The Structure of Consumer Taste Heterogeneity in Revealed vs. Stated Preference Data

by

Michael P. Keane
University of Oxford, Nuffield College

Nada Wasi
University of Michigan, Survey Research Center

February 4, 2013

Abstract: In recent years it has become common to use stated preference (SP) discrete choice experiments (DCEs) to study and/or predict consumer demand. SP is particularly useful when revealed preference (RP) data is unobtainable or uninformative (e.g., to predict demand for a new product with an attribute not present in existing products, to value non-traded goods). The increasing use of SP data has led to a growing body of research that compares SP vs. RP demand predictions (in contexts when both are available). The present paper goes further by comparing the structure of consumer taste heterogeneity in SP vs. RP data. Our results suggest the nature of taste heterogeneity is very different: In SP data consumers are much more likely to exhibit either (i) lexicographic preferences, or (ii) "random" choice behavior. And many consumers appear to be fairly insensitive to price. This suggests that caution should be applied before using SP to answer questions about the distribution of taste heterogeneity in actual markets.

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

Keywords: Discrete choice experiments, Stated preference data, Discrete choice models, Consumer demand, Consumer heterogeneity, Mixture models

Acknowledgments: Keane's work on this project was supported by ARC grant FF0561843.

I. Introduction

In recent years it has become increasingly common to use stated preference (SP) discrete choice experiments (DCEs) 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,[1] or to value non-traded (public) goods.

Given this trend, Arrow et al (1993) argue for the importance of validating SP by comparing demand predictions from SP vs. RP data (in contexts when both are available). There is a growing body of research in this area. For quasi-public goods, Carson et al. (1996) conducted a meta-analysis of 83 studies that contain estimates from both stated-preference methodologies and RP data. They found that RP and SP results were generally similar. For private goods, however, the results are mixed (see Carson et al, 1994). However, Louviere and Kocur (1983) and Horowitz and Louviere (1990) found that aggregate market share predictions from RP and SP data were similar for public transport options and college choices, respectively.

In cases where both SP and RP data are available, several authors have advocated pooling to estimate "richer" choice models (see, e.g., Swait and Andrews (2003), Morikawa, Ben-Akiva and McFadden (2002)). If one cannot reject the hypothesis that parameters are equal across the RP and SP data generating processes, pooling increases efficiency. Also, if the SP data contain variation along dimensions that are fixed in the RP data, additional parameters can be identified. Early papers exploiting these ideas were Ben-Akiva and Morikawa (1990), Hensher and Bradley (1993) and Swait and Louviere (1993).[2] Some more recent work that combines RP and SP data includes Brownstone et al (2000), who estimate demand for alternative-fuel vehicles, Swait and Andrews (2003), who estimate demand for laundry detergent, Small, Winston and Yan (2005, 2006), who model commuter demand for express vs. regular traffic lanes, and Earnhart (2002) who studies housing choice. As Fiebig et al (2010) note, many studies cannot reject that attribute weights in logit models estimated on RP and SP data are equal, *up to a common scale parameter*. This implies that mean preference parameters are often comparable across RP and SP data.

[1] For example, Goett (1998) studies consumer choice of electricity suppliers. SP data allows him to consider pricing plans that did not previously exist. Brooks and Lusk (2010) study demand for milk from cloned cows. Other well-known applications are Hensher (1994, 2001) and Hensher and Greene (2003), who use SP to model travel demand, and Revelt and Train (1998), who study demand for energy efficient appliances. A few more recent applications include Hall et al (2006) on demand for medical testing, Hjelmgren and Anell (2007) on demand for modes of primary health care in Sweden, and Berninger et al. (2010) on demand for a sustainable forest management program.

[2] See Carson et al (1994) or Keane (1997a) for more detailed discussion of early studies combining SP and RP data.

It is notable that prior work comparing results from RP vs. SP choice data has generally compared aggregate demand predictions (e.g., predicted market shares), or tested equality (up to scale) of mean utility weight vectors (estimated on RP vs. SP data). Yet, in many applications one is interested in knowing the whole distribution of consumer preference weights. The reasons are obvious: e.g., optimal pricing, new product design and consumer welfare calculations depend on the distribution of preferences, not just the mean preference vector.

To fill this gap, the present paper compares the structure of consumer taste heterogeneity in SP vs. RP data. We do this in two steps: First, we compare the performance of several of the most popular models of consumer taste heterogeneity on both RP and SP data. We determine which of these several models provides the best fit to the structure of taste heterogeneity in each type of data. The set of models we consider is described in detail in Section II. Second, using the best fitting models, we compare the structure of heterogeneity as inferred from RP vs. SP data. To our knowledge, there is no prior work that attempts to do this directly.

It is important to understand why we estimate many parametric models: Given the high dimensional nature of consumer choice data (i.e., many choices/attributes) flexible parametric models are the only feasible way to draw inferences about heterogeneity distributions. Thus, if we wish to compare heterogeneity distributions in RP vs. SP data, finding and comparing the best fitting parametric choice models for each type of data is the only feasible approach.

To compare patterns of heterogeneity in RP vs. SP data, we need SP and RP datasets that are similar in three ways. First, they should involve a similar product or service. Second, the set of attributes observed in each dataset should be fairly comparable, and the choice task should be of similar complexity.[3] Third, we need multiple observations on each consumer in both the SP and RP data. This lets us form accurate posterior distributions of individual level parameters (which we use to determine the shape of preference heterogeneity distributions in each dataset).

After assessing many available datasets, we feel the best opportunity to compare RP and SP data involves data on pizza choices. On the RP side, Information Resources Inc. (IRI), a large market research firm, began collecting household level data on supermarket purchases of frozen pizza in 2001. We use the data from 2001-3 for estimation of RP choice models. On the SP side, Louviere et al (2008) collected SP data on preferences for pizza delivery services. The data was

[3] For example, say the SP task is a simple choice between two products with only a few attributes. It is implausible that consumers' processing of such a task is comparable to that of choosing among the highly differentiated products available in a grocery store. Clearly, the choice objects in the SP task should have several attributes as well.

collected through PureProfile, a major survey research firm in Australia.

These data satisfy the three requirements listed above: (i) the products are comparable, (ii) the options are described by several attributes in each dataset, so the choice task is of comparable complexity, and (iii) in each dataset there are many choice occasions per respondent.

One may ask if preferences inferred from choices of pizza delivery service vs. choices among types of pizza are comparable. We argue they are for two reasons: First, we adopt an attributes based view of utility as in Lancaster (1966). That is, goods are bundles of attributes, and utility is defined over attributes, rather than goods *per se*. Fortunately, in the SP data, delivery services are described as offering different types and qualities of pizza, at different prices. For example, one service might offer vegetarian or thick crust pizza while another might not. Thus, subjects can avail themselves of different pizza attributes (at different prices), by choosing different delivery services. This enables us to identify subjects' willingness to pay for pizza attributes from delivery service choices, just as we can identify them from pizza choices.

If goods have attributes in common, we can infer willingness to pay (WTP) for those attributes by studying demand for either good. The same logic is used in public or environmental economics to identify WTP for school quality or environmental amenities from home location choices. In our case the goods are far more similar. That is, in both the RP and SP data, the final good that generates utility is pizza. There is no direct demand for pizza *delivery*, only the derived demand for the final product. It is reasonable to assume that the incremental utilities consumers receive from pizza attributes are the same *regardless* of whether it was delivered or bought in a store. (Put it simply, why would WTP for sausage depend on whether a pizza is delivered?).

Second, even absent Lancaster type assumptions, the marginal utility of consumption of the outside good should not vary across different inside goods (of similar cost) under standard economic assumptions. And the marginal utility of the outside good is given by the negative of the price coefficient (up to scale). Thus, the distribution of the price coefficient is comparable even for two different inside goods. We discuss this point in more detail in Section II.

Nevertheless, we agree it would be ideal if our RP and SP data were identical, in terms of the choice set, attributes and population under study. This would make the estimates comparable under weaker economic assumptions. But it is important to note that RP and SP data will, by construction, never be identical. The two types of data are collected for different purposes, and are constructed in very different ways. We expand on this point in the data section (Section III**)**.

A closely related point is that opportunities to compare RP and SP data are rare. Many household RP datasets are public. But SP data is usually collected for proprietary research.[4] This limits the availability of SP data for comparison purposes. Thus, while our RP and SP data are not identical, we believe they are as similar as one is likely to find. Hence, they represent a good opportunity to compare the structure of heterogeneity in RP vs. SP data.

Our results suggest that estimates of the distribution of tastes can diverge substantially between RP and SP data. In particular, individual consumers are more likely to exhibit "extreme" or lexicographic preferences in SP data (i.e., they base choices almost entirely on one or two attributes). Conversely, some subjects in SP experiments exhibit behavior that is close to "random," in the sense that product attributes (including prices) have only small effects on their choices. These polar behavioral patterns are much less common in RP data. We will show that this difference in estimates cannot be explained by scale differences between the RP and SP data. Thus, while SP data may be informative about aggregate demand, considerable caution should be applied if using it to infer detailed aspects of the preference heterogeneity distribution.

A finding that the heterogeneity distributions in SP data differ in important ways from those found in RP data may raise concern about the whole enterprise of using SP to elicit heterogeneity (or willingness to pay) distributions. At minimum, it might suggest that improved experimental designs are needed, so SP provides a better representation of the market choice task. Of course, advocates of SP would argue the reverse – i.e., that experimental control of attributes makes SP results more reliable by avoiding collinearity and endogeneity. But regardless of whether a researcher has more faith in RP or SP data, it is important to know which of the many available models of heterogeneity provides the best fit to each type of data.

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 RP and SP data sets that we use. In Section IV we present our main empirical results, and evaluate which models of heterogeneity provide the best fit in RP vs. SP data. Section V assesses how patterns of consumer behavior differ in the RP vs. SP data. This enables us to determine <u>why</u> different models fit better in different cases (in terms of which behavioral patterns are most prevalent in each dataset, and which model(s) can best capture those patterns). Section VI concludes.

[4] We have access to proprietary SP data on medical tests, cell phones, holiday packages, charge cards, etc., along with the pizza delivery data we use here. But none of the other products had comparable RP datasets available.

II. Alternative Models of Heterogeneity in Consumer Choice Behavior

We first discuss some general features of discrete choice demand models for frequently purchased consumer goods, and then turn to specific models. 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 also sensible) to assume the marginal utility of the outside good is roughly constant over the range of consumption levels C generated by different levels of spending on the inside good. This implies the indirect utility function (conditional on purchase of the inside good) is linear in C over that range. Let $u(j|p_j, I)$ be the indirect utility conditional on purchase of inside good j. Then we have $u(j|p_j, I) = D(j) + (u_c)(I - p_j)$, where u_c is the marginal utility of consumption of the outside good, which is assumed constant, and $D(j)$ is the direct utility from inside good j.

It is worth emphasizing that the price coefficient u_c is invariant to the option j that is chosen. There have indeed been studies that 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 necessary to identify error correlations in a discrete choice model, and the restriction $u_c = \alpha \; \forall \; j$, where α is a constant, is usually the most 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 goods, such as mode of transport (e.g., bus, train, car).[5]

It is standard to assume the direct utility from the inside good $D(j)$ depends on both observed attributes and a stochastic component unobserved by the econometrician. For instance, we may have $D(j) = \Gamma A_j + \varepsilon_j/\sigma$, 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. Such a specification is consistent with the attributes based approach of Lancaster (1966) and the random utility model (RUM) of McFadden (1974).

[5] 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.

Note that choice is deterministic for consumers in the RUM, but random for an observer.

Using 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 unobserved, 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 $(1/\sigma)$.

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:

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

Note that the attribute weights Γ and the price coefficient $(-u_c)$ are only identified up to the scale factor σ.[6] So the price coefficient will differ across goods, even with u_c constant, due to scale differences. Similarly, even under a Lancaster-type assumption that consumers get utility from attributes rather than a good itself, common attribute coefficients will differ across goods due to differences in scale.[7] But measures of willingness to pay for attributes, given by ratios of attribute coefficients to the price coefficient $-\Gamma/u_c$, are scale invariant. This makes it possible to compare WTP measures for the same attribute obtained by studying demand for different goods.[8]

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. In RP data we often see multiple purchase occasions per person, while in SP 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.

[6] 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.
[7] For example, suppose utility weights on an attribute like sausage topping are the same whether a pizza is bought in a store or delivered. But suppose scale differs between the two goods, because delivery adds an extra unobserved service attribute. In this case common attribute coefficients will be equal up to scale between the two goods.
[8] Of course, if $D(j)$ is not linear in attributes then WTP may be harder to calculate, but invariance still holds.

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 that 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:

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

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 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.[9]

It is worth commenting on interpretation of ASCs. If alternatives correspond to 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:

[9] 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.

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

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

It is useful to define $\Gamma_n^* = \{\Gamma_n, -u_{cn}\}$ so that $\beta_n = \sigma_n \Gamma_n^*$. 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, in the attribute-based approach, 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.[10] 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 three popular models that can be derived from (4):

1) First is the logit with normal mixing (**N-MIXL**), where β_n is distributed multivariate 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 $\beta_n = \{\sigma_n \Gamma, -\sigma_n u_c\}$, 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 SP data takes the form of scale heterogeneity. A key point of the present paper is to investigate whether scale heterogeneity is also important in RP data.

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

[10] This may arise from differences in income/wealth, or unobserved heterogeneity in preferences. Income and wealth information is scarce in both RP and SP demand data, so the unobserved heterogeneity approach is usually adopted.

3A) Third is the "generalized multinomial logit" (**G-MNL**) model which was 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.

It is useful to note how G-MNL restricts the general structure in equation (3). Say we define $\beta_n^* = \{\Gamma_n, -u_{cn}\}$ so $\beta_n = \sigma_n\beta_n^*$, 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 G-MNL, 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 have that β_n is the product of a (positive) scalar random variable and a vector random variable.

We next consider models that are <u>not</u> 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.[11] 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 three popular models that can be derived from (3):

3B) Fiebig et al (2010) considered an alternative version of the G-MNL model in which $\beta_n = [\sigma_n\bar{\beta} + \eta_n]$. This differs from model #3 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.

Feibig et al (2010) also noted that the two versions of G-MNL can be nested in a single model 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.

[11] In particular, the structure in (4) implies that variation in σ_n must affect the whole $\beta_n^* = \{\Gamma_n, -u_{cn}\}$ 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. However, as these restrictions on the discrete β_n vectors are not imposed in estimation, it is unlikely they will hold in practice.

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 heterogeneity 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.

Keane and Wasi (2012a) compare the performance of these five models on data from ten SP choice experiments. The data cover a wide range of products such as medical tests, cell phones, pizza delivery services, holiday packages and charge cards. They find that MM-MNL and/or G-MNL fit better than N-MIXL in 8 of 10 datasets. The reason for the strong performance of these models is they capture two common behaviors of subjects in SP data: (i) lexicographic or "extreme" behavior and (ii) nearly "random" behavior. By this we mean: (i) some subjects base choices on just on one or two attributes, and (ii) some subjects are little influenced by product attributes. The more restrictive N-MIXL model has difficulty generating these patterns.

Of course, there is no *a priori* reason to think such behaviors are more common in SP or RP data. When choosing in a "real" market, we might expect consumers to weigh attributes more

carefully, and follow something closer to a compensatory utility rule. On the other hand, real market choice contexts tend to be more complex than SP experiments (e.g., many more options are available). This may cause people to use of simplifying heuristics (like lexicographic rules), or even lead to fairly random choice behavior for inexpensive goods. Our work sheds light on these issues by examining the differences in choice behavior between the RP and SP contexts.

III. The SP and RP Data Sets

As we discussed in Section I, our best opportunity to compare RP and SP choice behavior is with data on pizza choices. On the RP side, we use data on household purchases of frozen pizza collected by IRI at a large store in Eau Claire, Wisconsin, from Jan. 2001-Dec. 2003. On the SP side, we use two pizza delivery service choice experiments previously studied in Louviere et al (2008). In both experiments, respondents are asked to choose between two generic Pizza delivery services (labeled 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." As we noted earlier, delivery services are described as offering different types of pizza (e.g., thick crust, vegetarian) at different prices. Thus, choice of service provides a way to choose different pizza attributes.

In the RP data, transactions are observed at the Universal Product Code (UPC) level. 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 "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 size, typical of RP data, creates computational problems, which explains why the literature on UPC-level choice modeling is small. Fader and Hardie (1996) and Andrews and Manrai (1999) are classic works in this area.

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. There are 129 panelists who meet our criteria. And the total number of shopping trips where frozen pizza is bought is 4,123.[12]

[12] 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 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.

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. 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).[13]

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 small "other" brands make up 17%. 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 the various 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.

In Section I we argued that, while our RP and SP datasets are not identical, they are as similar as one is likely to obtain. We argued that RP and SP data will, by construction, never be identical, because they are collected for different purposes, and are constructed in very different ways. We conclude this section by giving some examples to illustrate this point:

One of the best attempts to obtain similar SP and RP data is Swait and Andrews (2003). They use IRI scanner data on liquid detergent purchases by Chicago households over 112 weeks in 1995-7. Then, in October 2000, they mailed a liquid detergent SP choice task to 2000 Chicago households (roughly 400 responded). They use these data to estimate N-MIXL models and test whether preference weights for common attributes are equal across the RP and SP data. While Swait and Andrews went to great lengths to obtain similar data, several differences are notable:

First, in the actual market there are 84 choice options, while the SP data contain only 13 options. This smaller choice set size is almost inevitable – i.e., SP data will almost always have relatively small choice sets, in order to make the choice task manageable for respondents.

Second, as in our data, the attributes are not identical between the SP and RP data. The SP data contain additional attributes not observed in the RP data. This is not surprising, as the

[13] 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.

main raison de'etre for SP data is to assess demand for new attributes (not traded in markets). Conversely, the RP data contain display/feature information, which is hard to mimic in SP data.

Third, both data sets were collected in Chicago, but the households differ. The process of self-selection into the data may also differ. Also, the SP data was collected a few years later.

In summary, we see that, even in the best comparative studies, SP and RP data are rather different.[14] Furthermore, to put any data limitations in context, note that ours is the first study to compare heterogeneity in RP vs. SP data, not just mean utility weights or aggregate predictions.

IV. Empirical Results

In Sections IV.A and IV.B we compare choice models estimated on SP and RP data. 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 parameters (k).

IV.A. 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. (Note: We treat all attributes as exogenous, given that they are experimentally controlled).

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)

[14] Another attempt to obtain similar SP and RP data is Brownstone et al (2000), who estimated auto demand models. The SP choice sets had six hypothetical vehicles (3 cars, 3 trucks). Several months later, respondents were asked about actual vehicle purchases. The RP choice set contained 689 vehicle types. Notably, not only was the SP choice set much smaller, but it was designed so most (or all) of the 6 options were alternative fuel vehicles (as opposed to conventional gasoline vehicles). Thus, the choice objects are quite different in the two datasets. Finally, the RP data had only one choice occasion per household, so it would not be useful for our purpose of estimating heterogeneity.

Another good example is Small, Winston and Yan (2005, 6). The RP data is on choice of free vs. express lanes on California SR 91. The SP data gave subjects a choice of regular vs. express lanes on a *hypothetical* highway with stated characteristics. This study is unusual in that choice set size is the same in the RP and SP data (possible because there are only 2 choices). But a *hypothetical* highway is clearly not the same thing as SR 91. Our point is that all studies comparing RP and SP data require us to accept some differences in the nature of the choice sets.

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 Var(β) matrix. We report an N-MIXL model with uncorrelated coefficients (i.e., a diagonal Var(β) matrix), as it achieves a better BIC value than the version with correlations. The log-likelihood 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 <u>both</u> 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 <u>all</u> 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.[15] As we will see below (see p. 13), the ability to capture both these types of behavior *simultaneously* is why G-MNL fits better than N-MIXL on these data. Finally, the parameter γ is essentially 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 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

[15] 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 <u>common factor</u> 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).

must be larger for observed attributes to have the same effect on choice.[16]

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 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 sets of consumers who place great weight on some or all of these attributes, as well as some consumers who place little weight on all of them. This is a feature that the MM-MNL model 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).

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). The number of choice occasions per person is also larger (32 vs. 16), as is the number of subjects (328 vs. 178). Given the large number of attributes, it is not surprising that the structure of heterogeneity is more complex in Pizza B. The LC model now identifies 6 segments, and the MM-MNL model identifies 3 types.

In the MNL model, the price coefficient is very close to that in Pizza A (-0.17 vs. -0.16). The new variable for free delivery is positive as expected (0.12). The most important attributes are again freshness and hot delivery, but the magnitude of their coefficients drops roughly in half from the Pizza A case. This seems to be a consequence of introducing several new attributes. People put small but significant weights on several of these, including crust type, size options, local store, baking method, vegetarian option, guaranteed delivery time, and variety.

In the S-MNL and G-MNL models the τ parameters that govern the degree of scale heterogeneity are again large and highly significant (1.22 and 1.50, respectively). Again, G-MNL

[16] See Keane (1997b) 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.

achieves substantial likelihood and BIC improvements over both S-MNL and N-MIXL. For instance, the log-likelihood improves from -5892 to -5662 (230 points) in going from N-MIXL to G-MNL. As before, the parameter γ is essentially 0, suggesting the normal errors are scaled proportionately to the scale of the mean β vector. 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.

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. An alternative explanation is that, given 5 segments devoted to types who like specific attributes, the best fit is obtained simply by grouping everyone else into a portmanteau type that lacks strong preferences. The behavioral implications of these explanations are radically different, particular regarding how well SP data can predict behavior in real markets (e.g., Does the RP data also have a large "random" segment?). The comparison of RP and SP results in Sections IV.B and V will shed light on this issue.

Turning to the MM-MNL model, 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. In an overall comparison, the 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, while in Pizza A the situation was reversed.

IV.A.3. ***Summary: Behavioral Patterns in the SP Data***

To summarize, G-MNL and MM-MNL are preferred to N-MIXL in both datasets, while N-MIXL is preferred to the other models. To better understand why the G-MNL and MM-MNL models are preferred to N-MIXL, 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 ingredient freshness 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 has difficulty capturing "extreme" consumers who place great weight on price or freshness.

In contrast, the posteriors of G-MNL and MM-MNL depart substantially from normality. In the left panel of Figure 1, both models generate a mass of consumers in the left tail who care intensely about price. And in the right panel, we see both models generate a mass of consumers who care intensely about fresh ingredients. The G-MNL and MM-MNL posteriors also exhibit excess kurtosis – i.e., 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. These differences in the posteriors give us a clue as to why G-MNL and MM-MNL may fit better than N-MIXL. So now we consider if these patterns are consistent with patterns in the 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 quality (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 3) 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 random behavior: If a subject chooses randomly between options A and B, the mean attribute differences between the chosen and non-chosen options will be "close" to zero (except for sampling variation). Keane and Wasi (2012a) give a definition of randomness based on this idea. 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.

Combining results for the 113 lexicographic 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 over N-MIXL. 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 extremely strong preferences for particular attributes and (ii) consumers whose choices are little affected by the whole attribute vector. This is because both models use mixture-of-normal distributions that are more flexibile than the normal distribution assumed by N-MIXL.

A more subtle question is why G-MNL is preferred over MM-MNL in Pizza A, while MM-MNL is preferred in Pizza B. Note that MM-MNL has a superior log-likelihood to G-MNL in both datasets. But it also has many more parameters (i.e., 33 vs. 18 in Pizza A and 98 vs. 34 in Pizza B). As we discussed earlier, the structure of heterogeneity is considerably more complex in Pizza B than in A. According to BIC, the complexity of Pizza B justifies the extra parameters of MM-MNL. Still, G-MNL and MM-MNL make very similar behavioral predictions.

IV.B. The Revealed Preference Data

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. The data contain 4,123 purchase occasions. Variables included in the model are price, an indicator for promotion, 5 indicators for name brands ("other" is the omitted category), 6 indicators for different toppings ("combo" is the omitted category), 2 indicators for crust type (regular is the omitted category) and an indicator for microwaveable. As we noted in Section II, with many varieties it is natural (and practical) to use brand intercepts rather than ASCs. So, as in Fader and Hardie (1996), the attributes based approach implies there are many fewer parameters (15) than options (102).

We treat price as exogenous (contrary to a common practice in IO), under the assumption that brand intercepts capture latent quality of brands and/or brand equity – see Erdem, Imai and Keane (2003) for arguments in favor of our approach. The concept of brand equity incorporates both perceived quality and consumer familiarity with brands (Erdem and Keane (1996)).

The large size of the choice set ($J \approx 100$), combined with the large sample size, causes significant computational burden in estimating models that do not have closed form choice probabilities (i.e., the models with heterogeneity). To deal with this problem we use randomly chosen subsets of the full choice set in estimation. This procedure deserves further comment:

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 Web Appendix, McFadden (1978)'s result does not hold when MNL is extended to include heterogeneity. But in the Web Appendix we also report results of an extensive Monte Carlo study designed to assess the bias induced by using random subsets to estimate the N-MIXL, G-MNL and MM-MNL models. In our experiments the full choice set has 60 alternatives and the random subsets have 10 or 20 alternatives. We find little (if any) evidence that use of random subsets of the full choice set induces bias in estimates of multinomial logit models with heterogeneity.[17] (This finding may be useful in many contexts).

Based on our Monte Carlo results we decided to estimate the RP models using 3 different choice set sizes, 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 Web Appendix. A striking finding is that estimates are quite stable across choice set sizes for all 6 models (MNL, S-MNL, N-MIXL, G-MNL, LC, and MM-MNL) – see Web Appendix for a detailed discussion. This stability of the estimates adds to our confidence (already considerable in light of the Monte Carlo results) that any bias induced by using random subsets of the full choice set is negligible, even for models that include heterogeneity. Given this, we will focus our discussion on the RP results for choice sets of size 40.

Consider first the MNL estimates in Table 4. The price coefficient is -0.84 while the promotion dummy is 0.81. Thus, putting a type of pizza on promotion (making it more visible in the store) is equivalent to roughly a $1.00 price cut. It is easy to show that the price elasticity of demand for an alternative with market share s and price p is simply $e = (-.84)p(1-s)$. For a pizza of average price ($2.80) and market share ($s$=1/J) we have a price elasticity of -2.3. This is in the ballpark of prior RP estimates for supermarket goods (see Keane (2010)). Note that the elasticity formula depends on choice set size through the $(1-s)$ term. But for a typical option $(1-s) \approx (J-1)/J$. This is close to one and little affected by choice set size (if we use a reasonably large random subset, say, $J \geq 20$). So the random choice set size will have little impact on elasticity estimates.

[17] 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 Web Appendix.

It is interesting to compare the price elasticity in the RP data to what we found in SP: In the SP data, price is coded so that -1 corresponds to $13 and 1 corresponds to $17 (this is known as "effects coding" in the DCE literature). 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 0.34. This increases demand by roughly 17%. As the price decrease is 24%, the implied price elasticity is roughly -0.71. (For a $4 price *increase* we get -0.55). The Pizza A results are very similar.

So, the RP data implies a price elasticity of demand (-2.3) more than three times greater than in the SP data (-0.7). A monopolistically competitive firm operates in a region where the price elasticity of demand is ≤ -1.[18] And both pizza and pizza delivery are well described by the monopolistically competitive model (i.e., differentiated product, many sellers).[19] Thus, one might question using the small elasticity in the SP data to predict behavior in an actual market.

Several factors may account for the low SP elasticity estimate. It may be due to 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 choose randomly as discussed earlier.[20]

Alternatively, note that elasticities in SP data depend on preferences and experimental prices. But RP elasticities are a function of preferences and equilibrium market prices. Thus, RP and SP elasticities may differ because they are evaluated at different points on the demand curve.

As a concrete illustration of this point, consider symmetric Bertrand competition with differentiated products and linear demand curves $q_j = \beta_0 - \beta_p p_j + \beta_c p_k$ for j,k =1,2 where $\beta_p > \beta_c$, and a constant marginal cost of c. Then the equilibrium price is $p=(\beta_0 + \beta_p c)/(2\beta_p - \beta_c)$ and the price elasticity of demand is $e = -(\beta_0 + \beta_p c)/[\beta_0 - (\beta_p - \beta_c)c]$. Note that e depends not just on the preference parameters β_0 , β_p and β_c but also on marginal cost c and, more subtly, on the Bertrand assumption itself. Note also that $e \leq -1$ where $e = -1$ if c=0.

Now, suppose an experimenter sets price at a fraction $0<\Delta<1$ of the lowest possible market price, so $p= \Delta\beta_0/(2\beta_p - \beta_c)$. The price elasticity at this hypothetical experimentally set price is $e = -\Delta\beta_p/ [(1-\Delta)(2\beta_p - \beta_c) + \Delta\beta_p]$. This only depends on preference parameters and Δ.

[18] 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.

[19] In the Australian context there are many small pizza shops that deliver different types of pizza.

[20] 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. The set of controls for product attributes is as least as rich in the SP data as in the RP data. And adding eight new controls in going from Pizza A to B hardly changes the elasticity. Alternatively, a larger error scale in the SP data could occur because some subjects don't take the task seriously (as noted in the text).

For example, say β_0 =10, β_p = 2, β_c =1 and c=1. Then in market equilibrium price would be 4 and e = -1.33. But if the experimenter sets Δ=3/4, then price is 2.5 and e = -0.67. So a plausible explanation for small elasticities in SP data is if prices are set below market level. Note, however, that this may not invalidate the experiment as a way to uncover taste parameters.

Other aspects of the results are that the most "popular" brands, as indicated by the brand intercepts, are Tombstone and Jacks, while the least popular is Bernatello. (Recall that the intercepts capture latent quality and/or brand equity). 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 quite poorly in practice because 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 attributes like price, it is unrealistic to assume brand intercepts have the same signs for all consumers. Instead, we assume the brand intercepts are normally distributed.[21]

S-MNL achieves a large log-likelihood improvement over MNL (-11,930 vs. -13,602) with 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 intercepts. The estimate of τ is 0.86, which is highly significant, but smaller than in the SP data. The β vector is similar to that for MNL, although the (mean) price coefficient increases to -0.97.

Next, Table 4 reports the N-MIXL results. The version preferred by BIC imposes a one-factor structure on the covariance matrix of the β_n vector. As the model contains 16 variables (11 attributes and 5 brand intercepts) a full covariance matrix would have 136 parameters. The one-factor structure 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. The β vector is similar to both the MNL and S-MNL models, although the (mean) price coefficient increases to -1.21.

The G-MNL results are interesting, as they contrast sharply with those for the SP data. First, G-MNL leads to only a modest (31 point) log-likelihood improvement over N-MIXL. In percentage terms this is only 0.3%, compared to the gains of 2.1% and 3.4% in Pizza A and B.

[21] 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.

Second, the estimate of 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. These 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.[22] 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. Elrod and Keane (1995) found that LC models tend to understate heterogeneity in consumer preferences.

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. Clearly, MM-MNL has the best BIC of all models considered; i.e., 297 better than G-MNL and 344 better than N-MIXL.[23] In contrast, in the SP data, MM-MNL and G-MNL had fairly similar BIC values.

The two types 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. It is interesting to compare these RP elasticities with the SP results. In Pizza B we found three types, with elasticities (at the mean) of -.42 (41%), -2.96 (31%) and -.71 (28%). So, as in the MNL model, elasticities are much smaller in the SP data.

Figure 2 plots posterior distributions of person-level price and vegetarian parameters for N-MIXL, G-MNL and MM-MNL. The N-MIXL and G-MNL posteriors for price (left panel) look quite similar. Each departs modestly from normality, with a moderate degree of skewness to

[22] 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.).

[23] As we show in the Web Appendix, the use of reduced choice sets biases the BIC toward smaller models. Does this affect our results? MM-MNL is favored over G-MNL by 297 points using a random choice set of 40. If we instead use choice sets of 30 or 20, the margins are 252 and 230. A similar pattern holds for N-MIXL. So the finding that MM-MNL is preferred over G-MNL and N-MIXL is reinforced by using larger choice sets. This is as expected, as MM-MNL has the most parameters of the three models (66 vs. 50 vs. 48).

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 the G-MNL and N-MIXL models.

These RP results are very different from 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. We discuss implications of this difference in Section V.

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.

V. Differences between the SP and RP Data

We have seen some important differences between the SP and RP data. For instance, price elasticities are much smaller in SP, and both lexicographic and random choice behavior are much more common. In general, the shapes of the individual-level distributions of parameters differ substantially. Due to this difference in the structure of heterogeneity, the preferred model of heterogeneity differs as well. In this section we explore these issues in more detail.

In SP data, we found that G-MNL and MM-MNL fit better than N-MIXL mostly because they capture both: (i) "extreme" consumers who exhibit lexicographic behavior with respect to certain attributes like price or quality, and (ii) consumers who are insensitive to price and other attributes, and thus appear to make choices at random.[24]

These lexicographic and random behaviors we see in SP data are not prevalent in the RP data. For instance the MM-MNL model (which is the best fitting) implies that relatively few consumers place great weight on price (i.e., as we see in Figure 2, it puts much less mass on price coefficients less than -2.6 than do either N-MIXL or G-MNL). And all our models imply that relatively few consumers in the RP data have price coefficients near zero (see Figure 2).

An important question is whether the very different distributions of the price coefficient in SP vs. RP arise from fundamental differences between these types of data, or another factor. A potential issue is that the goods in the RP and SP data are similar but not identical. However, as we saw in equation (4), the u_c component of the price coefficient should be equal across data on

[24] In a true RUM choice is deterministic for consumers, given observed and unobserved (to the analyst) attributes. But Feibig et al (2010) find that subjects in SP data often change their choice if a scenario is repeated.

different *inexpensive* goods. Thus, 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 compared to the RP data, (i) the σ_n distribution must have much more mass near zero, 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. Why would many consumers put a huge weight on unobserved attributes of delivered pizza (substantial mass of σ_n near zero), but not of frozen pizza? Indeed, the controls for product attributes are at least as rich in the SP data as in the RP data.

Furthermore, we emphasize that σ_n is a scalar that shifts all attribute coefficients (see equation (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, the segment of consumers with very large σ_n should tend to have both large price coefficients and large weights on quality.

Figure 3 gives a scatter plot of posteriors of the (absolute) 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.[25]

Based on this evidence, the most plausible explanation for the very different distribution of the price coefficient in the SP data is that the u_{cn} differ from the RP data. It appears that many 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

[25] There is also no evidence that a negative correlation between the β_{nk} for price and quality in the SP data can explain this pattern. In the two models that estimate this correlation (G-MNL and N-MIXL) an uncorrelated coefficients model is preferred (see Table 3).

simplify the experimental task. This leads to much greater heterogeneity in σ_n in the SP data (i.e., higher values of τ). Finally, we stress the patterns we have described are not peculiar to SP data on frozen pizza. Feibig et al (2010) find similar patterns in SP data on several other products.

We now turn to the question of why the ranking of models differs in the RP vs. SP data. As we saw in Section IV.A, the main 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. Hence, the fit of G-MNL is similar to that of N-MIXL in the RP data.

But why does MM-MNL fit much better than both G-MNL and N-MIXL in the RP data? Table 5 sheds light on this issue. First, we group consumers in the SP and RP data 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 the SP data (Pizza B) by levels of price sensitivity. For each subject, we calculate the average difference in price between the chosen and rejected options (over all 32 choice tasks). 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. The more negative is the average price difference, the more price sensitive is the consumer.

Only 19 levels of the average price difference are observed in the SP data, so we cannot group subjects into 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 price 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.

Note 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 5 contains a similar analysis for the quality 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 are indifferent to fresh ingredients. Thus, the top two panels of Table 5 provide evidence 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.

The bottom panels of Table 5 repeat this type of analysis for the RP data. Here we look at the price and vegetarian attributes. In the RP data there are many choice options, so we 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. Unlike the SP data, the average attribute differences are not bunched at a few points, so we can group people into quintiles.

In the 3rd panel of Table 5, 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.[26] Another striking difference is that, in RP, G-MNL does not fit the most price sensitive consumers better than N-MIXL.[27] 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 5, 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.

[26] 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 5).
[27] Indeed, the average BIC difference for G-MNL vs. N-MIXL in this group is -0.55, so N-MIXL fits slightly better.

In Section IV.A we discussed why N-MIXL has trouble generating posteriors that depart substantially from normality (see also Allenby and Rossi (1998)). But 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 SP the modal price coefficient is close to zero, so increasing kurtosis and increasing mass near zero are equivalent. But in the RP data the mode is far left of zero (see Figure 2), so the two 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 4A). 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 4B). 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 4B) is a very poor representation.

The flexibility of MM-MNL is even more apparent for the vegetarian coefficient. In the 4th panel of Table 5 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 likes 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. The most important departure in the SP data (i.e., excess mass both near zero and in the tail) is well captured by a normal scaled by a lognormal. Thus, G-MNL fits about as well as MM-MNL in the SP data, 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. Hence, MM-MNL is clearly preferred to both. The MM-MNL model does well in both contexts because a discrete mixture-of-normals can approximate any density (Ferguson (1973)).

VI. Conclusion

In recent years it has become increasingly common to use stated preference (SP) data to study consumer demand. And prior work has compared aggregate market share predictions or mean parameter vector estimates from models based on SP vs. RP data.[28] But to our knowledge the present study is the first to compare heterogeneity in the distributions of tastes.

We have found substantial differences in the structure of heterogeneity in demand models estimated on RP vs. SP data. This has two important implications: (i) it affects which model of heterogeneity is best for each type of data, and (ii) it raises obvious questions about whether SP data is really useful for predicting demand (at least if heterogeneity matters).

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 care little about attributes), as well as a mass of consumers in the tail who care greatly about one or two attributes. 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 data reveals groups of subjects whose behavior is consistent with such rules.

Our results only apply to two pizza datasets, but Fiebig et al (2010) found similar patterns in eight SP datasets on very different products. Thus, these behaviors appear to be typical in SP data. In contrast, there is little evidence of lexicographic or "random" behavior in the RP data.

Comparing models, both G-MNL and MM-MNL can capture the typical heterogeneity pattern in SP data. This is because both involve mixtures-of-normals that can generate both kurtosis near zero (i.e., random behavior) and fat tails (i.e., lexicographic behavior). The popular N-MIXL model, which assumes normal heterogeneity, has difficulty fitting these patterns.

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. As a result, MM-MNL is clearly preferred. The obvious virtue of MM-MNL is its flexibility. It does well in both the RP and SP contexts because a discrete mixture-of-normals can approximate any density (Ferguson (1973)). An avenue for future research is to improve on G-MNL by using more flexible distributions for the scale factor.

Notably, if we had relied exclusively on the popular N-MIXL model, and not considered

[28] 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)) and other common types of normalization.

the more general models (MM-MNL and G-MNL), the difference in heterogeneity structures between the SP and RP data would not have been readily apparent. This is because of the strong tendency of N-MIXL to draw in all heterogeneity distributions towards its normal prior.

The very different heterogeneity structures in RP vs. SP data obviously raise questions about the usefulness of using SP data to predict demand. This is particularly true if we are not so interested in aggregate demand (i.e., market shares) as in questions where the whole distribution of consumer taste heterogeneity matters: price discrimination, welfare calculations, etc.. The surprisingly small price elasticities of demand in the SP data are also of concern.

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.

An alternative interpretation is that both the RP and SP data reveal valid information about preferences, and the different patterns of taste heterogeneity in the two types of data result from the different structure of the choice tasks. A key difference is the SP task presents subjects with one time choices among hypothetical objects. In the RP task, choices are repeated over time and there are many more options. These differences have at least three important implications:

(i) In the actual market setting consumers have the opportunity to learn about the different brands and varieties over time (see Erdem and Keane (1996)). This makes it easier to have some knowledge of the attributes of all the alternatives, even if there are many.

(ii) The market setting prevents consumers from using simple rules like "choose the cheapest option." This is because with many options it is not feasible to check all prices on any given shopping trip. Presumably the consumer relies largely on what he/she has learned about typical prices of various options in the past. The same logic applies to any other attribute.

(iii) In the market context choices take place over time, so there may be variety seeking. But subjects may interpret the experimental choices as one time choices. In general this may alter behavior (e.g., I usually prefer sausage to bacon, but on some shopping trips I may try bacon).

Points (i)-(iii) highlight that choice model parameters capture not only preferences but also (a) market structure and (b) the informational and other constraints under which the choice is made. Thus, it is not surprising that RP and SP results differ, as consumers face quite different

market structure and constraints in the two contexts. Indeed, in Section IV.B we discussed how differences in market structure could explain the lower price elasticity estimates in the SP data.

This discussion raises an obvious point: A potential criticism of our work is that the SP choice tasks were not designed to be similar enough to a market environment. But it is important to note that our SP data is representative of the SP data that is currently used in practice. Thus, our results suggest that, given the current state of practice, SP choice experiments may not enable one to make reliable inferences about consumer taste heterogeneity in actual markets.

In conclusion, our results suggest caution when using SP data to infer distributions of consumer preferences. This is important because use of SP data to model demand has become very common (in industry and academia). Future research should focus on improving the design of SP experiments so their results are more representative of actual consumer taste distributions. This may require making SP experiments more "realistic," as suggested by Arrow et al (1993).[29]

Of course, there is a limit to how far one can go in making SP data more realistic, as (i) subjects will not tolerate very large choice sets, or tasks that are too difficult in other ways, and (ii) the temporal aspect of choice is very hard to replicate in an experiment. One way to enhance realism is to give subjects a budget, and have them actually buy at least one of the items they choose (see Ding et al. (2005)). But a key point of SP is to estimate demand for prospective products and/or attributes that don't yet exist (so obviously they can't be bought). We are aware of proprietary research that tries to model behavior given large choice sets by using replica store environments. But again, these stores must be stocked with existing products. In a classic paper Roberts and Urban (1988) used experimental data to model consumer learning over time about a single new car model. But their approach is not easily applicable to frequently purchased goods with many varieties. Clearly, the development of more realistic SP choice experiments remains an important area for future research.

[29] Alternatively, Harris and Keane (1999) combine SP and RP data in a different way: They use the SP data to give noisy measures of the attribute weights in an RP discrete choice model with heterogeneous preferences. They find this extra information about preferences leads to a dramatic improvement in the fit of the RP choice model.

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.

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.

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R. and H. Schuman (1993), "Report of the NOAA Panel on Contingent Valuation," *Federal Register*, 58:10, 4601–4614.

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.

Berninger, K., W. Adamowicz, D. Kneeshwa and C. Messier (2010), "Sustainable Forest Management Preferences of Interest Groups in Three Regions with Different Levels of Industrial Forestry: An Exploratory Attribute-Based Choice Experiment," *Environmental Management* 46:117–133

Brooks. K. and J. Lusk (2010), "Stated and Revealed Preferences for Organic and Cloned Milk: Combining Choice Experiment and Scanner Data," *American Journal of Agricultural Economics*, 92:4, 1229-41.

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.

Carson, R.T., J.J. Louviere, D.A. Anderson, P.Arabie, D.S. Bunch, D.A.Hencher, R.M. Johnson, W.F. Kuhfeld, D.Steinberg, J. Swait, H. Timmermans and J.B.Wiley (1994), Experimental analysis of choice. *Marketing Letters* 5: 351-361.

Carson, R.T., N.E. Flores, K.M. Martin and J.L.Wright (1996), Contingent valuation and revealed preference methodologies: Comparing the estimates for quasi-public goods, *Land Econom*ics 72, 80-99.

Ding, M., R. Grewal and J. Liechty (2005), Incentive-aligned conjoint analysis, *Journal of Marketing Research*," 42(1), 67–82.

Earnhart, D. (2002), "Combining Revealed and Stated data to examine housing decisions using discrete choice analysis" Journal of Urban Economics 51, 143–169.

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. and M.P. Keane (1996), "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets," *Marketing Science,* 15(1), 1-20.

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

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.

Goatt, A. (1998), 'Estimating Customer Preferences for New Pricing Products', Electric Power Research Institute Report TR-111483, Palo Alto

Hall, J.P., D.G. Fiebig, M. King, I. Hossain and J.J. Louviere (2006), "What influences participation in genetic carrier testing? Results from a discrete choice experiment," *Journal of Health Economics*, 25, 520-537.

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.

Heckman, J. and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* , 52(2), 271-320

Hensher, D. and M. Bradley (1993), "Using stated response choice data to enrich revealed preference discrete choice models," *Marketing Letters*, 4(2), 139-151.

Hjelmgren, J. and A. Anell (2007), "Population preferences and choice of primary care models: A discrete choice experiment in Sweden," Health Policy, 83(2), 314-322.

Hensher, D.A. (1994), Stated preference analysis of travel choices: the state of practice. *Transportation* 21: 107-133.

Hensher, D.A. (2001), "The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications," *Transportation*, 28(2), 101-118.

Hensher D.A. and W.H. Greene (2003), The Mixed logit model: The state of practice. *Transportation* 30: 133-176.

Horowitz, J. and J.J. Louviere (1990). "The External Validity of Choice Models Based on Laboratory Choice Experiments," in M.M. Fischer, P. Nijkamp, and Y.Y. Papageorgiu, eds., Spatial Choices and Processes, North-Holland Publishing Co., 247-263.

Keane, Michael P. (1997a), "Current Issues in Discrete Choice Modelling," *Marketing Letters*, 8, 307-322.

Keane, Michael P. (1997b), "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 N. Wasi (2012a), "Comparing Alternative Models of Heterogeneity in Consumer Choice Behavior," *Journal of Applied Econometrics*, forthcoming.

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

Lerman, S. and C. Manski, C. (1981), 'On the use of simulated frequencies to approximate choice probabilities', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, 305–319.

Louviere, J.J. and G. Kocur (1983), "The Magnitude of Individual-level Variations in Demand Coefficients: A Xenia, Ohio case example," Transportation Research A, 17A(5), 363-373.

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.

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. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica* , 57(5), 995-1026.

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).

Morikawa, T., Ben-Akiva M. and D. McFadden (2002), "Discrete Choice Models Incorporating Revealed Preferences and Psychometric Data," in P.H. Franses and A.L. Montgomery (Eds.), Econometric Models in Marketing, Advances in Econometrics, Vol. 16, 27-53, Elsevier Science: Oxford.

Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica* , 54, (4), 755-784.

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.

Roberts, J. and G. Urban (1988), "Modeling Multiattribute Utility, Risk and Belief Dynamics for new Consumer Durable Brand Choice," *Management Science*, 34 (February), 167-185.

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

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

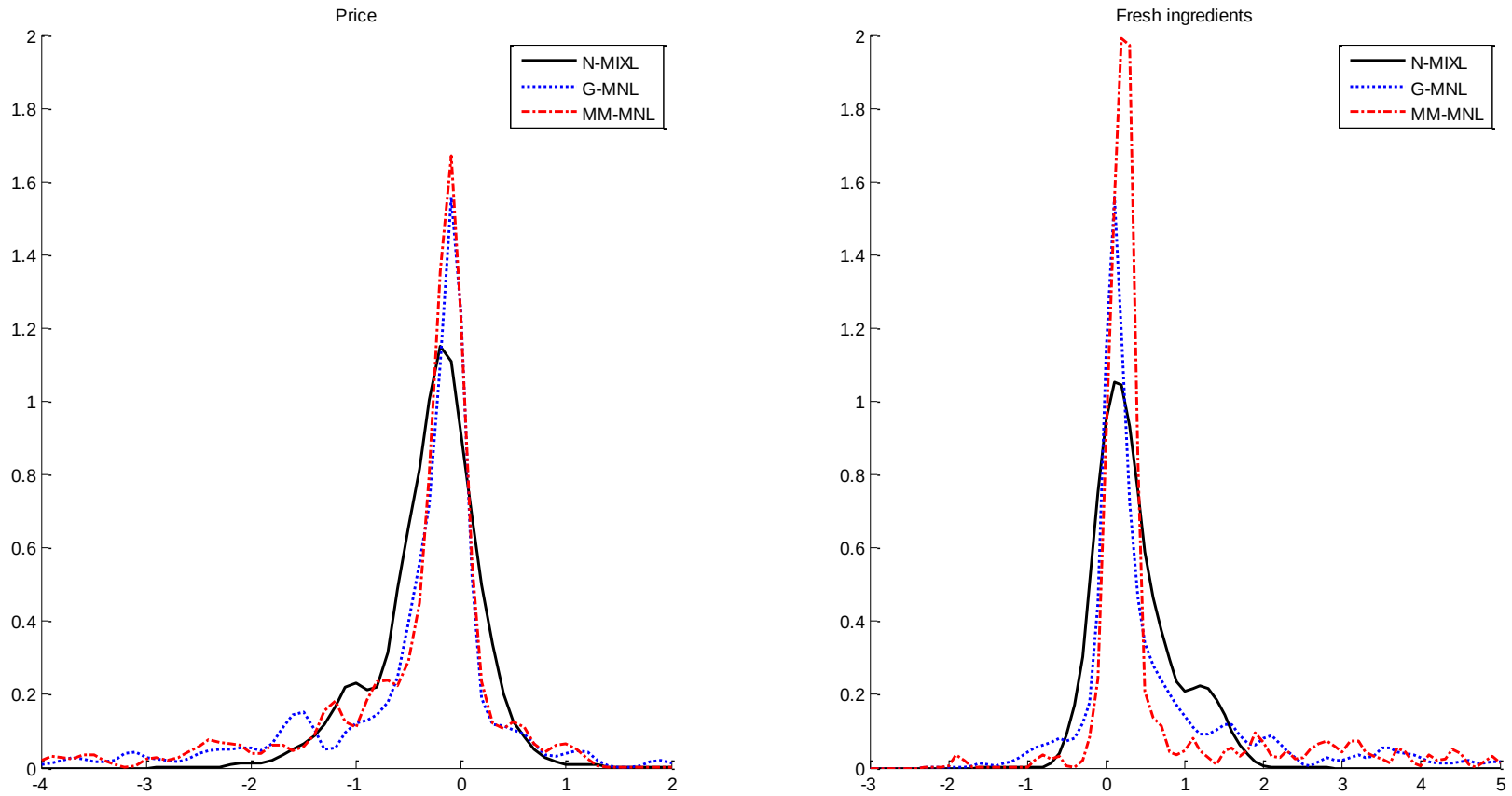
_____ (2006), Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design. *Brookings-Wharton Papers on Urban Affair*, 53-96.

Swait, J. and J.J. Louviere (1993), "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research* , 30(3), 305-314.

Swait, J. and R.L. Andrews (2003), "Enriching Scanner Panel Models with Choice Experiments," Marketing Science, 22(4), 442-260.

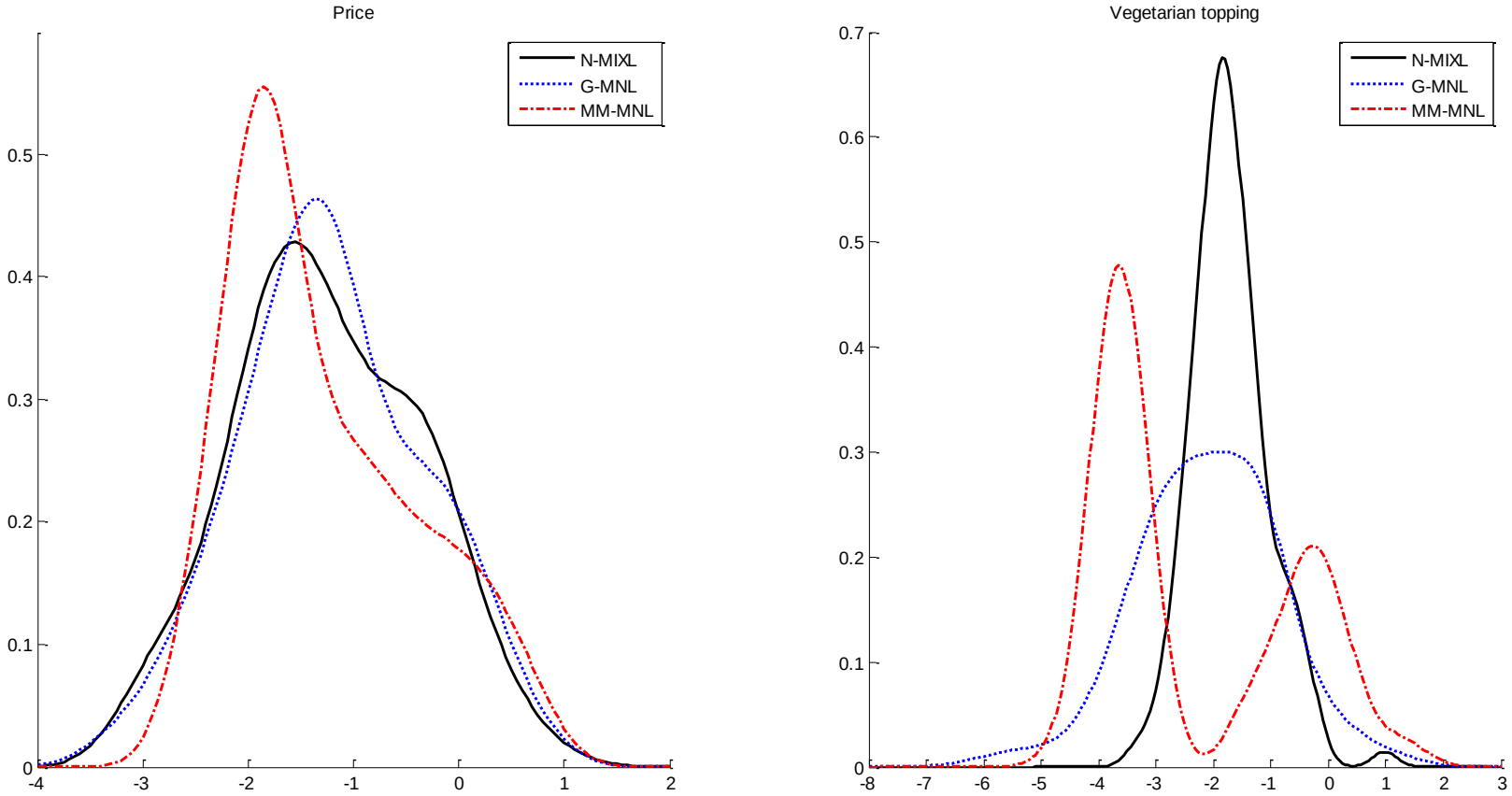
Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press.

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



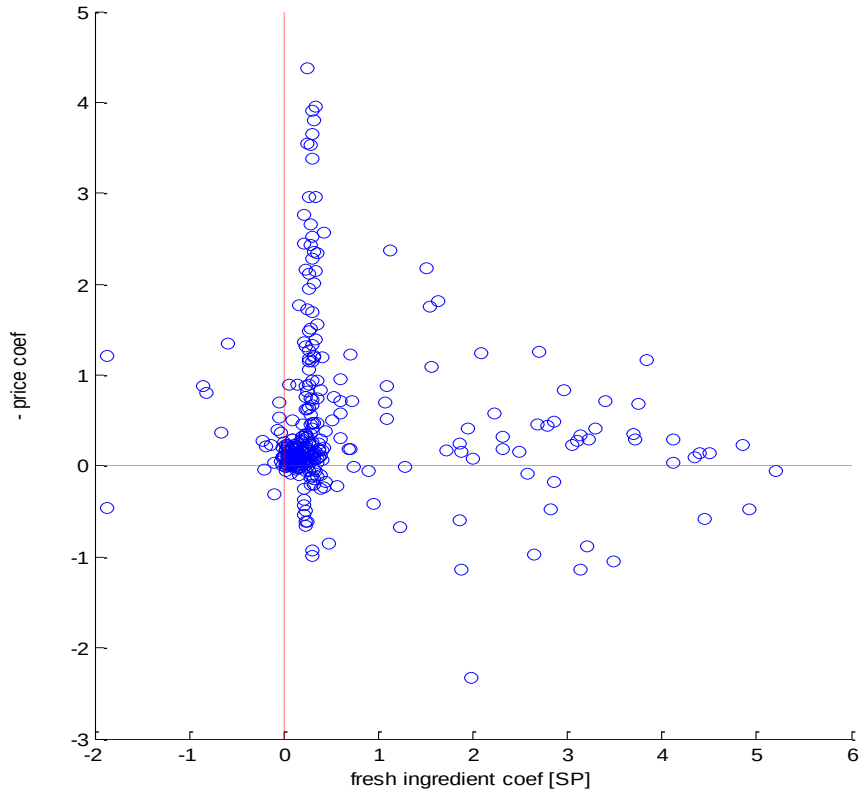
Note: Each kernel density estimate uses a normal kernel with an optimal bandwidth (h). The formula used is $h = \sigma(4/3N)^{1/5}$ where σ is the standard deviation and N is the number of observations. The optimal bandwidths for N-MIXL, G-MNL and MM-MNL for the price coefficients are .111, .095 and .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 $h = \sigma(4/3N)^{1/5}$ 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 .355, .346 and .303, respectively. For the vegetarian topping coefficients, the optimal bandwidths are .218, .513 and .380, respectively.

Figure 3: 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.

Figure 4: Comparison of the Normal Distribution and the Continuous Mixture-of-Scaled-Normals Distribution

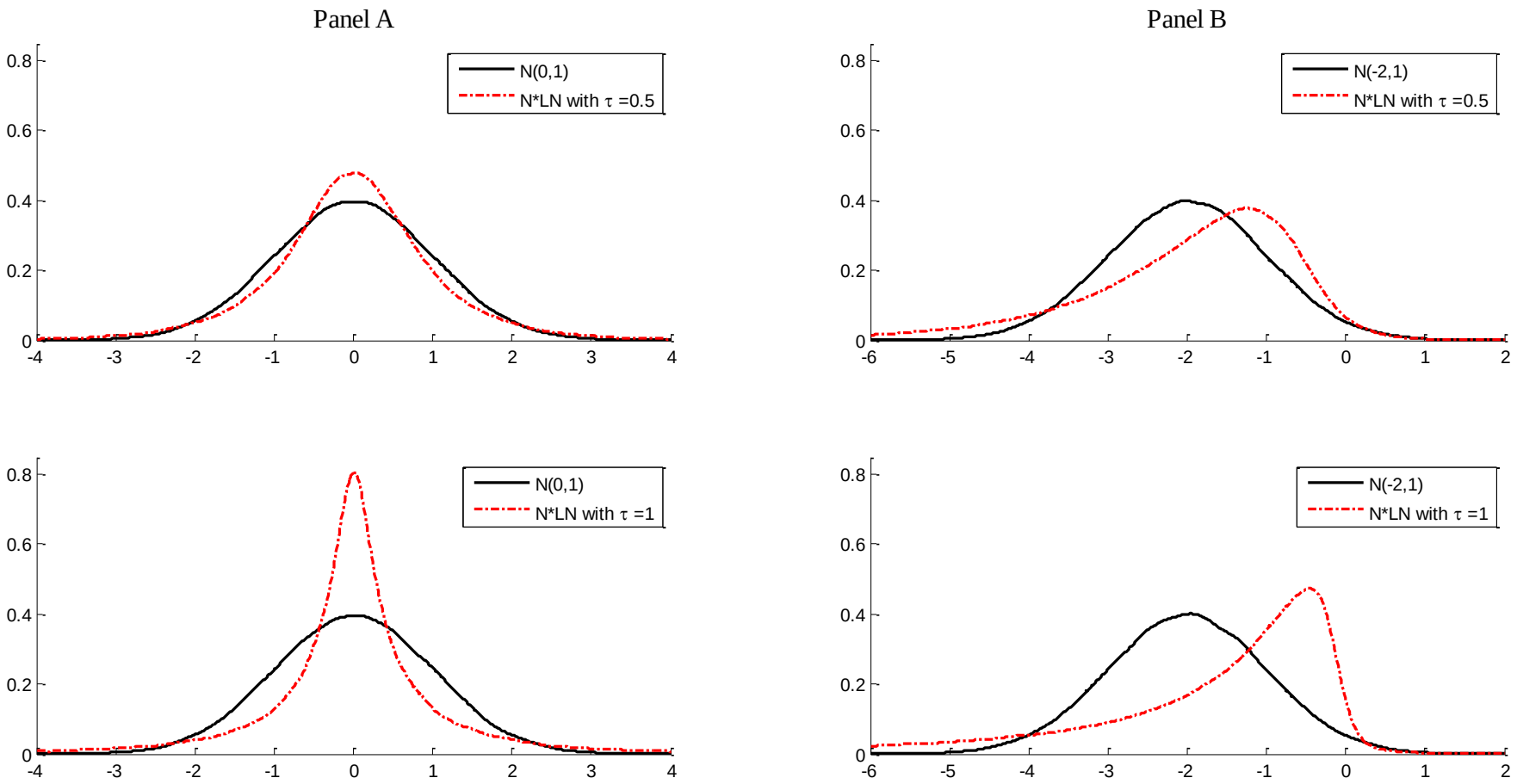


Table 1: Average Characteristics of Alternatives in the RP Scanner Data

	All	Tombstone	Roma	Jacks	Red baron	Bernatello	Others
Chosen choices		1085 (26%)	833 (20%)	844 (20%)	437 (11%)	238 (6%)	686 (17%)
Average characteristics of available choices (2001-2003)							
Price	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.) Entries with n/a indicate the option does not exist.

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. For details see Fiebig et al. (2010).

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).

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 5: 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	
			Range	Average		MM-MNL over N-MIXL	G-MNL over N-MIXL
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).