How to Model Consumer Heterogeneity?
Lessons from Three Case Studies on SP and RP Data

by

Michael P. Keane
Department of Economics, University of Oxford

Nada Wasi
Survey Research Center and Institute for Social Research, University of Michigan

February 24, 2016

Abstract: The structure of consumer taste heterogeneity in discrete choice demand models is important, as it drives the structure of own and cross-price elasticities of demand, and the pattern of competition between products. Here we compare performance of three leading discrete choice models, using three datasets with very different properties. The models are the mixed logit with normal heterogeneity (N-MIXL), the generalized multinomial logit (G-MNL) and the mixture-of-normals logit (MM-MNL). Which model is preferred depends on the context: G-MNL does an excellent job of capturing the sort of departures from normality that are prevalent in stated preference (SP) data. But MM-MNL can capture more general departures from normality that are prevalent in revealed preference (RP) data. The finding that the structure of consumer taste heterogeneity is very different in SP vs. RP data suggests that caution should be applied before using SP to answer questions about the distribution of taste heterogeneity in actual markets. In an application to RP data on demand for frozen pizza, we obtain the interesting result that when a variety of a brand raises its price, most of the lost market share goes to other brands (rather than alternative varieties of the same brand). This suggests modeling heterogeneity in tastes for varieties is quite important for understanding brand switching.

JEL Codes: D12, C35, C33, C91, M31

Keywords: Discrete choice models, Consumer demand, Consumer heterogeneity, Mixture models, Mixed Logit, Generalized multinomial logit, Discrete choice experiments, Stated preference data, Very large choice sets, Internal analysis of market structure

I. Introduction

The question of how best to model consumer preference heterogeneity in discrete choice demand analysis has been a subject of great interest (and controversy) for at least 40 years. The reason the issue is so important is that the taste heterogeneity distribution drives the pattern of competition between products, as well as the structure of own and cross-price elasticities of demand. These features of a market matter for two main reasons:

(1) In deciding whether to introduce a new product, a firm must predict both the market share of the product and where it will come from (e.g., will the new product steal customers away from other firms, or just from the firm's own existing products?). The same question arises with respect to the change in market share if a firm changes the price of an existing product;

(2) In deciding whether to allow a merger of two firms, an antitrust regulator is concerned with the degree of competition between their products. If their products have high cross-price elasticities of demand, a merger may generate large price increases in equilibrium.

Unfortunately, the discrete choice model that has traditionally been the most widely used, the simple multinomial logit (MNL) with a common coefficient vector for all consumers (i.e., no heterogeneity), cannot address issues like (1) and (2). This is because MNL imposes *a priori* that all market shares move proportionately when a new product is introduced in a market (or when an existing product changes its price). This is obviously unrealistic.

For example, if Toyota introduces a new small economy car, we would expect it to compete with other small economy cars, and have little impact on family cars or luxury cars. But MNL makes the unrealistic prediction that the new car will steal market share proportionally from all alternatives. This problem can be solved by developing models where consumers are *heterogeneous* in their tastes for product attributes (such as fuel economy, seating capacity, performance, etc.). This allows cross-price elasticities to be determined flexibly by the data.

Since about 2000, it would be fair to say that the most popular discrete choice demand model among sophisticated practitioners and academics has been the so-called mixed logit model (MIXL). This is an extension of MNL that allows for population heterogeneity in the parameter vector. In most applications it has been assumed to be normally distributed in the population, leading to the logit model with normal mixing, or "N-MIXL" model. Early papers that developed and advocated this approach were Berry (1994), Berry et al (1995), Revelt and Train (1998), Harris and Keane (1999) and McFadden and Train (2000). N-MIXL is popular because it allows for much more flexible patterns of substitution among alternatives than MNL, yet it is quite simple to estimate using simulation methods, and it is available in standard packages.

Another branch of the literature advocates the use of discrete mixtures-of-normals distributions to model consumer heterogeneity. In this model, the population is assumed to consist of a discrete set of types, each with its own normal heterogeneity distribution. Keane and Wasi (2013) refer to a mixed logit model where the heterogeneity distribution is specified as a discrete mixture-of-normals as the "mixed-mixed-logit" or MM-MNL model.

The virtue of the discrete mixtures-of-normals approach is both theoretical and practical. Theoretically, as shown by Ferguson (1973), a discrete mixture-of-normals can approximate any heterogeneity distribution arbitrarily closely. Practically, as shown in Geweke and Keane (1997, 2001) and Rossi, Allenby and McCulloch (2000), just a small number of mixture components (i.e., 2 or 3) can usually approximate even very complex heterogeneity distributions quite well.

Despite these appealing features, the discrete mixture-of-normals approach to modelling heterogeneity has not achieved anything close to the popularity of the normal-mixture-of-logits (N-MIXL) approach. We suspect this is because discrete mixture-of-normals models are more complex to estimate, tend to proliferate parameters, and require judgement in choosing the number of mixture elements. For these reasons, for better or worse, they have remained largely the preserve of advanced econometricians.

In a recent contribution, Feibig, Keane, Louviere and Wasi (2010) proposed a new model called the generalized multinomial logit or G-MNL model. This is a mixed logit model where the heterogeneity distribution is specified as a scaled mixture-of-normals. That is, the multivariate normal coefficient vector of the N-MIXL model is scaled by a positive scalar random variable, which Feibig et al (2010) assume is log normally distributed in the population. Of course, in the logit model, a scaling of the entire coefficient vector is observationally equivalent to an inverse scaling the idiosyncratic errors. Thus, one interpretation of the G-MNL model is that the idiosyncratic errors (or taste shocks) are more important (relative to the observed attributes) for some consumers than others.[1] This is often called "scale heterogeneity."

[1] Feibig et al (2010) interpret their model as incorporating both "parameter" and "scale" heterogeneity, although some may prefer to interpret the scaled mixture-of-normals as simply another parametric heterogeneity distribution, and not adopt this behavioral interpretation.

The Fiebig et al (2010) paper was influential, and the G-MNL model became very popular among practitioners almost as soon as it was introduced.[2] There are a number of reasons for this sudden popularity of G-MNL: First, the model nests N-MIXL, reducing to that model as the variance of the scale heterogeneity parameter goes to zero. Second, the model adds only two parameters to the N-MIXL model (as we discuss below), and it is hardly any more difficult to estimate.[3] Third, Fiebig et al (2010) showed in several applications that the G-MNL model usually gave a much better fit to consumer choice behavior than the N-MIXL model. If a model is no more difficult to estimate than its main competitor, is strictly more general, and often provides a substantially better fit, it is hard to see why it wouldn't be popular.

Keane and Wasi (2013) provided further support for the G-MNL model by comparing it to the theoretically more general discrete mixture-of-normals approach on ten example data sets. We found that in most instances G-MNL is preferred over MM-MNL by standard metrics of model fit that penalize proliferation of parameters. This is because the G-MNL model typically gives almost as good a fit as MM-MNL, but it does so with many fewer parameters.

The point of the present paper is to provide some guidance as to the typical context in which each of these three models – N-MIXL, G-MNL or MM-MNL – is the preferred approach to modelling heterogeneity. We present three case studies, and assess which model performs best in each case. Much more importantly, we explain why each model emerges as preferred in a particular context. Our results can be viewed as tempering the conclusions of Fiebig et al (2010) and Keane and Wasi (2013). That is, two of the case studies illustrate contexts where the MM-MNL model substantially outperforms N-MIXL and G-MNL, making the greater generality of the mixture-of-normals approach to modeling heterogeneity worth the extra computational effort. We also provide a substantive analysis of the different patterns of own and cross-price elasticities implied by the alternative models, using a large panel data set on demand for frozen pizza.

The outline of the paper is as follows: Section II describes the models of consumer taste heterogeneity that we consider, while Section III describes the three data sets that we examine. In Section IV we present our empirical results, and evaluate which models provide the best in each

[2] For example, Shugan (2014) ranked Fiebig et al (2010) as the 3rd most cited marketing article of the current decade, and the most cited methodological article. See *Shugan's Top 20 Marketing Meta-Journal*, April 2014, at: http://bear.warrington.ufl.edu/centers/MKS/.

[3] This contrasts with discrete mixture-of-normal models, where proliferation of parameters and choice of the number of elements of the mixture are both serious issues for the practitioner.

data set. Section V asks why each model fits best in each case. Specifically, we examine how patterns of consumer behavior differ in each of the three data sets, and assess which model(s) can best capture those patterns. Section VI presents our analysis of the different patterns of own and cross-price elasticities implied by the alternative models, thus returning to the substantive issues discussed at the outset. Section VII concludes. Finally, the Appendix contains interesting results on computational methods for estimating models with very large choice sets, which arise in our third dataset set. These results are of independent interest form many problems in economics.

II. Alternative Models of Consumer Heterogeneity

We begin by describing some general features of discrete choice demand models, focusing on the case of frequently purchased consumer goods. We then turn to the specific models of consumer heterogeneity that we compare.

It is common to treat the good under study as an inside good, and to group all other goods into a composite outside commodity. The budget constraint is $C = I - p_j d_j$ where C is consumption of the outside good, I is income, d_j is an indicator for purchase of discrete type/variety j of the inside good, and p_j is its price. We have j=1,…J, where J is the size of the choice set. Price of the outside good is normalized to one.

Frequently purchased consumer goods are relatively inexpensive. So it is common (and sensible) to assume the marginal utility of the outside good, which we denote as u_c, is constant over the (small) range of consumption levels C generated by different levels of spending on the inside good. This implies the utility function is <u>linear</u> in C over the relevant range. Let $u(j|p_j, I)$ be the indirect utility conditional on purchase of option j. Then $u(j|p_j, I) = D(j) + (u_c)(I - p_j)$ where $D(j)$ is the direct utility from good j.

It is standard to assume $D(j)$ depends on both observed attributes and a stochastic component unobserved by the econometrician. For instance, let $D(j) = \Gamma A_j + \varepsilon_j/\sigma$, where A_j is a vector of observed attributes of good j, with associated vector Γ of utility weights. The stochastic term ε_j captures unobserved attributes, and the scalar $(1/\sigma)$ captures their utility weight. This specification is consistent with the attributes-based approach of Lancaster (1966) and the random utility model (RUM) of McFadden (1974). Note that choice is deterministic for consumers in the RUM, as they see $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_J)$, but choice appears random to an outside observer.

Given this form for *D*(*j*), we have $u(j|p_j, I) = \Gamma A_j + \varepsilon_j/\sigma + (u_c)(I - p_j)$. The scale of the unobserved attribute ε_j is cannot be identified separately from the utility weight (1/*σ*), so it is common to assume ε_j is a *standard* random variable, often standard normal or standard extreme value. Thus its scale is subsumed in (1/*σ*).

Two key features of discrete choice are: (1) only utility differences determine choice and (2) the scale of utility is not identified. Fact (1) implies income *I* is irrelevant, as it is the same for all choices. Fact (2) implies scale normalization is needed. This is usually achieved by setting the scale of the errors to one (i.e., setting *σ*=1), which is equivalent to multiplying $u(j|p_j, I)$ through by *σ*. So define $U(j|p_j) = \sigma u(j|p_j, I = 0)$. Then we have the normalized utility function:

$$U(j|p_j) = (\sigma\Gamma)A_j - (\sigma u_c)p_j + \varepsilon_j \qquad j = 1, \dots, J \tag{1}$$

From (1) we see that the attribute weights Γ and the price coefficient $(-u_c)$ are only identified up to the scale factor *σ*.[4] Thus, the price coefficient may differ across different product categories or choice contexts, even with u_c constant, simply due to differences in scale. Similarly, even under Lancaster's attribute-based view of utility, common attribute coefficients may differ across goods due to differences in scale. But measures of willingness to pay for attributes, given by ratios of attribute coefficients to the price coefficient -Γ/u_c, are scale invariant. This makes it theoretically possible to compare WTP measures for the same attribute obtained by studying demand for different goods or services.[5]

We have argued that, for inexpensive goods, u_c is invariant to the option *j* that is chosen. Some studies allow u_c to differ by *j*, but this is hard for economic theory to rationalize given inexpensive or similarly priced goods. Also, as Keane (1992) noted, an exclusion restriction is needed to identify error correlations in a discrete choice model, and the restriction $u_c = \alpha \; \forall \; j$, where α is a constant, is a very natural way to achieve it. It is also worth emphasizing that it is similarity in price, not in the options *per se*, that implies a constant u_c. For instance, such an assumption is regularly invoked both in models where agents chose among different brands of a consumer product (e.g., pizza), and in applications where they choose among very different

[4] As we noted, the scale factor subsumes three things: the variability of the unobserved attributes, the scale of the unobserved attributes, and the utility weights on the unobserved attributes. It is positive without loss of generality.

[5] Of course, if *D*(*j*) is not linear in attributes then WTP may be harder to calculate, but invariance still holds.

goods, such as mode of transport (e.g., bus, train, car).[6] Of course, we would expect u_c to vary across people, generating heterogeneity in the price coefficient (as we discuss below).

With this background, we turn to specific choice models. To keep notation more compact we define $x_j = (A_j , p_j)$ and $\beta = \{\sigma\Gamma , -\sigma u_c\}$. We also introduce subscripts for person, *n*, and for time, or choice scenario, *t*. Later we will consider applications of discrete choice models both to revealed preference (RP) data and to stated preference (SP) choice experiments. In RP data we often see multiple purchase occasions per person, while in SP data we often have multiple experiments per subject. Both are captured by *t*. Thus, we write:

$$U_{njt} = \beta x_{njt} + \varepsilon_{njt} \qquad n = 1, \dots, N; \quad j = 1, \dots, J; \quad t = 1, \dots, T \tag{2}$$

The multinomial logit (MNL) model of McFadden (1974) is obtained by assuming the ε_{njt} are standard type I extreme value distributed (i.e., Gumbel), and *iid* across choices and over time. MNL was the primary basis for analysis of multinomial choice for many years, largely due to its computational simplicity. In particular, MNL gives simple closed form expressions for choice probabilities of the form $P(j|X_{nt}) = \exp(\beta x_{njt}) / \sum_{k=1}^{J} \exp(\beta\, x_{nkt})$, where $X_{nt} \equiv \{x_{n1t},\ldots,x_{nJt}\}$.

Unfortunately, the MNL assumptions of (i) homogeneous tastes for observed attributes, and (ii) an *iid* random component of utility, rule out some phenomena that are clearly present in the data, like strong "loyalty" to particular brands, or substitution patterns that imply some goods are more similar than others (in terms of unobserved attributes).

In the past 25 years a number of alternative models that extend MNL to allow for taste heterogeneity have been proposed. We examine several of the most important models here. All models we consider can be written in the following form: The utility to person *n* from choosing alternative *j* on purchase occasion (or choice scenario) *t* is given by:

$$U_{njt} = \beta_n x_{njt} + \varepsilon_{njt} \qquad n = 1, \dots, N; \quad j = 1, \dots, J; \quad t = 1, \dots, T \tag{3}$$

The only difference between (2) and (3) is that β_n is a vector of person *n* specific coefficients.

The model in (3), where the coefficient vector β_n is heterogeneous in the population, is known as the mixed logit (MIXL) model. As noted by McFadden and Train (2000), given proper

[6] A point worth noting is that that price elasticity of demand for a variety *j* of a good is not determined by the price coefficient alone, but also by how similar the good is to other goods in the attribute space.

choice of the mixing distribution the MIXL family nests (or can approximate) all random utility models. For example, if β_n is multivariate normal, and if the x_{njt} vector is specified to include alternative specific constants (ASCs), the MIXL model can approximate multinomial probit.[7]

It is worth commenting on interpretation of ASCs. If the alternatives are different brands, we can rationalize ASCs in a Lancaster framework if we view "brand" as a product attribute. But this seems artificial. Preferably, we may assume brands have different mean levels of *unobserved* attributes. Under this interpretation, the random coefficients on the ASCs capture heterogeneity in tastes for the unobserved attributes of each brand. In RP data there are often many more varieties than brands. Then it is natural (and practical) to use brand intercepts rather than ASCs.

In Section IV we will compare the fit of a range of different MIXL models to both RP and SP data. Of course, we cannot consider all possible MIXL models, so we limit ourselves to five that have been particularly important in the literature. The models are distinguished by different specifications of the mixing distribution (β_n). We now describe these five models:

We begin by considering models consistent with the structure in (1), which we now rewrite to include heterogeneity in tastes (*n* subscripts), and a time dimension (*t* subscripts):

$$U_{njt} = (\sigma_n \Gamma_\mathrm{n}) A_j - (\sigma_n u_{cn}) p_{njt} + \varepsilon_{njt} \qquad n = 1, \dots, N; \quad j = 1, \dots, J; \quad t = 1, \dots, T \qquad (4)$$

It is useful to define $\Gamma_n^* = \{\Gamma_n, -u_{cn}\}$ so that β_n = $\sigma_n \Gamma_n^*$. That is, Γ_n^* is an extended vector of attribute preference weights that also includes the marginal utility of consumption (i.e., the negative of the preference weight on the outside good). Then it is clear that (4) is a special case of (3), where β_n has a structure implied by the attribute-based approach.

It is clear from (4) that heterogeneity in the β_n vector may come from three sources. First is heterogeneity in the attribute weights Γ. Second is heterogeneity in the marginal utility of consumption, u_c. Third is heterogeneity in the scale factor, σ, where, in the attribute-based view, $(1/\sigma)$ is the utility weight on unobserved attributes.

We now consider five well-known discrete choice models that can be derived from (4):

1) First is the logit with normal mixing (**N-MIXL**), where β_n is distributed multivariate

[7] Intuitively, if we scale up the normal β_n vector, increasing both the means and variances of the normals, the type 1 extreme value errors become irrelevant to choice, and we approach the probit model. This only works if x_{njt} includes a vector of ASCs, as the elements of the normal β_n vector that multiply the ASCs play the role of the probit errors.

normal N(0,Σ) in the population. This model can be obtained from (4) by assuming σ_n=σ=1 for all n, and that Γ_n^* is distributed multivariate normal. As we noted earlier, N-MIXL can approximate MNP if ASCs are included, and it is very popular in applications. Some papers constrain the price coefficient to be *log*-normal, to enforce the theoretical sign constraint.

2) Next is the scale heterogeneity logit model (**S-MNL**) proposed in Fiebig et al (2010). This can be derived from (4) by assuming that all heterogeneity is in σ, while Γ and u_c are homogeneous in the population. Thus, we have that β_n = $\{\sigma_n\Gamma\,, -\sigma_n u_c\}$, where σ_n is a positive scalar that shifts the whole β vector up or down. The motivation for the S-MNL specification is work by Louviere et al (1999, 2002) and Meyer and Louviere (2007) that finds that much of the heterogeneity in discrete models (at least in SP data) takes the form of scale heterogeneity.

We assume σ_n has a lognormal distribution, $\ln(\sigma_n) \sim N(\bar{\sigma}, \tau^2)$. This assures σ_n >0. We also normalize E(σ_n)=1 for identification. That is, we estimate only β and τ. We *calibrate* $\bar{\sigma}$ so that E(σ_n)=1. Thus β is interpretable as the mean vector of the random preference weights β_n.

3) The third model is the generalized multinomial logit (**G-MNL**) model developed by Fiebig et al (2010). This model nests S-MNL and N-MIXL. Specifically, in (4) we assume σ_n is log-normal as in S-MNL, while the $\{\Gamma_n$, -$u_{cn}\}$ vector is multivariate normal as in N-MIXL. To obtain the N-MIXL special case, one sets the scale parameter σ_n=σ=1. To obtain the S-MNL special case one sets Trace(V(Γ_n^*))=0, so the variance-covariance matrix of Γ_n^* is degenerate. The G-MNL model has two interesting special cases and a general case that we consider in turn:

3A) We refer to the first special case as **G-MNL-I**. To motivate this model, it is useful to note how G-MNL restricts the general structure in equation (3). Define $\beta_n^* = \{\Gamma_n, -u_{cn}\}$ so β_n = $\sigma_n\beta_n^*$, and define $\beta_n^* = \bar{\beta} + \eta_n$. Here, $\bar{\beta}$ is the mean parameter vector in the population, and η_n is the person n specific deviation from the mean (a N(0,Σ) random vector). In the **G-MNL-I** model the coefficient vector β_n takes the form $\beta_n = \sigma_n[\bar{\beta} + \eta_n]$, which is a log-normal times a normal random variable; that is, a continuous mixture of scaled normals. Of course, one may choose other distributions for σ_n and η_n, but in any version of G-MNL we will have that β_n is the product of a (positive) scalar random variable and a vector random variable.

We next consider versions of G-MNL that are not consistent with the specialized structure in (4), but that are consistent with the more general structure in (3). Recall that (4) was derived from particular interpretations of Γ, u_c and σ based on Lancaster's attribute-based approach.[8] But the model in (3) is "reduced-form" in the sense that it is agnostic about the sources of heterogeneity in the β_n vector. Here we consider models that can be derived from (3):

3B) Fiebig et al (2010) considered an alternative version of the G-MNL model **(G-MNL-II)** in which $\beta_n = [\sigma_n\bar{\beta} + \eta_n]$. This differs from model #3A above in that the scale of the normal errors does not vary with that of the β vector. This model cannot be rationalized by the attribute-based approach. That is, if (1/σ) represents tastes for the unobserved attributes ε, then scaling by σ should affect the whole of the $\beta_n^* = \{\Gamma_n, -u_{cn}\} = \{\bar{\beta} + \eta_n\}$ vector, not just the mean vector $\bar{\beta}$. Nevertheless, this model is perfectly plausible as a stochastic specification in (3). But it is important to keep in mind that it is agnostic about the source of scale heterogeneity.

3C) Fiebig et al (2010) also noted that the two versions of G-MNL can be nested in a single model **(G-MNL)** if we write $\beta_n = [\sigma_n\bar{\beta} + \gamma\eta_n + (1-\gamma)\sigma_n\eta_n]$. Here γ is a parameter that determines how scale affects η_n. For instance, if $\gamma = 1$ then $\beta_n = [\sigma_n\beta + \eta_n]$ and the scale of the normal errors does not vary with that of the β vector. But if $\gamma = 0$ then $\beta_n = [\sigma_n\beta + \sigma_n\eta_n]$ and the normal errors are scaled proportionately to the scale of the β vector.

4) Next is the latent class **(LC)** model. This assumes there are *S* discrete segments of consumers, *s*=1,…,*S*. Each segment has its own β vector (β_s), but there is no heterogeneity within segments. That is, $\beta_n = \beta_s \forall n \in s$. Segments are latent and *S* is not known *a priori*. In our empirical work we choose the number of segments to maximize the Bayes information criterion (BIC). The LC model may be consistent with the structure in (4) under the assumption that σ_n=σ=1 and there are a finite number of β_n^* vectors in the population. In principle, we could also let σ_n take on a finite number of values, so that $\beta_{s(\sigma),s(\beta)} = \sigma_{s(\sigma)}\beta_{s(\beta)}^*$, where $s(\sigma)$ indexes the discrete σ segments and $s(\beta)$ indexes the discrete β_n vectors. These restrictions are testable.

For some time in the 80s and 90s the LC model was arguably the most popular model of consumer heterogeneity. Its popularity was driven by two main factors: First, it gives closed form expression for choice probabilities, which are just an average of MNL probabilities over the discrete types. Second, it provides an intuitively appealing "segmentation" of a market into consumer types. However, Elrod and Keane (1995) found that LC models tend to understate heterogeneity in consumer preferences, leading to poor fit relative to models with continuous heterogeneity distributions. And, with the advent of simulation methods, it became much less

[8] In particular, the structure in (4) implies that variation in σ_n must affect the whole $\beta_n^* = \{\Gamma_n, -u_{cn}\}$ vector.

important to have closed forms. Still, we find that the LC model is worth analyzing as it provides useful insights into the nature of taste heterogeneity in a given market.

5) The fifth and final model we consider is the “mixed-mixed-logit” or **MM-MNL** model. This generalizes N-MIXL by specifying β_n in (3) as a discrete *mixture*-of-multivariate normals. The motivation for this model is that a mixture-of-normals provides a very flexible heterogeneity distribution. Ferguson (1973) shows the mixture-of-normals can approximate any distribution arbitrarily well, and Geweke and Keane (1999, 2001, 2007) show that a small number of normals can approximate even highly non-normal distributions quite well in practice. Also, MM-MNL has been shown to fit choice behavior better than N-MIXL in some recent studies (Rossi, Allenby and McCulloch (2005), Burda, Harding and Hausman (2008)).

As a mixture-of-normals can approximate any distribution, it can obviously generate the special structure in (4), but it may generate more general structures as well. Note that G-MNL and MM-MNL are related, as G-MNL assumes β_n is a *continuous* mixture of scaled normals, while MM-MNL assumes it is a *discrete* mixture-of-normals.

III. The Three Data Sets

We will compare the performance of the choice models described in Section II on three data sets with quite different properties. In this section we describe the data sets. The first two data sets are taken from among the set of stated preference (SP) discrete choice experiments (DCEs) that were analyzed by Fiebig et al (2010) and Keane and Wasi (2013). The third data set is revealed preference (RP) data on demand for frozen pizza that has not previously been analyzed. We are particularly interested to see if the nature of consumer heterogeneity is very different in the RP vs. SP data, and if this leads to different conclusions about the relative performance of the models.

III.A. Two Stated Preference Data Sets on Pizza Delivery

In recent years it has become increasingly common to use SP data to study consumer demand, especially in contexts where revealed preference (RP) data is unobtainable or uninformative. Leading examples are to predict demand for a new product with an attribute that is not present in existing products, to predict the effect of changing an attribute that does not vary in the data, or to value non-traded (public) goods. SP data is also appealing because the investigator experimentally controls the covariates, thus avoiding the problems of endogeneity and collinearity that often arise in RP data.

The SP data that we examine consist of two pizza delivery choice experiments that were run in 2003-4 and that were originally studied in Louviere et al (2008). These data sets have become fairly common "testing grounds" for new modelling approaches. As they are fairly well-known we only describe them briefly:

In both DCEs, respondents are asked to choose between two generic Pizza delivery services (labelled A and B). The experiments differ in the number of attributes that characterize each choice (8 or 16), the number of choice occasions (16 or 32), and the number of respondents (178 or 328). We call the two datasets "Pizza A" and "Pizza B." The delivery services are described as offering different types of pizza (e.g., thick crust, vegetarian) at different prices. Thus, choice of service also provides a way to choose different pizza attributes. This allows us to estimate willingness to pay for some attributes of pizza, as well as of the delivery services.

III.B. Revealed Preference Data on Frozen Pizza

The RP data that we examine consist of data on household purchases of frozen pizza collected by IRI at a large store in Eau Claire, Wisconsin, from Jan. 2001-Dec. 2003. The transactions are observed at the Universal Product Code (UPC) level. In setting up the choice model we face the difficult problem that there are over 400 UPCs. In part, this is because UPC codes change due to trivial changes in attributes. Thus, we group very similar UPCs into broader "types" of pizza. Types are differentiated by more significant attributes such as brand, topping and type of crust. But even then, there are 102 types of pizza available sometime during the study period. (In any one week the number of types varies from 72 to 96).

This large choice set is typical of RP data, and it creates computational problems. That is why the literature on UPC-level choice modeling is small. Fader and Hardie (1996) and Andrews and Manrai (1999) are among the few key works in this area. Later, in Section IV.B and in the Appendix, we will discuss the methods we use to deal with such very large choice sets.

In creating the RP data set two sample screens are applied: First, only panelists who IRI classifies as consistent reporters are included. We also require that sample members be regular frozen pizza consumers. Specifically, they must make 15 to 60 purchases in the 3 years. These sort of criteria are commonly used in marketing studies to screen out households who may have left the city, ceased responding, etc.. There are 129 panelists who meet our criteria. The total number of shopping trips where frozen pizza is bought is 4,123.[9]

[9] On 50 percent of shopping trips people buy only one type of pizza. For trips where consumers buy K types, we treat each as a separate observation, but we take the geometric mean of the likelihood contributions $(\prod_k p_k)^{1/K}$. In

Of course, we also observe non-purchase trips, where panelists visit a store but do not buy frozen pizza. Pizza trips account for roughly 30% of all shopping trips. Here we will estimate models of demand for types of pizza *conditional* on frozen pizza purchase. That is, we do not model purchase timing decisions. This puts our RP and SP results on the same footing.

In the RP data, the characteristics we use to predict utility from each type of pizza are brand (5 major brands plus "other"), topping (7 types), crust (3 types), if the pizza is micro-waveable, price and if the pizza is on promotion (i.e., on feature or display).[10]

Table 1 presents descriptive statistics for the RP data. Note that the five major brands cover 83% of total sales. The largest is Tombstone (26%) followed by Roma and Jacks (20% each). The various "other" brands make up 17% (this group includes DiGiorno at 3%).[11]

Roma is the least expensive brand (average price $2.04 per pound) while Tombstone and Red Baron are the most expensive (around $3.00). It is interesting that average prices differ across types of pizza (e.g., the different toppings) in a way that differs by brand. Brands also differ considerably in their overall promotion intensity, and in how often they promote different pizza types. As we discuss in the note to Table 1, the average price per pizza is AUS $4.50.

It is important to note that the RP and SP data are quite different by construction. In the RP data there are roughly 100 choices available on each supermarket visit. But the SP subjects face a simple choice between two alternatives in each choice task. The hope of SP research is that use of manageably small choice sets and experimentally controlled attribute levels will give more reliable estimates of preferences than RP data.

In the RP data pizzas are characterized by 5 attributes, but three of these have multiple levels, giving 252 possible pizza profiles. But, as noted, only about 100 actually exist in the data. The two SP data sets have 8 or 16 attributes, each with two levels. So there are 256 possible profiles in Pizza A, and 2^{16} in Pizza B. Thus, the complexity of the attribute vector is similar in the RP data and Pizza A (i.e., 252 vs. 256 profiles). But in Pizza B it is more complex.

the MNL case this is the same as weighting the log-likelihood contributions by $1/K$. Less than 2 percent of trips involve purchase of more than 4 types. For these trips, we randomly select 4 types to include in the analysis.

[10] The price and promotion variables are constructed as follows: For the pizza a consumer buys, the recorded price and promotion information for that UPC is used. For non-chosen pizza types, the price and promotion variables are a weighted average over all UPCs within that type. The weights are the market shares for the whole sample period.

[11] At the time of the study, the Tombstone and Jacks brands were owned by Kraft. Red Baron is owned by Schwan's, and Roma and Bernatello are owned by Bernatello. While Kraft and Schwan's are multinational, Bernatello is regional, with a large market share in the Midwest. This may be why Kraft's leading national brand, DiGiorno, was not a big seller in Wisconsin.

IV. Empirical Results: Comparing the Fit of Alternative Models of Heterogeneity

In this section we compare results from choice models estimated on all three data sets. To estimate N-MIXL, G-MNL, S-MNL and MM-MNL we use simulated maximum likelihood with 500 draws. Standard errors are calculated using 5000 draws. We compare fit of the alternative models of heterogeneity using the Bayes Information Criterion, BIC = $-2LL + k\cdot\ln(N)$, where LL is the log-likelihood and the second term is the penalty for number of parameters (k).

IV.A. The Stated Preference Choice Experiments

IV.A.1. *The Pizza A Data Set* (*A Relatively Simple Choice Experiment*)

Table 2 presents estimation results for the Pizza A dataset, for each of the six models discussed in Section II: MNL, S-MNL, N-MIXL, G-MNL, LC and MM-MNL.

The simple MNL model generates reasonable estimates: the price coefficient is negative, while consumers have a strong preference for fresh ingredients and hot delivery. Other attributes are less important, although (quick) delivery time and size options are significant.

The use of effects coding (i.e., “no”=-1, “yes”=1) means the impact of each attribute on utility is double its coefficient. And, as the low and high price levels are set $4 apart, one must divide the price coefficient by 2 to obtain the effect of a $1 price increase. So, e.g., the MNL estimates imply the willingness to pay (WTP) for hot delivery is (0.38×2)/[(0.16)/2] = $9.50.

Initially this struck us as an implausibly large value, as $9.50 is roughly 2/3 of the price of the whole pizza. This seems excessive as, if a pizza is not delivered hot, it can be easily corrected by heating it in the oven. However, $9.50 is roughly the price premium of delivered over frozen pizza.[12] So the estimate implies that subjects are not willing to pay more for pizza delivered cold than for a frozen pizza, which makes sense.

In the next column we report the S-MNL model. This differs from MNL only in that it adds the τ parameter, which measures the degree of heterogeneity in the scale parameter σ_n. Note that τ is large (1.69) and highly significant. Thus, compared to the mean of β, the 90th (10th) percentile β_n is shifted up (down) by 174% (-95%). The log-likelihood improves from -1657 to -1581 when we go from MNL to S-MNL. Thus, there is clear evidence for scale heterogeneity.

The next column reports results for the N-MIXL model. To conserve on space we report only the mean β vector and not the Σ matrix. We report an N-MIXL model with a diagonal Σ matrix, as it achieves a better BIC value than the version with correlations. The log-likelihood

[12] A *Wall Street Journal* online article (blogs.wsj.com/numbersguy/how-much-does-pizza-delivery-cost-469) from 2008 reports an estimated delivery premium of roughly US$9.44. Pizza Hut in Sydney charges AUS $8 for delivery.

improves from -1657 to -1403 in going from MNL to N-MIXL, which is larger than the improvement achieved by adding scale heterogeneity.

Next we report the G-MNL model, which nests S-MNL and N-MIXL. (Again, we do not report the covariance parameters to conserve on space.) The G-MNL model achieves likelihood and BIC improvements over both S-MNL and N-MIXL. For instance, the log-likelihood improves from -1403 to -1372 when we go from N-MIXL to G-MNL, with addition of just the two parameters τ and γ. Thus, the G-MNL results suggest that both scale heterogeneity and the normally distributed random coefficients are important:

The scale heterogeneity parameter τ is quantitatively large (1.80). This, combined with the normal shocks to β_n, allows G-MNL to capture situations where: (i) some consumers have strong preferences for one or two attributes and care little about others, and (ii) some consumers place little weight on all attributes. The latter occurs for a draw of the scale parameter σ_n near zero. Notably, case (ii) is highly unlikely in the N-MIXL model, as the draws for all 8 elements of β must be near zero.[13] As we will see below, the ability to capture both these types of behavior *simultaneously* is why G-MNL fits better than N-MIXL on these data.

The parameter γ in the G-MNL model is close to 0, suggesting that $\beta_n \approx [\sigma_n \beta + \sigma_n \eta_n]$. So the normal errors are scaled proportionately to the scale of the β vector. This is consistent with model 3A in Section II, which is consistent with the attributes-based model (4).

Note that the mean β vectors for the S-MNL, N-MIXL and G-MNL models have similar patterns to the MNL estimates. The coefficient on price is negative, those on fresh ingredients and hot delivery are strongly positive, and those on quick delivery and size options are moderately positive. However, the scale of the mean β vector increases as heterogeneity is added to the model. This is because, with the variance of the logistic errors held fixed, the variance of the composite errors $(\beta_n - \beta)x_{nj} + \varepsilon_{nj}$ increases as heterogeneity is added. Thus, the β coefficients must be larger for observed attributes to have the same effect on choice.[14]

The next set of columns report estimates of the LC model. It identifies 4 segments. We report the β_s vectors only for the three largest to save space. Segment #1 (36%) places great

[13] N-MIXL with a high positive correlation among elements of β_n will not capture this pattern in general. Consider a correlation near one. Then, for each person n, each element of β_n is shifted by a common factor times the standard deviation of that element. The common factor can only move the whole β_n vector to near zero if standard deviations are proportional to means for each element of β_n (which is a very special case).

[14] See Keane (1997) for an alternative parameterization where the scale of the composite errors are held fixed while the fraction of the error variance due to idiosyncratic vs. person specific error components is allowed to vary.

weight on fresh ingredients, segment #2 (32%) has rather modest utility weights on all attributes, and segment #3 (23%) places great weight on hot delivery.

Finally, the far right columns report estimates of the MM-MNL model. The version preferred by BIC has two latent types, with a diagonal variance matrix of the β_s vector for each type. Type 1, estimated as 57% of the population, cares primarily about price, freshness and hot delivery. In fact, the coefficients for type 1 are similar to the MNL estimates. Type 2 (43%) is quite interesting. It has very large mean coefficients on price, freshness, hot delivery and quick delivery, but also very large variances on these coefficients (not reported). Thus, the model can generate consumers who place great weight on some or all of these attributes, and consumers who place little weight on all of them. This is a feature that MM-MNL shares with G-MNL.

The overall ranking of the six models by BIC is G-MNL (2886), MM-MNL (2919), N-MIXL (2933), LC (3115), S-MNL (3233) and MNL (3378). Thus, the G-MNL model is clearly preferred over the alternatives. MM-MNL is second and N-MIXL third, with other models trailing far behind. The good fit of the G-MNL and MM-MNL models relative to N-MIXL implies that the heterogeneity distribution departs substantially from normal.

IV.A.2. *The Pizza B Data Set* (*A More Complex Choice Experiment*)

Table 3 presents results for the Pizza B dataset. It is more complex than Pizza A because each option is characterized by more attributes (16 vs. 8). Louviere et al (2008) wanted to see how this increase in complexity affects choice behavior. The number of choice occasions per person is also larger (32 vs. 16), as is the number of subjects (328 vs. 178).

In the MNL model, the price coefficient is very close to that in Pizza A (-0.17 vs. -0.16). Recall that the SP data uses "effects coding," with a price of \$13 coded as -1 and a price of \$17 coded as 1. Thus, the MNL price coefficient of -0.17 in Pizza B implies that if price *decreases* by \$4 the deterministic part of utility increases by 2×0.17=0.34. This increases demand by roughly 17%. As the price decrease is 24%, the implied price elasticity is roughly -0.71. (If we instead consider a \$4 price *increase* we get -0.55). The Pizza A results are very similar.

The small price elasticity in the SP data is surprising. A monopolistically competitive firm should operate in a region where the price elasticity of demand is ≤ -1.[15] And pizza delivery is well described by the monopolistically competitive model (e.g., in Sydney there are hundreds

[15] For a constant marginal cost of c we have that the monopolist sets p so that $e = c[D'(p)/D(p)] - 1$ where $D(p)$ is the demand curve and e is the price elasticity. If c=0 then e = -1, while if c>0 then e <-1.

of delivery services, differentiated along dimensions like the style of pizza).[16] Would the small price elasticity in the SP data accurately predict behavior in a market setting?

Several factors may account for the low SP elasticity estimates. It may be problems with the SP data itself, e.g., if people do not take the budget constraint seriously when making hypothetical choices, or if some subjects do not take the task seriously and choose randomly.[17]

Alternatively, note that the elasticities in SP data depend on preferences and *experimental* prices. This contrasts with RP elasticities, which are a function of preferences and *equilibrium* market prices. Thus, the SP elasticities may be too small simply because they are evaluated at non-equilibrium prices (i.e., at the "wrong" point on the demand curve).

More precisely, MNL implies the price elasticity of demand for an option with market share s and price p is $e = \beta_p p(1\text{-}s)$, where β_p is the price coefficient. So the price elasticity is proportional to the price level itself. Thus, if the price elasticity evaluated at $17 is -0.71, then price must increase by 40% to $24 to reach the point on the demand curve where the price elasticity is $\leq$ -1. This seems plausible – e.g., Domino's in Sydney has a minimum delivery order of $24, and the average order is of course larger.

Turning to other parameters of interest, the most important attributes are again freshness and hot delivery, but the magnitude of their coefficients drops in half from Pizza A. This may be due to the introduction of 8 new attributes. The new indicator for free delivery is positive as expected (0.12). People put small but significant weights on several other attributes, including crust type, size options, local store, baking method, vegetarian option, delivery time, and variety.

In the S-MNL and G-MNL models the τ parameters that capture scale heterogeneity are again large and significant (1.22 and 1.50, respectively). As in Pizza A, the parameter γ is close to 0, so the normal errors are scaled proportionately to the scale of the mean β vector. This is consistent with the attribute-based model. Again, G-MNL achieves substantial likelihood and BIC improvements over S-MNL and N-MIXL. For instance, the log-likelihood improves by 230 points (from -5892 to -5662) in going from N-MIXL to G-MNL. And again, the mean β vectors for the S-MNL, N-MIXL and G-MNL models have similar patterns to the MNL estimates, except that the scale of the mean β vector increases as heterogeneity is added to the model.

[16] See, e.g., www.deliveryhero.com.au/takeaway-sydney/?categories=pizza.

[17] A larger scale of the errors (a smaller σ) also leads to a smaller price elasticity. Intuitively, demand is less elastic if unobservables are more important determinants of choice. But this cannot explain elasticities > -1. Furthermore, we are skeptical that higher scale due to greater importance of unobservables can explain the small elasticities in the SP data, because adding eight new controls in going from Pizza A to B hardly changes the elasticity.

We now turn to the LC estimates. It is interesting that by far the most common type (51%) has very small coefficients on all attributes. Thus, this type exhibits close to "random" choice behavior, in the sense that observed attributes have little influence on choices, which are driven primarily by the random shocks to utility. The 2nd most common type (14%) places great weight on price, while the 3rd largest places great weight on freshness. As for types not reported in the table, the 4th (10%) cares greatly about crust type, the 5th (9%) wants hot delivery, and the 6th (4%) likes vegetarian. So we have 5 small segments that care about different attributes.

The prevalence of "random" behavior in Pizza B (i.e., the 51% in type 1) may result from subjects being confronted with a very complex choice task (16 attributes), causing confusion or a lack of desire to take the task seriously. This implies that 16 attributes is too many to include in an experimental design. An alternative explanation is purely statistical: with 5 segments devoted to types who like specific attributes, perhaps the best fit is obtained by grouping everyone else into a portmanteau type that lacks strong preferences. In this case, the SP data itself is reliable, and the "random" type is merely an artifact of how the LC model attempts to fit the data using a finite mixture. These two explanations have very different implications regarding the internal and external validity of the results. We explore this issue further in Sections IV.B and V.

Turning to the MM-MNL model, the preferred version now how three components in the normal mixture, compared to only two in Pizza A. Comparing the mean β vector for each type, the most common type (41%) has small weights on all attributes (like the "random" type in the LC model). The 2nd type (31%) places substantial weight on price and the 3rd type (28%) places great weight on both freshness and hot delivery.

Both the MM-MNL and LC model estimates imply that the structure of consumer heterogeneity is more complex in Pizza B than it was in Pizza A. That is, LC model identifies 6 segments in B vs. only 4 in A, and the MM-MNL model identifies 3 types in B vs. only 2 in A. Given the larger number of attributes (16 vs. 8), it is not surprising that the structure of heterogeneity is more complex in dataset B. The overall ranking of the six models by BIC is MM-MNL (11527), G-MNL (11639), N-MIXL (12081), LC (12118), S-MNL (13372) and MNL (13641). Thus, in this dataset the MM-MNL model is preferred over G-MNL, whereas in Pizza A the situation was reversed. Based on these results, our preliminary hypothesis is that more complex heterogeneity structures tend to favor the discrete mixture-of-normals model. But we analyze this issue more carefully in Section V.

IV.B. The Revealed Preference Data on Frozen Pizza

In our models for the RP data the attributes we use to predict utility of each type of pizza are the price, an indicator for promotion (i.e., on feature or display), 5 brand indicators ("other" is the omitted category), 6 indicators for toppings ("combo" is the omitted category), 2 indicators for crust ("regular" is omitted) and an indicator for microwaveable. We assume the brand intercepts capture latent quality of brands and/or brand equity. Thus, price effects are identified from variation of prices (within brands) over time. We assume this variation is exogenous – see Erdem, Imai and Keane (2003) for arguments in favor of this assumption.

Recall from Section III that 102 types of pizza are available sometime during the sample period, and within any one week the number of types varies from 72 to 96. As we noted in Section II, with many varieties it is natural (and practical) to use brand intercepts rather than alternative specific constants (ASCs). By adopting an attributes-based approach, as in Fader and Hardie (1996), we obtain a model with many fewer parameters (15) than options (102).

The large choice set size ($J \approx 100$), along with the large sample size, causes significant computational burden in estimating models that lack closed form choice probabilities (i.e., the models with heterogeneity). To deal with this issue, we use randomly chosen subsets of the full choice set in estimation. This procedure deserves further comment, as it is independent interest:

McFadden (1978) showed that one can consistently estimate parameters of MNL using random subsets of the full choice set. For example, in a case with 100 alternatives, one might construct hypothetical choice sets that include the chosen alternative, plus 19 alternatives chosen randomly (without replacement) from the remaining 99, giving a reduced choice set size of 20.

Unfortunately, as we show in the Appendix, McFadden's result does not hold when MNL is extended to include heterogeneity. However, we also obtain results indicating the bias that results from using random subsets of the full choice set to estimate MIXL models is negligible.[18] This finding is of independent interest, as it may be useful in many other contexts in economics where choice sets are very large (e.g., school choice, housing choice, occupational choice).

Based on these results, we decided to estimate the RP models using choice set sizes of 20, 30 or 40, which correspond to roughly 25%, 37.5% or 50% of the full choice set. The results for 40 draws are reported in Table 4 while those for 20 and 30 draws are reported in the Appendix

[18] The basic intuition of McFadden's consistency result is that the use of a random subset of the full choice set shifts the (expected) log-likelihood up, but does not alter where it is maximized. Even though this result does not hold exactly with heterogeneity, it holds to a very good approximation, as we discuss in the Appendix.

for comparison purposes. Strikingly, the estimates are very stable across choice set sizes for all 6 models. This stability strongly suggests that any bias induced by using random subsets of the full choice set is negligible, even for models with heterogeneity.

Consider first the MNL estimates in Table 4. The price coefficient is -0.84. Thus, for MNL the price elasticity of demand for an alternative with market share *s* and price *p* is simply *e* = (-.84)*p*(1-*s*). For a variety of pizza with an average price ($2.80/lb) and an average market share (*s*=1/J) we have a price elasticity of -2.3. [19] This is in the ballpark of prior RP estimates for supermarket goods (see Keane (2010)).

It is worth emphasizing at this point that MNL only permits the price elasticity of demand to differ with market share. But the more general models that allow for consumer taste heterogeneity also let elasticities depend on product attributes. In this section our focus is primarily on model fit, but we examine price elasticities in more detail in Section VI.

Other aspects of the results are that Tombstone and Jacks (both owned by Kraft) have the largest intercepts, while Bernatello has the smallest. Recall that the intercepts capture latent quality and/or brand equity. It is not surprising that national brands have more brand equity than a regional brand. The most popular toppings are sausage and pepperoni, while vegetarian is very unpopular. Rising crust is the least popular crust type.

Next consider the S-MNL model. It includes scale heterogeneity in the observed attribute coefficients. But the brand intercepts are treated differently. As Fiebig et al (2010) note, scaling the intercepts works very poorly in practice due to the phenomenon of brand loyalty. A consumer loyal to a brand will have a positive intercept for that brand, and negative intercepts for other brands. Thus, in contrast to vertical attributes like price or quality, it is unrealistic to assume brand intercepts have the same signs for all consumers. Instead, we assume the brand intercepts are normally distributed.[20]

S-MNL achieves a large log-likelihood improvement over MNL (-11,930 vs. -13,602). It has 11 extra parameters: the scale heterogeneity parameter τ, and the variance matrix of the 5 random brand intercepts. BIC prefers a model where this matrix has a one-factor structure with 10 parameters (Elrod and Keane (1995)). Much of the log-likelihood gain is due to the random

[19] The elasticity depends on choice set size through the (1-*s*) term. But for a typical option we have (1-*s*) ≈ (J-1)/J. This is close to one and little affected by choice set size (if we use a reasonably large random subset, say, J ≥ 20).

[20] We interpret brand intercepts and ε as capturing unobserved attributes of brands and varieties (within brands), respectively. Thus, in this version of S-MNL, the observed factors (x_{njt}) are scaled while unobserved factors are not. This means observed factors are more important determinants of choice for some consumers than others.

intercepts. The estimate of τ is 0.86, which is highly significant, but much smaller than in the SP data. The mean β vector is similar to that for MNL.

Next, Table 4 reports the N-MIXL results. This model has 16 variables (11 attributes and 5 brand intercepts), so a full covariance matrix would have 136 parameters. The BIC selects a model with a one-factor structure on V(β_n). This model has only 32 parameters, so it is much more parsimonious. N-MIXL generates a log-likelihood that is superior to S-MNL by 882 points. But the mean β vector is still fairly similar to both the MNL and S-MNL models.

The G-MNL results are interesting, as they contrast sharply with those for the two SP data sets. First, G-MNL gives only a small (31 point) log-likelihood improvement over N-MIXL. In percentage terms this is only 0.3%, compared to gains of 2.1% and 3.4% in Pizza A and B. Second, the scale heterogeneity parameter τ is only 0.40, which is several times smaller than in the SP data. Third, γ is 0.79, which is close to the $\beta_n = [\sigma_n \beta + \eta_n]$ case where the scale of the normal errors does not vary with that of the β vector (model 3B). The values of τ and γ imply that: (i) scale heterogeneity is less important in the RP data, and (ii) the G-MNL model is much closer to its N-MIXL special case than in the SP data. We will see more evidence of this below.

Next, Table 4 reports the LC results. BIC prefers a model with 5 latent types. Note there is no large "random" type like we found in the SP data. The most notable difference between types is in price sensitivity. The largest type (28%) has a price coefficient of -1.85. That for the 2nd largest (24%) is -1.65, the 3rd largest (23%) is -1.07 and the 4th largest (14%) is -0.89. The smallest type (10%) has an insignificant price coefficient.[21] The log-likelihood for the LC model is 1486 better than MNL, but 186 worse than S-MNL, 1068 worse than N-MIXL and 1099 worse than G-MNL. The relatively poor fit of LC is consistent with the SP results.

Finally, the MM-MNL results are in last 4 columns of Table 4. The version preferred by BIC has two types, with proportional covariance matrices for the β_n vectors. The two types uncovered by the mixture-of-normals model are very different, particularly with regard to price elasticity. Type 1 (65%) has a mean price coefficient of -1.75, giving an elasticity of roughly e = (-1.75)p(1-s) = -4.7 for a "typical" brand with s=1/J. Type 2 (35%) has a mean price coefficient of -0.48, giving a much smaller price elasticity of roughly -1.3. [Of course, the whole point of MM-MNL is to let elasticities vary flexibly across items, an issue we return to in Section VI].

[21] There are also differences in how types value other attributes (e.g., type 1 has a stronger preference for Tombstone and Roma than other types, types 1, 2 and 5 have strong preferences for sausage/pepperoni while 3 and 4 do not, type 3 is not sensitive to promotion while other types are, etc.).

To summarize, MM-MNL has by far the best BIC of all models considered; i.e., 297 better than G-MNL and 344 better than N-MIXL.[22] These results contrast in a subtle but interesting way with the results for the two SP datasets. The fit comparison for all three data sets is summarized in Table 5. Notice that the N-MIXL model comes out 3rd in all three comparisons. But what is interesting is that, while G-MNL clearly outperforms N-MIXL in both SP data sets,[23] the fit of the G-MNL and N-MIXL models is very close in the RP data. This, combined with the clearly superior fit of MM-MNL, suggests that, in the RP data, the heterogeneity distribution does depart from normality, but not in a way that G-MNL can easily capture. We focus on this problem in Section V.

Table 5: Model Fit Comparison

	MM-MNL	G-MNL	N-MIXL
Pizza A	2919 +1.1%	2886 -----	2933 +1.6%
Pizza B	11527 -----	11639 +1.0%	12081 +4.8%
RP Data	10802 -----	11017 **+2.0%**	11048 **+2.3%**

Note: The table reports BIC values and the percentage by which each model has a higher BIC than the preferred model.

V. What Accounts for Which Model Fits Best in Each Case?

In this section we examine how patterns of consumer behavior differ in each of the three data sets, and assess which model(s) can best capture those patterns.

V.A. Behavioral Patterns in the SP Data

Here, in order to better understand why the G-MNL and MM-MNL models are preferred to N-MIXL in the SP data, we examine how well each model fits key patterns in the data. First, we determine what each model implies about the distribution of consumer taste heterogeneity. To obtain posteriors of the individual level parameters, we adopt what Allenby and Rossi (1998) call an "approximate Bayesian" approach: A model's estimated heterogeneity distribution is taken as the prior. We then calculate posterior means of the person-specific parameters conditional on each person's observed choices (see Train (2003) for details).

Figure 1 plots, for Pizza B, posterior distributions of the person-level price and fresh

[22] As we show in the Appendix, the use of reduced choice sets biases BIC toward smaller models. This does not affect our results as the preferred model, MM-MNL, is the largest of the three.

[23] As well as in most of the several other SP data sets studied by Fiebig et al (2010).

ingredient coefficients for N-MIXL, G-MNL and MM-MNL. These are kernel density estimates using a normal kernel. Notice the N-MIXL posteriors have a distinctly normal shape. As Allenby and Rossi (1998) point out, N-MIXL's normal prior has a strong tendency to draw in outliers, so it can't capture the behavior of consumers who place very great weight on a particular attribute.

In contrast, the posteriors of G-MNL and MM-MNL depart substantially from normality. In the left panel of Figure 1, we see that both models generate a mass of consumers in the left tail who care intensely about price. And in the right panel, we see that both models generate a mass of consumers who care intensely about fresh ingredients. The G-MNL and MM-MNL posteriors also exhibit excess kurtosis – a large mass of consumers with price or quality coefficients near zero. This enables them to generate consumers who are largely indifferent to observed attributes.

Are subjects who exhibit such preferences actually common in the SP data? In the Pizza B data, 27 of 328 subjects chose the cheaper pizza on all 32 choice occasions regardless of other attribute settings, while 24 always chose the fresh pizza. Thus, 51 subjects exhibit lexicographic preferences for price or freshness. For these 51 subjects, G-MNL and MM-MNL have BIC advantages over N-MIXL of 135 and 158 points, respectively. An additional 62 subjects exhibit lexicographic preference for some other attribute (e.g., hot delivery, vegetarian, crust type).

Among the total of 113 subjects who exhibit lexicographic preferences, G-MNL and MM-MNL have BIC advantages over N-MIXL of 296 and 529 points. Recall (Table 5) that the overall BIC advantages of these models over N-MIXL are 442 and 554 points. Thus, lexicographic subjects account for 67% and 95% of these overall BIC gains.

Next, consider consumers who appear to be largely indifferent to observed attributes. Keane and Wasi (2013) refer to this as "random" behavior and give a definition of randomness based on the following idea: If a subject chooses randomly between options A and B, so that the attributes do not affect choice, then the mean attribute differences between the chosen and non-chosen options should be "close" to zero (except for sampling variation). Given their definition, 31 subjects exhibit "random" behavior in Pizza B. For these subjects, G-MNL and MM-MNL have BIC advantages over N-MIXL of 107 and 58 points.

If we combine the results for the 113 lexicographic subjects and 31 random subjects, the BIC advantages of G-MNL and MM-MNL over N-MIXL are 403 and 587 points. Thus, the lexicographic and random consumers together account 91% and 106% of the overall BIC gains of these two models. Yet these consumers account for only 144/328 = 44% of the subjects.

Given these results, it is clear why G-MNL and MM-MNL are preferred to N-MIXL: Both models capture two important features of the SP data that N-MIXL does not: (i) consumers with very strong (or even lexicographic) preferences for particular attributes and (ii) consumers whose choices are little affected by the whole attribute vector. The G-MNL and MM-MNL models are able to capture these features of the data because both use mixture-of-normal distributions that are more flexible than the normal distribution assumed by N-MIXL.[24]

Table 6 presents another way to look at the data. Here, we group consumers into types (based on price sensitivity and other characteristics, as revealed by their choices). Then we report the mean BIC advantage of the MM-MNL and G-MNL models over N-MIXL for each type. In the top panel we group subjects in Pizza B by price sensitivity. For each subject, we calculate the average difference in price between the chosen and rejected options (over all 32 choice tasks).[25] The more negative the average price difference, the more price sensitive the consumer.[26]

Notice that MM-MNL and G-MNL have large average BIC advantages for consumers in both the very price sensitive group and the price insensitive group. For example, G-MNL has BIC advantages of 1.24 and 1.79 points (per subject) in these groups, respectively. Given the group sizes, this generates a BIC advantage of (1.24)(26)+(1.79)(178) = 351. This accounts for most of the overall BIC advantage of G-MNL over N-MIXL, which is 442 points (see Table 3).

The 2nd panel of Table 6 contains a similar analysis for the fresh ingredients variable. Again, most of the BIC advantage of both G-MNL and MM-MNL over N-MIXL comes from the groups that either care most or least about fresh ingredients.

Thus, the top two panels of Table 6 show that G-MNL and MM-MNL are preferred over N-MIXL in the SP data because they can *simultaneously* generate consumers who care very much and very little about certain attributes. This is because these models are flexible enough to generate heterogeneity distributions like those in Figure 1, with both skewness and excess kurtosis. The normality assumption makes it difficult for N-MIXL to generate such distributions.

[24] A more subtle question is why G-MNL is preferred over MM-MNL in Pizza A, while this is reversed in Pizza B. Note that MM-MNL has a better log-likelihood than G-MNL in both datasets. But it also has many more parameters (33 vs. 18 in Pizza A and 98 vs. 34 in Pizza B). But heterogeneity is more complex in B, and according to BIC this justifies the extra parameters of MM-MNL. Still, G-MNL and MM-MNL make very similar behavioral predictions.

[25] The prices in the experiment are $13 or $17, so the largest possible average price difference is 4 (always buy the more expensive pizza), while the smallest is -4 (always buy the cheapest pizza).

[26] Only 19 levels of the average price difference are observed in the SP data, so we cannot group subjects into even quantiles. Instead we sort them into 4 unequal sized groups. The most price sensitive contains 26 people with a mean price difference of -2.41. The least price sensitive contains 178 people with a mean difference of 0.20. This group is so large because 106 subjects have a mean price difference of exactly zero – i.e., they seem insensitive to price.

V.B. Behavioral Patterns in the RP Data

Next we examine the shape of parameter heterogeneity distributions in the RP data. Figure 2 plots posterior distributions of two selected person-level parameters, the price and vegetarian parameters, from three models: N-MIXL, G-MNL and MM-MNL. In sharp contrast to the SP results, the N-MIXL and G-MNL posteriors for price (left panel) look very similar. Each departs only modestly from normality, with a moderate degree of skewness to the right. The MM-MNL posterior for the price coefficient departs much more sharply from normality. It is highly leptokurtic, with a sharp peak near -1.9, and it is strongly skewed to the right, with much less mass to the left of -2.6 than for either the G-MNL and N-MIXL models.

These RP results are very different from what we saw in the SP data. There, the G-MNL and MM-MNL posteriors are similar (see Figure 1). Each departs sharply from normality, while the N-MIXL posterior has a distinctly normal shape.

Next, consider the posterior for vegetarian. As we saw earlier, most consumers dislike vegetarian. But in Figure 2 (right panel) the MM-MNL posterior is bi-modal, picking up a subset that prefers vegetarian. Both N-MIXL and G-MNL fail to capture this, generating close to normal distributions. G-MNL attempts to mimic the bi-modal pattern through greater dispersion.

As we discussed in the introduction, the G-MNL and N-MIXL models are very popular in applications, while MM-MNL is relatively little used. Thus, it is important to understand why MM-MNL dominates both models in the present context. As we saw in Section V.A, the reason G-MNL and MM-MNL fit better than N-MIXL in the SP data is their ability to fit lexicographic and random behaviors. But these behaviors are not prevalent in the RP data, so G-MNL loses its main source of advantage over N-MIXL.

But why does MM-MNL fit much better than both G-MNL and N-MIXL in the RP data? Table 6 sheds light on this issue. First, we group consumers in the RP data into types (based on price sensitivity and other characteristics, as revealed by their choices). In contrast to the SP data, where there are only two choices, in the RP data there are many choice options. Thus, we decided to compare the attribute of the chosen option with the *average* attributes of all non-chosen options. The difference is then averaged over all of a person’s purchases. We then group people into quintiles of attribute sensitivity. Then we report the mean BIC advantage of the MM-MNL and G-MNL models over N-MIXL for each quintile.

The results for the RP data are reported in the bottom panels of Table 6. Here we look at the price and vegetarian attributes. In the 3rd panel of Table 6, we see the most price sensitive consumers buy (on average) a pizza costing 67 cents less (per lb) than the non-chosen alternatives. The least price sensitive consumers spend (on average) 21 cents more than the cost of alternatives. A striking difference between the RP and SP results is that the percentage of consumers who are insensitive to price is much smaller in the RP data.[27] Another striking difference is that, in RP, G-MNL does not fit the most price sensitive consumers better than N-MIXL.[28] The reason is clear in Figure 2. In the RP data, the left tail of the price coefficient distribution is almost identical for G-MNL and N-MIXL. So it is unsurprising that the most price sensitive consumers behave similarly in the two models.

In contrast, the MM-MNL model still has a BIC advantage over N-MIXL for the most price sensitive consumers (1.03 per person). In fact, it has an even larger advantage over G-MNL (1.58 per person). In Figure 2, we see MM-MNL puts much less mass in the left tail of the price coefficient distribution (i.e., left of about -2.6) than do G-MNL and N-MIXL. Thus, MM-MNL implies the most price sensitive consumers are not as price sensitive as the other models suggest.

As we see in Table 6, MM-MNL not only beats N-MIXL and G-MNL on BIC for the most price sensitive consumers; it has a clear advantage in every quintile of price sensitivity. Consistent with this, in Figure 2 we see the distributions of person-level price coefficients are very similar for N-MIXL and G-MNL, while that of MM-MNL is distinctly different: The mode is shifted left (to about -1.9), and kurtosis is much greater (putting more mass near -1.9). So, despite the mode shifting left, there is less mass in the left tail. Furthermore, the flexibility of the mixture-of-normals enables it to also generate a fat right tail (price coefficients near zero). These features enable MM-MNL to give a better fit to consumers in all 5 quintiles of price sensitivity.

V.C. Why Can't G-MNL Capture the Departures from Normality in the RP Data?

The real puzzle here is why does G-MNL, which generated a substantial departure from normality in the SP case, have trouble doing so in the RP case? The answer involves the different nature of the departure from normality in the two cases:

In the SP data we see that G-MNL generates both more large price coefficients and more values near zero than a normal (Figure 1). In the SP data the modal price coefficient is close to zero, so increasing kurtosis and increasing mass near zero are equivalent. But in the RP data the

[27] In the RP data only 14/129 = 11% of consumers buy (on average) a pizza that costs the same or more than the average cost of alternatives. In the SP data the figure is 178/328 = 54% (see Table 6).

[28] Indeed, the average BIC difference for G-MNL vs. N-MIXL in this group is -0.55, so N-MIXL fits slightly better.

mode is far to the left of zero (see Figure 2), so increasing kurtosis and increasing mass near zero are not equivalent. This is the source of the problem.

In the RP data (Figure 2) MM-MNL generates a departure from normality involving both (i) excess kurtosis (more mass near -1.9) and (ii) a fat right tail (more mass near zero). G-MNL cannot generate this pattern because it relies on a normal scaled by a log-normal. If the normal is mean zero this distribution generates both kurtosis and fat tails (see Figure 3A). But if the normal has a negative mean, this distribution can only generate kurtosis by shifting the mode towards zero and creating a fat left tail (see Figure 3B). Assuming the posterior for the price coefficient generated by MM-MNL (see Figure 2) gives a fairly accurate picture of the true distribution, the prior distribution assumed by G-MNL (see Figure 3B) is a very poor representation.

The flexibility of MM-MNL is even more apparent for the vegetarian coefficient. In the 4th panel of Table 6 we see that only consumers in the 1st quintile like vegetarian, while all others avoid it to a degree. MM-MNL dominates both N-MIXL and G-MNL for all quintiles but the 3rd. In Figure 2, MM-MNL generates a plausible bi-modal shape for the taste distribution, with most consumers disliking vegetarian while a significant minority is either indifferent or does like it. N-MIXL cannot generate this pattern; it instead generates a close to normal distribution centered at -1.85. G-MNL also generates a close to normal distribution but with a higher variance.

In summary, the structure of heterogeneity is very different in the RP vs. SP data. Both exhibit substantial departures from normality, but their nature is different. Those in in the SP data (i.e., excess kurtosis and fat tails) are well captured by a normal scaled by a lognormal. Thus, G-MNL fits about as well as MM-MNL, and both are clearly preferred to N-MIXL. But in the RP data the departures from normality are more complex, and neither G-MNL nor N-MIXL is well suited to capture them. In contrast, the MM-MNL model does well in both contexts because a discrete mixture-of-normals can approximate any density (Ferguson (1973)).

V.D. Why Does the Distribution of Heterogeneity Differ in the SP and RP Data?

An interesting question is why the distributions of heterogeneity look so different in the SP and RP data. Here we have only looked at three representative data sets, but the "excess kurtosis and fat tails" heterogeneity pattern that we see here is very common in SP data sets across many different types of products and choices (see, e.g., Fiebig et al (2010), Keane and Wasi (2013)). In contrast, the (somewhat limited) literature that allows for very flexible heterogeneity distributions in RP data does not report this sort of pattern (see, e.g., Burda et al

(2008), Rossi et al (2005), Geweke and Keane (1999, 2001)).

The two obvious hypotheses for why the heterogeneity distributions would be different in the RP vs. SP data are: (1) that the goods are different (delivered vs. frozen pizza), so that there is no reason to expect the distribution to be similar, and (2) that consumers behave very differently in RP vs. SP data (e.g., they may fail to take the budget constraint seriously in SP). We will argue that the first hypothesis is highly implausible, which leads us to favor the second.

V.D.A. *Can the Difference in Products Explain the Results*?

For the moment let us grant that delivered and frozen pizza are completely different goods (a point we contest in the next section). A crucial point, that we developed in Section II, is that the distribution of the price coefficient is comparable even for two completely different inside goods, and even without Lancaster-type assumptions, provided the goods are low cost relative to total income. This is because the price coefficient is the marginal utility of the outside good (up to scale). All we need for this result is maximization subject to a budget constraint.

More precisely, as we saw in equation (4), the u_c component of the price coefficient is equal across data on different *inexpensive* goods. Thus, as long as consumers take the budget constraint seriously in each case, theory says the price coefficient should differ between the SP and RP data only because of differences in the scale σ_n. As we see in (4), a larger scale of the errors (smaller σ_n) leads to a smaller price coefficient. Intuitively, demand is less responsive to price if unobservables are more important determinants of choice.

In order for differences in the distribution of σ_n to account for the different distributions of the price coefficient in the SP data (Figure 1) vs. the RP data (Figure 2) two things are needed: In the SP data: (i) the σ_n distribution must have much more mass near zero (than in the RP data), and (ii) the σ_n distribution must have considerably more mass on large positive values.

In our view it is highly implausible that the scale parameter would differ between frozen and delivered pizza in this way. Put simply, why would large segments of consumers be either very sensitive or very insensitive to price when ordering pizza delivery, but not exhibit such behavior when buying pizza in the store?

A more rigorous test proceeds as follows: Recall that σ_n is a scalar that shifts all attribute coefficients (see (3)). Thus, if the σ_n parameter for delivered pizza has the type of distribution we just described, it has implications for the distributions of attribute coefficients as well. For instance, consider the coefficient on freshness (quality). Consumers with σ_n near zero should obviously have both price and quality coefficients near zero. Similarly, consumers with very large σ_n should tend to have both large price coefficients and large weights on quality.

Figure 4 gives a scatter plot of posteriors of the price and quality coefficients in Pizza B (based on the MM-MNL model). Note that subjects with large price coefficients invariably have small quality coefficients. And subjects with large quality coefficients all have relatively small price coefficients. These patterns are consistent with the data patterns we discussed earlier, whereby many subjects in the SP data exhibit lexicographic preferences for price or quality. They are not consistent with the positive correlation between price and quality coefficients we would expect if the σ_n distribution in the SP data had the form describe above.[29]

Based on this evidence, we conclude that the most plausible explanation for the very different distribution of price coefficients in the SP data is that the u_{cn} differ from those in the RP data. That is, some subjects do not take the budget constraint seriously in SP, and act as if u_{cn} is very small, leading to price coefficients near zero. And some subjects adopt lexicographic or random choice rules to simplify the experimental task. This leads to much greater heterogeneity in σ_n in the SP data (i.e., higher values of τ), and to very large price coefficients for a subset of consumers, and to the "excess kurtosis and fat tails" heterogeneity pattern more generally.

V.D.B. *Are the Products Really Different*?

The very different heterogeneity patterns that we find in the RP and SP data are all the more puzzling given that delivered pizza and frozen pizza are closely related products. In each case the final good that consumers demand is "convenient pizza at home" (thus saving time). Delivery and freezing are merely alternative methods to achieve this. Thus, it is reasonable to assume the utility that consumers receive from a pizza of given attributes is the same *regardless* of whether it was delivered or bought in a store.

In assessing this argument, it is important to remember that our frozen pizza data come from the 2001-2003 period. Prior to the late 1990s, frozen pizza was of such low quality that it was not comparable to pizzeria pizza. Instead, it was positioned as a cheap junk food, not a competitor for pizzeria-style pizza. But in 1995 Kraft introduced frozen yeast-leavened dough technology. This enabled them to make frozen pizza with a rising crust. As a result, it became possible to make frozen pizza that is arguably as good as pizzeria pizza. For discussions of this, see New York Times (2002, 2003). For example, consider this excerpt:

[29] There is also no evidence that negative correlation between the β_{nk} for price and quality can explain this pattern. This correlation is estimated in G-MNL and N-MIXL and a diagonal V(β_n) is preferred in each case (see Table 3).

> “For decades, frozen food meant ... cardboard crust pizzas ... most often eaten in front of the television -- which provided a welcome distraction from the taste. ....The changes in frozen food started with the rising crust pizza five years ago. The crust actually fluffed while it baked, giving it the taste and consistency of pies from a pizzeria.” (“Frozen Foods Show Upturn in Taste and Sales,” *The New York Times*, Feb. 23, 2002, Section C, p. 3).

The industry now views frozen pizza and delivered pizza as being in close competition, and this is reflected in firms’ advertising strategies.[30] This is discussed, for example, in Frozen Food Age (2007, 2008). The following excerpt illustrates the point:

> "The national print, broadcast and Internet campaign shows that families can experience restaurant quality and style of pizza from the convenience of their freezer and for a lot less money than take-out" (“Firing on All Cylinders,” *Frozen Food Age*, 2008, 56:8 (March), p. 16).

Given this industry background, we see that frozen and delivered pizza are now perceived as rather close substitutes. Given this, we find it highly implausible that people would behave extremely differently when buying the two products. For example, it seems implausible that a large fraction of consumers would simply forget about the budget constraint when buying delivered pizza. A far more plausible hypothesis is simply that (at least some) consumers simply behave differently in the RP vs. SP settings.

For example, while some subjects may take an SP task seriously, others may try to get through it quickly by adopting simple lexicographic rules like “always choose the lowest price option” or “always choose the highest quality option.” This does not invalidate SP data as a way to learn about demand, but it suggests we need to address this problem. One possible avenue is the “process heterogeneity” model developed in Keane and Thorp (2016), where a fraction of subjects make optimal choices, while a fraction follow one or more sub-optimal rules of thumb.

VI. Substantive Results from the RP Estimation

Finally, we return to the issue of how consumer taste heterogeneity influences own and cross-price elasticities of demand. Ultimately, this is really motivates our interest in how best to model consumer preference heterogeneity. There is a large literature in marketing that estimates price elasticities of demand at the brand level. If retailers move the prices of all varieties of a brand in tandem, then this is all that is of interest. But inspection of our RP data reveals the existence of substantial price variability for individual varieties *within* brands. Our use of a large

[30] See, e.g., https://www.youtube.com/watch?v=iE6T6pfV7i4, https://www.youtube.com/watch?v=lquKTqn0B2E, or https://www.youtube.com/watch?v=T9yH2VBuHT0 for some representative ads.

choice set, with varieties treated as separate options, enables us to compare brand vs. variety level elasticities, and to decompose variety level elasticities into brand vs. variety components.

To keep the exposition manageable, we aggregate varieties to the brand/topping level, where there are 37 options. Using our best fitting model, MM-MNL, Table 7 compares predicted and actual market shares for brand/topping combinations. Clearly, the fit is rather good.

VI.A. Brand Level Price Elasticities

Table 8 reports brand-level price elasticities for selected brands. In forming these elasticities, we assume prices of all varieties of a brand are increased proportionally. We report elasticities for all 6 models, estimated using random choice set sizes of 20, 30 and 40. The first thing to notice is that elasticities are quite stable across the different random choice set sizes.

For Tombstone, which is the largest brand (26% market share), the own price elasticity is -1.66 according the preferred MM-MNL model (estimated using random choice sets of size 40). The simple MNL model implies a nearly identical elasticity of -1.71. So, for this purpose, it is not clear why one would go to the trouble of estimating the MM-MNL model.

But the story is different if we look at Jacks, which has a smaller market share (20%). Here the MM-MNL model gives an elasticity of -2.14, while MNL gives -1.69. Thus, MM-MNL implies a larger price elasticity of demand for the smaller brand, and MNL implies the reverse. Indeed, this is a fairly general pattern. Across all brands/varieties, the correlation between the log market share and the price elasticity is .53 for MM-MNL (i.e., higher share, smaller elasticity), while the correlation implied by MNL is -.50 (i.e., higher share, larger elasticity). This is a fundamental behavioral difference between the models. Furthermore, all the models with heterogeneity generate a positive correlation (i.e., higher share, smaller elasticity).

VI.B. Variety Level Price Elasticities

In Table 9 we report variety-specific elasticities for brand/topping combinations. We only report results for the preferred MM-MNL model (estimated using a random choice set size of 40). First, in the top panel we consider Tombstone with the sausage/pepperoni topping. This is the most popular of all varieties, with a market share of 11%. The price elasticity of demand for this variety is -1.68, compared to -1.66 for the whole Tombstone brand.

The closeness of the brand and variety level elasticities at first seems very surprising. Conventional wisdom suggests that the price elasticity for one variety should be appreciably larger than that for a whole brand, because a variety has more close substitutes. That is, in response to a price increase for a single variety, a consumer who is loyal to the Tombstone brand can switch to other varieties of Tombstone, rather than having to switch a different brand.

What drives this result becomes apparent if we decompose the brand level elasticity at the variety level. When the whole Tombstone brand raises its price, the elasticity for the sausage variety is -1.35, while that for all other varieties combined is -1.93. So the price elasticity for the sausage variety is indeed larger with respect to its own price than with respect to the price of the whole brand (-1.68 vs. -1.35). But the less popular varieties have a large elasticity (-1.93), which brings the average elasticity for the whole brand up to that for the sausage variety.

The situation is different when we look at varieties with a smaller market share. The average market share in the data is 100/37 = 2.7%. So consider Jacks meat/supreme pizza, which has a "typical" market share of 2.6%. In this case the variety level elasticity is -3.34, which is much greater than the elasticity of -2.14 for the Jacks brand as a whole. The compositional effect we saw with Tombstone does not arise here, because the elasticity of the Jacks meat variety with respect to the Jacks brand price is almost identical to that of all the other varieties.

VI.C. The Brand/Variety Switching Patterns Induced by Price Changes

As we noted in the introduction, the ability to predict not just how a price change will affect market share, but also to predict where the change in market share will come from, is the true *raison d'etre* of discrete choice demand models. Thus, in this sub-section, we examine the decomposition of variety level price elasticities into their source components.

For instance, we simulate that a 10% increase in the price of Tombstone sausage/pepperoni would cause its sales to drop by 17%. We can decompose this drop into three components: 27% is due to consumers switching to other varieties of Tombstone, 41% is due to switching to other brands while still choosing the sausage/pepperoni topping, and 32% is due to switching to other varieties of other brands.

For Jacks meat/supreme pizza we simulate that a 10% price increase would cause sales to drop by 33%. Of this drop, 43% is due to consumers switching to other varieties of Jacks. Only 14% is due to consumers switching to other brands while still choosing the meat/supreme topping, and 43% is due to switching to other varieties of other brands.

Notice that in the case of the less popular topping (meat/supreme) we see much more switching to other toppings within the same brand (43% vs. 27%), and much less switching to other brands while staying with the same topping (14% vs. 41%). Both patterns seem intuitive.

We also see a higher level of switching to completely difference brands/varieties (43% vs. 32%).

Viewed another way, we can decompose the variety level elasticities into the parts due to switching varieties (within brand), switching brand (within variety) and switching brand and variety. The elasticity of -1.68 for Tombstone sausage/pepperoni decomposes into 0.45, 0.69 and 0.54, respectively. The elasticity of -3.34 for Jacks meat/supreme decomposes into 1.43, 0.46 and 1.45, respectively. Clearly, the larger elasticity for the smaller market share variety is due to both (i) a greater propensity to switch to other varieties within the same brand, and (ii) a greater propensity to switch to other brands/varieties.

For both the large and medium market share varieties, the preferred MM-MNL model implies that, after a price increase, *the majority of switchers go to a different brand, rather than a different variety of the same brand* (the fraction who completely switch brand are 73% and 57% in the two cases). This is also true for small market share brands and small varieties.[31] Thus, our results suggest brand loyalty is not strong enough to keep the majority of switching consumers within the same brand if one variety of that brand increases its price. This finding highlights the importance of considering not only brands but also varieties when studying consumer choice behavior. This, in turn, highlights the importance of methods to deal with very large choice sets.

VI. Conclusion

The structure of consumer heterogeneity in discrete choice demand models is crucial, as this determines the pattern of own and cross-price elasticities of demand. In recent years the most popular choice model among sophisticated practitioners has been the mixed logit (MIXL). This is an extension of the simple multinomial logit (MNL) that allows for population heterogeneity in the parameter vector. In most applications it is assumed to be normally distributed, leading to the logit model with normal mixing (N-MIXL). Fiebig et al (2010) introduced a generalization of N-MIXL that allows heterogeneity in the scale of the errors. This is known as the generalized multinomial logit (G-MNL). It has also become popular lately, presumably because it is hardly more difficult to estimate than N-MIXL yet often gives significant improvements in fit.

Another approach to modeling heterogeneity is the mixture-of-normals approach that has

[31] For instance, we also looked at Red Baron bacon/burger, which has a share of only 0.47%. We simulate a 10% price increase would cause sales to drop by 29%. Of this drop, 24% is due to consumers switching to other varieties of Red Baron. Only 7% is due to consumers switching to other brands while still choosing the bacon/burger topping, and 69% is due to switching to other varieties of other brands. The total fraction who switch to other brands is 76%.

been advocated by Geweke and Keane (1997, 2001) and Rossi, Allenby and McCulloch (2000). The appeal of this approach is that, as shown by Ferguson (1973), a discrete mixture-of-normals can approximate any heterogeneity distribution arbitrarily closely. But the mixture-of-normals approach – which we call MM-MNL – has not achieved the popularity of N-MIXL or G-MNL. We suspect this is because mixture-of-normals models are more complex to estimate, tend to proliferate parameters, and require judgement in choosing the number of mixture elements.

In this paper we have used case studies of three data sets in an attempt to assess the type of choice environments where the discrete mixture-of-normals approach may lead to substantial improvements in fit over G-MNL and N-MIXL. We find that the relative performance of the MM-MNL model tends to improve as the complexity of the heterogeneity structure increases:

First, we find that as the heterogeneity distribution departs more sharply from normality, the advantage of MM-MNL over N-MIXL becomes greater. This may seem obvious *ex post*, but *ex ante* many researchers have been skeptical of whether discrete choice data is rich enough to detect departures from normality, and hence whether generalization of N-MIXL is worth the effort (see discussion in Geweke and Keane (1999, 2001)).

Second, we find that G-MNL is very well suited to capture certain types of departure from normality. It can capture distributions that exhibit excess kurtosis near zero and fat tails. The very good performance of G-MNL relative to either MM-MNL or N-MIXL on SP data, documented by Fiebig et al (2010) and Keane and Wasi (2013), results from the fact that this type of departure from normality is very common in SP data.

Third, however, we also find there are many other types of departure from normality that G-MNL cannot capture. These include bi-modal distributions and distributions with kurtosis at a point away from the origin. Such distributions are apparent in the RP data that we examined. As a result, MM-MNL fits much better than either G-MNL or N-MIXL.

Our results suggest the mixture-of-normals approach should receive more attention from applied researchers. The obvious virtue of MM-MNL is its flexibility. It does well in all three datasets because a discrete mixture-of-normals can approximate any density (Ferguson (1973)).

Of course, the advantage of G-MNL over MM-MNL is ease-of-use. MM-MNL requires one to estimate a set of models with different numbers of mixture elements. One then chooses the best model from this set (e.g., using a criterion like BIC). But in G-MNL all parameters that determine the nature of heterogeneity are estimated simultaneously. Furthermore, as we have seen, the MM-MNL approach requires the researcher to put some structure on error covariance matrices to avoid severe proliferation of parameters as the number of elements of the mixture increases. This is a nontrivial exercise. An important avenue for future research is to develop software that automates MM-MNL estimation to make it more accessible for applied researchers.

Alternatively, another potential avenue for future research is to improve on G-MNL by using more flexible distributions for the scale factor. Then it may be possible to obtain a model that is competitive with MM-MNL in the sense that it can capture a wider range of departures from normality, but that still maintains the ease-of-use of the G-MNL framework.

Our paper also makes a substantive contribution by shedding light on differences in the structure of heterogeneity in stated preference (SP) vs. revealed preference (RP) data. This is an important issue because in recent years it has become very common to use SP data to study consumer demand. Prior work has compared aggregate market share predictions or mean parameter vector estimates from models based on SP vs. RP data.[32] But to our knowledge the present study is the first to compare heterogeneity in the distributions of tastes.

We find substantial differences in the nature of heterogeneity in choice models estimated on RP vs. SP data. This has two key implications: (i) as we noted above, it affects which model of heterogeneity is best for each type of data, (ii) it raises obvious questions about whether SP data is reliable for predicting demand, identifying consumer segments, calculating welfare, etc.

In the SP data, heterogeneity in coefficients on price, quality and other key attributes is characterized by a simple structure: relative to a normal distribution, there is excess kurtosis near zero (i.e., many consumers seem to care little about attributes), as well as a mass of consumers in the tail who care greatly about one or two attributes. Our SP results only apply to two datasets, but Fiebig et al (2010) found similar patterns in eight other SP datasets on different products. Thus, these behavioral patterns appear to be typical in SP data.

As Fiebig et al (2010) note, this is consistent with a scenario where subjects in the choice task use simple rules like: “always choose the cheaper option” or “always choose the high quality option” or “choose randomly.” Descriptive analysis of the SP data reveals groups of subjects whose behavior is consistent with such rules. In contrast to SP, there is little evidence of lexicographic or “random” behavior in the RP data that we examined.

[32] These studies have generally found that SP data is fairly reliable for such purposes – at least after adjustment of the scale of the error terms (see Ben-Akiva and Morikawa (1990), Adamowicz et al (1994), Cameron et al (2002)).

Perhaps the most obvious (and negative) interpretation of the SP results is that they reflect a failure of subjects to take the SP task seriously. If a subject is interested in minimizing his/her effort in the experiment, adopting a simple rule like "always choose the cheapest option" or "choose randomly" provides an easy way to get through the task. In RP data, where choices actually "matter" for utility, consumers will presumably weigh the options more carefully.

But we hesitate to adopt such a negative conclusion. There is a substantial literature showing that SP data can be useful for demand forecasting. Furthermore, our analysis of SP data on pizza delivery produced two insights of managerial interest:

First, we find consumers' willingness-to-pay for hot delivery is roughly equal to the price differential between delivered and frozen pizza (which is a very intuitive result given the close competition that now exists between frozen and delivered pizza). Thus, pizzerias must focus on quality delivery to maintain their competitive position relative to frozen pizza.

Second, our SP models indicate the optimal price for a typical delivered pizza should be at least AUS $24. This is consistent with the minimum price that most pizzerias in Sydney set for delivery. The fact that the SP data produce two such intuitive and useful results suggests that it does contain useful information.

Thus we think future research should focus on two ideas: (i) improving incentives for SP subjects to take the budget constraint seriously, and (ii) analysis of SP data using "process heterogeneity" models of the type proposed by Keane and Thorp (2016), where a fraction of subjects make optimal choices, while a fraction follow one or more sub-optimal rules of thumb.

Finally, our paper makes two other contributions: First, we show that use of random subsets of the full choice set is a reliable procedure to reduce computationally burden in discrete models with heterogeneity. This finding has potentially wide application in the many contexts where large choice sets are a problem, such as choice of homes or residential location, choice of colleges and majors, choice of TV shows or movies, choice of occupation at a detailed level, etc..

Second, in an application to choice among roughly 100 types of frozen pizza at the UPC level, we find that consideration of the full choice set, including the many varieties that exist within brands, can be very informative. For instance, we find that if one variety of a brand increases its price, most switching consumers actually go to other brands. This implies that accounting for "loyalty to varieties" (i.e., heterogeneity in tastes for varieties) is quite important for understanding brand switching behavior.

References

Adamowicz, W., J. Louviere and M. Williams (1994), “Combining revealed and stated preference methods for valuing environmental amenities,” Journal of Environmental Economics and Management, 26, 271-292.

Allenby and P. Rossi (1998), “Marketing Models of Consumer Heterogeneity,” *Journal of Econometrics*, 89(1-2), p. 57-78.

Anand, B. and R. Shachar (2011) “Advertising, the Matchmaker,” *RAND Journal of Economics*, 42 (2), 205-245.

Andrews, R.L. and A.K. Manrai (1999), “MDS Maps for Product Attributes and Market Response: An Application to Scanner Panel Data,” *Marketing Science,* 18(4), 584-604.

Arcidiacono, P. (2015). “Affirmative Action in Higher Education: How do Admission and Financial Aid Rules Affect Future Earnings?" *Econometrica,* 73(5), 1477-1524.

Ben-Akiva, M. and T. Morikawa (1990), “Estimation of switching models from revealed preferences and stated intentions,” Transportation Research Part A, 24(6), 485-495.

Berry, Steven (1994). “Estimating Discrete Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242-262.

Berry, Steven, James Levinsohn and Ariel Pakes (1995). “Automobile Prices in Market Equilibrium," Econometrica, 60(4), 889-917.

Brownstone, D., D.S. Bunch and K. Train (2000), “Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles,” *Transportation Research Part B,* 34(5), 315-338.

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

Cameron, T.A., G.L. Poe, R.G. Ethier and W.D. Schulze (2002), “Alternative non-market value elicitation methods: are the underlying preferences the same?,” *Journal of Environmental Economics and Management,* 44, 391:425.

Domanski, A. and R.H. von Haefen (2011), “Estimating Mixed Logit Recreation Demand Models with Large Choice Sets,” Working paper, North Carolina State University.

Elrod, Terry and Michael P. Keane (1995), “A Factor Analytic Probit Model for Representing the Market Structure in Panel Data,” *Journal of Marketing Research*, 32, 1-16.

Erdem, T., Imai, S. and M. Keane (2003), “Brand and Quantity Choice Dynamics under Price Uncertainty,” *Quantitative Marketing and Economics*, 1:1, 5-64.

Fader, P.S. and B.G.S. Hardie (1996), “Modeling Consumer Choice among SKUs,” *Journal of Marketing Research*, 33(4), 442-452.

Ferguson, T.S. (1973), A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1: 209-230.

Fiebig, D., M. Keane, J. Louviere and N. Wasi (2010), The Generalized Multinomial Logit Model: Accounting for scale and coefficient heterogeneity, *Marketing Science* 29: 393-421.

Frozen Food Age (2007), “Pizza Nation,” 55:10 (May), p. 22.

Frozen Food Age (2008), “Firing on All Cylinders,” 56:8 (March), p. 16.

Guevara, C.A. and Ben-Akiva, M. (2013), Sampling of alternatives in logit mixture models, *Transportation Research Part B*, 58, 185-198.

Geweke, J. and M. Keane (1999), Mixture of Normals Probit Models. in *Analysis of Panels and Limited Dependent Variable Models*, Hsiao, Lahiri, Lee and Pesaran (eds.), Cambridge University Press, 49-78.

Geweke, J. and M. Keane (2001), Computationally Intensive Methods for Integration in Econometrics. In *Handbook of Econometrics: Vol. 5*, J.J. Heckman and E.E. Leamer (eds.), Elsevier Science B.V., 3463-3568.

Geweke, J. and M. Keane (2007), Smoothly Mixing Regressions. *Journal of Econometrics* 138: 291-311.

Harris, K. and M. Keane (1999), “A Model of Health Plan Choice: Inferring Preferences and Perceptions from a Combination of Revealed Preference and Attitudinal Data,” *Journal of Econometrics*, 89: 131-157.

Keane, Michael P. (1992), “A Note on Identification in the Multinomial Probit Model,” *Journal of Business and Economic Statistics*, 10:2, 193-200.

Keane, Michael P. (1997), “Modelling Heterogeneity and State Dependence in Consumer Choice Behaviour,” *Journal of Business and Economic Statistics*, 15:3, 310-327.

Keane, Michael P. (2010), “A Structural Perspective on the Experimentalist School,” *Journal of Economic Perspectives*, 24(2), 47-58.

Keane, M.P. and S. Thorp (2016). Complex Decision Making: The Roles of Cognitive Limitations, Cognitive Decline and Ageing, *The Handbook of Population Ageing*, Elsevier, J. Piggott and A. Woodland (eds), forthcoming.

Keane, M.P. and N. Wasi (2013), “Comparing Alternative Models of Heterogeneity in Consumer Choice Behavior,” *Journal of Applied Econometrics*, 28:6, 1018-1045.

Lancaster, Kelvin J. (1966), "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132-157.

Louviere, Jordan J., Robert J. Meyer, David S. Bunch, Richard Carson, Benedict Dellaert, W. Michael Hanemann, David Hensher and Julie Irwin (1999), “Combining Sources of Preference Data for Modelling Complex Decision Processes,” *Marketing Letters*, 10:3, 205-217.

Louviere, J.J., R.T. Carson, A. Ainslie, T. A. Cameron, J. R. DeShazo, D. Hensher, R. Kohn, T. Marley and D.J. Street (2002), “Dissecting the random component of utility,” *Marketing Letters*, 13, 177-193.

Louviere, J.J., T. Islam, N. Wasi, D.J. Street and L. Burgess (2008), “Designing discrete choice experiments: Do optimal designs come at a price?,” *Journal of Consumer Research*, 35, 360-375.

McConnell, K.E. and W.Tseng (2000), "Some Preliminary Evidence on Sampling of Alternatives with the Random Parameters Logit," *Marine Resource Economics*, 14, 317–332.

McFadden, D. (1974), Conditional Logit Analysis of Qualitative Choice Behavior, in *Frontiers in Econometrics*, in P. Zarembka (ed.), New York: Academic Press, 105-42.

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

McFadden, D. and K. Train (2000), "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447-470.

Meyer, Robert. J. and Jordan J. Louviere (2007), "Formal Choice Models of Informal Choices: What Choice Modelling Research Can (and Can't) Learn from Behavioral Theory", *Review of Marketing Research*, 4, (in press).

Narella, S. and C. Bhat (2004), "A Numerical Analysis of the Effect of Sampling of Alternatives in Discrete Choice Models," *Transportation Research Record*, 1894, pp. 11-19.

Nevo, A. (2000). "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *The RAND Journal of Economics*, 31(3), 395-421.

Revelt, D. and K. Train (1998), "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level," *Review of Economics and Statistics* 80(4), 647-657.

Rossi, P., Allenby, G. and R. McCulloch (2005), *Bayesian Statistics and Marketing*, John Wiley and Sons, Hoboken, N.J..

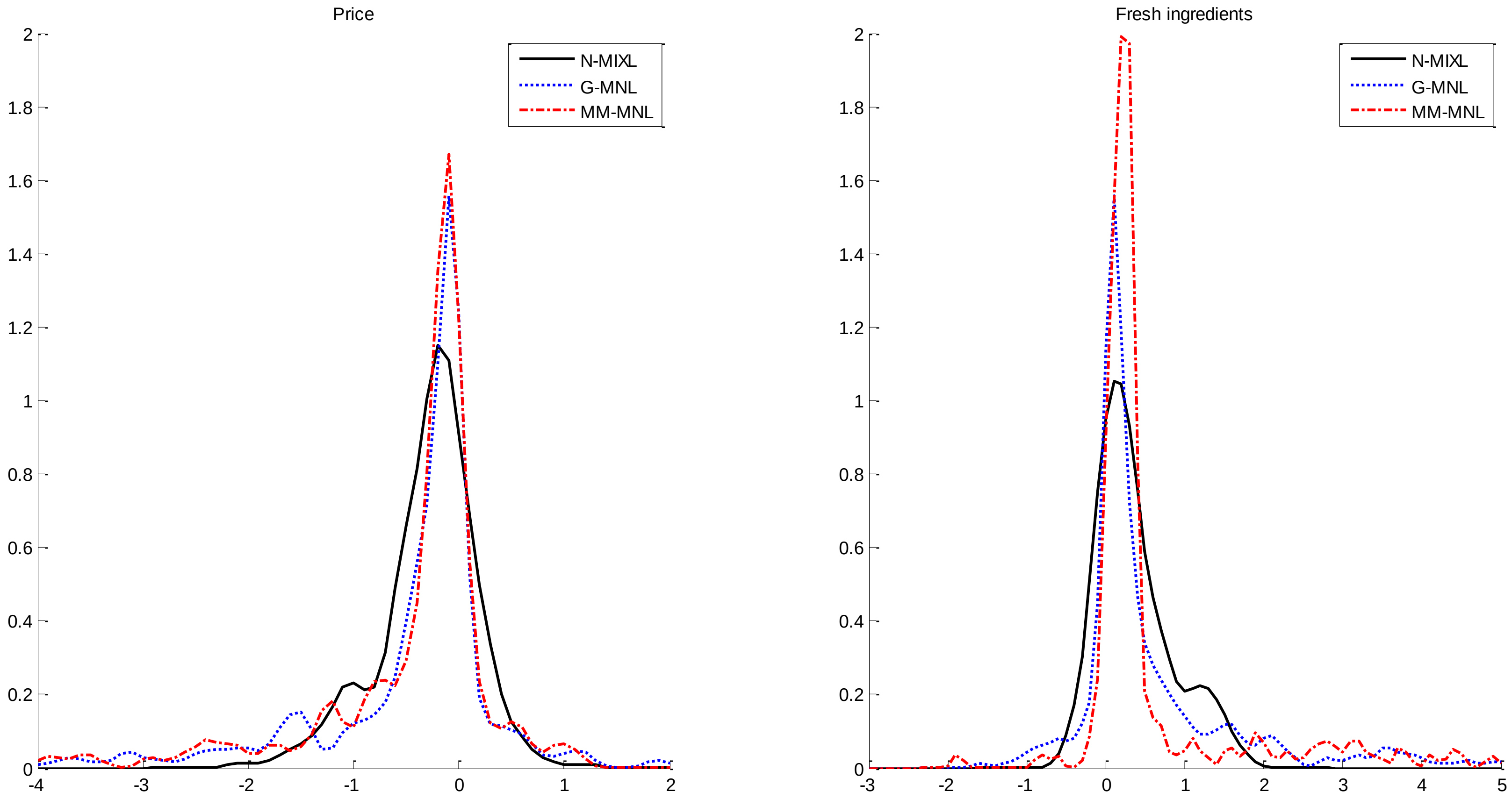
Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press.

The New York Times (2002), "Frozen Foods Show Upturn in Taste and Sales," Feb. 23, p. C3.

The New York Times (2004), "How Frozen Pizza Got Hot," May 2, Section 3, p. 2.

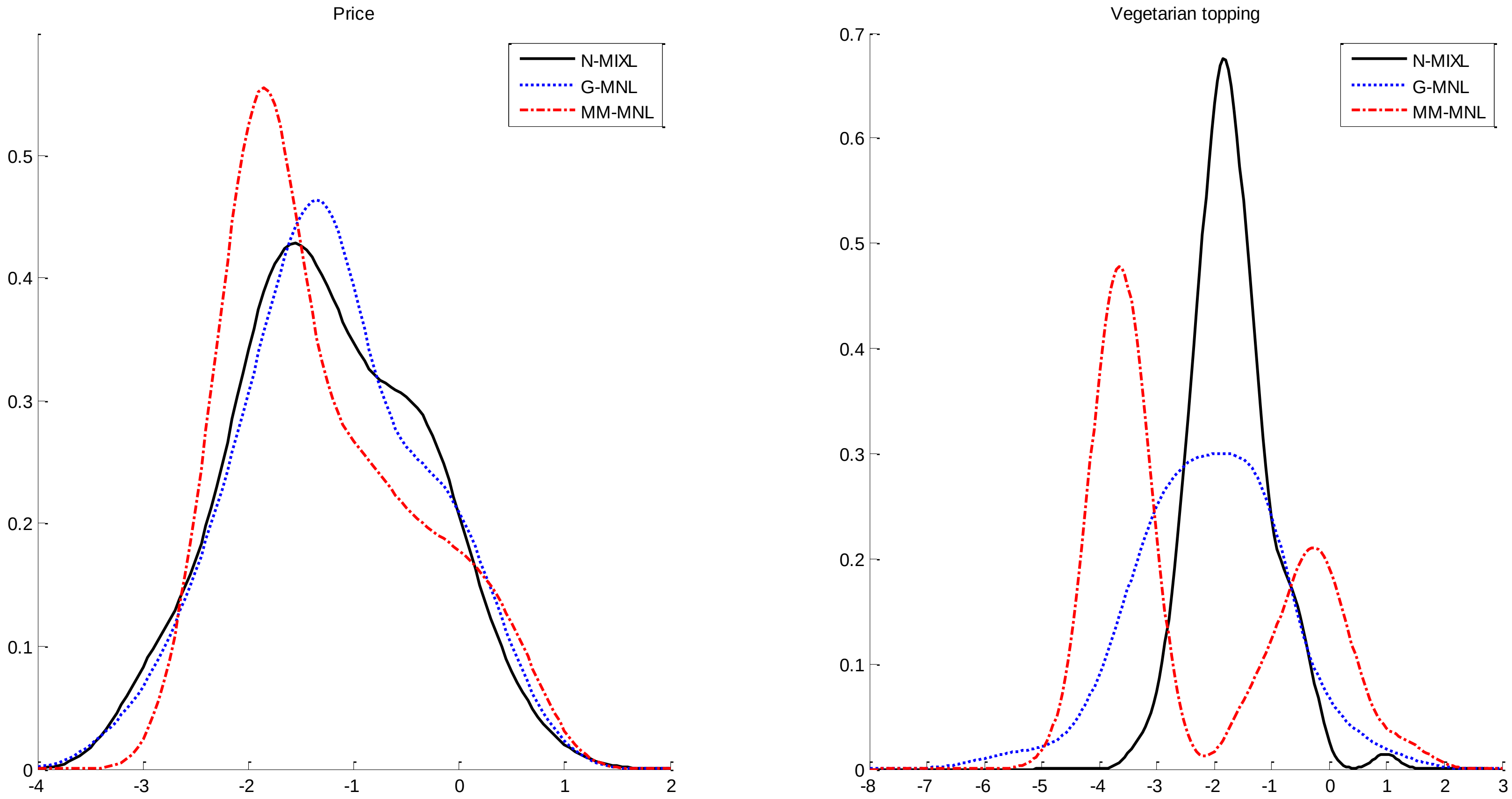
von Haefen, R. and A. Domanski (2013) "Estimating Mixed Logit Models with Large Choice Sets," Paper presented at the *3rd International Choice Modelling Conference*, Sydney, www.icmconference.org.uk/index.php/icmc/ICMC2013/paper/viewFile/755/215.

Figure 1: Posterior Distribution of Individual-level PRICE and FRESH INGREDIENT Coefficients from Pizza B Dataset



Note: Each kernel density estimate uses a normal kernel with an optimal bandwidth (h). The formula used is $h = \sigma(4/3N)^{1/5}$ where σ is the standard deviation and N is the number of observations. The optimal bandwidths for N-MIXL, G-MNL and MM-MNL for the price coefficients are .111, .095 and .086, respectively. For the fresh ingredient coefficients, the optimal bandwidths used are .135, .096 and .063, respectively.

Figure 2: Posterior Distribution of Individual-level PRICE and Vegetarian Topping Coefficients from Scanner Data



Note: Each kernel density estimate uses a normal kernel with an optimal bandwidth (h). The formula used is h = σ(4/3N)$^{1/5}$ where σ is the standard deviation and N is the number of observations. The optimal bandwidths for N-MIXL, G-MNL and MM-MNL for the price coefficients are .355, .346 and .303, respectively. For the vegetarian topping coefficients, the optimal bandwidths are .218, .513 and .380, respectively.

Figure 3: Comparison of the Normal Distribution and the Continuous Mixture-of-Scaled-Normals Distribution

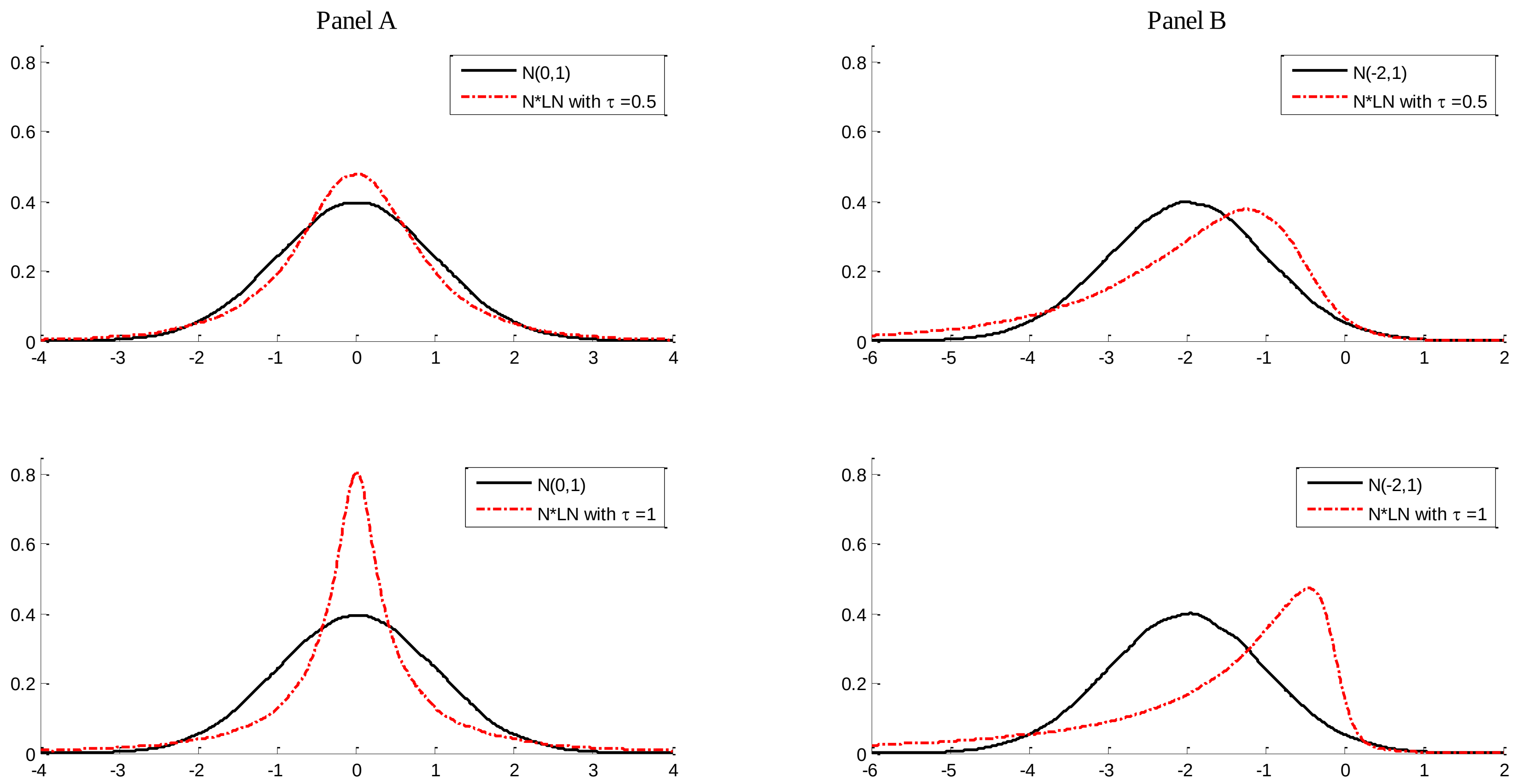
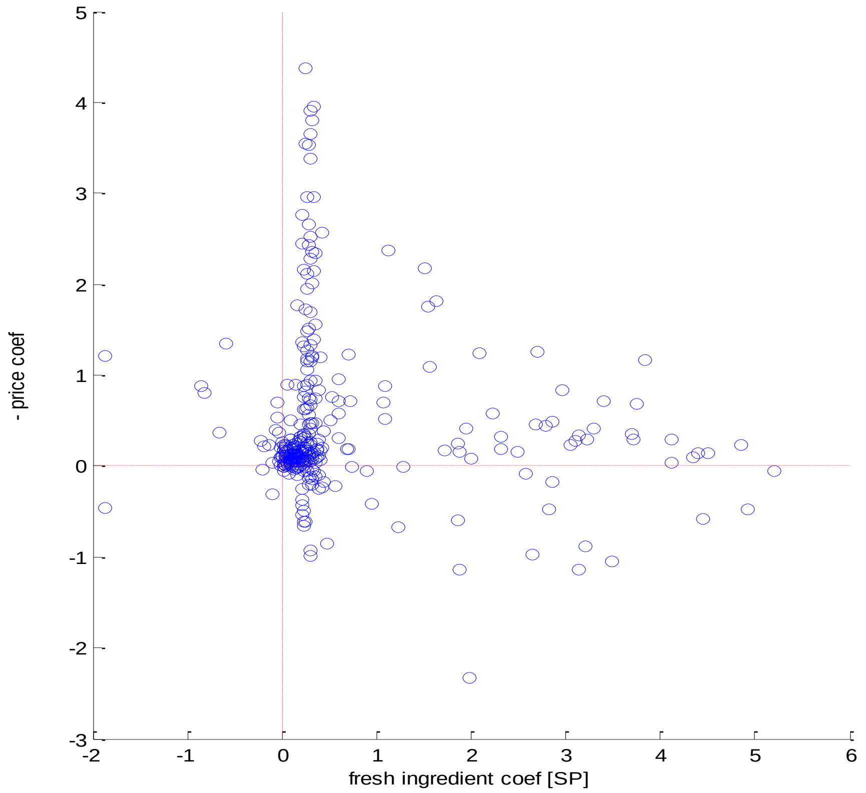


Figure 4: Price vs. Freshness Coefficients in the Pizza B data (MM-MNL model posteriors)



Note: The figure reports posterior mean parameter values for the price and freshness coefficients in the Pizza B dataset, as derived from the MM-MNL model. The figure plots the negative of the price coefficient.

Table 1: Average Characteristics of Alternatives in the RP Scanner Data

	All	Tombstone	Roma	Jacks	Red baron	Bernatello	Others
Choice Frequency		1085 (26%)	833 (20%)	844 (20%)	437 (11%)	238 (6%)	686 (17%)
Average characteristics of available choices (2001-2003)							
Price (*per pound*)	2.80	2.95	2.04	2.66	3.14	2.4	3.17
Price by topping							
Cheese only	2.95	3.25	2.21	2.9	3.17	2.53	3.27
Sausage/pepperoni	2.73	2.89	1.98	2.65	3.31	2.48	2.95
Meat/supreme	2.71	2.68	2.2	2.54	2.87	2.85	2.86
Bacon/burger	2.59	2.97	1.79	2.61	2.77	2.19	5.74
Chicken/Mexican	3.02	3.59	n/a	2.51	2.93	2.71	3.07
Vegetables	3.59	2.59	n/a	n/a	n/a	n/a	4.26
Combination/other	2.67	2.83	1.92	2.81	5.06	2.04	2.98
Price by crust							
Rising	2.59	2.59	1.87	2.54	2.65	1.72	3.4
Thin/crispy	2.73	2.96	2.31	2.71	2.99	2.46	3.02
Regular/other	2.98	3.06	1.8	2.76	3.39	2.5	3.11
Price by microwavable							
No	2.71	2.98	2.01	2.66	2.87	2.27	3.04
Yes	3.41	2.03	2.17	n/a	3.99	3.56	3.65
Promotion	0.019	0.018	0.023	0.029	0.004	0.008	0.027
Promotion by topping							
Cheese only	0.015	0.016	0.022	0.024	0.002	0.008	0.019
Sausage/pepperoni	0.017	0.014	0.017	0.031	0.002	0.005	0.032
Meat/supreme	0.013	0.015	0.036	0.029	0.002	0	0.006
Bacon/burger	0.017	0.012	0.007	0.022	0.014	0.019	0.038
Chicken/Mexican	0.023	0.037	n/a	0.045	0.013	0.007	0.014
Vegetables	0.014	0.027	n/a	n/a	n/a	n/a	0.005
Combination/other	0.036	0.015	0.028	0.041	0	0.006	0.07
Promotion by crust							
Rising	0.010	0.006	0.008	0.018	0	0	0.014
Thin/crispy	0.016	0.016	0.045	0.03	0	0	0.003
Regular/other	0.026	0.023	0.006	0.038	0.006	0.017	0.046
Promotion by microwavable							
No	0.017	0.017	0.022	0.029	0.003	0.009	0.02
Yes	0.029	0.043	0.026	n/a	0.006	0	0.052

Note: Means are taken over all available alternatives (not only the purchased alternative). An n/a indicates the option does not exist. From 2001-2003 the purchasing power parity exchange rate between the US$ and AUS$ was roughly 1.34 (see http://stats.oecd.org/Index.aspx?datasetcode=SNA_TABLE4#). So US$ figures can be increased by roughly 1/3 to put them in AUS$ terms. Also, the average pizza size is 1.2 pounds. So the mean price per pizza is 2.80×1.34×1.2= AUS$4.50.

Table 2: Pizza A (Stated Preference Data)

	MNL		S-MNL		N-MIXL[a]		G-MNL[a]		Latent class[b]						MM-MNL[c]			
									class 1		class 2		class 3		class 1		class 2	
	est	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Gourmet	0.02	0.02	0.03	0.04	0.03	0.05	**0.49**	0.30	-0.01	0.09	0.02	0.08	0.08	0.08	0.02	0.08	0.14	0.62
Price	**-0.16**	0.02	**-0.19**	0.05	**-0.35**	0.06	**-1.82**	0.85	**-0.20**	0.07	**-0.16**	0.07	**-0.39**	0.14	**-0.18**	0.07	-4.63	2.80
Ingredient freshness	**0.48**	0.03	**1.45**	0.29	**0.96**	0.10	**5.06**	2.17	**1.57**	0.16	**0.12**	0.06	0.30	0.08	**0.59**	0.10	13.47	7.47
Delivery time	**0.09**	0.03	**0.16**	0.08	**0.16**	0.05	**0.81**	0.43	0.10	0.11	**0.10**	0.05	**0.32**	0.15	0.06	0.05	3.95	2.43
Crust	0.02	0.03	0.01	0.04	0.02	0.07	0.48	0.33	**-0.12**	0.11	0.01	0.07	**-0.30**	0.18	-0.06	0.09	1.18	1.11
Sizes	**0.09**	0.03	**0.12**	0.06	**0.20**	0.05	**0.88**	0.44	**0.15**	0.08	0.06	0.06	**0.23**	0.08	**0.23**	0.07	0.92	0.92
Steaming hot	**0.38**	0.03	**1.02**	0.24	**0.87**	0.09	**4.86**	2.10	**0.50**	0.12	**0.12**	0.08	**1.60**	0.26	**0.50**	0.09	9.85	5.49
Late open hours	**0.04**	0.02	0.08	0.06	0.07	0.05	0.30	0.21	0.09	0.08	**0.06**	0.05	0.02	0.10	**0.12**	0.06	-0.97	0.95
τ			**1.69**	0.18			**1.80**	0.27										
γ							-0.02	0.02										
Class probability									**0.36**	0.04	**0.32**	0.06	**0.23**	0.05	**0.57**	0.05	**0.43**	0.05
No. of parameters	8		9		16		18		35						33			
LL	-1657		-1581		-1403		-1372		-1418						-1328			
BIC	3378		3233		2933		**2886**		3115						2919			

Note: [a] estimates from uncorrelated coefficient specification; [b] estimates from LC with 4 classes; [c] estimates from MM-MNL with 2 independent normals. Bold estimates are statistically significant at 5%. S-MNL, N-MIXL, G-MNL and MM-MNL are estimated by simulated maximum likelihood with 500 draws. The standard errors are calculated using 5000 draws. All attributes have 2-levels and are coded using the "effects coding" method that is common in DCE studies. For example, gourmet = 1 or -1; a price of $13 is coded as -1, and a price of $17 is coded as 1; free delivery is coded as 1 and a delivery charge of $2 is coded as -1. For further details see Fiebig et al. (2010).

According to G-MNL, hot delivery increases utility by 4.86x2=9.72, while a $4 price reduction increases utility by 2x1.82 = 3.64. Thus, the WTP for hot delivery is (9.72/3.64)x4 = $10.68. Note that the coefficient on a $1 price change would be (-1.82x2)/4 = -0.91, and the coefficient on hot delivery if it were coded 1/0 would be 9.72. So we have 9.72/0.91=$10.68.

Table 3: Pizza B (Stated Preference Data)

	MNL		S-MNL		N-MIXL[a]		G-MNL[a]		Latent class[b]						MM-MNL[c]					
									class 1		class 2		class 3		class 1		class 2		class 3	
	est	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Gourmet	0.01	0.01	**0.05**	0.01	0.01	0.02	0.02	0.05	0.01	0.03	0.02	0.10	0.09	0.15	-0.03	0.04	-0.12	0.08	**0.37**	0.09
Price	**-0.17**	0.01	**-0.25**	0.02	**-0.30**	0.04	**-0.94**	0.14	-0.04	0.03	**-1.71**	0.32	0.24	0.30	-0.10	0.04	**-0.86**	0.13	-0.17	0.15
Ingredient freshness	**0.21**	0.01	**0.36**	0.03	**0.34**	0.04	**1.22**	0.18	**0.10**	0.02	**0.46**	0.18	**2.17**	0.31	**0.12**	0.03	**0.29**	0.08	**1.02**	0.24
Delivery time	**0.03**	0.01	0.04	0.02	0.05	0.02	**0.21**	0.07	0.02	0.02	0.14	0.06	-0.03	0.16	0.02	0.03	0.19	0.10	0.14	0.09
Crust	**0.08**	0.01	**0.09**	0.01	**0.08**	0.04	**0.64**	0.12	**-0.04**	0.02	-0.05	0.09	**0.31**	0.56	-0.03	0.03	**0.62**	0.20	0.15	0.09
Sizes	**0.07**	0.01	**0.08**	0.02	**0.11**	0.02	**0.21**	0.05	**0.05**	0.02	**0.19**	0.08	**0.28**	0.10	0.06	0.03	**0.31**	0.09	**0.26**	0.11
Steaming hot	**0.20**	0.01	**0.35**	0.03	**0.34**	0.04	**1.42**	0.19	**0.10**	0.03	**0.22**	0.08	**0.67**	0.15	**0.11**	0.03	**0.37**	0.06	**1.43**	0.23
Late open hours	**0.04**	0.01	0.02	0.02	**0.08**	0.02	0.10	0.05	**0.04**	0.02	0.06	0.06	0.07	0.07	0.01	0.03	**0.29**	0.11	**0.19**	0.07
Free delivery charge	**0.12**	0.01	**0.15**	0.02	**0.20**	0.02	**0.69**	0.11	**0.11**	0.03	**0.56**	0.13	0.15	0.11	**0.22**	0.05	**0.26**	0.08	**0.28**	0.07
Local store	**0.08**	0.01	**0.06**	0.02	**0.15**	0.02	**0.60**	0.11	**0.14**	0.03	-0.01	0.06	0.10	0.07	**0.09**	0.03	**0.43**	0.13	0.08	0.09
Baking Method	**0.07**	0.01	**0.07**	0.02	**0.11**	0.02	**0.27**	0.05	**0.06**	0.02	0.16	0.06	**0.29**	0.10	0.01	0.03	**0.32**	0.07	**0.35**	0.17
Manners	0.01	0.01	-0.004	0.02	0.02	0.02	0.02	0.06	0.03	0.02	0.03	0.06	-0.06	0.11	0.03	0.03	-0.06	0.10	0.11	0.14
Vegetarian availability	**0.09**	0.01	**0.06**	0.01	**0.13**	0.04	**0.42**	0.12	0.02	0.02	**0.15**	0.07	0.04	0.32	0.04	0.03	**0.35**	0.16	0.04	0.07
Delivery time guaranteed	**0.07**	0.01	**0.07**	0.02	**0.11**	0.02	**0.15**	0.05	**0.08**	0.02	**0.17**	0.09	0.12	0.06	**0.14**	0.04	0.07	0.08	**0.19**	0.08
Distance to the outlet	**0.06**	0.01	0.04	0.02	**0.09**	0.02	0.08	0.05	**0.09**	0.02	0.11	0.05	-0.12	0.10	**0.11**	0.04	0.09	0.07	0.06	0.08
Range/variety availability	**0.06**	0.02	0.04	0.02	**0.09**	0.02	0.15	0.06	**0.07**	0.02	0.03	0.05	0.07	0.05	**0.10**	0.03	0.03	0.09	0.19	0.08
τ			**1.22**	0.08			**1.50**	0.09												
γ							-0.05	0.03												
Class probability									**0.51**	0.04	**0.14**	0.03	**0.12**	0.02	**0.41**	0.03	**0.31**	0.03	**0.28**	0.03
No. of parameters	16		17		32		34		101						98					
LL	-6747		-6607		-5892		-5662		-5591						-5310					
BIC	13641		13372		12081		11639		12118						**11527**					

Note: [a] estimates from uncorrelated coefficient specification; [b] estimates from LC with 6 classes; [c] estimates from MM-MNL with 3 independent normals. Bold estimates are statistically significant at 1%. S-MNL, N-MIXL, G-MNL and MM-MNL are estimated by simulated maximum likelihood with 500 draws. The standard errors are calculated using 5000 draws. All attributes have 2-levels and are coded using the "effects coding" method that is common in DCE studies. For example, gourmet = 1 or -1. For details see Fiebig et al. (2010). According to G-MNL, hot delivery increases utility by 1.42x2=2.84, while a $4 price reduction increases utility by 2x0.94 = 1.88. Thus, the WTP for hot delivery is (2.84/1.88)x4 = $6.04. Note that the coefficient on a $1 price change would be (-0.94x2)/4 = -0.47, and the coefficient on hot delivery if it were coded 1/0 would be 2.84. So we have 2.84/0.47=$6.04.

Table 4: Estimates from Revealed Preference Data

	MNL		**S-MNL**[a]		**N-MIXL**[b]		**G-MNL**[b]		**Latent class**[c]						**MM-MNL**[d]			
									class 1		class 2		class 3		class 1		class 2	
	est	**s.e.**	**est**	**s.e.**	**est**	**s.e.**	**est**	**s.e.**	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Brand [omitted others]																		
Tombstone	**0.87**	0.05	**0.68**	0.21	**0.52**	0.10	**0.48**	0.11	**2.04**	0.31	**0.64**	0.31	1.12	0.59	**1.71**	0.16	-0.45	0.23
Roma	**0.24**	0.05	-0.25	0.17	**-0.26**	0.11	**-0.34**	0.10	**-0.69**	0.30	**0.91**	0.36	-0.89	1.20	-0.30	0.17	**-0.58**	0.25
Jacks	**0.66**	0.05	0.25	0.21	**0.29**	0.10	0.01	0.12	**2.30**	0.32	0.17	0.36	0.24	0.47	**1.52**	0.18	**-1.98**	0.30
Red Baron	**0.13**	0.06	-0.12	0.20	-0.17	0.12	-0.17	0.10	-0.30	0.28	0.37	0.33	**1.09**	0.28	**0.55**	0.17	-0.43	0.23
Bernatello	**-1.02**	0.06	**-0.98**	0.15	**-0.98**	0.08	**-1.08**	0.10	**-0.68**	0.30	**-1.17**	0.27	**-1.40**	0.50	**-0.57**	0.10	**-1.35**	0.13
Price	**-0.84**	0.04	**-0.97**	0.06	**-1.21**	0.06	**-1.20**	0.06	**-1.85**	0.14	**-1.65**	0.24	**-1.07**	0.15	**-1.75**	0.07	**-0.48**	0.09
Promotion	**0.81**	0.10	**0.77**	0.13	**0.86**	0.17	**0.76**	0.15	**1.11**	0.22	**1.18**	0.21	-0.04	0.79	**0.92**	0.23	0.65	0.41
Toppings [omitted combo, others]																		
Cheese only	**0.16**	0.05	-0.01	0.05	-0.28	0.13	**-0.42**	0.14	-0.38	0.29	**-1.34**	0.39	0.52	0.28	-0.24	0.17	**-1.01**	0.30
Sausage/pepperoni	**1.02**	0.04	**0.85**	0.07	**1.02**	0.07	**1.04**	0.06	**1.17**	0.18	**1.14**	0.19	0.42	0.27	**0.61**	0.10	**1.53**	0.21
Meat/supreme	-0.06	0.05	-0.08	0.04	-0.04	0.08	-0.05	0.07	0.27	0.21	-0.05	0.16	-0.43	0.37	**-0.38**	0.12	-0.26	0.26
Bacon/burger	**-0.36**	0.06	**-0.44**	0.06	**-0.64**	0.10	**-0.83**	0.12	-0.21	0.24	-0.31	0.25	-0.89	0.50	**-0.62**	0.15	-0.22	0.27
Chicken/Mexican	**-0.53**	0.08	**-0.76**	0.09	**-0.71**	0.10	**-0.89**	0.14	**-1.02**	0.23	0.08	0.53	-0.55	0.54	**-1.19**	0.22	0.18	0.22
Vegetarian	**-1.07**	0.15	**-1.66**	0.17	**-1.75**	0.23	**-2.19**	0.34	**-3.08**	0.92	**-1.74**	0.63	-0.42	0.76	**-3.76**	0.45	-0.18	0.37
Crust [omitted regular, others]																		
Rising	**-0.85**	0.04	**-1.00**	0.06	**-1.16**	0.08	**-1.11**	0.10	**-1.45**	0.26	**-1.08**	0.23	**-1.10**	0.23	**-1.62**	0.14	**-1.04**	0.19
Thin/crispy	0.03	0.03	0.02	0.02	-0.05	0.07	-0.01	0.05	**-0.62**	0.12	**0.74**	0.18	-0.46	0.79	**-0.32**	0.11	0.24	0.17
Microwavable	0.07	0.06	**0.16**	0.04	0.02	0.08	0.01	0.10	0.64	0.37	**0.90**	0.26	-1.04	1.51	**0.67**	0.16	**-0.89**	0.23
τ			**0.86**	0.05			**0.40**	0.04										
γ							**0.79**	0.07										
class prob.									**0.28**	0.04	**0.24**	0.05	0.23	0.15	**0.65**	0.08	**0.35**	0.08
	16		27		48		50		84						66			
LL	-13602		-11930		-11048		-11017		-12116						-10802			
BIC	27337		24085		22497		22450		24931						**22153**			

Note: [a] estimates from S-MNL with random correlated (one-factor) intercepts; [b] estimates from correlated coefficients (imposing 1-factor structure on the covariance matrix); [c] estimates from LC with 5 classes; [d] estimates from MM-MNL with 2 proportional covariance matrices. Bold estimates are statistically significant at 5%. S-MNL, N-MIXL, G-MNL and MM-MNL are estimated by simulated maximum likelihood with 500 draws. The standard errors are calculated using 5000 draws.

Table 6: Model Fit for Different Consumer Types (RP vs. SP Data)

			Difference in Mean Attribute Levels between Chosen and Non-chosen Alternatives		Frequency	Average BIC Gain per Subject	
						MM-MNL over N-MIXL	G-MNL over N-MIXL
			Range	Average			
Stated Preference Data							
Price	Most sensitive	1	[-4, -1.75]	-2.41	26	2.39	1.24
[$13,$17]		2	[-1.5,-.75]	-1.03	58	-1.37	0.53
		3	[-.5,-.25]	-0.32	66	0.76	0.92
	Least sensitive	4	[0,2]	0.2	178	2.93	1.79
Ingredients	Most sensitive	1	[1,2]	1.26	31	3.09	2.31
1 = fresh		2	[.5,.875]	0.66	44	-2.1	0.65
-1 = canned		3	[.125,.375]	0.22	101	0.23	0.36
	Least sensitive	4	[-.5,0]	-0.08	152	3.47	2.01
Revealed Preference Data							
Price ($/lb)							
	Most sensitive	1	[-0.96,-0.56]	-0.67	26	1.03	-0.55
		2	[-0.55,-0.43]	-0.5	26	1.76	-0.4
		3	[-0.43,-0.30]	-0.37	26	2.18	-0.84
		4	[-0.30,-0.12]	-0.21	26	3.63	1.1
	Least sensitive	5	[-0.11,0.99]	0.21	25	4.79	2.56
Vegetarian topping							
1 if vegetarian topping	Most sensitive	1	[-0.032,0.1655]	0.02	25	4.9	1.13
0 otherwise		2	[-0.040,-0.0330]	-0.037	26	2.77	0.48
		3	[-0.0428,-0.0401]	-0.041	26	-1.89	-0.3
		4	[-0.0454,-0.0431]	-0.044	24	3.52	0.49
	Least sensitive	5	[-0.0538,-0.0455]	-0.048	28	4.07	0.05

Note: In the RP data, of the 25 consumers in the least price sensitive quintile, 14 have price coefficients that are positive (i.e., the wrong sign).

Table 7: MM-MNL Model Fit to Market Shares by Brand/Topping

Observed Market Shares

	cheese	sausage/ pepperoni	meat/ supreme	bacon/ burger	chicken/ mexican	Veg	Other	Total
Tombstone	1.71	11.03	4.21	2.48	1.76	0.33	4.79	**26**
Roma	3.76	10.97	1.95	0.69	-----	-----	2.84	**20**
Jacks	2.26	10.18	2.59	3.43	0.70	-----	1.31	**20**
Red Baron	2.71	4.54	2.34	0.47	0.45	-----	0.09	**11**
Bernatello	1.14	2.85	0.31	0.58	0.03	-----	0.86	**6**
Others	2.45	9.07	2.46	0.02	0.22	0.91	1.50	**17**
Total	**14**	**49**	**14**	**8**	**3**	**1**	**11**	**100**

Predicted Market Shares by MM-MNL Model

	cheese	sausage/ pepperoni	meat/ supreme	bacon/ burger	chicken/ mexican	Veg	Other	Total
Tombstone	2.56	11.55	4.02	3.03	0.89	0.66	3.49	**26**
Roma	3.86	11.48	1.49	0.48	-----	-----	2.72	**20**
Jacks	2.17	10.08	2.93	2.87	0.96	-----	1.45	**20**
Red Baron	2.34	4.42	2.71	0.64	0.54	-----	0.11	**11**
Bernatello	0.73	2.81	0.42	0.72	0.11	-----	1.07	**6**
Others	2.59	7.96	2.25	0.04	0.73	0.50	2.61	**17**
Total	**14**	**48**	**14**	**8**	**3**	**1**	**11**	**100**

Note: There are seven brand/topping combinations that do not exist in the data, leaving 37 options.

Table 8: Brand Level Price Elasticities Across Models and Choice Set Sizes

	Random Choice Set Size for Estimation	MNL	S-MNL	LC	MXL	G-MNL	MM-MNL
Tombstone	20	-1.66	-1.49	-1.75	-1.46	-1.40	-1.69
	30	-1.69	-1.48	-2.05	-1.49	-1.49	-1.72
	40	-1.71	-1.50	-2.09	-1.52	-1.42	-1.66
Jacks	20	-1.64	-1.52	-2.44	-1.81	-1.63	-2.06
	30	-1.67	-1.55	-2.45	-1.82	-1.69	-2.05
	40	-1.69	-1.55	-2.37	-1.84	-1.66	-2.14

Table 9: Decomposition of Variety Level Price Elasticities

A. Tombstone sausage/pepperoni (Share = 11%)

	Change in Market Share (percentage points)	Distribution of Switchers	Decomposition of elasticity
Tombstone Sausage/Pepperoni	-1.81		-1.68
Tombstone Other Varieties	+0.48	27%	0.45
Other Brands Sausage/Pepperoni	+0.75	41%	0.69
Other Brands Other Varieties	+0.58	32%	0.54

B. Jacks meat/supreme (Share = 2.9%)

	Change in Market Share (percentage points)	Distribution of Switchers	Decomposition of elasticity
Jacks Meat/Supreme	-0.97		-3.34
Jacks Other Varieties	+0.41	43%	1.43
Other Brands Meat/Supreme	+0.13	14%	0.46
Other Brands Other Varieties	+0.42	43%	1.45

Appendix: Using Random Subsets to Deal with Very Large Choice Sets

This appendix considers the problem of estimating discrete choice models with very large choice sets. Large choice sets are a common problem in empirical work: One example is the case of roughly 100 varieties of frozen pizza considered in the main text. Other examples include choice of homes or residential location (McFadden, 1978), choice of breakfast cereal (Nevo, 2000), choice of colleges and majors (Arcidiacono, 2015), choice of TV shows or movies (Anand and Shachar, 2011), choice of occupation at a detailed level, and so on.

Large choice sets create computational problems in estimating discrete choice models that allow for consumer taste heterogeneity. This is because choice probabilities in these models take the form of integrals whose dimension is comparable to the size of the choice set, and that usually have no closed form. In recent years, simulation methods have made estimation of such models feasible.[33] Nevertheless, if choice sets are large enough, the estimation of discrete choice models can still be very computationally burdensome, particularly given the very large sample sizes often available in modern datasets on consumer behavior.

The use of random subsets of the full choice set in estimation is a potential solution to the problem of very large choice sets. In a classic paper, McFadden (1978) showed that the use of random subsets of the full choice set has no effect on the consistency of parameter estimates in the multinomial logit model (MNL).[34] However, as we discuss below, McFadden's result does not go through in models with unobserved consumer taste heterogeneity. Nevertheless, in this Appendix we show that use of random subsets can be a useful device to reduce computational burden even in models with heterogeneity. The structure of the Appendix is as follows:

In Appendix A.1 we present the result of McFadden (1978), who proved that the use of randomly selected subsets of the full choice set has no effect on consistency of MNL. We then show why this result does not go through for models with heterogeneous preferences.

In Appendix A.2 we present Monte Carlo evidence on the bias from using randomly selected subsets of the full choice set in the N-MIXL, G-MNL and MM-MNL models. Our results suggest the bias from this procedure is negligible.[35] We give an intuition for this result.

In Appendix A.3 we present results using the actual RP data on demand for frozen pizza

[33] See McFadden (1989), Pakes (1986), Keane (1994), McCulloch and Rossi (1994) and Geweke and Keane (2001).

[34] Given the great speed of modern computers, it is now feasible to estimate MNL even with very large choice sets and very large datasets. This is because MNL generates closed form expressions for the choice probability integrals.

[35] However, we show that use of reduced choice sets does lead to bias in information criteria like BIC that are used to compare models (such that BIC is biased towards models with fewer parameters).

used in the text. The full choice set has roughly 100 alternatives, and we present results using random subsets of 20, 30 or 40 alternatives. We find parameter estimates are very stable across the three choice set sizes. This gives us confidence that use of random subsets of the full choice set is a reliable procedure.

A.1. Previous Studies on Logit with Many Alternatives

McFadden (1978) showed that one can consistently estimate parameters of MNL using randomly selected subsets of the full choice set (i.e., hypothetical choice sets that include the chosen alternative plus a random subset of the non-chosen alternatives). To see why, we need some notation. Let j_n denote the observed choice of person n. Let C denote the full choice set which has J elements, and let D_n denote a subset of C. This subset is randomly constructed *except* that it must contain j_n. Let $\pi(D_n|j_n)$ denote the probability that subset D_n is constructed from all the possible subsets of C that contain j_n. Finally, let $P(j|\theta^*,C,x_n)$ be the probability that j is chosen from C, where θ^* is the true parameter vector and x_n is the matrix of attributes of the J alternatives.

Note that $\pi(D_n|j_n)$ is a function chosen by the researcher. For example, if J=100, then one choice is the following rule: include j_n in D_n and then chose 19 addition alternatives by sampling without replacement from the remaining 99 elements of C. Then D_n is a hypothetical choice set with 20 elements. Here $\pi(D_n|j_n)$ is a constant over all possible D_n, and $\pi(D_n|k)=\pi(D_n|j)$ for $k,j\in D_n$. Also note that, as the chosen alternative is always included in D_n, we have $\pi(D|\, j_n)=0$ if $j_n \notin \mathrm{D}$.

Now consider the hypothetical log-likelihood we would construct if each consumer n=1,…,N chose from the hypothetical choice set D_n. Let $U_j(\theta,x_{nj})$ denote the "deterministic" part of utility in the logit model, which excludes the additive *iid* extreme value error terms. Then the likelihood is given by the simple MNL formula:

$$LL_N(\theta)=\frac{1}{N}\sum_{n=1}^{N}\ln\frac{\exp[U_{j_n}(\theta,x_{n,j_n})]}{\sum_{j\in D_n}\exp[U_j(\theta,x_{nj})]} \qquad \text{(A1)}$$

Here we have suppressed the t subscripts to conserve on notation (i.e., we consider a single choice occasion per consumer). Taking the expectation of (A1) over realizations from the data generating process (and viewing the x_n as fixed) we obtain:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k\in C}\sum_{D\subset C} P(k\,|\,\theta^*, C, x_n)\pi(D\,|\,k)\ln\frac{\exp[U_k(\theta, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta, x_{nj})]} \tag{A2}$$

It is important to note that $\pi(D|k)$=0 if $k\notin$ D, and that the third summation is over all possible $D\subset C$, regardless of whether they have positive probability. Next, we multiply and divide $P(k\,|\,\theta^*, C, x_n) = \exp[U_k(\theta^*, x_{nk})] \Big/ \sum_{j\in C}\exp[U_j(\theta^*, x_{nj})]$ by $\sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]$. This gives:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k\in C}\sum_{D\subset C} \frac{\sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]}{\sum_{j\in C}\exp[U_j(\theta^*, x_{nj})]}\frac{\exp[U_k(\theta^*, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]}\pi(D\,|\,k)\ln\frac{\exp[U_k(\theta, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta, x_{nj})]} \tag{A3}$$

It is convenient to define $R_n(D, C, \theta^*) \equiv \sum_{j\in D}\exp[U_j(\theta^*, x_{nj})] \Big/ \sum_{j\in C}\exp[U_j(\theta^*, x_{nj})]$.[36] Now, as $R_n(D,C,\theta^*)$ does not depend on k (but only on D) we can bring the sum over k inside to obtain:[37]

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{D\subset C} R_n(D, C, \theta^*)\cdot\sum_{k\in C}\left\{\frac{\exp[U_k(\theta^*, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]}\pi(D\,|\,k)\ln\frac{\exp[U_k(\theta, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta, x_{nj})]}\right\} \tag{A4}$$

Next we multiply and divide by $\pi(D|j)$, utilizing what McFadden (1978) calls the "uniform conditioning" property of π – i.e., the fact that $\pi(D|k)=\pi(D|j)=\pi(D)$ if $k,j\in D$ – to obtain:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{D\subset C} R_n(D, C, \theta^*)\pi(D)\cdot\sum_{k\in C}\left\{\frac{\exp[U_k(\theta^*, x_{nk})]\pi(D\,|\,k)}{\sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]\pi(D\,|\,j)}\ln\frac{\exp[U_k(\theta, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta, x_{nj})]}\right\}$$

Now focus on the term $\exp[U_k(\theta^*, x_{nk})]\pi(D\,|\,k) \Big/ \sum_{j\in D}\exp[U_j(\theta^*, x_{nj})]\pi(D\,|\,j)$. Because $\pi(D|k)$=0 if $k\notin$ D, the numerator vanishes in all cases except where $k\in D$. Also, because $\pi(D|k)=\pi(D|j)$ if

[36] Note that $R_n(D,C,\theta^*)$ is the ratio of the probability of choosing $k\in D$ from full choice set C to that of choosing k from the reduced choice set D (That is, the ratio of the denominator of the logit choice probability for the reduced choice set D to the full choice set C). Clearly $R_n(D,C,\theta^*)<1$ as $D\subset C$.

[37] At this point in the proof it is crucial to note that the third summation in (A3) is over all possible $D\subset C$, regardless of whether they have positive probability. Specifically, the set of D that is summed over here is not constrained by the fact that any D that does not contain k has zero probability. This is dealt with via the $\pi(D|k)$ term.

$k,j \in D$, the π terms cancel out and we are left with:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{D\subset C} R_n(D,C,\theta^*)\pi(D)\cdot\sum_{k\in D}\left\{\frac{\exp[U_k(\theta^*,x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta^*,x_{nj})]}\ln\frac{\exp[U_k(\theta,x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta,x_{nj})]}\right\} \quad \text{(A5)}$$

Notice that the third summation term has the form $\sum_k \{P_k(\theta^*)\ln P_k(\theta)\}$. For any sequences of numbers $\{P_1,\ldots,P_J\}$ and $\{P_1^*,\ldots,P_J^*\}$ such that $\sum_k P_k = 1$, $\sum_k P_k^* = 1$, and $P_k > 0$, $P_k^* > 0$ for all k, where the P_k^* are given and the P_k are to be chosen, it is simple to show that $\sum_k \{P_k^* \ln P_k\}$ is maximized by setting $P_k = P_k^*$ for all k.[38] In equation (A5) the P_k and P_k^* correspond to the logit probability expressions, and we achieve equality (for all k) by setting $\theta = \theta^*$. This completes the proof.

Returning to (A1), note that the correct log-likelihood for the MNL model can be written:

$$LL_N^C(\theta) = \frac{1}{N}\sum_{n=1}^{N}\ln\frac{\exp[U_{j_n}(\theta,x_{n,j_n})]}{\sum_{j\in D_n}\exp[U_j(\theta,x_{nj})]}R_n(D,C,\theta) = \frac{1}{N}\sum_{n=1}^{N}\ln\frac{\exp[U_{j_n}(\theta,x_{n,j_n})]}{\sum_{j\in D_n}\exp[U_j(\theta,x_{nj})]} + \frac{1}{N}\sum_{n=1}^{N}\ln R_n(D,C,\theta)$$

And so the pseudo-log likelihood based on the reduced choice set is:

$$LL_N(\theta) = LL_N^C(\theta) - \frac{1}{N}\sum_{n=1}^{N}\ln R_n(D,C,\theta) = LL_N^C(\theta) + \frac{1}{N}\sum_{n=1}^{N}\ln\left[R_n(D,C,\theta)^{-1}\right] \quad \text{(A6)}$$

Thus, the basic intuition of McFadden's result is that the (1/N) $\sum \ln R_n(D,C,\theta)^{-1}$ term shifts the (expected) log-likelihood up, but it does not alter where it is maximized. Notice that:

$$\frac{1}{N}\sum_{n=1}^{N}\ln R_n(D,C,\theta)^{-1} \equiv \frac{1}{N}\sum_{n=1}^{N}\ln \sum_{j\in C}\exp[U_j(\theta,x_{nj})] \Big/ \sum_{j\in D_n}\exp[U_j(\theta,x_{nj})] \quad \text{(A7)}$$

As the set D_n contains the true choice (along with randomly selected other options from C), the expectation of (A7), i.e. the positive divergence between $LL_N(\theta)$ and $LL_N^C(\theta)$, is minimized at θ^*.

[38] For example, take $y = p^*\ln p + (1-p^*)\ln(1-p)$. Then the maximum is found by setting $dy/dp = p^*/p - (1-p^*)/(1-p) = 0$ which implies that $p=p^*$.

Unfortunately, McFadden (1978)'s method of proof does not go through when MNL is extended to include heterogeneity. To see why, consider a case with T types of consumers with type proportions $p^*(\tau)$ and parameters θ_τ^* for τ=1,…,T. In that case equation (A2) becomes:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k\in C}\sum_{D\subset C} P(k\,|\,\theta^*, p^*, C, x_n)\pi(D\,|\,k)\ln\left\{\sum_{\tau=1}^{T} p(\tau)\frac{\exp[U_k(\theta_\tau, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau, x_{nj})]}\right\} \tag{A8}$$

The term in brackets is the unconditional choice probability for person n, obtained by taking a weighted sum of the choice probabilities conditional on type τ, and weighting by the estimated type proportions, which we denote by $p(\tau)$. The term $P(k\,|\,\theta^*, p^*, C, x_n)$ is the true unconditional probability that option k is chosen. This depends on the vector of true type proportions, denoted by p^*, and the vector of true parameters for each type $\theta^* = (\theta_1^*,\ldots, \theta_T^*)$. Specifically, we have:

$$P(k\,|\,\theta^*, p^*, C, x_n) = \sum_{\tau=1}^{T} p^*(\tau)\exp[U_k(\theta_\tau^*, x_{nk})] \Big/ \sum_{j\in C}\exp[U_j(\theta_\tau^*, x_{nj})]$$

Now consider the key step of the previous proof where we multiply and divide this object by $\sum_{j\in D}\exp[U_j(\theta_\tau^*, x_{nj})]$, where now the θ^* are type specific. This gives:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k\in C}\sum_{D\subset C}\sum_{\tau=1}^{T} R_n(D,C,\theta_\tau^*)p^*(\tau)\frac{\exp[U_k(\theta_\tau^*, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau^*, x_{nj})]}\pi(D\,|\,k)\ln\left\{\sum_{\tau=1}^{T} p(\tau)\frac{\exp[U_k(\theta_\tau, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau, x_{nj})]}\right\}$$

where $R_n(D,C,\theta_\tau^*) \equiv \sum_{j\in D}\exp[U_j(\theta_\tau^*, x_{nj})] \Big/ \sum_{j\in C}\exp[U_j(\theta_\tau^*, x_{nj})]$ is the ratio of the probability of choosing $k\in D$ from the full choice set C to that of choosing it from the reduced choice set D. In general this is type τ specific. However, in the special case $R_n(D,C,\theta_\tau^*) = R_n(D,C)$ for all τ, we could use manipulations like those leading from (A3) to (A5) to obtain:

$$E[LL_N(\theta)] = \frac{1}{N}\sum_{n=1}^{N}\sum_{D\subset C} R_n(D,C)\pi(D)\cdot\sum_{k\in D}\left\{\sum_{\tau=1}^{T} p^*(\tau)\frac{\exp[U_k(\theta_\tau^*, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau^*, x_{nj})]}\ln\sum_{\tau=1}^{T} p(\tau)\frac{\exp[U_k(\theta_\tau, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau, x_{nj})]}\right\}$$

In that case, the third summation term would have the form $\sum_{k}\left\{P_k(\theta^*, p^*)\ln P_k(\theta, p)\right\}$, and by the same logic as above the maximum is achieved by setting $\theta=\theta^*$ and $p=p^*$.

But, in general, when $R_n(D,C,\theta_\tau^*)$ is type specific, we cannot bring this term outside the summation over τ, and so the form $\sum_k \left\{P_k(\theta^*, p^*)\ln P_k(\theta, p)\right\}$ cannot be achieved. Instead we obtain:

$$E\left[LL_N(\theta)\right] = \frac{1}{N}\sum_{n=1}^{N}\sum_{D\subset C}\pi(D)\sum_{k\in D}\left\{\sum_{\tau=1}^{T}R_n(D,C,\theta_\tau^*)\,p^*(\tau)\frac{\exp[U_k(\theta_\tau^*, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau^*, x_{nj})]}\ln\sum_{\tau=1}^{T}p(\tau)\frac{\exp[U_k(\theta_\tau, x_{nk})]}{\sum_{j\in D}\exp[U_j(\theta_\tau, x_{nj})]}\right\}$$

So, unlike in (A5), the necessary symmetry of the term in curly brackets is not achieved.

The problem is that subset *D* is not chosen completely at random, as it must include the chosen option *k*. Thus, $R_n(D,C,\theta_\tau^*)$ differs across types because $\exp[U_k(\theta_\tau^*, x_{nk})]$ differs across types. For example, suppose one type has a strong preference for option *k*, and chooses it with probability near one. For that type $R_n(D,C,\theta_s^*) \approx 1$. On the other hand, suppose a type chooses randomly, picking each option with probability 1/(#C). For them $R_n(D,C,\theta_s^*) = (\#D)/(\#C)$.[39]

However, note that as the size of the randomly selected subset *D* approaches that of the full choice set *C*, we have that $R_n(D,C,\theta_\tau^*) \rightarrow 1$ for all τ. This suggests that if the random subsets *D* are sufficiently large, then any bias induced by using random subsets of the full choice set *C* will tend to be negligible. In the next section we examine this issue using a Monte Carlo study. We find that finite sample bias is indeed negligible even for modest sized subsets.

To our knowledge only a few studies examine the impact of randomly reducing choice set size in MINL models with heterogeneity: Brownstone et al (2000), McConnell and Tseng (2000), Narella and Bhat (2004), Domanski and von Haefen (2011), von Haefen and Domanski (2013) and Guevara and Ben-Akiva (2013). However, it is difficult to apply these authors' results in our context because we consider different types of models and data. Only Domanski and von Haefen (2011) consider RP panel data, and they only consider the LC model.[40] Thus, we turn to our own Monte Carlo study of N-MIXL, G-MNL and MM-MNL in the panel data case.

A.2. Monte Carlo Experiments

Here, we report results of Monte Carlo experiments to assess the bias induced by using randomly selected subsets of the full choice set to estimate N-MIXL, G-MNL and MM-MNL

[39] If the sampling scheme could be designed so $R_n(D,C,\theta_\tau) = (\#D)/(\#C)$ for all types, then we would have $R_n(D,C,\theta_\tau) = R_n(D,C)$ and McFadden's proof goes through. Narella and Bhat (2004) also discuss this (see their equation (8)).

[40] In their paper the full choice set consists of 569 lakes in Wisconsin. They have an unbalanced panel of 513 respondents whose number of fishing trips range from 1 to 50. There are 15 attributes. They estimate the LC model via the EM algorithm using both the full choice set and randomly sampled subsets ranging from 50% to 1% of the full set. They find one can obtain reasonably reliable willingness to pay estimates using 5% subsets.

models. In each experiment, there are 200 hypothetical respondents and 20 choice occasions per respondent (4,000 observations). There are 60 alternatives in the full choice set, each with 4 attributes. The first and second attributes are dummy variables.[41] The third and fourth are drawn from standard normal distributions. The experiments differ in the specification of heterogeneity. We generate 20 artificial datasets for each case.

We estimate each model using both the full choice set (60 alternatives) and randomly sampled subsets with 20 or 10 alternatives. The chosen alternative is always included. Then, either 19 or 9 additional alternatives are randomly drawn from the remaining 59 alternatives. The random choice sets are drawn independently for each observation (person). Thus, people with the same observed choices will have random choice sets with different sampled alternatives. In the estimation, 500 draws are used to simulate the likelihood.

First, consider N-MIXL applied to artificial datasets of different design. Table A1 reports a case where only the first of the four attributes has a random coefficient, with a mean and standard deviation of one. The table reports the true parameter values, the mean estimates across the 20 Monte Carlo data sets, the empirical standard deviation of the estimates, and the mean of the asymptotic standard errors using both the BHHH (outer product) and robust (sandwich) formulas. An asterisk indicates bias in an estimated parameter is significant at the 5% level.

In Table A1 there is no evidence of significant bias using the full choice set of 60 or a random subset of 20 alternatives. If we use a random subset of only 10 alternatives there is significant bias only for β_3. But the bias is quantitatively small, as the true value is 1.0 and the mean estimate is 1.01. Also, the estimated and empirical standard errors align closely. Table A2 reports results for a case with (correlated) random coefficients on the first two attributes. The results are almost identical to those in Table A1.

Table A3 considers N-MIXL with a full variance-covariance matrix. Here, only a few covariance matrix parameters exhibit significant bias when the choice set is reduced from the full set of 60 to either 20 or 10. Again the magnitude of the bias is quantitatively small. Here, however, the BHHH standard errors are small relative to the empirical standard errors, regardless of whether the full or reduced choice set is used. Robust standard errors are more accurate.

We consider the G-MNL model with a full covariance matrix in Table A4. Here there is little evidence of bias, regardless of whether we use the full choice set or subsets of 20 or 10. In

[41] Each of these dummies is equal to 1 with probability 0.5.

each case, we see significant (but quantitatively small) bias for just one covariance parameter. Again however, BHHH standard errors using only 500 draws are too small, and robust standard errors are more accurate.

Finally, Table A5 reports results for the MM-MNL model. We consider a case where there are two consumer types, each with its own mean vector and covariance matrix for the parameter vector (the covariance matrices are assumed diagonal). There is no evidence of bias when using the full choice set. If we use a subset of 20 choices there is a significant but quantitatively small bias for the standard deviation of the β_2 parameter for type 1 (0.94 vs. 1.0). And if we use a subset of 10 choices there is a significant but small bias for the standard deviation of the β_2 parameter for type 2 (0.51 vs. 0.60). The BHHH standard errors are again too small. However, we re-calculated the BHHH standard errors in Tables A3-A5 using 5000 draws, and the results were much more accurate.

Thus, in the more complicated cases of Tables A3-A5, we find that 500 draws is too few to give reliable estimates of the standard errors. But using 5000 draws the empirical and asymptotic standard errors align well. Given these results, we decided to use 5000 draws to construct the standard errors reported in the main text. That is, we use 500 draws to simulate the likelihood in the estimation process, but then shift to using 5000 draws when we calculate the standard errors at the final estimates.

In summary, we find little evidence that use of randomly selected subsets of the full choice set induces bias in estimates of the N-MIXL, G-MNL or MM-MNL models. This is despite the fact that the proof of consistency of logit models estimated on reduced choice sets does not go through with heterogeneity (see Appendix A.1). Why does this occur? We conjecture that the consistency condition that $R_n(D,C,\theta_\tau^*) = R_n(D,C)$ for all types τ (see Section A.1), while not holding exactly, still holds to a good approximation. And as a result, the bias from using reduced choice sets is negligible.

The accuracy of the approximation is illustrated in Figure A1. It is based on the experiment in Table A1 that considered an N-MIXL model where only attribute #1 has a random coefficient. The dashed line is the log-likelihood contour as we vary $\hat{\beta}_1$ holding the other parameters at the true values. The solid line is the pseudo-log-likelihood contour if we use a random subset of 10 alternatives instead of the full choice set. (The log-likelihood value at each $\hat{\beta}_1$ is an average from 20 simulated datasets). The two likelihood contours are plotted on different scales (see the left and right axes) so they can fit in the same graph. The bottom panel of Figure A1 plots the difference between the log-likelihoods based on the reduced vs. full choice sets.

As we noted earlier, the basic intuition of McFadden's consistency result is that the use of a random subset of the full choice set shifts the (expected) log-likelihood up, but does not alter where it is maximized. Although this result does not hold exactly with heterogeneity, it holds to a very good approximation, as can be seen in Figure A1. The true and pseudo-likelihood contours have very similar shapes and are maximized at almost the same point (1.02 vs. 1.01). The difference between them (bottom panel) varies by only about ½ point over a very wide range around the optimum. Hence the use of a reduced choice set has little effect on the optimum.

An important point is that, even if using reduced choice sets does not induce significant bias in the estimates, it will lead to bias in information criteria used to compare models. Consider the Bayes Information Criterion, BIC = $-2LL + k \cdot \ln(\mathrm{N})$. The second term is the penalty for parameters (k), which depends on the number of observations N. If we (randomly) reduce the size of the choice set, the penalty term remains the same, but the (pseudo)-log-likelihood will improve. So the BIC difference between two models is only invariant to the size of the random choice set under the stringent condition that the log-likelihood difference between the two models is the same for all random choice set sizes. But in our simulations this is not the case.

To illustrate this point, we take the simulated data sets used to study the N-MIXL model in Table A2 and estimate simple MNL models on all of them. In Table A6, top panel, first two columns, we report the average log-likelihood for MNL (the misspecified model) and N-MIXL (the true model). The next column reports the difference between the log-likelihoods. Note that the difference is smaller for smaller random choice set sizes. As a result, the smaller the random choice set, the smaller the BIC gain from estimating the true model. We see the same pattern in the bottom panel, which compares MNL and MM-MNL (using data sets from Table A5). Thus, we see how BIC is biased toward smaller models when only a subset of alternatives is used in the estimation. We continue to use BIC, while keeping in mind that it is conservative in this sense.

Appendix A.3. Estimation Results for Revealed Preference (RP) Data

Based on our Monte Carlo results we decided to estimate the RP choice models for frozen pizza using 3 different choice set sizes, 20, 30 or 40, which correspond to roughly 25%, 37.5% or 50% of the full choice set. Table 4 in the main text reports results using a choice set size of 40, while Appendix Tables A7 and A8 report results for choice sets of 20 and 30.

A striking finding is that estimates are very stable across all three choice set sizes. This is true for all 6 models (MNL, S-MNL, N-MIXL, G-MNL, LC, and MM-MNL). This is not surprising for MNL, which is consistent for randomly chosen subsets of the full choice set. But it is surprising for models with heterogeneity.

For example, for N-MIXL, the price coefficients are -1.22 (standard error = 0.07), -1.18 (0.06) and -1.21 (0.06) using choice sets of 20, 30 and 40, respectively. So, variation of the estimates across the three choice set sizes is less than one standard error. Similarly, in G-MNL, the price coefficients are -1.23 (0.07), -1.16 (0.06) and -1.20 (0.06), respectively. The MM-MNL model identifies two types in each case. Price coefficients for type 1 are -1.62 (0.11), -1.66 (0.11) and -1.75 (0.07), while those for type 2 are -0.26 (0.14) -0.33 (0.11) and -0.48 (0.09). Results for LC and S-MNL are similar. In all cases, the price coefficient is stable across choice set sizes.

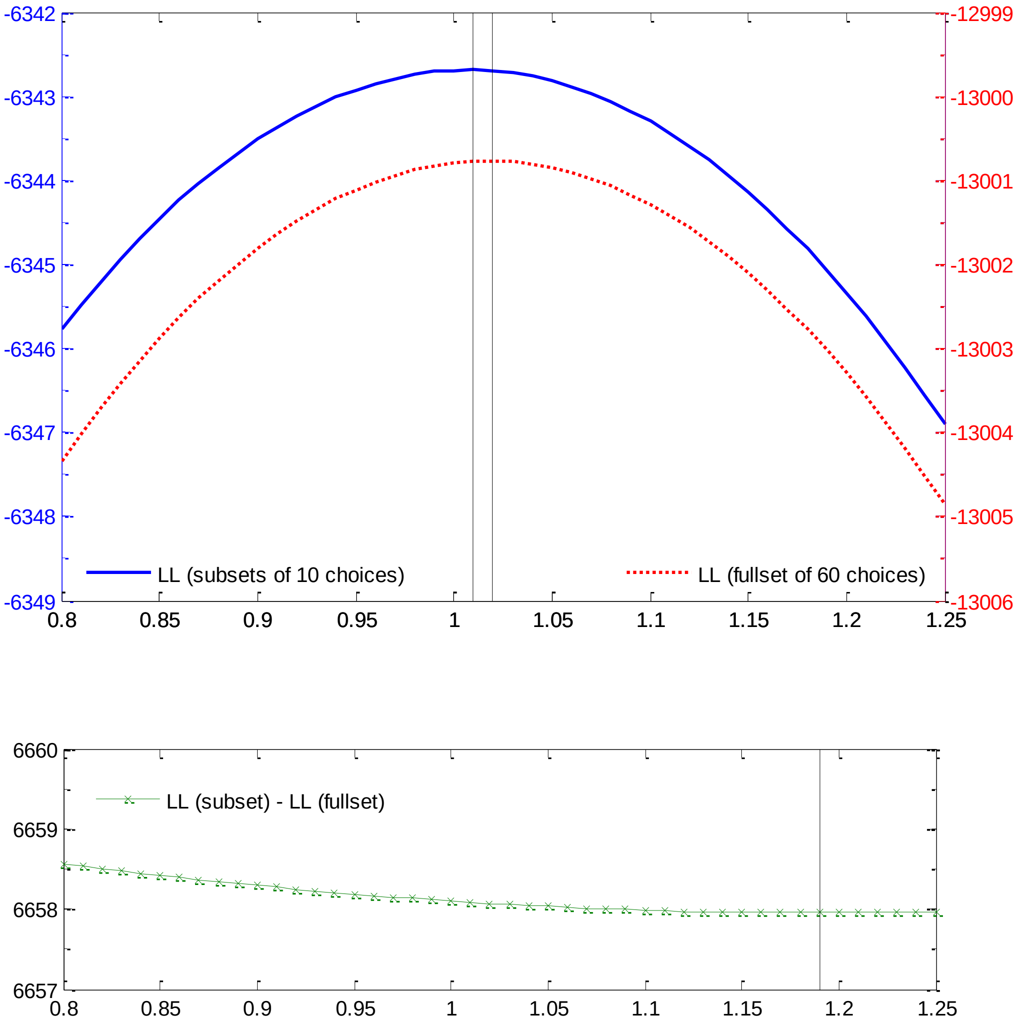
The reader can verify that the other parameters besides price are also quite stable across the three choice set sizes. This similarity adds to our confidence (already considerable in light of the earlier Monte Carlo results) that any bias induced by using random subsets of the full choice set is negligible, even for models that include heterogeneity.

A.4. Conclusion

We have found evidence that using randomly reduced sets does not induce significant bias in estimating logit choice models with heterogeneity on very large choice sets. We have presented both Monte Carlo evidence and results from actual data to support this claim. Also, based on McFadden’s (1978) original proof for the MNL case, we have given some intuition why his result is likely to hold approximately once heterogeneity is introduced.

Our finding that the use of random subsets of the full choice set is a reliable procedure is not only relevant for choice of supermarket goods. There are many contexts where large choice sets are a problem, such as choice of homes or residential location, choice of colleges and majors, choice of TV shows or movies, choice of occupation at a detailed level, etc.. So this finding may be useful in many contexts.

Figure A1: Simulated MIXL Likelihood Profiles



Note: In the top panel the red line is the log-likelihood profile for $\hat{\beta}_1$ based on the full choice set of J = 60. The blue line is the log-likelihood profile based on a random subset of 10 choices (including the chosen alternative). The scale for the blue line is on the left axis while that for the red line is on the right axis. The vertical lines indicate the maximum for each profile. The green line in the bottom panel is the (positive) difference between the two profiles.

Table A1: Monte Carlo Results for Mixed Logit with One Random Coefficient

		Complete choiceset (60 choices)				Random subset (20 choices)				Random subset (10 choices)			
		mean	empi.	mean ASE		mean	empi.	mean ASE		mean	empi.	mean ASE	
	TRUE	est	s.e.	BHHH	Robust	est	s.e.	BHHH	Robust	est	s.e.	BHHH	Robust
β_1	**1**	**1.02**	0.09	0.08	0.08	**1.01**	0.09	0.08	0.08	**1.01**	0.09	0.09	0.08
β_2	**1**	**1.00**	0.04	0.04	0.04	**1.00**	0.04	0.04	0.04	**1.01**	0.05	0.04	0.04
β_3	**1**	**1.01**	0.02	0.02	0.02	**1.01**	0.02	0.02	0.02	**1.01**	0.02*	0.03	0.03
β_4	**-1**	**-1.00**	0.03	0.02	0.02	**-1.01**	0.03	0.02	0.02	**-1.01**	0.03	0.03	0.03
$\sqrt{\sigma_{11}}$	**1**	**1.01**	0.10	0.07	0.07	**1.00**	0.10	0.08	0.08	**0.99**	0.09	0.08	0.08

Note: $\sqrt{\sigma_{11}}$ denotes standard deviation of β_1. A * indicates significant bias at the 5% level.

Table A2: Monte Carlo Results for Mixed Logit with Two Correlated Random Coefficients

		Complete choiceset (60 choices)				Random subset (20 choices)				Random subset (10 choices)			
		mean	empi.	mean ASE		mean	empi.	mean ASE		mean	empi.	mean ASE	
	TRUE	est	s.e.	BHHH	Robust	est	s.e.	BHHH	Robust	est	s.e.	BHHH	Robust
β_1	**1**	**1.02**	0.10	0.08	0.08	**1.01**	0.10	0.08	0.09	**1.00**	0.09	0.08	0.09
β_2	**1**	**1.02**	0.07	0.08	0.09	**1.02**	0.07	0.08	0.09	**1.01**	0.07	0.09	0.09
β_3	**1**	**1.00**	0.03	0.02	0.02	**1.01**	0.03	0.03	0.02	**1.02**	0.03*	0.03	0.03
β_4	**-1**	**-1.00**	0.03	0.02	0.02	**-1.01**	0.03	0.02	0.03	**-1.01**	0.03	0.03	0.03
σ_{11}	**1**	**1.00**	0.17	0.15	0.15	**0.99**	0.19	0.15	0.16	**0.99**	0.17	0.16	0.16
σ_{22}	**1**	**1.02**	0.13	0.16	0.15	**1.00**	0.12	0.16	0.15	**1.00**	0.12	0.16	0.16
σ_{21}	**0.60**	**0.60**	0.10	0.11	0.11	**0.59**	0.10	0.11	0.11	**0.58**	0.09	0.11	0.11

Note: σ_{ii} denotes variance of β_i and σ_{ij} denotes covariance of β_i and β_j. A * indicates significant bias at the 5% level.

Table A3: Monte Carlo Results for Mixed Logit with Four Correlated Random Coefficients

		Complete choiceset (60 choices)				**Random subset (20 choices)**				**Random subset (10 choices)**			
		mean	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**	
	TRUE	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**
β_1	**1**	**1.00**	0.12	0.07	0.10	**1.01**	0.11	0.08	0.09	**1.02**	0.12	0.08	0.11
β_2	**1**	**1.01**	0.10	0.07	0.11	**1.00**	0.12	0.08	0.10	**1.01**	0.11	0.08	0.10
β_3	**1**	**1.00**	0.11	0.06	0.10	**1.01**	0.12	0.07	0.09	**1.01**	0.10	0.07	0.09
β_4	**-1**	**-1.02**	0.12	0.06	0.10	**-1.01**	0.10	0.07	0.09	**-1.01**	0.11	0.07	0.10
σ_{11}	**1**	**1.00**	0.23	0.14	0.16	**1.01**	0.22	0.15	0.17	**0.98**	0.21	0.16	0.17
σ_{22}	**1**	**1.04**	0.18	0.15	0.17	**1.04**	0.16	0.15	0.17	**1.07**	0.14*	0.17	0.18
σ_{33}	**1**	**1.08**	0.20	0.11	0.17	**1.07**	0.18	0.12	0.15	**1.05**	0.17	0.13	0.14
σ_{44}	**1**	**1.07**	0.19	0.10	0.13	**1.09**	0.18*	0.11	0.13	**1.11**	0.18*	0.13	0.17
σ_{21}	**0.6**	**0.57**	0.13	0.09	0.12	**0.58**	0.12	0.10	0.13	**0.59**	0.13	0.11	0.13
σ_{32}	**0.6**	**0.61**	0.12	0.08	0.14	**0.62**	0.11	0.09	0.12	**0.63**	0.13	0.10	0.12
σ_{43}	**0**	**0.01**	0.07	0.07	0.10	**0.00**	0.10	0.08	0.09	**-0.01**	0.10	0.09	0.12
σ_{31}	**0**	**-0.02**	0.08	0.06	0.09	**-0.02**	0.10	0.07	0.10	**0.00**	0.08	0.08	0.12
σ_{42}	**0**	**-0.01**	0.07	0.08	0.12	**0.01**	0.09	0.08	0.11	**0.00**	0.10	0.09	0.13
σ_{41}	**0.6**	**0.63**	0.16	0.07	0.10	**0.66**	0.16	0.08	0.11	**0.65**	0.14	0.09	0.17

Note: σ_{ii} denotes variance of β_i and σ_{ij} denotes covariance of β_i and β_j. A * indicates significant bias at the 5% level.

Table A4: Monte Carlo Results for G-MNL

		Complete choiceset (60 choices)				**Random subset (20 choices)**				**Random subset (10 choices)**			
		mean	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**	
	TRUE	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**
β_1	**1**	**1.02**	0.14	0.08	0.10	**1.02**	0.12	0.08	0.11	**1.03**	0.10	0.09	0.11
β_2	**1**	**1.01**	0.10	0.07	0.10	**1.02**	0.09	0.08	0.12	**1.03**	0.09	0.09	0.11
β_3	**1**	**1.00**	0.10	0.06	0.11	**1.01**	0.10	0.07	0.11	**1.04**	0.08*	0.08	0.10
β_4	**-1**	**-1.00**	0.14	0.07	0.10	**-1.01**	0.11	0.07	0.10	**-1.02**	0.12	0.08	0.10
σ_{11}	**1**	**1.04**	0.16	0.15	0.20	**1.02**	0.20	0.16	0.20	**1.06**	0.23	0.19	0.20
σ_{22}	**1**	**1.03**	0.20	0.16	0.19	**0.96**	0.22	0.16	0.18	**0.99**	0.14	0.18	0.20
σ_{33}	**1**	**1.08**	0.17*	0.12	0.20	**1.02**	0.19	0.13	0.16	**1.02**	0.20	0.14	0.17
σ_{44}	**1**	**1.05**	0.19	0.11	0.16	**1.02**	0.16	0.12	0.15	**1.06**	0.15	0.14	0.17
σ_{21}	**0.6**	**0.64**	0.14	0.09	0.15	**0.59**	0.17	0.10	0.16	**0.57**	0.13	0.11	0.15
σ_{32}	**0.6**	**0.64**	0.16	0.09	0.14	**0.55**	0.22	0.09	0.13	**0.57**	0.17	0.11	0.14
σ_{43}	**0**	**0.04**	0.10	0.08	0.13	**0.05**	0.08*	0.08	0.12	**0.02**	0.07	0.09	0.12
σ_{31}	**0**	**0.07**	0.11*	0.07	0.13	**0.02**	0.13	0.07	0.12	**-0.03**	0.13	0.08	0.12
σ_{42}	**0**	**0.02**	0.16	0.08	0.12	**0.03**	0.11	0.08	0.12	**0.04**	0.10	0.10	0.11
σ_{41}	**0.6**	**0.61**	0.18	0.07	0.13	**0.64**	0.16	0.09	0.13	**0.66**	0.12*	0.10	0.13
τ	**0.5**	**0.48**	0.14	0.05	0.10	**0.51**	0.12	0.06	0.09	**0.52**	0.10	0.07	0.09
γ	**0.5**	**0.47**	0.35	0.12	0.18	**0.60**	0.29	0.12	0.16	**0.58**	0.25	0.13	0.14

Note: σ_{ii} denotes variance of β_i and σ_{ij} denotes covariance of β_i and β_j. A * indicates significant bias at the 5% level.

Table A5: Monte Carlo Results for MM-MNL

		Complete choiceset (60 choices)				**Random subset (20 choices)**				**Random subset (10 choices)**			
		mean	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**		**mean**	**empi.**	**mean ASE**	
	TRUE	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**	**est**	**s.e.**	**BHHH**	**Robust**
class 1													
β_{11}	**1**	**0.98**	0.13	0.10	0.13	**0.98**	0.14	0.11	0.14	**0.98**	0.16	0.12	0.15
β_{21}	**1**	**1.00**	0.13	0.10	0.14	**1.02**	0.14	0.11	0.17	**1.02**	0.16	0.11	0.16
β_{31}	**1**	**1.01**	0.15	0.07	0.12	**1.02**	0.14	0.08	0.11	**1.04**	0.15	0.09	0.12
β_{41}	**-1**	**-0.97**	0.14	0.07	0.13	**-0.97**	0.16	0.08	0.12	**-0.97**	0.15	0.09	0.13
$\sqrt{\sigma_{11,1}}$	**1**	**1.01**	0.16	0.10	0.12	**1.02**	0.19	0.11	0.12	**1.04**	0.15	0.12	0.12
$\sqrt{\sigma_{22,1}}$	**1**	**0.94**	0.15	0.10	0.12	**0.94**	0.13*	0.11	0.16	**0.96**	0.16	0.11	0.15
$\sqrt{\sigma_{33,1}}$	**1**	**0.99**	0.11	0.06	0.09	**1.00**	0.09	0.07	0.09	**1.01**	0.10	0.08	0.11
$\sqrt{\sigma_{44,1}}$	**1**	**0.99**	0.10	0.06	0.10	**0.98**	0.11	0.06	0.10	**1.03**	0.09	0.07	0.10
class 2													
β_{12}	**-1**	**-0.95**	0.15	0.12	0.15	**-0.92**	0.19	0.13	0.17	**-0.95**	0.17	0.14	0.18
β_{22}	**-1**	**-1.00**	0.13	0.10	0.12	**-1.02**	0.15	0.11	0.13	**-1.02**	0.15	0.13	0.14
β_{32}	**1**	**0.98**	0.13	0.11	0.14	**0.99**	0.12	0.12	0.14	**0.99**	0.16	0.13	0.15
β_{42}	**-1**	**-1.01**	0.18	0.11	0.13	**-0.99**	0.17	0.12	0.14	**-1.02**	0.19	0.13	0.15
$\sqrt{\sigma_{11,2}}$	**0.6**	**0.65**	0.28	0.15	0.17	**0.65**	0.28	0.15	0.19	**0.62**	0.34	0.18	0.24
$\sqrt{\sigma_{22,2}}$	**0.6**	**0.49**	0.26	0.13	0.13	**0.47**	0.29	0.14	0.14	**0.51**	0.17*	0.18	0.19
$\sqrt{\sigma_{33,2}}$	**1**	**1.02**	0.15	0.10	0.11	**1.01**	0.15	0.11	0.13	**1.02**	0.15	0.12	0.14
$\sqrt{\sigma_{44,2}}$	**1**	**1.04**	0.15	0.08	0.11	**1.05**	0.13	0.09	0.12	**1.03**	0.16	0.11	0.12
class prob.	**0.64**	**0.62**	0.04	0.05	0.05	**0.62**	0.05	0.05	0.05	**0.63**	0.05	0.06	0.06

Note: $\sqrt{\sigma_{ii,s}}$ denotes standard deviation of β_{is}. A * indicates significant bias at the 5% level.

Table A6: The Bias in Information Criteria when Using a Subset of Alternatives

N-MIXL	Average LL		$LL_{N\text{-}MIXL}$-LL_{MNL}	Average BIC gain
	MNL	N-MIXL		
Choiceset size 60	-13336	-12953	383	741
Choiceset size 20	-9113	-8778	335	646
Choiceset size 10	-6602	-6323	279	533

MM-MNL	Average LL		$LL_{MM\text{-}MNL}$-LL_{MNL}	Average BIC gain
	MNL	MM-MNL		
Choiceset size 60	-14814	-12741	2073	4037
Choiceset size 20	-10502	-8633	1868	3629
Choiceset size 10	-7846	-6242	1604	3100

Note: N-MIXL and MM-MNL use datasets from Table 3 and Table 6, respectively.

Table A7: Estimates from RP Data (Based on a Random Subset of 20 Choices)

	MNL		S-MNL[a]		N-MIXL[b]		G-MNL[b]		Latent class[c]						MM-MNL[d]			
									class 1		class 2		class 3		class 1		class 2	
	est	s.e.	Est	s.e.	est	s.e.	est	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Brand [omitted others]																		
Tombstone	**0.83**	0.05	**0.65**	0.24	**0.48**	0.12	**0.57**	0.12	**2.19**	0.39	0.23	0.53	0.30	0.29	**1.44**	0.18	-0.35	0.50
Roma	**0.24**	0.05	-0.26	0.16	**-0.33**	0.12	-0.20	0.11	-0.18	0.32	0.19	0.27	**-1.48**	0.47	0.10	0.19	**-0.86**	0.42
Jacks	**0.63**	0.05	0.23	0.25	0.22	0.12	**0.27**	0.11	**2.53**	0.25	**0.83**	0.34	**-1.15**	0.51	**1.42**	0.16	**-2.07**	0.46
Red Baron	0.11	0.06	-0.08	0.20	-0.13	0.11	-0.16	0.11	0.32	0.36	0.58	0.31	-0.04	0.24	**0.45**	0.15	**-0.82**	0.35
Bernatello	**-1.03**	0.06	**-0.99**	0.16	**-1.01**	0.09	**-1.07**	0.12	-0.47	0.25	-0.72	0.40	-1.67	0.94	**-0.44**	0.10	**-2.01**	0.39
Price	**-0.82**	0.04	**-1.02**	0.09	**-1.22**	0.07	**-1.23**	0.07	**-1.89**	0.19	**-1.66**	0.23	**-1.04**	0.14	**-1.62**	0.11	-0.26	0.14
Promotion	**0.81**	0.10	**0.85**	0.16	**0.91**	0.20	**0.88**	0.19	**1.15**	0.26	**1.19**	0.16	0.88	0.45	**1.08**	0.29	0.78	0.52
Toppings [omitted combo]																		
Cheese only	**0.15**	0.05	-0.04	0.05	-0.29	0.16	**-0.41**	0.17	-0.09	0.48	0.47	0.46	0.15	0.64	-0.24	0.22	**-1.31**	0.52
Sausage/pepperoni	**1.02**	0.04	**0.89**	0.08	**0.99**	0.09	**1.00**	0.07	**1.13**	0.19	**0.82**	0.40	0.30	0.47	**0.71**	0.13	**1.37**	0.30
Meat/supreme	-0.06	0.05	-0.07	0.05	-0.12	0.09	-0.12	0.09	0.12	0.24	**-0.48**	0.18	0.29	0.37	-0.22	0.15	-0.05	0.23
Bacon/Burger	**-0.34**	0.06	**-0.46**	0.07	**-0.67**	0.11	**-0.75**	0.15	-0.21	0.23	-0.96	0.52	-0.35	0.70	**-0.40**	0.15	-0.44	0.41
Chicken/Mexican	**-0.51**	0.08	**-0.79**	0.12	**-0.78**	0.11	**-0.85**	0.15	**-0.91**	0.31	**-1.55**	0.62	0.01	0.64	**-0.74**	0.18	-0.26	0.30
Vegetarian	**-0.99**	0.15	**-1.71**	0.22	**-1.40**	0.23	**-2.07**	0.35	**-2.65**	1.08	**-2.66**	0.96	0.13	0.87	**-2.55**	0.44	-0.43	0.63
Crust [omitted regular, others]																		
Rising	**-0.84**	0.04	**-1.04**	0.09	**-1.20**	0.12	**-1.13**	0.11	**-1.50**	0.37	**-0.90**	0.32	-0.29	0.36	**-1.88**	0.18	**-0.81**	0.25
Thin/crispy	0.03	0.03	0.01	0.03	-0.02	0.06	-0.03	0.05	**-0.41**	0.14	0.09	0.30	**-0.57**	0.24	0.06	0.09	-0.22	0.17
Microwavable	0.03	0.06	**0.14**	0.05	-0.06	0.08	-0.06	0.10	0.26	0.61	**0.84**	0.32	-1.06	1.07	0.29	0.19	-0.73	0.39
τ			**0.92**	0.07			**0.37**	0.04										
γ							**0.80**	0.10										
class prob.									**0.30**	0.06	**0.25**	0.08	**0.19**	0.06	0.73	0.08	**0.27**	0.08
	16		27		48		50		84						66			
LL	-10814		-9285		-8538		-8524		-9417						**-8343**			
BIC	21761		18795		17476		17464		19533						**17234**			

Note: [a] estimates from S-MNL with random correlated (one-factor) intercepts; [b] estimates from correlated coefficients (imposing 1-factor structure on the covariance matrix); [c] estimates from LC with 5 classes; [d] estimates from MM-MNL with 2 proportional covariance normal. Bold estimates are statistically significant at 5%. S-MNL, N-MIXL, G-MNL and MM-MNL are estimated by simulated maximum likelihood with 500 draws. The standard errors are calculated using 5000 draws.

Table A8: Estimates from RP Data (Based on a Random Subset of 30 Choices)

	MNL		**S-MNL**[a]		**N-MIXL**[b]		**G-MNL**[b]		**Latent class**[c]						**MM-MNL**[d]			
									class 1		class 2		class 3		class 1		class 2	
	est	**s.e.**	**est**	**s.e.**	**est**	**s.e.**	**est**	**s.e.**	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Brand [omitted others]																		
Tombstone	**0.86**	0.05	**0.71**	0.21	**0.40**	0.11	**0.48**	0.11	0.87	1.32	**2.64**	0.47	0.08	0.36	**1.63**	0.16	-0.34	0.31
Roma	**0.23**	0.05	-0.21	0.17	**-0.30**	0.10	**-0.29**	0.10	-0.25	0.28	-0.61	0.64	**1.51**	0.34	-0.05	0.19	**-0.74**	0.21
Jacks	**0.64**	0.05	0.26	0.22	**0.20**	0.10	0.06	0.12	0.43	0.40	**2.67**	0.87	-0.03	0.44	**1.47**	0.14	**-2.12**	0.37
Red Baron	**0.12**	0.06	-0.07	0.21	**-0.25**	0.11	**-0.34**	0.12	0.48	0.48	0.61	0.72	0.20	0.33	**0.47**	0.14	**-0.83**	0.30
Bernatello	**-1.03**	0.06	**-0.95**	0.15	**-0.99**	0.09	**-1.07**	0.10	**-0.92**	0.29	-0.51	0.51	**-0.98**	0.41	**-0.42**	0.11	**-1.81**	0.34
Price	**-0.83**	0.04	**-1.01**	0.07	**-1.18**	0.06	**-1.16**	0.06	**-1.62**	0.44	**-1.85**	0.30	**-0.80**	0.18	**-1.66**	0.11	**-0.33**	0.11
Promotion	**0.83**	0.10	**0.84**	0.15	**0.88**	0.18	**0.79**	0.18	**1.16**	0.24	**0.98**	0.34	**0.94**	0.18	**1.05**	0.27	0.63	0.50
Toppings [omitted combo]																		
Cheese only	**0.16**	0.05	-0.03	0.05	**-0.32**	0.15	**-0.31**	0.15	-0.54	0.41	-0.03	0.34	0.28	0.42	-0.32	0.19	**-1.09**	0.35
Sausage/pepperoni	**1.02**	0.04	**0.88**	0.07	**0.99**	0.07	**1.03**	0.07	**0.58**	0.17	**1.21**	0.31	**1.24**	0.25	**0.73**	0.11	**1.45**	0.26
Meat/supreme	-0.06	0.05	-0.07	0.05	-0.002	0.08	0.02	0.08	0.07	0.78	0.10	0.91	-0.09	0.21	-0.20	0.13	-0.13	0.22
Bacon/burger	**-0.36**	0.06	**-0.47**	0.06	**-0.60**	0.10	**-0.72**	0.14	-0.31	0.23	-0.46	0.30	-0.13	0.31	**-0.40**	0.16	-0.49	0.31
Chicken/Mexican	**-0.52**	0.08	**-0.79**	0.10	**-0.61**	0.11	**-0.80**	0.17	-0.29	0.25	**-1.19**	0.46	-0.01	0.91	**-0.67**	0.17	-0.20	0.33
Vegetarian	**-1.04**	0.15	**-1.70**	0.18	**-1.56**	0.26	**-2.13**	0.35	-1.43	2.09	-2.93	1.67	-0.66	0.80	**-2.73**	0.42	-0.34	0.42
Crust [omitted regular, others]																		
Rising	**-0.85**	0.04	**-1.03**	0.07	**-1.13**	0.12	**-1.03**	0.10	**-1.10**	0.37	**-1.62**	0.33	**-1.17**	0.34	**-1.83**	0.14	**-0.66**	0.21
Thin/crispy	0.03	0.03	0.02	0.02	-0.03	0.06	-0.06	0.05	-0.14	0.63	-0.67	0.28	**0.98**	0.24	0.02	0.09	-0.12	0.15
Microwavable	0.05	0.06	**0.17**	0.05	-0.08	0.09	-0.07	0.10	0.12	0.61	0.52	0.82	0.02	0.32	**0.46**	0.18	**-0.70**	0.30
τ			**0.88**	0.06			**0.43**	0.04										
γ							**0.59**	0.07										
class prob.									**0.33**	0.15	0.26	0.17	**0.20**	0.05	**0.72**	0.08	**0.28**	0.08
	16		27		48		50		84						66			
LL	-12441		-10820		-9969		-9944		-10933						-9751			
BIC	25015		21865		20338		20304		22565						**20052**			

Note: [a] estimates from S-MNL with random correlated (one-factor) intercepts; [b] estimates from correlated coefficients (imposing 1-factor structure on the covariance matrix); [c] estimates from LC with 5 classes; [d] estimates from MM-MNL with 2 proportional covariance normal. Bold estimates are statistically significant at 5%. S-MNL, N-MIXL, G-MNL and MM-MNL are estimated by simulated maximum likelihood with 500 draws. The standard errors are calculated using 5000 draws.