Essays on Integrating Product Design and Marketing Research

by

Eleanor McDonnell Feit

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration)
in The University of Michigan
2009

Doctoral Committee:

Professor Fred M. Feinberg, Chair
Professor Peter J. Lenk
Professor Puneet Manchanda
Professor Panos Y. Papalambros
Staff Researcher Mark A. Beltramo, General Motors Company

Acknowledgements

Mark Beltramo, **Peter Lenk**, and **Sue McDonnell**, each in their own way, played a critical role in persuading me to undertake this project. As I first contemplated applying to PhD programs, I was concerned about managing my responsibilities as a new mother with the rigorous life of a graduate student. Mark confirmed my fears, telling me that it would be “near-impossible” to balance a family and dissertation, even as he agreed to write a recommendation for me. Later, when I visited Michigan and met Peter, he assured me that this wouldn’t be a problem at all because women can manage any number of things. Peter’s words reminded me of my mother, Sue, who completed her dissertation when I was 12 and gave me a concrete example of how to be both a great researcher and a great mother. My father, **Frank McDonnell III,** also helped in his own way, offering his sensible and well-thought out advice, but only when solicited. Without the advice and example of these mentors, I would have never undertaken or completed this project.

As I worked my way though the program, I could not have asked for a more caring and empathetic advisor, **Fred Feinberg.** Fred embodies the word *advisor*, never using his students as a means to his own ends, but always helping them to find their own voice. (Literally, he once helped me overcome a writer’s block by drafting a few paragraphs in my writing style, which he was able to mimic perfectly.) He has always listened carefully to my ideas, encouraged me to develop my own interests and propped me back up when I felt I was failing. His co-authorship on the first essay of the dissertation (Chapter 2) is hardly compensation for the endless hours he has spent, editing this dissertation and chatting with me about research. Fred also served as co-PI on the NSF/GM grant that funded the first essay.

Mark Beltramo deserves the title of honorary co-advisor for the many hours he spent on the phone with me, coaching me through the challenges of developing and debugging MCMC samplers. Unlike most industry collaborators, he participated fully as a co-author of the first essay (Chapter 2), fully sharing in the research, writing and editing. In addition, he served as co-PI on the NSF/GM grant that funded the project and acted as the liaison with GM organizations that provided data. I also greatly admire his ability to subtly manage the people around him; I'm sure he knew me well enough to know that telling me that something is "near-impossible" would motivate me to stick to it.

I feel lucky to have had a dissertation committee made up of researchers who each have a deep command of an area related to my research. Mark and Fred were joined on the committee by **Puneet Manchanda**, **Peter Lenk** and **Panos Papalambros** who all contributed to the research in important ways. Peter was always there to answer technical questions and his answers led me to think more clearly about Bayesian inference, Puneet helped me to see how my work fits into the overall landscape of research in Marketing, and Panos never let me forget my roots in product design.

The rest of the faculty in the Marketing department at the Ross School contributed to my education and to this dissertation in numerous ways that can not all be recalled or recounted here. But the quantitative faculty, **Michel Wedel**, **S. Sriram** and **Anocha Aribarg**, deserve special mention as they were all subject to my random queries in the hallways of the Ross School, for example, "One of my data sets is missing a critical variable. What should I do?" Everyone provided accurate and encouraging answers. Michel, in particular, answered the question about missing variables, providing a critical suggestion for the second essay.

I also found great support in the broader faculty at Michigan, especially **Steven Pollock** who provided feedback on an early version of the third essay (Chapter 4) and **Roderick Little**, whose book taught me critical skills used in the first essay (Chapter 2). Both of them should also be commended for teaching the two classes that I most enjoyed during my time at Michigan.

The first essay of the dissertation was supported in part by the National Science Foundation and General Motors under Grant No. DMI-0541610. Any opinions, findings and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the GM or the NSF. In addition to their financial support, many GM staff provided helpful feedback on the project including **John Cafeo, Joe Donndelinger, Harvey Bell**, and **Alan Taub.**

Finally, the last five years would never have been possible without the understanding of my daughter, **Zoe McDonnell Feit,** whose empathy exceeds her years, and the steady support of my husband, **Jeff Feit**. I am lucky that I have family who was willing to recognize my ambitions and stand behind me as I pursued them (even pushing a little on occasion). I hope that I am able to return the favor as Zoe and Jeff pursue their goals in the future. While many warned that Jeff and Zoe would be liability during my graduate studies, I found them to be a foundation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Practicing engineers and product designers have developed sophisticated models to predict the technical performance of their designs, yet they seldom quantify how performance affects the desirability of the resulting products to consumers. For instance, car companies routinely predict the fuel economy of proposed designs early in the design process, but often have little information about what level of fuel economy will be acceptable to consumers. In order to make decisions, managers need quantitative estimates about how changes in product performance will impact sales. Without this information, managers have difficulty judging which products are most likely to succeed in the marketplace and may unduly emphasize the detailed estimates of cost and product performance that are available, while underestimating the impact of their decisions on market share and revenue.

A well-established approach to quantifying consumer preferences is to develop choice models based on choice experiments. Choice experiments are a marketing research method where consumers are exposed to hypothetical product profiles and are asked to choose the most desirable product from a set (Green and Rao 1971). The resulting data can be analyzed using discrete choice models such as multinomial logit and multinomial probit (cf., Train 2003). These models can be used to estimate the relationship between product attributes and consumer choices. By observing many product choices, it is possible to statistically estimate the relative value of each attribute. These preference measures can be used to drive design decisions, giving managers a quantitative estimate of how changes in product attributes will affect the desirability and potential sales of the product. Such tools can exert real impact on the products and services ultimately offered by a company and have become widely used in practice (Wittink and Cattin 1989).

However, there are still a number of problems that arise when these models are used in the design of complex products, i.e., products that require high levels of engineering, and in this dissertation we address three of those problems. The first essay addresses the issue that product designers often do not trust model predictions because the model estimates are based on observations of consumers making hypothetical purchase decisions in a survey setting. Asking consumers to make hypothetical decisions allows the researcher to use experimental design methods to estimate preferences for product features that do not yet exist in the market - a key interest of product designers. However, parameters estimated from experimental data often show marked inconsistencies with those inferred from the market, reducing their usefulness in forecasting and decision making. We propose an approach for combining choice-based conjoint data with individual-level purchase data to produce estimates that are more consistent with the market. Unlike prior approaches for calibrating conjoint models so that they correctly predict aggregate market shares for a 'baseline' market, the proposed approach is designed to produce parameters that are more consistent with those that can be inferred from individual-level market data.

The method proposed in the first essay relies on a new general framework for combining two or more sources of individual-level choice data to estimate a hierarchical discrete choice model. Past approaches to combining choice data assume that the population mean for the parameters is the same across both data sets and require that data sets are sampled from the same population. In contrast, we incorporate in the model individual characteristic variables, and assert only that the mapping between individuals' characteristics and their preferences is the same across the data sets. This allows the model to be applied even if the sample of individuals observed in each data set is not representative of the population as a whole, so long as appropriate product-use variables are collected that can explain the systematic deviations between them. The framework also explicitly incorporates a model for the individual characteristics, which allows us to use Bayesian missing-data techniques to handle the situation where each data set contains different demographic variables. This makes the method useful in practice for a wide range of existing market and conjoint data sets. We apply the method to a set of conjoint and market data for minivan choice and find that the proposed method predicts holdout market choices better than a model estimated from conjoint data alone or a model that does not include demographic variables. This joint model can be used by product designers to make design decisions with greater confidence that model predictions are consistent with the market.

In the second essay we turn to the issue of which model forms are best suited to data from a choice experiment and find evidence that models that account for heterogeneity in logit error scale fit better to some choice experiment data sets then the frequently used hierarchical multinomial logit model. Using a model with heterogeneity in error scale, we find preliminary evidence that respondents who have greater expertise in the product category will have lower estimated logit error scale, a finding which suggests that there may be individual-level differences in the consistency with which respondents answer questions in a choice experiment. Essays 1 and 2 are closely related to each other methodologically as both essays explore the role of error scale in the multinomial logit model. By using the models described in essays 1 and 2, market researchers will be able to provide more reliable estimates of potential market share to product designers.

In the third essay, which is more speculative than the first two, we turn to the question of whether information about a particular product design problem can be used to inform market research. We propose a new framework for designing choice experiments that chooses the questions to include in the choice task so as to maximize the profitability of the product design that will ultimately be produced for the market. Unlike other approaches to designing choice experiments, the proposed approach incorporates specific data on the cost and feasibility of developing alternative product designs. This data, which is often readily available in practice, can be used as part of the loss function in a decision-theoretic experimental design framework and would serve to focus the market research questions to provide information most relevant to the product design decision at hand. We speculate on how this framework could be further developed to provide product designers with market research that is tailored to their particular product design problems.

The three essays are described in chapters 2, 3 and 4 of the dissertation. Each essay stands alone as a complete work; references and appendices specific to each essay are presented within the chapter. In chapter 5, which concludes the dissertation, we summarize how each of the three essays contribute to the development of market research methods that can be used to inform product design.

References

Green, P.E. and V.R. Rao (1971) Conjoint Measurement for Quantifying Judgmental Data, *Journal of Marketing Research*, 8:3, 355-363.

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

Wittink, D.R. and P. Cattin (1989) Commercial Use of Conjoint Analysis: An Update, *Journal of Marketing*, 53:3, 91-96.

Chapter 2

Essay 1. Reality Check: Combining choice experiments with market data to estimate the importance of product attributes

Introduction

For many companies, decisions made today about which products should be developed will drive profitability for years or even decades to come (Krishnan and Ulrich 2001). A rich array of methods has been devised to guide managers in their product design decisions. Such methods typically attempt to measure the importance that consumers place on various product attributes and use those measurements to make design trade-offs. One of the most successful and widely-applied among these is conjoint analysis or choice experiments (Green and Rao 1971), a set of experimental techniques that present consumers with various combinations of product attributes and statistically estimate the effects of those attributes on choice. These survey-based methods have proven their worth in a remarkable variety of contexts (for examples see Green, Krieger and Wind 2001), yet they are not without their drawbacks. Most notably, the attribute effects estimated from choice experiments are sometimes inconsistent with those inferred from market data, an indication that respondents do not make hypothetical survey choices exactly as they make purchase decisions (c.f., Brownstone, Bunch and Train 2000, Blamey and Bennett 2001).

One naïve response to this failing is to abandon choice experiments and estimate models exclusively from market data. Unfortunately, in practice, there are frequently *key, managerially-relevant attributes* for which there is insufficient variation in the products offered in the market to estimate parameters. For example, in product design, managers often want to understand preferences for new attributes that are not yet available in the market, which is a practical impossibility using market data alone. There is also often extensive collinearity between observed and unobserved attributes in the products on offer in the market, due either to physical design limitations or similar marketing strategies among firms, which can lead to high levels of parameter uncertainty and even parameter estimates that are biased due to omitted variables (c.f., Brownstone, Bunch and Train 2000). In many markets, there is simply not sufficient information in the market data to estimate a heterogeneous choice model. In this paper, we develop a general method for combining different sources of choice data and apply it to the specific problem of combining data from choice experiments with market data. The incorporation of individual level market data adjusts some of the parameter estimates to be more consistent with those that we might infer from the market, thereby serving as a "reality check" on the conjoint data. However, as we will show, the proposed approach will not adjust any parameters that are not well-identified by the market data alone. For instance, parameters for new product features that are not yet available in the market will not be adjusted.

This approach is complementary to methods designed to improve the conjoint task itself, with the goal of generating hypothetical choices that are more consistent with choices observed in the market place. For example, Ding, Grewal and Liechty (2005) showed that when conjoint respondents are required to pay for and consume one of the conjoint profiles that they chose (selected at random after the conjoint task is completed), the conjoint model estimated from these "incentive-aligned" choices makes predictions that are much more consistent with observed market choices. The approach proposed here is complementary in that it focuses on augmenting the conjoint data in the estimation phase, rather than improving the data collection method. The two approaches could be used together, and incorporating market data in estimation might serve to adjust for any inconsistencies remaining between incentive-aligned conjoint choices and market choices.

A new approach to combining choice data

Methods for combining sources of preference data to estimate a *homogeneous* discrete choice model have been demonstrated in a number of applications in transportation research and environmental economics (see Ben-Akiva, Bradley, Morikawa, et al. 1994 and Louviere, Meyer, et al. 1999 for reviews). A key modeling insight in past work is that combining two sets of choice data requires a scale parameter to accommodate differences in error scaling between them (Ben-Akiva and Morikawa 1990, Swait and Louviere 1993). While there are many demonstrated benefits to combining sources of preference data, there remain a number of unresolved modeling issues concerning how to relate the two data sets together in the presence of consumer *heterogeneity* (Swait and Andrews 2003). In past work, researchers estimating heterogeneous choice models from two data sources required that each individual decision maker be observed making choices in both settings. With such data, they could impose the constraint that each individual maintained his or her preferences across the two choice contexts (Brownstone, Bunch and Train 2000, Bhat and Castelar 2002). However, in GM's experience, collecting such matched data would require effort in planning and recruitment that is impractical in commercial market research. Companies that regularly use choice experiments to estimate relative attribute importance typically also have access to individual-level market data that could readily be used to estimate joint models, but they seldom have this data for the *same individuals* that have completed the choice experiment. The available data sources often have strengths that are complementary, and our goal is to build a flexible modeling framework that can be readily used with commonly collected forms of consumer choice data.

A key element of our approach is a hierarchical choice model in which an individual's preferences depend on his or her personal characteristics (c.f., Allenby and Ginter 1995). Incorporating individual characteristics in the joint modeling framework confers several advantages. Most importantly, if consumers taking part in a conjoint study differ systematically in a relevant way from those in the marketplace, the analyst should not impose the restriction that the distributions of preferences in these two groups are identical; doing so is an overt misspecification. Instead of constraining the expected value of *preferences themselves* to be the same across the two data sets (as in Swait and Andrews 2003), we posit that the *relationship between individual-level characteristics and preferences* holds at the population level, and so is the same across data sets drawn from that population. Individual-level characteristics can include not only relevant socioeconomic variables but also information from the consumer about what needs he or she desires the product to fulfill. We refer to the latter type of data collectively as *product-use variables.* When available, product-use variables are typically much more informative about preferences than are commonly available demographic data (Fennell et al. 2003, De Bruyn et al. 2008).

Because we assume that the underlying relationships between the product-use variables and preferences are the same across data sets, the model structure can accommodate systematic differences in choice behavior—between individuals observed in the market and those observed in the choice experiment—that can be related to the observed product-use variables. Thus, the approach can be applied even if the sample of individuals observed in each data set is not representative of the population as a whole, so long as appropriate product-use variables are collected that can explain the systematic deviations. Past approaches to combining choice data are restricted to data sets that are random samples from the same population (Swait and Andrews 2003), which is difficult to achieve or ascertain in practice. For example, market research often samples a group that has a somewhat different distribution of demographic and product-use variables than the population as a whole (e.g., market research respondents are often older and sometimes more enthusiastic and knowledgeable about the category than the general population of buyers).

One potential disadvantage of incorporating individual characteristics using standard hierarchical choice models is that it requires that these variables be collected for all the decision makers observed in each choice data set. To maintain the assumption that the relationship between individual characteristics and attribute preferences is the same in each data set, it is critical that the same set of individual characteristics is accounted for in the regression of individual preferences on individual characteristics in both data sets. If the individual characteristics are correlated with one another, then omitting one of those individual characteristics may produce a bias in the remaining coefficients. Thus if an individual characteristic is omitted from one data set and not the other, the equality restriction on the parameters cannot be maintained. We overcome this disadvantage by incorporating a likelihood-based approach to missing characteristics (Little and Rubin 2002), which allows us to account for any individual characteristic(s) observed in at least one of the data sets. Bayesian estimation proceeds in a natural fashion using data augmentation for the missing characteristics.

Benefits of combining market and experimental choice data

The method we develop can be applied when combining *any* available sources of choice data, such as choices observed at different retailers or in different research studies, but there are particular advantages to combining experimental choices with market data. Hypothetical conjoint choices are typically collected following an orthogonal experimental design which can compensate for lack of variation and collinearity in market data (c.f. Louviere, Hensher and Swait 2000, chapter 8). There are additional benefits to combining conjoint and market data in the context of *heterogeneous* models. In market data for many product categories, we observe only one, or at most a few, choices for each individual (Urban, Hauser and Roberts 1990). Without a sufficient number of choice observations per individual, it is difficult to estimate the amount of unexplained heterogeneity in preferences, even if the parametric models employed by the analyst are formally identified (Andrews, Ainslie and Currim 2002, Rossi, Allenby and McCulloch 2005). For example, when we apply this method using market data on minivan purchases, we observe just one purchase for each household, so that a heterogeneous model cannot be estimated using this market data alone. By contrast, in a conjoint task it is relatively easy to collect multiple hypothetical choices for each respondent, and survey designs well-suited to estimating heterogeneous models can be readily developed (Sandor and Wedel 2005). When the data sets are combined, the conjoint data can serve to identify the distribution of heterogeneity, while still leveraging the observed preferences in the market data.

Combining data versus calibrating to aggregate market shares

There are several widely-used approaches that allow an analyst to calibrate a conjoint model by making *post hoc* adjustments to the estimated parameters so that the predicted shares closely match aggregate shares from the market (Orme and Johnson 2006). These approaches are relatively straightforward to implement, as they don't necessitate changes in estimation software and only require aggregate share data from the market. The resulting model, by design, makes market share predictions that closely match real-world market shares, a feature that helps build confidence with users. However, with these sorts of approaches, the analyst has to decide how closely to match the observed market data and which parameters of the model to adjust. When alternative-specific constants are included in the specification and are adjusted in the calibration, it is possible to calibrate the model so that it predicts observed shares as closely as the analyst wishes. Gilbride, Lenk and Brazell (2008) avoid most of these difficulties by specifying a loss function to choose a set of parameters that accurately predicts the aggregate market shares (in addition to fitting the likelihood well.)

Approaches that calibrate conjoint models to aggregate market share result in models that are well suited to sales forecasting and other decision problems where accurately predicting market share is a key goal. However, in product development applications, it is more critical that the *relative importance of product attributes is accurately estimated*, so that managers can determine which product attributes will have the greatest impact on sales. For example, a product designer is more interested in whether consumers place more weight on seating capacity or styling, rather than in the total market share for a particular brand. It has been shown that, in the realm of choice models, it is generally not possible to infer parameter accuracy on the basis on how well the model predicts market shares (Andrews Ainslie and Currim 2002). Further empirical work is required to determine how approaches that calibrate to aggregate market share work in practice, when the goal is accurately estimating the importance of product attributes.

By contrast, we propose combining individual level market-data with conjoint data directly in the model estimation. The resulting parameter for each attribute is influenced by both the market and the conjoint data in proportion to the relative degree of information available in the data for that attribute. Thus parameters, including alternative specific constants, that are not well-identified by the market data will not be influenced by supplementing the conjoint data with market data. The resulting model may not predict observed market shares as well as those deliberately calibrated to those market shares, but the parameter estimates will be consistent with the effects that can be identified by the individual-level market choices[1].

In the next section, we formally develop the model to combine individual-level choice data from different data sources. We then present an application of the model to experimental and market choice data for the US minivan market collected by General Motors. In the final section, we summarize our conclusions and discuss future research directions.

Model Development

Our model leverages the hierarchical discrete choice framework, where the part-worths of attributes are specified as a function of individual characteristics plus some error (c.f., Allenby and Ginter 1995). We assume that each individual's choices are related to a vector of attribute preferences, β_n , $n \in \{1, \ldots, N\}$. These preferences follow a multivariate normal linear model, i.e.,

$$\beta_n = \beta_0 + \Delta z_n + \nu_n \qquad \nu_n \sim \mathrm{MVN}(\mathbf{0}, \Sigma_\nu) \tag{1}$$

where β_0 is a vector of intercepts, z_n is a vector of observed characteristics of the individual, and Δ is an estimated matrix of regression parameters relating z_n to β_n. The error term, ν_n , is distributed multivariate normal with mean vector **0** and covariance matrix Σ_ν . The data sets are interrelated by taking the parameters β_0 and Δ to be common across data sets.

By relating the data sets through β_0 and Δ, we gain a great deal of flexibility. Past approaches to combining choice data do not include individual characteristics in this

[1] In the minivan application, the parameters we estimate jointly from conjoint and market data are by construction more consistent with the market data than a model estimated from conjoint data alone; a comparison of homogeneous models estimated using the minivan conjoint and market data independently bore this out. Direct evidence of this in the heterogeneous models is unavailable, as a heterogeneous model is inestimable from the minivan market data alone.

regression (i.e., they assume $E(\beta_n) = \beta_0$, which implies that the mean of β_n is the same across the two choice contexts. This assumption requires that the two data sets both represent random samples from the target population, which can be attempted through careful sampling (Swait and Andrews 2003) or by collecting experimental choices for the same group of individuals as is observed in the market (Brownstone, Bunch and Train 2000, Bhat and Castelar 2002). In the proposed approach we can avoid assuming that all data sets are a random sample from the target population, if we have appropriate individual characteristic variables, such as product-use variables, that can account for the relevant systematic differences between the two samples of decision makers. If there are differences in the distribution of z_n between the data sets, then our model will predict that the distribution of β_n and the resulting choices will be different across the two data sets. Because selection bias is prevalent in commercial marketing research, it seems prudent to accommodate any observed differences in the distribution of individual characteristics between the two samples. Of course, data with different empirical distributions of z_n should only be combined in situations where the researcher is confident that the specification of equation (1) is reasonable across both data sets. We do not recommend taking this approach to the extreme and combining, say, survey choices from 5 year-olds with market data from 55 year-olds in a situation where the relationship between age and preferences may be difficult to model, especially linearly.

Conditional on a vector of preferences, β_n, we assume that the likelihood of observing a particular choice follows the standard random utility formulation. Specifically, we assume that on each choice occasion, $t \in \{1,...T_n\}$, individual n will choose the alternative, $j \in \{1,...J_{nt}\}$, with greatest utility, u_{njt}, where

$$u_{njt} = \begin{cases} x_{njt}\beta_n + \varepsilon_{njt} & \text{if } n \in \text{data set } 1 \\ \mu x_{njt}\beta_n + \varepsilon_{njt} & \text{if } n \in \text{data set } 2 \end{cases} \qquad \textbf{(2)}$$

and x_{njt} is the row vector of attributes for alternative j faced by decision maker n on occasion t. The error term, ε_{njt}, is distributed IID according to the standard Extreme Value distribution. The resulting model takes the familiar multinomial logit specification. We chose a logit specification following the tradition in conjoint models; however, it would be possible to use a probit specification, though doing so would require some way of accommodating that each question in the choice experiment includes only a subset of the alternatives (c.f., Zeithammer and Lenk 2006). Equation (2) describes the approach for combining two data sets; extensions to three or more data sets would incorporate an additional scale parameter for each additional data set and are straightforward.

The Swait-Louviere scaling parameter, μ in equation (2), accounts for the possibility that the scale of the unexplained variation in u_{njt} is different across the data sets (Ben-Akiva and Morikawa 1990, Swait and Louviere 1993). This can arise for a number of reasons. Individuals' choice consistency is known to vary across choice contexts (Bradley and Daly 1994): for instance, consumers may make more consistent choices when making real purchase decisions than when making hypothetical decisions. The scale parameter can also be used to accommodate situations where proportion of variation explained by the attributes differs across data sources, e.g., when the set of observed attributes is different (Swait and Louviere 1993). Thus, the scale parameter makes it possible to combine data sources that have different, but overlapping, sets of product attributes. The choice of attributes to include in the specification of equation (1) is important. Since many product attributes tend to be correlated in the market, omitting an influential attribute from market data can lead to an omitted variables bias in β_n, which would be inconsistent with the scaling assumption in equation (2). The inclusion or exclusion of attributes in the conjoint data is less critical, assuming that the conjoint experiment manipulates the attributes orthogonally. In our example application, we are careful to include appropriate control variables in x_{njt} to mitigate omitted attributes bias in the market data.

If there is a difference in the parameters, β_0 and Δ, across data sets (for example, due to some context effect in the conjoint setting), the parameter estimates based on equation (2) will be informed by both data sets. Whether a particular parameter estimate is more consistent with the conjoint data or the market data is largely a function of the relative Fisher information for that particular parameter for each data set. Since the Fisher information depends on the sample size, the amount of the adjustment away from the conjoint data and toward the market data is related to the relative sample sizes for the two data sets. Prior work on combining choice data has not addressed the issue of how to determine the relative sample sizes or relative Fisher information of the two data sets and fully addressing that issue is beyond the scope of this paper. If the objective is to produce a model that will fit the *market* choices as well as possible, then the analyst should secure enough market data to 'overwhelm' the conjoint data for the parameters that are estimable from market data. Parameters that are not well informed by the market (due to lack of variation or collinearity) will not be adjusted and will remain informed primarily by the conjoint data. However, we caution against relying too heavily on the market data, as it may also inaccurately reveal attribute preference due to errors in the measurement of attribute values or unobserved supply-side constraints in the market.

Missing individual characteristics

While the model described by equations (1) and (2) allows for a great deal of flexibility by incorporating individual characteristics, it cannot be applied as-is when there are missing individual characteristics in one data set. This limitation poses a serious challenge. Even within a company with a systematic marketing research program, like GM's, it is very seldom the case that precisely the same set of individual characteristics is available in both data sets. This is especially so when these individual characteristics are product-use questions (which are most likely to be informative about attribute preferences). The need for consistency in ongoing market research programs (such as the survey of recent buyers that we use in our application) often limits opportunities for including new questions.

Because the method would be far less widely applicable if it required the same set of individual characteristics in both data sets, it is critical to overcome this apparent limitation. But this is made difficult by the intrinsically correlated nature of the individual characteristics themselves. If the regressors are correlated and each data set includes a different subset of the regressors, then regression coefficients for separately estimated models will both be biased in possibly different ways. This means the analyst cannot assert that the regression coefficients are the same for those variables that are common across the two data sets (c.f. Dominici et al. 1997). And if we cannot assert that the regression coefficients – β_0 and Δ, in (1) – are the same across both data sets, then we no longer have a way to relate the two data sets together. Thus the model in equations (1) and (2) can not be applied in its stated form if different individual characteristics are available in each data set.

To address this problem, we adopt a likelihood-based approach to missing data. Unlike imputation approaches to missing data, likelihood-based approaches simply define the likelihood of the observed data as the marginal of the complete data likelihood, integrating over the distribution of the missing data. This marginal likelihood can be maximized or used in Bayesian inference. This approach assumes that the process that caused the data to be missing is ignorable, which in turn requires that the characteristics are missing at random (MAR); that is, the probability that a covariate is missing does not depend on the realized value for that covariate, but may depend on other observed data. This is a reasonable assumption in the case where the covariate data simply was not collected in one of the data sources. It may not be appropriate in situations where particular respondents choose not to complete a particular survey question (e.g., if high income respondents are less likely to report their income). Importantly, the missing characteristics need not be missing completely at random (MCAR) for the approach to apply, and so the value of the missing characteristics *can* depend on the values of other characteristics or on the observed choices. (See Little and Rubin 2002 for a complete discussion.)

In the case of missing regressors, defining the complete data likelihood requires supplementing the usual likelihood for the dependent variables with a model for the regressors (Little and Rubin 2002, Dominici et al. 1997). The particular form of this model will depend on the nature of the regressors. For now, we will simply allow $[z_n \mid \varphi]$ to denote the likelihood of observing a covariate vector z_n dependent on some parameters, φ. Given an expression for the likelihood of the individual characteristics and the assumption that they are each MAR, we can write the likelihood of the observed choices as

$$[y, Z^{obs} \mid \mu, \beta_0, \Delta, \Sigma_v, \varphi, X] = \prod_n \left(\int_{z_n^{mis}} \left(\int_{\beta_n} \left(\prod_{t \in \{1,\ldots,T_n\}} [y_{nt} \mid X_{nt}, \beta_n, \mu] \right) [\beta_n \mid z_n, \beta_0, \Delta, \Sigma_v] d\beta_n \right) [z_n \mid \varphi] dz_n^{mis} \right), \qquad \textbf{(3)}$$

where $y \equiv \{y_{11}, \ldots, y_{nt}, \ldots y_{NT_N}\}$ is the set of all observed choices, $Z \equiv \{z_1, \ldots, z_N\}$ is the (complete) set of individual characteristics, $X_{nt} \equiv \{x_{n1t}, \ldots, x_{njt}, \ldots, x_{nJ_{nt}t}\}$ is the attribute data for a particular choice observation, and $X \equiv \{X_{11}, \ldots, X_{nt}, \ldots, X_{NT_N}\}$ is the set of all attribute data. The superscripts *obs* and *mis* indicate the observed or missing portion of the variable. Once the model $[z_n \mid \varphi]$ is specified, the likelihood in equation (3) can be used in maximum likelihood or Bayesian estimation.

The model for z_n can be any model that appropriately captures the relationships among the characteristics. If the z_n are continuous with full support, they can be modeled via a multivariate normal model (Dominici et al. 1997). However, covariate data used in marketing is often discrete or measured using a discrete scale (e.g., employment status, income ranges). In our particular case study, the covariate data used was binary, so we illustrate the approach for binary data only; extensions to other common survey data types are analogous and straightforward. To model the vector of correlated binary characteristics, we use a multivariate binary probit model (Chib and Greenberg 1998)[2]. We assume that the vector of zeros and ones that are observed, w_n, arises from an underlying multivariate normal vector, z_n, as follows,

$$w_{nl} = \begin{cases} 1 & \text{if } z_{nl} > 0 \\ 0 & \text{if } z_{nl} \leq 0 \end{cases} \quad \text{where, } z_n \sim \mathrm{MVN}(\mu_z, \Sigma_z), \tag{4}$$

where l indexes the elements of z_n and w_n. The covariance matrix, Σ_z, is restricted so that the variance of each element of z_n is one. When all of the individual characteristics are observed, the parameters of this model, μ_z and Σ_z, can be estimated separately for each data set (i.e., conjoint versus market), to account for selection differences between the two groups of decision makers. When individual characteristics are completely missing from one data set, as is the case in the minivan data, it is necessary to assume that μ_z and Σ_z are common across the two data sets. We model an individual's preference vector, β_n, as a function of the latent continuous vector, i.e., $\beta_n = \Delta z_n + \nu_n$. This structure allows preferences to vary continuously as a function of the underlying constructs that gave rise to their binary responses, and also preserves conjugacy in the estimation algorithm.

[2] Alternatively, Ibrahim, Lipsitz and Chen (1999) propose to model a vector of missing discrete regressors as a series of related univariate generalized linear models.

Estimation

Our approach to estimation is Bayesian, using diffuse but proper priors on all parameters (c.f. Rossi, Allenby and McCulloch 2005). The integrals over β_n and z_n^{mis} in equation (3) are handled using data augmentation (Tanner and Wong 1987). The resulting Gibbs sampler draws sequentially from the posterior of the parameters β_0, Δ, Σ_ν, μ, μ_z, and Σ_z, and the unobserved latent variables β_n and z_n . The parameters of the multivariate probit model, μ_z and Σ_z, are sampled over the unidentified space and posterior distributions for the identified parameters are obtained by marginalizing over the posterior draws (McCulloch and Rossi 1994). The full conditional densities of all parameters are standard distributions, with the exception of β_n and μ, which were drawn using Metropolis-Hastings steps. Because β_n is potentially a long vector, we used a normal random-walk proposal with an adaptive covariance matrix based on the covariance of all previous draws for individual *n*. This proposal density has been shown to maintain the convergence properties of the MCMC chain (Haario, Saksman and Tamminen 2001). The algorithm is described in detail in the Appendix A.

Insight into how the data informs the posterior for individual-level parameters

In a parameter-recovery study with simulated data (reported in Appendix B) we demonstrate two important features of the model. First, unsurprisingly, individual β_n are better recovered when relatively more of the variance is explained by z_n. Thus if β_n is poorly identified by the observed choices, as is common in market data where we observe few choices, including informative variables in z_n (e.g., product-use variables) improves posterior inference for β_n. More importantly, we also find that *increasing the number of observed choices for an individual improves inference about any missing individual characteristics for that individual*. Because we use a likelihood-based approach to the missing individual characteristics, our inference about a particular individual's latent characteristics (z_n) is informed both by what we know about other individuals' characteristics *and* the choices we have observed for that individual. In fact, the likelihood of z_n, conditional on the observed data and the other parameters, depends on both the model for the characteristics *and* on the choice parameters, β_n, as follows:

$$[z_n \mid w_n, \mu_z, \Sigma_z, \beta_0, \Delta, \Sigma_\nu, \beta_n] \propto [\beta_n \mid z_n, \beta_{0,} \Delta, \Sigma_\nu][z_n \mid w_n, \mu_z, \Sigma_z] \quad \textbf{(5)}$$

This stands in contrast to some other approaches to missing data, such as hot-deck, where missing characteristics would be imputed based only on information about other respondents' characteristics, ignoring the observed choices of the respondent in question.

Application: US Minivan Market

General Motors is among the many companies that regularly use choice models based exclusively on conjoint data to predict how new products will perform in the market. Because GM managers use these models to make critical product development decisions, they are keenly interested in improving the accuracy with which these models recover product attribute preferences. Methods that can be applied to existing conjoint data, with minimal additional data collection, are extremely valuable to GM and other practitioners, as they can be applied in situations where a conjoint study has been fielded, and the resulting parameter estimates are found to lack face validity. Because our model can accommodate data where different groups of respondents are observed in the conjoint setting and in the market setting, we can readily augment an existing conjoint data set with existing purchase data from a different set of consumers.

In this section we describe how our method was used to adjust the parameters of a conjoint model for minivan purchase. The goal of this section is to demonstrate how the proposed approach can be used to estimate a model jointly from market and conjoint choice data. We also explore the value of including the product-use variables, z_n, in the formulation by comparing our model to a model estimated without individual-level characteristics.

Conjoint data

The conjoint data for this application is a subset of data collected for a large conjoint study that was designed and fielded for GM in summer 2003. Our subset consists of 12 choice responses for each of 199 respondents who were selected based on their interest in purchasing a new minivan. In each choice task the respondents chose from among three alternatives with three attributes: price (levels: $20,000, $23,000, $26,000, $29,000, $32,000, $35,000), styling appeal (levels: very unappealing, unappealing, neutral, somewhat appealing, very appealing) and brand (14 levels which we label A-N at the request of GM). Respondents were randomly assigned to one of two fixed designs and made forced choices from among three alternatives. The choice questions were designed by GM's conjoint vendor using a proprietary method that allows for efficient estimation of a heterogeneous multinomial logit model. Although product-use variables were not systematically collected for each respondent in the conjoint study, the demographic profile did include one variable that is related to minivan product needs: number of children in the household.

Market data

To assemble market data that could be combined with the original conjoint study, we drew on an ongoing GM-proprietary survey of new vehicle buyers. This mail-out survey is sent quarterly to a sample of all new vehicle registrants. We selected from this survey all of the 7078 respondents who purchased a minivan during the 2004 model year (September 2003 – August 2004). For each respondent we observed one choice (the minivan purchase that qualified them for the survey) from among the 12 minivans that were on the market in 2004. The attribute data for the 12 minivans on the market were assembled from several sources.

The average consumer price paid (negotiated price less consumer rebates) for each of the minivan models was estimated based on the price reported by other buyers in the same survey. Although the prices faced by a particular individual (who may have been a particularly good negotiator, or shopping for a minivan with many extra features) could be different from the average prices we use in estimation, we assume that the average prices reasonably reflect the *relative* prices faced by each respondent.[3] These averages were computed by month to reflect seasonal price variation in the market data. We also assembled data from another GM source on the average consumer-rated styling appeal of each van (on the same scale as used for the conjoint study).

Table 1. Attribute data for alternatives available in the market (Brands D and M were not available in the market).

Brand		**A (old)**	**A (new)**	**B (old)**	**B (new)**	**C**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**N**
Styling Appeal		4.03	4.03	3.59	3.59	3.47	3.67	3.31	3.27	3.47	3.46	3.29	3.52	3.65	3.48
Price ($K)	Oct 03	20.7	NA	16.3	NA	22.0	18.8	12.7	17.5	22.8	21.4	20.4	18.8	21.2	16.8
	Nov 03	18.2	NA	17.5	NA	21.3	19.0	12.8	16.3	22.3	20.5	20.4	19.7	22.0	16.4
	Dec 03	17.8	NA	15.2	NA	20.4	18.1	12.6	15.4	20.8	21.1	19.8	18.1	20.8	16.3
	Jan 04	16.9	NA	16.4	NA	20.4	19.3	11.8	14.5	23.2	20.9	20.3	19.4	21.8	17.6
	Feb 04	16.6	NA	15.3	NA	18.5	18.5	11.7	14.4	22.9	22.5	20.8	19.6	22.0	17.1
	Mar 04	20.1	NA	16.0	NA	20.7	18.1	12.0	15.9	19.4	21.4	19.0	18.5	20.8	16.6
	Apr 04	NA	21.0	NA	16.6	17.8	18.3	10.3	12.9	21.5	22.0	22.2	18.8	18.9	17.6
	May 04	NA	18.7	NA	16.3	19.0	17.1	10.4	13.9	19.6	21.1	20.6	18.4	21.4	18.2
	Jun 04	NA	19.4	NA	17.7	18.7	17.5	10.2	13.7	24.4	22.3	19.1	19.2	21.1	16.0
	Jul 04	NA	18.7	NA	16.3	16.3	18.8	11.2	15.1	19.6	20.6	19.6	18.8	20.8	15.3
	Aug 04	NA	18.0	NA	15.4	19.6	17.4	11.9	13.4	17.8	22.9	17.8	16.4	21.4	14.5
	Sep 04	NA	19.3	NA	14.6	12.5	18.8	11.7	15.3	15.7	16.8	18.6	17.0	22.9	16.4
Last Design Refresh		2001	2005	2001	2005	2004	1999	2002	2000	2004	2004	1997	1997	2004	1997

As discussed above, it is critical that appropriate control variables be included in the product attributes in order to prevent an omitted variables bias in β_n. In the automotive market, products that have had a recent design refresh tend to have better features and command higher prices. So, we included the product attribute "date of last design refresh" in the product

[3] Because vehicle prices in the US are privately negotiated between the buyer and the dealer, it is difficult for a manufacturer like GM to get transaction data from the dealer for the particular individual that has been surveyed. Transaction data collected by third parties, such as J.D. Power's Power Information Network (PIN) data, do not contain the informative product-use variables that were included in the GM survey of recent buyers.

attributes for the market data. By controlling for the age of the design, we hope to prevent any omitted variables bias in the estimate of the importance of price in the market data. The attribute data for the minivans on the market is summarized in Table 1. Brands D and M are excluded from the market data, since there were no minivans on the market from brands D and M. Brands A and B launched re-designed products in April 2004 so we have used different attribute data for the old and the new designs.

Because we have only one choice observation for each respondent in the market, it is helpful to incorporate individual characteristics for the buyers in the market data. To improve individual-level parameter recovery, such characteristics should be correlated with attribute preferences and observed choices. Although past research has found relatively little correlation between standard demographic variables and attribute preferences (Fennell et al. 2003), variables that capture information about intended product usage or product needs have been found to be highly correlated with product choices (De Bruyn et al. 2008). We were able to construct similarly informative individual characteristics using a section from the market survey where respondents could check any of 78 potential "reasons for purchase", such as "Luggage/cargo capacity" and "Family oriented". GM developed these questions over several years of fielding the survey; the reasons were designed to be an exhaustive set and GM had found that they were related to brand choice. GM grouped these 78 items into 25 blocks of similar reasons using a clustering approach that resulted in groups with high face validity (see Table 2.) Using these blocks, we coded a binary variable for each respondent indicating whether the respondent had selected any item in the block. These 25 binary reasons-for-purchase variables entered the model as individual characteristics (z_n) that were covariates of the brand parameters. (We excluded from the data 208 buyers who did not check any of the 76 reasons, indicating that they failed to respond to that section of the survey.) We also included as covariates of the price parameters binary variables for whether the household had income less than $75K per year (roughly the median in this sample) and for whether or not the household had children.

Because the reasons-for-purchase data and the income data were not collected for the conjoint respondents, the likelihood-based missing-data approach was used to account for these missing variables. As there is but one individual characteristic common across the two data sets, this represents a fairly extreme instance of missing characteristics. Because there is no information available to estimate μ_z and Σ_z separately for the conjoint data set, we took them to be common across the two data sets. Note, however, that we observe a relatively large number of well-designed choices for the conjoint respondents and, based on these choices, the posterior for β_n is often quite tight. (The conjoint study was, after all, designed to infer β_n from the choices.) When the posterior for β_n is tight, the posterior for z_n^{mis} may be as well, and the posterior distribution of z_n for the conjoint respondents may be different from the distribution implied by μ_z and Σ_z.

Table 2. Summary of individual characteristics.

Covariate	% of Respondents		Covariate	% of Respondents	
	Field	Conjoint		Field	Conjoint
Household Income <75K	54.2%	-	Towing / Hauling	6.2%	-
Household with Children	48.8%	48.4%	Accident Safety	65.3%	-
Usability	75.0%	-	Collision Avoidance	35.3%	-
Dependability	63.6%	-	Kid Features	51.4%	-
Rugged / AWD / RWD	12.0%	-	Exterior Styling	61.0%	-
Dealer	57.0%	-	Fun to Drive	33.7%	-
Warranty	44.8%	-	Country of Origin	20.2%	-
Roominess	78.5%	-	Practical	51.8%	-
Cargo / Versatility	52.2%	-	Environment	15.4%	-
Fuel Economy / Value	58.7%	-	Manufacturer Reputation	58.9%	-
Incentives	43.6%	-	Interior styling	59.2%	-
Driving Performance	59.8%	-	Willing to Negotiate	30.6%	-
No Negotiation	34.8%	-	Cargo Loading	35.6%	-
Luxury	25.2%	-			

In the interest of parsimony, we placed restrictions on which individual characteristics were included in the regression for each attribute preference. For instance, whether or not a respondent has children or high income is excluded from the model of a respondent's preferences for particular minivan brands; the effect on brand preference of the former is captured by the “Kid Features” and other reasons-for-purchase variables and the effect of the latter operates through its effect on price sensitivity. Also, we assumed that the reasons for purchase are not related to the respondent’s price sensitivity.

The ongoing GM survey from which we collected the market data samples respondents on the basis of their chosen vehicle. The survey is mailed to a stratified sample of owners who have registered a new vehicle during the year, with the goal of receiving a fixed number of returns for each vehicle model. All returned surveys (typically around 20-25% of those mailed out) are included in the data set. To approximately adjust this choice-based sample to known market shares from national registration data, we adjust the likelihood of each individual’s choice following Manski and Lerman (1977), as follows

$$[y_{nt} \mid X_{nt}, \beta_n, \mu] = \frac{\exp\left(x_{n,y_{nt},t}\beta_n + \log(s_{y_{nt},t})\right)}{\sum_j \exp\left(x_{njt}\beta_n + \log(s_{jt})\right)}, \quad \textbf{(6)}$$

where s_{jt} is the sales-to-sample ratio for alternative *j* at time *t.* (Manski and Lerman developed this correction for homogeneous choice models.) Sales-to-sample ratios were computed for each month based on the number of survey responses and national sales data.

The resulting market data consisted of 6870 respondents for whom we observed one purchase and the 27 individual characteristics. We divided this data set into 2356 randomly selected individuals (roughly equal to the number of choices observed in the choice experiment) to be used for estimation with the remainder (4514 purchases) reserved as a holdout sample.

Fisher information for conjoint and market data

By using individual-level market data directly in the estimation, the proposed approach allows the joint estimation to adjust the parameters of the conjoint model so that they can account for empirical features of the market data not reflected in the conjoint data alone. It is important to note that this procedure does not necessarily influence all parameters to the same degree (or, in fact, at all), only those that are substantially informed by the market data itself. The relative influence of one data set versus the other for a particular parameter will reflect the relative information between the two data sets for that parameter. Table 3 reports the Fisher information for the population mean of each of the parameters, computed separately for the market and conjoint data. The conjoint data, with its well-designed choice questions, is reasonably informative about all of the parameters that were included in the conjoint study.[4] Thus, when the market data contains little information about a parameter, the conjoint data 'fills-in' the missing information.

For example, the Fisher information is zero for the brand parameters for D and M; since these were not offered in 2004, there is literally no information in the market data about the parameters for these brands. In contrast, the experimentally designed conjoint data has positive Fisher information for these two parameters allowing estimation of these parameters from conjoint data. Because the information is zero in the market data, the joint model estimate for brands D and M will be informed entirely by the conjoint data.

[4] We should point out that the conjoint data still has some differences in the information between brand parameters. Because the task was designed under the presumption of equal preferences for all brands, the conjoint data contain more information about the brands that were most frequently chosen in the conjoint task: A, J and M, and less information about the less frequently chosen brands. More balance across brand parameters could be achieved in the conjoint data if prior information about the relative preferences of the brands were used in the design of the conjoint task. If existing market data were used to form the prior, the result would be a conjoint study specifically designed to complement the market data.

Table 3. Fisher information for conjoint and market data.[5]

Population Mean of	Expected Fisher Information: Market (1 choice x 2356 respondents)	Conjoint (12 choices x 199 respondents)	Total
Styling	6.1	297.5	303.6
Price (linear)	94.6	281.4	376.0
Price (squared)	208.7	505.5	714.2
A	173.2	120.3	293.5
B	166.8	87.1	253.9
C	156.9	94.8	251.7
D	0.0	95.3	95.3
E	144.3	85.6	229.9
F	107.0	78.8	185.9
G	89.2	84.3	173.5
H	125.9	87.6	213.5
I	121.8	77.3	199.1
J	107.3	103.2	210.5
K	135.6	78.9	214.5
L	155.4	83.7	239.1
M	0.0	106.8	106.8
Design Age	379.9	0.0	379.9

Even if a variable is *observed* in the market data, the corresponding parameter may not be well *informed* by the market data. For example the market data information for the styling parameter is 6.1, relatively near zero, reflecting the lack of variation in styling among minivans on the market (all of the minivans have styling appeal slightly above neutral on a 5-point scale and, with the exception of two brands, styling did not change during the 2004MY—see Table 1.) Thus, it would be difficult to estimate a parameter for styling from the market data alone, which severely limits the usefulness of the market data for making product planning decisions with respect to styling. By contrast, the information for the styling parameter in the conjoint data is 297.5. Thus the joint model estimate for the styling parameter is primarily informed by the conjoint data. For brands such as A and B, where the information is large for both data sets, the parameter will be informed by both data sets.

We should point out that the lack of information in the market data about some of the parameters cannot be corrected simply by increasing the amount of market data used in

[5] We do not report the Fisher information for Brand N, as it was the base level in the effects coding.

estimation. The attributes of the products available, i.e. the ‘design matrix’, in each monthly market is fixed across all consumers, so observing, say, twice as many purchases in the market data will roughly double the information for each of the parameters, still leaving us unable to estimate parameters for styling and for brands D and M from the market data alone. Similarly, when we reduce the market data by half (as reported in the model estimates below) the information for each parameter in the market data is reduced roughly by half. Thus the conjoint data will always be critical for estimating managerially important parameters that could not be estimated using any quantity of market data.

Model estimates

The parameters of the model estimated from the minivan conjoint and market data are shown in Table 4. All reported estimates are based on 100,000 draws from each of two chains thinned to every 10th draw. Convergence was assessed by comparing the two chains and nearly all of the monitored parameters achieved Gelman-Rubin potential scale reduction factors below 1.1 (Brooks and Gelman 1998). Trace plots comparing the log-likelihood of the draws also indicated that the two chains had converged.

The central panel in Table 4 shows the parameters in Δ describing the relationship between the individual characteristics and the choice parameters. Consistent with intuition, the intercept for the Styling Appeal parameter is positive and the intercept for the Price (linear) parameter is negative. Households with lower than median income and households with children have higher price sensitivity. Many of the other parameters in Δ are consistent with GM managers’ intuition; for example, the estimate for the relationship between the “Warranty” reason for purchase and preference for brand “F” is high, indicating that respondents for whom warranty is important are more likely than others to choose brand F, consistent with brand F’s industry-leading warranty program. We also find that buyers who indicated “Country of Origin” as a reason for purchase had significantly lower preferences for brands E, F, G, I and L, which were the only non-US brands in the sample.

Table 4. Estimated parameters for joint choice experiment/market data model.

mu **0.55**		***Delta*** Other Attributes (X)				Brands (X)													
		Styling Appeal	Price (linear)	Price (squared)	Year of Last Refresh	A	B	C	D	E	F	G	H	I	J	K	L	M	***mu.z***
Covariates (z)	Intercept	**2.74**	**-1.43**	**-0.24**	**1.08**	**3.92**	**2.07**	0.40	**0.35**	-0.28	**-6.17**	**-3.43**	-0.04	-0.91	**2.95**	0.77	0.12	**-1.69**	**NA**
	Household.Income.<75		**-1.00**	0.25															**0.13**
	Household.with.Children		**-0.65**	0.05															-0.04
	Reason.Usability					0.38	**1.02**	0.06		**-1.26**	-0.01	-0.59	0.02	-0.58	0.69	0.39	-0.88		**0.67**
	Reason.Dependability					**-0.93**	-0.27	-0.05		**1.59**	0.36	0.64	-0.60	0.63	-0.35	**-1.18**	**1.48**		**0.37**
	Reason.Rugged_AWDRWD					-0.33	-0.58	-0.25		**-0.87**	**-1.04**	-0.27	-0.79	**-1.05**	**1.33**	**1.57**	0.25		**-1.17**
	Reason.Dealer					-0.12	-0.05	**1.11**		**-0.81**	-0.09	0.02	**1.48**	**-1.16**	0.08	0.43	**-1.45**		**0.17**
	Reason.Warranty					0.16	**0.73**	**-0.92**		-0.09	**3.24**	**0.85**	**-1.02**	-0.25	0.08	**-1.42**	**-0.85**		**-0.13**
	Reason.Roominess					-0.38	-0.59	**-0.73**		0.28	0.16	**0.85**	-0.28	0.82	-0.70	0.39	0.08		**0.79**
	Reason.Cargo_Versatility					0.30	0.12	**0.73**		0.02	**-1.34**	-0.30	0.48	-0.16	0.23	0.10	0.08		**0.06**
	Reason.FuelEcon_Value					**-0.72**	-0.08	**-1.07**		0.19	0.80	**1.41**	**-0.86**	0.14	**-0.54**	0.23	0.51		**0.22**
	Reason.Incentives					0.49	0.47	**1.57**		**-2.50**	-0.05	**-0.69**	**1.57**	**-2.68**	**1.70**	**1.81**	**-3.07**		**-0.18**
	Reason.DrivePerform					0.61	-0.13	0.74		**-1.06**	-0.10	-0.68	0.36	**-1.28**	0.70	0.76	-0.65		**0.25**
	Reason.No_Negotiation					0.07	**-0.65**	-0.23		0.43	0.75	-0.21	-0.11	0.40	-0.37	0.18	-0.10		**-0.39**
	Reason.Luxury					**0.74**	-0.45	-0.18		-0.37	-0.47	-0.32	**0.85**	**1.54**	-0.45	**-1.28**	**1.00**		**-0.66**
	Reason.Tow_Haul					0.42	0.32	-0.25		0.08	0.81	0.18	-0.34	0.32	-0.52	-0.64	0.07		**-1.52**
	Reason.Safety_Security					-0.62	-0.54	0.80		0.47	0.72	0.18	0.39	0.00	**-1.03**	**-1.41**	**1.02**		**0.40**
	Reason.AvoidCollision					**-1.01**	**-1.11**	0.61		-0.08	-0.64	-0.67	**1.49**	0.74	0.49	**1.07**	-0.04		**-0.37**
	Reason.KidFeatures					-0.14	0.21	-0.34		**0.96**	-0.56	0.15	**-1.20**	**0.92**	**-0.55**	-0.03	**0.59**		0.03
	Reason.ExteriorStyling					-0.12	0.08	-0.11		-0.57	-0.49	**0.65**	-0.03	0.46	0.10	0.52	**-0.71**		**0.28**
	Reason.FuntoDrive					**-1.37**	0.02	**-1.17**		0.28	0.75	**1.35**	-0.60	**2.40**	-0.49	-0.08	-0.04		**-0.41**
	Reason.CountryofOrigin					0.38	**0.63**	0.14		**-1.08**	**-1.74**	**-1.02**	0.19	**-0.85**	**1.25**	**1.44**	**-0.96**		**-0.83**
	Reason.Practical					0.32	0.30	0.11		-0.29	0.38	-0.52	-0.11	**-0.89**	0.20	**0.64**	-0.48		0.04
	Reason.Environment					**0.91**	0.39	0.34		0.75	-0.59	-0.69	0.03	-0.80	-0.10	-0.40	0.39		**-1.01**
	Reason.MfgReputation					0.35	0.08	-0.71		**2.23**	**-1.75**	-0.40	-0.76	0.16	0.03	**-1.09**	**1.96**		**0.24**
	Reason.InteriorStyling					0.23	-0.48	-0.20		0.11	**1.57**	0.00	0.25	-0.12	-0.35	-0.16	0.43		**0.23**
	Reason.WillingtoNegotiate					0.31	**0.54**	-0.43		0.37	-0.25	0.01	-0.17	0.20	-0.06	0.00	-0.24		**-0.50**
	Reason.CargoLoading					0.01	0.25	0.19		-0.01	-0.51	0.30	-0.23	0.31	-0.22	-0.53	0.58		**-0.37**
Sigma.nu		**3.59**	**5.88**	**0.90**	**3.08**	**7.52**	**11.13**	**4.94**	**1.29**	**2.48**	**2.68**	**3.71**	**2.07**	**3.78**	**2.64**	**6.29**	**1.91**	**10.88**	

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.

The last row in Table 4 lists the estimated variances of the unexplained population heterogeneity for the attributes. There is more unexplained heterogeneity in preferences for styling and the linear term for price and less unexplained heterogeneity in preferences for age of vehicle design and the squared term for price. Unexplained heterogeneity in brand preferences varies widely depending on the brand. Some brands seem to be more universally liked or disliked while others appear to have more dispersion across individuals. The right-most column in 4 shows the estimated population means for the multivariate probit model that was used to estimate missing individual characteristics.

The value of incorporating market data

The key benefit of the proposed modeling framework is that it allows us to incorporate both market and conjoint data to estimate a choice model. By combining these sources of data, the resulting model still benefits from the well-conditioned attribute data in the conjoint study, yet should make more accurate predictions about choices in the market. To gain some insight into the effect of incorporating the market data, we compare the joint model to a model estimated from the conjoint data alone. We compare the ability of the joint model and the conjoint model to predict market data based on the posterior predictive likelihood of the estimation and holdout choices. We compute the log posterior predictive likelihood (*lppl*) of an observed choice by individual *n* on occasion *t* as

$$\log\left(\int_{\beta,z,\mu,\beta_0,\Delta,\Sigma_\eta,\mu_z,\Sigma_z} [y_{nt} \mid X_{nt},\beta,\mu][\beta \mid z,\beta_0,\Delta,\Sigma_\nu][z \mid w_n,\mu_z,\Sigma_z] [\mu,\beta_0,\Delta,\Sigma_\nu,\mu_z,\Sigma_z \mid \text{data}]\, d\beta dz d\beta_0 d\Delta d \right. \tag{7}$$

where $[\mu, \beta_0, \Delta, \mu_z, \Sigma_z \mid \text{data}]$ is the posterior distribution of the population parameters. (Note that *lppl* is proportional to the deviance averaged over the posterior distribution of the population parameters.) This measure of model fit conditions on w_n and reflects how well the estimated model is able to predict an individual's choices given his or her characteristics. To give managers a more intuitive measure of the predictive performance, we convert *lppl* to an "average hit rate," computed as exp(*lppl*/*N*), where *N* is the number of choice observations. Note that the conjoint and the joint models have essentially the same *structure* and differ primarily on what *data* is used in estimation. Thus it would be less appropriate to compare the models using Bayes Factors, which compare how well two models with different structure fit the same data set.

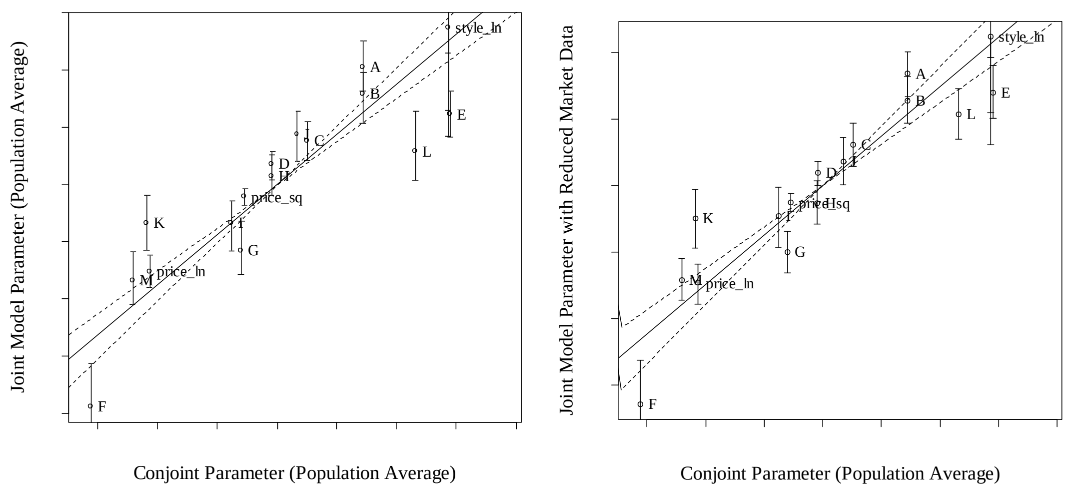
In this example, we find that the joint model does a substantially better job at predicting market choices than the conjoint model. The *lppl* of the market choices used in estimation is -4711 (average hit rate = 13.5%) for the joint model versus -6333 (average hit rate = 6.8%) for the conjoint model. In fact, the conjoint model does worse at predicting market data than a model that predicts according to the aggregate shares in the market data (*lppl*=-5830, average hit rate = 8.4%), an indication that the preferences expressed in the conjoint study were inconsistent with market shares. Clearly, incorporating market data yields more accurate predictions about individual purchases in the market. We find a similar pattern of results in the *lppl* of the holdout market data (see Table 5). Interestingly, we also find that the joint model only does slightly worse than the conjoint model at predicting conjoint choices (*lppl*=-2325 versus *lppl*=-2249.) Thus, the joint model can predict market data much better than the conjoint model, yet still makes reasonably good predictions for the conjoint data.[6]

To understand why the conjoint model makes poor predictions for the market data relative to the joint model, we compare the parameters of both models. Figure 1(a) shows the differences between the population mean of β_n estimated from the conjoint model versus the joint model. (When conjoint data are analyzed alone, the specifications of β_0, Δ and Σ_ν differ from that in the joint model due to the missing characteristics in the conjoint data. We therefore compare the distributions of individual β_n implied by each model, but we do not compare β_0, Δ and Σ_ν directly. The conjoint-only parameter estimates for β_0, Δ and Σ_ν are included in the Appendix C.) The population means of the elements of β_n are plotted for the joint model versus the conjoint model and the error bars indicate the 5th and 95th percentile for the joint model estimates.

[6] Our focus here is on *comparing* the predictive performance of the models and so do not view the *generally* low hit rates as cause for concern. In practice, GM has found that it can achieve substantially higher choice models fits, both in sample and out-of-sample, by including additional attributes that are predictive of automotive choice (such as fuel economy, passenger capacity, driving performance, etc.), when they are available.

Figure 1. Comparison of estimated parameters for conjoint and joint models.

(a) (b)



The solid diagonal line in Figure 1(a) has a slope equal to the estimated scale difference (μ) between the conjoint and the market data and the dashed lines show the 5th and 95th percentiles of the posterior distribution of μ. When a parameter falls below this range, it suggests that the attribute level is less preferred in the joint model relative to the conjoint model. Parameters above the range are associated with attribute levels that are more preferred in the joint model. Although the joint model can change the parameters for styling and price relative to the conjoint-only model, substantial adjustments to these parameters do not seem to be supported by the data. The joint model does substantially adjust preference for many of the brands, which we attribute to differences between the general “knee-jerk” brand attitudes that consumers express in the conjoint setting versus how the brands are perceived when the customer is shopping for a minivan and is actively engaged in collecting information about the brands and specific products. While it is impossible to say which set of parameter estimates are closer to the “true” effects in the market, the effects are by construction more consistent with what can be inferred from the market data. Important product planning decisions, such as which brands should offer minivans in the future, would be misinformed using the model estimated from conjoint data alone.

The effects of relative sample sizes

In situations, where the effects of the attributes may differ across the two data sets, the estimated parameters of the joint model will depend on the relative information in the two data sets. Reducing the sample size of the market data will, all else equal, reduce the Fisher information for the market data proportionately, resulting in a model that is more consistent with the conjoint data. To demonstrate this, we re-estimated the joint model using approximately half as much market data (1194 individuals.) (Parameter estimates are included in the Appendix C.) (b) compares the population means of the parameters for the conjoint model to this second joint model. Note that the dispersion in Figure 1(b) is relatively less than in Figure 1(a), indicating that the joint model with reduced market data is more similar to the conjoint model than the joint model with more market data. The model estimated with reduced market data also fits the conjoint data better; the *lppl* of the conjoint data for this model is -2290 (versus -2325 for the original joint model). Similarly, the *lppl* of the holdout data for this second joint model is -9786 (versus -9689 for the original joint model) indicating that the joint model with less market data is not as good at predicting market choices. Thus the degree of adjustment achieved is dependent on the amount of market data used in estimation. In general, the analyst should consider his or her relative confidence in the two data sources when choosing how much data to use from each source.

The value of incorporating individual characteristics

To assess the benefits of incorporating individual characteristics, we also compare our model to one where all elements of Δ are fixed to zero, leaving only the intercepts. We will refer to this specification as the No-Individual-Characteristics (NIC) formulation. Similar to the model of Swait and Andrews (2003), this model forces the means of the part worths to be the same across the two data sets. We find overwhelming support for our model over the NIC model. The Newton-Raftery estimator (Newton and Raftery 1994) of the marginal likelihood of the NIC model is -4216 versus -3273 for the proposed model, indicating a log Bayes Factor of -943 in favor of our proposed formulation. (Because of the unbounded sampling variance of the Newton-Raftery estimator, we repeated this calculation individually for each chain. We found the estimated integrated log-likelihood to be similar across the chains with a difference of no more than 40 points, still clearly favoring our proposed formulation.) We also found that the *lppl* of the holdout data is significantly better for our formulation (*lppl*=-9689 for the joint model versus -11100 for NIC model). In other data sets where the differences in the distribution of individual characteristics between the market and the conjoint data are better observed, we would expect our formulation to be even more strongly favored.

Table 5 summarizes the *lppl* of the conjoint data, the market data used in estimation and the holdout market data relative to the joint model and our comparison models. Overall, we find that the joint model does a better job than the conjoint model or the NIC model at predicting holdout market choices and thus it is better suited to making predictions about product planning decisions that will play out in the market.

Table 5. Log posterior predictive likelihoods and average hit rates[7]

Model	Fit to Estimation Data				Fit to Holdout Data	
	Survey		Market			
	lppl	average hit rate	*lppl*	average hit rate	*lppl*	average hit rate
Joint Model	-2325	37.8%	-4711	13.5%	-9689	11.7%
Joint Model (reduced market data)	-2290	38.3%	-4823	12.9%	-9786	11.4%
Conjoint Model	-2249	39.0%	-6333	6.8%	-12019	7.0%
NIC model	-2245	39.1%	-5738	8.8%	-11100	8.6%
Aggregate Shares	-2618	33.4%	-5830	8.4%	-11180	8.4%
N	2388		2356		4514	

[7] Note that, for ease of comparison, each method's *lppl* for the estimation market data is computed with the 2356 purchases used to estimate the joint model—even for methods that use only some or none of that data.

Market share predictions

The focus of this paper was on adjusting the parameters of a conjoint model so that the estimated parameters are more consistent with observed market choices, not on accurately predicting aggregate market shares. While precisely matching market shares for a base case may raise managers' confidence in the model, it does not ensure that the *effects* are consistent with those that can be inferred from market data. Studies with synthetic data have shown that for choice models it is generally not possible to infer parameter accuracy on the basis of how well the model predicts market shares alone (Andrews Ainslie and Currim 2002). A popular approach proposed by Orme and Johnson (2006) calibrates the conjoint model by adjusting alternative specific constants, which correspond to the brand parameters in the minivan model, so that the model almost perfectly matches observed market shares. By contrast, in a model estimated jointly from individual-level market and conjoint data, it is the data that determines which parameters are adjusted. The resulting model may not predict observed market shares as well as a model calibrated to exactly match those market shares, but the parameter estimates will be consistent with the effects that can be identified by the market data.

Table 6. Comparison of aggregate market share predictions.

Brand	MY04 Annual Market Shares	Predicted Annual Market Share (95% credible interval)		
		Conjoint Model	Joint Model	NIC Model
A	13.0%	9.5% (6.5%, 13.2%)	15.1% (11.5%, 19%)	15.3% (11.5%, 19.6%)
B	15.4%	13.7% (8.7%, 19.5%)	13.5% (9.7%, 17.1%)	14.9% (10.2%, 19.5%)
C	7.2%	6.0% (3.6%, 9%)	10.7% (7.4%, 14.5%)	9.6% (5.9%, 13.4%)
E	19.9%	20.6% (14.5%, 26.7%)	10.4% (6.9%, 13.7%)	12.0% (8.3%, 16.6%)
F	6.6%	3.4% (1.8%, 6%)	5.2% (2.9%, 7.6%)	5.6% (2.5%, 9.5%)
G	1.9%	5.4% (3.1%, 8.6%)	2.4% (1.1%, 3.8%)	1.8% (1%, 3.1%)
H	1.8%	3.0% (1.6%, 4.7%)	5.6% (3.4%, 8%)	3.4% (2%, 5.2%)
I	6.1%	2.6% (1.4%, 4.2%)	6.9% (4.4%, 9.8%)	6.0% (3.5%, 8.9%)
J	0.9%	4.6% (2.6%, 7.2%)	3.2% (1.8%, 4.9%)	2.4% (1.3%, 4%)
K	2.3%	2.6% (1.4%, 4.3%)	7.1% (4.5%, 10.5%)	6.2% (3%, 9.9%)
L	20.5%	14.9% (10%, 20.8%)	12.9% (9%, 17%)	16.0% (11.2%, 21.3%)
N	4.2%	13.7% (8.4%, 20.2%)	7.0% (4%, 10.2%)	6.8% (3.5%, 11.2%)

For completeness, however, Table 6 reports the aggregate market share predictions for the joint model as well as the conjoint model and the NIC model. These predictions were obtained by computing the monthly shares based on the attribute data reported in Table 1 using the population parameters of the model (β_0, Δ, Σ_v, μ_z and Σ_z). Because market-share forecasts should reflect the behavior of the population as a whole, the shares were computed by integrating over the estimated population distributions of β_n and z_n. (The integrals were approximated using 200 random draws of β_n and z_n from the population distribution.) So, unlike the *lppl* statistics reported in Table 5, the share estimates do not reflect the likelihood of an individual with a particular set of individual characteristics choosing a certain minivan. Rather, the shares represent the overall prediction from the model about the propensity of the *population* to choose a particular minivan. The monthly share predictions were aggregated to annual shares by weighting the share predictions by the actual sales volume for each month. Posterior uncertainty in the population parameters was propagated by repeating the market share prediction for 1000 draws from the joint posterior distribution of the population parameters to produce a predictive distribution for annual market shares. Table 6 reports the mean of this distribution along with the 2.5^{th} and 97.5^{th} percentiles of the distribution. We compare these market share predictions to the actual annual market shares observed in the data.

Overall, as was our expectation, we don't find substantial differences in the ability of the joint model and the conjoint model to predict market shares. While the joint model and the conjoint model do make different market-share forecasts, neither of them is particularly accurate in predicting the actual market share, and both have fairly wide prediction intervals. This analysis demonstrates how difficult it can be to judge whether the estimated effects in a conjoint model are consistent with the market based solely on single set of historic aggregate market shares. Even if the aggregate market share predictions indicate that the conjoint model makes an inaccurate prediction for a particular brand (e.g., Brand N), we cannot determine whether this is due to an inaccurate measure of preference for Brand N or inaccuracy in other estimated effects, such as price or styling. But by comparing the conjoint model to a model that is estimated using individual-level market choices, as in **Error! Reference source not**

ound., we can see directly which effects may have been inaccurately measured in the conjoint study. The joint model, by design, adjusts the estimated parameters to be more consistent with what can be inferred about those parameters in the market data.

Conclusions

Companies need to understand the relative importance of product attributes to consumers, and so have traditionally turned to methods from marketing science to measure it. Chief among these methods are discrete choice models estimated from experimental choice data. But it is well-known that certain critical quantities can be inaccurately measured by even the most scrupulous conjoint design, for example, reactions to price changes or socially-desirable attributes. Conversely, market data does not allow product designers to assess the impact of attributes that are truly new, or do not vary sufficiently among products on the market. The two types of data have complementary strengths, yet prior work attempting to meld them had data requirements so stringent as to render most existing data sources unusable. In this article, we developed a flexible framework for combining existing sources of conjoint and individual-level market data in estimation to produce a model that is more consistent with attribute effects observed in the market.

Using conjoint and market data for minivans, we found that the model estimated jointly from both conjoint and market data predicts holdout market choices better than one estimated from conjoint data alone, demonstrating the benefits of pooling information from multiple data sources. A particularly useful aspect of the minivan market data was that it included each individual's 'reasons for purchase', which—in contrast to prior findings in the empirical modeling literature about demographic variables—turn out to be effective at explaining the relationship between product attributes and choice. Using our framework, we were able to incorporate these characteristics, which (along with the conjoint data) help to inform the distribution of heterogeneity for the choice parameters even though we only observe one choice for each individual in the market data. Our joint model also fits the estimation data better and predicts holdout purchases better than a model that does not include the individual characteristics, demonstrating that including these individual characteristics not only allows for flexibility, but also improves prediction.

It would be valuable to test this method using other data sets, as the minivan data represented a rather extreme case of missing characteristics. In situations where there is more overlap between the characteristics, we would expect the likelihood-based missing-data method to perform even better. The present application also did not allow for a strong test of the ability of the model to accommodate situations where there are observed differences in the distributions of the individual characteristics. It would also be useful to apply the method to “incentive-aligned” conjoint data (Ding, Grewal and Liechty 2005) to explore the complementary strengths of approaches that improve the design of the conjoint data collection versus approaches that incorporate other data sources in model estimation.

There are a number of extensions to this model that could be considered to accommodate different data. Although we have chosen a logit specification for tractability and to conform to typical conjoint practice, a probit specification could be used. In situations where there is selection bias based on the outcomes, it would also be possible to incorporate a model that accounts for selection on β_n (Heckman 1979), such as might happen if the market data contained buyers and the conjoint data sampled both buyers and non-buyers. In durables, it is also common to collect data on the consumer’s second choice product, and this second-choice could be formally incorporated in the likelihood. Researchers can readily accommodate such extensions within the Bayesian MCMC sampler.

In the present work, we have also assumed that the parameters that relate an individual’s characteristics to his or her attribute preferences are common across choice data sets. However, it is possible that there are systematic differences in choices made in different choice contexts. One way to capture these differences in behavior across choice contexts is to introduce another layer in the hierarchical model to allow for shrinkage across data sets. If a large number of choice contexts were observed (say market data across different retailers or a number of conjoint studies fielded at different locations) it would be possible to explicitly model the distribution of β_0 and Δ across data sets, adding another level to the hierarchical model. Dominici et al. (1997) applied a similar idea in the context of hierarchical linear models for meta-analysis of regression studies. We also encourage consumer behavior researchers to further explore the underlying mechanisms that cause differences in choice behavior between real markets and hypothetical conjoint experiments. A better understanding of context effects that frequently occur in the conjoint setting could inspire a structural model that explicitly accounts for the behavioral differences across the two choice contexts.

Appendix A. Priors & sampler algorithm

Throughout Appendix A, we will assume that z_n includes the value 1 as its leading element for all *n* and that Δ includes an initial column to multiply these initial ones. Thus the intercept β_0 is incorporated into Δ.

Priors

We use proper but diffuse conditionally conjugate priors. Specifically,

$$[\Delta] = \mathrm{MVN}\left(\mathrm{vec}(\Delta) \middle| \mathbf{0}, diag(vec(F))\right)$$

$$[\Sigma_v] = \mathrm{I}W(\Sigma_v \mid K+2, I)$$

$$[\mu_z] = MVN(\mu_z \mid \mathbf{0}, 1000I)$$

$$[\Sigma_z] = IW(\Sigma_z \mid L+2, I)$$

$$[\mu] = \mathrm{Gamma}(\mu \mid 1/1000, 1000)$$

where *I* is the identity matrix, *K* is the number of attributes and *L* is the number of individual characteristics plus one for the intercept. Note that we are not using the more common fully conjugate prior for Δ and Σ (c.f. Rossi, Allenby and McCulloch 2005, p. 71). The non-standard form of the prior on Δ allows us to specify tight priors on particular elements of Δ, which we use to restrict which individual characteristics relate to particular attribute preferences. *F* is a matrix of the same dimension as Δ that is used to determine which elements of Δ have tight priors near zero and which have diffuse priors. For a tight prior, the corresponding element of *F* had a value of 10^{-12}. For a diffuse prior, the corresponding element of *F* had a value of 1000.

Sampler Algorithm

Step 0. Initialize values for μ, Δ, Σ_υ , μ_z and Σ_z, and for β_n and z_n^{mis} for all *n*.

The scale ratio μ is initialized to the maximum likelihood estimate from a homogeneous joint model. The parameters in Δ are initialized at their maximum likelihood estimates from a homogeneous logit model that includes interactions between elements of w_n (coded -1,1) and x_{njt} and was estimated from the market data. The vector μ_z is initialized to zero. The matrices Σ_υ and Σ_z are initialized to identity matrices. Starting values of z_n are drawn based on μ_z, Σ_z and any observed w_n. Then, we generate starting values for β_n and z_n^{mis} according to the model.

Step 1. For each individual, *n*, draw β_n.

$$[\beta_n|\{y_{nt}\},\{X_{nt}\},z_n,\mu,\Delta,\Sigma_\nu] \propto \left(\prod_t [y_{nt} \mid X_{nt},\beta_n,\mu]\right)[\beta_n \mid z_n,\Delta,\Sigma_\nu]$$

$$\propto \left(\prod_{t\in\text{data set 1}} \frac{\exp(\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\beta_n x_{njt})}\right)\left(\prod_{t\in\text{data set 2}} \frac{\exp(\mu\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\mu\beta_n x_{njt})}\right) N_K(\beta_n \mid \Delta z_n,\Sigma_\nu)$$

This distribution is not a standard distribution and we use a Metropolis-Hastings step to complete the draw. The proposal is a multivariate normal random-walk from the most recent draw where the covariance of the random walk for individual *n* is based on the covariance of all previous draws for individual *n*. Haario, Saksman and Tamminen (2001) show that if all previous draws (not just a window) are used to compute the covariance used in the proposal, the ergodic properties of the chain are preserved. To simplify computation, a recursive formula is used to update the covariance for the proposal with each draw.

Step 2. Draw *μ*.

$$[\mu | y, X, \{\beta_n\}] \propto \left(\prod_{(n,t)\in \text{data set 2}} [y_{nt} \mid X_{nt}, \beta_n] \right) [\mu]$$
$$\propto \left(\prod_{t\in \text{data set 2}} \frac{\exp(\mu\beta_n x_{n,y_{nt},t})}{\sum_j \exp(\mu\beta_n x_{njt})} \right) \text{Gamma}(\mu \mid 1/1000, 1000)$$

We make this draw using a Metropolis-Hastings step with a normal random walk proposal. Note that this draw depends only on the choice observations for the conjoint data.

Step 3. Draw Δ.

$$[\text{vec}(\Delta) | Z, \{\beta_n\}, \Sigma_\nu] \propto \left(\prod_n [\beta_n \mid z_n, \Delta, \Sigma_\nu] \right) [\Delta]$$
$$\propto N_{KL}(\bar{\mu}_\Delta, \bar{\Sigma}_\Delta)$$

where,

$$\bar{\Sigma}_\Delta = \left(Z'Z \otimes \Sigma_\nu^{-1} + \text{diag}(\text{vec}(F))^{-1} \right)^{-1}$$
$$\bar{\mu}_\Delta = \bar{\Sigma}_\Delta \left((Z' \otimes \Sigma_\nu^{-1}) \text{vec}(\beta') + \text{diag}(\text{vec}(F))^{-1} (\mathbf{0})' \right)$$

where Z is the matrix obtained by stacking the row vectors z_n' and β is the matrix obtained by stacking the row vectors β_n'.

Step 4. Draw Σ_ν.

$$[\Sigma_\nu | Z, \{\beta_n\}, \Sigma_\nu] \propto \left(\prod_n [\beta_n \mid z_n, \Delta, \Sigma_\nu] \right) [\Sigma_\nu]$$
$$\propto IW(K + 2 + N, (K+2)I^{-1} + (\beta - Z\Delta')'(\beta - Z\Delta'))$$

Recall that N is the number of individuals in the sample and K is the number of attributes.

Step 5. For each n, draw z_n.

$$[z_n \mid w_n, \mu_z, \Sigma_z] \propto [\beta_n \mid z_n\Delta, \Sigma_\nu][z_n \mid \mu_z, \Sigma_z]$$
$$\propto MVN(\beta_n \mid z_n\Delta, \Sigma_\nu) MVN(z_n \mid \mu_z, \Sigma_z) I(z_n \in B_n)$$
$$= MVN(z_n \mid \tilde{\mu}, \tilde{\Sigma}) I(z_n \in B_n)$$

$$B_n = B_{n1} \times B_{n2} \times \cdots \times B_{nL} \text{ and } B_{nl} = \begin{cases} (-\infty, 0] & \text{if } w_{nl} = 0 \\ (0, \infty) & \text{if } w_{nl} = 1 \\ (-\infty, \infty) & \text{if } w_{nl} \text{ is missing} \end{cases}$$

$$\tilde{\mu} = \mu_z + \Sigma_z \tilde{\Delta}^T \left(\Sigma_\nu + \tilde{\Delta} \Sigma_z \tilde{\Delta}^T \right)^{-1} \left(\beta_n - \tilde{\Delta} \mu_z \right)$$

$$\tilde{\Sigma} = \Sigma_z - \tilde{\Delta} \Sigma_z^T \left(\Sigma_\nu + \tilde{\Delta} \Sigma_z \tilde{\Delta}^T \right)^{-1} \Sigma_z \tilde{\Delta}^T$$

where $\tilde{\Delta}$ is columns 2 through L of Δ. We sample from this truncated normal distribution by sequentially drawing univariate truncated normal Gibbs samples of each element of z_n (McCulloch and Rossi 1994).

Step 6. Draw μ_z.

$$\begin{aligned} [\mu_z | Z, \Sigma_z] &\propto \left(\prod_n [z_n \mid \mu_z, \Sigma_z] \right) [\mu_z] \\ &\propto N_L(\mu_z \mid \bar{\mu}_z, \bar{\Sigma}_z) \end{aligned}$$

where,

$$\bar{\Sigma}_z = \left(N \Sigma_z^{-1} + (1000I)^{-1} \right)^{-1}$$

$$\bar{\mu}_z = \bar{\Sigma}_z \left(\Sigma_z^{-1} \sum_n z_n + (1000I)^{-1} \mathbf{0} \right)$$

Step 7. Draw Σ_z.

$$\begin{aligned} [\Sigma_z | Z, \mu_z] &\propto \left(\prod_n [z_n \mid \mu_z, \Sigma_z] \right) [\Sigma_z] \\ &\propto IW(v_{\Sigma_z}, S_{\Sigma_z}) \end{aligned}$$

where,

$$v_{\Sigma_z} = L + 2 + N$$

$$S_{\Sigma_z} = (L+2)I + \sum_n (z_n - \mu_z)'(z_n - \mu_z)$$

Appendix B. Parameter recovery study

Data generation

To understand the parameter recovery properties of the model, we simulated data according to the model. The simulated data set had 1100 individuals: 100 from a hypothetical 'conjoint data' set and 1000 from a hypothetical 'market data' set. We generated a vector of 5 latent continuous characteristics for each individual (z_n) according to a multivariate normal distribution with mean $\mu_z = \mathbf{0}$. The variance for each element was 1 and the second and third characteristics had a correlation of 0.4. All other characteristics were independent. We generated choice parameters (β_n) for each individual according to the model $\beta_n = \beta_0 + \Delta z_n + \nu_n$, where z_n is the original vector of length 5. The population parameters Δ and Σ_ν were assumed to be:

$$\beta_0 = \begin{pmatrix} 3 \\ -3 \end{pmatrix} \quad \Delta = \begin{pmatrix} -1 & 1 & 1 & 1 & 1 \\ 0.5 & 2 & 0 & 0 & 0 \end{pmatrix} \quad \Sigma_\nu = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

For each individual in the conjoint study, we generated 100 choice observations from choice sets with three alternatives. The product attributes (x_{njt}) used in each choice observation were generated from independent standard normals, so each choice observation has a unique set of attribute values. We assumed that the scale ratio between data sets was $\mu = 1.5$ and we generated choice observations for each individual in the market data using the choice parameters $\mu\beta_n$. For the market data, we generated 1 choice observation for each individual from a choice set of size twelve. The product attributes for the market data were also generated from independent standard normals. (Note that this is unlikely to be true in real market data where there are often significant correlations between attributes. In this way, our synthetic market data is more informative than real market data is likely to be.)

Base case

To show that we are able to recover the population parameters in circumstances similar to that in the GM minivan data, we ran our estimation algorithm for a 'base case'. In this base case, only 10 choices were observed for each conjoint respondent and the one choice was observed for each market respondent. For the 1000 market respondents, the researcher observed a vector of two binary variables (w_n) indicating whether the first two characteristics (z_{n1} and z_{n2}) were positive or negative. None of the other three continuous characteristics were observed. For the conjoint respondents, we assumed that the researcher observed the binary indicator for just the second characteristic. So, for the conjoint respondents, the first binary indicator was 'missing' and is imputed based on the observed choices and the distribution of the covariates in the market data.

We used diffuse, but proper priors. Inference was based on 20,000 draws from our MCMC algorithm with a burn-in of 6,000 draws. The draws were thinned to every twentieth draw to reduce data storage. Trace plots indicated that the chain had clearly converged after 6,000 draws.

Table 7 shows that the recovery of the population level parameters is quite good for this base case. We should point out that the distribution of heterogeneity for β_n is identified primarily by the conjoint data, where we observe 10 choices for each respondent. With just one observation per respondent, the heterogeneity distribution would be difficult to recover from the market data alone (Andrews, Ainslie and Currim 2002). The joint modeling method is quite useful in situations like this where the conjoint data can provide substantial information about the heterogeneity in β_n that is not available in the market data.

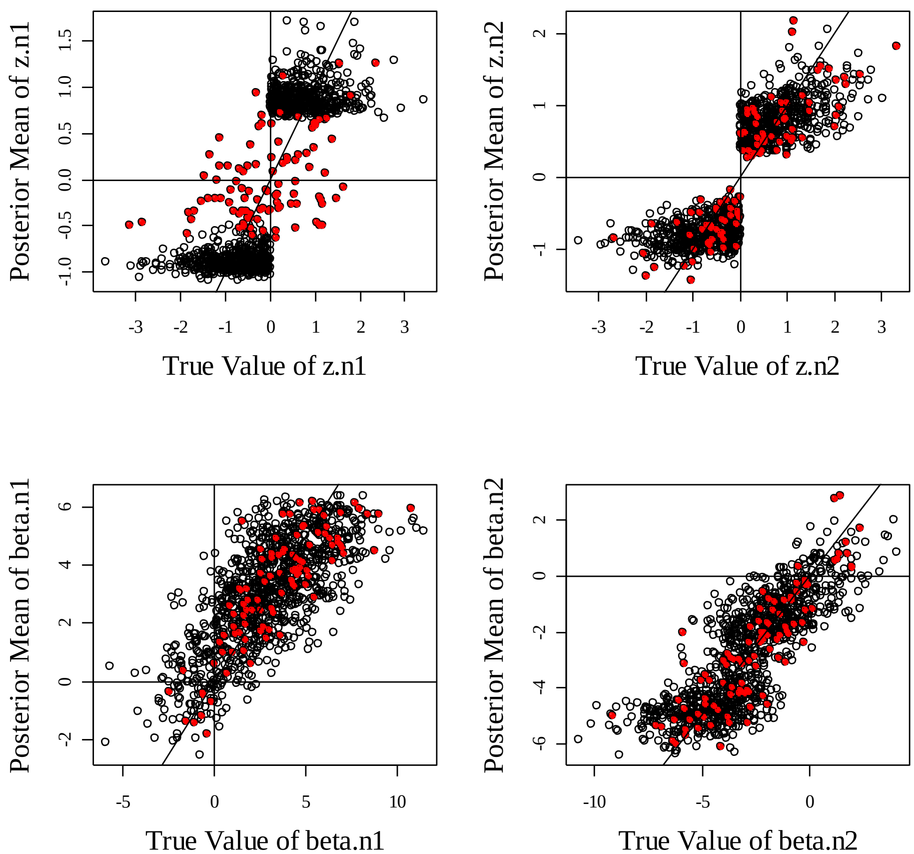
Table 7. Recovery of population-level parameters (Δ, Σ_ν, μ_z and Σ_z).

Parameter	True Value	Posterior Mean	Posterior SD
Delta.11	3.00	3.07	0.18
Delta.21	-3.00	-3.02	0.17
Delta.12	-1.00	-1.08	0.16
Delta.22	0.50	0.54	0.15
Delta.13	1.00	1.34	0.16
Delta.23	2.00	2.23	0.16
mu	0.75	0.74	0.06
Sigma.11	4.00	3.60	0.62
Sigma.12	0.50	0.44	0.35
Sigma.22	1.00	0.83	0.27
mu.w.1	0.00	0.01	0.04
mu.w.2	0.00	0.03	0.04

To understand recovery of the individual-level parameters, z_n and β_n, we computed the posterior means of these parameters for each individual. Figure 2 shows a plot of these posterior means against the true values that were used to generate the data. The closed red circles represent conjoint individuals and the open circles represent market individuals. The top panels in Figure 2 show recovery of z_n. The binary indicator for the second characteristic is observed for all individuals, so the model always predicts the correct sign for z_{n2}. For the first characteristic, z_{n1}, which is not observed for conjoint individuals, there are a number of conjoint individuals for which the mean of the posterior distribution is not the same sign as the true value.

The bottom panels in Figure 2 shows the recovery of β_n. For the conjoint individuals (indicated with closed red circles) the posterior mean is a very good estimate of the true choice parameters. However, we are also able to get reasonably good recovery of the individual-level parameters for the market respondents (open circles), even with just one choice observation per respondent.

Figure 2. Recovery of individual-level characteristics (z_n) and choice parameters (β_n) across all respondents.



Value of w_n in recovering individual estimates of β_n

To demonstrate the value of observing individual characteristics, w_n, we re-estimated the model using none, two (the base case reported above) or five individual-characteristics. Table 8 shows the improvement in individual parameter recovery that is gained by using individual characteristics. For each data set, we compute the mean squared error between individuals' true parameter value and their posterior mean for that parameter. We also compute the average standard error around each individual's posterior mean, indicating how diffuse the individual posteriors are. Table 8 shows that, as the number of individual characteristics included in the estimation is increased, the posterior means of β_n for the market respondents approach their true values, even though we only observe one choice for each market respondent. Thus if β_n is poorly identified by the observed choices, including informative characteristics improves posterior inference for β_n.

Table 8. Recovery of individual-level betas for conjoint and market respondents.

Number of binary individual characteristics observed in market data	Recovery of β_{n1}				Recovery of β_{n2}			
	Conjoint		Market		Conjoint		Market	
	MSE	Avg. SE	MSE	Avg. SE	MSE	Avg. SE	MSE	Avg. SE
0	2.37	1.19	4.04	1.75	1.53	1.05	3.33	1.64
2	2.09	1.30	3.34	1.75	1.40	1.05	1.98	1.42
5	1.71	1.21	2.15	1.47	1.37	1.04	1.92	1.42

Value of observed choices in recovering z_n and missing w_n

More importantly, we also find that *increasing the number of observed choices improves inference about any missing individual characteristics*. Because we use a likelihood-based approach to the missing individual characteristics, our inference about a particular individual's latent characteristics (z_n) is informed both by what we know about other individuals' characteristics *and* the choices we have observed for that individual. In fact, the likelihood of z_n, conditional on the observed data and the other parameters, depends on both the model for the characteristics *and* on the choice parameters, β_n, as follows:

$$[z_n \mid w_n, \mu_z, \Sigma_z, \beta_0, \Delta, \Sigma_\nu] \propto [\beta_n \mid z_n, \beta_{0,} \Delta, \Sigma_\nu][z_n \mid w_n, \mu_z, \Sigma_z]$$

This stands in contrast to other approaches to missing data, such as multiple imputation and hot-deck, where missing characteristics would be imputed based only on information about other respondents' characteristics, ignoring the observed choices of the respondent in question.

To demonstrate the importance of observed choices in imputing the latent z_n and missing elements of w_n, re-estimated the model four times, with three, ten, fifty and one hundred observed choices for each conjoint individual. (Consistent with the base case, the two binary individual characteristics were observed for each market respondent, w_{n1} and w_{n2}, and just one characteristic, w_{n2}, was observed for the conjoint respondents.) Table 9 shows that as the number of observed choices increases, recovery of the latent continuous variable z_{n1} improves for the conjoint respondents. As the number of choice observations increases, the posterior distributions for z_{n1} becomes less diffuse with means closer to the true values. Inference for the second covariate, z_{n2}, for which we observe w_{n2} for the conjoint respondents, also improves slightly.

Table 9. Recovery of individual characteristics for conjoint and market respondents.

Number of choices observed for conjoint respondents	Recovery of z_{n1}				Recovery of z_{n2}			
	Conjoint		Market		Conjoint		Market	
	MSE	Avg. SE	MSE	Avg. SE	MSE	Avg. SE	MSE	Avg. SE
3	0.87	1.08	0.36	0.75	0.35	0.56	0.32	0.59
10	0.79	1.00	0.36	0.70	0.31	0.52	0.32	0.62
50	0.74	1.08	0.37	0.83	0.25	0.56	0.33	0.71
100	0.69	0.93	0.36	0.72	0.22	0.49	0.32	0.63

Appendix C. Parameter estimates for alternative model specifications

Below we provide the posterior means of the population parameters for the models which are presented as alternatives to the joint model in Table 4.

Table 10. Estimated parameters for conjoint model.

		Delta																	
		Other Attributes (X)				Brands (X)													
		Styling Appeal	Price (linear)	Price (squared)	Design Age	A	B	C	D	E	F	G	H	I	J	K	L	M	***mu.z***
Cov	Intercept	**1.38**	**-1.06**	**-0.31**		**0.76**	**0.65**	0.19	-0.09	**1.26**	**-1.61**	-0.26	-0.15	-0.24	0.23	**-0.98**	**1.11**	**-1.04**	NA
	Household.with.Children		-0.05	-0.01															-0.06
Sigma.nu		**1.14**	**2.26**	**0.45**		**2.45**	**3.43**	**2.76**	**2.03**	**3.93**	**2.76**	**1.89**	**2.28**	**3.34**	**2.31**	**3.06**	**6.21**	**6.31**	

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.

Table 11. Estimated parameters for No Individual Characteristics (NIC) formulation.

mu	*Delta*																
0.76	Other Attributes (X)				Brands (X)												
	Styling Appeal	Price (linear)	Price (squared)	Design Age	A	B	C	D	E	F	G	H	I	J	K	L	M
Intercept	**1.86**	**-1.18**	**-0.23**	**1.07**	**1.27**	**1.18**	0.23	-0.10	**1.55**	**-1.94**	**-1.01**	-0.22	-0.21	0.15	-0.51	**0.94**	**-1.43**
Sigma.nu	**2.03**	**3.88**	**0.71**	**4.54**	**6.12**	**6.72**	**4.26**	**2.51**	**8.08**	**6.59**	**4.06**	**2.51**	**5.06**	**2.44**	**5.24**	**7.38**	**6.89**

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.

Table 12. Estimated parameters for Joint model with less market data used in estimation.

mu **0.68**	*Delta*: Other Attributes (X)				*Delta*: Brands (X)													
Covariates (z)	Styling Appeal	Price (linear)	Price (squared)	Design Age	A	B	C	D	E	F	G	H	I	J	K	L	M	***mu.z***
Intercept	**2.25**	**-1.40**	**-0.27**	**0.86**	**3.55**	**2.03**	0.38	0.20	-0.33	**-5.19**	**-2.56**	-0.01	-1.88	**2.13**	0.71	-0.32	**-1.42**	**NA**
Household.Income.<75		**-1.12**	0.13															**0.11**
Household.with.Children		**-0.62**	-0.02															**-0.08**
Reason.Usability					-0.33	0.38	0.38		-0.80	0.54	-0.20	0.31	-0.34	0.29	-0.27	-0.23		**0.67**
Reason.Dependability					-0.58	-0.60	-0.45		**1.81**	0.30	**0.94**	-0.35	0.94	**-0.98**	-0.67	**1.34**		**0.35**
Reason.Rugged_AWDRWD					-0.20	-0.16	-0.38		-0.58	-0.66	-0.25	-0.40	-0.66	**0.97**	**1.05**	-0.05		**-1.18**
Reason.Dealer					-0.04	0.13	**0.84**		-0.65	0.15	-0.08	**0.91**	-0.72	-0.02	**0.75**	**-1.45**		**0.14**
Reason.Warranty					0.18	**0.91**	**-0.90**		-0.04	**2.59**	0.52	**-1.12**	-0.23	0.38	**-1.03**	**-0.81**		**-0.11**
Reason.Roominess					-0.24	-0.37	-0.69		0.48	0.14	**0.92**	-0.44	0.53	-0.53	0.24	0.08		**0.73**
Reason.Cargo_Versatility					0.71	0.47	0.53		-0.30	**-1.19**	-0.58	-0.04	-0.18	0.22	0.46	0.15		0.05
Reason.FuelEcon_Value					**-0.61**	-0.05	**-1.18**		0.45	0.71	**0.88**	**-0.67**	0.46	-0.43	0.18	0.60		**0.22**
Reason.Incentives					0.44	0.46	**1.51**		**-2.02**	0.06	**-0.67**	**1.48**	**-2.31**	**1.37**	**1.13**	**-2.55**		**-0.18**
Reason.DrivePerform					0.54	-0.08	0.63		-0.51	-0.95	-0.58	0.15	-0.93	**0.88**	0.04	0.02		**0.23**
Reason.No_Negotiation					0.19	**-0.83**	-0.24		0.18	0.58	0.29	0.02	0.41	-0.46	-0.09	-0.36		**-0.39**
Reason.Luxury					0.47	-0.37	-0.44		0.04	-0.17	0.15	0.05	**0.97**	-0.28	-0.61	0.66		**-0.67**
Reason.Tow_Haul					0.26	0.12	0.32		-0.45	0.43	-0.12	0.16	0.03	-0.21	-0.66	-0.12		**-1.53**
Reason.Safety_Security					-0.66	-0.46	**0.97**		-0.14	0.08	-0.41	0.79	0.34	**-0.91**	-0.75	0.64		**0.38**
Reason.AvoidCollision					-0.62	-0.88	0.15		0.21	-0.26	-0.37	**1.06**	0.52	0.31	0.43	0.28		**-0.40**
Reason.KidFeatures					-0.06	0.39	-0.42		0.51	-0.11	0.09	**-1.15**	**0.91**	-0.52	-0.14	0.64		-0.01
Reason.ExteriorStyling					-0.23	-0.20	0.00		**-0.94**	0.29	0.17	-0.12	0.05	0.43	0.61	-0.42		**0.30**
Reason.FuntoDrive					**-0.91**	-0.37	-0.40		-0.19	0.74	**0.85**	-0.31	**1.48**	-0.22	0.46	-0.48		**-0.41**
Reason.CountryofOrigin					0.09	0.51	-0.30		-0.51	**-1.48**	-0.69	0.04	-0.55	**0.93**	**1.16**	-0.30		**-0.88**
Reason.Practical					-0.10	-0.17	0.28		-0.19	0.54	0.03	-0.05	-0.50	0.21	0.57	-0.60		**0.09**
Reason.Environment					**0.86**	0.43	0.42		0.41	-0.67	-0.49	0.19	**-1.02**	-0.19	-0.18	0.13		**-0.98**
Reason.MfgReputation					0.27	0.47	-0.31		**1.22**	**-1.79**	-0.79	-0.46	-0.25	0.62	-0.64	**1.26**		**0.21**
Reason.InteriorStyling					0.34	-0.12	-0.18		-0.16	0.81	-0.08	0.47	0.36	-0.28	-0.38	0.11		**0.22**
Reason.WillingtoNegotiate					0.26	**0.58**	-0.26		0.22	0.11	0.05	-0.08	-0.33	0.00	-0.15	-0.16		**-0.48**
Reason.CargoLoading					0.05	0.31	-0.18		0.15	-0.18	0.60	-0.29	0.39	-0.52	-0.43	0.23		**-0.36**
Sigma.nu	**2.23**	**3.15**	**0.67**	**1.49**	**4.83**	**6.45**	**2.15**	**0.86**	**1.85**	**1.80**	**2.31**	**1.40**	**2.25**	**1.33**	**3.44**	**1.34**	**5.59**	

* Values in boldface have a posterior mean more than two posterior standard errors different than zero.

References

Allenby, Greg M. and James L. Ginter. 1995. Using extremes to design products and segment markets. Journal of Marketing Research, 32(4), pp 392-349.

Andrews, Rick L., Andrew Ainslie and Imran S. Currim. 2002. An Empirical Comparison of Logit choice Models with Discrete Versus Continuous Representations of Heterogeneity. Journal of Marketing Research, 39(11), pp 479-487.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. 1994. Combining Revealed and Stated Preferences Data. Marketing Letters, 5(4), pp 335-350.

Ben-Akiva, Moshe and Takayuki Morikawa. 1990. Estimation of switching models from revealed preferences and stated intentions. Transportation Research A, 24A(6), pp 485-495.

Bhat, Chandra R. and Saul Castelar. 2002. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. Transportation Research B, 36, pp 593-616.

Blamey, Russell and Jeff Bennett. 2001. Yea-saying and Validation of a Choice Model of Green Product Choice. in Jeff Bennett and Russell Blamey, eds., The Choice Modeling Approach to Environmental Valuation. Edward Elgar, Northampton, MA, pp178-201.

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

Brooks, Stephen P. and Andrew Gelman. 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7(4), pp 434-455.

Brownstone, David, David S. Bunch and Kenneth Train. 2000. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research B, 34, pp 315-338.

Chib, Siddhartha and Edward Greenberg. 1998. Analysis of Multivariate Probit Models. Biometrika, 85(2), pp 347-361.

Ding, Min, Raj Grewal and John Liechty. 2005. Incentive-Aligned Conjoint Analysis. Journal of Marketing Reearch, 42(1), pp 67-82.

De Bruyn, Arnaud, John C. Liechty, Eelko K.R.E. Huizingh and Gary L. Lilien. 2008. Offering Online Recommendations with Minimum Customer Input through Conjoint-Based Decision Aids, Marketing Science, articles in advance.

Dominici, Francesca, Giovanni Parmigiani, Kenneth H. Reckhow and Robert L. Wolpert. 1997. Combining Information from Related Regressions. Journal of Agricultural, Biological and Environmental Statistics, 2(3), pp 313-332.

Fennell, Geraldine, Greg M. Allenby, Sha Yang and Yancy Edwards. 2003. The Effectiveness of Demographic and Psychographic Variables for Explaining Brand and Product Category Use. Quantitative Marketing and Economics, 1(2), pp 223-244.

Gilbride, Timothy J., Peter J. Lenk and Jeff D. Brazell. 2008. Market Share Constraints and the Loss Function in Choice Based Conjoint Analysis. Marketing Science, 27(6), 995-1011.

Green, Paul E., Abba M. Krieger and Yoram (Jerry) Wind. 2001. Thirty Years of Conjoint Analysis: Reflections and Prospects. Interfaces, 31(3), pp S56 – S73.

Green, Paul E. and Vithala R. Rao. 1971. Conjoint Measurement for Quantifying Judgmental Data. Journal of Marketing Research, 8(3), pp 355-363.

Haario, Heikki, Eero Saksman, and Johanna Tamminen. 2001. An adaptive Metropolis algorithm. Bernoulli, 7(2), pp 223-242.

Heckman, James J. 1979. Sample Bias As A Specification Error. Econometrica, 1979, 47(1), pp 153-162.

Ibrahim, Joseph G., Stuart R. Lipsitz and Ming-Hui Chen. 1999. Missing characteristics in generalized linear models when the missing data mechanism is non-ignorable. Journal of the Royal Statistical Society, 61(1), pp 173-190.

Krishnan, V. and Karl T. Ulrich. 2001. Product Development Decisions: A Review of the Literature. Management Science, 47(1), pp 1-21.

Little, Roderick J.A. and Donald B. Rubin. 2002. Statistical Analysis with Missing Data, Second Edition. Wiley Interscience, Hoboken, New Jersey.

Louviere, Jordan J., David A. Hensher and Joffre D. Swait. 2000. Stated Choice Methods: Analysis and Application. Cambridge University Press, Cambridge, UK.

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

Manski, Charles F., Steven R. Lerman. 1977. The Estimation of Choice Probabilities from Choice Based Samples. Econometrica, (45)8, pp 1977-1988.

McCulloch, Robert and Peter E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics, 64, pp 207-240.

Newton, Michael A. and Adrian 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.

Orme, Bryan and Rich Johnson. 2006. External Effects Adjustments in Conjoint Analysis. Sawtooth Software Research Paper Series.

Rossi, Peter E., Greg M. Allenby and Rob McCulloch. 2005. Bayesian Statistics and Marketing. John Wiley and Sons, West Sussex, England.

Sandor, Zsolt and Michel Wedel. 2005. Heterogeneous Conjoint Choice Designs. Journal of Marketing Research, 42(2), pp 210-218.

Swait, Joffre and Rick L. Andrews. 2003. Enriching scanner panel models with choice experiments. Marketing Science, 22(4), pp 442-460.

Swait, Joffre and Jordan Louviere. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research, 30, pp 305-314.

Tanner, Martin A. and Wing Hung Wong. 1987. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82(398), pp 528-540.

Urban, Glen L., John R. Hauser, John H. Roberts. 1990. Prelaunch Forecasting of New Automobiles. Management Science, 36(4), pp 401-421.

Zeithammer, Robert and Peter Lenk. 2006. Bayesian estimation of multivariate-normal models when dimensions are absent. Quantitative Marketing and Economics, 4(3), pp 241-265

Chapter 3

Essay 2: Error scale heterogeneity in the hierarchical multinomial logit model

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.

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 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 the 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.

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,

$$u_j = x_j{}'\beta + \varepsilon_j , \qquad \textbf{(8)}$$

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

$$\left[y_{it} = j^* \mid \beta, \lambda, \{x_j\}\right] = \frac{\exp\left(x'_{j^*}\left(\beta / \lambda\right)\right)}{\sum_j \exp\left(x'_j\left(\beta / \lambda\right)\right)} \qquad \textbf{(9)}$$

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 (8) 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 (9) 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 thedimension 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,

$$\begin{aligned} \left[\beta_i \mid \Delta, \Sigma\right] &= N_K\left(\beta_i \mid \beta_0 + z_i'\Delta, \Sigma\right) \\ \left[\log(\lambda_i) \mid \delta, \sigma\right] &= N\left(\log(\lambda_i) \mid z_i'\delta, \sigma^2\right) \end{aligned} \qquad \textbf{(10)}$$

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 (10) 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 is obtained (cf., Train 2003).

We will refer to the model proposed in equations (9) and (10) 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 (10).

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,

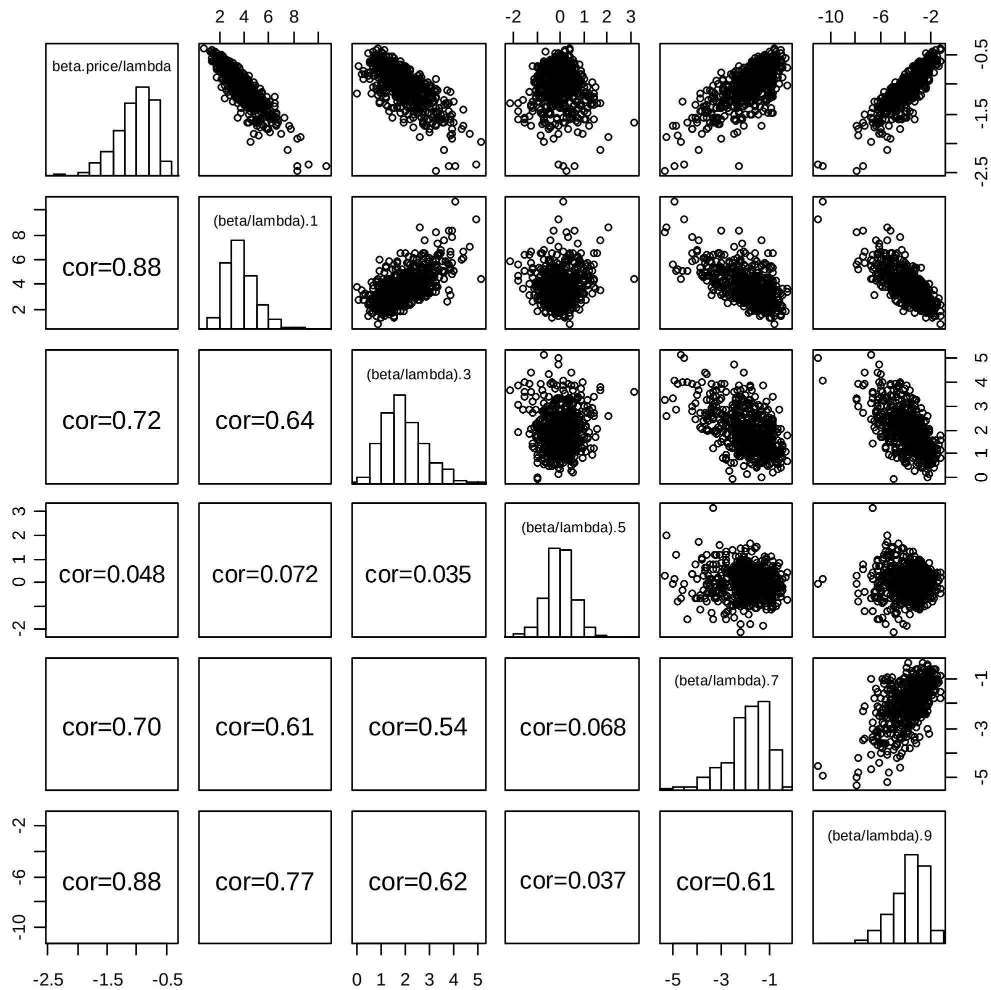
$$\operatorname{var}\left(\frac{\beta_i}{\lambda_i}\right) = E\left(\operatorname{var}\left(\frac{\beta_i}{\lambda_i}\middle|\lambda_i\right)\right) + \operatorname{var}\left(E\left(\frac{\beta_i}{\lambda_i}\middle|\lambda_i\right)\right)$$

$$= E\left(\frac{1}{\lambda_i^2}\right)\operatorname{var}(\beta_i) + \operatorname{var}\left(\frac{1}{\lambda_i}\right)E(\beta_i)E(\beta_i)' \tag{11}$$

The second term in the last line of equation (11) 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 3 shows an example set of synthetic values for β_i / λ_i generated according to the population distribution in equation (10) 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, β_i. For Figure 3, 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 3 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 3. 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 (10) 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 3, if not the skewness and, as we will show, may represent a reasonable prior on individual-level parameters. But, it is clear from Figure 3, that if λ_i is heterogeneous across the population, the distribution of β_i / λ_i would not be well modeled with specification that does not allow for correlations between the attribute coefficients, i.e., a model where $\lambda_i = 1 \ \forall i$ and Σ is restricted to be diagonal. We will refer to this model with diagonal covariance as N-MNL.

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 indivdiual-level coefficients and error scale values.

Related Models

Type I and type II generalized multinomial logit model. The model proposed in equations (9) and (10) 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.

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 (9) and the joint multivariate normal distribution for $(\log(\lambda_i), \beta_i)$. Full details of the sampling algorithm are included in Appendix A.

Identification of the model

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[8]. Individual-level parameters β_i were generated as in Figure 3.[9]

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,

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

[9] 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.

information for each individual is increased by increasing the number of choices for each individual (cf., Louviere, et al. 2008b). Table 13 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.)

We also found a general pattern that the posterior modes of the population parameters are biased outwards; that is, elements of β_0 and $\mathrm{diag}(\Sigma)$ have posterior modes that are greater in absolute value than the values used to generate the data. We summarize this "outward bias" as $\left(\mathrm{mode}([\beta_{0,i} \mid \mathrm{data}]) - \beta_{0,i}^{true}\right)\mathrm{sgn}(\beta_{0,i}^{true})$ in Table 13. 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 $\mathrm{diag}(\Sigma)$. 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 13. 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 13 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 13 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.

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

$$L^I = \prod_i \left(\prod_t \left[y_{it} \mid \beta_i, \lambda_i, \{x_{ijt}\} \right] \right) \quad \textbf{(12)}$$

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 (12).

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

$$L^P = \prod_i \int_{\beta_i} \int_{\lambda_i} \left(\prod_t \left[y_{it} \mid \beta_i, \lambda_i, \{x_{ijt}\} \right] \right) \left[\beta_i \mid \beta_0, \Delta, \Sigma, z_i \right] \left[\lambda_i \mid \delta, \sigma^2, z_i \right] d\lambda_i d\beta_i \quad \textbf{(13)}$$

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[10]. 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 (12). 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

[10] 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.

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).

Table 14 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.[11] 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.

---

[11] Note that each cell in Table 14 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 14. Model fit statistics computed based on the individual-level likelihood do not distinguish different specifications of the population-level model.

Data generaged according to			Model Estimated			
			S-MNL	**N-MNL**	**MVN-MNL**	**G-MNL**
	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 14 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 15. 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 (13). The integral in equation (13) 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.[12] Average 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 15 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 (11) and Figure 3, 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 16). We suggest that researchers who estimate MVN-MNL models should check estimates of Σ, to

[12] 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.

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 16. 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.)

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.

Bathroom Scale Choice Experiment

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).

The parameter recovery studies reported in the previous section suggested this data set would be a sufficient size to estimate the G-MNL model (see Table 13), however 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[13]. 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: $\Sigma \sim \mathrm{IW}(100, I) \quad \sigma^2 \sim \Gamma^{-1}(0.05,100)$. All posterior estimates are based on chains of length 200,000 with a burn-in of 10,000.

[13] 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.

Model comparisons. Table 17 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 G-MNL model only has 65 population-level parameters 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 17. 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 18, Table 19, and Table 20 compare the population-level parameters for the four estimated models. Table 18 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 19 and Table 20 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 error scales for the estimate of the population means of the attribute preferences.

Table 18. 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 19. 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 20. 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 17 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 21, 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 21. 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

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: $\Sigma \sim \mathrm{IW}(K+2, I) \quad \sigma^2 \sim \Gamma^{-1}(0.01, 100)$. 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 22). 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 22. 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 23 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 23. 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 24, 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 24. 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

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.

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.

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

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

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

$$=\prod_i\left(\prod_t\frac{\exp\left(x'_{ijt}\left(\beta_i/\lambda_i\right)\right)}{\sum_{\tilde{j}}\exp\left(x'_{\tilde{i}\tilde{j}t}\left(\beta_i/\lambda_i\right)\right)}\right)N_k(\beta_i\mid\beta_0+z_i'\Delta,\Sigma)N(\lambda_i\mid z_i'\delta,\sigma^2)$$

Priors

$$\left[(\beta_0,\Delta)\right]=N_{k(l+1)}\left(vec((\beta_0,\Delta))\mid\mu_\Delta^{pr},\Sigma_\Delta^{pr}\right)$$

$$\left[\Sigma\right]=IW\left(\Sigma\mid v_\Sigma^{pr},S_\Sigma^{pr}\right)$$

$$\left[\delta\right]=N_l(\delta\mid\mu_\delta^{pr},\Sigma_\delta^{pr})$$

$$\left[\sigma^2\right]=\Gamma^{-1}(\sigma^2\mid s_\sigma^{pr},S_\sigma^{pr})$$

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.

$$[(\beta_0, \Delta) \mid \Sigma, \{\beta_i\}, \{z_i\}] \propto \left(\prod_i [\beta_i \mid z_i, \beta_0, \Delta, \Sigma] \right) [(\beta_0, \Delta)]$$

$$= N_k\left(vec((\beta_0, \Delta)) \mid \mu', \Sigma'\right)$$

$$\Sigma' = \left(\left(Z' Z \otimes \Sigma^{-1} \right) + S_{\Delta}^{pr^{-1}} \right)^{-1}$$

$$\mu' = \Sigma' \left(\left(Z' \otimes \Sigma^{-1} \right) vec(\alpha) + \Sigma_{\Delta}^{pr^{-1}} \mu_{\Delta}^{pr} \right)$$

$$Z = \begin{pmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix}$$

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

$$[\Sigma \mid \beta_0, \Delta, \{\beta_i\}, \{z_i\}] = \left(\prod_i [\beta_i \mid z_i, \Delta, \Sigma] \right) [\Sigma]$$

$$= IW\left(\Sigma \mid v_{\Sigma}^{pr} + n, v_{\Sigma}^{pr} S_{\Sigma}^{pr} + \left(\beta - Z(\beta_0, \Delta)\right)' \left(\beta - Z(\beta_0, \Delta)\right)\right)$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

3. Draw σ from

$$[\sigma^2 \mid \delta, \{\lambda_i\}, \{z_i\}] = \left(\prod_i [\lambda_i \mid \delta, \sigma^2, z_i] \right) [\sigma^2]$$

$$= \Gamma^{-1}\left(\sigma^2 \mid n + v_{\sigma}^{pr}, (\lambda - Z'\delta)'(\lambda - Z'\delta) + v_{\sigma}^{pr} S_{\sigma}^{pr}\right)$$

$$\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \quad Z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

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

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

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

$$= \frac{\exp\left(x'_{ijt}\left(\beta_i \big/ \lambda_i\right)\right)}{\sum_{\tilde{j}} \exp\left(x'_{i\tilde{j}t}\left(\beta_i \big/ \lambda_i\right)\right)} N_k\left(\beta_i \mid \beta_0 + z_i'\Delta, \Sigma\right) N(\lambda_i \mid z_i'\delta, \sigma^2)$$

$$= \frac{\exp\left(x'_{ijt}\left(\beta_i \big/ \lambda_i\right)\right)}{\sum_{\tilde{j}} \exp\left(x'_{i\tilde{j}t}\left(\beta_i \big/ \lambda_i\right)\right)} N_{k+1}\left(\alpha_i = (\log(\lambda_i), \beta_i) \mid \begin{pmatrix} z_i'\delta \\ \beta_0 + z_i'\Delta \end{pmatrix}, \begin{matrix} \sigma^2 & 0 \\ 0 & \Sigma \end{matrix}\right)$$

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 13. 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 25) versus G-MNL (Table 13), 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 25. 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_t	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_t	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_t	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 26. 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	
Intercept	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
	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
PC Owner	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
	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
Gender	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
	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
Age	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
	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
Expert Buyer	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
	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 27. 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

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.

Chapter 4

Essay 3: Designing choice experiments to maximize the profitability of new product designs

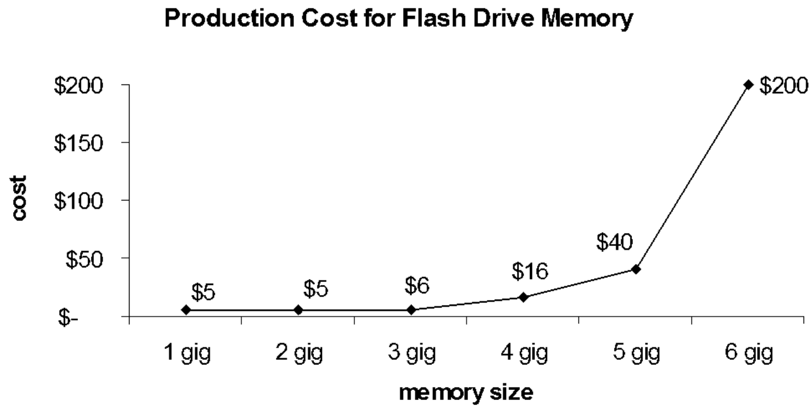
Introduction

Despite the steady advances in the design of choice experiments (Kuhfeld, Tobias and Garratt 1994, Zwerina, Huber and Kuhfeld 1996, Huber and Zwerina 1996, Arora and Huber 2001, Sandor and Wedel 2001, 2002 and 2005, Kessels, Goos and Vandebroek 2006, Liu, Dean and Allenby 2007, Liu, Dean, Bakken and Allenby 2009), this literature has focused on determining the set of choice tasks that will maximize statistical efficiency measures. These efficiency measures summarize the amount of information that is collected in the experiment about the model parameters. For example, the commonly used D-efficiency measure is defined as

$$\mathrm{D}_{\mathrm{eff}} = \left| I(\beta; X) \right|^{1/K} \quad \textbf{(14)}$$

where I is the expected Fisher Information matrix for the model parameters that will be estimated using the data. In the case of non-linear models like the multinomial logit and multinomial probit typically used to estimate choice models, this matrix depends on assumptions about the model parameters, β, and on the attributes of the alternatives in the survey questions, X. Like nearly all statistical efficiency measures, D-efficiency assigns equal weight to the information obtained (i.e., expected posterior variance) for each of the model parameters, regardless of whether that is appropriate for the problem the analyst wishes to address.

Figure 4. Hypothetical production cost information.



Although statistical efficiency has been the workhorse of experimental design, it ignores the fact that managerial decisions are not based directly on the estimates of the parameters and the quality of managerial decisions is not directly related to the accuracy of parameter estimates (Toubia and Hauser 2007). Statistical efficiency also ignores valuable information about the costs of producing alternative products, information that is typically available to the firm. For example, consider a firm that is designing a flash-memory drive. The firm will typically have access to a production cost function similar to that in Figure 4. Intuitively, since the cost of producing a 2 gig drive is the same as the cost of producing a 1 gig drive, there is no managerial need to accurately measure the preference for 1 gig drives relative to 2 gig drives (assuming we have prior information that larger drives are always preferred). Similarly, we may be able to rule out a 6 gig drive as prohibitively expensive to produce, depending on prior information about which prices might be reasonable for consumers. Choice experiments that are designed to optimize statistical efficiency criteria like D-efficiency ignore valuable cost information like that in Figure 4 and favor measuring including an equal number of profiles at each proposed level of the attribute (Huber and Zwernia 1996). A statistically efficient design would likely ask the same number of questions about 1 and 6 gig drives as about 4 and 5 gig drives, ignoring the fact that *a priori* the firm is much more likely to arrive at a 4 or 5 gig design than a 1 or 6 gig design. Although managers sometimes exclude some irrelevant attribute levels from the choice experiment in an *ad hoc* way, there is no formal method that incorporates this cost information to guide the design of a choice experiment towards attribute levels that are likely to be in the optimal product design.

Further, there are very few approaches to designing choice experiments that comprehend that there are *combinations* of attributes that are expensive or infeasible to produce (e.g., a computer that is both fast and lightweight or a car that is both roomy and fuel efficient). Experimental designs based on statistical efficiency will often present respondents with product alternatives that are impossible to produce. This potentially wastes valuable respondent time with questions about infeasible alternatives. Respondent time would be better spent exploring regions of the design space that are more likely to produce feasible, profitable product designs. Worse yet, presenting respondents with product profiles that are radically different than the products that they are familiar with from the marketplace (as will happen with designs based on statistically efficiency) may cause the respondent to react differently to them than they would in the marketplace, potentially leading to a model that does not predict market behavior well.

Bayesian decision-theory (cf. Berger 1980) provides an alternative framework for designing an experiment that incorporates design constraints and cost information in a natural way. When firms undertake marketing research, their goal is to gather information that will inform their decisions. Given this, it seems reasonable to gauge the quality of a particular choice experiment based on how well the results of the experiment will inform the decision. Following a decision theoretic approach, a firm designing a product should quantify their profit function (or loss function as it is called in the decision theory terminology) and then choose the questions to maximize expected profit, presuming that they will make the best decision they can after recieving the results of the experiment. In this essay, we recast the design of choice experiments as a decision theoretic optimization problem and outline some of the computational challenges related to solving this optimization problem. We conclude by discussing the potential benefits of this new approach to the design of choice experiments and outlining future research aimed at solving the computational challenges associated with the optimization problem.

Decision theoretic conjoint design criterion

To make the decision theoretic approach concrete, we introduce some notation. Assume a firm is choosing a vector of product attributes, *x,* to be sold at a price, *p*. The profit of this product design depends on the demand and the costs associated with a particular set of attributes, *x*. Suppose that we have a per-unit production cost function, $C(x)$, and a statistically estimated demand function, $D(p, x \mid \beta)$, with unknown parameters *β*. The profit that the firm hopes to maximize is

$$\prod(p,x \mid \beta) = D(p, x \mid \beta) \times (p - C(x)) \qquad \textbf{(15)}$$

The functions $D(p, x \mid \beta)$ and $C(x)$ could take a variety of forms. If the attributes are restricted to take on discrete levels, as is typically the case in choice experiments, $C(x)$ could be a simple look-up table indicating the production cost for any combination of attributes. Infeasible combinations of attributes could be assigned prohibitively high costs (i.e., following a "big-*M*" approach). This table could be populated by running existing engineering feasibility models (c.f, Michalek 2005) or simply by asking engineers to make subjective judgments about the costs of alternative configurations. The demand function could take on any number of forms, but for simplicity we will assume that $D(p, x \mid \beta)$ is an aggregate multinomial logit model, conditional on parameter *β,* scaled by an estimate of the total population of buyers. For simplicity, the profit function in (15) ignores differences in investment costs across alternative product designs (e.g., additional engineering, testing or tooling associated with particular combinations of attributes), however these costs could readily be included.

Given the profit function, the choice experiment should be designed to maximize the expected profit posterior to the experiment. For example, if we were designing a single question where respondents would choose between two products with prices and attributes (p_1, x_1) and (p_2, x_2), then p_1, x_1, p_2 and x_2 should be chosen to maximize

$$\sum_{i=1}^{2}\left(E_{\beta}(y=i)\times\left(\max_{p,x} E_{\beta|y=i,p_1,x_1,p_2,x_2}\left(\Pi(p,x|\beta)\right)\right)\right) \qquad \textbf{(16)}$$

where $E_{\beta}(y=i)$ is the prior expectation that alternative i will be chosen (where the expectation is take over *prior* beliefs about β) and $\max_{p,x} E_{\beta|y=i,p_1,x_1,p_2,x_2}\left(\Pi(p,x|\beta)\right)$ is the expected profit associated with the best product design (where the expectation is taken over the *posterior* distribution of β conditional on i being chosen). The criterion in (16) is an alternative to D-efficiency that logically incorporates the cost information available to the firm along with the firm's prior beliefs about the parameters.

An illustrative example

We illustrate the implications of the decision theoretic design criteria using a hypothetical product design problem. The simplified structure of this problem makes it possible to solve the decision theoretic conjoint design problem by brute force.

Assume that an automotive manufacturer is trying to determine whether a vehicle that it is designing should include a DVD player (production cost = \$100), a navigation system (\$100), or both. Since DVD and navigation systems can share the video display, the production cost of including both of them is assumed to be \$105. The base price of the vehicle will be increased by some amount (between \$0 and \$400) to account for the additional content. Assume that the demand follows a multinomial logit model with dummy codes for the four possible configurations of the product (no additional content, DVD player, navigation system, or navigation system and DVD player) and dummy codes for five levels of price ($0, $100, $200, $300, or $400). To estimate total take rates for the target product, the model includes an outside good with utility normalized to zero and assumes that the total population of potential buyers is 10,000. Following equation (15) the ultimate goal of the manufacturer is to determine the configuration and price that will maximize total profit. A choice experiment will be conducted to learn more about the parameters of the demand function, β, before choosing the final product design.

For this illustration, we simplify the computation by using a discrete prior over the parameters of the multinomial logit model. Assume that there are two possible states of nature, one where consumers prefer vehicles with DVD players and one where consumers actually dislike DVD players prefer vehicles without DVD players. The parameters associated with each state of nature are given in Table 28. Under the prior, both states of nature are assumed to be equally likely.

Table 28. Hypothetical states of nature for illustrative example

Attribute	**Level**	**Part-Worth**	
		State of Nature A	**State of Nature B**
Features	Base	0	0
	DVD	2	-2
	Navigation	2	2
	Navigation & DVD	4	0
Price	$0	0	0
	$100	-1	-1
	$200	-2	-2
	$300	-3.5	-3.5
	$400	-6	-6

Prior to observing any market research, the design that optimizes expected profit is a vehicle with navigation and DVD player for an additional price of $300. This design would produce a profit of $1,210K under state of nature A (people like DVD players) but would only produce a profit of $57K under state of nature B. The resulting expected profit is $635K.

For the purposes of illustration, we assume that the firm is planning to conduct a simplified choice experiment where one respondent will answer a single binary choice question. There are six questions the firm could use that optimize the decision theoretic criterion in (16). One of those optimal questions is a choice between alternative 1 with navigation system for $300 and alternative 2 with a DVD player for $100. Intuitively, by offering a product with a cheap DVD player we can discern whether people like or dislike DVD players. If the respondent chooses alternative 1, then the conditional posterior probability that respondents like DVD players (state A) is 0.08. If the respondent chooses alternative 2, then this probability is 0.84. Thus the question clearly differentiates between the two states of nature. This question also displays a feature that is similar to the “utility balance” criterion that has been proposed in the literature on design of choice experiments (Huber and Zwerina 1996); the prior probability of choosing alternative 1 over 2 is 0.45, very close to equal probabilities for the two alternatives.

In computing the optimal question design, we also determine the optimal product designs conditional on the two possible outcomes of the experiment. If the respondent chooses alternative 1 then the optimal design includes navigation for $200 with an expected profit of $500K, but if the respondent chooses alternative 2 then the optimal design includes navigation and DVD for $300 with an expected profit of $1,023K. Thus, it is explicit exactly how the results of the experiment will be used to drive decision making. Without conducting the experiment, the expected profit was $635K, but conditional on conducting the experiment (and behaving optimally based on the outcome) the expected profit is 0.45*$500K +0.55*$1,023K = $788K.

In this illustrative example, we can see some of the benefits of the decision theoretic approach. First, using the objective in (16), we can actually estimate how much the experiment will improve expected profit. In this case, the experiment is expected to improve expected profit by $153K. A manager could use this information to determine whether value of the study exceeds the cost. In contrast, statistical efficiency criteria, which focus on parameter estimates, don't provide any estimate of the *economic* value of the choice experiment. Second, statistical efficiency measures may not make distinctions between alternative questions when only one or a few questions will be included in the choice experiment. In the above example, we found many questions that have the same D-efficiency as six questions that are optimal according to the decision-theoretic criteria, but have no impact on the expected profit of the ultimate product design. This limitation of D-efficiency and other statistical criteria is likely to be even more pronounced in more realistic problems, especially under the diffuse priors often used in practice. A potential strength of the decision theoretic approach is that it can be used to design a very short questionnaire suitable for web-based surveys. One might even envision extending the approach to a sequential survey design process (either using a greedy heuristic or a full dynamic programming approach), where subsequent questions are designed based on prior responses resulting in an adaptive survey that can be exited at any time by the respondent.

This example also demonstrates some potential problems with the decision theoretic approach. First, it is highly dependent on the prior beliefs about the model parameters. In this example, there was greater uncertainty in beliefs about DVD player preferences, so the decision theoretic design focused on gathering information about preferences for DVD players. While some may see this as a weakness in the approach, it can also be viewed as a potential strength. In practice managers often *do* have prior information about consumer preferences and our approach provides a natural way to include that information in the market research and product design process. Second, by going through the calculations involved in this simple problem, we can see that there is a great deal of computation required to evaluate survey designs against the decision theoretic criteria. We discuss those computational issues next.

Computational approaches for experimental design problem

While conceptually straightforward, the criterion in equation (16) is computationally challenging to evaluate. Simply evaluating (16) for a particular conjoint design involves sampling over the posterior of β and solving the product design optimization problem, both of which are research problems in their own right (cf., Rossi Allenby and McCulloch 2005 and Simpson, Saddique and Jiao 2006 respectively). Figure 5 summarizes the steps involved in solving decision theoretic conjoint design problem as stated in equation (16). By going through the steps in detail, we hope to illuminate some of the computational challenges.

Figure 5. Algorithm for solving the decision theoretic conjoint design problem

For a candidate conjoint design, e.g., (x_1, p_1), (x_2, p_2)	1
Compute the expected profit conditional on the conjoint design	2
For each outcome y = i, e.g., y = 1, 2	3
Compute the sampling probability $E_\beta(y=i)$	4
Compute the posterior of β conditional on $y=i$, *i.e,.* $\beta \vert y=i$, (x_1, p_1), (x_2, p_2)	5
For a candidate product design (x, p)	6
Compute expected profit (expectation over $\beta \vert y=i$, (x_1, p_1), (x_2, p_2))	7
Search candidate product designs to find maximum	8
Next outcome of conjoint experiment	9
Take the weighted sum of expected profit over the sampling probabilities	10
Search candidate conjoint designs to find maximum	11

Conjoint design problem

The outer optimization loop (lines 1 and 11) involves searching over the space of potential survey questions for the design that maximizes profit. The total possible number of conjoint designs is $\left(\prod_k l_k\right)^{qa}$ where q is the number of questions in the experiment, a is the number of alternatives in each question, k indexes the attributes (including price), and l_k is the number of levels for attribute k. Even for our example in the previous section, the total number of possible designs was $(4*5)^{1*2} = 400$ and this number grows rapidly with more realistic numbers of question and attributes. Despite the fact that this is a large discrete optimization problem, a number of efficient heuristics have been proposed for solving this optimization problem including cycling and swapping heuristics (Huber and Zwerina 1996) and the modified Federov algorithm (Sandor and Wedel 2005). Since evaluating the decision theoretic criteria is extremely computationally intensive, this outer optimization heuristic must be efficient in terms of the number of objective function evaluations required.

Estimating the sampling distribution

More serious computational challenges lie in evaluating the expected profit (lines 2-10) rather than the simpler to compute D-efficiency metric. To compute the expected profit associated with a particular conjoint design, we have to consider in turn each of the potential outcomes of the experiment (line 3). The number of possible outcomes of the experiment is a^q and thus grows very large for reasonable size surveys. For each of these expected outcomes, we must evaluate the sampling distribution, which is the likelihood of the outcome conditional on the prior (line 4). Computing the sampling distribution is straightforward, but it does involve integrating the likelihood function for the proposed questions over the prior distribution. One approach to reducing the computational burden is to continue on with computing expected profit only for those outcomes with a reasonably high probability and ignore the contribution of low-probability outcomes of the experiment to expected profit.

Estimating the conditional posterior

For each outcome of the experiment, the posterior distribution of the parameters must be computed (line 5). For a multinomial logit model, the standard way of determining the posterior conditional on an observed set of choices is to use an MCMC sampler (cf. Rossi, Allenby and McCulloch 2005). For a simple aggregate model, this may take several minutes of computation and it may be difficult to automate due to the need to check for convergence of the MCMC sampler. The posterior may be very diffuse particularly when the prior is diffuse and the proposed conjoint design is not very informative about the parameters. Heterogeneous models may require hours or days for this step alone (cf., Rossi, Allenby and McCulloch 2005). Methods for automated algorithm tuning, e.g., adaptive MCMC approaches (cf. Atchade and Rosenthal 2005) and convergence checking, e.g., monitoring Gelman-Rubin potential scale reduction statistics to determine when to stop the sampler (Gelman et al. 2003), could be used to automate the process of running the sampling algorithm for each potential outcome of the experiment.

Product design optimization

Once a sample of draws from the posterior conditional on a potential outcome is obtained, we can compute the expected profit for any given product design based on that conditional posterior. This calculation involves integrating the profit function over the conditional posterior of the parameters (line 7), but the integral can easily be estimated using a sample from the conditional posterior. However, for each potential outcome of the experiment, the design that maximizes expected profit must be identified (lines 6-8). This product design optimization problem remains an open area of research (cf., Simpson, Saddique and Jiao 2006 for a review) and although many product design algorithms have been proposed, few of the existing algorithms are efficient enough to be used repeatedly as in this context. Exiting algorithms may require on the order of several minutes, which may not be reasonable given that this optimization problem may need to be solved a^q times for each evaluation of the conjoint design objective function.

A potential approach to solving the design algorithm is to use a branch and bound algorithm based on the production-line design algorithm proposed in Feit and Wu (2001). This algorithm involves constructing bounds on the expected profit for an entire set of possible product designs by constructing 'super designs' consisting of the product design with the best of all attributes of across all product designs in the set. Those bounds are then used to efficiently search for the best possible design following the usual branch and bound procedures. In the conjoint design context, the branch and bound has some advantage over proposed discrete optimization heuristics (e.g., genetic algorithms) because it does not involve any algorithm tuning.

The final step in solving the decision theoretic experimental design problem is to compute the total expected profit for the design as the average of the expected profit conditional on each outcome over the sampling distribution for the outcomes (line 10). This step is computationally straightforward.

Conclusions and future research

In this essay we have proposed a new approach to the design of choice experiments that provides several important benefits over approaches based on statistical efficiency. First, unlike prior methods, the decision-theoretic approach provides a natural way to incorporate available information about production costs and other aspects of the product design problem into the design of the choice experiment. While managers have long recognized the importance of using this information in the design of choice experiments, there have been no formal methods for incorporating this information into the design algorithm. Second, the proposed approach estimates the economic value of the choice experiment itself and this value can be easily compared the cost of fielding the choice experiment to determine whether it makes sense to proceed with the market research at all (or to simply proceed with a decision based on available prior information). Third, in our illustrative example, we found that the decision theoretic approach is particularly well suited, relative to statistical efficiency approaches, to designing very short questionnaires suitable for web-based surveys and recommendation engines.

Much work remains to develop this approach into a method that is useable in practical marketing research applications. The critical next step is to develop an algorithm to solve the decision theoretic conjoint design problem, or some simplification of it, so that differences between experimental designs based on statistical efficiency versus those based on the decision theoretic criteria can be explored for problems more practical in size. Ultimately, we hope to demonstrate that the decision theoretic approach leads to experimental designs that are more efficient, in that they produce more decision-relevant information given a similar number of questions. Since the decision-theoretic approach is computationally expensive, it would also be valuable to investigate under what conditions the more tractable D-efficient design is the same as or similar to the decision theoretic design.

There are also a number of issues that could be explored as part of a broader research agenda to develop the decision theoretic approach to conjoint design. For example, the formulation we propose here assumes that cost data is known, however the decision-theoretic framework could also easily accommodate prior uncertainty in the cost data resulting in choice experiments and product designs that are robust to these uncertainties. The decision theoretic framework could also be used as the basis for a sequential approach to designing choice experiments that could be used to create adaptive choice surveys, by recasting the decision theoretic design as a dynamic program.

References

Atachade, Yves and Jeffrey S. Rosenthal (2005) On Adaptive Markov Chain Monte Carlo Algorithms, *Bernoulli*, 11, 815-828.

Arora, Neeraj and Joel Huber (2001) Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments, *Journal of Consumer Research*, 28, 273-283.

Berger, James O. (1980) Statistical Decision Theory: Foundations, Concepts and Methods. Springer-Verlag, New York.

Feit, Eleanor M. and S. David Wu (2001) Transfer line design with uncertain machine performance information, IEEE Transactions on Robotics and Automation, 16:5, 581-587.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin (2003) Bayesian Data Analysis, second edition, CRC Press.

Huber, Joel and Klaus Zwerina (1996) The Importance of Utility Balance in Efficient Choice Designs, Journal of Marketing Research, 33.

Kessels, Roselinde, Peter Goos and Martina Vandebroek (2006) A Comparison of Criteria to Design Efficient Choice Experiments, Journal of Marketing Research, 43, 409-419.

Kuhfeld, Warren F., Randall D. Tobias and Mark Garratt (1994) Efficient Experimental Design with Marketing Research Applications, Journal of Marketing Research, 31:4.

Liu, Qing, Angela M. Dean, and Greg M. Allenby (2007) Design for Hyperparameter Estimation in Linear Models, Journal of Statistical Theory and Practice, 1(3), 311-328.

Liu, Qing, Angela M. Dean, David Bakken and Greg M. Allenby (2009) Studying the Level-Effect in Conjoint Analysis: Application of Efficient Experimental Designs for Hyperparameter Estimation, Quantitative Marketing and Economics, 7, 69-93.

Michalek, Jeremy J., Fred M. Feinberg, Feray Adiguzel, Peter Ebbes, Panos Y. Papalambros (2008WP) Realizable Product Line Design Optimization: Coordinating Marketing and Engineering Models via Analytical Target Cascading, working paper.

Toubia, Olivier, and John R. Hauser (2007) On Managerial Efficiency for Experimental Designs, Marketing Science, 26:6.

Rossi, Peter E., Greg M. Allenby and Robert McCulloch (2005) Bayesian Statistics and Marketing, John Wiley & Sons, Ltd.

Simpson, Timothy W., Zahed Saddique, and Jianxin Jiao (2006) Product platform and product family design: methods and applications. Birkhauser.

Sandor, Zsolt and Michel Wedel (2001) Designing Conjoint Choice Experiments Using Managers Prior Beliefs, Journal of Marketing Research, 38.

Sandor, Zsolt and Michel Wedel (2002) Profile Construction in Experimental Choice Designs for Mixed Logit Models, Marketing Science, 21:4.

Sandor, Zsolt and Michel Wedel (2005) Heterogeneous Conjoint Designs, Journal of Marketing Research, 42:2.

Zwerina, Klaus, Joel Huber and Warren F. Kuhfeld (1996) A General Method for Constructing Efficient Choice Designs, working paper.

Chapter 5

Conclusions

In the previous three essays, we have developed three distinct marketing research methods, each designed to address issues that arise in the context of product design. Because market data is of limited use in the context of product design, all three essays employ choice experiments and each essay addresses a different weakness of existing methods for collecting and analyzing data from a choice experiment. In the first essay we address the potential for *dependent variables* observed in a choice experiment, i.e. hypothetical choices, to be inconsistent with the market behavior that product designers would like to predict. We address this problem by including additional data from the market in the estimation of the model. The third essay focuses on how one can carefully design the *independent variables* to maximize the usefulness of the data to product designers. In the second essay we investigate how assumptions about the *error term* in a choice model affect predicted outcomes in a choice experiment. Each essay contributes to the goal of developing accurate, practical models that can predict market acceptance of new product designs and ultimately inform product design decisions.

Although all three essays were motivated by problems in the intersection of market research and product design, they each also contribute to the larger literature in Marketing. For instance, the method for combining different sources of choice data proposed in the first essay is just as useful in the context of marketing mix planning, where data from a choice experiment is equally suspect in terms of predicting outcomes in the marketplace. In essay 2, we explore a basic question about how consumers' expertise in a category changes their decision making behavior and a complete understanding of this relationship would be useful for predicting the development of any market where the expertise of consumers varies. Finally, the concept of using a decision-theoretic approach to designing market research surveys, which requires the analyst to formalize the marketing decision at hand in terms of a loss function, is broadly applicable beyond the context of product design and could be applied to numerous marketing research applications.