11-2009

Extending the Generalized Multinomial Logit Model: Error Scale and Decision-Maker Characteristics

Eleanor M. Feit
University of Pennsylvania

Extending the Generalized Multinomial Logit Model: Error Scale and Decision-Maker Characteristics

Abstract
This essay contributes to the development of models that allow for heterogeneity across respondents in the error scale of the multinomial logit model. The potential to explain respondent heterogeneity by differences in error scale has been recognized for some time (Louviere 2001), but models that allow for continuous error scale heterogeneity have only recently been developed (Sonnier, Ainslie and Otter 2007, Keane et al. 2009). The most general of these models is the "Generalized Multinomial Logit Model" (G-MNL), which allows for heterogeneity both in error scale and all attribute preferences, including the price attribute (Keane et al. 2009). We further develop the G-MNL by proposing a Bayesian estimation strategy, allowing for straightforward incorporation of decision-maker characteristics as covariates to individual-level error scale, in a way that is computationally tractable. In a data set on personal computer (PC) choices in a survey setting (Lenk et al. 1996), we find that respondents who are older have higher average error scale indicating that they make less reliable decisions than those who are younger. Respondents who perceive themselves to be expert when it comes to making PC choices have lower average error scale, indicating that they make more reliable choices. These findings are consistent with recent theorizing on the relationship between cognitive resources and error scale (Swait and Adamowicz 2001). We also facilitate the use of G-MNL in practice by empirically exploring the data requirements for obtaining accurate estimates of the G-MNL and find that estimating this model requires a somewhat larger number of respondents and a larger number of observed choices per respondent than is typical in commercial market research.

Keywords
error scale, generalized multinomial logit, discrete choice, Bayesian MCMC

Disciplines
Behavioral Economics | Business | Business Analytics | Marketing | Statistics and Probability

Comments
This is an unpublished manuscript.

Extending the generalized multinomial logit model:

Error scale and decision-maker characteristics

Eleanor McDonnell Feit

The Modellers, LLC

PRELIMINARY DRAFT: 23 November 2009

Abstract

This essay contributes to the development of models that allow for heterogeneity across respondents in the error scale of the multinomial logit model. The potential to explain respondent heterogeneity by differences in error scale has been recognized for some time (Louviere 2001), but models that allow for continuous error scale heterogeneity have only recently been developed (Sonnier, Ainslie and Otter 2007, Keane et al. 2009). The most general of these models is the "Generalized Multinomial Logit Model" (G-MNL), which allows for heterogeneity both in error scale and all attribute preferences, including the price attribute (Keane et al. 2009). We further develop the G-MNL by proposing a Bayesian estimation strategy, allowing for straightforward incorporation of decision-maker characteristics as covariates to individual-level error scale, in a way that is computationally tractable. In a data set on personal computer (PC) choices in a survey setting (Lenk et al. 1996), we find that respondents who are older have higher average error scale indicating that they make less reliable decisions than those who are younger. Respondents who perceive themselves to be expert when it comes to making PC choices have lower average error scale, indicating that they make more reliable choices. These findings are consistent with recent theorizing on the relationship between cognitive resources and error scale (Swait and Adamowicz 2001). We also facilitate the use of G-MNL in practice by empirically exploring the data requirements for obtaining accurate estimates of the G-MNL and find that estimating this model requires a somewhat larger number of respondents and a larger number of observed choices per respondent than is typical in commercial market research.

1. Introduction

Bayesian methods have allowed marketing researchers to develop complex model specifications, including hierarchical specifications that allow for heterogeneity across individual decision makers in the parameters of the multinomial logit model. Early modeling efforts typically placed convenient and tractable specifications on the distribution of heterogeneity, most commonly a normal or lognormal distribution on the parameters of the multinomial logit model (cf., Allenby and Ginter 1995, Rossi, Allenby and McCulloch 2005). However, recent work has proposed alternative specifications of the population distribution that, it is argued, apply shrinkage to parameters that are economically meaningful, and provide better fit to the types of data sets typically found in marketing (cf., Sonnier, Ainslie and Otter 2007).

This essay contributes specifically to the development of models that allow for heterogeneity across respondents in the error scale of the multinomial logit model. The potential to explain respondent heterogeneity by differences in error scale has been recognized for some time (Louviere 2001), but models that allow for continuous error scale heterogeneity have only recently been developed (Sonnier, Ainslie and Otter 2007, Keane et al. 2009). The most general of these models is the "Generalized Multinomial Logit Model" (G-MNL), which allows for heterogeneity both in error scale and all attribute preferences, including the price attribute. Using a simulated maximum likelihood estimation framework, Keane et al. (2009) show that this model is identified (using synthetic data designed to be similar to typical data sets from choice experiments) and that it provides superior fit (measured by BIC), in a number of empirical applications, over the commonly used specification that places a multivariate normal distribution on the coefficients of the multinomial logit model (MVN-MNL), implicitly assuming error scale homogeneity. In this essay, we further develop the G-MNL by proposing a Bayesian estimation strategy, allowing for straightforward incorporation of covariates to individual-level error scale, such as demographics.

We use our proposed approach to explore the relationship between decision maker characteristics and error scale. In particular, we investigate the relationship between error scale and the decision maker's

age and expertise. In a data set on personal computer (PC) choices in a survey setting (Lenk et al. 1996), we find that respondents who are older have higher average error scale indicating that they make less consistent decisions and respondents who perceive themselves to be expert when it comes to making PC choices have lower average error scale, indicating that they make more consistent choices.

1.1 Related literature

Since the importance of considering error scale in multinomial logit models was first pointed out (Swait and Louviere 1993), a number of different modeling approaches have been proposed for investigating error scale differences. Covariance heterogeneity models and heteroscedastic multinomial logit models allow researchers to explore the *aggregate* effect of manipulations on error scale and have been used to explore how the design of choice experiments affects error scale (Swait and Adamowicz 2001a, 2001b, Hensher Louviere and Swait 1999, DeShazo and Fermo 2002, Dellaert Brazell and Louviere 1999) and to measure the aggregate effects of context or framing manipulations on error scale (Salisbury and Feinberg 2009).

To understand heterogeneity in error scale across *individual respondents*, a variety of methods have been used. A brute-force approach is to collect more data, and more informative data, from each respondent as in Louviere et al. (2008b), allowing individual-level, fixed-effects models to be estimated for each respondent. When individual-level fixed-effects are identified by the data, error scale differences can be explored using the approach proposed by Swait and Louviere (1993), treating each person as a "data set". When sufficient data to estimate individual-level models is not available, researchers have proposed using latent class models that allow for differences in error scale, but not other parameters, across discrete groups (Magidson and Vermunt 2007, Kanetkar, Islam and Louviere 2005). However, the latent class framework is limiting in that it restricts the distribution of error scale across the population to be multinomial with a relatively small number of support points and does not readily allow the incorporation of observed decision-maker characteristics of as covariates to error scale.

A few researchers have proposed hierarchical random coefficients models that allow for error scale to vary continuously across the population. In this paper we adopt a slight variation of G-MNL model proposed by Keane et al. (2009), which nests the model proposed by Sonnier, Ainslie and Otter (2007). We develop a Bayesian approach to estimating the G-MNL model, which, unlike other estimation approaches, allows us to easily investigate covariates to individual error scale differences. We use this model to explore the relationship between an individual's error scale and his or her expertise with the purchase category and find that error scale is negatively correlated with expertise and positively correlated with age. While preliminary, these findings suggest that experts make more consistent choices while older respondents make less consistent choices. In the conclusions, we discuss implications of these findings for behavioral research on choice consistency.

In addition, we facilitate the use of G-MNL in practice by empirically exploring the data requirements for obtaining accurate estimates of the G-MNL and find that estimating this model requires a larger number of respondents and a larger number of observed choices per respondent than is typical in commercial market research. Even so, collecting appropriate data to estimate the model seems to be feasible. We also explore the ability of different Bayesian model fit statistics, in particular log marginal likelihood and deviance information criteria (DIC), to identify when the true model used to generate the data is G-MNL versus the traditional MVN-MNL specification. We find that whether the researcher's focus is on the individual- or population-level likelihood (Trevisani and Gelfand 2003) is important when identifying the correct population-level model.

2. A model for heterogeneity in error scale

Under the random utility interpretation of the multinomial logit model, consumers are assumed to choose the product that offers them the greatest utility, where utility is an unobserved random variable,

$$u_j = x_j'\beta + \varepsilon_j, \qquad \textbf{(1)}$$

where x_j is a vector of K attributes for alternative j, β is an unknown K-vector of parameters, and ε_j is an IID error term distributed according to the double exponential distribution with scale parameter λ. This results in the multinomial logit likelihood

$$\left[y_{it} = j^* \mid \beta, \lambda, \{x_j\}\right] = \frac{\exp\left(x'_{j^*}\left(\beta / \lambda\right)\right)}{\sum_j \exp\left(x'_j\left(\beta / \lambda\right)\right)} \qquad \textbf{(2)}$$

It is well-understood that the parameters of this model, the vector β and the scalar λ, are not separately identified (cf., Ben-Akiva and Lerman 1985, Louviere, Hensher and Swait 2000), and λ is typically normalized to 1, resulting in the familiar multinomial logit likelihood with parameters β.

When the multinomial logit model is used as the unit-level likelihood in a hierarchical model specification, it is standard practice to maintain the assumption that λ is 1 across all consumers and to specify that β_i follows a multivariate normal distribution (cf., Rossi, Allenby and McCulloch 2005) across consumers (indexed by *i*). However, equation (1) suggests that there may also be heterogeneity across consumers in the error scale parameter, λ_i (Louviere, et al. 2008b, Keane et al. 2009). For a given vector of preferences, β_i, if the scalar λ_i is small for a particular consumer then *all* elements of the vector β_i / λ_i will be larger and the model in equation (2) will predict that 1) the consumer will make more consistent choices when repeatedly faced with the same set of alternatives (i.e., the model will predict more extreme purchase likelihoods for a given set of alternatives), and 2) the consumer will react more strongly than others with the same β_i to changes in *any* of the attributes. As we will discuss in more detail, these differences in predicted choices for different levels of λ_i can serve to identify the error scale of one consumer relative to another, even though the absolute level of error scale is unidentified. Thus it seems reasonable to explore specifications of the population distribution for the multinomial logit parameters that allow for heterogeneity in λ_i as well as β_i. (Note that heterogeneity in λ_i is better

identified the greater the dimension of β_i; in fact heterogeneity in β_i can not be distinguished from heterogeneity in λ_i when the dimension of β_i is 1.)

We should point out that differences across individuals in error scale are not merely a phenomenon of theoretical interest; differences in error scale lead to fundamentally different predictions about what consumers will choose, given a new set of alternatives. Salisbury and Feinberg (2008) show that when error scale is larger, choice probabilities for less desirable options increase while choice probabilities for more desirable options decrease, and that an increase in error scale can lead to respondents choosing a more diverse range of options, even as relative preferences for the alternatives remain constant. Similarly, sequences of choices from individuals with high error scale will appear more varied or "diversified" than choices from individual with low error scale, even when those two individuals have the same preferences for the alternatives. Estimates of economically meaningful quantities, like price elasticity and willingness-to-pay, may also be different, depending on whether heterogeneity in error scale is accommodated in the model (Sonnier, Ainslie and Otter 2007).

There are a variety of ways one might specify a joint distribution for β_i and λ_i; one might consider any distribution with positive support for λ_i. For computational simplicity, we specify the population distributions for β_i and λ_i as multivariate normal and log-normal respectively, specifically,

$$\begin{aligned}\left[\beta_i \mid \Delta, \Sigma\right] &= N_K\left(\beta_i \mid \beta_0 + z_i'\Delta, \Sigma\right) \\ \left[\log(\lambda_i) \mid \delta, \sigma\right] &= N\left(\log(\lambda_i) \mid z_i'\delta, \sigma^2\right)\end{aligned} \qquad \textbf{(3)}$$

where z_i is a vector of variables describing consumer *i*, which has been mean-centered. By mean centering z_i, the mean of $\log(\lambda_i)$ is fixed at zero and the median of λ_i is fixed at 1. This constraint is required for identification; without it, there would be multiple pairs of distributions for β_i and λ_i that would result in the same implied distribution on β_i / λ_i and therefore the same likelihood. Under the restriction, the estimated parameter, λ_i, can be interpreted as a measure of consumer *i*'s logit error

relative to the median. (An alternative identification constraint, which we do not explore here, is to fix the error scale for one consumer to 1 and assume that the remaining λ_i follow some population distribution. This would result in a model form similar to those that have been proposed for combining different sources of choice data, where each consumer represents a unique "data source" (Louviere, et al. 2008b). The joint distribution proposed in equation (3) does not allow for correlation between β_i and λ_i, as allowing for correlations would lead to a similar identification problem in practice . The proposed model nests within it the usual specification of the hierarchical multinomial logit (MVN-MNL) model (i.e., $\lambda_i = 1$ for all *i*) when $\delta = \sigma^2 = 0$ (cf., Rossi, Allenby and McCulloch 2005). When, additionally, $\Delta = 0$, the mixed logit model (N-MNL) is obtained (cf., Train 2003).

We will refer to the model proposed in equations (2) and (3) as the generalized hierarchical multinomial logit model (G-MNL). It is similar to the type II generalized multinomial logit model proposed by Keane et al. 2009; however, our formulation and Bayesian estimation approach allows for the inclusion of individual characteristic variables (e.g., age, gender, category experience) as covariates to the individual-level error scale parameters, allowing us to explore potential drivers of individual differences in choice error scale. Keane et al. (2009) discuss the possibility of including such covariates in the formulation, but their simulated maximum likelihood estimation approach limits the feasibility of estimating models with these covariates and they do not present any model estimates with covariates. The other minor difference is in how they choose to fix the location of the error scale distribution; they propose to fix the mean of the lognormal distribution for λ_i at 1, rather than the median as in equation (3).

2.1 Implied distribution of β_i / λ_i.

While the standard MVN-MNL model (i.e., $\lambda_i = 1 \ \forall i$) results in a distribution of β_i / λ_i, that is normally distributed, the G-MNL model implies that β_i / λ_i is the ratio of a multivariate normal and a

univariate normal. This ratio, β_i / λ_i, has a specific pattern of correlation between the elements even when β_i has a diagonal covariance matrix. Specifically,

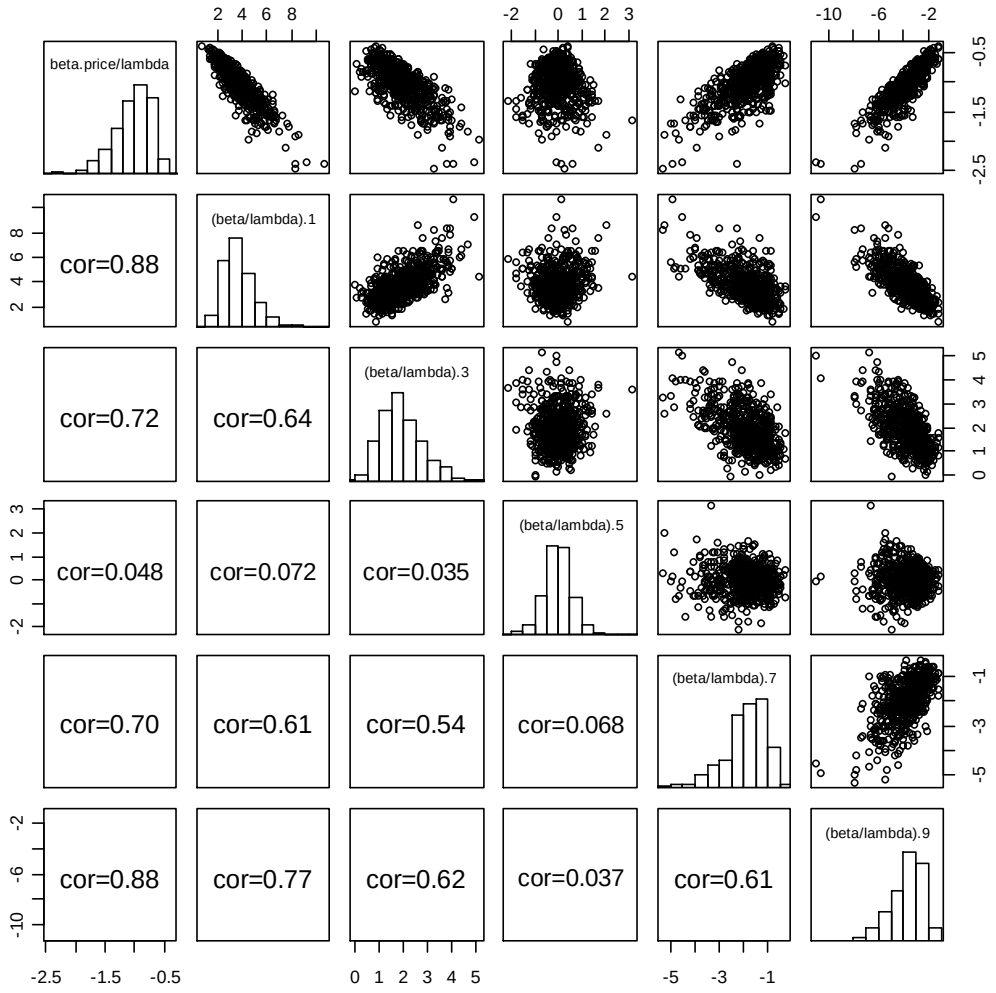
$$\begin{aligned} \text{var}\left(\frac{\beta_i}{\lambda_i}\right) &= E\left(\text{var}\left(\frac{\beta_i}{\lambda_i}\middle|\lambda_i\right)\right) + \text{var}\left(E\left(\frac{\beta_i}{\lambda_i}\middle|\lambda_i\right)\right) \\ &= E\left(\frac{1}{\lambda_i^{\,2}}\right)\text{var}(\beta_i) + \text{var}\left(\frac{1}{\lambda_i}\right)E(\beta_i)E(\beta_i)' \end{aligned} \qquad \textbf{(4)}$$

The second term in the last line of equation (4) corresponds to a common correlation across the elements of β_i / λ_i that is induced solely by heterogeneity in λ_i. This correlation is proportional to $E(\beta_i)E(\beta_i)'$. If two elements of β_i both have large, positive expectations, then the corresponding elements of β_i / λ_i will be have a large, positive correlation. If one element of β_i has a large, positive expectation and another a large, negative expectation, then the corresponding elements of β_i / λ_i will have a large, negative correlation. Intuitively, this structured nature of the covariance of β_i / λ_i that is induced by heterogeneity in λ_i helps to identify heterogeneity in λ_i based only on observed choices.

Figure 1 shows an example set of synthetic values for β_i / λ_i generated according to the population distribution in equation (3) with $\Sigma = diag(0.3)$ and $\sigma^2 = 0.1$. When Σ is restricted to be diagonal, i.e., there are no correlations between elements of the attribute preference vector, we will refer to the model as the diagonal G-MNL. For Figure 1, the mean of β_i was set at (-3.65, -2.74, -1.83, -0.91, 0, 0.91, 1.83, 2.74, 3.65). The resulting distribution for λ_i has a mean of 1.051 and a variance of 0.116. The scatterplots in Figure 1 show 600 draws of β_i / λ_i. Even with this modest amount of variation in error scaling across the population, the resulting distribution for β_i / λ_i shows a distinct pattern: elements

of β_i / λ_i with means further from zero have skewed distribution and are more highly correlated with other elements of β_i / λ_i, even though the elements of β_i are uncorrelated.

Figure 1. The distribution of β_i / λ_i for the diagonal G-MNL model shows strong correlations and skewness even when elements of β_i are uncorrelated.



It is an empirical question as to whether the distribution of β_i / λ_i implied by equation (3) can be well approximated by the standard MVN-MNL model, which imposes a multivariate normal distribution on β_i / λ_i. Certainly, the MVN-MNL can capture the correlations described in Figure 1, if not the skewness and, as we will show, may represent a reasonable prior on individual-level parameters. But, it is clear from Figure 1, that if λ_i is heterogeneous across the population, the distribution of β_i / λ_i would

not be well modeled with the diagonal G-MNL specification, i.e., a model where $\lambda_i = 1 \ \forall i$ and Σ is restricted to be diagonal.

In the next section, we will investigate the sample size required to distinguish data generated according G-MNL from data generated according to MVN-MNL or N-MNL. To the extent that β_i / λ_i can be identified from observed choices for individual *i*, we expect to be able to empirically identify which specification of the population distribution fits best to a particular data set. These three population distributions (G-MNL, MVN-MNL and N-MNL) will also lead to different patterns of shrinkage, and we would expect to get the best individual-level parameter recovery when the higher-level model used in estimation corresponds to the one used to generate the individual-level coefficients and error scale values.

2.2 Related Models

Type I and type II generalized multinomial logit model. The model proposed in equations (2) and (3) is closely related to the type II generalized multinomial logit model proposed by Keane et al. (2009). They also propose an alternative G-MNL model (type I) that allows for an error scaling term to multiply the population means, but not the unexplained heterogeneity. Their estimation strategy allows for continuous mixing between type I and type II scaling and, empirically, they find support for type II scaling in most of the choice-based conjoint data sets they investigate. They also show that a G-MNL model will fit better to data generated according to G-MNL with sample sizes similar to a typical choice-based conjoint study. In ten choice-based conjoint data sets, they find empirical support for models that allow for heterogeneity in error scale (G-MNL and a model where *only* λ_i is heterogeneous, which we will refer to as S-MNL), versus models that do not allow for heterogeneity in error scale (N-MNL and MVN-MNL). Their analysis suggests that comprehending heterogeneity in error scale is particularly important in data sets that involve more complex choice objects (i.e., objects with more complex attributes).

Surplus or willingness-to-pay multinomial logit model. The G-MNL model nests within it the surplus or willingness-to-pay (WTP-MNL) model proposed by Sonnier, Ainslie and Otter (2007) when log(price) is included as an attribute with a coefficient restricted to $-1/\lambda_i$ for all consumers. This results in a model where the willingness-to-pay for any attribute (i.e., the ratio of the coefficient for an attribute relative to the coefficient for price) is normally distributed. The implied distribution for β_i / λ_i is quite similar to that for the G-MNL, except that the WTP-MNL model results in large correlations between the coefficient for price and the other coefficients, induced by heterogeneity in λ_i, even when Σ is diagonal. Applying the model to data from choice experiments on midsize sedans and cameras, they find that the WTP-MNL model fits the data sets better (as measured by the posterior predictive likelihood of holdout tasks) and that the resulting estimated distribution of willingness-to-pay has greater face validity, relative to the MVN-MNL model. This suggests that the G-MNL, which nests the willingness-to-pay MNL model, would also provide better fit to this data than the standard MVN-MNL specification.

2.3 Estimation

Our approach to estimation is Bayesian with conditionally-conjugate, diffuse, proper priors for β_0, Δ, Σ, δ, and σ^2, which allows us to use the usual Metropolis-within-Gibbs sampler for the hierarchical multinomial logit model (cf., Rossi, Allenby and McCulloch 2005) with only minor modifications to accommodate the additional error scale parameter. The parameters are drawn in four blocks: 1) β_0, Δ, and, δ are drawn from their joint full-conditional distribution, which is multivariate normal, 2) Σ and σ^2 are drawn from their joint full-conditional, which is Inverted-Wishart, 3) β_i are drawn individually for each *i* using a Metropolis-Hastings step, 4) λ_i are drawn individually for each *i* using a similar Metropolis-Hastings step. There are two factors in the full-conditional likelihood for λ_i and β_i, the multinomial logit likelihood in equation (2) and the joint multivariate normal distribution for $\left(\log(\lambda_i), \beta_i\right)$. Full details of the sampling algorithm are included in Appendix A.

3. Identification of the model

3.1 Data requirements for estimation of G-MNL

Parameter recovery in hierarchical model specifications is a complicated function of the structure of the data and the prior, so to shed light on what type of data is necessary to estimate the model with reasonable precision, we estimated the G-MNL model using a number of synthetic data sets with systematically varying structure. As a baseline, we estimated the G-MNL model using data generated according to the G-MNL model, with 600 respondents, 50 choice tasks per respondent, 3 alternatives per choice task (with no 'none' option) and 9 attributes. Attribute data was generated independently for each attribute according to a standard normal distribution[1]. Individual-level parameters β_i were generated as in Figure 1.[2]

Although hierarchical specifications like G-MNL do not require that individual-level fixed-effects parameters are identified, recovery of individual-level and population-level parameters, particularly the parameters that involve second and higher-order moments, requires substantial information in the data about each of the individuals. In this context, information for each individual is increased by increasing the number of choices for each individual (cf., Louviere, et al. 2008b). Table 1 shows that as we vary the number of choices observed for each respondent from 20 to 100 (holding other characteristics of the data at the base level), we find increasingly tighter posteriors for both individual and population-level parameters (as measured by the average posterior standard error) and posteriors that are more consistent with the values used to generate the data (as measured by the root mean squared error between the true values and the posterior modes.)

[1] Data from a designed experiment where attribute data is manipulated to maximize information would likely be more informative than our synthetic data.

[2] Keane et al. (2009) also use synthetic data to show that the G-MNL model is formally identified by the likelihood for two data sets similar in structure to a typical choice experiment and that reasonable recovery of true parameters in synthetic data is possible, however their study is limited to two synthetic data sets: one with 79 respondents, 32 choices per respondent, 2 alternatives per choice task, and 6 attributes; and one with 331 respondents, 16 choice tasks per respondent, 2 alternatives per choice task, and 8 attributes.

We also found a general pattern that the posterior modes of the population parameters are biased outwards; that is, elements of β_0 and $\text{diag}(\Sigma)$ have posterior modes that are greater in absolute value than the values used to generate the data. We summarize this "outward bias" as $\left(\text{mode}([\beta_{0,i} \mid \text{data}]) - \beta_{0,i}^{true}\right)\text{sgn}(\beta_{0,i}^{true})$ in Table 1. We find that when there are few choices observed for each respondent or when there are a large number of parameters, there is substantial posterior support for extremely high absolute values of β_i / λ_i among a small number of individuals, *i*, whose choices are perfectly predicted by the model. This leads to high estimates of the level of heterogeneity in error scale, i.e., outward bias in σ^2, which can be compensated for by outward bias in β_0 and $\text{diag}(\Sigma)$. This problem with outward bias in finite samples is not unique to the G-MNL model. A similar study with the MVN-MNL model (reported in Appendix B) shows that MVN-MNL estimates are also subject to outward bias when the number of observed choices is small, although the bias is less than for the G-MNL model.

When we decreased the number of choice tasks to 10 (keeping other characteristics of the data at the base levels), we found that the posterior of the individual-level error scale became so diffuse that the MCMC sampler did not traverse the space well and we were unable to obtain parameter estimates. (Note this did not happen with the MVN-MNL model estimates reported in Appendix B.) Consequently, caution should be used when estimating the G-MNL model with low number of choice tasks per individual relative to the number of parameters in β_i.

Similarly, decreasing the number of attributes from 9 to 3 (holding other characteristics of the data at the base level) decreases the number of parameters, thereby increasing the information available for each individual about the parameters. The reverse is true when the number of attributes is increased and when we attempted to estimate the model with 21 attributes we again had difficulty traversing the highly diffuse posterior.

Increasing the number of alternatives from 3 to 10 also modestly improves inference. Intuitively, when more alternatives are included in the choice task, the utility of the chosen alternative is more clearly bounded and the individual-level parameters are better identified.

To explore how the amount of information available at the population-level improves inference, we also varied the number of individuals observed. We find that inference about population-level parameters is substantially improved as the number of respondents is increased from 200 to 1000, with RMSE and outward bias both notably reduced. However, inference about individual-level parameters does not improve much between 200 and 600 and seems to level off between 600 and 1000.

Table 1. Recovery of G-MNL parameters improves as the information available for each individual increases and as the total sample size increases.

		Number of Choice Tasks				Number of Alternatives		Number of Respondents			Number of Attributes		
		10**	20	**50***	100	**3***	10	200	**600***	1000	3	**9***	21**
RMSE	β_0	-	1.11	0.31	0.19	0.31	0.13	0.81	0.31	0.06	0.06	0.31	-
	diag(Σ)	-	0.58	0.24	0.13	0.24	0.09	0.48	0.24	0.09	0.07	0.24	-
	σ^2	-	0.23	0.05	0.02	0.05	0.03	0.14	0.05	0.02	0.01	0.05	-
	β_i	-	0.93	0.46	0.39	0.46	0.33	0.95	0.46	0.46	0.79	0.46	-
	λ_i	-	0.36	0.28	0.24	0.28	0.21	0.33	0.28	0.25	0.26	0.28	-
Average Posterior SD	β_0	-	0.14	0.05	0.04	0.05	0.04	0.12	0.05	0.03	0.07	0.05	-
	diag(Σ)	-	0.16	0.06	0.04	0.06	0.04	0.14	0.06	0.04	0.13	0.06	-
	σ^2	-	0.05	0.02	0.01	0.02	0.01	0.04	0.02	0.01	0.01	0.02	-
	β_i	-	0.76	0.44	0.36	0.44	0.36	0.56	0.44	0.40	0.51	0.44	-
	λ_i	-	0.49	0.29	0.22	0.29	0.22	0.35	0.29	0.24	0.25	0.29	-
Average Outward Bias	β_0	-	0.94	0.26	0.17	0.26	0.11	0.69	0.26	0.05	0.04	0.26	-
	diag(Σ)	-	0.56	0.22	0.12	0.22	0.09	0.45	0.22	0.08	-0.05	0.22	-
	σ^2	-	0.23	0.05	0.02	0.05	0.03	0.14	0.05	-0.02	0.01	0.05	-

* Base level. Other columns represent parameter recovery when one feature of the data is changed and all others are held at base levels.

** We were unable to obtain parameter estimates with a diffuse prior.

Overall, the results presented in Table 1 suggest that recovery of the parameters of the G-MNL model is possible in data sets similar to those produced in commercial choice experiments, although it is preferable to use data sets with somewhat more respondents and more choice tasks than is typical. However, the estimates reported in Table 1 correspond to the case where there is no model misspecification, i.e., the data was generated according to the G-MNL model and the G-MNL model was

used in estimation. In the next section, we investigate the issue of misspecification of the population distribution in hierarchical MNL models.

3.2 Empirically identifying the specification of the population distribution

To determine whether or not it is possible to identify which model specification is most appropriate for a given data set using model fit statistics, we generated data according a particular specification of the population-level model (N-MNL, S-MNL or diagonal G-MNL) and then estimated alternative specifications using the synthetic data. Based on the results of the previous section, we generated data sets that consisted of 600 individuals completing 50 choice tasks out of three alternatives with 9 continuous attributes, as this quantity of data seems sufficient to get reasonable recovery of G-MNL parameters and is similar in size to commercial choice experiments (albeit on the large side).

In computing model fit statistics for hierarchical models, it is helpful to make a distinction between two sets of parameters: the parameters of the population distribution and the individual-level parameters. When we compare two hierarchical models, it is important to consider whether the researcher's focus is on the population-level parameters or the individual-level parameters and model comparison statistics, including marginal likelihoods and deviance information criteria (DIC) will differ depending on which parameters are in focus (Trevisani and Gelfand 2003, Spiegelhalter et al. 2002). This distinction is an important one to make, particularly when comparing findings across studies that use a classical estimation framework (i.e., maximum simulated likelihood) and those that take a Bayesian perspective and employ MCMC methods. Although, as we will describe, it is possible to compute model comparisons with either a population or individual-level focus under both estimation paradigms, the computational methods used to estimate parameters make it more computationally convenient for Bayesian researchers to take an individual-level focus and classical researchers to take a population-level focus.

Despite what may be computationally convenient, if the managerial goal is to make accurate predictions for the individuals in the estimation sample, as is often the case in direct marketing contexts,

then the researcher's focus should be on the individual-level parameters, $\{(\beta_i, \lambda_i)\}$. In this context, model fits should be computed with respect to the individual-level likelihoods, conditional on the posterior of the individual-level parameters. For instance, in the case of the hierarchical MNL model, the individual-level likelihood is

$$L^I = \prod_i \left(\prod_t \left[y_{it} \mid \beta_i, \lambda_i, \{x_{ijt}\} \right] \right) \qquad \textbf{(5)}$$

When we use L^I as the likelihood when computing model comparisons, the population-level distributions can be interpreted as a complex, adaptive prior on the focal parameters, $\{(\beta_i, \lambda_i)\}$. The "model" is then simply L^I. MCMC samplers for the hierarchical MNL compute L^I on each pass of the sampler, thus it is computationally convenient for those who use a Bayesian estimation approach to compute model-fit statistics with an individual-level parameter focus (cf., Rossi, Allenby and McCulloch 2005). However it is possible to estimate individual-level parameters using classical methods (Train 2003) and individual-level fit statistics can be computed using these individual-level estimates and the likelihood in equation (5).

While individual-level focus may be appropriate in applications where inference about the individuals in the sample is important, researchers often intend to make inference *beyond* the individuals in the sample. For instance, those who use choice models in product design typically view the individuals used in estimation as a sample of a larger population and will often use the population-level parameters to make predictions about total market share (cf., Michalek 2005). It is also common in academic research to interpret the population parameters in order to make statements about the general nature of consumer choice, for instance, whether or not heterogeneity across decision makers can be explained more parsimoniously by differences in error scale (Keane et al. 2009). If the modeling goal is to interpret the population-level parameters or make predictions about individuals outside the sample, then it is appropriate to take a population-level focus. In the case of the G- MNL model, the likelihood used to compute model fits with a population-level focus is

$$L^P = \prod_i \int_{\beta_i} \int_{\lambda_i} \left(\prod_t \left[y_{it} \mid \beta_i, \lambda_i, \{x_{ijt}\} \right] \right) \left[\beta_i \mid \beta_0, \Delta, \Sigma, z_i \right] \left[\lambda_i \mid \delta, \sigma^2, z_i \right] d\lambda_i d\beta_i \quad \textbf{(6)}$$

In this context, both the population-level and individual-level likelihoods are both considered components of the model and the prior is simply the prior on the population-level parameters. Maximum simulated likelihood algorithms maximize an estimate of L^P, so it is computationally convenient for those who use a classical estimation framework to report model fit statistics with a population-level focus (cf., Train 2005). Note that this calculation is equivalent to computing the likelihood of the observed choices in the data set using the posterior predictive distribution of $\{(\beta_i, \lambda_i)\}$ for *new* individuals not observed in the sample and so is appropriate when considering how well the population-level model estimated from a sample characterizes the population.

To investigate the ability of individual-level and population-level fit statistics to detect the correct specification of the population distribution, we generated data according to three different models: N-MNL, S-MNL and diagonal G-MNL. The population-level parameters were the same as those used in the previous section. We then estimate four alternative models using this data: S-MNL, N-MNL, MVN-MNL, and G-MNL[3]. Both the deviance information criteria (Spiegelhalter et al. 2002) and the log marginal likelihood (cf., Rossi Allenby and McCulloch 2005) were computed using the individual-level likelihood in equation (5). The log marginal likelihood was estimated using the harmonic mean of L^I over the 14,000 draws from the posterior of the individual-level parameters (Newton and Raftery 1994). Average deviance, model complexity, pD, and DIC were also was computed based on L^I for these 14,000 draws. For comparison with Keane et al. (2009) we also report BIC, which was estimated by taking the maximum of L^I over the 14,000 draws (that is, we did not run a procedure to maximize L^I, but presume that the maximum over the MCMC draws is a close approximation to the actual maximum).

[3] Estimation of the MVN-MNL and G-MNL model with the S-MNL data proved to be difficult, so we do not report results. When diffuse priors are used for the population-level parameters these models allow extensive over-fitting of the individual-level choice data.

Table 2 reports the model comparison statistics computed with an individual-level parameter focus, which generally do a poor job at determining the true model. The log-marginal likelihood favors the G-MNL model regardless of what the true model is. The log marginal likelihood for G-MNL and MVN-MNL are also quite close, indicating that these two models are difficult to distinguish using the log marginal density computed with an individual-level focus.[4] This is somewhat unsurprising; when there is sufficient data available for each individual and when the population-level likelihood is interpreted as a component of a complex prior on the individual-level parameters, then both of these models provide sufficient flexibility to fit any data set well.

The S-MNL model, in contrast, does not seem to be able to fit the data well when there is heterogeneity in the preference parameters. This suggests that when the true data generating process is unknown and the goal is individual-level prediction (e.g., CRM database scoring), models that allow for heterogeneity in preferences, such as G-MNL and MVN-MNL, may provide better individual-level fits than models like S-MNL that do not. If sufficient data is available, it may not be critical which population distribution (MVN-MNL versus G-MNL) is used as the individual-level estimates will be more influenced by the individual-level likelihood than by the specification of the population distribution which forms a prior on the individual-level parameters.

[4] Note that each cell in Table 2 represents a single set of data and a single estimation run, and findings may change were the experiment run repeatedly. There is also the potential for inaccuracy in our estimate of the log marginal likelihood and further investigation with more accurate estimates of log marginal likelihood (Gelfand and Dey 1994, Chib and Jeliazkov 2001) is warranted.

Table 2. Model fit statistics computed based on the individual-level likelihood do not distinguish different specifications of the population-level model.

			Model Estimated			
			S-MNL	**N-MNL**	**MVN-MNL**	**G-MNL**
Data generaged according to	**S-MNL**	log marginal likelihood (NR)	-6,183	-6,062	**-5,402**	NA
		deviance	12,260	12,118	10,486	NA
		pD	599	1,668	3,330	
		DIC	**12,859**	13,786	13,816	
		maximum ll	-6,071	-5,944	-5,098	NA
		parameters	609	5,400	5,400	
		BIC	**16,037**	46,431	44,740	
	N-MNL	log marginal likelihood (NR)	-9,361	-5,982	-5,820	**-5,724**
		deviance	18,643	11,505	11,293	11,078
		pD	343	6,007	5,353	5,795
		DIC	18,987	17,512	**16,646**	16,873
		maximum ll	-9,275	-5,580	-5,472	-5,331
		parameters	609	5,400	5,400	6,000
		BIC	**22,446**	45,704	45,488	49,043
	diagonal G-MNL	log marginal likelihood (NR)	-9,343	-6,357	-6,119	**-5,866**
		deviance	18,600	12,365	11,827	11,342
		pD	419	5,164	4,468	6,479
		DIC	19,019	17,529	**16,295**	17,821
		maximum ll	-9,248	-5,993	-5,748	-5,461
		parameters	609	5,400	5,400	6,000
		BIC	**22,393**	46,529	46,039	49,303

Table 2 also shows that when the DIC and BIC statistics are computed with individual-level focus, they are also seldom unable to identify the true model. The BIC statistic, in particular, seems poorly suited when the individual-level parameters are the focus; it strongly favors the more parsimonious S-MNL model regardless of the true data generating process, perhaps because the BIC adjustment for number of parameter is inappropriate for the large numbers of individual-level parameters.

Table 3. Model fit statistics computed based on the population-level likelihood more clearly distinguish different specifications of the population-level model.

			Model Estimated			
			S-MNL	**N-MNL**	**MVN-MNL**	**G-MNL**
Data generaged according to	**S-MNL**	log marginal likelihood (NR)	**-6,360**	-6,529	-7,278	NA
		deviance	12,697	13,041	14,338	
		pD	10	17	120	NA
		DIC	**12,707**	13,059	14,458	
		maximum ll	-6,342	-6,500	-7,042	
		parameters	11	18	45	NA
		BIC	**12,754**	13,115	14,372	
	N-MNL	log marginal likelihood (NR)	-9,479	**-8,657**	-8,680	-8,735
		deviance	18,937	17,093	17,166	17,254
		pD	12	70	47	166
		DIC	18,949	**17,164**	17,212	17,420
		maximum ll	-9,460	-8,464	-8,446	-8,531
		parameters	11	18	45	47
		BIC	18,989	**17,044**	17,181	17,363
	diagonal G-MNL	log marginal likelihood (NR)	-9,484	**-8,759**	-8,872	-8,886
		deviance	18,951	17,368	17,514	17,535
		pD	12	-23	76	101
		DIC	18,963	**17,345**	17,590	17,637
		maximum ll	-9,466	-8,622	-8,648	-8,664
		parameters	11	18	45	47
		BIC	19,003	**17,360**	17,583	17,628

We also computed the log marginal likelihood, DIC and BIC using the population-level likelihood in equation (6). The integral in equation (6) was estimated using 100 draws from $\left[\beta_i \mid \beta_0, \Delta, \Sigma, z_i\right]$ and $\left[\lambda_i \mid \delta, \sigma^2, z_i\right]$ for each respondent and this calculation was repeated for 500 draws from the posterior of the population-level parameters taken from the MCMC sampler. (This takes a similar amount of computational time as running 50,000 iterations of the MCMC sampler.) The log marginal likelihood was estimated using the harmonic mean of L^P over the 500 draws.[5] Average

[5] For the individual-level focus, the log marginal density was estimated as the harmonic mean of 14,000 draws from the posterior of $\{(\beta_i, \lambda_i)\}$ and so may be less noisy than the population-level estimates of log marginal density which were based only on 500 posterior draws.

deviance, pD, and DIC were also computed based on these 500 draws and BIC was estimated based on the maximum of L^P over the 500 draws. Table 3 reports the model comparison statistics computed with a population-level focus.

Unsurprisingly, when the model fit statistics are computed at the population-level, there is a clearer distinction between the fit of MVN-MNL and G-MNL. When the true data generating process is N-MNL or S-MNL, DIC and BIC both agree with the log marginal likelihood and correctly identify the true model. However, we did find that when the true data generating process was diagonal G-MNL, the population-level statistics incorrectly identified the N-MNL model as the best model. In this case, the N-MNL, MNV-MNL and G-MNL models all have very similar log marginal likelihoods and it is possible that if the experiment were repeated, the model identified as best might change. The population-level models that allow for heterogeneity in β_i (N-MNL, MVN-MNL and G-MNL) are all quite similar in fit and it seems that a fair amount of data is required to distinguish them, even when a population-level focus is used.

Because of the availability of software to estimate MVN-MNL, it is possible that researchers are estimating MVN-MNL when the only source of heterogeneity in the data is error scale heterogeneity (Keane et al. 2009). When MVN-MNL is estimated with S-MNL data, the estimated covariances for β_i show a distinct pattern that is consistent with equation (4) and Figure 1, specifically, the estimated correlations between elements of β_i in the MVN-MNL specification are related to the estimated population means of β_i (Table 4). We suggest that researchers who estimate MVN-MNL models should check estimates of Σ, to see if this pattern of correlations is present and, if it is, a model that accommodates error scale (S-MNL or G-MNL) should be tested.

Table 4. When MVN-MNL is estimated with S-MNL data, estimates for Σ show a distinct pattern of correlations.

2.444	1.612	0.995	0.513	0.013	-0.462	-0.952	-1.500	-2.056
0.793	**1.689**	0.796	0.389	0.011	-0.385	-0.751	-1.167	-1.612
0.662	0.637	**0.924**	0.236	0.005	-0.235	-0.491	-0.771	-1.048
0.433	0.396	0.325	**0.572**	0.002	-0.098	-0.226	-0.379	-0.494
0.013	0.012	0.009	0.005	**0.444**	0.010	-0.024	0.019	-0.004
-0.397	-0.398	-0.329	-0.173	0.019	**0.555**	0.241	0.350	0.471
-0.637	-0.605	-0.534	-0.313	-0.038	0.339	**0.913**	0.722	0.965
-0.768	-0.719	-0.642	-0.401	0.022	0.377	0.605	**1.559**	1.560
-0.822	-0.776	-0.681	-0.409	-0.004	0.395	0.631	0.781	**2.559**

*The mean of β_i was set at (-3.65, -2.74, -1.83, -0.91, 0, 0.91, 1.83, 2.74, 3.65.)

4. Heterogeneity in error scale among choice experiment respondents

In this section we explore, empirically, the extent to which there is evidence for heterogeneity in error scale in observed consumer choices, i.e. evidence for S-MNL and G-MNL models over MVN-MNL models. Our empirical investigations use data collected in choice experiments, where it is possible to collect the larger numbers of choices for each individual required to identify G-MNL. In two choice experiments, one on bathroom scales and a second on personal computers, we find that the G-MNL model is the best fitting population-level model, suggesting that there is heterogeneity in error scale. Additionally, the Bayesian estimation approach we propose readily allows the incorporation of covariates to λ_i and β_i, and so we incorporate several covariates to error scale. In the data set on personal computer choice, we find that error scale is negatively correlated with expertise and positively correlated with age, suggesting that people who feel they are expert PC buyers make more consistent choices in a choice experiment and that those who are older make less consistent choices.

{*Ed. Note: In both applications of the model, I am currently reporting all results including model fits, parameter estimates for all of the models I attempted to estimates. I envision cutting this down to a smaller set of tables designed to illustrate the key points. But for now, I wanted those who read this to be able to see all the results that we might draw from.*}

4.1 Bathroom Scale Choice Experiment

{*Ed. Note: This study may be eliminated entirely. The only thing it contributes is another data set where G-MNL is the best-fitting model and there are many objectionable things I had to do to get the model to estimate.}*

Data and estimation. The bathroom scale data consisted of responses from 184 student subjects, who each completed 50 choice tasks from a set of three different bathroom scale profiles and a "none" option. The bathroom scale profiles had six attributes, which could each take one of 5 discrete levels. The attributes were manipulated according to an experimental design that was fixed across respondents. Effects codes for the attributes are used in estimation. In addition to the choice responses, the data set included each consumer's response to the question, "Have you purchased a [bathroom] scale in the past 2 years?", which we incorporated in the model as a covariate to error scale. The experiment is described in more detail in Michalek (2005). Initial MCMC runs suggested that the posterior with the G-MNL specification including effects codes for all six attributes was too diffuse to properly traverse with the MCMC sampler[6]. So, the least important attribute, "Area", was dropped from the model specification. We also eliminated 32 respondents who selected the cheapest alternative more than half the time or selected the "none" option more than half the time, as these respondents had extremely poorly identified error scale. For these respondents, the error scale, λ_i, is confounded with the price parameter or the "none" parameter. To provide additional shrinkage for the remaining respondents, we used a moderately informative prior on the population variance parameters: $\Sigma \sim \text{IW}(100, I)$ $\sigma^2 \sim \Gamma^{-1}(0.05,100)$. All posterior estimates are based on chains of length 200,000 with a burn-in of 10,000.

Model comparisons. Table 5 shows that the population-level log marginal likelihood and the DIC statistics suggest that the G-MNL is most consistent with the bathroom scale data. In fact, all of the models that allow for heterogeneity in error scale (S-MNL, diagonal G-MNL and G-MNL) have better log-marginal density than the MVN-MNL model, which does not. This strongly suggests that there *is* error scale heterogeneity in this data set. The fact that diagonal G-MNL is favored over the MVN-MNL model is remarkable given that the diagonal G-MNL model only has 65 population-level parameters

[6] Note that the bathroom scale data is likely to be somewhat more informative per observed choice than the data used in the parameter recovery study, because the bathroom scale data followed an experimental design, while the parameter recovery study used randomly generated (but orthogonal) data. However, we are increasing the demands on the data relative to the simulation studies by incorporating the covariate to λ_i and β_i.

versus 252 for the MVN-MNL model. However, we also find that the G-MNL model is also preferred over the S-MNL and diagonal G-MNL models indicating that there is also evidence of heterogeneity in β_i, in addition to λ_i and that accommodating correlations between elements of β_i improves fit.

Table 5. Population-level model fit statistics for bathroom scale data suggest that G-MNL is most consistent with this data.

	S-MNL	diagonal G-MNL	MVN-MNL	G-MNL
log marginal density (N-R)	-9,060	-9,106	-9,807	**-8,871**
pD	9.9	46.7	-396.0	-50.2
average deviance	18,096	17,623	18,859	17,039
DIC	18,106	17,670	18,463	**16,989**
maximum ll (from draws)	-9,038	-8,485	-9,001	-8,212
parameters	23	65	252	254
observed choices	7,600	7,600	7,600	7,600
BIC	18,282	**17,552**	20,254	18,693

Table 6, Table 7, and Table 8 compare the population-level parameters for the four estimated models. Table 6 shows the estimated population-level parameters related to error scale. All of the models that allow for heterogeneity in error scale find support for substantial heterogeneity in error scale. Additionally, we find no relationship between error scale and the covariate "purchased a bathroom scale in the past 2 years". Note that in Table 7 and Table 8 the estimated parameters Δ and Σ for G-MNL appear to be re-scaled relative to the MVN-MNL and diagonal G-MNL models. In light of the outward bias we found with G-MNL in the parameter recovery study, it seems quite possible that the G-MNL estimates have some outward bias. The S-MNL estimates for bathroom scale data also seem to be scaled down relative to the other models, which is consistent with an inward bias that we found when the S-MNL model is estimated to G-MNL or MVN-MNL models (details available from the author upon request). This suggests that caution should be used when comparing the population-level parameter estimates across different specifications of the population distribution. It seems that models that allow for different levels of flexibility in error scale can lead to different scales for the estimate of the population means of the attribute preferences.

Table 6. Comparison of estimates of (δ, σ^2) for the bathroom scale data.

	S-MNL			diagonal G-MNL			MVN-MNL			G-MNL		
	median	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
Purchased in last 2 years	0.05	-0.01	0.14	-0.03	-0.15	0.08				-0.05	-0.20	0.08
Variance	**0.33**	0.26	0.45	**0.36**	0.26	0.52				**0.79**	0.47	1.10

Table 7. Comparison of estimates of Δ for the bathroom scale data.

			S-MNL			diagonal G-MNL			MVN-MNL			G-MNL		
			median	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
Intercept	none		0.01	-0.06	0.07	0.40	-0.13	0.85	**0.74**	0.32	1.16	**1.42**	0.78	2.15
	capacity	250 lbs.	**0.11**	0.06	0.16	**0.29**	0.13	0.48	**0.31**	0.11	0.50	**0.50**	0.17	0.88
		300 lbs.	**0.22**	0.18	0.28	**0.62**	0.44	0.79	**0.67**	0.49	0.85	**1.14**	0.85	1.46
		350 lbs.	**0.10**	0.06	0.16	**0.34**	0.17	0.54	**0.36**	0.15	0.54	**0.63**	0.32	0.97
		400 lbs	**0.08**	0.02	0.14	**0.27**	0.08	0.44	**0.27**	0.06	0.49	**0.43**	0.13	0.74
	aspect ratio	0875	**0.27**	0.23	0.32	**0.69**	0.52	0.84	**0.64**	0.47	0.84	**1.19**	0.79	1.59
		1.00	**0.25**	0.20	0.33	**0.70**	0.52	0.89	**0.70**	0.54	0.86	**1.15**	0.85	1.49
		1.143	**-0.06**	-0.12	0.00	0.08	-0.13	0.26	0.05	-0.14	0.24	0.03	-0.30	0.35
		1.333	**-0.49**	-0.56	-0.43	**-1.18**	-1.46	-0.87	**-1.23**	-1.51	-0.97	**-2.20**	-2.88	-1.67
	gap	0.094"	**-0.15**	-0.20	-0.11	**-0.28**	-0.43	-0.11	**-0.33**	-0.51	-0.14	**-0.51**	-0.80	-0.21
		0.125"	**0.18**	0.13	0.23	**0.64**	0.48	0.82	**0.64**	0.44	0.84	**1.08**	0.81	1.38
		0.156"	**0.17**	0.12	0.23	**0.58**	0.37	0.79	**0.53**	0.32	0.75	**0.90**	0.57	1.21
		0.188"	**0.17**	0.12	0.23	**0.37**	0.11	0.65	**0.44**	0.21	0.66	**0.71**	0.36	1.09
	number size	1.00"	**-0.26**	-0.32	-0.21	**-0.44**	-0.62	-0.27	**-0.52**	-0.76	-0.32	**-0.91**	-1.30	-0.60
		1.25"	**0.28**	0.23	0.36	**0.73**	0.55	0.93	**0.76**	0.59	0.95	**1.39**	1.08	1.74
		1.50"	**0.36**	0.28	0.42	**0.91**	0.75	1.07	**0.93**	0.66	1.13	**1.74**	1.39	2.22
		1.25"	**0.47**	0.41	0.53	**1.02**	0.81	1.22	**1.00**	0.79	1.29	**1.91**	1.44	2.52
	price	$15	**0.43**	0.38	0.49	**0.97**	0.81	1.14	**0.96**	0.75	1.17	**1.61**	1.31	2.00
		$20	**0.10**	0.05	0.15	**0.37**	0.21	0.52	**0.40**	0.28	0.55	**0.71**	0.48	0.97
		$25	**-0.28**	-0.35	-0.21	**-0.54**	-0.73	-0.37	**-0.57**	-0.79	-0.37	**-0.97**	-1.37	-0.64
		$30	**-0.82**	-0.91	-0.74	**-1.87**	-2.16	-1.58	**-1.88**	-2.22	-1.57	**-3.25**	-3.96	-2.75
Purchased in last 2 years	none					-0.17	-0.52	0.20	-0.04	-0.43	0.33	-0.06	-0.77	0.39
	capacity	250 lbs.				0.06	-0.12	0.32	0.09	-0.12	0.30	0.15	-0.18	0.44
		300 lbs.				0.03	-0.11	0.19	0.03	-0.14	0.22	0.03	-0.23	0.29
		350 lbs.				-0.02	-0.24	0.17	-0.03	-0.22	0.16	-0.07	-0.36	0.24
		400 lbs				-0.03	-0.21	0.14	-0.02	-0.23	0.17	-0.01	-0.27	0.28
	aspect ratio	0875				**0.21**	0.08	0.35	**0.26**	0.06	0.44	**0.35**	0.03	0.65
		1.00				0.04	-0.12	0.21	0.05	-0.11	0.21	0.11	-0.14	0.43
		1.143				-0.11	-0.29	0.11	-0.11	-0.29	0.08	-0.20	-0.49	0.12
		1.333				**-0.25**	-0.53	-0.02	**-0.36**	-0.61	-0.06	-0.39	-0.89	0.13
	gap	0.094"				-0.05	-0.20	0.09	-0.08	-0.24	0.11	-0.19	-0.48	0.12
		0.125"				-0.10	-0.25	0.04	-0.07	-0.23	0.08	-0.01	-0.25	0.24
		0.156"				0.17	-0.02	0.39	**0.20**	0.02	0.39	0.27	-0.07	0.56
		0.188"				-0.03	-0.22	0.18	0.05	-0.19	0.29	0.00	-0.32	0.39
	number size	1.00"				0.10	-0.03	0.23	0.13	-0.05	0.31	0.22	-0.06	0.49
		1.25"				-0.04	-0.20	0.11	-0.04	-0.22	0.14	-0.14	-0.43	0.13
		1.50"				-0.14	-0.31	0.03	-0.13	-0.32	0.05	**-0.34**	-0.66	-0.05
		1.25"				-0.09	-0.32	0.12	-0.07	-0.34	0.15	-0.21	-0.59	0.17
	price	$15				**-0.17**	-0.34	-0.01	-0.10	-0.33	0.10	**-0.37**	-0.65	-0.07
		$20				-0.10	-0.25	0.02	-0.10	-0.26	0.04	**-0.22**	-0.44	-0.01
		$25				0.13	-0.05	0.32	0.14	-0.07	0.36	0.23	-0.09	0.51
		$30				**0.34**	0.09	0.54	0.22	-0.11	0.51	**0.72**	0.19	1.14

Table 8. Comparison of estimates of diagonal (∑) for the bathroom scale data.

		S-MNL		diagonal G-MNL			MVN-MNL			G-MNL		
		median	2.5-97.5 %-tile	**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
none				**5.73**	4.35	8.45	**5.08**	4.23	7.40	**13.47**	9.81	20.00
capacity	250 lbs.			**1.05**	0.74	1.51	**1.76**	1.24	2.09	**2.96**	2.23	4.30
	300 lbs.			**0.54**	0.39	0.98	**1.00**	0.79	1.32	**1.80**	1.31	2.77
	350 lbs.			**0.85**	0.59	1.23	**1.21**	0.85	1.50	**2.81**	2.07	4.09
	400 lbs			**0.74**	0.54	1.13	**1.26**	1.03	1.67	**2.51**	1.89	3.68
aspect ratio	0875			**0.53**	0.36	0.78	**1.24**	0.83	1.55	**3.90**	2.76	6.04
	1.00			**0.63**	0.46	0.88	**0.82**	0.68	1.07	**1.92**	1.46	2.86
	1.143			**1.16**	0.79	1.55	**1.30**	1.11	1.75	**3.55**	2.67	5.39
	1.333			**2.16**	1.64	3.01	**2.78**	2.10	3.73	**9.21**	6.71	14.89
gap	0.094"			**0.45**	0.30	0.65	**0.99**	0.80	1.36	**2.42**	1.68	3.29
	0.125"			**0.67**	0.50	1.04	**0.89**	0.68	1.21	**1.75**	1.31	2.52
	0.156"			**0.98**	0.72	1.49	**1.41**	1.00	1.72	**3.11**	2.29	4.22
	0.188"			**1.62**	1.28	2.43	**1.86**	1.36	2.62	**4.24**	3.36	6.49
number size	1.00"			**0.42**	0.30	0.70	**1.05**	0.85	1.86	**2.84**	2.01	4.23
	1.25"			**0.58**	0.40	1.07	**0.94**	0.71	1.10	**2.42**	1.65	3.52
	1.50"			**0.55**	0.33	0.79	**1.21**	0.88	1.42	**3.32**	2.45	5.27
	1.25"			**1.40**	1.10	1.94	**1.94**	1.60	2.67	**6.53**	4.82	10.11
price	$15			**0.44**	0.32	0.61	**1.24**	1.00	1.72	**2.87**	1.95	4.28
	$20			**0.25**	0.18	0.42	**0.50**	0.43	0.64	**1.03**	0.76	1.43
	$25			**0.79**	0.58	1.24	**1.44**	1.10	2.01	**2.99**	2.22	4.62
	$30			**1.97**	1.38	2.76	**2.85**	2.45	3.69	**6.45**	5.13	11.67

Although the population-level comparisons in Table 5 suggest that the G-MNL model is most consistent with this data, estimation software for MVN-MNL is widely available and is regularly used by practitioners. The estimated individual-level parameters are then used to make market share predictions by averaging over individual-level share predictions (cf., Sawtooth Software 2005). Table 9, which reports the model fit statistics for the individual-level parameters, suggests that when this approach is used, it may not be critical which population-level specification is used. Individual-level log-marginal density and DIC are quite close for the MVN-MNL and G-MNL model and favor the MVN-MNL specification, suggesting that when the MVN-MNL model serves as a prior on the individual-level parameters, it provides sufficient flexibility to fit the individual-level parameters well.

Table 9. Individual-level model fit statistics for bathroom scale data suggest that the MVN-MNL serves as an adequate prior on individual-level parameter estimates.

	S-MNL	diagonal G-MNL	MVN-MNL	G-MNL
log marginal density (N-R)	-8,951	-5,289	**-5,038**	-5,072
pD	193.1	3,668.6	2,258.1	2,329.1
average deviance	17,855	10,326	9,854	9,924
DIC	18,048	13,995	**12,112**	12,254
maximum ll (from draws)	-8,895	-5,050	-4,811	-4,835
parameters	173	3,344	3,192	3,344
BIC	**19,337**	39,983	38,144	39,551

4.2 PC Buy/No-Buy Data

Data and estimation. The PC data consisted of 201 subjects, who each made 20 binary choices between a PC profile and "don't buy." The PC profiles each had 14 binary attributes, which were near-orthogonally manipulated according to a design that was fixed across respondents. In addition, we considered four potential individual-level characteristics to include in the model: gender, age, PC ownership and whether the respondent considered him/herself to be an "expert at buying PCs." Similar to the bathroom scale data, initial MCMC runs suggested that the posterior for this data and the G-MNL specification with 14 attributes was too diffuse to properly traverse, so we dropped the three attributes that had insignificant parameter estimates (based on preliminary estimates for a MVN-MNL model). We also used a diffuse prior on the population variance parameters: $\Sigma \sim \text{IW}(K+2, I)$ $\sigma^2 \sim \Gamma^{-1}(0.01,100)$. All posterior estimates are based on chains of length 200,000 with a burn-in of 10,000.

Model comparisons. We estimated S-MNL, MVN-MNL, and G-MNL specifications for the PC buy/no-buy data. Focusing on the population-level parameters, we find that all four models have similar log marginal likelihood, with the G-MNL model favored (Table 10). Notably, the log-marginal likelihood for the S-MNL model is nearly the same as for the MVN-MNL, suggesting that a model that includes heterogeneity in error scale (but not in β_i) produces a model that describes the data nearly as well as one that includes full-covariance heterogeneity in β_i.

Table 10. Population-level log marginal densit for PC buy/no-buy data favors the G-MNL specification.

	S-MNL	MVN-MNL	G-MNL
log marginal density (N-R)	-1,774	-1,772	**-1,762**
average deviance	3,537.6	3,449.5	3,404.3
pD	16.8	60.3	120.4
DIC	3,554	**3,510**	3,525
maximum ll	-1,761	-1,673	-1,637
parameters	16	110	115
BIC	**3,655**	4,258	4,228

Table 11 reports the individual-level model comparison statistics, which favor the MVN-MNL model. Similarly to the bathroom scale data, we find that the MVN-MNL and the G-MNL models both produce individual-level parameter estimates that fit the data quite well, while the S-MNL model, which does not allow as much flexibility in the individual-level parameters, does not fit the individual-level data well. This suggests that if individual-level parameters are the object of inference and are used in prediction then it is reasonable to use a MVN-MNL model.

Table 11. Individual-level model fit statistics for PC buy/no-buy data indicate that the MVN-MNL specification produces the best-fitting individual-level parameters.

	S-MNL	MVN-MNL	G-MNL
log marginal density (N-R)	-1,704	**-926**	-987
average deviance	3,368.9	1,681.1	1,774.9
pD	188.8	1,252.7	2,286.0
DIC	3,558	**2,934**	4,061
maximum ll	-1,649	-743	-771
parameters	212	2,211	2,412
BIC	**5,058**	19,828	21,553

Parameter estimates. In Table 12, we report the relationship between error variance, λ_i, and the respondents' age, gender, current PC ownership and self-reported expertise in buying a computer. In the G-MNL formulation, we find significant relationships between error variance and PC ownership, age and

purchasing expertise. We find that those who claim to have expertise in purchasing PCs have lower error variance. As we will discuss further, this is consistent with the hypothesis that respondents who have greater expertise make more consistent decisions when faced with similar choice tasks. The G-MNL model estimates also indicate that that those who currently own a PC have *higher* error scale, which may be indicative of their being more conflicted about the task in general, e.g., "Why should I buy a PC if I already own one anyway?" or could be due to owners placing more weight on terms left out of the utility specification such as omitted attributes or interactions. We also find that those who are older make less consistent choices, which seems reasonable given that older people have less expertise in the category in general and may devote fewer cognitive resources to answering the survey questions. Estimates for the remaining population-level parameters are included in Appendix C.

Table 12. Estimates of (δ, σ^2) for the PC buy/no buy data indicate that respondents who do not own a PC, who are younger and who are more experienced in the category make more consistent choices.

	S-MNL			MVN-MNL			G-MNL		
	median	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
PC Owner	-0.02	-0.08	0.05				**0.33**	0.23	0.43
Gender	-0.04	-0.29	0.24				-0.32	-0.99	0.27
Age	0.01	-0.01	0.03				**0.04**	0.02	0.07
Expert Buyer	-0.07	-0.14	0.02				**-0.29**	-0.52	-0.06
Variance	**0.34**	0.29	0.44				**0.58**	0.41	0.85

5. Interpretation of individual differences in error scale

In the bathroom scale and PC data we found preliminary evidence of heterogeneity in the error scale in a choice model, consistent with what has been found by Keane et al. 2009. There are many potential contributors to heterogeneity in error scale. In market data on choices, a major potential source of individual differences in error scale is differences in the importance of omitted attributes across respondents; for example, if a subgroup of consumers pays close attention to aesthetic appeal of the alternatives, but aesthetic appeal is not included in the model, then these consumers will have greater

estimated error scale. The G-MNL model will accommodate these differences and may provide better predictive ability if heterogeneity in the importance of omitted attributes exists in the data.

In choice experiments like those reported on here, the researcher controls the presentation of the choice task and there are no systematically varying attributes other than those presented and modeled, so the potential for differences in omitted attributes is reduced. However, even in choice experiments, heterogeneity in error scale may still remain due to misspecification errors; for example, if a significant interaction has been left out of the specification of the deterministic portion of the utility, then consumers who place the greatest weight on the interaction will have higher estimated error scale, *ceteris paribus*, than respondents who don't place high weight on this interaction. Respondents for whom the linear specification of the deterministic portion of the utility is inaccurate may also have greater estimated error scale. Similarly, respondents who are making more inferences about attributes that have been left out of the choice task may have greater estimated error scale. It is important to keep in mind when interpreting estimates of error variance that differences in misspecification across respondents will lead to differences in estimated error scale.

However, it has been suggested that even in the complete absence of specification errors, we would likely still find individual differences in error scale and that the remaining variation in error scale can be interpreted as a characteristic of the decision maker and choice context that might be dubbed "choice consistency" (Deallert, Brazell and Louviere 1999, Louviere 2001). Choice consistency can be defined as the respondent's propensity to make the same decision when faced with the same choice scenario repeatedly. Indeed, error scale increases when consumers make decisions about future consumption versus decisions about immediate consumption (Salisbury and Feinberg 2009) and error scale increases as the complexity of a choice task increases (cf. Louviere, al. 2008b). These observations are difficult to explain entirely by misspecification and suggest that some portion of what we estimate as the error scale in the G-MNL model corresponds to the consistency with which individual consumers answer choice questions. While our modeling approach does not permit us to disentangle choice consistency from other contributors to heterogeneity in error scale, the concept of choice consistency

motivates our interest in the relationship between characteristics of the individual such as age and expertise and error scale.

In particular, we find in the PC data that older respondents have greater error scale. This is consistent with the hypothesis that respondents who have fewer *cognitive resources* to devote to a choice task, for example due to age, will make less consistent choices (de Palma, Myers and Papageorgiou 1994, Swait and Adamowicz 2001). This hypothesis has been substantiated in other studies for instance, fatigue effects have been found to occur in choice experiments (Bradley and Daly 1994), where respondents make less consistent choices during the second half of a choice experiment versus the first. It has also been shown that choice experiments with more taxing designs (e.g., more attributes, more attributes that differ between alternatives) result in greater error scale (Louviere, et al. 2008a, Dellaert Brazell and Louviere 1999). Our finding on the relationship between age and error scale in the PC data contributes to the growing body of evidence that any situation that decreases a respondent's cognitive resources (e.g., distraction, aging) will, all else equal, result in less consistent decisions and greater error scale.

Similarly, one might hypothesize that respondents with high *expertise* making decisions in the target category require fewer cognitive resources to make a decision and will make more consistent decisions than those with less expertise, contributing to lower estimated error scale for respondents with high expertise. Our findings in the PC data are consistent with this hypothesis; respondents with high stated expertise have significantly lower estimated error scale. Although our modeling methods cannot shed light on what differentiates the thought processes of "experts" from non-experts, we would expect that an expert will have developed a rich schema around the product category, including the benefits of various product features and how he values those features.

Note that our findings on the relationship between error scale and expertise are *not* consistent with what one would expect were differences in error scale driven by differences in the extent of misspecification between experts and non-experts. One would expect that experts are more likely to have considered all attributes and potential interactions (e.g., "cell phones with 4G service really should have larger displays") and to the extent that we leave these interactions out of the model, error scale should be

higher for these expert individuals. In contrast, we find empirically that experts have *lower* levels of error scale, suggesting that the relationship between expertise and error scale is mediated through an effect of expertise on choice consistency, rather than the effect of expertise on misspecification (although it is possible that both effects are operative in our data set).

Our preliminary findings on the relationship between error scale and expertise would be complemented by additional experiments designed to confirm and flesh out our preliminary findings. Ideally, these experiments should be designed to have more choices observed for each decision maker, so that individual-level error scale is better identified. This can be achieved either by presenting more tasks to each decision maker or by asking the decision maker to make more choices within each task, e.g., by using dual response choice tasks (Brazell et al. 2006) or by asking respondents to choose most and least preferred alternatives (Louviere, et al. 2008b). Such experiments should also employ simpler choice alternatives so that heterogeneity in misspecification can reduced by estimating interaction terms and non-linear specifications in the utility function. We could then interpret error scale estimates more clearly as "choice consistency". In an experimental setting we can also manipulate the independent variables that we hypothesize may influence choice consistency; for instance, we could manipulate the amount of cognitive resources the subject can devote to the task (e.g., through distraction) or to change their experience in the product category (e.g., by asking them to read neutral product reviews before completing the choice task) to more fully flesh out the causal relationships between choice consistency, cognitive capacity and experience with the product category. Such experiments could also be used to identify other moderators of choice consistency.

Beyond expertise, there are a number of other covariates to error scale that could be included in G-MNL models. For example, people with lower need for cognition (Petty and Cacioppo 1986) might be expected to have greater error scale. It has also been suggested that response latencies are related to error scale (Haaijer, Kamakura and Wedel 2000). Similarly, increasing the complexity of the choice task may increase error scale and practitioners should consider experimental designs that anticipate this effect (Louviere, et al. 2008a). Designs explicitly based on the information matrix for the G-MNL model,

integrating over prior distributions for the relationship between error scale and other covariates, should be considered.

6. Conclusions and future research

This essay contributes to the development of the G-MNL model in a number of ways. We propose a Bayesian estimation procedure for the G-MNL, which readily accommodates characteristics of individual decision makers as covariates to error scale. We then test that procedure using two data sets and find that the G-MNL model does provide better fit to both data sets than the standard MVN-MNL, as measured by the log marginal likelihood focused on the population-level parameters, suggesting that the population-level model in G-MNL is more consistent with the data. This finding suggests that there is heterogeneity in error scale that is not properly accounted for by the structure of the MVN-MNL model. We note, however, that individual-level parameters estimated under MVN-MNL and G-MNL models seem to perform equally well and the MVN-MNL model is likely sufficient for applications where individual-level prediction is the goal and there is sufficient data available for each individual (e.g., CRM applications). We also find little support at the population-level for the S-MNL specification, suggesting that there is heterogeneity in β_i in these data sets. The inflexibility of the S-MNL model at the individual-level also severely limits the ability of that model to fit individual-level parameters well.

We also facilitate the use of G-MNL in practice by empirically exploring the data requirements for obtaining accurate estimates of the G-MNL and find that estimating this model requires a larger number of respondents and a larger number of observed choices per respondent than is typical in commercial market research, but even so, seems to be feasible.

There are a number of outstanding methodological issues related to G-MNL that remain to be addressed. In particular, given the widespread availability of software to estimate the MVN-MNL model, it would be valuable to practitioners to develop a method to detect error scaling effects directly from MVN-MNL model estimates, without estimating the G-MNL or S-MNL models.

Finally, our preliminary experience with the MCMC sampler for G-MNL suggests that introducing the heterogeneous error scale parameter improves mixing. This is consistent with the recent findings in Bayesian estimation that suggest that introducing weakly or unidentified "working parameters" improves mixing (see Gelman et al. 2008 for a review). Further research comparing algorithm performance could lead to substantially improved sampling algorithms for both G-MNL and the traditional MVN-MNL model.

Expertise and error scale. In addition to contributing to the development of the G-MNL model, we also use the model and a Bayesian estimation approach to explore the relationship between an individual's error scale and several covariates. In the PC data, we find that an individual's error scale is positively related to age and negatively correlated with his self-stated expertise at making purchases in the category. Both findings are suggestive: age is negatively related to cognitive resources, so it is not surprising that older respondents would make less consistent choices when faced with the same set of alternatives. Respondents who believe they have greater expertise are likely to have more stable preferences and more confidence in their choices, and so would be expected to make more consistent choices. These findings contribute to the growing body of literature that suggests that some of the variation in error scale across respondents can be interpreted as differences in "choice consistency" (Louviere 2001).

These preliminary findings suggest a new opportunity for the study of the marketing dynamics of consumer expertise in an emerging category. As the category develops, we would expect that experienced buyers, who are likely to have lower error variance, will represent a growing portion of the market. If expertise is related to error scale, then the product attributes will explain more and more of the choice behavior in the market over time, even if the underlying value respondents place on those attributes remains constant. This, in turn, would lead to less "diversification" in market shares as the category develops; the product with the best set of features will gain market share over time relative to products with less desirable features (even if the products remain unchanged). If we ignore the relationship between error scale and expertise when developing choice models for emerging products, we risk making

inaccurate predictions about how the market will develop. Similarly, we might make different predictions based on how the age of the consumer base evolves over time and influences the distribution of error scale.

References

Allenby, G.M. and J.L. Ginter (1995) Using Extremes to Design Products and Market Segments, *Journal of Marketing Research*, 32(4), 392-403.

Ben-Akiva, M. and S.R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press, Cambridge, MA.

Bradley, M. and A. Daly. 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, (21)2, 167-184.

Brazell, J.D., C.G. Diener, E. Karniouchina, W.L.Moore, V. Severin and P.-F. Uldry (2006) The no-choice option and dual response choice designs, *Marketing Letters*, 17, 255-268.

Chib S. and I. Jeliazkov (2001) Marginal Likelihood From the Metropolis–Hastings Output, *Journal of the American Statistical Association*, 96(453), 270-281.

Dellaert, G.C., Brazell, J.D. and J.J. Louviere (1999) The Effect of Attribute Variation on Consumer Choice Consistency, *Marketing Letters*, 10(2), 139-147.

dePalma, A., G.M. Myers and P.I. Papageorgiou (1994) Rational Choice Under an Imperfect Ability to Choose, *American Economic Review*, 84(3), 419-440.

DeShazo, J.R. and G. Fermo (2002) Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice consistency, *Journal of Environmental Economics and Management*, 44(1), 123-143.

Gelfand, A.E. and D.K. Dey (1994) Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of the Royal Statistical Society Series B*, 56(3), 501-514.

Gelman, A., D.A. van Dyk, Z. Huang and W.J. Boscardin (2008) Using Redundant Parameterizations to Fit Hierarchical Models, *Journal of Computational and Graphical Statistics*, 17(1), 95-122.

Haaijer, R. W. Kamakura and M. Wedel (2000) Response Latencies in the Analysis of Conjoint Choice Experiments, *Journal of Marketing Research*, 37(3), 376-382.

Hensher, D. A., J. J. Louviere, and J. Swait (1998) Combining Sources of Preference Data, *Journal of Econometrics*, 89(1-2), 197–221.

Kanetkar, V., T. Islam and J. Louviere (2005) Latent Segments or Scale Variations: A Simple Choice Model to Incorporate Heterogenety, working paper.

Keane, M.P., J. Louviere, N. Wasi and D.G. Fiebig (2009) The Generalized Multinomial Logit Model, *Marketing Science*, forthcoming.

Lenk, P.J., W.S. DeSarbo, P.E. Green and M.R. Young (1996) Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs, *Marketing Science*, 15(2), 173-191.

Louviere, J.J. (2001) What If Consumer Experiments Impact Variances as well as Means? Response Variability as a Behavioral Phenomenon, *Journal of Consumer Research*, 28(3), 506-511.

Louviere, J.J. and T. Eagle (2008) Confound it! That Pesky Little Scale Constant Messes Up Our Convenient Assumptions!, *Proceedings of the 2006 Sawtooth Software Conference*.

Louviere, J.J., D.A. Hensher and J.D. Swait (2000) *Stated Choice Methods*, Cambridge University Press, Cambridge, UK.

Louviere, J.J., Islam, T., Wasi, N., Street, D. & Burgess, L.B. (2008a) Designing Discrete Choice Experiments: Do Optimal Designs Come At A Price?, *Journal of Consumer Research*, 35(2), 360-375.

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

Magidson, J., and Vermunt, J.K. (2007). Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. *October 2007 Sawtooth Software Conference Proceedings.*

Michalek, J.J. (2005) Preference Coordination in Engineering Design Decision-Making, Ph.D. Dissertation, Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA.

Newton, M.A. and A.E. Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society Series B (Methodological), 56(1), pp 3-48.

Petty, R. E., & J.T. Cacioppo (1986) *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

Rossi, P.E., G.M. Allenby, R. McCulloch (2005) *Bayesian Statistics and Marketing*, John Wiley and Sons, Chichester, UK.

Sawtooth Software (2005) The CBC/HB System for Hierarchical Bayes Estimation: Version 4.0 Technical Paper.

Salisbury, L.C. and F. Feinberg (2008) Future preference uncertainty and diversification: The Role of Temporal Stochastic Inflation, *Journal of Consumer Research*, 35(2), 349-359.

Salisbury, L.C. and F. Feinberg (2009) Alleviating the Constant Stochastic Variance Assumption in Marketing Research: Theory, Measurement, and Experimental Test, *Marketing Science*, forthcoming.

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

Swait, J. and W. Adamowicz (2001) The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching, *Journal of Consumer Research*, 28(1), 135-148.

Swait, J. and W. Adamowicz (2001) Choice Environment, Market Complexity and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice, *Organizational Behavior and Human Decision Processes*, 86(2), 141-167.

Swait, J. and 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.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin and A. van der Linde (2002) Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society Series B*, 64(4), 583-639.

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

Trevisani M. and A.E. Gelfand (2003) Inequalities between Expected Marginal Log-Likelihoods, with Implications for Likelihood-Based Model Complexity and Comparison Measures, *The Canadian Journal of Statistics*, 31(3), 239-250.

Appendix A. Details of the MCMC sampling algorithm

Model Likelihood

$$\prod_{it}\left[y_{it}=j \mid \beta_0,\Delta,\Sigma,\delta,\sigma^2,\{x_{ijt}\},\{z_i\}\right]$$
$$=\prod_i \int_{\beta_i}\int_{\lambda_i}\left(\prod_t\left[y_{it}\mid\beta_i,\lambda_i,\{x_{ijt}\}\right]\right)\left[\beta_i\mid\beta_0,\Delta,\Sigma,z_i\right]\left[\lambda_i\mid\delta,\sigma^2,z_i\right]d\lambda_i d\beta_i$$

Data augmentation likelihood

$$\prod_{it}\left[y_{it}=j \mid \beta_0,\Delta,\Sigma,\delta,\sigma^2,\{\beta_i\},\{\lambda_i\},\{x_{ijt}\},\{z_i\}\right]$$
$$=\prod_i\left(\prod_t\left[y_{it}\mid\beta_i,\lambda_i,\{x_{ijt}\}\right]\right)\left[\beta_i\mid\beta_0,\Delta,\Sigma,z_i\right]\left[\lambda_i\mid\delta,\sigma^2,z_i\right]$$
$$=\prod_i\left(\prod_t\frac{\exp\left(x_{ijt}'\left(\beta_i/\lambda_i\right)\right)}{\sum_{\tilde{j}}\exp\left(x_{\tilde{i}jt}'\left(\beta_i/\lambda_i\right)\right)}\right)N_k(\beta_i\mid\beta_0+z_i'\Delta,\Sigma)N(\lambda_i\mid z_i'\delta,\sigma^2)$$

Priors

$$\left[(\beta_0,\Delta)\right]=N_{k(l+1)}\left(vec((\beta_0,\Delta))\mid\mu_\Delta^{pr},\Sigma_\Delta^{pr}\right)$$
$$\left[\Sigma\right]=IW\left(\Sigma\mid\nu_\Sigma^{pr},S_\Sigma^{pr}\right)$$
$$\left[\delta\right]=N_l(\delta\mid\mu_\delta^{pr},\Sigma_\delta^{pr})$$
$$\left[\sigma^2\right]=\Gamma^{-1}(\sigma^2\mid s_\sigma^{pr},S_\sigma^{pr})$$

Posterior

$$\left[\beta_0,\Delta,\Sigma,\delta,\sigma^2,\{\beta_i\},\{\lambda_i\}\mid\{y_{it}\},\{x_{ijt}\},\{z_i\}\right]$$
$$=\left(\prod_i\left(\prod_t\left[y_{it}\mid\beta_i,\lambda_i,\{x_{ijt}\}\right]\right)\left[\beta_i\mid\beta_0,\Delta,\Sigma,z_i\right]\left[\lambda_i\mid\delta,\sigma^2,z_i\right]\right)[\Delta][\Sigma][\delta][\sigma^2]$$

Full Conditionals

1. Draw (β_0,Δ) per the usual full conditional for the multivariate normal model.

$$\begin{aligned}
\left[(\beta_0,\Delta)\mid\Sigma,\{\beta_i\},\{z_i\}\right] &\propto \left(\prod_i\left[\beta_i\mid z_i,\beta_0,\Delta,\Sigma\right]\right)\left[(\beta_0,\Delta)\right] \\
&= N_k\left(vec((\beta_0,\Delta))\mid\mu',\Sigma'\right) \\
\Sigma' &= \left(\left(Z'Z\otimes\Sigma^{-1}\right)+S_\Delta^{pr^{-1}}\right)^{-1} \\
\mu' &= \Sigma'\left(\left(Z'\otimes\Sigma^{-1}\right)vec(\alpha)+\Sigma_\Delta^{pr^{-1}}\mu_\Delta^{pr}\right) \\
Z &= \begin{pmatrix}1 & z_1\\ \vdots & \vdots\\ 1 & z_n\end{pmatrix}
\end{aligned}$$

2. Draw Σ per the usual full conditional for the multivariate normal model.

$$\begin{aligned}
\left[\Sigma\mid\beta_0,\Delta,\{\beta_i\},\{z_i\}\right] &= \left(\prod_i\left[\beta_i\mid z_i,\Delta,\Sigma\right]\right)\left[\Sigma\right] \\
&= IW\left(\Sigma\mid v_\Sigma^{pr}+n, v_\Sigma^{pr}S_\Sigma^{pr}+\left(\beta-Z(\beta_0,\Delta)\right)'\left(\beta-Z(\beta_0,\Delta)\right)\right) \\
\beta &= \begin{pmatrix}\beta_1\\ \vdots\\ \beta_n\end{pmatrix}
\end{aligned}$$

3. Draw σ from

$$\begin{aligned}
\left[\sigma^2\mid\delta,\{\lambda_i\},\{z_i\}\right] &= \left(\prod_i\left[\lambda_i\mid\delta,\sigma^2,z_i\right]\right)\left[\sigma^2\right] \\
&= \Gamma^{-1}\left(\sigma^2\mid n+v_\sigma^{pr},\left(\lambda-Z'\delta\right)'\left(\lambda-Z'\delta\right)+v_\sigma^{pr}S_\sigma^{pr}\right) \\
\lambda &= \begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix} \quad Z = \begin{pmatrix}z_1\\ \vdots\\ z_n\end{pmatrix}
\end{aligned}$$

5. For each *i*, draw $\alpha_i = (\log(\lambda_i), \beta_i)$ from

$$\begin{aligned}
&\left[\alpha_i = (\log(\lambda_i), \beta_i) \mid \beta_0, \Delta, \Sigma, \delta, \sigma^2, \{y_{it}\}, \{x_{ijt}\}, z_i\right] \\
&= \prod_i \left(\prod_t \left[y_{it} = j \mid \beta_i, \lambda_i, \{x_{ijt}\} \right] \right) \left[\beta_i \mid \beta_0, \Delta, \Sigma, z_i\right] \left[\lambda_i \mid \delta, \sigma^2, z_i\right] \\
&= \frac{\exp\left(x'_{ijt} \left(\beta_i / \lambda_i \right) \right)}{\sum_{\tilde{j}} \exp\left(x'_{\tilde{i}jt} \left(\beta_i / \lambda_i \right) \right)} N_k\left(\beta_i \mid \beta_0 + z_i'\Delta, \Sigma\right) N(\lambda_i \mid z_i'\delta, \sigma^2) \\
&= \frac{\exp\left(x'_{ijt} \left(\beta_i / \lambda_i \right) \right)}{\sum_{\tilde{j}} \exp\left(x'_{\tilde{i}jt} \left(\beta_i / \lambda_i \right) \right)} N_{k+1}\left(\alpha_i = (\log(\lambda_i), \beta_i) \mid \begin{pmatrix} z_i'\delta \\ \beta_0 + z_i'\Delta \end{pmatrix}, \begin{matrix} \sigma^2 & 0 \\ 0 & \Sigma \end{matrix} \right)
\end{aligned}$$

which can be done by the usual M-H step for the multinomial logit.

Appendix B. Parameter recovery study for MVN-MNL model.

To explore the potential for outward bias in the MVN-MNL model, we generated data according to the true MVN-MNL model and then estimated the MVN-MNL model using this data. The design of the study mirrors that reported in Table 1. As expected, we find that posterior standard errors decrease as we increase the amount of data or decrease the number of parameters of the model. Recovery of true parameters, as measured by the root mean squared error between the true parameters and the modes of the posterior distributions, also improves as we increase data or increase parameters. *More importantly, we find that as the number of choices per unit is decreased, there is substantial outward bias in the posterior distributions of both the individual and population-level parameters.* When we observe 10 choices for each of 600 units with 3 alternatives per task and 9 attributes (data that is quite typical for commercial choice experiments), we find an average outward bias in β_0 of 0.43. Although this may not affect the predictive performance of the model a great deal (i.e., estimated shares may be reasonably accurate), it does suggest that caution should be used when interpreting MVN-MNL parameters. In comparing parameter recovery results for MVN-MNL (Table 13) versus G-MNL (Table 1), we find that both models are subject to this outward bias, however the flexibility of the G-MNL to seems to make the bias more pronounced (i.e. more choice observations per unit are required to eliminate the bias). Our experience estimating both MVN-MNL and G-MNL with real data sets bore this out.

Table 13. Recovery of the parameters of the MVN-MNL model shows substantial outward bias as the number of observations per respondent is reduced.

		Number of Observations				**Number of Alternatives**		**Number of Units**			**Number of Attributes**		
		10	20	**50**	100	**3**	10	200	**600**	1000	3	**9**	21
RMSE	β_0	0.44	0.18	0.04	0.03	0.04	0.03	0.11	0.04	0.04	0.09	0.04	0.04
	diag(Σ)	0.27	0.13	0.02	0.03	0.02	0.03	0.08	0.02	0.04	0.11	0.02	0.04
	β_1	0.62	0.52	0.41	0.28	0.41	0.40	0.37	0.41	0.37	0.00	0.41	0.33
Average Posterior SD	β_0	0.11	0.06	0.04	0.03	0.04	0.03	0.07	0.04	0.03	0.08	0.04	0.03
	diag(Σ)	0.13	0.08	0.04	0.04	0.04	0.03	0.08	0.04	0.03	0.11	0.04	0.02
	β_1	0.65	0.55	0.37	0.33	0.37	0.32	0.42	0.37	0.43	0.00	0.37	0.30
Average Outward Bias	β_0	0.43	0.17	-0.01	0.02	-0.01	0.00	0.09	-0.01	0.02	-0.10	-0.01	0.03
	diag(Σ)	0.24	0.10	0.01	0.01	0.01	0.00	0.05	0.01	0.02	-0.07	0.01	0.03
	β_1	0.31	0.25	-0.02	-0.02	-0.02	0.00	0.05	-0.02	-0.02	0.00	-0.02	-0.07

Appendix C. Parameter Estimates for PC buy/no buy data

Table 14. Comparison of estimates of Δ for the PC buy/no buy data.

		S-MNL			MVN-MNL			G-MNL		
		median	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
	Constant	**-1.80**	-2.00	-1.61	**-3.45**	-3.87	-3.09	**-5.54**	-6.24	-5.04
	Hot Line	**0.16**	0.05	0.27	**0.23**	0.03	0.42	0.10	-0.20	0.41
	Ram	**0.14**	0.02	0.30	**0.60**	0.38	0.82	**0.92**	0.65	1.20
	Screen	**0.23**	0.14	0.33	**0.47**	0.28	0.70	**0.91**	0.60	1.22
	CPU Speed	**0.37**	0.26	0.49	**0.89**	0.66	1.11	**0.96**	0.70	1.26
Intercept	Hard Disk	**0.21**	0.06	0.31	**0.33**	0.13	0.54	**0.42**	0.05	0.76
	CD	**0.40**	0.30	0.53	**1.08**	0.87	1.29	**1.16**	0.82	1.46
	Color	-0.10	-0.20	0.02	-0.17	-0.38	0.07	**-0.42**	-0.75	-0.11
	Channel	**0.24**	0.12	0.34	**0.56**	0.31	0.80	**1.22**	0.92	1.51
	Guarantee	**0.12**	0.04	0.22	**0.30**	0.08	0.52	**0.65**	0.35	0.95
	Price	**-1.51**	-1.68	-1.34	**-2.92**	-3.23	-2.66	**-4.03**	-4.57	-3.60
	Constant				**-0.37**	-0.65	-0.11	**-3.47**	-4.14	-2.82
	Hot Line				0.06	-0.10	0.21	-0.33	-0.81	0.07
	Ram				-0.03	-0.21	0.14	**0.39**	0.07	0.73
	Screen				**0.20**	0.06	0.38	**1.03**	0.56	1.47
	CPU Speed				-0.08	-0.27	0.09	-0.22	-0.59	0.13
PC Owner	Hard Disk				0.09	-0.12	0.29	0.13	-0.21	0.49
	CD				-0.09	-0.26	0.08	-0.23	-0.64	0.15
	Color				-0.10	-0.28	0.06	**-0.77**	-1.19	-0.33
	Channel				**0.20**	0.02	0.39	**1.32**	0.95	1.75
	Guarantee				**0.36**	0.19	0.52	**0.85**	0.41	1.28
	Price				-0.08	-0.27	0.11	**-0.85**	-1.29	-0.38
	Constant				-0.26	-0.84	0.44	0.09	-0.94	1.06
	Hot Line				0.44	-0.09	1.02	0.38	-0.19	0.97
	Ram				-0.17	-0.71	0.38	-0.40	-1.06	0.24
	Screen				-0.05	-0.59	0.43	-0.22	-0.75	0.42
	CPU Speed				0.07	-0.39	0.56	-0.03	-0.61	0.54
Gender	Hard Disk				-0.17	-0.72	0.41	-0.40	-0.96	0.20
	CD				-0.28	-0.81	0.22	-0.43	-1.01	0.24
	Color				0.26	-0.24	0.87	0.37	-0.30	1.06
	Channel				-0.35	-0.87	0.25	-0.61	-1.19	0.02
	Guarantee				0.00	-0.55	0.54	-0.14	-0.75	0.46
	Price				-0.13	-0.71	0.44	0.16	-0.83	1.02
	Constant				-0.05	-0.10	0.00	**-0.30**	-0.42	-0.19
	Hot Line				-0.01	-0.05	0.03	**-0.11**	-0.19	-0.03
	Ram				0.00	-0.03	0.04	**0.07**	0.00	0.15
	Screen				-0.01	-0.05	0.03	0.02	-0.05	0.10
	CPU Speed				0.00	-0.05	0.03	-0.02	-0.10	0.06
Age	Hard Disk				**-0.04**	-0.08	0.00	**-0.09**	-0.17	-0.02
	CD				-0.03	-0.07	0.01	**-0.13**	-0.21	-0.05
	Color				0.00	-0.04	0.04	0.00	-0.09	0.07
	Channel				0.02	-0.01	0.06	**0.14**	0.06	0.22
	Guarantee				-0.03	-0.07	0.01	0.05	-0.04	0.14
	Price				**0.05**	0.00	0.09	-0.06	-0.18	0.04
	Constant				0.09	-0.15	0.37	**0.78**	0.39	1.19
	Hot Line				-0.05	-0.23	0.14	-0.14	-0.39	0.11
	Ram				-0.01	-0.20	0.19	-0.06	-0.36	0.20
	Screen				0.03	-0.17	0.23	-0.10	-0.30	0.11
	CPU Speed				**0.30**	0.12	0.49	**0.39**	0.15	0.63
Expert Buyer	Hard Disk				0.14	-0.06	0.36	0.01	-0.24	0.26
	CD				0.02	-0.16	0.21	-0.08	-0.37	0.20
	Color				0.02	-0.17	0.22	0.05	-0.18	0.28
	Channel				-0.19	-0.40	0.04	**-0.36**	-0.66	-0.08
	Guarantee				0.03	-0.17	0.26	-0.15	-0.40	0.12
	Price				**-0.28**	-0.51	-0.07	0.20	-0.19	0.57

Table 15. Comparison of estimates of Σ for the PC buy/no buy data.

	S-MNL		MVN-MNL			G-MNL		
	median	2.5-97.5 %-tile	**median**	2.5-97.5 %-tile		**median**	2.5-97.5 %-tile	
Constant			**2.99**	2.33	3.73	**2.57**	1.88	3.57
Hot Line			**0.94**	0.76	1.19	**0.96**	0.78	1.27
Ram			**0.94**	0.78	1.13	**1.00**	0.81	1.25
Screen			**0.97**	0.83	1.22	**1.07**	0.90	1.31
CPU Speed			**1.16**	0.93	1.42	**1.22**	1.00	1.56
Hard Disk			**1.04**	0.85	1.24	**1.12**	0.82	1.45
CD			**1.11**	0.89	1.40	**1.13**	0.93	1.46
Color			**0.85**	0.70	1.08	**0.96**	0.75	1.20
Channel			**0.91**	0.71	1.24	**1.01**	0.81	1.24
Guarantee			**0.91**	0.75	1.18	**1.04**	0.73	1.30
Price			**1.85**	1.49	2.28	**1.85**	1.46	2.32