A New Generalized Heterogeneous Data Model (GHDM) to Jointly Model Mixed Types of Dependent Variables

Chandra R. Bhat
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712
Phone: 512-471-4535; Fax: 512-475-8744
Email: bhat@mail.utexas.edu
and
King Abdulaziz University, Jeddah 21589, Saudi Arabia

August 1, 2014

ABSTRACT

This paper formulates a generalized heterogeneous data model (GHDM) that jointly handles mixed types of dependent variables—including multiple nominal outcomes, multiple ordinal variables, and multiple count variables, as well as multiple continuous variables—by representing the covariance relationships among them through a reduced number of latent factors. Sufficiency conditions for identification of the GHDM parameters are presented. The maximum approximate composite marginal likelihood (MACML) method is proposed to estimate this jointly mixed model system. This estimation method provides computational time advantages since the dimensionality of integration in the likelihood function is independent of the number of latent factors. The study undertakes a simulation experiment within the virtual context of integrating residential location choice and travel behavior to evaluate the ability of the MACML approach to recover parameters. The simulation results show that the MACML approach effectively recovers underlying parameters, and also that ignoring the multi-dimensional nature of the relationship among mixed types of dependent variables can lead not only to inconsistent parameter estimation, but also have important implications for policy analysis.

Keywords: Latent factors, big data analytics, high dimensional data, MACML estimation approach, mixed dependent variables, structural equations models, integrated land use-transportation modeling, factor analysis.

1. INTRODUCTION

The joint modeling of data with mixed types of dependent variables (including ordered-response or ordinal variables, unordered-response or nominal variables, count variables, and continuous variables) is of interest in several fields, including biology, developmental toxicology, finance, economics, epidemiology, social science, and transportation (see a good synthesis of applications in De Leon and Chough, 2013). For instance, in the clinical biology field, alternative treatments for a specific condition are assessed based on binary, ordered, and continuous indicators of the treatment's after-effects; this approach has been used to assess the effectiveness of depression medication in reducing the occurrence, frequency, and intensity of depression (such as in Gueorguieva and Sanacora, 2006). In the health field, in addition to binary, count, and continuous variables related to the occurrence, frequency, and intensity, respectively, of specific health problems, it is not uncommon to obtain ordinal information on quality of life outcomes/perceptions. In the toxicology field, the focus is on regulating the use of chemical and pharmaceutial drugs (Abrams *et al*., 2000). Typically, varying quantities of a drug are administered to mice; the effects on their offspring are studied in terms of combinations of discrete outcomes (such as the presence of congenital deformations) and continuous outcomes (such as birth weight). In the transportation field, households that are not auto-oriented are likely to locate in transit- and pedestrian-friendly neighborhoods that are characterized by mixed and high-density land use; pedestrian-oriented design in such communities may also further structurally reduce motorized vehicle miles of travel. If that is the case, then it is likely that the choices of residential location (nominal variable), vehicle ownership (count), and vehicle miles of travel (continuous) are being made jointly as a bundle (see, for example, Bhat *et al.*, 2014a).

The interest in mixed model systems has been spurred particularly by the recent availability of high-dimensional heterogeneous data with complex dependence structures, thanks to technology that allows the collection and archival of voluminous amounts of data ("big data"). Unlike standard correlated linear data that can be analyzed using traditional multivariate linear regression models, the presence of non-commensurate outcomes creates difficulty because of the absence of a convenient multivariate distribution to jointly (and directly) represent the relationship between discrete and continuous outcomes. Several approaches have been developed to handle such situations. The first and simplest is, of course, to simply ignore the dependence and estimate separate models. However, such an approach is inefficient in estimating covariate

effects for each outcome because it fails to borrow information on other outcomes, and is limiting in its ability to answer intrinsically multivariate questions such as the effect of a covariate on a multidimensional outcome (Teixeira-Pinto and Harezlak, 2013). Besides, joint analysis of mixed outcomes obviates the need for multiple tests and facilitates global tests, offering superior power in testing and better control of type I error rates (De Leon and Zhu, 2008). But, more importantly, if some endogenous outcomes are used to explain other endogenous outcomes (such as examining the effect of density of residence on auto-ownership model), and if the outcomes are not modeled jointly to recognize the presence of unobserved exogenous variable effects, the result is inconsistent estimation of the effects of one endogenous outcome on another (see Bhat and Guo, 2007, and Mokhtarian and Cao, 2008). A second common approach to joint mixed outcome modeling originates in the general location model (GLOM), which assumes an arbitrary marginal distribution for the discrete outcomes and a conditional (on the discrete component) normality assumption for the continuous outcomes (De Leon and Chough, 2013). However, the GLOM is not suitable for ordinal outcome variables and does not accommodate dependence between nominal and ordinal outcomes. A third "reverse-factorization" approach is to employ a latent variable representation for binary/ordinal outcomes, and assume a multivariate normal (MVN) distribution for the continuous outcomes and the latent variables underlying the binary/ordinal outcomes. Then, the joint distribution is derived using a marginal distribution of the continuous outcomes and the conditional distribution of the latent variables (given the continuous variables) underlying the binary/ordinal outcomes. This approach is referred to as the conditional grouped continuous model (CGCM) by De Leon and Chough (2013). However, this approach cannot be directly extended to the case of nominal outcomes, since nominal outcomes do not arise from the partitioning of a single latent variable using thresholds (as is the case for binary/ordinal outcomes). So, De Leon and Carriere (2007) and De Leon *et al.* (2011) proposed an extended factorization approach, which they label as the general mixed data model (GMDM), to accommodate nominal outcomes. They use a GLOM for the joint distribution of the nominal and continuous outcomes, and a CGCM for the joint distribution of the ordinal and continuous outcomes. Specifically, the GMDM uses a multinomial distribution for the marginal distribution of the possible multidimensional discrete states obtained from the combinatorics of a set of nominal outcomes, followed by a conditional MVN distribution for the latent variables (underlying the ordinal outcomes) and the continuous outcomes. The mean

vector for this latter conditional MVN distribution is specified to be a function of the multidimensional discrete state, engendering an association between the nominal discrete outcomes and the ordinal/continuous outcomes. However, the covariance matrix of the conditional MVN distribution is constant across the nominal discrete states. A further problem with the GMDM is that the number of multidimensional discrete states explodes as the number of nominal discrete outcomes increases, and as the number of elemental categories within each nominal discrete outcome increases. Besides, the GMDM (like the GLOM) resorts to a factorization approach in which an artificial hierarchy is implicitly assumed. In this hierarchy, the multidimensional discrete outcomes are intermediate responses and the ordinal/continuous outcomes are the ultimate responses (see Wu *et al*., 2013).

Independent from the work discussed above, a fourth approach originates in the economics and transportation fields, wherein mixed models with nominal outcomes are based on latent variable representations of nominal outcomes. Surprisingly, such studies are rarely mentioned in papers in the statistical field that deal with mixed outcomes. The studies in this strand may be viewed as extensions of the CGCM approach to the case of nominal outcomes, except that each nominal outcome is represented by a series of latent variables. An early example of such a multivariate model may be found in Keane (1992), who considered one nominal variable and one continuous variable. However, only relatively recently has this methodology been extended to include mixed nominal, binary, ordinal, count, and continuous variables (for example, see Paleti *et. al.*, 2013 and Bhat *et al.*, 2014a). The resulting mixed models may be viewed as an alternative to the GMDM, and have the advantage that all outcomes are tied based on their latent or observed continuous variable representations (rather than using different types of linkages for different types of outcomes, as in the GMDM). Further, these models treat the mixed outcomes symmetrically rather than imposing any form of hierarchy. The models typically assume an MVN distribution over the entire set of latent and observed continuous variables characterizing the many types of outcomes. A variant of this methodology uses a Gaussian copula function to tie the latent and observed continuous variables if the variables have different marginal distributions, though this approach has been confined to scenarios without a nominal outcome (see, for example, Wu *et al.*, 2013). Another variant introduces random error terms linearly in the latent and observed continuous variable equations associated with the discrete outcomes and continuous outcomes, respectively. The underlying continuous variables

are considered to be independent, conditional on these random error terms. Then, if these random error terms are common or correlated, the result is an association structure among the mixed outcomes. Such a specification falls under the label of a multivariate generalized linear latent and mixed model (GLLAMM), and is particularly helpful when considering clustering effects (due to multiple observations from the same person or due to spatial dependency) in addition to correlation across mixed outcomes (see, for example, Faes *et al*., 2009 and Bhat *et al*., 2014a) . An extension of this approach that accommodates clustering as well as an association structure among mixed outcomes (that is, mixed outcomes are independent, conditional on appropriately specified latent variables) is referred to as the item response theory (IRT) model in the literature (see Bartholomew *et al*., 2011 and Feddag, 2013). However, again, these GLLAMM and IRT models have been predominantly used for cases with no nominal variables, though similar approaches can be used to generate dependence between a nominal variable and other kinds of variables too (see, for example, Bhat and Guo, 2007 and Pinjari *et al*., 2008).

A fifth approach, originating from the social sciences, implicitly generates dependence among mixed outcomes by writing the latent and observed continuous variables as a function of unobserved psychological constructs. These relationships are characterized as measurement equations, in that the psychological constructs are manifested in the larger combination of mixed outcomes. The constructs themselves are related to exogenous variables and may be correlated with one another in a structural relationship. In this approach, the unobserved psychological constructs serve as latent factors that provide a structure to the dependence among the many mixed indicator variables. Seen from this perspective, the approach can also be viewed as a parsimonious attempt to explain the covariance relationship among a large set of mixed outcomes through a much smaller number of unobservable latent factors. Sometimes referred to as *factor analysis*, the approach represents a powerful dimension-reduction technique to analyze high-dimensional heterogeneous outcome data by representing the covariance relationship among the data through a smaller number of unobservable latent factors. An entire field of structural equations modeling (SEM) has been developed around this psychological construct-based dependence modeling, originating in some of the early works of Jöreskog (1977). However, the SEM field has focused almost exclusively on non-nominal outcome analysis (see Gates *et al.*, 2011 and Hoshino and Bentler, 2013). Indeed, traditional SEM software (such as LISREL, MPLUS, and EQS) is either not capable of handling nominal indicators or at least are

not readily suited to handle nominal indicators (see Temme *et al.*, 2008). But when this approach is extended to include a nominal indicator, it essentially takes the form of an integrated choice and latent variable (ICLV) model (Ben-Akiva *et al.*, 2002, and Bolduc *et al*., 2005). Also, while traditional SEM techniques typically adopt normally distributed latent factors along with normally distributed measurement error terms (leading to probit models in the presence of binary/ordered outcomes), ICLV models tend to use normally distributed latent factors mixed with logistically distributed errors in the measurement equations for ordinal variables and type-1 extreme value errors in the nominal outcome utility functions (leading to a probability expression that involves a multivariate integral over the product of logit-type probabilities for the outcomes). In both the SEM and ICLV cases, the standard estimation methodology is the method of maximum likelihood estimation. When there are many binary/ordered-response outcomes (indicators) and/or a nominal variable, the integrals in the overall probability expression are computed using simulation techniques. As indicated by Hoshino and Bentler (2011), this can "be difficult to impossible when the model is complex or the number of variables is large." This is particularly the case with the traditional mixture formulation of ICLV models in general, and particularly when there are several latent factors (see Daziano and Bolduc, 2013).

Recently, Bhat and Dubey (2014) proposed a different way of formulating ICLV models, in which they use a SEM-like probit approach while also accommodating a single nominal variable. Essentially, this approach combines the power and parsimony of the dimension-reduction factor analysis structure of SEMs (as just discussed above) with the extended CGCM approach that uses a symmetric, latent continuous variable representation for all non-continuous outcomes (as in Paleti *et al*., 2013 and Bhat *et al*., 2014a). In this paper, we generalize Bhat and Dubey's approach to the case of multiple nominal outcomes, multiple ordinal variables, multiple count variables, and multiple continuous variables. The resulting model, which we label simply as the *generalized heterogeneous data model* (GHDM), is general enough to accommodate other models in the literature as special cases. Straightforward extensions of the model are available to accommodate longitudinal and spatial clustering, though we focus on the non-clustered mixed outcome model in the current paper. We propose the estimation of the GHDM using Bhat's maximum approximate composite marginal likelihood (MACML) inference approach. In particular, in our approach, the dimensionality of integration in the composite marginal likelihood (CML) function that needs to be maximized to obtain a consistent estimator (under

standard regularity conditions) for the GHDM parameters is independent of the number of latent factors and easily accommodates general covariance structures for the structural equation and for the utilities of the discrete alternatives for each nominal outcome. Further, the use of the analytic approximation in the MACML approach to evaluate the multivariate cumulative normal distribution (MVNCD) function in the CML function simplifies the estimation procedure even further so that the proposed MACML procedure requires the maximization of a function that has no more than bivariate normal cumulative distribution functions to be evaluated.

2. THE GHDM FORMULATION

There are two components to the model: (1) the latent variable SEM, and (2) the latent variable measurement equation model. These components are discussed in turn below. In the following presentation, for ease in notation, we will consider a cross-sectional model. As appropriate and convenient, we will suppress the index q for decision-makers (q=1,2,… Q) in parts of the presentation, and assume that all error terms are independent and identically distributed across decision-makers.

2.1. Latent Variable SEM

Let l be an index for latent variables (l=1,2,…,L). Consider the latent variable z_l^* and write it as a linear function of covariates:

$$z_l^* = \boldsymbol{\alpha}_l' \boldsymbol{w} + \eta_l, \tag{1}$$

where $\boldsymbol{w}$ is a $(\widetilde{D} \times 1)$ vector of observed covariates (excluding a constant), $\boldsymbol{\alpha}_l$ is a corresponding $(\widetilde{D} \times 1)$ vector of coefficients, and η_l is a random error term assumed to be standard normally distributed for identification purposes (see Stapleton, 1978). Next, define the $(L \times \widetilde{D})$ matrix $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_L)'$, and the $(L \times 1)$ vectors $\boldsymbol{z}^* = (z_1^*, z_2^*, ..., z_L^*)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \ldots, \eta_L)'$. Unlike much of the earlier research in ICLV modeling, we allow an MVN correlation structure for $\boldsymbol{\eta}$ to accommodate interactions among the unobserved latent variables: $\boldsymbol{\eta} \sim MVN_L[\mathbf{0}_L, \boldsymbol{\Gamma}]$, where $\mathbf{0}_L$ is an $(L \times 1)$ column vector of zeros, and $\boldsymbol{\Gamma}$ is $(L \times L)$ correlation matrix. In matrix form, we may write Equation (1) as:

$$\boldsymbol{z}^* = \boldsymbol{\alpha} \boldsymbol{w} + \boldsymbol{\eta}. \tag{2}$$

It is not uncommon in the SEM literature to have latent variables affecting each other in the SEM. However, it is not easy to justify *a priori* inter-relationships between unobserved variables, and so we prefer a general covariance structure for the latent variables as in Equation (2). Alternatively, Equation (2) may be viewed as an unrestricted reduced form representation of the actual inter-relationships between latent variables.

2.2. Latent Variable Measurement Equation Model Components

We will consider a combination of continuous, ordinal, count, and nominal outcomes (indicators) of the underlying latent variable vector $\boldsymbol{z}^*$. However, these outcomes may be a function of a set of exogenous variables too.

Let there be H continuous outcomes $(y_1, y_2, ..., y_H)$ with an associated index h $(h = 1, 2, ..., H)$. Let $y_h = \boldsymbol{\gamma}_h' \boldsymbol{x} + \boldsymbol{d}_h' \boldsymbol{z}^* + \varepsilon_h$ in the usual linear regression fashion, where $\boldsymbol{x}$ is an $(A \times 1)$ vector of exogenous variables (including a constant) as well as possibly the observed values of other endogenous continuous variables, other endogenous ordinal variables, other endogenous count variables, and other endogenous nominal variables (introduced as dummy variables). $\boldsymbol{\gamma}_h$ is a corresponding compatible coefficient vector.[1] $\boldsymbol{d}_h$ is an $(L \times 1)$ vector of latent variable loadings on the h^{th} continuous outcome, and ε_h is a normally distributed measurement error term. Stack the H continuous outcomes into an $(H \times 1)$ vector $\boldsymbol{y}$, and the H error terms into another $(H \times 1)$ vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_H)'$. Also, let $\boldsymbol{\Sigma}$ be the covariance matrix of $\boldsymbol{\varepsilon}$, which is restricted to be diagonal. This helps identification because there is already an unobserved latent variable vector $\boldsymbol{z}^*$ that serves as a vehicle to generate covariance between the outcome variables (as we discuss in the next section). Define the $(H \times A)$ matrix $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ...\gamma_H)'$ and the

[1] In joint limited-dependent variable systems in which one or more dependent variables are not observed on a continuous scale, such as the joint system considered in the current paper that has discrete dependent and count variables (which we will more generally refer to as limited-dependent variables), the structural effects of one limited-dependent variable on another can only be in a single direction. That is, it is not possible to have correlated unobserved effects underlying the propensities determining two limited-dependent variables, as well as have the observed limited-dependent variables themselves structurally affect each other in a bi-directional fashion. This creates a logical inconsistency problem (see Maddala, 1983, page 119 for a good discussion). It is critical to note that, regardless of which directionality of structural effects among the endogenous variables is specified (or even if no relationships are specified), the system is a joint bundled system because of the correlation in unobserved factors impacting the underlying propensities.

$(H \times L)$ matrix of latent variable loadings $\boldsymbol{d} = \left(\boldsymbol{d_1}, \boldsymbol{d_2}, ..., \boldsymbol{d_H}\right)'$. Then, one may write, in matrix form, the following measurement equation for the continuous outcomes:

$$\boldsymbol{y} = \boldsymbol{\gamma x} + \boldsymbol{d z}^* + \boldsymbol{\varepsilon}. \tag{3}$$

Next, consider N ordinal outcomes (indicator variables) for the individual, and let n be the index for the ordinal outcomes $(n = 1, 2, ..., N)$. Also, let J_n be the number of categories for the n^{th} ordinal outcome $(J_n \geq 2)$ and let the corresponding index be j_n $(j_n = 1, 2, ..., J_n)$. Let $\tilde{y}_n^*$ be the latent underlying variable whose horizontal partitioning leads to the observed outcome for the n^{th} ordinal variable. Assume that the individual under consideration chooses the a_n^{th} ordinal category. Then, in the usual ordered response formulation, for the individual, we may write:

$$\tilde{y}_n^* = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{n}}' \boldsymbol{x} + \tilde{\boldsymbol{d}}_{\boldsymbol{n}}' \boldsymbol{z}^* + \tilde{\varepsilon}_n, \text{ and } \tilde{\psi}_{n,a_n-1} < \tilde{y}_n^* < \tilde{\psi}_{n,a_n}, \tag{4}$$

where $\boldsymbol{x}$ is a vector of exogenous and possibly endogenous variables as defined earlier, $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{n}}$ is a corresponding vector of coefficients to be estimated, $\tilde{\boldsymbol{d}}_{\boldsymbol{n}}$ is an $(L \times 1)$ vector of latent variable loadings on the n^{th} continuous outcome, the $\tilde{\psi}$ terms represent thresholds, and $\tilde{\varepsilon}_n$ is the standard normal random error for the n^{th} ordinal outcome. For each ordinal outcome, $\tilde{\psi}_{n,0} < \tilde{\psi}_{n,1} < \tilde{\psi}_{n,2} ... < \tilde{\psi}_{n,J_n-1} < \tilde{\psi}_{n,J_n}$; $\tilde{\psi}_{n,0} = -\infty$, $\tilde{\psi}_{n,1} = 0$, and $\tilde{\psi}_{n,J_n} = +\infty$. For later use, let $\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}} = (\tilde{\psi}_{n,2}, \tilde{\psi}_{n,3} ..., \tilde{\psi}_{n,J_n-1})'$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1', \tilde{\boldsymbol{\psi}}_{\boldsymbol{2}}' ..., \tilde{\boldsymbol{\psi}}_{\boldsymbol{N}})'$. Stack the N underlying continuous variables $\tilde{y}_n^*$ into an $(N \times 1)$ vector $\tilde{\boldsymbol{y}}^*$, and the N error terms $\tilde{\varepsilon}_n$ into another $(N \times 1)$ vector $\tilde{\boldsymbol{\varepsilon}}$. Define $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1, \tilde{\boldsymbol{\gamma}}_2 ..., \tilde{\boldsymbol{\gamma}}_H)'$ [$(N \times A)$ matrix] and $\tilde{\boldsymbol{d}} = \left(\tilde{\boldsymbol{d}}_{\boldsymbol{1}}, \tilde{\boldsymbol{d}}_{\boldsymbol{2}}, ..., \tilde{\boldsymbol{d}}_{\boldsymbol{N}}\right)$ [$(N \times L)$ matrix], and let $\mathbf{IDEN_N}$ be the identity matrix of dimension N representing the correlation matrix of $\tilde{\boldsymbol{\varepsilon}}$ (so, $\tilde{\boldsymbol{\varepsilon}} \sim MVN_N(\mathbf{0}_N, \mathbf{IDEN_N})$; again, this is for identification purposes, given the presence of the unobserved $\boldsymbol{z}^*$ vector to generate covariance. Finally, stack the lower thresholds for the decision-maker $\tilde{\psi}_{n,a_n-1}(n = 1, 2, ..., N)$ into an $(N \times 1)$ vector $\tilde{\boldsymbol{\psi}}_{\boldsymbol{low}}$ and the upper thresholds $\tilde{\psi}_{n,a_n}(n = 1, 2, ..., N)$ into another vector $\tilde{\boldsymbol{\psi}}_{\boldsymbol{up}}$. Then, in matrix form, the measurement equation for the ordinal outcomes (indicators) for the decision-maker may be written as:

$$\tilde{\boldsymbol{y}}^* = \tilde{\boldsymbol{\gamma}} \boldsymbol{x} + \tilde{\boldsymbol{d}} \boldsymbol{z}^* + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\psi}}_{\boldsymbol{low}} < \boldsymbol{y}^* < \tilde{\boldsymbol{\psi}}_{\boldsymbol{up}}. \tag{5}$$

Let there be C count variables for a household, and let c be the index for the count variables $(c = 1, 2, ..., C)$. Let the count index be k_c $(k_c = 0, 1, 2, ..., \infty)$ and let r_c be the actual observed count value for the household. Then, following the recasting of a count model in a generalized ordered-response probit formulation (see Castro, Paleti, and Bhat, or CPB, 2012 and Bhat *et al*., 2014b), a generalized version of the negative binomial count model may be written as:

$$\breve{y}_c^* = \breve{\boldsymbol{d}}_c' \boldsymbol{z}^* + \breve{\varepsilon}_c, \breve{\psi}_{c,r_c-1} < \breve{y}_c^* < \breve{\psi}_{c,r_c}, \tag{6}$$

$$\breve{\psi}_{c,r_c} = \Phi^{-1}\left[\frac{(1-\upsilon_c)^{\theta_l}}{\Gamma(\theta_c)}\sum_{t=0}^{r_c}\left(\frac{\Gamma(\theta_c+t)}{t!}(\upsilon_c)^t\right)\right] + \varphi_{c,r_c},\ \upsilon_c = \frac{\lambda_c}{\lambda_c+\theta_c}, \text{ and } \lambda_c = e^{\breve{\gamma}_c'\boldsymbol{x}}. \tag{7}$$

In the above equation, $\breve{y}_c^*$ is a latent continuous stochastic propensity variable associated with the count variable c that maps into the observed count r_c through the $\breve{\boldsymbol{\psi}}_c$ vector (which is a vertically stacked column vector of thresholds $(\breve{\psi}_{c,-1}, \breve{\psi}_{c,0}, \breve{\psi}_{c,1}, \breve{\psi}_{c,2}, ...)'$. $\breve{\boldsymbol{d}}_c$ is an $(L \times 1)$ vector of latent variable loadings on the c^{th} count outcome, and $\breve{\varepsilon}_c$ is a standard normal random error term. $\breve{\boldsymbol{\gamma}}_c$ is a column vector corresponding to the vector $\boldsymbol{x}$. Φ^{-1} in the threshold function of Equation (7) is the inverse function of the univariate cumulative standard normal. θ_c is a parameter that provides flexibility to the count formulation, and is related to the dispersion parameter in a traditional negative binomial model $(\theta_c > 0 \forall c)$. $\Gamma(\theta_c)$ is the traditional gamma function; $\Gamma(\theta_c) = \int_{\tilde{t}=0}^{\infty} \tilde{t}^{\,\theta_c - 1} e^{-\tilde{t}}\, d\tilde{t}$. The threshold terms in the $\breve{\boldsymbol{\psi}}_c$ vector satisfy the ordering condition (*i.e.*, $\breve{\psi}_{c,-1} < \breve{\psi}_{c,0} < \breve{\psi}_{c,1} < \breve{\psi}_{c,2} .... < \infty \ \forall\ c$) as long as $\varphi_{c,-1} < \varphi_{c,0} < \varphi_{c,1} < \varphi_{c,2} .... < \infty$. The presence of the φ_c terms in the thresholds provides substantial flexibility to accommodate high or low probability masses for specific count outcomes without the need for cumbersome traditional treatments using zero-inflated or related mechanisms in multi-dimensional model systems (see Castro *et al*., 2011 for a detailed discussion). For identification, we set $\varphi_{c,-1} = -\infty$ and $\varphi_{c,0} = 0$ for all count variables c. In addition, we identify a count value e_c^* $(e_c^* \in \{0, 1, 2, ...\})$ above which φ_{c,k_c} $(k_c \in \{1, 2, ...\})$ is held fixed at φ_{k,e_c^*}; that is, $\varphi_{c,k_c} = \varphi_{c,e_c^*}$

if $k_c > e_c^*$, where the value of e_c^* can be based on empirical testing. Doing so is the key to allowing the count model to predict beyond the range available in the estimation sample. For later use, let $\boldsymbol{\varphi}_c = (\varphi_{c,1}, \varphi_{c,2}, \ldots \varphi_{c,e_c^*})'$ ($e_c^* \times 1$ vector) (assuming $e_c^* > 0$), $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1', \boldsymbol{\varphi}_2', \ldots \boldsymbol{\varphi}_C')'$ $\left[\left(\sum_c e_c^*\right) \times 1 \text{ vector}\right]$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_C)'$ $[\text{C} \times 1 \text{ vector}]$. Also, stack the C latent variables $\breve{y}_c^*$ into a $(C \times 1)$ vector $\breve{\boldsymbol{y}}^*$, and the C error terms $\breve{\varepsilon}_c$ into another $(C \times 1)$ vector $\breve{\boldsymbol{\varepsilon}}$. Let $\breve{\boldsymbol{\varepsilon}} \sim MVN_C(\mathbf{0}_C, \mathbf{IDEN_C})$ from identification considerations, and stack the lower thresholds of the individual $\breve{\psi}_{c,r_c-1}(c = 1, 2, ..., C)$ into a $(C \times 1)$ vector $\breve{\boldsymbol{\psi}}_{\boldsymbol{low}}$, and the upper thresholds $\breve{\psi}_{c,r_c}(c = 1, 2, ..., C)$ into another $(C \times 1)$ vector $\breve{\boldsymbol{\psi}}_{\boldsymbol{up}}$. Define $\breve{\boldsymbol{\gamma}} = (\breve{\boldsymbol{\gamma}}_1, \breve{\boldsymbol{\gamma}}_2, ..., \breve{\boldsymbol{\gamma}}_C)'$ [$(C \times A)$ matrix] and $\breve{\boldsymbol{d}} = \left(\breve{\boldsymbol{d}}_1, \breve{\boldsymbol{d}}_2, ..., \breve{\boldsymbol{d}}_C\right)'$ [$(C \times L)$ matrix]. With these definitions, the latent propensity underlying the count outcomes may be written in matrix form as:

$$\breve{\boldsymbol{y}}^* = \breve{\boldsymbol{d}}\boldsymbol{z}^* + \breve{\boldsymbol{\varepsilon}}, \quad \breve{\boldsymbol{\psi}}_{\boldsymbol{low}} < \boldsymbol{h}^* < \breve{\boldsymbol{\psi}}_{\boldsymbol{up}} \tag{8}$$

Note also that the interpretation of the generalized ordered-response recasting is that consumers have a latent "long-term" propensity $\breve{y}_c^*$ associated with the demand for each product/service represented by the count *c,* which is a linear function of the latent variable vector $\boldsymbol{z}^*$ (see CPB for a discussion of the interpretation of the generalized ordered-response recasting of count models). Such a specification enables covariance across the count outcomes (through the propensity variables $\breve{y}_c^*$) and between the count outcomes and other mixed outcomes. On the other hand, there may be some specific consumer contexts and characteristics (embedded in $\boldsymbol{x}$) that may dictate how the long-term propensity is manifested in a count demand at any given *instant of time*. Our implicit assumption is that the latent variable vector $\boldsymbol{z}^*$ affects the "long-term" latent demand propensity $\breve{y}_c^*$, but does not play a role in the instantaneous translation of propensity to actual manifested count demand. This allows us to easily incorporate count outcomes within a mixed outcome model, and estimate the resulting model using Bhat (2011) MACML approach. Similarly, an implicit assumption in Equation (8) is that the factors/constraints that are responsible for the instantaneous translation of propensity to manifested count demand (that is, the elements of the $\boldsymbol{x}$ vector) do not affect the "long-term"

demand propensity, though this is being imposed purely for parsimony purposes. Relaxing this assumption does not complicate the model system or the estimation process in any way.

Finally, let there be G nominal (unordered-response) variables for an individual, and let g be the index for the nominal variables (g = 1, 2, 3, …, G). Also, let I_g be the number of alternatives corresponding to the g^{th} nominal variable ($I_g \geq 3$) and let i_g be the corresponding index (i_g = 1, 2, 3, …, I_g). Consider the g^{th} nominal variable and assume that the individual under consideration chooses the alternative m_g. Also, assume the usual random utility structure for each alternative i_g.

$$U_{gi_g} = \boldsymbol{b}'_{gi_g}\boldsymbol{x} + \boldsymbol{\vartheta}'_{gi_g}(\boldsymbol{\beta}_{gi_g}\boldsymbol{z}^*) + \varsigma_{gi_g}, \tag{9}$$

where $\boldsymbol{x}$ is as defined earlier, $\boldsymbol{b}_{gi_g}$ is an $(A \times 1)$ column vector of corresponding coefficients, and ς_{gi_g} is a normal error term. $\boldsymbol{\beta}_{gi_g}$ is an $(N_{gi_g} \times L)$-matrix of variables interacting with latent variables to influence the utility of alternative i_g, and $\boldsymbol{\vartheta}_{gi_g}$ is an $(N_{gi_g} \times 1)$-column vector of coefficients capturing the effects of latent variables and its interaction effects with other exogenous variables. If each of the latent variables impacts the utility of the alternatives for each nominal variable purely through a constant shift in the utility function, $\boldsymbol{\beta}_{gi_g}$ will be an identity matrix of size L, and each element of $\boldsymbol{\vartheta}_{gi_g}$ will capture the effect of a latent variable on the constant specific to alternative i_g of nominal variable g. Let $\boldsymbol{\varsigma_g} = (\varsigma_{g1}, \varsigma_{g2}, \ldots \varsigma_{gI_g})'$ ($I_g \times 1$ vector), and $\boldsymbol{\varsigma_g} \sim MVN_{I_g}(\mathbf{0}, \boldsymbol{\Lambda_g})$. Taking the difference with respect to the first alternative, the only estimable elements are found in the covariance matrix $\breve{\boldsymbol{\Lambda}}_g$ of the covariance matrix of the error differences, $\breve{\boldsymbol{\varsigma}}_{\boldsymbol{g}} = (\breve{\varsigma}_{g2}, \breve{\varsigma}_{g3}, \ldots, \breve{\varsigma}_{gI_g})$ (where $\breve{\varsigma}_{gi} = \breve{\varsigma}_{gi} - \breve{\varsigma}_{g1}$, $i \neq 1$).[2] Further, the variance term at the top left diagonal of $\breve{\boldsymbol{\Lambda}}_g$ (g=1,2,….,G) is set to 1 to account for scale invariance. $\boldsymbol{\Lambda_g}$ is constructed from $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ by adding a row on top and a column to the left. All elements of this additional row and column are filled with values of zero. In addition, the usual identification

[2] Also, in multinomial probit models, identification is tenuous when only individual-specific covariates are used in the vector $\boldsymbol{x}$ (see Keane, 1992 and Munkin and Trivedi, 2008). In particular, exclusion restrictions are needed in the form of at least one individual characteristic being excluded from each alternative's utility in addition to being excluded from a base alternative (but appearing in some other utilities). But these exclusion restrictions are not needed when there are alternative-specific variables.

restriction is imposed such that one of the alternatives serves as the base when introducing alternative-specific constants and variables that do not vary across alternatives (that is, whenever an element of $\boldsymbol{x}$ is individual-specific and not alternative-specific, the corresponding element in $\boldsymbol{b}_{gi_g}$ is set to zero for at least one alternative i_g). To proceed, define $\boldsymbol{U}_g = (U_{g1}, U_{g2}, ..., U_{gI_g})'$ ($I_g \times 1$ vector), $\boldsymbol{b}_g = (\boldsymbol{b}_{g1}, \boldsymbol{b}_{g2}, \boldsymbol{b}_{g3}, ..., \boldsymbol{b}_{gI_g})'$ ($I_g \times A$ matrix), and $\boldsymbol{\beta}_g = (\boldsymbol{\beta}'_{g1}, \boldsymbol{\beta}'_{g2}, ..., \boldsymbol{\beta}'_{gI_g})'$ $\left(\sum_{i_g=1}^{I_g} N_{gi_g} \times L\right)$ matrix. Also, define the $\left(I_g \times \sum_{i_g=1}^{I_g} N_{gi_g}\right)$ matrix $\boldsymbol{\vartheta}_g$, which is initially filled with all zero values. Then, position the $(1 \times N_{g1})$ row vector $\boldsymbol{\vartheta}'_{g1}$ in the first row to occupy columns 1 to N_{g1}, position the $(1 \times N_{g2})$ row vector $\boldsymbol{\vartheta}'_{g2}$ in the second row to occupy columns $N_{g1}+1$ to $N_{g1} + N_{g2}$, and so on until the $(1 \times N_{gI_g})$ row vector $\boldsymbol{\vartheta}'_{gI_g}$ is appropriately positioned. Further, define $\boldsymbol{\varpi}_g = (\boldsymbol{\vartheta}_g \boldsymbol{\beta}_g)$ ($I_g \times L$ matrix), $\ddot{G} = \sum_{g=1}^{G} I_g$, $\widetilde{G} = \sum_{g=1}^{G} (I_g - 1)$, $\widetilde{T} = \sum_{g=1}^{G} T_g$, $\boldsymbol{U} = \left(\boldsymbol{U}'_1, \boldsymbol{U}'_2, ..., \boldsymbol{U}'_G\right)'$ ($\ddot{G} \times 1$ vector), $\boldsymbol{\varsigma} = (\varsigma_1, \varsigma_2, ... \varsigma_G)'$ ($\ddot{G} \times 1$ vector), $\boldsymbol{b} = (\boldsymbol{b}'_1, \boldsymbol{b}'_2, ..., \boldsymbol{b}'_G)'$ ($\ddot{G} \times A$ matrix), $\boldsymbol{\varpi} = (\boldsymbol{\varpi}'_1, \boldsymbol{\varpi}'_2, ..., \boldsymbol{\varpi}'_G)'$ ($\ddot{G} \times L$ matrix), and $\boldsymbol{\vartheta} = Vech(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, ..., \boldsymbol{\vartheta}_G)$ (that is, $\boldsymbol{\vartheta}$ is a column vector that includes all elements of the matrices $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, ..., \boldsymbol{\vartheta}_G$). Then, in matrix form, we may write Equation (9) as:

$$\boldsymbol{U} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\, \boldsymbol{z}^* + \boldsymbol{\varsigma}, \tag{10}$$

where $\boldsymbol{\varsigma} \sim MVN_{\ddot{G}}(\boldsymbol{0}_{\ddot{G}}, \boldsymbol{\Lambda})$. As earlier, to ensure identification, we specify $\boldsymbol{\Lambda}$ as follows:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_2 & \boldsymbol{0} & \boldsymbol{0} \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Lambda}_3 & \boldsymbol{0} \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \ \ldots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \ \ldots & \boldsymbol{\Lambda}_G \end{bmatrix} (\ddot{G} \times \ddot{G} \text{ matrix}), \tag{11}$$

In the general case, this allows the estimation of $\sum_{g=1}^{G}\left(\frac{I_g * (I_g - 1)}{2} - 1\right)$ terms across all the G nominal variables, as originating from $\left(\frac{I_g * (I_g - 1)}{2} - 1\right)$ terms embedded in each $\breve{\boldsymbol{\Lambda}}_g$ matrix; g=1,2,…G) .

3. MODEL SYSTEM IDENTIFICATION AND ESTIMATION

Let $E = (H + N + C)$. Define $\vec{y} = \left(\boldsymbol{y}', \left[\widetilde{\boldsymbol{y}}^*\right]', \left[\breve{\boldsymbol{y}}^*\right]' \right)'$ [$E \times 1$ vector], $\vec{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}', \widetilde{\boldsymbol{\gamma}}', \mathbf{0}_{AC})'$ [$E \times A$ matrix], $\vec{\boldsymbol{d}} = (\boldsymbol{d}', \widetilde{\boldsymbol{d}}', \breve{\boldsymbol{d}}')'$ [$E \times L$ matrix], and $\vec{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}', \widetilde{\boldsymbol{\varepsilon}}', \breve{\boldsymbol{\varepsilon}}')'$ ($E \times 1$ vector), where $\mathbf{0}_{AC}$ is a matrix of zeros of dimension $A \times C$. Let $\boldsymbol{\delta}$ be the collection of parameters to be estimated: $\boldsymbol{\delta} = [\text{Vech}(\boldsymbol{\alpha}), \text{Vech}(\boldsymbol{\Sigma}), \text{Vech}(\vec{\boldsymbol{\gamma}}), \text{Vech}(\vec{\boldsymbol{d}}), \text{Vech}(\breve{\boldsymbol{\gamma}}), \boldsymbol{\varphi}, \boldsymbol{\theta}, \text{Vech}(\boldsymbol{b}), \boldsymbol{\vartheta}, \text{Vech}(\boldsymbol{\Lambda})]$, where the operator "Vech(.)" vectorizes all the non-zero elements of the matrix/vector on which it operates. We will assume that the error vectors $\boldsymbol{\tau}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\xi}$, and $\boldsymbol{\varsigma}$ are independent of each other. While this assumption is not strictly necessary (and can be relaxed in a very straightforward manner within the estimation framework of our model system as long as the resulting model is identified), the assumption aids in developing general sufficiency conditions for identification of parameters in a mixed model when the latent variable vector $\boldsymbol{z}^*$ already provides a mechanism to generate covariance among the mixed outcomes.

With the matrix definitions above, the continuous components of the model system may be written compactly as:

$$\boldsymbol{z}^* = \boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta} \tag{12}$$

$$\vec{\boldsymbol{y}} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}\boldsymbol{z}^* + \vec{\boldsymbol{\varepsilon}}, \text{ with } \text{Var}(\vec{\boldsymbol{\varepsilon}}) = \vec{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{IDEN}_{\mathbf{N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{IDEN}_{\mathbf{C}} \end{bmatrix} \ (E \times E \text{ matrix}) \tag{13}$$

$$\boldsymbol{U} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\boldsymbol{z}^* + \boldsymbol{\varsigma} \tag{14}$$

To develop the reduced form equations, replace the right side of Equation (12) for $\boldsymbol{z}^*$ in Equations (13) and (14) to obtain the following system:

$$\vec{\boldsymbol{y}} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}\boldsymbol{z}^* + \vec{\boldsymbol{\varepsilon}} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}(\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}) + \vec{\boldsymbol{\varepsilon}} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}\boldsymbol{\alpha}\boldsymbol{w} + \vec{\boldsymbol{d}}\boldsymbol{\eta} + \vec{\boldsymbol{\varepsilon}} \tag{15}$$

$$\boldsymbol{U} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\boldsymbol{z}^* + \boldsymbol{\varsigma} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}(\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}) + \boldsymbol{\varsigma} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\varpi}\boldsymbol{\eta} + \boldsymbol{\varsigma}$$

Now, consider the $[(E + \vec{G}) \times 1)]$ vector $\boldsymbol{yU} = \left[\vec{\boldsymbol{y}}', \boldsymbol{U}'\right]'$. Define

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} = \begin{bmatrix} \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}\boldsymbol{\alpha}\boldsymbol{w} \\ \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\,\boldsymbol{\alpha}\boldsymbol{w} \end{bmatrix} \text{ and } \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \boldsymbol{\Omega}_{12}' \\ \boldsymbol{\Omega}_{12} & \boldsymbol{\Omega}_2 \end{bmatrix} = \begin{bmatrix} \vec{\boldsymbol{d}}\boldsymbol{\Gamma}\vec{\boldsymbol{d}}' + \vec{\boldsymbol{\Sigma}} & \vec{\boldsymbol{d}}\boldsymbol{\Gamma}\boldsymbol{\varpi}' \\ \boldsymbol{\varpi}\boldsymbol{\Gamma}\vec{\boldsymbol{d}}' & \boldsymbol{\varpi}\boldsymbol{\Gamma}\boldsymbol{\varpi}' + \boldsymbol{\Lambda} \end{bmatrix} \tag{16}$$

Then $\boldsymbol{yU} \sim \boldsymbol{MVN}_{E+\vec{G}}(\boldsymbol{B}, \boldsymbol{\Omega})$.

The question of identification relates to whether all the elements of $\boldsymbol{\delta}$ are estimable from the elements of $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ (that is, from $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \boldsymbol{\Omega}_{12}$). A simple approach would be to develop easy-to-apply sufficiency conditions for identification (even if they may lead to over-identification and may be more restrictive than needed). A starting point for this is Stapleton (1978), who develops sufficiency conditions for multiple-indicator multiple-cause (MIMIC) models, and whose discussion is applicable to SEM-based models with no nominal variablesv (see also Reilly and O'Brien, 1996). Conforming with the setup of Stapleton and earlier MIMIC models, we will assume in our mixed model that the number of measurement equations with non-nominal variables exceeds the number of latent factors (this will typically be the case, and indeed forms the backbone of modeling a high-dimensional mixed data model through a lower dimensional factor analytic structure). That is, we will assume that $E > L$. However, in constrast to Stapleton, in our study we have nominal variables and also allow the variable vector $\boldsymbol{x}$ to appear in the measurement equations. In this situation, we can develop sufficiency conditions in four steps as follows.

(1) First, if the exogenous covariates do not appear in the measurement equations, one can use Stapleton's (1978) exposition for MIMIC models with no nominal variables (that is, for the sub-model given by Equations (12) and (13) with $\vec{\boldsymbol{y}} = \boldsymbol{0}$) to show that the elements of this sub-model (*i.e.*, $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$, $\vec{\boldsymbol{d}}$, and $\vec{\boldsymbol{\Sigma}}$) are all identifiable as long as:

   (a) diagonality is maintained across the elements of the error term vector $\vec{\boldsymbol{\varepsilon}}$ (that is, $\vec{\boldsymbol{\Sigma}}$ is diagonal),

   (b) $\boldsymbol{\Gamma}$ in the structural equation is specified to be a correlation matrix, and

   (c) for each latent variable, there is at least one outcome variable that loads only on that latent variable and no other latent variable (that is, there is at least one factor complexity one outcome variable for each latent variable) (see also Reilly and O'Brien, 1996).

   The first two of these conditions have already been imposed in the development of our mixed model formulation (the specification that the covariance matrices of $\widetilde{\boldsymbol{\varepsilon}}$ and $\breve{\boldsymbol{\varepsilon}}$ are identity matrices is a result of imposing diagonality combined with a scaling restriction for ordinal and count outcomes). The third condition can be imposed through the empirical specification based on theoretical/intuitive considerations.

(2) Next, we consider the result from the first step, but now relax the constraint that $\vec{\gamma} = \mathbf{0}$, and allow some exogenous variables to influence the non-nominal variables. In this situation, there is an identification problem in Equation (13) if the same exogenous variable is allowed to have a direct impact through the $\boldsymbol{x}$ vector as well as an indirect impact through a latent variable. That is, in general, it is not possible to disentangle the separate effects of the same variable through the direct $\vec{\gamma}$ effect and through the indirect $\vec{d}$ effect. A sufficient identification condition is then to ensure that the element corresponding to the effect of each exogenous variable is zero in either the $\vec{\gamma}$ vector or the $\boldsymbol{\alpha}$ vector. In other words, a sufficient condition for identification of the parameters in the structural equation and the measurement equations for non-nominal outcomes (that is, $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$, $\vec{\gamma}$, $\vec{d}$, and $\vec{\boldsymbol{\Sigma}}$) is:

(a) the three conditions from the first step hold, plus

(b) the condition holds that each element of $\vec{y}$ in Equation (13) is either

(i) directly related to an exogenous variable without being a function of any latent variable that itself has the exogenous variable as a covariate in the structural equation, or

(ii) loaded onto latent variables, but then not directly related to any exogenous variable that itself impacts any of the latent variables on which the outcome variable loads.

Of course, an exogenous variable may not impact an element of $\vec{y}$ both directly and indirectly.

(3) Third, we proceed to the choice model components. Following Bhat and Dubey (2014), we ignore the information available from the covariance matrix $\boldsymbol{\Omega}_{12} = \boldsymbol{\varpi}\boldsymbol{\Gamma}\vec{\boldsymbol{d}}'$. While one can effectively use this covariance matrix to identify parameters in specific situations, we develop a simpler (albeit more restrictive than needed) and general sufficiency condition for identification of the measurement equation parameters corresponding to the nominal outcomes based only on the mean element of the utilities $\boldsymbol{B}_2 = \boldsymbol{bx} + \boldsymbol{\varpi}\boldsymbol{\alpha w}$ (but we retain a general covariance matrix $\boldsymbol{\Lambda}_g$ across alternative utilities for each nominal outcome g). Specifically, all the parameters in the nominal measurement equation part in Equation (14) (that is, elements of $\boldsymbol{b}$, the elements of $\boldsymbol{\vartheta}_g$ (g=1,2,…,G) embedded in $\boldsymbol{\varpi}$, and $\boldsymbol{\Lambda}$) are estimable if all latent variables appear only as interactions and not as direct shifters of utility.

In this case, there are effectively no common exogenous variables in the $\boldsymbol{x}$ effect and the $\boldsymbol{w}$ effect, and so identification of the elements of $\boldsymbol{b}_{\boldsymbol{g}}$ and $\boldsymbol{\vartheta}_g$ is immediate for each nominal variable *g* through estimation of the mean $\boldsymbol{B}_2$. But identification becomes more challenging in the case when the latent variables appear by themselves in the choice models (with or without additional interaction effects of the latent variables). In this case, if an element of $\boldsymbol{b}_{gi_g}$ corresponding to a specific variable in the vector $\boldsymbol{x}$ is non-zero, a sufficient condition for identification is that the utility of alternative i_g not depend on any latent variable that contains that specific variable as a covariate in the structural equation system. This is the most common way that identification has been achieved in most earlier ICLV studies. In fact, most ICLV studies do not even seem to discuss this identification issue. Alternatively, one may include common elements (including alternative-specific attributes in the utilities of the alternatives of nominal variables and those same variables in the structural model for latent variables that impact the utilities), but appropriate restrictions have to be imposed (for example, a latent variable may affect the utility of one of three alternatives for a nominal variable, and a covariate affecting that latent variable may also impact the utility of the same alternative but the coefficient on the covariate may be constrained to be the same as a covariate appearing in the utility of one of the other two alternatives). However, given the sheer number of such specific situations, we leave an in-depth study of identification issues in the context of the overlapping explanatory variables in the structural equation and in the utilities of nominal variables for a later date.

(4) Finally, as indicated in footnote 1, endogenous variable effects can be specified only in a single direction.

To estimate the model, note that, under the utility maximization paradigm, $U_{gi_g} - U_{gm_g}$ must be less than zero for all $i_g \neq m_g$ corresponding to the *gth* nominal variable, since the individual chose alternative m_g. Let $u_{gi_gm_g} = U_{gi_g} - U_{gm_g} (i_g \neq m_g)$, and stack the latent utility differentials into a vector $\boldsymbol{u}_{\boldsymbol{g}} = \left[\left(u_{g1m_g}, u_{g2m_g}, ..., u_{gI_gm_g} \right)' ; i_g \neq m_g \right]$. Also, define $\boldsymbol{u} = \left(\left[\boldsymbol{u}_1 \right]', \left[\boldsymbol{u}_2 \right]', ..., \left[\boldsymbol{u}_{\boldsymbol{G}} \right]' \right)'$. We now need to develop the distribution of the vector $\boldsymbol{yu} = \left(\vec{\boldsymbol{y}}', \boldsymbol{u}' \right)'$ from that of $\boldsymbol{yU} = \left[\vec{\boldsymbol{y}}', \boldsymbol{U}' \right]'$. To

do so, define a matrix $\mathbf{M}$ of size $\left[E+\widetilde{G}\right]\times\left[E+\vec{G}\right]$. Fill this matrix with values of zero. Then, insert an identity matrix of size E into the first E rows and E columns of the matrix $\mathbf{M}$. Next, consider the rows from $E+1$ to $E+I_1-1$, and columns from $E+1$ to $E+I_1$. These rows and columns correspond to the first nominal variable. Insert an identity matrix of size (I_1-1) after supplementing with a column of '-1' values in the column corresponding to the chosen alternative. Next, rows $E+I_1$ through $E+I_1+I_2-2$ and columns $E+I_1+1$ through $E+I_1+I_2$ correspond to the second nominal variable. Again position an identity matrix of size (I_2-1) after supplementing with a column of '-1' values in the column corresponding to the chosen alternative for the second nominal variable. Continue this procedure for all G nominal variables. With the matrix $\mathbf{M}$ as defined, we can write $\boldsymbol{yu} \sim \boldsymbol{MVN}_{\boldsymbol{E+\widetilde{G}}}(\widetilde{\boldsymbol{B}},\widetilde{\boldsymbol{\Omega}})$, where $\widetilde{\boldsymbol{B}}=\mathbf{M}\boldsymbol{B}$ and $\widetilde{\boldsymbol{\Omega}}=\mathbf{M}\boldsymbol{\Omega}\mathbf{M}'$. Next, partition the vector $\widetilde{\boldsymbol{B}}$ into components that correspond to the mean of the vectors $\boldsymbol{y}$ (for the continuous variables), $\ddot{\boldsymbol{y}}=\left(\left[\widetilde{\boldsymbol{y}}^*\right]',\left[\breve{\boldsymbol{y}}^*\right]'\right)'$ [$(N+C)\times 1$ vector], (for the ordinal and count outcomes), and $\boldsymbol{u}$ (for the nominal outcomes), and the matrix $\widetilde{\boldsymbol{\Omega}}$ into the corresponding variances and covariances:

$$\widetilde{\boldsymbol{B}}=\begin{bmatrix}\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}\\ \widetilde{\boldsymbol{B}}_{\ddot{\boldsymbol{y}}}\\ \widetilde{\boldsymbol{B}}_{\boldsymbol{u}}\end{bmatrix}\ (E+\widetilde{G})\times 1 \text{ vector and } \widetilde{\boldsymbol{\Omega}}=\begin{bmatrix}\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}\ddot{\boldsymbol{y}}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{yu}}\\ \widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{y}\ddot{\boldsymbol{y}}} & \widetilde{\boldsymbol{\Omega}}_{\ddot{\boldsymbol{y}}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{yu}}\\ \widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{yu}} & \widetilde{\boldsymbol{\Omega}}'_{\ddot{\boldsymbol{y}}\boldsymbol{u}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{u}}\end{bmatrix}(E+\widetilde{G})\times(E+\widetilde{G}) \text{ matrix} \qquad (17)$$

Define $\widetilde{\boldsymbol{u}}=\left(\ddot{\boldsymbol{y}}',\boldsymbol{u}'\right)'$, so that $\boldsymbol{yu}=(\boldsymbol{y}',\widetilde{\boldsymbol{u}}')'$. Re-partition $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{\Omega}}$ in a different way such that:

$$\widetilde{\boldsymbol{B}}=\begin{bmatrix}\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}\\ \widetilde{\boldsymbol{B}}_{\widetilde{\boldsymbol{u}}}\end{bmatrix},\ \text{where } \widetilde{\boldsymbol{B}}_{\widetilde{\boldsymbol{u}}}=\begin{bmatrix}\widetilde{\boldsymbol{B}}_{\ddot{\boldsymbol{y}}}\\ \widetilde{\boldsymbol{B}}_{\boldsymbol{u}}\end{bmatrix}(N+C+\widetilde{G})\times 1 \text{ vector, and} \qquad (18)$$

$$\widetilde{\boldsymbol{\Omega}}=\begin{bmatrix}\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}\widetilde{\boldsymbol{u}}}\\ \widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{y}\widetilde{\boldsymbol{u}}} & \widetilde{\boldsymbol{\Omega}}_{\widetilde{\boldsymbol{u}}}\end{bmatrix},\ \widetilde{\boldsymbol{\Omega}}_{\widetilde{\boldsymbol{u}}}=\begin{bmatrix}\widetilde{\boldsymbol{\Omega}}_{\ddot{\boldsymbol{y}}} & \widetilde{\boldsymbol{\Omega}}_{\ddot{\boldsymbol{y}}\boldsymbol{u}}\\ \widetilde{\boldsymbol{\Omega}}'_{\ddot{\boldsymbol{y}}\boldsymbol{u}} & \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{u}}\end{bmatrix}(N+C+\widetilde{G})\times(N+C+\widetilde{G}),\ \text{and}\ \widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{y}\widetilde{\boldsymbol{u}}}=\begin{bmatrix}\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}\ddot{\boldsymbol{y}}}\\ \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{yu}}\end{bmatrix}.$$

The conditional distribution of $\widetilde{\boldsymbol{u}}$, given $\boldsymbol{y}$, is MVN with mean $\ddot{\boldsymbol{B}}_{\widetilde{\boldsymbol{u}}}=\widetilde{\boldsymbol{B}}_{\widetilde{\boldsymbol{u}}}+\widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{y}\widetilde{\boldsymbol{u}}}\,\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y}-\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}\right)$ and variance $\ddot{\boldsymbol{\Omega}}_{\widetilde{\boldsymbol{u}}}=\widetilde{\boldsymbol{\Omega}}_{\widetilde{\boldsymbol{u}}}-\widetilde{\boldsymbol{\Omega}}'_{\boldsymbol{y}\widetilde{\boldsymbol{u}}}\,\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}^{-1}\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}\widetilde{\boldsymbol{u}}}$. Next, define threshold vectors as follows:

$\ddot{\boldsymbol{\psi}}_{\boldsymbol{low}} = \left[\widetilde{\boldsymbol{\psi}}'_{\boldsymbol{low}}, \breve{\boldsymbol{\psi}}'_{\boldsymbol{low}}, \left(-\boldsymbol{\infty}_{\widetilde{G}}\right)', \right]'$ ([$(N + C + \widetilde{G}) \times 1$] vector) and $\ddot{\boldsymbol{\psi}}_{\boldsymbol{up}} = \left[\widetilde{\boldsymbol{\psi}}'_{up}, \breve{\boldsymbol{\psi}}'_{up}, \left(\mathbf{0}_{\widetilde{G}}\right)' \right]'$ ([$(N + C + \widetilde{G}) \times 1$] vector), where $-\boldsymbol{\infty}_{\widetilde{\boldsymbol{G}}}$ is a $\widetilde{G} \times 1$-column vector of negative infinities, and $\mathbf{0}_{\widetilde{G}}$ is another $\widetilde{G} \times 1$-column vector of zeros. Then the likelihood function may be written as:

$$L(\boldsymbol{\delta}) = f_H(\boldsymbol{y} \mid \widetilde{\boldsymbol{B}}_{\boldsymbol{y}}, \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}) \times \Pr\left[\ddot{\boldsymbol{\psi}}_{low} \le \widetilde{\boldsymbol{u}} \le \ddot{\boldsymbol{\psi}}_{up}\right], \tag{19}$$

$$= f_H(\boldsymbol{y} \mid \widetilde{\boldsymbol{B}}_{\boldsymbol{y}}, \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}) \times \int_{D_r} f_{N+C+\widetilde{G}}(\boldsymbol{r} \mid \ddot{\boldsymbol{B}}_{\widetilde{\boldsymbol{u}}}, \ddot{\boldsymbol{\Omega}}_{\widetilde{\boldsymbol{u}}}) dr,$$

where the integration domain $D_r = \{\boldsymbol{r} : \ddot{\boldsymbol{\psi}}_{low} \le \boldsymbol{r} \le \ddot{\boldsymbol{\psi}}_{up}\}$ is simply the multivariate region of the elements of the $\widetilde{\boldsymbol{u}}$ vector determined by the observed ordinal indicator outcomes, and the range $(-\boldsymbol{\infty}_{\widetilde{\boldsymbol{G}}}, \mathbf{0}_{\widetilde{G}})$ for the utility differences is taken with respect to the utility of the observed choice alternative for the nominal outcome. $f_H(\boldsymbol{y} \mid \widetilde{\boldsymbol{B}}_{\boldsymbol{y}}, \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}})$ is the MVN density function of dimension H with a mean of $\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}$ and a covariance of $\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{y}}$, and evaluated at $\boldsymbol{y}$. The likelihood function for a sample of Q decision-makers is obtained as the product of the individual-level likelihood functions.

The above likelihood function involves the evaluation of an $N + C + \widetilde{G}$-dimensional rectangular integral for each decision-maker, which can be computationally expensive. Thus, the MACML approach of Bhat (2011) is used.

3.1. The Joint Mixed Model System and the MACML Estimation Approach

Consider the following (pairwise) composite marginal likelihood (CML) function formed by taking the products (across the N ordinal variables, the C count variables, and G nominal variables) of the joint pairwise probability of the chosen alternatives for a decision-maker, and computed using the analytic approximation of the multivariate normal cumulative distribution (MVNCD) function.

$$L_{CML}(\boldsymbol{\delta}) = f_H(\boldsymbol{y} \mid \widetilde{\boldsymbol{B}}_y, \widetilde{\boldsymbol{\Omega}}_y) \times \left(\prod_{n=1}^{N-1} \prod_{n'=n+1}^{N} \Pr(j_n = a_n, j_{n'} = a_{n'}) \right) \times \left(\prod_{c=1}^{C-1} \prod_{c'=c+1}^{C} \Pr(k_c = r_c, k_{c'} = r_{c'}) \right) \times$$
$$\left(\prod_{n=1}^{N} \prod_{c=1}^{C} \Pr(j_n = a_n, k_c = r_c) \right) \times \left(\prod_{n=1}^{N} \prod_{g=1}^{G} \Pr(j_n = a_n, i_g = m_g) \right) \times \qquad (20)$$
$$\left(\prod_{c=1}^{C} \prod_{g=1}^{G} \Pr(k_c = r_c, i_g = m_g) \right) \times \left(\prod_{g=1}^{G-1} \prod_{g'=g+1}^{G} \Pr(i_g = m_g, i_{g'} = m_{g'}) \right).$$

In the above CML approach, the MVNCD function appearing in the CML function is of dimension equal to (1) two for the second component (corresponding to each pair of observed ordinal outcomes), (2) two for the third component (corresponding to each pair of count outcomes), (3) two for the fourth component (corresponding to each pair of an ordinal outcome and a count outcome), (4) I_g for the fifth component (corresponding to each pair of a nominal variable and an ordinal variable), (5) I_g for the sixth component (corresponding to a nominal variable and a count variable), and (6) $I_g + I_{g'} - 2$ for the seventh component (corresponding to a pair of nominal outcomes g and g'). The net result is that the pairwise likelihood function now only needs the evaluation of a cumulative normal distribution function of dimension that is utmost equal to the sum of the alternatives associated with the pair of nominal variables with the two highest number of alternatives.

To explicitly write out the CML function in terms of the standard and bivariate standard normal density and cumulative distribution function, define $\boldsymbol{\omega}_\Delta$ as the diagonal matrix of standard deviations of matrix $\boldsymbol{\Delta}$, using $\phi_R(.;\boldsymbol{\Delta}^*)$ for the multivariate standard normal density function of dimension R and correlation matrix $\boldsymbol{\Delta}^*$ ($\boldsymbol{\Delta}^* = \boldsymbol{\omega}_\Delta^{-1} \boldsymbol{\Delta} \boldsymbol{\omega}_\Delta^{-1}$), and $\Phi_E(.;\boldsymbol{\Delta}^*)$ for the multivariate standard normal cumulative distribution function of dimension E and correlation matrix $\boldsymbol{\Delta}^*$. Define a set of two selection matrices as follows: (1) $\boldsymbol{D}_{vg}$ is an $I_g \times (N + C + \widetilde{G})$ selection matrix with an entry of '1' in the first row and the v^{th} column, an identity matrix of size $I_g - 1$ occupying the last $I_g - 1$ rows and the $N + C + \left[\sum_{j=1}^{g-1} (I_j - 1) + 1 \right]^{th}$ through $N + C + \left[\sum_{j=1}^{g} (I_j - 1) \right]^{th}$ columns (with the convention that $\sum_{j=1}^{0} (I_j - 1) = 0$), and entries of '0'

everywhere else, (2) $\boldsymbol{R}_{gg'}$ is a $(I_g + I_{g'} - 2) \times (N + C + \widetilde{G})$ selection matrix with an identity matrix of size ($I_g - 1$) occupying the first $I_g - 1$) rows and the $N + C + \left[\sum_{j=1}^{g-1}(I_j - 1) + 1\right]^{th}$ through $N + C + \left[\sum_{j=1}^{g}(I_j - 1)\right]^{th}$ columns (with the convention that $\sum_{j=1}^{0}(I_j - 1) = 0$), and another identity matrix of size $(I_{g'} - 1)$ occupying the last $(I_{g'} - 1)$ rows and the $N + C + \left[\sum_{j=1}^{g'-1}(I_j - 1) + 1\right]^{th}$ through $N + C + \left[\sum_{j=1}^{g'}(I_j - 1)\right]^{th}$ columns; all other elements of $\boldsymbol{R}_{gg'}$ take a value of zero. Also, let $\widehat{\boldsymbol{\Omega}}_{vg} = \boldsymbol{D}_{vg}\ddot{\boldsymbol{\Omega}}_{\tilde{u}}\boldsymbol{D}'_{vg}$, $\vec{\boldsymbol{\Omega}}_{gg'} = \boldsymbol{R}_{gg'}\ddot{\boldsymbol{\Omega}}_{\tilde{u}}\boldsymbol{R}'_{gg'}$, $\mu_{v,up} = \dfrac{\left[\ddot{\boldsymbol{\psi}}_{\boldsymbol{up}}\right]_v - \left[\ddot{\boldsymbol{B}}_{\tilde{u}}\right]_v}{\sqrt{\left[\ddot{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}}\right]_{vv}}}$, $\mu_{v,low} = \dfrac{\left[\ddot{\boldsymbol{\psi}}_{\boldsymbol{low}}\right]_v - \left[\ddot{\boldsymbol{B}}_{\tilde{u}}\right]_v}{\sqrt{\left[\ddot{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}}\right]_{vv}}}$, $\rho_{vv'} = \dfrac{\left[\ddot{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}}\right]_{vv'}}{\sqrt{\left[\ddot{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}}\right]_{vv}\left[\ddot{\boldsymbol{\Omega}}_{\tilde{\boldsymbol{u}}}\right]_{v'v'}}}$, where $\left[\ddot{\boldsymbol{\psi}}_{\boldsymbol{up}}\right]_v$ represents the v^{th} element of $\ddot{\boldsymbol{\psi}}_{\boldsymbol{up}}$ (and similarly for other vectors), and $\left[\ddot{\boldsymbol{\Omega}}_{\tilde{u}}\right]_{vv'}$ represents the vv'^{th} element of the matrix $\ddot{\boldsymbol{\Omega}}_{\tilde{u}}$. Then,

$$
\begin{aligned}
L_{CML}(\boldsymbol{\delta}) = & \left(\prod_{\mathrm{h}=1}^{\mathrm{H}} \boldsymbol{\omega}_{\tilde{\Omega}_y}\right)^{-1} \phi_H\left(\left[\boldsymbol{\omega}_{\tilde{\Omega}_y}\right]^{-1}\left[\boldsymbol{y} - \widetilde{\boldsymbol{B}}_{\boldsymbol{y}}\right]; \widetilde{\boldsymbol{\Omega}}^*_{\boldsymbol{y}}\right) \times \\
& \left(\prod_{v=1}^{N+C-1}\prod_{v'=v+1}^{N+C}\begin{bmatrix} \Phi_2(\mu_{v,up}, \mu_{v',up}, \rho_{vv'}) - \Phi_2(\mu_{v,up}, \mu_{v',low}, \rho_{vv'}) \\ -\Phi_2(\mu_{v,low}, \mu_{v',up}, \rho_{vv'}) + \Phi_2(\mu_{v,low}, \mu_{v',low}, \rho_{vv'}) \end{bmatrix}\right) \times \\
& \left(\prod_{v=1}^{N+C}\prod_{g=1}^{G}\Phi_{I_g}\left[\boldsymbol{\omega}^{-1}_{\widehat{\Omega}_{vg}}\boldsymbol{D}_{vg}\left\{\ddot{\boldsymbol{\psi}}_{\boldsymbol{up}} - \ddot{\boldsymbol{B}}_{\tilde{u}}\right\}; \widehat{\boldsymbol{\Omega}}^*_{vg}\right] - \Phi_{I_g}\left[\boldsymbol{\omega}^{-1}_{\widehat{\Omega}_{vg}}\boldsymbol{D}_{vg}\left\{\ddot{\boldsymbol{\psi}}_{\boldsymbol{low}} - \ddot{\boldsymbol{B}}_{\tilde{u}}\right\}; \widehat{\boldsymbol{\Omega}}^*_{vg}\right]\right) \times \\
& \left(\prod_{g=1}^{G-1}\prod_{g'=1}^{G}\Phi_{I_g+I_{g'}-2}\left[\boldsymbol{\omega}^{-1}_{\vec{\Omega}_{gg'}}\boldsymbol{R}_{\boldsymbol{gg'}}\left\{-\ddot{\boldsymbol{B}}_{\tilde{u}}\right\}; \vec{\boldsymbol{\Omega}}^*_{gg'}\right]\right).
\end{aligned}
\tag{21}
$$

In Equation (21), the first component corresponds to the marginal likelihood of the continuous outcomes, the second component corresponds to the likelihood of pairs of outcomes across all ordinal and count outcomes (essentially this combines the second, third, and fourth components of Equation (20)), the third component corresponds to the pairwise likelihood for ordinal/count outcomes and nominal outcomes (this combines the fifth and sixth components of Equation (20)), and the last component corresponds to the pairwise likelihood for the nominal outcomes (this is also the last component of expression (20)). In the MACML approach, all

MVNVD function evaluations greater than two dimensions are evaluated using an *analytic approximation* method rather than a simulation method. This combination of the CML with an analytic approximation for the MVNCD function is effective because the analytic approximation involves only univariate and bivariate cumulative normal distribution function evaluations. The MVNCD analytic approximation method used here is based on linearization with binary variables (see Bhat, 2011). As has been demonstrated by Bhat and Sidharthan (2011), the MACML method has the virtue of computational robustness in that the approximate CML surface is smoother and easier to maximize than are traditional simulation-based likelihood surfaces. We can write the resulting equivalent of Equation (21) computed using the analytic approximation for the MVNCD function as $L_{MACML,q}(\boldsymbol{\delta})$, after introducing the index q for individuals. The MACML estimator is then obtained by maximizing the following function:

$$\log L_{MACML}(\boldsymbol{\delta}) = \sum_{q=1}^{Q} \log L_{MACML,q}(\boldsymbol{\delta}). \tag{22}$$

The covariance matrix of the parameters $\boldsymbol{\delta}$ may be estimated by the inverse of Godambe's (1960) sandwich information matrix (see Zhao and Joe, 2005; Bhat, 2014).

$$V_{MACML}(\boldsymbol{\delta}) = \frac{\left[\hat{\boldsymbol{G}}(\boldsymbol{\delta})\right]^{-1}}{Q} = \frac{\left[\hat{\boldsymbol{H}}^{-1}\right]\left[\hat{\boldsymbol{J}}\right]\left[\hat{\boldsymbol{H}}^{-1}\right]'}{Q}, \tag{23}$$

$$\text{with } \hat{\boldsymbol{H}} = -\frac{1}{Q}\left[\sum_{q=1}^{Q} \frac{\partial^2 \log L_{MACML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\right]_{\hat{\boldsymbol{\delta}}_{MACML}}$$

$$\hat{\boldsymbol{J}} = \frac{1}{Q}\sum_{q=1}^{Q} \left[\left(\frac{\partial \log L_{MACML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\right)\left(\frac{\partial \log L_{MACML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}\right)\right]_{\hat{\boldsymbol{\delta}}_{MACML}} \tag{24}$$

An alternative estimator for $\hat{\boldsymbol{H}}$ may be obtained by computing the quantity below for each decision-maker, and averaging across decision-makers:

$$\hat{\boldsymbol{H}} \text{ for each } q = \left(\begin{array}{l} \left[\frac{\partial \log\left[f_H(\boldsymbol{y}\,|\,\widetilde{\boldsymbol{B}}_y, \widetilde{\boldsymbol{\Omega}}_y)\right]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[f_H(\boldsymbol{y}\,|\,\widetilde{\boldsymbol{B}}_y, \widetilde{\boldsymbol{\Omega}}_y)\right]}{\partial \boldsymbol{\delta}'}\right] + \\ \sum_{n=1}^{N-1}\sum_{n'=n+1}^{N}\left[\frac{\partial \log\left[\Pr(j_n = a_n, j_{n'} = a'_n)\right]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[\Pr(j_n = a_n, j_{n'} = a_{n'})\right]}{\partial \boldsymbol{\delta}'}\right] + \\ \sum_{c=1}^{C-1}\sum_{c'=c+1}^{C}\left[\frac{\partial \log[\Pr(k_c = r_c, k_{c'} = r_{c'})]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[\Pr(k_c = r_c, k_{c'} = r_{c'})\right]}{\partial \boldsymbol{\delta}'}\right] + \\ \sum_{n=1}^{N}\sum_{c=1}^{C}\left[\frac{\partial \log[\Pr(j_n = a_n, k_c = r_c)]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[\Pr(j_n = a_n, k_c = r_c)\right]}{\partial \boldsymbol{\delta}'}\right] + \\ \sum_{n=1}^{N}\sum_{g=1}^{G}\left[\frac{\partial \log\left[\Pr(j_n = a_n, i_g = m_g)\right]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[\Pr(j_n = a_n, i_g = m_g)\right]}{\partial \boldsymbol{\delta}'}\right] + \\ \sum_{c=1}^{C}\sum_{g=1}^{G}\left[\frac{\partial \log\left[\Pr(k_c = r_c, i_g = m_g)\right]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log[\Pr(k_c = r_c, i_g = m_g)]}{\partial \boldsymbol{\delta}'}\right] \\ \sum_{g=1}^{G-1}\sum_{g'=g+1}^{G}\left[\frac{\partial \log\left[\Pr(i_g = m_g, i_{g'} = m_{g'})\right]}{\partial \boldsymbol{\delta}}\right]\left[\frac{\partial \log\left[\Pr(i_g = m_g, i_{g'} = m_{g'})\right]}{\partial \boldsymbol{\delta}'}\right] + \end{array} \right) \tag{25}$$

3.2. Positive Definiteness

The matrix $\widetilde{\boldsymbol{\Omega}}$ for each household has to be positive definite. The simplest way to guarantee this in our mixed model system is to ensure that the $(L \times L)$ correlation matrix $\boldsymbol{\Gamma}$ is positive definite, and each matrix $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ (g=1,2,…,G) is also positive definite. An easy way to ensure the positive-definiteness of these matrices is to use a Cholesky decomposition and parameterize the CML function in terms of the Cholesky parameters. Further, because the matrix $\boldsymbol{\Gamma}$ is a correlation matrix, we write each diagonal element (say the aa^{th} element) of the lower triangular Cholesky matrix of $\boldsymbol{\Gamma}$ as $\sqrt{1-\sum_{j=1}^{a-1} p_{aj}^2}$, where the p_{aj} elements are the Cholesky factors that are to be estimated. In addition, note that the top diagonal element of each $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ matrix has to be normalized to one (as discussed in Section 2.2), which implies that the first element of the Cholesky matrix of each $\breve{\boldsymbol{\Lambda}}_{\mathbf{g}}$ is fixed to the value of one.

4. SIMULATION EXPERIMENT

In this section, we present the design of, and results from, a simulation experiment to evaluate the performance of the MACML approach to recover parameters in a GHDM system from different finite sample sizes. For ease in interpretation and understanding, the simulation design is motivated from an integrated land use-transportation context. Specifically, consider the situation where an analyst wants to examine residential location choices and travel choices of an individual using a cross-sectional data set, with a specific interest on whether (and how much) a neo-urbanist design (compact built environment design, high bicycle lane and roadway street density, good land-use mix, and good transit and non-motorized mode accessibility/facilities) would help in reducing motorized auto ownership of the household of which the individual is a part, and in influencing the individual's commute mode in a way that reduces solo auto mode use. In doing so, the analyst should consider what is commonly labeled as residential self-selection; that is, cross-sectional data reflect residential location preferences co-mingled with the travel preferences of individuals. For example, individuals who have an overall travel freedom and privacy orientation (typically associated with auto inclination) may locate themselves in suburban/rural neighborhoods (low population density, low bicycle lane and roadway street density, primarily single use residential land use, and auto-dependent urban design), own many motorized autos, and favor driving alone to work and other activities. On the other hand, a household whose members have a green and active lifestyle propensity may seek out urban neighborhoods so they can pursue their activities using non-motorized and transit modes of travel. If such self-selection effects in residence choices are ignored, when actually present, the result can be a "spurious" causal effect of neighborhood attributes on auto ownership and travel, and potentially misinformed BE design policies (see a detailed discussion in Bhat *et al*., 2014a) . But the self-selection may not be based solely on residential choice, and can also be based on auto ownership choice. Thus, individuals with a travel freedom and privacy orientation may both prefer more autos as well as be predisposed to traveling in motorized vehicles to work and other activities. As a consequence, any effect of the number of motorized vehicles on auto travel will be moderated by the travel freedom and privacy orientation of the individual.

The potential self-selection effects above can be acknowledged by considering workers' decisions associated with residential location, auto ownership, commute travel mode choice, and some quantification of non-commute travel as a multi-dimensional bundle. It is in this context

that our simulation design is set. Residential location choice is represented as a nominal discrete choice among a multinomial set of three different types of BE designs as captured by designations as urban, suburban, and rural neighborhoods (these designations can be combinations of housing density and employment density; see Kim and Brownstone, 2013, Paleti *et al.*, 2013, Cao and Fan, 2012, and Bhat *et al*., 2014a, who all use such a density-based classification scheme as a representation of residential location choice as this simplifies the representation of residential choice alternatives and also alleviates the problem of strong multi-collinearity of density with other built environment attributes). In addition, we also use a second continuous outcome, the (logarithm of) commute distance for the individual, to characterize residential location choice. This is because it has been well established in the literature that commute distance is one of the most important determinants of residential location (see, for example, Clark *et al*., 2003, Rashidi *et al*., 2012).[3] Auto ownership is a count outcome, while commute travel mode choice is represented as a second nominal choice in the system from among three different modes of transportation – non-motorized transportation (NM), public transportation (PT), and motorized (private) transportation or MT (either as a driver or a passenger). Non-commute travel is quantified as a multi-dimensional bundle of three ordinal variables that relate to intensities (occurrences) of weekly non-commute travel by NM, by PT, and by MT. However, since most household travel surveys capture only daily travel, we suppose that use of alternative modes over longer periods of time (as would be important particularly for NM and PT use) is obtained through an ordinal categorical indicator response from among three possibilities: (1) Never or about once a week, (2) about 2-3 times a week, and (3) four or more times in a week (see Sener *et al.*, 2009 for a survey that captures non-commute travel in such ordinal categories). In all, our system has seven endogenous outcomes/indicators, with one continuous outcome (commute distance), three ordinal indicators (non-commute travel occurrences by NM, PT, and MT), one count outcome (auto ownership), and two nominal outcomes (residential choice location based on density categorization and commute mode choice). While modeling all of these as a joint bundle, we also accommodate structural relationships among the endogenous outcomes/indicators. In particular, we specify that commute distance and auto ownership will affect commute mode choice, and the geographic area of

[3] The implicit assumption here is that work location choices precede residential choice. While it is certainly possible that residential moves may motivate job moves, earlier research using panel data suggests that a vast majority (85% or more) of residential relocations follow a job move (see Rashidi *et al*., 2012).

residential location (urban, suburban, or rural) will affect auto ownership, commute distance, and non-commute travel occurrences by NM and PT.

4.1. Experimental Design

Consider a multi-dimensional choice bundle of residential location and activity-travel behavior, as dicussed in the previous section. In previous studies on the integration of land-use patterns and activity-travel behavior, such as Pinjari *et al.* (2011) and Bhat *et al*. (2014a), correlated unobserved effects among multiple (but limited) choice dimensions were captured through the error terms of the many individual dimensions, resulting in a relatively large dimensional covariance matrix. The difference between these earlier studies and this simulation study is that, as dicussed in Section 1, the covariance in a large number of choice dimensions is captured in a parsimonious manner through a factor-analytic structure where the choice dimensions are a function of a smaller dimension of correlated latent constructs. In addition, such a specification provides structure to the jointness among the choice dimensions by appealing to theoretical psychological constructs.

4.2. The Structural Equation System

Four latent variables associated with lifestyle and attitudes are employed as psychological constructs impacting the multi-dimensional choice bundle of residential location and activity-travel behavior (we use several latent variables here to examine the ability of the MACML approach to recover parameters even in the presence of quite a few latent constructs). The latent variables are shown in Figure 1, where the ovals represent the latent constructs, while rectangles represent observed explanatory variables. The first latent factor is *green lifestyle propensity* (z_1^*) or the individual's level of environmental consciousness, which is specified to be a function of whether the individual has a Bachelor's degree or higher ($w_1; w_1 = 1$ if individual has a Bachelor's degree or higher and 0 otherwise). This reflects the finding from earlier studies that individuals with a Bachelor's degree or higher tend to be more active proponents and followers of ecologically friendly lifestyles (Paleti *et al.*, 2013). The specified value of this effect (embedded within the $\boldsymbol{\alpha}_1$ vector) is 0.8. The second factor is *activity seeking personality* (z_2^*), or the individual's propensity to partake in various non-work activities. This latent factor is

specified as a function of whether or not a person lives alone ($w_2; w_2 = 1$ if individual lives alone and 0 otherwise). It captures the well-established finding that individuals living alone are more likely to partake in non-work activities outside the home than are individuals living in family settings (see Kim, 2011, and Champion, 2011). The specified value of this effect in the simulation design (as embedded within the $\boldsymbol{\alpha}_2$ vector) is 0.3. The third factor is *travel freedom affinity* (z_3^*), generally associated with travel comfort/convenience and a sense of control over the travel experience. This latent variable is specified to be associated with men (w_3; $w_3 = 1$ if individual is male and 0 otherwise), and high income individuals ($w_4; w_4 = 1$ if individual earns a high income and zero otherwise), as found in Schwanen and Mokhtarian (2007). The design values of these effects in the simulation (as embedded within the $\boldsymbol{\alpha}_3$ vector) are 0.2 and 0.5, respectively. The final latent factor corresponds to *privacy desire* (z_4^*), with the expectation that high income individuals generally value privacy more than their lower income peers (see Jansen, 2012, Shiftan *et al*., 2008, and Day, 2000). The specified value of this effect is 0.2. In the vector notation of Equation (2), the effects in Figure 1 may be written as follows:

$$\begin{bmatrix} z_1^* = \text{GLP} \\ z_2^* = \text{ASP} \\ z_3^* = \text{TFA} \\ z_4^* = \text{PD} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} w_1 = \text{Bachelor's degree or higher or not} \\ w_2 = \text{Living alone or not} \\ w_3 = \text{Male or not} \\ w_4 = \text{high income or not} \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix},$$

where GLP is *green lifestyle propensity* , ASP is *activity seeking personality*, TFA is *travel freedom affinity*, and PD is *privacy desire*. The parameters in the matrix $\boldsymbol{\alpha}$ to be estimated can be stacked into a vector $\text{Vech}(\boldsymbol{\alpha}) = [\alpha_{11} = 0.8,\ \alpha_{22} = 0.3,\ \alpha_{33} = 0.2,\ \alpha_{34} = 0.5,\ \alpha_{44} = 0.2]$. The correlation matrix of the error vector $\boldsymbol{\eta}$ is specified as follows:

$$Var(\boldsymbol{\eta}) = \boldsymbol{\Gamma} = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.6 \\ 0.0 & 0.0 & 0.6 & 1.0 \end{bmatrix} = \mathbf{L}_{\boldsymbol{\Gamma}} \mathbf{L}_{\boldsymbol{\Gamma}}' = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.6 & 1.0 \end{bmatrix} \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.6 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

In the matrix above, we allow a correlation (entry of 0.6) between the latent personality constructs of *travel freedom affinity* (z_3^*) and *privacy desire* (z_4^*), to reflect the existence of the unobserved underlying value of individuality that affects both of these personality constructs. To ensure the positive definiteness of $\boldsymbol{\Gamma}$, a Cholesky decomposition is conducted. In our specification, a single element is to be estimated in matrix $\boldsymbol{\Gamma}$: $l_{\Gamma 33} = 0.6$.

4.3. The Measurement Equation System

The measurement equation system includes the non-nominal equation system $\vec{y} = \vec{\gamma}x + \vec{d}z^* + \vec{\varepsilon}$ (Equation (13) earlier) as well as the nominal equation system $U = bx + \varpi z^* + \varsigma$ (Equation (14) earlier). Within each of these systems, there are exogenous and endogenous outcome effects (embedded in $\vec{\gamma}$ and $\breve{\gamma}$ for the non-nominal system and in b for the nominal system), as well as latent construct effects (embedded in $\vec{d}$ and ϖ). The simulation design effects specified for the non-nominal equation system (including both the exogenous and latent construct effects) are presented in Figure 2a, while the corresponding effects for the nominal equation system are presented in Figure 2b. Finally, the endogenous variable effects (that is, the inter-relationships between the endogenous outcomes/indicators, which can only be recursive as discussed in Section 2.2), are presented in Figure 2c. Each of these effects is discussed in turn in the subsequent sections, while Section 4.2.4 brings all parameters to be estimated together in the measurement equation system. Note that the design considers four exogenous variables: (1) whether the individual is an immigrant or not (a dummy variable "immigrant" taking the value of 1 if the individual is born in the US and 0 otherwise), (2) whether the individual owns or rents her/his household (a dummy variable "owns hh" taking the value of 1 if the individual owns her/his household and 0 otherwise), (3) number of children less than 11 years of age, and (4) number of young active adults (to represent the presence of the so-called millenials born between 1981 and 1996).

4.2.1. Non-Nominal Equation System with Exogenous and Latent Construct Effects

This system is shown diagrammtically in Figure 2a. Immigrant status positively influences (log) commute distance, as it has been observed that immigrants have longer commutes than do non-immigrants (see Paleti *et al*., 2013). Further, individuals with young children are less likely to

travel by non-motorized modes and more likely to travel by motorized vehicles (as they undertake pick up/drop off activities; see Sener *et al*., 2009). Also, in the simulation design, we specify the number of young active adults in the individual's household to negatively influence travel by motorized vehicles, as households with millenials tend to undertake their out-of-home activities less using private vehicles (see Bhat *et al*., 2014a). A total of four exogenous variable effects are specified above. However, there are also constants to be specified in the (log) commute distance equation, and for the latent propensities for the ordinal indicators. The constant in the (log) commute disance equation is arbitrarily set to 1.0, while the constant effects for the ordinal indicators are all specified to be -1.0.

A total of six latent construct effects are also specified. As expected, a green lifestyle propensity increases non-commute travel occurrences by the non-motorized (NM) mode, while an activity seeking personality leads to more non-commute travel occurrences by motorized transportation (MT). As alluded to earlier, we expect travel freedom affinity to be positively related to auto ownership, with a similar positive effect of privacy desire on auto ownership. Finally, privacy desire is also specified to negatively impact travel occurrences by public transportation (PT), while travel freedom is positively related to commute distance (see, for example, Schwanen and Mokhtarian, 2007).

As presented in Equation (13), the covariance matrix $\ddot{\boldsymbol{\Sigma}}$ of random error $\ddot{\boldsymbol{\varepsilon}}$ for non-nominal indicators is restricted to be diagonal, with elements corresponding to ordinal and count indicators being normalized to 1. This leaves the variance component for the continuous outcome (logarirthm of commute distance), which is specified to be 1.25 in the simulatin design. Thus, one element is to be estimated in the matrix $\ddot{\boldsymbol{\Sigma}}$ is 1.25, which we write as $l_{\tilde{\Sigma}11} = 1.25$.

There are three ordinal outcomes (non-commute travel occurrences by NM, PT, and MT), in the simulation design, which leads to the need to specify $\widetilde{\psi}_{n,2}$ for each ordinal outcome n $(n = 1,\ 2,\ 3)$ (see discussion in Section 2.2). All of these threshold values are set to 1.5. In addition, we need to specify the parameters in the threshold function for the count outcome (corresponding to auto ownership). This refers to the coefficient vector $\breve{\boldsymbol{\gamma}}$, the flexibility parameter vector $\boldsymbol{\varphi}_c = (\varphi_{c,1}, \varphi_{c,2}, \ldots \varphi_{c,e_c^*})'$, and the dispersion parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_C)'$. For the $\breve{\boldsymbol{\gamma}}$ coefficient vector, we include only a constant effect and another endogenous effect (the latter is discussed in the next section). The coefficient on the constant is

specified to be 1.0. For the flexibility vector, we will drop the index *c* since we have only one outcome in the simulation design. We also specifiy a single flexibility parameter $\varphi_1 = 0.75$. For the dispersion parameter vector (which collapses to a scalar because there is only a single count outcome), we specify $\theta = 2.0$.

4.2.2. Nominal Equation System with Exogenous and Latent Construct Effects

Five exogenous effects and nine latent construct effects are specified here (see Figure 2b). All of the exogenous effects specified have been reasonably well established in earlier studies. Immigrants tend to cluster in urban neighborhoods (see Bhat *et al*., 2013), while those who own households are less likely to reside in urban neighborhoods. There is also evidence that individuals with children tend to favor suburban neighorhoods due to the open spaces and good quality schools (Aditjandra *et al*., 2012), as do households with many young active adults (Brownstone and Golob, 2009). Further, as has been found in many earlier studies, immigrants, more so than US-born individuals, tend to use public transportation for their commute. In addition to the variable effects above, we also allow constants in two of the utilities for residential location and two of the utilities for commute mode. Specifically, we use a constant effect of 0.2 in the urban location utility and 0.5 in the suburban location utility (with the rural constant specified to be zero for identification). Also, we use a constant effect of -0.5 for the PT mode, and -0.2 for the NM mode (with the MT mode constant specified to be zero for identification).

The latent construct effects specified are rather intuitive. These are specified to shift the utility of specific alternatives of the nominal variables. Essentially, then, in the notation of Section 2.2, $\boldsymbol{\varpi}_g = \boldsymbol{\vartheta}_g$, because $\boldsymbol{\beta}_g$ is an identity matrix. Thus, for convenience, we will refer to the parameters to be estimated as being elements of $\boldsymbol{\varpi}_g$, which is the same as the elements of $\boldsymbol{\vartheta}_g$. For the residential location nominal outcome, individuals with a green lifestyle propensity and activity seeking personality tend to reside in urban neighborhoods, so that they can pursue their desired lifestyles due to greater opportunities for social interaction and the buzz of city life (Schwanen and Mokhtarian, 2007). Individuals with a travel freedom affinity prefer suburban and rural neighborhoods over urban neighborhoods, while those who have a privacy desire are likely to locate themselves in rural neighborhoods. For the commute mode nominal outcome,

green lifestyle propensity is specified to negatively affect MT mode utility and positively affect use of NM modes. On the other hand, travel freedom affinity increases the propensity to use the MT mode, and privacy desire should reduce PT mode use.

The covariance matrix of $\boldsymbol{\varsigma}$ is specified as follows.

$$\text{Var}(\boldsymbol{\varsigma}) = \boldsymbol{\Lambda} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.70 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.70 & 1.49 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.60 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.60 & 1.36 \end{bmatrix} \tag{30}$$

$$= \mathbf{L}_{\boldsymbol{\Lambda}} \mathbf{L}'_{\boldsymbol{\Lambda}} = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.70 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.60 & 1.00 \end{bmatrix} \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.70 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.60 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

In the matrix $\boldsymbol{\Lambda}$, four elements are to be estimated ($l_{\Lambda 32} = 0.70, l_{\Lambda 33} = 1.49, l_{\Lambda 65} = 0.60, l_{\Lambda 66} = 1.36$).

4.2.3. Endogenous Outcome Effects

These effects correspond to recursive effects among the endogenous outcomes, as discussed just before Section 4.1. These are parts of the $\vec{\gamma}$ matrix (for the continuous/ordinal outcomes), the $\breve{\gamma}$ matrix (for the count outcomes), and the $\boldsymbol{b}$ matrix (for the nominal outcomes). The important point is that these are "cleansed" effects after accommodating unobserved covariance effects among the endogenous outcomes engendered by the presence of latent constructs, as discussed in the previous two sections. Figure 2c provides a pictorial representation for these endogenous effects. For the continuous/ordinal outcomes, we specify that urban dwelling leads to a shorter commute distance, and more non-commute travel occurrences by the NM and PT modes (see Paleti *et al.*, 2013). For the auto count variable, several earlier studies have established that urban dwellers tend to own fewer vehicles even after accounting for any residential self-selection effects (see, for example, Bhat and Guo, 2007). This effect is specified through the threshold in the count model; that is, in the $\boldsymbol{x}$ vector with a corresponding coefficient vector $\breve{\gamma}$ (the $\breve{\gamma}$

matrix becomes a vector in our simulation design because there is only one count variable). In particular, in our formulation of the count model, a positive coefficient element in $\breve{\gamma}$ implies that an increase in the corresponding element of $\boldsymbol{x}$ shifts all the thresholds toward the left of the auto ownership propensity scale (see Castro *et al.*, 2011), which has the effect of reducing the probability of zero cars, while a negative coefficient in $\breve{\gamma}$ implies that an increase in the corresponding element of $\boldsymbol{x}$ shifts all the thresholds toward the right of the auto ownership propensity scale, which has the effect of increasing the probability of zero cars. In our simulation design, we impose a negative coefficient of -0.5.

For the nominal variables, our design specifies a positive effect of urban dwelling on the propensity to use PT as the commute mode, and a negative effect of car ownership and commute distance on the use of the NM mode for the commute.

4.2.4. Overall Measurement Equation System

The overall measurement equation for the vector $\boldsymbol{yU} = \left[\tilde{\boldsymbol{y}}', \boldsymbol{U}'\right]'$ takes the following mathematical form:

$$\begin{bmatrix} y_1 = \log(commute\ dist) \\ \tilde{y}_1^* = NC\ \ propensity\ by\ NM \\ \tilde{y}_2^* = NC\ \ propensity\ by\ PT \\ \tilde{y}_3^* = NC\ propensity\ by\ MT \\ \breve{y}_1^* = auto\ own.\ propensity \\ U_{1,urban} \\ U_{1,suburban} \\ U_{1,rural} \\ U_{2,MT} \\ U_{2,PT} \\ U_{2,NMT} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.3 \\ -1.0 & 0.0 & 0.0 & -0.2 & 0.0 & 0.0 & 0.0 & 0.6 \\ -1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 \\ -1.0 & 0.0 & 0.0 & 0.4 & -0.3 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.4 & -0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.0 & 0.2 & 0.3 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.5 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 \\ -0.2 & 0.0 & 0.0 & 0.0 & 0.0 & -0.6 & -0.4 & 0.0 \end{bmatrix} \begin{bmatrix} \text{Constant} \\ \text{Immigrant household} \\ \text{Own hh} \\ \#\ \text{of Children} < 11\ \text{yrs} \\ \#\ \text{of young active adults} \\ \text{Commute distance} \\ \text{auto ownership} \\ \text{urban dwelling} \end{bmatrix}$$

$$+ \begin{bmatrix} 0.0 & 0.0 & 0.2 & 0.0 \\ 0.6 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.5 \\ 0.0 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.5 \\ 0.2 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.5 \\ -0.4 & 0.0 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.2 \\ 0.6 & 0.0 & 0.0 & 0.0 \end{bmatrix} \begin{bmatrix} z_1^* = \text{GLP} \\ z_2^* = \text{ASP} \\ z_3^* = \text{TFA} \\ z_4^* = \text{PD} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \tilde{\boldsymbol{\varepsilon}}_1 \\ \tilde{\boldsymbol{\varepsilon}}_2 \\ \tilde{\boldsymbol{\varepsilon}}_3 \\ \breve{\boldsymbol{\varepsilon}}_1 \\ \boldsymbol{\varsigma}_{11} \\ \boldsymbol{\varsigma}_{12} \\ \boldsymbol{\varsigma}_{13} \\ \boldsymbol{\varsigma}_{21} \\ \boldsymbol{\varsigma}_{22} \\ \boldsymbol{\varsigma}_{23} \end{bmatrix}$$

Based on the above, and using the notations employed in Section 2.2., the parameters to be estimated in the measurement equation above include the following:

$\text{Vech}(\vec{\boldsymbol{\gamma}}) = [\gamma_{11} = 1,\ \gamma_{12} = 0.5,\ \gamma_{18} = \text{-}0.3,\ \tilde{\gamma}_{11} = \text{-}1,\ \tilde{\gamma}_{14} = \text{-}0.2,\ \tilde{\gamma}_{18} = 0.6,\ \tilde{\gamma}_{21} = \text{-}1,\ \tilde{\gamma}_{28} = 0.2,\ \tilde{\gamma}_{31} = \text{-}1,$ $\tilde{\gamma}_{34} = 0.4,\ \tilde{\gamma}_{35} = \text{-}0.3]$,

$\text{Vech}(\breve{\boldsymbol{\gamma}}_1) = [\breve{\gamma}_{11} = 1, \breve{\gamma}_{18} = -0.5]$ (this is the vector corresponding to the coefficients on the constant and the urban dwelling variable embedded in the threshold in the auto ownership count model), $\text{Vech}(\boldsymbol{b}) = [b_{111} = 0.2, b_{112} = 0.4, b_{113} = -0.5, b_{121} = 0.5,\quad b_{123} = 0.2, b_{124} = 0.3, b_{221} = -0.5,$ $b_{222} = 0.3, b_{228} = 0.2,\qquad b_{231} = -0.2, b_{236} = -0.6, b_{237} = -0.4]$,

$\text{Vech}(\vec{\boldsymbol{d}}) = [d_{13} = 0.2, \tilde{d}_{11} = 0.6, \tilde{d}_{24} = -0.5, \tilde{d}_{32} = 0.3, \breve{d}_{13} = 0.4, \breve{d}_{14} = 0.5]$, and

$\text{Vech}(\boldsymbol{\varpi}) = [\varpi_{111} = 0.2, \varpi_{112} = 0.5, \varpi_{123} = 0.3, \varpi_{133} = 0.4, \varpi_{134} = 0.5, \varpi_{211} = -0.4, \varpi_{213} = 0.2,$
$\varpi_{224} = -0.2, \varpi_{231} = 0.6]$.

In addition, we have the variance component for the continuous outcome $l_{\tilde{\Sigma}11} = 1.25$. the flexibility parameter $\varphi_1 = 0.75$ and the dispersion parameter vector $\theta = 2.0$ for the auto ownership count outcome, the single element $(l_{\Gamma 33} = 0.6)$ in the covariance matrix of the error terms in the structural equation system, and the parameters for the covariance matrix of the nominal outcomes: $l_{\Lambda 32} = 0.70, l_{\Lambda 33} = 1.49, l_{\Lambda 65} = 0.60, l_{\Lambda 66} = 1.36$.

4.4. Data Generation Process

To generate the simulated dataset, the first step is to develop values for the exogenous variables in the vectors $\boldsymbol{w}$ and $\boldsymbol{x}$. There are six dummy variables in these two vectors, corresponding to bachelor's degree or higher (w_1), person lives alone (w_2), male (w_3), high income (w_4), immigrant (x_1), and own household (x_2). To construct these dummy variables, independent values were drawn from the standard uniform distribution. If the value drawn was less than 0.5, the value of '0' was assigned for the dummy variable. Otherwise, the value of '1' was assigned. For the two count exogenous variables corresponding to the number of children less than 11 years of age and the number of young active adults, a maximum value for each variable was first assigned (three for the first, and five for the second). Then, the range of the uniform distribution (0 to 1) was divided into as many equal ranges as the maximum value for the count plus one. Independent draws for the two count variables were made from the uniform distribution, and the value assigned of the count was based on the range in which a draw fell. For example, for the "number of children less than 11 years" variable, four equal intervals were created: [0.00, 0.25), [0.25, 0.50), [0.50, 0.75), or [0.75, 1.00]. If a draw was between 0.00 and 0.25 (but not including 0.25 exactly), a value of 0 was assigned for the variable; if a draw was between 0.25 and 0.5 (but not including 0.50 exactly), a value of 1 was assigned and so on.

The procedure above is used to construct a synthetic sample of Q=1000, 2000, and 3000 realizations of the exogenous variables. We consider different samples sizes to assess the accuracy and appropriateness of the asymptotic properties of the MACML estimator for finite sample sizes. Once drawn, the exogenous variables are held fixed for the rest of the simulation exercise. In the rest of this section, we will discuss the procedure to generate the data set

assuming Q=1000 observations (the same procedure may be applied for Q=2000 and Q=3000 observations). For each of the 1000 observations, a specific realization of the vector $(\vec{\varepsilon}', \varsigma')'$ $[(E+\ddot{G})\times 1)]$ is drawn from the multivariate distribution with mean $\mathbf{0_{11}}$ (a column vector of zero values of dimension 11) and covariance structure given by $\boldsymbol{\Omega}$ in Equation (16). The sub-vector of the mean vector $\boldsymbol{B_2}$ that corresponds to the utilities of the three residential choice alternatives is also computed using the expression in Equation (16). Then, the realization corresponding to $\boldsymbol{\varsigma}_1 = (\varsigma_{11}, \varsigma_{12}, \varsigma_{13})'$ (the error terms drawn for the three residential choice alternatives) is added to the mean vector for the three residential choice alternatives to obtain the realization of $\boldsymbol{U}_1 = (U_{1,urban}, U_{1,suburban}, U_{1,rural})'$ for each observation. The alternative with the highest utility value is then picked, and identified as the chosen residential choice alternative for each observation. Next, the continuous outcome y_1 is generated based on the exogenous variables, the design parameters, and the realization of the value of ε_1 from earlier. Similarly, the latent continuous values for the ordinal indicators are also generated, and then translated into ordinal outcomes based on comparison with the corresponding design thresholds. For the auto ownership count outcome, the latent continuous value is generated exactly as for the ordinal indicators. However, the thresholds also need to be computed based on the design parameters as well as the realized actual value of the urban residential choice outcome. Then, the latent continuous value for the count outcome is translated into an actual count outcomes based on a comparison with the computed thresholds. Finally, the utilities for the commute mode choice alternatives are computed based on exogenous variables, all realized values of the other endogenous outcomes, as well as the realization corresponding to $\boldsymbol{\varsigma}_2 = (\varsigma_{21}, \varsigma_{22}, \varsigma_{23})'$ from earlier (the error terms drawn for the three commute mode choice alternatives).

The above data generation process is undertaken 200 times with different realizations of the random errom components to generate 200 datasets for each sample size. The MACML estimator is applied to each dataset to estimate the 57 underlying parameters. A single random permutation is generated for each individual (the random permutation varies across individuals, but is the same across iterations for a given individual) to decompose the MVNCD function into a product sequence of marginal and conditional probabilities (see Section 2.1 of Bhat, 2011)[4]. In

[4]Technically, the MVNCD approximation should improve with a higher number of permutations in the MACML approach. However, when we investigated the effect of different numbers of random permutations per individual, we

order to obtain a sense of the approximation error (explained in the following subsection), 10 datasets are randomly selected from the 200 datasets for each sample size (*i.e.*, *N*=1000, 2000, and 3000). Then the estimator is applied to each dataset 10 times with different permutations. Based on the 100 estimations (10 datasets × 10 runs with different permutations per dataset) for each sample size, the estimates of approximation error are derived.

4.5. Performance Evaluation

The performance of the MACML inference approach in estimating the parameters of the GHDM and the corresponding standard errors is evaluated as follows (the discussion below is for a specific sample size; the same procedure is applied for evaluating performance with the different sample sizes of 1000. 2000, and 3000.

(1) Estimate the MACML parameters for the 200 datasets. Estimate the standard errors using the Godambe (sandwich) estimator.

(2) Compute the mean for each model parameter across the 200 datasets to obtain a mean estimate. Compute the **absolute percentage (finite sample) bias** (APB) of the estimator as:

$$APB = \left| \frac{\text{mean estimate - true value}}{\text{true value}} \right| \times 100 \tag{31}$$

(3) Compute the standard deviation of the mean estimate across the 200 datasets, and label this as the **finite sample standard error or FSSE** (essentially, this is the empirical standard error).

(4) Compute the mean standard error for each model parameter across the 200 datasets, and label this as **the asymptotic standard error or ASE** (essentially this is the standard error of the distribution of the estimator as the sample size gets large). Compute the ASE as a percentage of the mean estimate.

(5) Next, to evaluate the accuracy of the ASE formula as computed using the MACML inference approach for the finite sample size used, compute the relative efficiency of the estimator as:

$$\text{Relative efficiency} = \frac{\text{ASE}}{\text{FSSE}} \tag{33}$$

In general, the relative efficiency values should be less than 1, since we expect the ASE to be less than the FSSE. But, because we are using only a limited number of datasets to compute

---

noticed little difference in the estimation results between using a single permutation and higher numbers of permutations, and hence we settled with a single permutation per individual.

the FSSE, values higher than 1 can also occur. The more important point is to examine the closeness between the ASE and FSSE, as captured by the relative efficiency value.

(6) For each of the randomly selected 10 datasets (out of the 200 datasets), compute the mean estimate (10ME) for each model parameter across the 10 random permutations used for that dataset (to evaluate the MVNCD function). Then, for each of the 10 datasets, compute the standard deviation of the parameter values (across permutations) around the 10ME value. Take the mean of the standard deviation value across all the 10 datasets, and label this as the **approximation error** (**APERR**).

4.6. Simulation Results

The simulation results for Q=1000, 2000, and 3000 are presented in Tables 1, 2, and 3, respectively. The tables provide the true value of the parameters (second column), followed by the parameter estimate results and the standard error estimate results.

A number of observations may be made from the tables. First, the ability of the MACML approach to recover the parameters underlying the GHDM model is pretty good, as may be observed from the magnitude of the absolute percentage bias (APB) values. In particular, the mean APB value (see the bottom row of the third column under "Parameter Estimates") is 10.55% with 1000 observations, reducing to 8.01% with 2000 observations and further to 4.40% with 3000 observations. Overall, the difference between 1000 and 2000 observations in more accurately recovering parameters is moderate. The real difference in the APB values appears when moving from 2000 observations to 3000 observations, suggesting that there are critical thresholds in the number of observations. Second, the parameters corresponding to the effects of exogenous variables on the latent variables (that is, the elements of $\text{Vech}(\boldsymbol{\alpha})$), the effects of the latent variables on the non-nominal outcomes (that is, the elements of $\text{Vech}(\vec{\boldsymbol{d}})$), and the effects of the latent variables on the nominal outcomes (that is, the elements of $\text{Vech}(\boldsymbol{\varpi})$) are generally relatively more difficult to estimate compared to other parameters. Thus, for the case of Q=1000 observations, the APB value for the $\text{Vech}(\boldsymbol{\alpha})$ elements range from 11.30% to 19.01% with a mean APB of 15.36), the APB value for the $\text{Vech}(\vec{\boldsymbol{d}})$ elements range between 12.39% and 24.67% (with a mean APB of 19.44%), and the APB values for the $\text{Vech}(\boldsymbol{\varpi})$ elements range from 3.960% to 27.31% (with a mean of 13.82%). For datasets with 1000, 2000, and 3000

observations, the mean APB values for (a) the Vech($\boldsymbol{\alpha}$) elements are 15.36%, 11.38%, and 3.12%, respectively, (b) the Vech($\vec{\boldsymbol{d}}$) elements are 19.44%, 15.66%, and 6.52%, respectively, and (c) for the Vech($\boldsymbol{\varpi}$) elements are 13.82%, 12.38%, and 3.91%, respectively. The relatively less accurate recovery of these sets of parameters is intuitive. As one can notice from Equations (15) and (16), the only way to disentangle the effects of the $\vec{\boldsymbol{d}}$ matrix and the $\boldsymbol{\alpha}$ matrix in the first (non-nominal) part of Equation (15) is through the identification of the $\vec{\boldsymbol{d}}$ matrix elements from the covariance matrix $\boldsymbol{\Omega}$. Similarly, the only way to disentangle the effects of the ϖ matrix and the $\boldsymbol{\alpha}$ matrix in the second (nominal) part of Equation (15) is through the identification of the $\boldsymbol{\varpi}$ matrix elements from the covariance matrix $\boldsymbol{\Omega}$. As such, the $\vec{\boldsymbol{d}}$ matrix elements and the $\boldsymbol{\varpi}$ matrix elements enter into the covariance matrix $\boldsymbol{\Omega}$ in a non-linear fashion (see Equation 16), and $\boldsymbol{\Omega}$ itself enters into the composite likelihood function (Equation 21) in a complex manner. It is also interesting to note that the improvement in the accuracy of recovery is dramatic for the Vech($\boldsymbol{\alpha}$), Vech($\vec{\boldsymbol{d}}$), and Vech($\boldsymbol{\varpi}$) parameters as one goes from 2000 to 3000 observations, which is essentially driving the substantially overall improved performance with 3000 observations relative to 2000 observations as pointed out earlier. Indeed, with 3000 observations, the APB for these parameters is in the same range as for all other model parameters. Third, moving on to the standard error estimates, the entries in the "finite sample standard error (FSSE)" column indicate that the empirical ability of the MACML estimator to pin down parameters (that is, the precision of parameter recovery) is quite good. In particular, as a percentage of the true values, the mean FSSE values across all parameters are 47.33, 32.60, and 17.86 for 1000, 2000, and 3000 observations, respectively (see the last row of the sub-column entitled "% of true value" under the FSSE column). However, once again, and for the same reason that it is difficult to accurately recover the parameters of Vech($\boldsymbol{\alpha}$), Vech($\vec{\boldsymbol{d}}$), and Vech($\boldsymbol{\varpi}$), the FSSE values are relatively higher for these sets of parameters than for other parameters. For datasets with 1000, 2000, and 3000 observations, the FSSE values as a percentage of the true values for (a) the Vech($\boldsymbol{\alpha}$) elements are 96%, 56.7%, and 32.7%, respectively, (b) the Vech($\vec{\boldsymbol{d}}$) elements are 59.8%, 39.2%, and 17.5%, respectively, and (c) for the Vech($\boldsymbol{\varpi}$) elements are 61.7%, 55.7%, and 26.1%, respectively. Overall, it is difficult to both accurately and precisely recover the effects of exogenous variables on the latent variables (in the

structural equation system) as well as the effects of the latent variables on the outcomes (in the measurement equation system). The suggestion is the exercise of caution when GHDM models with many latent variables are being estimated with few observations. Our results suggest that there may be a need for 3000 observations or so for good accuracy and precision in the estimated coefficients. Of course, the situation is likely to be context-specific, but our simulation analysis does provide some guidance, given that it involves more latent variables than are typically used in extant GHDM models. Fourth, the asymptotic formula of the CML approach performs very well in estimating the FSSEs, based on the relative efficiency (RE values). The mean RE values are 0.857, 0.891, and 0.992 for datasets with observations of 1000, 2000, and 3000, respectively. In general, the FSSE and the ASE values are close to one another regardless of sample size, indicating that the asymptotic formula is performing well in estimating the finite sample standard error even for a sample size of the order of 1000. Finally, the APERR in the last column of all three tables indicates that even a single permutation (for each observation) of the approximation approach used to evaluate the MVNCD function provides adequate precision. For the case with 1000 observations, the values of the APERR range between 0.0002 and 0.046, and the mean APERR is 0.0058. At Q=2000, the minimum and maximum values of the APERR are 0.0002 and 0.025, with the mean APERR decreasing to 0.0052. When Q=3000, the APERR values are in the range of 0.0001 and 0.016, with the mean APERR further decreasing to 0.0022. More importantly, the approximation error (as a percentage of the FSSE), averaged across all the parameters, is of the order of 7%, 4%, and 3% for 1000, 2000, and 3000 observations, respectively. This is clear evidence that the convergent values are about the same for a given data set regardless of the permutation used for the decomposition of the multivariate probability expression.

4.6.1 Effects of Ignoring Latent Construct Effects

This section presents the results of the estimation when the latent variables are ignored, and the resulting dependencies among the multidimensional outcomes are not considered. As discussed earlier in the first part of Section 4, this is equivalent to ignoring all potential self-selection effects, which then should corrupt all endogenous variable effects discussed in Section 4.2.3, and lead to inaccurate and inefficient estimation of other parameters as well. Ignoring the presence of latent variables is tantamount to the restriction in the GHDM model that all elements of the $\vec{d}$

matrix and the $\boldsymbol{\varpi}$ matrix in Equation (15) are zero (no effects of latent variables on any (and all) outcome(s)). But doing so immediately renders all elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ unidentifiable, because the only way these elements are identified is by the relationship between the latent variable vector $\boldsymbol{z}^*$ and the observed outcomes. Thus, we also essentially are setting all elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$ to zero in the restricted model. The resulting equivalent of Equation (15), which we will refer to as the independent model for ease, can be compared with the GHDM model using the adjusted composite log-likelihood ratio test (ADCLRT) value (see Pace *et al*., 2011 and Bhat, 2011 for more details on the ADCLRT statistic, which is the equivalent of the log-likelihood ratio test statistic when a composite marginal likelihood inference approach is used; this statistic has an approximate chi-squared asymptotic distribution).

For the comparison of the GHDM and independent model coefficient estimates (vis-à-vis the true values of the experimental design), we estimate the independent model on the same 200 datasets as we estimated the GHDM model on earlier. Based on the results for the GHDM model, we decided to undertake this comparison only for the case of Q=3000 observations. For each of the 200 data sets, we use the same set of permutations for the joint model and the independent model, so that we are able to appropriately compare the ability to recover parameters from the two models. The simulation results for the independent model are presented in Table 4. For comparison purposes, we also present the results of the GHDM model for those coefficients estimated in the independent model. The GHDM model mean APB is 4.15 relative to the independent model mean APB of 17.85. In particular, the APB values for the estimated coefficients on the endogenous effects (γ_{18}=-0.3, $\widetilde{\gamma}_{18}$=0.6, $\widetilde{\gamma}_{28}$=0.2, $\breve{\gamma}_{18} = -0.5$, $b_{228} = 0.2$, $b_{236} = -0.6, b_{237} = -0.4$) are very high in the independent model relative to the GHDM model. This is to be expected. For instance, consider the effect of urban dwelling on auto ownership (that is, the coefficient $\breve{\gamma}_{18} = -0.5$). The probability of residing in an urban area and the propensity to own autos are negatively correlated because of the latent travel freedom affinity (TFA) latent construct (note that, in Figure 2b, TFA has a positive effect on the utilities of residing in suburban and rural areas, implying a negative effect on the probability of residing in an urban area, and, in Figure 2a, TFA has a positive effect on auto ownership propensity). If this TFA construct is ignored (as in the independent model), the result is a transfer of the negative covariance due to the TFA construct to a much higher negative (and biased) effect of urban

dwelling on auto ownership count. This is what we observe in Table 4 for the $\breve{\gamma}_{18}$ coefficient, where the independent model estimate is much more negative than the true value and the joint model estimate. Thus, accounting for endogeneity effects is not simply of academic interest, but can have substantial real implications for variable effects and subsequent policy analysis.

In addition to an APB comparison between the joint model and the independent model, we also compare the performance of the two models using the ADCLRT test. The ADCLRT statistic for the test between the two models has an approximate chi-squared distribution with 21 degrees of freedom. The corresponding table value for the chi-squared distribution is 41.401 at the 0.5% level of significance. In this paper, we identify the number of times (corresponding to the 200 data sets) that the ADCLRT value rejects the independent model in favor of the joint model. The result, presented toward the bottom of Table 4 clearly indicates that the joint model rejects the independent model in all the 200 data sets, further reinforcing the need to consider the GHDM model.

4.7. Procedure for Treatment Effects Based on Residential Choice

The estimation results from the simulation experiment may be used to examine the differences between the GHDM and independent models as they relate to the implied effects of one outcome variable on another. To demonstrate the potential problems of ignoring latent variables, we examine the impact of residential location choice on auto ownership (other outcome effects may also be computed, but, because this is only a simulation effort, we focus on one effect to demonstrate the potential biases accruing from ignoring jointness). This is helpful to obtain insights regarding whether, and how much, an independent model can bias the influence of an urban-like high density design on travel-related behaviors. An important approach to do so is the Average Treatment Effect (ATE) (see Heckman and Vytlacil, 2000 and Heckman *et al*., 2001).

In the context of motorized vehicle ownership, the ATE measure provides the expected difference in motorized vehicle ownership for a random individual if s/he were located in a specific density configuration i as opposed to another density configuration $i' \neq i$. The measure is estimated as follows:

$$\hat{\text{ATE}}_{ii'} = \frac{1}{Q}\sum_{q=1}^{Q}\left(\sum_{j_1=0}^{\infty} k_{q1}\cdot\left[P(k_{q1}\mid a_{qi}=1) - P(k_{q1}\mid a_{qi'}=1)\right]\right)$$

where a_{qi} is the dummy variable for the density category i for the individual q, and k_{q1} is an index for auto ownership k_{q1} ($k_{q1} = 0,1,2,....\infty$) (the subscript '1', consistent with the notation used earlier, indicates that auto ownership is the first count variable in the model system). Although the summation in the equation above extends until infinity, we consider counts only up to k_{q1} = 10. This should not affect the computations because the probabilities associated with higher motorized vehicle ownership levels are very close to zero.

The analyst can compute the ATE measures for all the pairwise combinations of residential density category relocations. Here, we focus on the case when an individual in a rural location is transplanted to an urban location. The standard error of the ATE measure is obtained using bootstraps from the sampling distributions of the estimated parameters. The GHDM model estimates an ATE of -0.194 (standard error of 0.038), which implies that a random household that is shifted from a rural location to an urban location will, on average, reduce its motorized vehicle ownership level by 0.194 vehicles. The corresponding independent model estimate is much higher with an ATE of -0.345 (standard error of 0.020), which indicates a much higher reduction in auto ownership because of a household move from a rural area to an urban area. This overestimation in the independent model is because of the explanation provided in Section 4.6.1.

5. CONCLUSIONS

This paper proposes a new model formulation, the generalized heterogeneous data model (GHDM), to jointly model data containing mixed types of dependent variables, including multiple continuous variables, multiple ordinal variables, multiple count variables, and multiple nominal variables. Within this integrated model system, the covariance relationships among high-dimensional heterogeneous outcomes are explained by a much smaller number of latent continuous factors. The paper proposes and develops a comprehensive blueprint for estimating the GHDM model using Bhat's maximum approximate composite marginal likelihood (MACML) approach. With this approach, the dimensionality of integration in the function that needs to be maximized to obtain a consistent estimator (under standard regularity conditions) is independent of the number of latent factors and easily accommodates general covariance structures for the structural equation and for the utilities of the discrete alternatives for each

nominal outcome. Further, the use of the analytic approximation in the MACML approach to evaluate the multivariate cumulative normal distribution (MVNCD) function in the CML function simplifies the estimation procedure even further, so that the proposed MACML procedure requires the maximization of a function that has no more than bivariate normal cumulative distribution functions to be evaluated.

A simulation experiment within the virtual context of the integrated modeling of residential location choice and travel behavior is undertaken to evaluate the ability of the MACML approach to recover parameters in the GHDM from finite samples. The simulation results show that the MACML estimation approach does reasonably well in recovering the parameters, regardless of the sample size (N=1000, 2000, and 3000) used in estimation. The MACML estimator exhibits good empirical efficiency since the asymptotic standard errors (ASEs) (and the finite sample standard errors, or FSSEs) are only a small proportion of the true values, and the ASEs (derived based on the inverse of the Godambe information matrix) perform well in estimating the FSSEs. Further, it is remarkable that the approximation error due to the use of only a single permutation for approximating the MVNCD function is extremely small. However, the results also indicate that it is relatively more difficult to both accurately and precisely recover the effects of exogenous variables on the latent variables (in the structural equation system) as well as the effects of the latent variables on the outcomes (in the measurement equation system), relative to effects of exogenous variables on the outcomes in the measurement equation system and the inter-relationships between the endogenous variables. The suggestion is the exercise of caution when GHDM models with latent variables are being estimated with few observations. Our results suggest that there may be a need for 3000 observations or so for good accuracy and precision in the estimated coefficients when there are more than 2-3 psychological constructs used.

The simulation experiment also examines the implications of ignoring the presence of latent variables, so that the unobserved covariance among the multidimensional outcomes are not considered. In the virtual integrated land use-transportation modeling context used in the simulation, this is equivalent to ignoring all potential self-selection effects, which then should corrupt the endogenous variable effects, and lead to inaccurate and inefficient estimation of other parameters as well. The results indeed reveal a substantial degradation of parameter recovery across the board if the latent constructs are ignored away, and especially those associated with

the endogenous variable effects (see Figure 2c). In addition, land use effects (residential built environment in the current paper) on travel choices can be substantially biased if the multi-dimensional bundled nature of residential and travel-related choices is not considered, which can lead to potentially inappropriate policy decisions regarding infrastructure investment. Overall, the simulation design and results do emphasize the fact that integrated land use-transportation (LU-T) modeling is not simply of academic interest, but can have substantial real implications for variable effects and subsequent policy analysis. The GHDM model proposed and used in the current paper can serve as a valuable tool for such integrated LU-T modeling efforts. More generally, the GHDM model should be widely applicable in numerous empirical contexts due to its ability to accommodate data with mixed types of dependent variables, including multiple ordinal variables, multiple continuous variables, multiple count variables, and multiple nominal variables.

ACKNOWLEDGEMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. The author would also like to acknowledge support from a Humboldt Research Award from the Alexander von Humboldt Foundation, Germany. Finally, the author is grateful to Lisa Macias for her help in formatting this document, and to Subodh Dubey and Xuemei Fu for help with the simulation runs.

REFERENCES

Abrams, K.R., Jones, D.R., Jones, D.R., Sheldon, T.A., and Song, F. (2000). *Methods for meta-analysis in medical research*. J. Wiley.

Aditjandra, P. T., Cao, X. J., and Mulley, C. (2012). Understanding neighbourhood design impact on travel behaviour: An application of structural equations model to a British metropolitan data. *Transportation Research Part A: Policy and Practice,* 46(1), 22-32.

Bartholomew, K.J., Ntoumanis, N., Ryan, R.M., Bosch, J.A., and Thøgersen-Ntoumani, C. (2011). Self-determination theory and diminished functioning the role of interpersonal control and psychological need thwarting. *Personality and Social Psychology Bulletin*, 37(11), 1459-1473.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., . . . Bunch, D.S. (2002). Hybrid choice models: Progress and challenges. *Marketing Letters*, 13(3), 163-175.

Bhat, C.R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B*, 45(7), 923-939.

Bhat, C.R. (2014). The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. *Foundations and Trends in Econometrics*, 7(1), 1-117.

Bhat, C.R., and Dubey, S.K. (2014). A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B*, 67, 68-85.

Bhat, C.R., and Guo, J.Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B*, 41(5), 506-526.

Bhat, C.R., and Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B*, 45(7), 940-953.

Bhat, C.R., Astroza, S., Sidharthan, R., Jobair Bin Alam, M., and Khushefati, W.H. (2014a). A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B*, 68, 31-51.

Bhat, C.R., Paleti, R., and Singh, P. (2014b). A spatial multivariate count model for firm location decisions. *Journal of Regional Science*, 54(3), 462-502.

Bhat, C.R., Paleti, R., Pendyala, R.M., Lorenzini, K., and Konduri, K.C. (2013). Accommodating immigration status and self-selection effects in a joint model of household auto ownership and residential location choice. *Transportation Research Record: Journal of the Transportation Research Board,* 2382(1), 142-150.

Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A., 2005. Hybrid choice models with logit kernel: applicability to large scale models. In: Lee-Gosselin, M., Doherty, S. (eds.) *Integrated Land-Use and Transportation Models: Behavioral Foundations*, Elsevier, Oxford, 275-302.

Brownstone, D., and Golob, T.F. (2009). The impact of residential density on vehicle usage and energy consumption. *Journal of Urban Economics,* 65(1), 91-98.

Cao, X., and Fan, Y. (2012). Exploring the influences of density on travel behavior using propensity score matching. *Environment and Planning-Part B,* 39(3), 459.

Castro, M., Paleti, R., and Bhat, C.R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B*, 46(1), 253-272.

Castro, M., Eluru, N., Bhat, C.R., and Pendyala, R.M. (2011). Joint model of participation in nonwork activities and time-of-day choice set formation for workers. *Transportation Research Record*, 2254, 140-150.

Champion, A.G. (2001). A Changing demographic regime and evolving poly centric urban regions: Consequences for the size, composition and distribution of city populations. *Urban Studies,* 38(4), 657-677.

Clark, W. A., Huang, Y., and Withers, S. (2003). Does commuting distance matter?: Commuting tolerance and residential change. *Regional Science and Urban Economics,* 33(2), 199-221.

Day, L.L. (2000). Choosing a house: the relationship between dwelling type, perception of privacy and residential satisfaction. *Journal of Planning Education and Research,* 19(3), 265-275.

Daziano, R.A., and Bolduc, D. (2013). Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica A: Transport Science*, 9(1), 74-106.

De Leon, A.R., and Carrière, K. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4), 533-548.

De Leon, A.R., and Chough, K.C. (2013). *Analysis of Mixed Data: Methods & Applications*, CRC Press.

De Leon, A.R., and Zhu, Y. (2008). ANOVA extensions for mixed discrete and continuous data. *Computational Statistics & Data Analysis*, 52(4), 2218-2227.

De Leon, A., Soo, A., and Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistic*s, 38(5), 1021-1032.

Faes, C., Geys, H., and Catalano, P. (2009). Joint models for continuous and discrete longitudinal data. *Longitudinal Data Analysis*, 327-348.

Feddag, M.-L. (2013). Composite likelihood estimation for multivariate probit latent traits models. *Communications in Statistics-Theory and Methods*, 42(14), 2551-2566.

Gates, K.M., Molenaar, P., Hillary, F.G., and Slobounov, S. (2011). Extended unified SEM approach for modeling event-related fMRI data. *NeuroImage*, 54(2), 1151-1158.

Gueorguieva, R., and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25(8), 1307-1322.

Heckman, J., Tobias, J.L., and Vytlacil, E. (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 211-223.

Heckman, J.J., and Vytlacil, E.J. (2000). The relationship between treatment parameters within a latent variable framework. *Economics Letters*, 66(1), 33-39.

Hoshino, T., and Bentler, P.M. (2011). Bias in factor score regression and a simple solution. In: De Leon, A.R., and Chough, K.C. (eds.) *Analysis of Mixed Data: Methods & Applications*, CRC Press, 43-61.

Jansen, S.J. (2012). What is the worth of values in guiding residential preferences and choices? *Journal of Housing and the built Environment*, 27(3), 273-300.

Jöreskog, K.G. (1977). Factor analysis by least squares and maximum likelihood methods. In: Enslein, K., Ralston, A., and Wilf, H.S. (eds), *Statistical Methods for Digital Computers*, John Wiley & Sons, New York.

Keane, M.P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2), 193-200.

Kim, J., and Brownstone, D. (2013). The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples. *Energy Economics*, 40, 196-206.

Kim, S. (2011). Intra-regional residential movement of the elderly: testing a suburban-to-urban migration hypothesis. *The Annals of Regional Science*, 46(1), 1-17.

Maddala, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, UK.

Mokhtarian, P.L., and Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B*, 42(3), 204-228.

Munkin, M.K., and Trivedi, P.K. (2008). Bayesian analysis of the ordered probit model with endogenous selection. *Journal of Econometrics*, 143(2), 334-348.

Pace, L., Salvan, A., and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21(1), 129.

Paleti, R., Bhat, C.R., and Pendyala, R.M. (2013). Integrated Model of Residential Location, Work Location, Vehicle Ownership, and Commute Tour Characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, 2382, 162-172.

Pinjari, A. R., Eluru, N., Bhat, C.R., Pendyala, R.M., and Spissu, E. (2008). Joint model of choice of residential neighborhood and bicycle ownership: accounting for self-selection and unobserved heterogeneity. *Transportation Research Record: Journal of the Transportation Research Board*, 2082(1), 17-26.

Pinjari, A.R., Pendyala, R.M., Bhat, C.R., and Waddell, P.A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), 933-958.

Rashidi, T.H., Auld, J., and Mohammadian, A.K. (2012). A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection. *Transportation Research Part A*, 46(7), 1097-1107.

Reilly, T., and O'Brien, R.M. (1996). Identification of confirmatory factor analysis models of arbitrary complexity the side-by-side rule. *Sociological Methods & Research*, 24(4), 473-491.

Schwanen, T., and Mokhtarian, P.L. (2007). Attitudes toward travel and land use and choice of residential neighborhood type: Evidence from the San Francisco bay area. *Housing Policy Debate*, 18(1), 171-207.

Sener, I.N., Eluru, N., and Bhat, C.R. (2009). An analysis of bicycle route choice preferences in Texas, US. *Transportation*, 36(5), 511-539.

Shiftan, Y., Outwater, M. L., and Zhou, Y. (2008). Transit market research using structural equation modeling and attitudinal market segmentation. *Transport Policy*, 15(3), 186-195.

Stapleton, D.C. (1978). Analyzing political participation data with a MIMIC Model. *Sociological Methodology*, 52-74.

Teixeira-Pinto, A., and Harezlak, J. (2013). Factorization and latent variable models for joint analysis of binary and continuous outcomes. *Analysis of Mixed Data*, pp. 81-91, Chapman and Hall/CRC.

Temme, D., Paulssen, M., and Dannewald, T. (2008). Incorporating latent variables into discrete choice models-A simultaneous estimation approach using SEM software. *Business Research*, 1(2).

Wu, B., de Leon, A., and Withanage, N. (2013). Joint analysis of mixed discrete and continuous outcomes via copulas. In: De Leon, A.R., and Chough, K.C. (eds.) *Analysis of Mixed Data: Methods & Applications*, CRC Press, 139-156.

Zhao, Y., and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3), 335-356.

LIST OF FIGURES

Figure 1: Diagrammatic representation of the structural equation

Figure 2a: Diagrammatic representation of the measurement equation for the non-nominal variables

Figure 2b: Diagrammatic representation of the measurement equation for the nominal variables

Figure 2c: Endogenous effects

LIST OF TABLES

Table 1. Simulation Results for the 1000-Observations Case with 200 Datasets

Table 2. Simulation Results for the 2000-Observations Case with 200 Datasets

Table 3. Simulation Results for the 3000-Observations Case with 200 Datasets

Table 4. Effect of Ignoring Self-Selection Effects (for the 3000-Observation case)

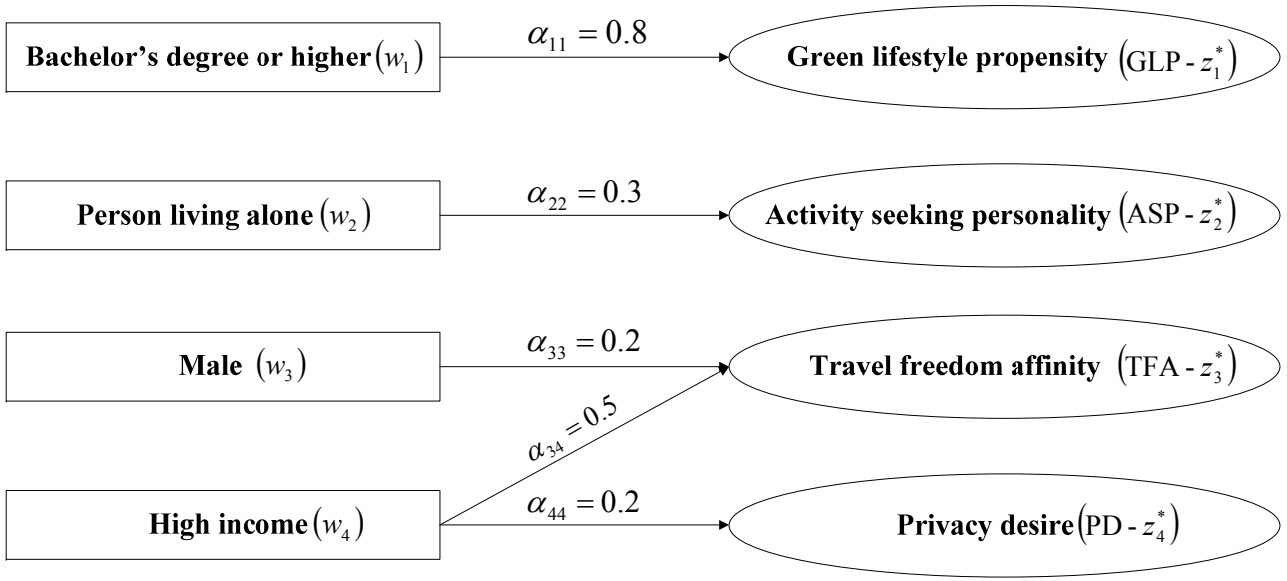

Figure 1: Digramatic representation of the structural equation

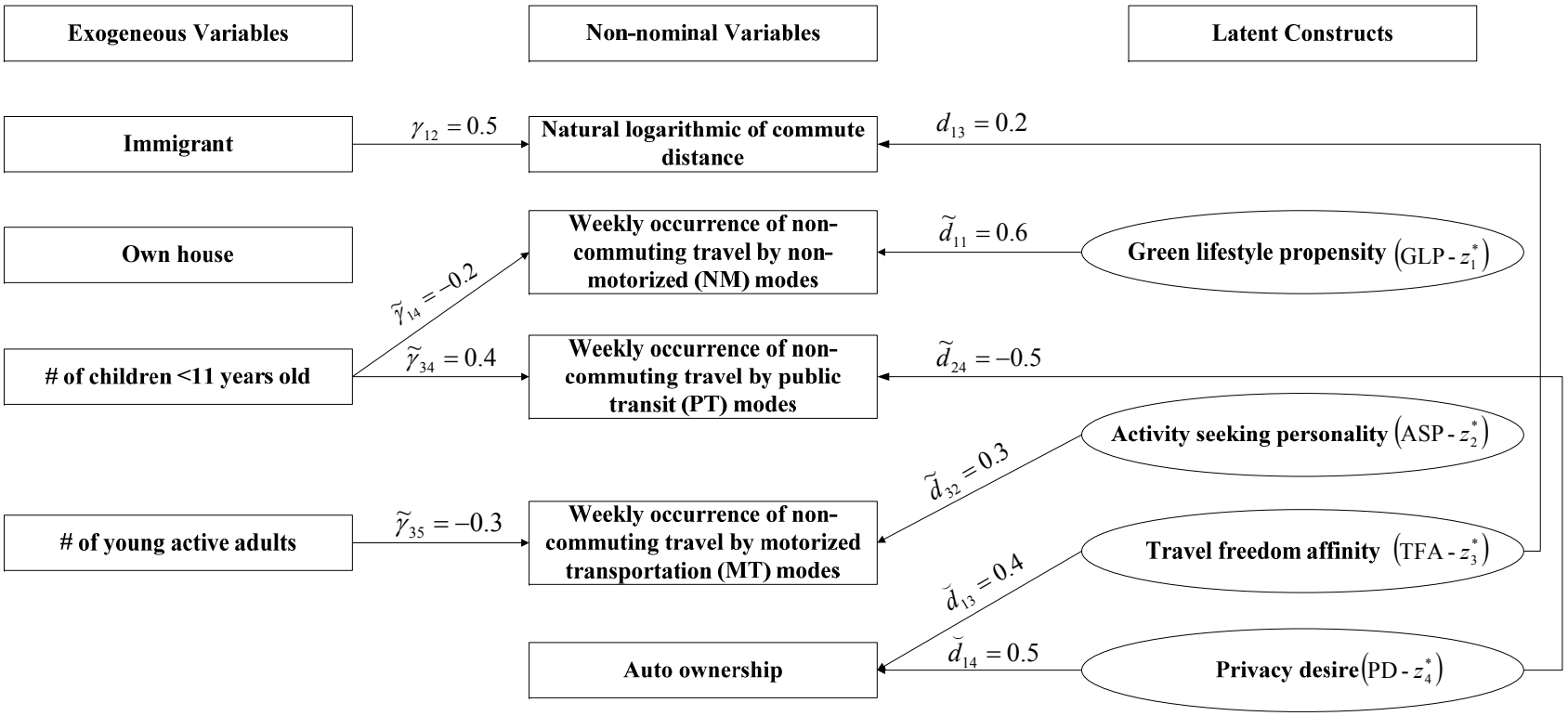

Figure 2a: Digramatic representation of the measurement equation for the non-nominal variables

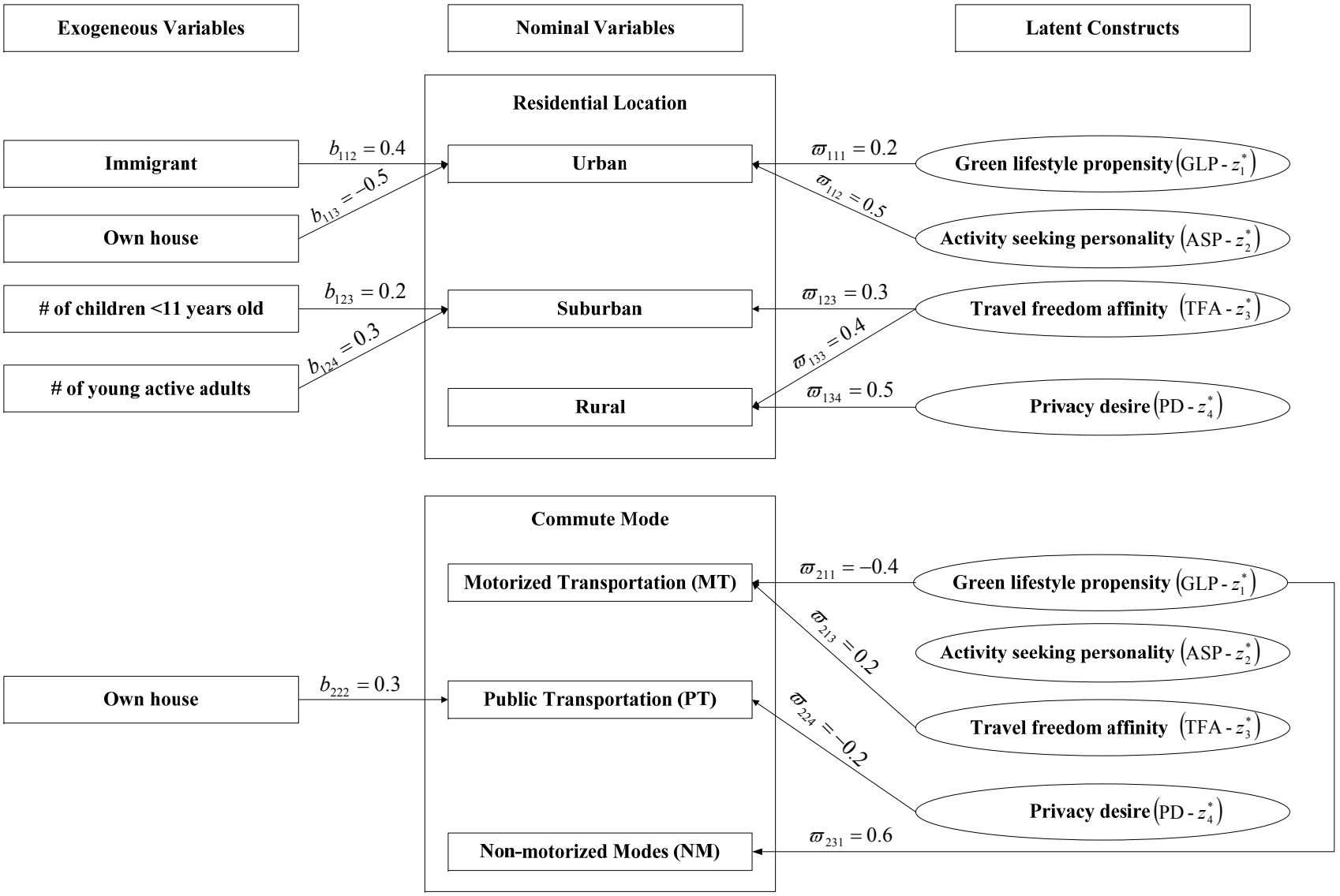

Figure 2b: Digramatic representation of the measurement equation for the nominal variables

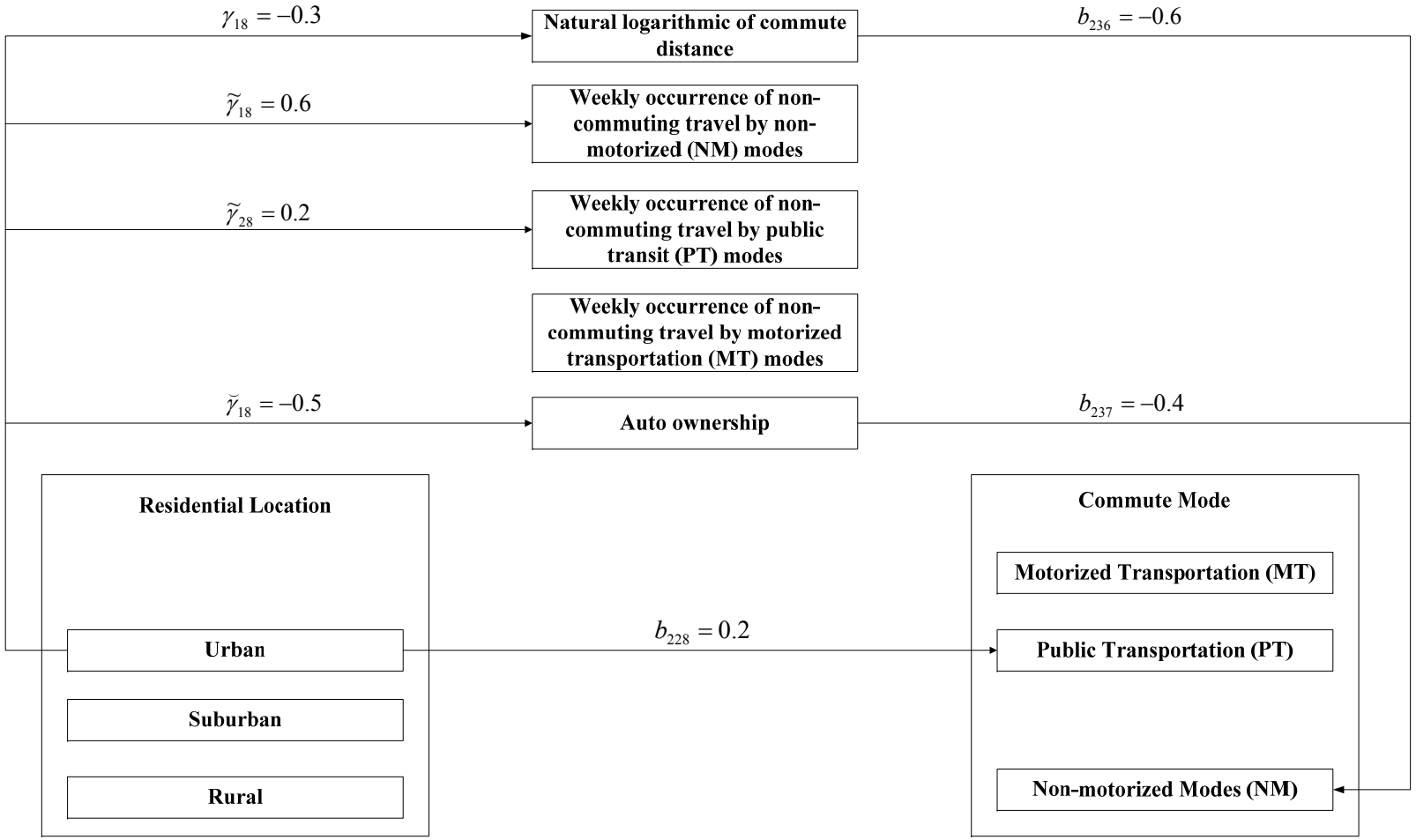

Figure 2c: Endogeneous effects

Table 1. Simulation Results for the 1000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
α_{11}	0.80	0.678	0.122	15.306	0.184	23.000	0.197	24.625	1.072	0.0046
α_{22}	0.30	0.355	0.055	18.421	0.265	88.333	0.356	118.667	1.345	0.0058
α_{33}	0.20	0.226	0.026	12.775	0.324	162.000	0.197	98.500	0.607	0.0055
α_{34}	0.50	0.557	0.057	11.301	0.396	79.200	0.239	47.800	0.602	0.0114
α_{44}	0.20	0.238	0.038	19.010	0.255	127.500	0.195	97.500	0.767	0.0055
$l_{\Gamma 33}$	0.60	0.575	0.025	4.226	0.287	47.833	0.253	42.167	0.884	0.0042
γ_{11}	1.00	0.913	0.087	8.671	0.252	25.200	0.196	19.600	0.776	0.0014
γ_{12}	0.50	0.464	0.036	7.109	0.137	27.400	0.084	16.800	0.612	0.0004
γ_{18}	-0.30	-0.267	0.033	11.085	0.120	40.000	0.072	24.000	0.603	0.0005
$\widetilde{\gamma}_{11}$	-1.00	-1.172	0.172	17.200	0.116	11.600	0.097	9.700	0.836	0.0030
$\widetilde{\gamma}_{14}$	-0.20	-0.179	0.021	10.526	0.096	48.000	0.112	56.000	1.163	0.0005
$\widetilde{\gamma}_{18}$	0.60	0.571	0.029	4.895	0.146	24.333	0.118	19.667	0.808	0.0014
$\widetilde{\gamma}_{21}$	-1.00	-1.121	0.121	12.083	0.376	37.600	0.252	25.200	0.670	0.0007
$\widetilde{\gamma}_{28}$	0.20	0.180	0.020	9.947	0.084	42.000	0.078	39.000	0.929	0.0002
$\widetilde{\gamma}_{31}$	-1.00	-1.188	0.188	18.848	0.118	11.800	0.162	16.200	1.372	0.0030
$\widetilde{\gamma}_{34}$	0.40	0.401	0.001	0.220	0.075	18.750	0.090	22.500	1.204	0.0009
$\widetilde{\gamma}_{35}$	-0.30	-0.255	0.045	15.045	0.166	55.333	0.230	76.667	1.389	0.0007
$\breve{\gamma}_{11}$	1.00	0.857	0.143	14.300	0.184	18.400	0.163	16.300	0.884	0.0019
$\breve{\gamma}_{18}$	-0.50	-0.492	0.008	1.614	0.178	35.600	0.143	28.600	0.800	0.0006
b_{111}	0.20	0.162	0.038	19.122	0.318	159.000	0.300	150.000	0.943	0.0049
b_{112}	0.40	0.388	0.012	2.981	0.183	45.750	0.114	28.500	0.625	0.0027
b_{113}	-0.50	-0.479	0.021	4.167	0.263	52.600	0.194	38.800	0.738	0.0042
b_{121}	0.50	0.478	0.022	4.364	0.273	54.600	0.206	41.200	0.754	0.0083
b_{123}	0.20	0.194	0.006	2.937	0.077	38.500	0.066	33.000	0.858	0.0013
b_{124}	0.30	0.290	0.010	3.172	0.100	33.333	0.087	29.000	0.876	0.0020
b_{221}	-0.50	-0.471	0.029	5.827	0.209	41.800	0.152	30.400	0.725	0.0039
b_{222}	0.30	0.287	0.013	4.496	0.119	39.667	0.080	26.667	0.676	0.0011
b_{228}	0.20	0.186	0.014	6.817	0.094	47.000	0.059	29.500	0.628	0.0010
b_{231}	-0.20	-0.211	0.011	5.409	0.012	6.000	0.008	4.000	0.649	0.0030

Table 1 (Cont.). Simulation Results for the 1000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
b_{236}	-0.60	-0.588	0.012	1.987	0.226	37.667	0.137	22.833	0.604	0.0032
b_{237}	-0.40	-0.390	0.010	2.376	0.290	72.500	0.202	50.500	0.698	0.0026
d_{13}	0.20	0.234	0.034	17.000	0.147	73.500	0.116	58.000	0.790	0.0016
$\widetilde{d}_{11}$	0.60	0.674	0.074	12.393	0.208	34.667	0.278	46.333	1.334	0.0053
$\widetilde{d}_{24}$	-0.50	-0.393	0.107	21.400	0.164	32.800	0.173	34.600	1.051	0.0013
$\widetilde{d}_{32}$	0.30	0.374	0.074	24.667	0.217	72.333	0.255	85.000	1.177	0.0073
$\breve{d}_{13}$	0.40	0.304	0.096	24.000	0.280	70.000	0.190	47.500	0.677	0.0054
$\breve{d}_{14}$	0.50	0.586	0.086	17.200	0.377	75.400	0.291	58.200	0.772	0.0057
ϖ_{111}	0.20	0.213	0.013	6.392	0.146	73.000	0.152	76.000	1.040	0.0096
ϖ_{112}	0.50	0.617	0.117	23.402	0.273	54.600	0.351	70.200	1.287	0.0111
ϖ_{123}	0.30	0.272	0.028	9.179	0.217	72.333	0.139	46.333	0.641	0.0043
ϖ_{133}	0.40	0.421	0.021	5.250	0.106	26.500	0.068	17.000	0.644	0.0067
ϖ_{134}	0.50	0.552	0.052	10.473	0.150	30.00	0.107	21.400	0.713	0.0185
ϖ_{211}	-0.40	-0.416	0.016	3.960	0.053	13.250	0.050	12.500	0.941	0.0041
ϖ_{213}	0.20	0.145	0.055	27.308	0.183	91.500	0.136	68.000	0.747	0.0048
ϖ_{224}	-0.20	-0.231	0.031	15.297	0.203	101.500	0.131	65.500	0.644	0.0113
ϖ_{231}	0.60	0.739	0.139	23.113	0.555	92.500	0.482	80.333	0.869	0.0092
$l_{\ddot{\Sigma}11}$	1.25	1.116	0.134	10.759	0.335	26.800	0.260	20.800	0.777	0.0005
ψ_{12}	1.50	1.497	0.003	0.168	0.083	5.533	0.094	6.267	1.127	0.0038
ψ_{22}	1.50	1.339	0.161	10.733	0.282	18.800	0.319	21.267	1.129	0.0004
ψ_{32}	1.50	1.409	0.091	6.048	0.347	23.133	0.374	24.933	1.078	0.0035
φ_1	0.75	0.702	0.048	6.446	0.216	28.800	0.197	26.267	0.912	0.0028
θ	2.00	1.784	0.216	10.800	0.340	17.000	0.219	10.950	0.644	0.0122
$l_{\Lambda 32}$	0.70	0.765	0.065	9.286	0.248	35.429	0.155	22.143	0.625	0.0198
$l_{\Lambda 33}$	1.49	1.610	0.120	8.054	0.225	15.101	0.135	9.060	0.600	0.0460
$l_{\Lambda 65}$	0.60	0.622	0.022	3.616	0.010	1.667	0.008	1.333	0.817	0.0132
$l_{\Lambda 66}$	1.36	1.461	0.101	7.426	0.160	11.765	0.137	10.074	0.858	0.0284
Overall mean value across parameters			0.059	10.55	0.204	47.32	0.172	39.71	0.857	0.0058

Table 2. Simulation Results for the 2000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
α_{11}	0.80	0.697	0.103	12.930	0.096	12.000	0.128	16.000	1.334	0.0028
α_{22}	0.30	0.271	0.029	9.692	0.154	51.333	0.180	60.000	1.173	0.0047
α_{33}	0.20	0.220	0.02	10.000	0.193	96.500	0.120	60.000	0.621	0.0115
α_{34}	0.50	0.569	0.069	13.830	0.234	46.800	0.141	28.200	0.601	0.0080
α_{44}	0.20	0.221	0.021	10.456	0.154	77.000	0.118	59.000	0.766	0.0065
$l_{\Gamma 33}$	0.60	0.581	0.019	3.228	0.237	39.500	0.207	34.500	0.871	0.0030
γ_{11}	1.00	0.977	0.023	2.348	0.054	5.400	0.051	5.100	0.945	0.0016
γ_{12}	0.50	0.500	0.000	0.006	0.055	11.000	0.049	9.800	0.897	0.0011
γ_{18}	-0.30	-0.296	0.004	1.460	0.049	16.333	0.049	16.333	1.005	0.0006
$\widetilde{\gamma}_{11}$	-1.00	-1.091	0.091	9.100	0.140	14.000	0.146	14.600	1.044	0.0014
$\widetilde{\gamma}_{14}$	-0.20	-0.198	0.002	0.972	0.031	15.500	0.031	15.500	0.998	0.0002
$\widetilde{\gamma}_{18}$	0.60	0.586	0.014	2.352	0.085	14.167	0.077	12.833	0.905	0.0006
$\widetilde{\gamma}_{21}$	-1.00	-1.129	0.129	12.900	0.055	5.500	0.051	5.100	0.927	0.0005
$\widetilde{\gamma}_{28}$	0.20	0.175	0.025	12.338	0.061	30.500	0.055	27.500	0.894	0.0002
$\widetilde{\gamma}_{31}$	-1.00	-1.154	0.154	15.400	0.113	11.300	0.115	11.500	1.013	0.0075
$\widetilde{\gamma}_{34}$	0.40	0.389	0.011	2.763	0.037	9.250	0.051	12.750	1.390	0.0017
$\widetilde{\gamma}_{35}$	-0.30	-0.290	0.010	3.370	0.033	11.000	0.034	11.333	1.050	0.0006
$\breve{\gamma}_{11}$	1.00	0.895	0.105	10.500	0.108	10.800	0.072	7.200	0.663	0.0027
$\breve{\gamma}_{18}$	-0.50	-0.533	0.033	6.681	0.063	12.600	0.039	7.800	0.616	0.0009
b_{111}	0.20	0.226	0.026	13.208	0.207	103.500	0.132	66.000	0.636	0.0073
b_{112}	0.40	0.395	0.005	1.327	0.106	26.500	0.072	18.000	0.680	0.0022
b_{113}	-0.50	-0.502	0.002	0.434	0.142	28.400	0.136	27.200	0.955	0.0024
b_{121}	0.50	0.520	0.020	3.991	0.211	42.200	0.131	26.200	0.619	0.0086
b_{123}	0.20	0.199	0.001	0.402	0.038	19.000	0.023	11.500	0.597	0.0009
b_{124}	0.30	0.303	0.003	1.058	0.046	15.333	0.031	10.333	0.682	0.0015
b_{221}	-0.50	-0.522	0.022	4.391	0.106	21.200	0.087	17.400	0.820	0.0040
b_{222}	0.30	0.314	0.014	4.733	0.066	22.000	0.053	17.667	0.794	0.0019
b_{228}	0.20	0.199	0.001	0.739	0.065	32.500	0.059	29.500	0.904	0.0013
b_{231}	-0.20	-0.182	0.018	9.000	0.206	103.000	0.153	76.500	0.744	0.0035

Table 2 (Cont.). Simulation Results for the 2000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
b_{236}	-0.60	-0.614	0.014	2.271	0.216	36.000	0.158	26.333	0.733	0.0026
b_{237}	-0.40	-0.411	0.011	2.671	0.140	35.000	0.107	26.750	0.762	0.0020
d_{13}	0.20	0.226	0.026	13.023	0.052	26.000	0.040	20.000	0.785	0.0029
$\widetilde{d}_{11}$	0.60	0.682	0.082	13.703	0.149	24.833	0.177	29.500	1.186	0.0024
$\widetilde{d}_{24}$	-0.50	-0.413	0.087	17.400	0.093	18.600	0.089	17.800	0.956	0.0033
$\widetilde{d}_{32}$	0.30	0.343	0.043	14.333	0.177	59.000	0.247	82.333	1.392	0.0093
$\breve{d}_{13}$	0.40	0.312	0.088	21.896	0.174	43.500	0.144	36.000	0.828	0.0173
$\breve{d}_{14}$	0.50	0.568	0.068	13.608	0.316	63.200	0.210	42.000	0.665	0.0076
ϖ_{111}	0.20	0.211	0.011	5.648	0.099	49.500	0.105	52.500	1.059	0.0057
ϖ_{112}	0.50	0.630	0.130	25.904	0.229	45.800	0.187	37.400	0.814	0.0050
ϖ_{123}	0.30	0.274	0.026	8.534	0.128	42.667	0.078	26.000	0.609	0.0043
ϖ_{133}	0.40	0.413	0.013	3.258	0.317	79.250	0.243	60.750	0.767	0.0057
ϖ_{134}	0.50	0.490	0.010	1.942	0.313	62.600	0.225	45.000	0.718	0.0150
ϖ_{211}	-0.40	-0.468	0.068	16.979	0.128	32.000	0.117	29.250	0.913	0.0045
ϖ_{213}	0.20	0.168	0.032	16.137	0.136	68.000	0.111	55.500	0.819	0.0073
ϖ_{224}	-0.20	-0.236	0.036	18.000	0.151	75.500	0.102	51.000	0.678	0.0138
ϖ_{231}	0.60	0.690	0.090	14.988	0.278	46.333	0.220	36.667	0.793	0.0048
$l_{\bar{z}11}$	1.25	1.238	0.012	0.923	0.049	3.920	0.043	3.440	0.876	0.0012
ψ_{12}	1.50	1.462	0.038	2.547	0.115	7.667	0.131	8.733	1.132	0.0048
ψ_{22}	1.50	1.319	0.181	12.048	0.083	5.533	0.082	5.467	0.993	0.0016
ψ_{32}	1.50	1.462	0.038	2.524	0.122	8.133	0.137	9.133	1.126	0.0014
φ_1	0.75	0.719	0.031	4.080	0.124	16.533	0.092	12.267	0.742	0.0030
θ	2.00	1.853	0.147	7.350	0.191	9.550	0.192	9.600	1.005	0.0172
$l_{\Lambda 32}$	0.70	0.730	0.030	4.286	0.150	21.429	0.138	19.714	0.920	0.0132
$l_{\Lambda 33}$	1.49	1.613	0.123	8.237	0.192	12.886	0.223	14.966	1.164	0.0248
$l_{\Lambda 65}$	0.60	0.558	0.042	7.000	0.114	19.000	0.131	21.833	1.152	0.0093
$l_{\Lambda 66}$	1.36	1.229	0.131	9.634	0.122	8.971	0.140	10.294	1.148	0.0103
Overall mean value across parameters			0.046	8.01	0.134	32.60	0.115	26.885	0.891	0.0052

Table 3. Simulation Results for the 3000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
α_{11}	0.80	0.784	0.016	2.005	0.068	8.500	0.086	10.750	1.261	0.0014
α_{22}	0.30	0.303	0.003	1.088	0.086	28.667	0.098	32.667	1.139	0.0038
α_{33}	0.20	0.194	0.006	2.977	0.125	62.500	0.101	50.500	0.806	0.0014
α_{34}	0.50	0.465	0.035	6.999	0.121	24.200	0.075	15.000	0.624	0.0034
α_{44}	0.20	0.205	0.005	2.506	0.079	39.500	0.064	32.000	0.800	0.0025
$l_{\Gamma 33}$	0.60	0.531	0.069	11.540	0.038	6.333	0.049	8.167	1.293	0.0009
γ_{11}	1.00	1.006	0.006	0.608	0.043	4.300	0.040	4.000	0.931	0.0005
γ_{12}	0.50	0.495	0.005	0.943	0.045	9.000	0.040	8.000	0.887	0.0002
γ_{18}	-0.30	-0.302	0.002	0.731	0.043	14.333	0.042	14.000	0.961	0.0001
$\widetilde{\gamma}_{11}$	-1.00	-1.120	0.120	12.000	0.080	8.000	0.102	10.200	1.277	0.0007
$\widetilde{\gamma}_{14}$	-0.20	-0.186	0.014	7.073	0.024	12.000	0.023	11.500	0.992	0.0001
$\widetilde{\gamma}_{18}$	0.60	0.555	0.045	7.533	0.057	9.500	0.058	9.667	1.018	0.0003
$\widetilde{\gamma}_{21}$	-1.00	-1.020	0.020	2.000	0.038	3.800	0.036	3.600	0.967	0.0002
$\widetilde{\gamma}_{28}$	0.20	0.172	0.028	13.784	0.052	26.000	0.044	22.000	0.836	0.0001
$\widetilde{\gamma}_{31}$	-1.00	-1.089	0.089	8.900	0.073	7.300	0.084	8.400	1.147	0.0015
$\widetilde{\gamma}_{34}$	0.40	0.374	0.026	6.536	0.025	6.250	0.024	6.000	0.978	0.0003
$\widetilde{\gamma}_{35}$	-0.30	-0.277	0.023	7.797	0.022	7.333	0.019	6.333	0.830	0.0002
$\breve{\gamma}_{11}$	1.00	0.963	0.037	3.700	0.071	7.100	0.053	5.300	0.752	0.0016
$\breve{\gamma}_{18}$	-0.50	-0.540	0.040	8.074	0.052	10.400	0.046	9.200	0.887	0.0003
b_{111}	0.20	0.188	0.012	5.750	0.123	61.500	0.085	42.500	0.690	0.0019
b_{112}	0.40	0.393	0.007	1.748	0.068	17.000	0.051	12.750	0.751	0.0013
b_{113}	-0.50	-0.499	0.001	0.250	0.079	15.800	0.058	11.600	0.738	0.0014
b_{121}	0.50	0.491	0.009	1.773	0.092	18.400	0.083	16.600	0.901	0.0034
b_{123}	0.20	0.199	0.001	0.727	0.024	12.000	0.018	9.000	0.735	0.0005
b_{124}	0.30	0.298	0.002	0.784	0.026	8.667	0.020	6.667	0.756	0.0008
b_{221}	-0.50	-0.507	0.007	1.380	0.074	14.800	0.053	10.600	0.714	0.0014
b_{222}	0.30	0.301	0.001	0.458	0.050	16.667	0.039	13.000	0.777	0.0007
b_{228}	0.20	0.197	0.003	1.432	0.047	23.500	0.035	17.500	0.753	0.0005
b_{231}	-0.20	-0.190	0.010	5.206	0.103	51.500	0.111	55.500	1.083	0.0021

Table 3 (Cont.). Simulation Results for the 3000-Observations Case with 200 Datasets

Para-meters	True Value	Parameters Estimates			Standard Error Estimates					
		Mean Est.	Abs. Bias	APB	FSSE		ASE		RE	APERR
					Value	% of true value	Value	% of true value		
b_{236}	-0.60	-0.606	0.006	1.072	0.048	8.000	0.054	9.000	1.107	0.0007
b_{237}	-0.40	-0.404	0.004	1.024	0.042	10.500	0.046	11.500	1.099	0.0006
d_{13}	0.20	0.206	0.006	2.783	0.038	19.000	0.034	17.000	0.894	0.0012
$\widetilde{d}_{11}$	0.60	0.574	0.026	4.416	0.069	11.500	0.080	13.333	1.172	0.0012
$\widetilde{d}_{24}$	-0.50	-0.461	0.039	7.800	0.041	8.200	0.053	10.600	1.301	0.0011
$\widetilde{d}_{32}$	0.30	0.279	0.021	6.849	0.080	26.667	0.092	30.667	1.149	0.0023
$\breve{d}_{13}$	0.40	0.343	0.057	14.275	0.111	27.750	0.109	27.250	0.984	0.0058
$\breve{d}_{14}$	0.50	0.515	0.015	3.000	0.059	11.800	0.064	12.800	1.078	0.0023
ϖ_{111}	0.20	0.194	0.006	2.756	0.056	28.000	0.065	32.500	1.170	0.0053
ϖ_{112}	0.50	0.515	0.015	3.012	0.061	12.200	0.077	15.400	1.277	0.0033
ϖ_{123}	0.30	0.290	0.010	3.481	0.087	29.000	0.066	22.000	0.756	0.0019
ϖ_{133}	0.40	0.432	0.032	8.099	0.116	29.000	0.113	28.250	0.976	0.0025
ϖ_{134}	0.50	0.468	0.032	6.419	0.085	17.000	0.096	19.200	1.131	0.0064
ϖ_{211}	-0.40	-0.412	0.012	3.056	0.086	21.500	0.101	25.250	1.177	0.0024
ϖ_{213}	0.20	0.209	0.009	4.432	0.089	44.500	0.063	31.500	0.713	0.0026
ϖ_{224}	-0.20	-0.199	0.001	0.394	0.076	38.000	0.075	37.500	0.993	0.0059
ϖ_{231}	0.60	0.621	0.021	3.573	0.092	15.333	0.101	16.833	1.098	0.0038
$l_{\bar{\Sigma}11}$	1.25	1.245	0.005	0.421	0.037	2.960	0.035	2.800	0.931	0.0005
ψ_{12}	1.50	1.403	0.097	6.451	0.062	4.133	0.072	4.800	1.163	0.0009
ψ_{22}	1.50	1.390	0.110	7.333	0.061	4.067	0.061	4.067	1.010	0.0005
ψ_{32}	1.50	1.386	0.114	7.595	0.063	4.200	0.064	4.267	1.009	0.0007
φ_1	0.75	0.787	0.037	4.896	0.057	7.600	0.060	8.000	1.060	0.0015
θ	2.00	1.917	0.083	4.150	0.110	5.500	0.121	6.050	1.100	0.0061
$l_{\Lambda 32}$	0.70	0.746	0.046	6.627	0.120	17.143	0.135	19.286	1.125	0.0052
$l_{\Lambda 33}$	1.49	1.540	0.050	3.342	0.262	17.584	0.306	20.537	1.170	0.0157
$l_{\Lambda 65}$	0.60	0.614	0.014	2.416	0.161	26.833	0.153	25.500	0.952	0.0044
$l_{\Lambda 66}$	1.36	1.368	0.008	0.609	0.236	17.353	0.401	29.485	1.699	0.0068
Overall mean value across parameters			0.027	4.40	0.076	17.86	0.077	16.93	0.992	0.0022

Table 4: Effect of Ignoring Self-Selection Effects (for the 3000-Observation case)

Para-meters	True Value	Joint			Independent		
		Mean Est.	Abs. Bias	Absolute Percentage Bias (APB)	Mean Est.	Abs. Bias	Absolute Percentage Bias (APB)
γ_{11}	1.00	1.006	0.006	0.608	1.073	0.073	7.330
γ_{12}	0.50	0.495	0.005	0.943	0.496	0.004	0.704
γ_{18}	-0.30	-0.302	0.002	0.731	-0.304	0.004	1.404
$\widetilde{\gamma}_{11}$	-1.00	-1.120	0.120	12.000	-0.941	0.059	5.925
$\widetilde{\gamma}_{14}$	-0.20	-0.186	0.014	7.073	-0.159	0.041	20.255
$\widetilde{\gamma}_{18}$	0.60	0.555	0.045	7.533	0.476	0.124	20.612
$\widetilde{\gamma}_{21}$	-1.00	-1.020	0.020	2.000	-1.218	0.218	21.780
$\widetilde{\gamma}_{28}$	0.20	0.172	0.028	13.784	0.169	0.031	15.577
$\widetilde{\gamma}_{31}$	-1.00	-1.089	0.089	8.900	-1.194	0.194	19.421
$\widetilde{\gamma}_{34}$	0.40	0.374	0.026	6.536	0.359	0.041	10.291
$\widetilde{\gamma}_{35}$	-0.30	-0.277	0.023	7.797	-0.266	0.034	11.221
$\bar{\gamma}_{11}$	1.00	0.963	0.037	3.700	0.946	0.054	5.417
$\bar{\gamma}_{18}$	-0.50	-0.540	0.040	8.074	-0.592	0.092	18.400
b_{111}	0.20	0.188	0.012	5.750	0.165	0.035	17.406
b_{112}	0.40	0.393	0.007	1.748	0.337	0.063	15.715
b_{113}	-0.50	-0.499	0.001	0.250	-0.426	0.074	14.715
b_{121}	0.50	0.491	0.009	1.773	0.333	0.167	33.307
b_{123}	0.20	0.199	0.001	0.727	0.176	0.024	12.226
b_{124}	0.30	0.298	0.002	0.784	0.263	0.037	12.407
b_{221}	-0.50	-0.507	0.007	1.380	-0.402	0.098	19.611
b_{222}	0.30	0.301	0.001	0.458	0.273	0.027	9.083
b_{228}	0.20	0.197	0.003	1.432	0.180	0.020	9.787
b_{231}	-0.20	-0.190	0.010	5.206	-0.112	0.088	44.000
b_{236}	-0.60	-0.606	0.006	1.072	-0.668	0.068	11.333
b_{237}	-0.40	-0.404	0.004	1.024	-0.479	0.079	19.750
$l_{\bar{\Sigma}11}$	1.25	1.245	0.005	0.421	1.292	0.042	3.325
ψ_{12}	1.50	1.403	0.097	6.451	1.346	0.154	10.261
ψ_{22}	1.50	1.390	0.110	7.333	1.245	0.255	16.987
ψ_{32}	1.50	1.386	0.114	7.595	1.172	0.328	21.862

Table 4 (Cont.): Effect of Ignoring Self-Selection Effects (for the 3000-Observation case)

Para-meters	True Value	Joint			Independent		
		Mean Est.	Abs. Bias	Absolute Percentage Bias (APB)	Mean Est.	Abs. Bias	Absolute Percentage Bias (APB)
φ_1	0.75	0.787	0.037	4.896	0.522	0.228	30.357
θ	2.00	1.917	0.083	4.150	0.739	1.261	63.056
$l_{\Lambda 32}$	0.70	0.746	0.046	6.627	0.780	0.080	11.500
$l_{\Lambda 33}$	1.49	1.540	0.050	3.342	1.640	0.150	10.081
$l_{\Lambda 65}$	0.60	0.614	0.014	2.416	0.771	0.171	28.497
$l_{\Lambda 66}$	1.36	1.368	0.008	0.609	2.053	0.693	50.991
Overall mean value across parameters			0.031	4.146		0.146	17.846
Mean log composite marginal likelihood at convergence			-66862.90		-67545.95		
Number of times the adjusted composite likelihood ratio test (ADCLRT) statistics favors the Joint model		All 200 times when compared with the value of $\chi^2_{21,0.005}=4140$] at any level of significance (mean ADCLRT statistics is 173.91)					