in every data set we examine. The models with scale heterogeneity are able to generate the sort of extreme (or lexicographic) behaviour that is common in choice experiments, whereas MIXL cannot. They are also able to capture random choice behaviour (i.e., low responsiveness to attribute settings) better than MIXL. The reason for both advantages is transparent. Models that include scale can generate random behaviour by setting the scale large and can generate lexicographic behaviour by setting the scale small (while also letting one attribute have a large idiosyncratic component of its preference weight).

5.4. Comparing the Importance of Heterogeneity Across Data Sets

Table 17 summarizes results across the 10 data sets.[23] One interesting pattern is the extent to which including heterogeneity of all types leads to improvement

[23] The mobile phone, Pizza B, Holiday B, and charge card data sets contained very many attributes (16 to 18), so it was not feasible to estimate a full variance-covariance matrix in these cases. Instead, we restricted them to have a one-factor structure. Such an approach may be worth pursuing as a compromise between the two extreme options commonly applied in practice: imposing no correlation or estimating full variance-covariance matrix.

in model fit. That is, what is the percentage improvement in the log-likelihood when we go from the simple MNL model to the full-fledged G-MNL model? Strikingly, this differs greatly by data set, ranging from only 12% to 16% in the mobile phone, Pizza B, and Holiday B data sets to as much as 33% to 40% in the Tay Sachs and Pap smear data sets. Another metric (not reported in Table 17) is the improvement in pseudo-R^2 when heterogeneity is included. This ranges from 0.27 to 0.35 in the three medical data sets, but from only 0.10 to 0.17 in the other data sets.[24]

Thus, the extent of preference heterogeneity in the three data sets involving medical decisions is roughly twice as great as in those for consumption goods (phones, pizza delivery, holidays, charge cards). There are a number of possible explanations for this pattern. People may have stronger feelings about medical procedures than the more mundane attributes of consumer products. People may have very different attitudes towards risk. Perhaps medical decision making is a more complex or higher involvement task, and taste heterogeneity in general (and perhaps scale heterogeneity in particular) increases with task complexity.[25,26]

A second interesting pattern is how the importance of scale heterogeneity differs across data sets. We cannot calculate the fraction of heterogeneity that the G-MNL model assigns to residual versus scale, because the improvement in the likelihood when these are included is not additive. However, we can get a sense of the importance of scale heterogeneity by asking, Of the total improvement in the likelihood achieved by adding all forms of heterogeneity, what fraction is attained by adding scale heterogeneity alone?[27] Table 17 reports this figure for the S-MNL models with and without random ASCs. The more relevant figure is that for models with fixed ASCs, because the likelihood improvement from adding random intercepts is more appropriately ascribed to residual taste heterogeneity.

Differences in results across data sets are striking. For mobile phones, 71% of the log-likelihood improvement that can be achieved by introducing all heterogeneity is achieved by introducing scale heterogeneity alone. For the three medical tests (Pap smear, Tay Sachs Jewish sample, Tay Sachs general population sample) the values range from 40% to 66%, but in the other data sets scale heterogeneity appears to be less important. In the pizza delivery and holiday destination data sets the fraction of the total log-likelihood improvement that can be achieved just by introducing scale is only 13% to 23%, and in the two charge card data sets, the values are only 21% to 22%.

Another way to gauge the importance of scale heterogeneity is by the improvement in pseudo-R^2 when it is added to the model. This ranges from 0.07 to 0.23 in the medical and mobile phone data sets[28] but from only 0.02 to 0.05 in the other data sets.

Thus, by both metrics, scale heterogeneity appears to be much more important in the medical and mobile phone data sets than for pizza, holidays, and charge cards. What accounts for this contrast? One hypothesis is that scale heterogeneity increases with task complexity. It is intuitive that choices about medical tests are complex, because they involve making decisions about risks and probabilities, which humans have difficultly understanding. Similarly, mobile phones are high-tech goods with attributes like WiFi connectivity, voice commands, USB connections, etc., which consumers may also find difficult to assess. In contrast, attributes of simple consumer goods like pizza and holidays (e.g., thick crust, quality of hotels) may be easier to evaluate. Thus, the results appear consistent with a view that scale heterogeneity is more important in more complex choice contexts.

In summary, we find that two hypotheses seem at least consistent with the observed patterns across data sets. First, heterogeneity (in general) is more important in data sets that involve high-involvement decisions (e.g., medical tests). Second, scale heterogeneity is more important in data sets that involve more complex choice objects (i.e., objects with more complex attributes). Of course, we view both of these hypotheses as merely preliminary, but they suggest interesting avenues for future work.

[24] Pseudo-R^2 for a discrete-choice model is defined as $1-\mathrm{LL(m)}/\mathrm{LL(0)}$, where LL(m) is the log-likelihood of the model, and LL(0) is that of a "null" model that assigns equal probability to each choice.

[25] Medical decisions are also relatively "unfamiliar" tasks compared to choice among common consumer goods. It is plausible that choice in such unfamiliar contexts is more difficult.

[26] It may simply be that taste heterogeneity is more important in data sets with ASCs, but this is contradicted by the mobile phone and credit card data, which contain ASCs but exhibit a relatively low degree of heterogeneity.

[27] Obviously this is not the same as decomposing the overall likelihood improvement in going from MNL to G-MNL into fractions because of residual versus scale heterogeneity, because the log-likelihood improvements from including them are not additive. However, examining incremental likelihood improvements in going from simple to more complex models is in the same spirit as information criteria like BIC. Intuitively, if the large majority of the likelihood improvement that can be obtained from adding both residual and scale heterogeneity to MNL can be achieved by adding scale heterogeneity *alone*, then BIC will tend to prefer the more parsimonious S-MNL model. In making this determination, BIC does not even consider how the overall likelihood improvement in going to the most general model may be decomposed into parts because of residual versus scale heterogeneity.

[28] The low end of this range (0.07) comes from the mobile phone data set, but this improvement appears more substantial when one considers that heterogeneity in general only improves pseudo-R^2 by 0.10 for mobile phones.

Finally, we tried using observed covariates to explain differences in scale across subjects, as in Equation (12), but we had little success and hence do not report the results. Our limited data on subject characteristics did not help to explain scale, nor did our measures of task complexity (number of attributes, number of alternatives, number of attributes that differ among alternatives, number of scenarios). Clearly, more work is needed on this topic.

5.5. Comparing Willingness to Pay Calculations in the MIXL and G-MNL Models

An important issue that arises in choice modelling is the calculation of consumer willingness to pay (WTP) for changes in product attributes. How best to do this in random coefficient models has recently been an active area of research (see, e.g., Sonnier et al. 2007). To understand the issue, consider a general model with heterogeneity in (1) the attribute weights, (2) the price coefficient, and (3) the scale parameter:

$$U_{njt} = \beta_n x_{njt} - \phi_n p_{njt} + \varepsilon_{njt}/\sigma_n$$
$$n = 1, \ldots, N;\ j = 1, \ldots, J;\ t = 1, \ldots, T. \quad (13)$$

This model is not identified, and the most common normalization is, of course, to set the scale parameter $\sigma_n = 1$ for all n. This gives what is called a model in "utility space." However, an alternative is to normalize the price coefficient $\phi_n = 1$ for all n. This gives what is called a model in "WTP space." It is useful to write out the two models explicitly.

$$U_{njt} = \beta_n x_{njt} - \phi_n p_{njt} + \varepsilon_{njt}; \quad \text{"Utility space"}$$
$$U^*_{njt} = \beta^*_n x_{njt} - p_{njt} + \varepsilon_{njt}/\sigma_n. \quad \text{"WTP space"}$$

In the model in Utility space, WTP for an additional unit of attribute k is β_{nk}/ϕ_n, and for the model in WTP space it is simply β^*_n.[29] These two models will give identical fits to the data, and identical estimates of WTP, provided the specification and estimation methods maintain the restrictions that $\beta^*_n = \beta_n/\sigma_n$ and $\phi_n = \sigma_n$.

Practitioners have reported, however, that the two models give very different estimates of WTP, and in particular that estimates obtained in utility space are often unreasonably large. The source of these differences is that, in practice, it is difficult (or inconvenient) to specify the utility space and WTP space models in such a way that they are equivalent.[30] This does not mean, however, that the two models are not equivalent if properly specified.

It is interesting to examine the issue of estimating WTP in the context of the G-MNL and S-MNL models. Rewriting (7) so the price coefficient is explicit, we have

$$U_{njt} = [\sigma_n\beta + \gamma\eta_n + (1-\gamma)\sigma_n\eta_n]x_{njt} - [\sigma_n\phi + \gamma\eta_{\phi n} + (1-\gamma)\sigma_n\eta_{\phi n}]p_{njt} + \varepsilon_{njt}, \quad (14)$$

where ϕ denotes the mean price coefficient in the population. If only scale heterogeneity matters, WTP for a unit of attribute k reduces to just (β_k/ϕ), where β_k denotes the kth element of the β vector. This illustrates a strong property of the S-MNL model—there is heterogeneity in coefficients but not in WTP.[31] However, this does not mean there is no heterogeneity in price sensitivity. For example, in the pure S-MNL model the derivative of the choice probability with respect to price is $\partial P_n(j \mid X_{nt})/\partial p_{njt} = -P_n(j \mid X_{nt})[1 - P_n(j \mid X_{nt})] \cdot \sigma_n \cdot \phi$. Thus, as $\sigma_n \to 0$, only unobserved attributes ε matter for choice, and price sensitivity goes to zero.

This illustrates an odd aspect of the "willingness to pay" concept in choice models. A consumer's WTP for an attribute increase is defined as the price increase which, combined with the attribute increase, leaves the deterministic part of his utility for a brand unchanged—and hence the choice probability unchanged. However, consider the same unit increase in the attribute holding price fixed. Given heterogeneity, consumers with the same WTP for the attribute will not in general have the same increase in their choice probability for the brand (even given the same initial probability). Consumers with larger σ_n in the WTP space model, or larger $\beta_n = \beta^*_n\sigma_n$ in the utility space model, will have a larger increase in demand. In other words, it is perfectly compatible that some consumers have a large WTP for an attribute but that introducing it leads to little increase in their probability of choosing the brand.

[29] Analogously, in contingent valuation data, Hanemann (1984) estimated WTP as the ratio of the intercept (representing the hypothetical program) to the price coefficient, whereas Cameron (1988) uses the expenditure function to estimate WTP directly.

[30] In the utility space model it is common to assume β_n is normal and that the price coefficient ϕ_n is normal or log-normal (to keep it positive). Regardless, the distribution of WTP for attribute k is the ratio (β_{nk}/ϕ_n), where the numerator is normal and the denominator normal or log-normal. In contrast, in the WTP space model, the WTP distribution is simply that of β^*_{nk}, which is typically specified as normal (see, e.g., Sonnier et al. 2007). Hence, it is not surprising that the two models—as typically specified—give very different answers, because there is no reason to expect (β_{nk}/ϕ_n) to be approximately normal. Furthermore, it is not surprising that, in a utility space model, WTP sometimes takes on extreme values; we are taking the ratio of two random variables, where the denominator is normal or log-normal, and the ratio can "explode" because the denominator is close to zero.

[31] This is because people with larger attribute weights have a proportionately larger price coefficient.

Table 18 Willingness to Pay Estimates vs. Aggregate Choice Probabilities

	Percent choosing A when A's attribute changes and charges \$4 more	Percent people with WTP \$4 or more	WTP distribution						
			10th	20th	25th	50th	75th	80th	90th
MIXL									
Traditional to gourmet	39.43	23.17	−Inf	−2.15	−1.27	0.16	2.96	5.85	Inf
Canned to fresh ingredient	52.08	49.70	−3.19	−0.34	0.25	3.69	23.08	53.37	Inf
Warm to steaming hot	52.05	52.13	−1.87	0.33	0.79	4.22	26.08	74.75	Inf
No veg. to veg avail.	43.68	33.23	−6.04	−1.17	−0.46	1.33	9.55	21.41	Inf
G-MNL									
Traditional to gourmet	38.98	20.12	−8.66	−1.36	−0.78	0.31	2.39	4.10	Inf
Canned to fresh ingredient	50.80	48.48	−1.38	0.25	0.64	3.80	29.50	96.80	Inf
Warm to steaming hot	53.25	61.59	0.55	1.41	1.85	6.23	39.57	1,308.71	Inf
No veg. to veg avail.	43.50	30.79	−8.77	−1.24	−0.85	1.06	6.40	11.57	Inf

In general, WTP in the G-MNL model is given by the ratio

$$[\sigma_n\beta_k+\gamma\eta_{kn}+(1-\gamma)\sigma_n\eta_{kn}]/[\sigma_n\phi+\gamma\eta_{\phi n}+(1-\gamma)\sigma_n\eta_{\phi n}].$$

Although seemingly complicated, this is no more difficult to simulate than (β_{nk}/ϕ_n) in the MIXL model. To guarantee "reasonable" WTP estimates one must choose distributions for σ_n and $\eta_{\phi n}$ so that the price coefficient $[\sigma_n\phi + \gamma\eta_{\phi n} + (1 - \gamma)\sigma_n\eta_{\phi n}]$ is bounded away from zero. However, in light of our previous comments, we argue that WTP calculations are overemphasized and that more emphasis should be placed on simulating demand. This will become clear below.

In Table 18, we compare demand and WTP predictions of the G-MNL and MIXL models, again focusing on the Pizza B data set. The top and bottom panels report results for MIXL and G-MNL, respectively. In the first row we consider an experiment where delivery service A switches from traditional to gourmet pizza while raising the price by \$4 (holding other attributes equal between the two services). Both MIXL and G-MNL predict that roughly 39% of consumers would choose service A under this experiment (as opposed to 50% under the baseline where all attributes are equal). Thus, the demand curves generated by both models imply that roughly 39% of consumers are willing to pay \$4 extra for gourmet pizza.

A similar pattern holds when we look (see the next three rows of each panel) at the demand predictions for fresh ingredients, guaranteed hot pizza, and vegetarian pizza. In each case, the demand predictions from G-MNL and MIXL are almost identical. This pattern holds across all data sets and a wide range of prediction scenarios: aggregate demand predictions from G-MNL and MIXL are almost identical. Recall that the key difference between the two models arises not from their predictions about aggregate demand but in their predictions about the distribution of demand across individual types of people (see §5.3).

We turn next to the distribution of WTP implied by each model. For both G-MNL and MIXL this is simulated in the conventional way, as described above. At the 50th percentile, the WTP for gourmet pizza is close to zero according to both models. At the 75th percentile it is about \$3 according to MIXL and \$2.40 according to G-MNL. Thus, given the simulated WTP distributions, both models imply less than 25% of consumers are willing to pay \$4 extra for gourmet pizza. Here, we see immediately how WTP distributions generated by both models (calculated in the conventional way) seriously contradict the demand predictions, as we have already seen that both models predict that roughly 39% of consumers would be willing to pay \$4 extra for the gourmet pizza. Indeed, based on the overall WTP distribution, the MIXL model implies that only 23% of consumers would buy the gourmet pizza at a \$4 price premium, and the G-MNL model implies that only 20% of consumers would do so.

6. Conclusion

Consumer taste heterogeneity is of central importance for many issues in marketing and economics. For at least 25 years there has been a large ongoing research program on how best to model heterogeneity. This research program has produced a large number of alternative modelling approaches. One of the most popular is the MIXL model. In most applications of MIXL, the vector of consumer utility weights is assumed to have a multivariate normal distribution in the population. However, Louviere et al. (1999, 2008b) have recently argued, based on estimation of individual-level models, that much of the heterogeneity in attribute weights is better described as a pure scale effect (i.e., across consumers, weights on all attributes are scaled up or down in tandem). This implies that choice behaviour is simply more random for some consumers than others (i.e., holding attribute coefficients fixed, the scale of their error term

is greater). This leads to what we have called an S-MNL model.

In this paper we have developed a G-MNL that nests S-MNL and MIXL. By estimating the G-MNL model on 10 data sets, we provide empirical evidence on the importance of scale heterogeneity and on the relative ability of the MIXL, S-MNL, and G-MNL models to fit the data. Our main results show that, based on BIC and CAIC, the G-MNL model is preferred in seven data sets, whereas the S-MNL model is preferred in the other three. This is striking evidence of the importance of scale heterogeneity—and of the ability of models that include scale heterogeneity to outperform MIXL.

We also show why G-MNL fits better than MIXL. Specifically, it can better explain the behaviour of extreme consumers who exhibit near lexicographic preferences (i.e., consumers who nearly always choose the option with a particular attribute, such as lowest price or highest quality, regardless of the attributes of other alternatives). G-MNL is also better able to explain highly random consumers whose choices are relatively insensitive to product attributes (i.e., consumer with a large scale of the idiosyncratic error terms).

We went on to show that the G-MNL model allows more flexibility in the posterior distribution of individual-level parameters than does MIXL. From an "approximate Bayesian" perspective, the MIXL model with a normal heterogeneity distribution imposes a normal prior on the distribution of individual-level parameters. However, G-MNL imposes a much more flexible continuous mixture of scaled normals prior. Thus, even given a large amount of data per person, MIXL posteriors depart only modestly from normality, but G-MNL posteriors exhibit sharp departures. These include both multimodality with spikes in the tails (people who care greatly about particular attributes) and excess kurtosis (people who have small attribute weights or, conversely, a large scale of the error term).

An important avenue for future research is to compare G-MNL with alternative models that also allow a more flexible distribution of individual-level parameters, such as mixture of normals logit and probit models. The potential advantage of G-MNL is that it achieves a flexible distribution while adding only two parameters to the normal model.

Our analysis also yielded two interesting empirical findings. First, taste heterogeneity in general was far more important for medical decisions than for consumer goods. Second, scale heterogeneity was more important in data sets that involve more complex choice objects. Of course, these empirical findings are quite preliminary because they involve only 10 data sets.

Acknowledgments

The second author's work on this project was supported by Australian Research Council Grant FF0561843 and the first author's by National Health and Medical Research Council Program Grant 25402.

References

Allenby, G. M., P. Rossi. 1998. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(1–2) 57–78.

Bartels, R., D. G. Fiebig, A. van Soest. 2006. Consumers and experts: An econometric analysis of the demand for water heaters. *Empirical Econom.* **31**(2) 369–391.

Ben-Akiva, M., D. McFadden, M. Abe, U. Böckenholt, D. Bolduc, D. Gopinath, T. Morikawa et al. 1997. Modelling methods for discrete choice analysis. *Marketing Lett.* **8**(3) 273–286.

Burda, M., M. Harding, J. Hausman. 2008. A Bayesian mixed logit-probit model for multinomial choice. *J. Econometrics* **147**(2) 232–246.

Cameron, T. A. 1988. A new paradigm for valuing non-market goods using referendum data: Maximum likelihood estimation by censored logistic regression. *J. Environ. Econom. Management* **15**(3) 335–379.

Cameron, T. A., G. L. Poe, R. G. Ethier, W. D. Schulze. 2002. Alternative non-market value-elicitation methods: Are the underlying preferences the same? *J. Environ. Econom. Management* **44**(3) 391–425.

DeShazo, J. R., G. Fermo. 2002. Designing choice sets for stated preference methods: The effects of complexity on choice consistency. *J. Environ. Econom. Management* **44**(1) 123–143.

Dubé, J.-P., P. Chintagunta, A. Petrin, B. Bronnenberg, R. Goettler, P. B. Seetharaman, K. Sudhir, R. Thomadsen, Y. Zhao. 2002. Structural applications of the discrete choice model. *Marketing Lett.* **13**(3) 207–220.

Elrod, T., M. P. Keane. 1995. A factor-analytic probit model for representing the market structure in panel data. *J. Marketing Res.* **32**(1) 1–16.

Fang, H., M. Keane, D. Silverman. 2006. Advantageous selection in the Medigap insurance market. *J. Political Econom.* **116**(2) 303–350.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**(2) 209–230.

Fiebig, D. G., J. P. Hall. 2005. Discrete choice experiments in the analysis of health policy. *Productivity Commission Conf., November 2004: Quant. Tools for Microeconomic Policy Anal., Canberra, ACT, Australia*, 119–136.

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

Geweke, J., M. Keane. 2001. Computationally intensive methods for integration in econometrics. J. J. Heckman, E. E. Leamer, eds. *Handbook of Econometrics*, Vol. 5. Elsevier Science B.V., Amsterdam, 3463–3568.

Geweke, J., M. Keane. 2007. Smoothly mixing regressions. *J. Econometrics* **138**(1) 291–311.

Geweke, J., M. Keane, D. Runkle. 1994. Alternative computational approaches to statistical inference in the multinomial probit model. *Rev. Econom. Statist.* **76**(4) 609–632.

Geweke, J. F., M. P. Keane, D. E. Runkle. 1997. Statistical inference in the multinomial multiperiod probit model. *J. Econometrics* **80**(1) 125–165.

Hall, J. P., D. G. Fiebig, M. T. King, I. Hossain, J. J. Louviere. 2006. What influences participation in genetic carrier testing? Results from a discrete choice experiment. *J. Health Econom.* **25**(3) 520–537.

Hanemann, W. M. 1984. Welfare evaluations in contingent valuation experiments with discrete responses. *Amer. J. Agricultural Econom.* **66**(15) 332–341.

Harris, K., M. Keane. 1999. A model of health plan choice: Inferring preferences and perceptions from a combination of revealed preference and attitudinal data. *J. Econometrics* **89**(1–2) 131–157.

Jiang, W., M. Tanner. 1999. Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Ann. Statist.* **27**(3) 987–1011.

Kamakura, W. A., G. J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* **26**(4) 379–390.

Keane, M. P. 1994. A computationally practical simulation estimator for panel data. *Econometrica* **62**(1) 95–116.

Keane, M. P. 1997a. Current issues in discrete choice modelling. *Marketing Lett.* **8**(3) 307–322.

Keane, M. P. 1997b. Modeling heterogeneity and state dependence in consumer choice behaviour. *J. Bus. Econom. Statist.* **15**(3) 310–327.

Keane, M. 2006. The generalized logit model: Preliminary ideas on a research program. Presentation, Motorola-CenSoC Hong Kong Meeting, October 22, Motorola, Hung Hom, Kowloon, Hong Kong.

Keane, M., R. Moffitt. 1998. A structural model of multiple welfare program participation and labor supply. *Internat. Econom. Rev.* **39**(3) 553–589.

Louviere, J. J., T. Eagle. 2006. Confound it! That pesky little scale constant messes up our convenient assumptions! *Proc. 2006 Sawtooth Software Conf.* Sawtooth Software, Sequem, WA, 211–228.

Louviere, J. J., R. J. Meyer. 2007. Formal choice models of informal choices: What choice modeling research can (and can't) learn from behavioral theory. N. K. Malhotra, ed. *Review of Marketing Research*. M. E. Sharpe, New York, 3–32.

Louviere, J. J., T. Islam, N. Wasi, D. J. Street, L. Burgess. 2008a. Designing discrete choice experiments: Do optimal designs come at a price? *J. Consumer Res.* **35**(2) 360–375.

Louviere, J. J., D. Street, L. Burgess, N. Wasi, T. Islam, A. A. J. Marley. 2008b. Modelling the choices of individuals decision makers by combining efficient choice experiment designs with extra preference information. *J. Choice Model.* **1**(1) 128–163.

Louviere, J. J., R. J. Meyer, D. S. Bunch, R. Carson, B. Dellaert, W. M. Hanemann, D. Hensher, J. Irwin. 1999. Combining sources of preference data for modelling complex decision processes. *Marketing Lett.* **10**(3) 205–217.

Louviere, J. J., D. J. Street, R. T. Carson, A. Ainslie, J. R. DeShazo, T. A. Cameron, D. Hensher, R. Kohn, T. Marley. 2002. Dissecting the random component of utility. *Marketing Lett.* **13**(3) 177–193.

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

McFadden, D., K. Train. 2000. Mixed MNL models for discrete response. *J. Appl. Econometrics* **15**(5) 447–470.

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

Small, K. A., C. Winston, J. Yan. 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* **73**(4) 1367–1382.

Sonnier, G., A. Ainslie, T. Otter. 2007. Heterogeneity distributions of willingness-to-pay in choice models. *Quant. Marketing Econom.* **5**(3) 313–331.

Swait, J., W. Adamowicz. 2001. Incorporating the effect of choice environment and complexity into random utility models. *Organ. Behav. Human Decision Processes* **86**(2) 141–167.

Thurstone, L. 1927. A law of comparative judgment. *Psych. Rev.* **34** 273–286.

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

Train, K. 2007. A recursive estimator for random coefficient models. Working paper, University of California Berkeley, Berkeley.

Villani, M., R. Kohn, P. Giordani. 2007. Nonparametric regression density estimation using smoothly varying normal mixtures. Sveriges Riksbank Research Paper Series 211, Stockholm.

Wedel, M., W. Kamakura. 1998. *Market Segmentation: Concepts and Methodological Foundations*. Kluwer Academic Publishers, Boston.