A New Flexible Generalized Heterogeneous Data Model (GHDM) with an Application to Examine the Effect of High Density Neighborhood Living on Bicycling Frequency

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA
Phone: 1-512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Aupal Mondal
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Email: aupal.mondal@utexas.edu

ABSTRACT

There is growing interest in multivariate dependent outcome models that include a mixture of different kinds of discrete and continuous variables. This may be attributed to at least two reasons. The first is the ability to generate multivariate distributions through the use of relatively flexible copula-based methods and/or effective factorization techniques for the covariance matrices. The second is the development of computationally efficient ways to estimate models based on variational methods for Bayesian inference or maximum approximate composite marginal likelihood methods for frequentist inference. However, there are two important assumptions in earlier mixed data models: (i) marginal normality of unobserved factors that generate jointness among the main outcome variables of interest, and (ii) independence between the unobserved factors and the propensity equations underlying the main outcomes of interest.

In the current paper, we simultaneously relax both these assumptions and develop a flexible Generalized Heterogeneous Data Model (GHDM) for mixed data modeling. We then propose a hybrid MSL-MACML inference approach for estimation. We demonstrate an application of our proposed model in the context of individuals' high-density residential neighborhood living choice and monthly bicycling frequency. The empirical results highlight the benefits of our proposed methodology, both from a policy standpoint as well as a predictive data fit standpoint.

Keywords: Multidimensional mixed data models; latent variables; maximum approximate composite marginal likelihood (MACML) estimation; GHDM; built environment; bicycling.

1. INTRODUCTION

The literature on multivariate dependent outcome models that include a mixture of different kinds of discrete and continuous variables has increased in recent years. This may be attributed to at least two reasons. The *first* is the ability to generate multivariate distributions through the use of relatively flexible copula-based methods (see, for example, Müller and Czado, 2018, Wei et al., 2019, Jiryaie and Khodadadi, 2019, and Kwon et al., 2022) and/or the use of effective factorization techniques for the covariance matrices that reduce the number of covariance parameters to be estimated (see, for example, Ong et al., 2018, Dias et al., 2020, Moore et al., 2020, and Kang et al., 2021). The *second* is the development of computationally efficient ways to estimate models based on variational methods for Bayesian inference (where a target distribution, which may be the posterior or an augmented posterior, is approximated with a simpler distribution; see, for example, Smith et al., 2020) or maximum approximate composite marginal likelihood methods for frequentist inference (where the likelihood function is replaced by a surrogate likelihood function, combined with efficient ways to estimate the multivariate normal cumulative distribution function; see, for example, Bhat, 2014, and Bhat, 2018). The effectiveness of such developments is evidenced in the sampling of studies identified above that focus on applications that span several fields, including transportation, clinical biology, environmental science, actuarial science, and economics, just to name a few.

The not-so-uncommon approach in many fields in the face of multivariate mixed data, until the developments just discussed, was to simply ignore the presence of any common underlying unobserved factors (attitudes, values, and lifestyle factors) of decision-makers that impact the dependent outcomes simultaneously (that is, to *a priori* assume that the covariance matrix involved in the modeling of the dependent variables are purely diagonal). But, doing so can lead to inefficient estimation of covariate effects because of throwing out valuable dependency information across the outcomes (Teixeira-Pinto and Harezlak, 2013), poor statistical power in testing and poor control of type I error rates (De Leon and Zhu, 2008), and (perhaps most importantly) inconsistent estimates of the structural effect of one endogenous variable on another (see Bhat, 2015). The last of these issues refers to the general concept of endogeneity bias as typically labeled in the econometric literature. Specifically, ignoring that an "explanatory" covariate may be endogenous to another dependent variable of interest is tantamount to a sequential decision-making process (rather than recognizing the bundled nature of the multiple outcomes), and can inflate or deflate the effect of the endogenous "explanatory" covariate on the other dependent variable of interest. The important point of note is that such mis-estimated effects are not simply an econometric (esoteric) nuisance, but can lead to mis-informed policy actions. Examples from the transportation literature in this regard include the classic residential "self-selection" effect in activity-pattern analysis (see, for example, the recent studies by Lu et al., 2018, Leung et al., 2019, and Zang et al., 2019) and traffic control placement "self-selection" effect in crash analysis (see, for example, Bhat et al., 2014 and Heydari et al., 2020). In the former case, the issue is that if households that are environmentally conscious (say an unobserved variable) choose to locate in transit and pedestrian friendly "neo-urbanist" neighborhoods that are characterized by

high land use density as well as have a strong preference for non-motorized modes of transportation, then at least some of the structural effect of living in neo-urbanist environments on the preference for non-motorized mode use may be associative (caused by environmental consciousness) rather than causal. Ignoring this possibility in a cross-sectional analysis can inflate the "true" effect of living in a neo-urbanist neighborhood on non-motorized mode use, which can, in turn, lead to misinformed land-use policies. In the latter case of crash analysis, it is possible that the location for the placement of a specific traffic control measure based on what traffic engineers identify as a high-risk crash site (say due to limited sight distance or some other topographical feature) may also be immediately identified by motorists as being relatively dangerous (that is, motorists may be more careful when they encounter a specific topographic feature, leading to low crashes at the site). Ignoring this crash risk perception of traffic engineers and motorists (that essentially creates an unobserved association such that the placement of the control measure is intrinsically correlated with a low intensity of crashes) may lead to an inflated estimate of the "true" effectiveness of the traffic control measure (see Mannering and Bhat, 2014 for a more detailed discussion of such issues in the safety literature).

In the general context of covariate endogeneity as discussed above, there has been a substantial amount of work in the econometric literature on the simultaneous modeling of multiple continuous variables. However, there has been relatively little emphasis on multiple non-continuous variables. Bhat (2015), Jiryaie et al. (2016), and Jiryaie and Khodadadi (2019) provide reviews of the many different approaches for modeling multiple and mixed data outcomes.[1] Of these, a whole class of structural equations models (SEMs), Generalized linear latent and mixed models (GLLAMM) and item response theory (IRT) models (all of which use factorization approaches to make the covariance matrix more parsimonious) have received particular attention because of their parsimonious specification. In this category of factorization-based mixed models, there are two types of latent variables. The first type is the latent continuous variable assumed to underly non-continuous observed outcomes (and tied to the observed outcomes through either a utility maximization rule for nominal outcomes or through a thresholding structure for ordinal, grouped, and count outcomes). A second type corresponds to latent constructs (factors) that develop the efficient factorization structure and the parsimonious covariance structure among the

[1] Such methods include (a) 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, but is not suitable for ordinal outcome variables and does not accommodate dependence between nominal and ordinal outcomes; see De Leon and Chough, 2013), (b) the conditional grouped continuous model (CGCM) (which is based on a "reverse-factorization" approach to employ a latent variable representation for binary/ordinal outcomes, and assumes a multivariate normal (MVN) distribution for the continuous outcomes, but cannot be directly extended to the case of nominal outcomes; see by De Leon (2005), (c) the general mixed data model (GMDM) (which extends the CGCM to include nominal varibles by using 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, which, like the GLOM, resorts to a factorization approach in which an artificial hierarchy is assumed in which the the multidimensional discrete outcomes are intermediate responses and the ordinal/continuous outcomes are the ultimate responses; see De Leon and Carriégre (2007) and Wu et al. (2013), (d) Methods that tie all the mixed outcomes 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; such methods assume an multivariate normal distribution or a Gaussian copula over the error terms of the entire set of latent and observed continuous variables characterizing the many types of outcomes, but become cumbersome when used with many nominal variables or many dependent outcomes due to an explosion in the number of covariance parameters to be estimated; see for example, Bhat et al., 2014 and Dias et al., 2020), and the method discussed in the text.

error terms of the first type of latent variables. In the rest of this paper, we will strictly use the term "latent variables" to refer to the first type discussed above and the term "latent constructs" to refer to the second factor-type variables.

The factorization-based approaches discussed above are predominantly used for cases with no nominal dependent variables, and the proposed maximum likelihood estimation approach gets unwieldy with many dependent variables (see Bhat and Guo, 2007, Pinjari et al., 2008, Feddag, 2013, Ong et al., 2018, and Smith et al., 2020). Within the transportation literature, Ben-Akiva et al., 2002 and Bolduc et al., 2005 used an SEM model structure with a nominal outcome (usually referred to as an integrated choice and latent variable or ICLV model). But the focus was on a single nominal outcome, rather than a mix of different types of dependent variables outcomes. The ICLV model uses type-I extreme value errors in the nominal outcome utility, mixed with normally distributed stochastic latent factors (Ben-Akiva et al., 2002, Bolduc et al., 2005, Vij and Walker, 2014). Bhat and Dubey (2014) proposed a different SEM-type model formulation for the ICLV model, based on a multivariate probit (MNP) kernel for the nominal outcome. They also propose a different estimation approach compared to the maximum simulated likelihood estimation (MSLE) estimation of the ICLV model. Specifically, they employ Bhat's (2011) maximum approximate composite marginal likelihood (MACML) inference approach, which leads to well-behaved surfaces for the gradient/hessian functions and shows an order-of-magnitude reduction in computational time.

Bhat (2015) further generalized Bhat and Dubey's probit kernel and MACML estimation approach by proposing the *generalized heterogeneous data model* (GHDM), which, while retaining the factorization-based latent construct approach, extends to the case of multiple nominal outcomes, multiple ordinal variables, multiple count variables, and multiple continuous variables. Bhat et al. (2016a) further extended the GHDM to include the case of dependent variables that are of the multiple-discrete-continous type. The GHDM is general enough to accommodate many other models in the literature as special cases, and has been used in many different applications, including recently by Blazanin et al., (2022), Asmussen et al. (2022), and Dannemiller et al. (2021). Straightforward variants of the model are available to accommodate longitudinal and spatial/social clustering, such as implemented in Vinayak et al. (2018) and Bhat et al. (2016b).

Despite the general features of the GHDM model, there are two important assumptions embedded in almost all implementations of the GHDM and other mixed models.

1.1. Multivariate Normality

The first assumption of almost all mixed models to date is that of marginal normality of the latent variable underlying each outcome of the multivariate mixed data, conditional on explanatory variables. However, using a normal distribuion may not be appropriate when the actual distribution is skewed (asymmetric) and/or has more tail risk (that is, the normal distribution may not assign adequate probabilities at the extreme left end or right end of the marginal data, conditional on exogenous variables). For instance, in the transportation field, the propensity to participate in recreational activities may be skewed to the right and have a fat right tail, with a small number of

individuals having a very large propensity to recreate. Similarly, in the finance/actuarial science field, the returns from portfolio investments may be skewed toward the left, with a fat left tail, implying a larger possibility of losses than that implied by the normal distribution (see, for example, Chin et al., 2016). Further, ignoring such issues for any single dependent outcome will, in general, lead to inconsistent estimates in not only the individual equation for that mixed dependent outcome, but also potentially to inconsistent estimates for all other equations as the error gets compounded and leads to an underestimation of the likelihood of extreme joint events. In the general mixed data literature, there have been a handful of studies that have considered non-normal marginal distributions through the use of copula distributions (a copula is essentially a multivariate functional form for the joint distribution of random variables derived purely from pre-specified (but arbitrary) parametric marginal distributions of each random variable; see Bhat and Eluru, 2009). To be sure, such copula models have a long history, with the classic estimation of two-equation sample selection models with a multinomial logit selection model that uses an extreme value error term in the utility function being a very specific case (see Lee, 1983, and applications of this technique in Hamed and Mannering, 1993, Bhat, 1996, and Bhat, 1998). However, the implementation of this copula approach in a broader sense to two-equation models came only a little later (see, for example, Bhat and Eluru, 2009, Spissu et al., 2009, Pinjari et al., 2009) and the implementation in a general multivariate context with many mixed outcomes has been much more recent (see, for example, Müller and Czado, 2018 and Wei et al., 2019). The problem with the implementation of this approach with a large number of dependent outcomes, though, is that the number of covariance parameters explodes quadratically with an increase in dependent variable outcomes. To resolve this issue, some recent studies have suggested a factorization approach with a copula structure imposed on the latent constructs (or factors) rather than the error terms of the latent variables underlying the dependent outcomes, thereby allowing a flexible marginal distribution for each latent construct. This reduces the number of parameters to be estimated in the model, similar to the GHDM model (but which assumes an *a priori* multivariate normal distribution for the latent constructs). Doing so allows non-normality in the overall error term of each latent variable (because the latent constructs impact the latent variables of the dependent outcomes; see Bhat et al., 2015, Jiryaie et al., 2016, Oh and Patton, 2017, and Loaiza-Maya and Smith, 2019). In this paper, we follow this approach of introducing non-normality, but different from these earlier efforts, do so through a special flexible type of copula approach for the latent constructs of the GHDM model that facilitates stability in estimation while also not profligate in parameters.

1.2. No Dependence Among Latent Constructs and Kernel Error Terms

The second assumption in the GHDM and every other mixed model we are aware of is that the latent constructs are completely independent of the error terms in the latent variables underlying the dependent outcomes (in econometric parlance, all of the factor-based models *a priori* assume that the factors are independent of the dependent outcomes). However, this assumption may be violated in a number of application contexts, including the case where the factors (latent

constructs) represent attitudes and the dependent outcomes being studied represent behavioral outcomes. In such a case, it is typical to assume that attitudes influence manifested behavior (Wicker, 1969, Ajzen, 1991). Specifically, individual attitudes interact with (are moderated by) subjective social norms (unwritten rules or expectations, or an individual's perceptions of how others in their life would perceive specific events/actions) and perceived behavioral control (perception of an individual's confidence in the ability to pursue a specific action) to form clear behavioral intentions (or readiness to implement a certain behavior) that then gets translated into actual behavior when the next opportunity arises to realize the behavior. This so-called "theory of planned behavior" (or TPB), however, assumes a one-directional flow from attitudes to intention to behavior (Ajzen, 1991). In reality, behavior itself can lead to a change in attitudes. Thus, consider the case that, in general, environmentally-conscious households do live in high density neighborhoods and use car modes less. But some households low on the scale of environmental consciousness may find themselves (due to a variety of circumstances, say because of the desire to stay close to other family members) locating in a high density neighhborhood and using non-car modes of transportation. Over time, such households may become more aware of the benefits of the use of non-car modes and become more environmentally conscious, which implies that not only can attitudes affect behavior, but behavioral choices can also affect attitudes. Similarly, some environmentally conscious households may decide to live in a low density, car-oriented neighborhood to have more space in the home for children, which can cause cognitive dissonance and a sense of feeling like a hypocrite. According to the social psychology literature (Festinger, 1957), this unpleasant sensation may be shed quickly in one of many ways (such as seeking new, and conveniently selected, information that attempts to play down the effects of car-use on the environment or trivializing the attitude by deciding that avoiding car use is not going to change the world). While these ways of dealing with cognitive dissonance have been well documented in the social pyschology field, they also have been found in transportation-related studies such as van de Coevering et al. (2016, 2018), Wang and Lin (2019), Kroesen (2019), De Vos et al. (2018), and De Vos and Singleton (2020) using before-after studies of travel-related behavior/attitudes in the context of a residential relocation. This is another case of behavioral choices influencing attitudes. Of course, this bi-directional relationship between attitudes and behavioral choices has a temporal component to it, where attitudes affect behavior and then the behavioral actions influence attitudes over time. However, many publicly collected and available multivariate data sets are cross-sectional, and so any analysis with such cross-sectional data should consider the attitude-behavior data as a package observation. Thus, if the relationship between attitudes and residential location/car use were as discussed above, this would lead to a positive correlation (due to unobserved factors) affecting the environmental conscious (EC) construct on the one hand and high density living/non-car mode use on the other (shown by bidirectional arrows marked as (1) in Figure 1a). We characterize this dimension of correlation effects as the *EC endogeneity effect*. Of course, after controlling for the *EC endogeneity effect*, there is likely an additional stochastic EC effect on high density neighborhood (HDN) living as well as on non-motorized mode use (shown by unidirectional arrows marked as (2) in Figure 1a). These effects result in *residential*

self-selection when capturing effects of HDN living on non-motorized mode use (marked also by "+", showing a moderate intensity of self-selection). Finally, there is the *"true" causal effect* of HDN living on non-motorized mode use (shown by the unidirectional arrow marked as (3) in Figure 1a, and by "++", assuming a high intensity of this effect). Now assume that the *EC endogeneity effect* is completely ignored as in Figure 1b. This will overestimate the effect of the environmental consciousness attitude on HDN/non-motorized mode use, thus exagerrating the error correlation between high density living and non-motorized mode use (that is, will inflate the contribution of the *residential self-selection effect*; see Figure 1b, where this effect is marked by a "++"). This, in turn will underestimate the *"true" causal effect* of HDN living on non-motorized mode use (marked by '+' on the unidirectional arrow (3) in Figure 1b). Of course, how things turn out in a specific empirtical context may vary, but the example above actually reflects the empirical results obtained later in our study.

1.3. The Current Paper

In the current paper, we simultaneously relax both the assumptions discussed in the previous section. First, we use the important insight that the analyst can consider flexible non-normal parametric distributions in each mixed outcome of a multivariate model by considering marginal non-normal parametric distributions for each latent construct and then consider a copula distribution to connect these non-normal marginal distributions for the latent constructs into a single multivariate distribution. Then, since the mixed outcomes are specified to be a function of a much smaller set of the unobserved latent constructs in measurement equations, it immediately recognizes the possibility of non-normality of the latent variable underlying each mixed dependent outcome. While a tantalizingly simple concept, this is a powerful and efficient means to introduce non-normality in the latent variables of the dependent outcomes (see Bhat et al., 2015). Different from Bhat et al. (2015), in allowing for marginal non-normality in the latent constructs, we use a Yeo-Johnson (YJ) transformation (Yeo and Johnson, 2000) that converts each latent construct into a marginal univariate normal distribution, and then enjoins the resulting univariate normal marginals using an implicit Gaussian copula. This approach is referred to as a variational approximation method in Bayesian inference (see, for example, Smith et al., 2020), because it is based on approximating the (unknown) target or augmented posterior density of the parameters of a relationship, given data, by a tractable parametric family of densities). This approach accounts for fat tails if present, while also accommodating asymmetry in the marginal distributions. Unlike commonly used copula approaches that assume a specific *a priori* parametric form for each margin (each latent construct in our case), the YJ transformation for each margin allows the data to determine the distribution of each margin during the process of estimation. Additionally, stability is introduced in our approach of using a YJ transformation first and then tying the YJ-transformed marginal normal constructs using a multivariate Gaussian Copula (that is, a simple multivariate normal distribution for the transformed constructs) because the scale elements of the multivariate Gaussian copula are exactly the scales of the transformed marginal constructs. Thus, no additional constraints are needed in our approach for identification of the copula parameters. Finally, our

approach makes it computationally efficient to compute the multivariate density function of the latent constructs, relative to computing the corresponding density function with general copula methods (see Smith, et al. 2020 for a detailed discussion).

Second, and differently from all earlier mixed models based on a factorization approach that ignore dependence between the factors (latent constructs in our case) and the marginal error terms underlying the latent variables (or, for short, the kernel error terms) of the main dependent outcomes of interest, we expressly consider this dependence in our flexible GHDM model. These kernel error terms are tied to the factors using the same copula as used to tie the latent constructs (factors) themselves together, as discussed in the previous paragraph. To do so, and also for ease in estimation and empirical analysis, conditional on the factors and exogenous variables, we assume the kernel error terms to be independent and normally distributed across outcomes. This keeps the number of parameters to a limited number in the multivariate analysis, while still allowing a rich dependence structure across the outcomes in the usual spirit of a factor-based approach.[2] It is in enjoining the kernel error terms and latent constructs that the Gaussian copula comes in particularly handy, because, rather than a formal copula-based approach, one can take a direct multivariate density approach as the entire set of dependent kernel errors and transformed (flexibly non-normal) latent constructs (factors) essentially take a multivariate normal distribution.

Third, the generated multivariate normal distribution of kernel errors and transformed constructs is closed under any affine transformations and under marginalization. Additionally, conditioning on a subset of non-normal random variables (latent constructs in our specification) leads to a multivariate normal distribution for the remaining subset of normally-distributed variates (kernel error terms of the dependent outcomes in our specification). These properties enable the use of very fast analytic approximations developed by Bhat (2018) for the multivariate normal cumulative distribution (MVNCD) function evaluations involved in the composite marginal likelihood of the outcomes conditional on the flexibly non-normal latent constructs, over which the distributions of the latent constructs can be integrated out through a maximum simulated likelihood (MSL) approach. That is, the estimation is achieved using a combination of the maximum simulated likelihood (MSL) technique (to accommodate the non-normal latent constructs) and Bhat's MACML inference approach (to accommodate the kernel normal error structure; see Bhat, 2011 and Bhat, 2014). The combination harnesses the advantages of each of these approaches. The MSL approach is general and can be used to estimate models with any distribution. However, the approach can be computationally expensive to ensure good asymptotic estimator properties, and can be prohibitive and literally infeasible (in the context of the computation resources available and the time available for estimation) as the number of latent constructs increases (see Bhat, 2011). On the other hand, the MACML approach is simple, computationally efficient, and simulation-free. It easily and accurately is able to accommodate even a high number of multivariate normally distributed variates, providing both more accuracy

[2] Note, however, that the overall error for the latent variables underlying the dependent outcomes is still not normal, because of the mixing of the non-normally distributed latent constructs (factors) over the normally distributed kernel error terms. Thus, the overall error terms of the dependent outcomes capture both tail risk at the marginal level as well as extremal dependence at the sub-asymptotic level with the YJ transformation approach.

(smaller bias in parameters) and an order-of-magnitude improvement in computational efficiency relative to the MSL inference approach (see Paleti and Bhat, 2013, and Patil et al., 2017). However, it is not suitable with non-normal factors or latent constructs. The combination of the MSL and MACML is especially well suited for the case when there are relatively few factors (so that the simulation does not involve very high dimensions) and many dependent outcomes (so that the MACML computational accuracy and efficiency can be realized). In the current paper, we use the fast analytic approximations developed by Bhat (2018) for the multivariate normal cumulative distribution (MVNCD) function.

Fourth, unobserved taste variations (i.e., unobserved heterogeneity in sensitivity to response variables in each dependent outcome) can also be introduced efficiently and in a non-normal fashion through interactions of explanatory variables with the latent variables. This issue of unobserved parameter heterogeneity has rarely been discussed in multivariate mixed data modeling, but, ignoring such unobserved heterogeneity can lead to inconsistent estimation of all parameters in the multivariate system. Importantly, in our approach, the random coefficients across all dependent outcomes are based on a handful of latent constructs, substantially cutting down on the number of random parameters and also providing a more intuitive interpretation of the random parameters (as originating because of latent construct variations) rather than as arbitrary individual variable-specific random coefficients. At the same time, should there be a need to specify a naturally bounded distribution (such as a power log-normal distribution or a Rayleigh distribution for cost and time coefficients in a travel choice model) for a specific variable for one or more dependent outcomes, it can easily be imposed over the proposed model by folding the relevant random coefficient as another element of the proposed copula formulation, and using a simple mixing approach to estimation (similar to Bhat and Lavieri's (2018) approach in the case of a unidimensional nominal choice model).

To summarize, in this paper, we develop a flexible and general GHDM model for mixed data modeling and propose a hybrid MSL-MACML inference approach for estimation. To our knowledge, this is the first study to propose such a flexible methodological structure for multiple mixed outcomes modeling. The rest of this paper is structured as follows. The next section provides a brief discussion on the YJ transformation mechanism. Section 3 presents the proposed model formulation and estimation procedure. Section 4 presents an empirical application of the proposed model in the context of individual's bicycling propensity and residential choice decision. Finally, concluding remarks are provided in Section 5.

2. THE YJ TRANSFORMATION

The Yeo and Johnson (2000) or YJ transformation has been widely used to transform data to near normality and symmetry for the marginal distributions. While other transformations are also available (see Goerg, 2015 for a discussion), the YJ transformation has the advantage of being a single-parameter transformation (so that only one additional parameter needs to be estimated to transform a non-normal distribution to a normal distribution for each margin, which leads to

parsimony in mixed data modeling). It has also been found to be robust and effective, and simplifies computations in econometric analysis (Smith et al., 2020).

The YJ transformation extends the well known Box-Cox transformation of a random variable/parameter Z_l to an assumed normal random variable/parameter G_l (with a mean parameter of μ_l and variance of σ_l^2) on the real line as follows, with an additional parameter $0 < \lambda_l < 2$:

$$G_l \sim N(\mu_l, \sigma_l^2) = t_{\lambda_l}(Z_l) = \begin{cases} -\dfrac{(-Z_l + 1)^{2-\lambda_l} - 1}{2 - \lambda_l} & \text{if } Z_l < 0 \\ \dfrac{(Z_l + 1)^{\lambda_l} - 1}{\lambda_l} & \text{if } Z_l > 0 \end{cases} \tag{1}$$

The transformation above is from the non-normal variable/parameter to the normal. The implied reverse transformation is as follows:

$$Z_l = t_{\lambda_l}^{-1}(G_l) = \begin{cases} 1 - \left[1 - (2 - \lambda_l) G_l\right]^{\left(\frac{1}{2-\lambda_l}\right)} & \text{if } G_l < 0 \\ \left[1 + G_l \lambda_l\right]^{\left(\frac{1}{\lambda_l}\right)} - 1 & \text{if } G_l > 0 \end{cases} \tag{2}$$

The transformation above allows for an asymmetric distribution for Z_l as well as thicker tails relative to the traditional normal distribution. To illustrate, Figure 2 plots Z_l for μ_l=0 and σ_l^2=1, and for different values of λ_l ($0 < \lambda_l < 2$). When $0 < \lambda_l < 1$, Z_l is skewed to the right with a thicker right tail, while if $1 < \lambda_l < 2$, Z_l is skewed to the left with a thicker left tail. When λ_l=1, the normal distribution is returned for Z_l. Thus, the YJ transformation allows for skew and fat tails. Across different variables/parameters $Z_1, Z_2, ..., Z_L$, the direction and intensity of skew/tail can vary. For future use, we define $\mathbf{Z} = \left(Z_1, Z_2, ..., Z_L\right)'$ ($L \times 1$ vector), $\mathbf{G} = \left(G_1, G_2, ..., G_L\right)'$ ($L \times 1$ vector), and $\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\mathbf{G}) = \left[t_{\lambda_1}^{-1}(G_1), t_{\lambda_2}^{-1}(G_2), ..., t_{\lambda_L}^{-1}(G_L)\right]$ ($L \times 1$ vector).

Construction of a Multivariate Gaussian Copula

Next, we construct an implicit Gaussian copula, by considering a multivariate normal distribution for the transformed G_l variables/parameters with another set of S_e ($e = 1, 2, ..., E$) variables/parameters. The S_e variables/parameters are considered normally distributed in their univariate margin, with mean μ_{e+L} and variance of σ_{e+L}^2). Let $\mathbf{S} = \left(S_1, S_2, ..., S_E\right)'$ ($E \times 1$ vector) Then, we have the following result for the multivariate distribution for the untransformed Z_l variables and the normally distributed S_e variables:

$$
\begin{aligned}
H(\mathbf{z},\mathbf{s}) = \ & \text{Prob}(\mathbf{Z}<\mathbf{z},\mathbf{S}<\mathbf{s}) = \text{Prob}\left[Z_1<z_1, Z_2<z_2,...,Z_L<z_L, S_1<s_1, S_2<s_2,...,S_L<s_L\right] \\
&= \text{Prob}\left[G_1<t_{\lambda_1}(z_1), G_2<t_{\lambda_2}(z_2),...,G_L<t_{\lambda_L}(z_L), S_1<s_1, S_2<s_2,...,S_L<s_L\right] \\
&= \text{Prob}(\mathbf{G}<\mathbf{g},\mathbf{S}<\mathbf{s}),\ \mathbf{g}=(g_1,g_2,...,g_L)',\ g_l = t_{\lambda_l}(z_l),\ l=1,2,...,L \\
&= F_{L+E}\left[\mathbf{g},\mathbf{s};\boldsymbol{\mu},\boldsymbol{\Omega}\right],\ \boldsymbol{\mu}=(\mu_1,\mu_2,...,\mu_L,\mu_{L+1},\mu_{L+2},...,\mu_{L+E})' = (\boldsymbol{\mu}_G,\boldsymbol{\mu}_S)'
\end{aligned}
\tag{3}
$$

and the diagonal elements of $\boldsymbol{\Omega}$ are the variances of the marginal elements (and thus are immediately identifiable). $F_{L+E}\left[.;\boldsymbol{\mu},\boldsymbol{\Omega}\right]$ in the equation above is the multivariate normal distribution function of dimension $L+E$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$ Thus, the multivariate distribution of $\mathbf{Y}=(\mathbf{Z}',\mathbf{S}')'$ is now translated to the multivariate normal distribution of $\tilde{\mathbf{Y}}=(\mathbf{G}',\mathbf{S}')'$. Next, partition the covariance matrix $\boldsymbol{\Omega}$ as follows: $\boldsymbol{\Omega}=\begin{bmatrix}\boldsymbol{\Omega}_G & \boldsymbol{\Omega}_{GS}' \\ \boldsymbol{\Omega}_{GS} & \boldsymbol{\Omega}_S\end{bmatrix}$. Immediately then, using the conditional distribution properties of the multivariate normal distribution, we are able to write the conditional distribution of the vector **S** conditional on **Z** as follows:

$$
\begin{aligned}
&\mathbf{S}\,|\,(\mathbf{Z}=\mathbf{z}) = \mathbf{S}\,|\,(\mathbf{G}=\mathbf{g}) \sim MVN_E(\tilde{\boldsymbol{\mu}}_S, \boldsymbol{\Psi}), \\
&\tilde{\boldsymbol{\mu}}_S = \boldsymbol{\Omega}_{GS}\boldsymbol{\Omega}_G^{-1}(\mathbf{g}-\boldsymbol{\mu}_G)+\boldsymbol{\mu}_S \text{ and } \boldsymbol{\Psi} = \boldsymbol{\Omega}_S - \boldsymbol{\Omega}_{GS}\boldsymbol{\Omega}_G^{-1}\boldsymbol{\Omega}_{GS}'.
\end{aligned}
\tag{4}
$$

This conditional distribution for **S** given **Z**, while accommodating the dependence between the two random vectors, plays a central role in the estimation of the proposed flexible GHDM model, as discussed in the next section.

3. THE FLEXIBLE 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. 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. Also, for ease in presentation, we will consider a combination of continuous, ordinal, and nominal outcomes in the presentation, though extension to include count variables and multiple discrete-continuous variables also is straightforward (see, for example, Bhat, 2015).

3.1. 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 follows in terms of its YJ-transformed counterpart G_l^* (based on Equation (1)).

$Z_l^* = t_{\lambda_l}^{-1}(G_l^*)$, where $G_l^* \sim N(\boldsymbol{\alpha}_l'\mathbf{w},1)$.

where **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 G_l^* is a normal random error term with variance one (the

normalization is for identification purposes; see Stapleton, 1978).[3] 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 $\boldsymbol{\mu}_{G^*} = (\boldsymbol{\alpha}_1'\mathbf{w}, \boldsymbol{\alpha}_2'\mathbf{w}, ..., \boldsymbol{\alpha}_L'\mathbf{w})'$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_L)'$, realizations of the random vector $\mathbf{Z}^* = (Z_1^*, Z_2^*, ..., Z_L^*)'$ as $\mathbf{z}^* = (z_1^*, z_2^*, ..., z_L^*)'$, and corresponding realizations of the random vector $\mathbf{G}^* = (G_1^*, G_2^*, ..., G_L^*)'$ as $\mathbf{g}^* = (g_1^*, g_2^*, ..., g_L^*)'$ $(g_l^* = t_{\lambda_l}(z_l^*),\ l = 1, 2, ..., L)$. Then, $\mathbf{G}^* \sim MVN(\boldsymbol{\mu}_{G^*}, \boldsymbol{\Omega}_{G^*})$, where the latent constructs are correlated through the non-diagonal correlation terms in $\boldsymbol{\Omega}_{G^*}$. The vector of latent constructs $\mathbf{Z}^*$ is thus characterized by the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$, and the correlation terms embedded in $\boldsymbol{\Omega}_{G^*}$ (say stacked into a $\left[L \times (L-1)/2\right]$-vector $\bar{\boldsymbol{\Omega}}_{G^*}$). In the typical GHDM, the vector $\boldsymbol{\lambda}$ is a priori assumed to be equal to $\mathbf{1}_L$ (that is a vector of length *L* with all elements equal to 1).

3.2. Latent Variable Measurement Equation Model (MEM) Component

We will consider a combination of continuous, ordinal, and nominal outcomes. For compactness in notation, we will confine the presentation to a single nominal outcome, though extension to multiple nominal outcomes is very straighforward as in Bhat (2015).

Let there be H continuous outcomes $(y_1, y_2, ..., y_H)$ with an associated index h $(h = 1, 2, ..., H)$. Let $y_h = \boldsymbol{\gamma}_h'\boldsymbol{x} + \boldsymbol{d}_h'(\ddot{\boldsymbol{x}}_h\mathbf{Z}^*) + \varepsilon_h$ 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 variables. $\boldsymbol{\gamma}_h$ is a corresponding compatible coefficient vector.[4] $\ddot{\boldsymbol{x}}_h$ is an $(\ddot{N}_h \times L)$-matrix of variables interacting with the latent constructs to influence y_h (with appropriately placed zero elements in specific columns), $\boldsymbol{d}_h$ is an $(\ddot{N}_h \times 1)$-column vector of coefficients capturing the effects of latent constructs and their interaction effects with other exogenous variables, and ε_h is a normally distributed measurement error term (if there are no interaction terms and the latent constructs have only a direct non-interactive effect, $\ddot{N}_h = L$). 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 in our factor model (where the latent vector $\mathbf{Z}^*$ represents the factor vector and serves as the vehicle to generate covariance between the outcome variables. Define the $(H \times A)$ matrix

[3] We exclude the constant in the covariate vector $\boldsymbol{w}$ for identification purposes. As we discuss in Section 3.2, we consider a constant in the vector of exogenous variables $\boldsymbol{x}$ in the continuous outcomes, as well as in the underlying continuous functions related to the ordinal indicators/outcomes and nominal outcomes; therefore, allowing for a separate constant in the covariate vector $\boldsymbol{w}$ simply will shift the constant value in these other measurement equations with no material effect in estimation/prediction.

[4] In joint limited-dependent variable systems in which one or more dependent outcome variables are not observed on a continuous scale, such as the joint system considered in the current paper that has discrete dependent variables, the structural effects of one dependent variable on another can only be in a single direction. This applies to all occurrences of the vector $\boldsymbol{x}$ in other parts of the model.

$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_H)'$, and the $\left(\sum_{h=1}^{H} \ddot{N}_h \times L \right)$ matrix $\ddot{\boldsymbol{x}} = (\ddot{\boldsymbol{x}}_1', \ddot{\boldsymbol{x}}_2', ..., \ddot{\boldsymbol{x}}_H')'$ matrix. Also, define the $\left(H \times \sum_{h=1}^{H} \ddot{N}_h \right)$ matrix $\boldsymbol{d}$, which is initially filled with all zero values. Then, position the $(1 \times \ddot{N}_1)$ row vector $\boldsymbol{d}_1'$ in the first row to occupy columns 1 to $\ddot{N}_1$, position the $(1 \times \ddot{N}_2)$ row vector $\boldsymbol{d}_2'$ in the second row to occupy columns $\ddot{N}_1 + 1$ to $\ddot{N}_1 + \ddot{N}_2$, and so on until the $(1 \times \ddot{N}_H)$ row vector $\boldsymbol{d}_H'$ is appropriately positioned. Further, define $\boldsymbol{\pi} = \boldsymbol{d}\ddot{\boldsymbol{x}}$ ($H \times L$ matrix) and $\bar{\boldsymbol{\pi}} = \text{Vech}(\boldsymbol{\pi})$ (that is, $\bar{\boldsymbol{\pi}}$ is a column vector that includes all elements of the matrix $\boldsymbol{\pi}$).Then, one may write, in matrix form, the following measurement equation for the continuous outcomes:

$$\boldsymbol{y} = \boldsymbol{\gamma x} + \boldsymbol{\pi Z}^* + \boldsymbol{\varepsilon}. \tag{5}$$

Next, consider N ordinal outcomes for the individual, and let n be the index for the ordinal outcomes $(n = 1, 2, ..., N)$ (some of these will be purely indicators of latent constructs, facilitating the estimation of the SEM component of the model, that is, estimation of the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ in Section 3.1) . 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:

$$\tilde{y}_n^* = \tilde{\boldsymbol{\gamma}}_n \boldsymbol{x} + \tilde{\boldsymbol{d}}_n' \boldsymbol{Z}^* + \tilde{\varepsilon}_n, \text{ and } \tilde{\psi}_{n,a_n - 1} < \tilde{y}_n^* < \tilde{\psi}_{n,a_n}, \tag{6}$$

where $\tilde{\boldsymbol{d}}_n$ is an $(L \times 1)$ vector of latent variable loadings on the underlying propensity of the n^{th} ordinal 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 $\tilde{\boldsymbol{\psi}}_n = (\tilde{\psi}_{n,2}, \tilde{\psi}_{n,3} ..., \tilde{\psi}_{n,J_n - 1})'$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1', \tilde{\boldsymbol{\psi}}_2', ..., \tilde{\boldsymbol{\psi}}_N)'$. 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 $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1, \tilde{\boldsymbol{\gamma}}_2, ..., \tilde{\boldsymbol{\gamma}}_N)'$ [$(N \times A)$ matrix] and $\tilde{\boldsymbol{d}} = \left(\tilde{\boldsymbol{d}}_1, \tilde{\boldsymbol{d}}_2, ..., \tilde{\boldsymbol{d}}_N \right)$ [$(N \times L)$ matrix], and let $\mathbf{IDEN}_N$ be the identity matrix of dimension N representing the correlation matrix of $\tilde{\boldsymbol{\varepsilon}}$ (so, $\tilde{\boldsymbol{\varepsilon}} \sim MVN_N \left(\mathbf{0}_N, \mathbf{IDEN}_N \right)$; again, this is for parsimony purposes, given the presence of the unobserved $\boldsymbol{Z}^*$ vector to generate covariance. For the pure ordinal indicators of the latent constructs, all elements of the vector $\tilde{\boldsymbol{\gamma}}_n$ are set to zero. 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:

$$\tilde{\boldsymbol{y}}^{*} = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{Z}^{*} + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\psi}}_{low} < \tilde{\boldsymbol{y}}^{*} < \tilde{\boldsymbol{\psi}}_{up}. \tag{7}$$

Finally, consider the nominal (unordered-response) variable. Let I be the number of alternatives let i be the corresponding index $(i = 1,\ 2,\ 3,...,\ I)$. Assume that the individual under consideration chooses the alternative m. Also, assume the usual random utility structure for each alternative i.

$$U_i = \boldsymbol{b}_i'\boldsymbol{x} + \boldsymbol{\vartheta}_i'(\tilde{\boldsymbol{x}}_i\boldsymbol{Z}^{*}) + \varsigma_i, \tag{8}$$

where $\boldsymbol{x}$ is as defined earlier, $\boldsymbol{b}_i$ is an $(A \times 1)$ column vector of corresponding coefficients, and ς_i is a normal error term. $\tilde{\boldsymbol{x}}_i$ is an $(N_i \times L)$-matrix of variables interacting with latent variables to influence the utility of alternative *i*, and $\boldsymbol{\vartheta}_i$ is an $(N_i \times 1)$-column vector of coefficients capturing the effects of latent variables and their interaction effects with other exogenous variables.[5] The fact that $\boldsymbol{Z}^{*}$ is stochastic immediately implies the presence of random coefficients on the exogenous variables appearing in the matrix $\tilde{\boldsymbol{x}}_i$. Besides, because we are allowing the latent constructs to have asymmetry and tails, the random coefficients are allowed to take a flexible distributional shape. Importantly, the random coefficients on multiple exogenous variables are generated through only a handful of latent constructs, making the specification very parsimonious. Let $\boldsymbol{\varsigma} = (\varsigma_1, \varsigma_2, ...\varsigma_I)'$ ($I_g \times 1$ vector), and $\boldsymbol{\varsigma} \sim MVN_I(\mathbf{0}, \mathbf{IDEN}_I)$. The usual identification 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 $\boldsymbol{b}_i$ is set to zero for at least one alternative *i*). To proceed, define $\mathbf{U} = (U_1, U_2, ..., U_I)'$ ($I \times 1$ vector), $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, ..., \boldsymbol{b}_I)'$ ($I \times A$ matrix), and $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1', \tilde{\boldsymbol{x}}_2', ..., \tilde{\boldsymbol{x}}_I')'$ $\left(\sum_{i=1}^{I} N_i \times L\right)$ matrix. Also, define the $\left(I \times \sum_{i=1}^{I} N_i\right)$ matrix $\boldsymbol{\vartheta}$, which is initially filled with all zero values. Then, position the $(1 \times N_1)$ row vector $\boldsymbol{\vartheta}_1'$ in the first row to occupy columns 1 to N_1, position the $(1 \times N_2)$ row vector $\boldsymbol{\vartheta}_2'$ in the second row to occupy columns N_1+1 to $N_1 + N_2$, and so on until the $(1 \times N_I)$ row vector $\boldsymbol{\vartheta}_I'$ is

[5] Of course, as in any discrete choice model, only differences in utilities matter, which implies that any individual-specific covariates in the vector $\boldsymbol{x}$ can appear in the utilities of utmost *I*–1 alternatives. This also applies to the latent construct elements in $\mathbf{Z}^{*}$. Any element of $\mathbf{Z}^{*}$ that is purely a function of individual-specific covariates in the SEM component can be introduced in the utilities of utmost *I*–1 alternatives. Also, in multinomial probit models, identification is tenuous if only individual-specific covariates appear in the vector $\boldsymbol{x}$ and if elements of $\mathbf{Z}^{*}$ are a function of purely individual-specific variables (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.

appropriately positioned. Further, define $\boldsymbol{\varpi} = (\boldsymbol{\vartheta}\tilde{\boldsymbol{x}})$ ($I \times L$ matrix) and $\bar{\boldsymbol{\vartheta}} = \text{Vech}(\boldsymbol{\vartheta})$ (that is, $\bar{\boldsymbol{\vartheta}}$ is a column vector that includes all elements of the matrix $\boldsymbol{\vartheta}$). Then, in matrix form, we may write Equation (8) as:

$$\boldsymbol{U} = \boldsymbol{bx} + \boldsymbol{\varpi}\,\mathbf{Z}^* + \boldsymbol{\varsigma}. \tag{9}$$

To proceed further, let $E = (H + N + I)$. Define $\vec{\boldsymbol{y}} = \left(\boldsymbol{y}', \left[\tilde{\boldsymbol{y}}^* \right]', \mathbf{U}' \right)'$ [$E \times 1$ vector], $\vec{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}', \tilde{\boldsymbol{\gamma}}', \boldsymbol{b}')'$ [$E \times A$ matrix], $\vec{\boldsymbol{d}} = (\boldsymbol{\pi}', \tilde{\boldsymbol{d}}', \boldsymbol{\varpi}')'$ [$E \times L$ matrix], and $\vec{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}', \tilde{\boldsymbol{\varepsilon}}', \boldsymbol{\varsigma}')'$ ($E \times 1$ vector), where $\mathbf{0}_{AN}$ is a matrix of zeros of dimension $A \times N$. With these matrix definitions, the MEM latent variable component of the model system may be written compactly as:

$$\vec{\boldsymbol{y}} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\boldsymbol{d}}\mathbf{Z}^* + \vec{\boldsymbol{\varepsilon}}, \text{ with } \vec{\boldsymbol{\varepsilon}} \sim \text{MVN}(\mathbf{0}_E, \boldsymbol{\Omega}_{\vec{\varepsilon}}),\ \boldsymbol{\Omega}_{\vec{\varepsilon}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{IDEN}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{IDEN}_I \end{bmatrix} (E \times E \text{ matrix}). \tag{10}$$

Define $\mathbf{Y} = (\mathbf{Z}^*, \vec{\boldsymbol{\varepsilon}})'$. The correspondence of our notations with the previous section should now be clear, with $\mathbf{Z}^*$ taking the place of $\mathbf{Z}$, and $\vec{\boldsymbol{\varepsilon}}$ taking the place of $\mathbf{S}$. Then, following the previous section, we may write the joint cumulative multivariate distribution of $\mathbf{Y}$ exactly as in Equation (3) after translating it into an equivalent joint cumulative multivariate standard normal distribution of $\tilde{\mathbf{Y}} = (\mathbf{G}^*, \vec{\boldsymbol{\varepsilon}})'$. Next, partition the covariance matrix $\boldsymbol{\Omega}$ of $\tilde{\mathbf{Y}}$ as follows: $\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{G^*} & \boldsymbol{\Omega}'_{G^*\vec{\varepsilon}} \\ \boldsymbol{\Omega}_{G^*\vec{\varepsilon}} & \boldsymbol{\Omega}_{\vec{\varepsilon}} \end{bmatrix}$. The elements of the covariance matrix $\boldsymbol{\Omega}'_{G^*\vec{\varepsilon}}$ corresponding to the correlations between the transformed vector of the latent constructs, $\mathbf{G}^*$, and the pure indicator-related kernel error terms in the vector $\tilde{\boldsymbol{\varepsilon}}$ are normalized to zero for identification. Also, without any loss of generality and for identification considerations, we will normalize the correlation between the elements of $\mathbf{G}^*$ and the utility of the first alternative in the nominal variable to zero. Immediately then, using the conditional distribution properties of the multivariate normal distribution, we are able to write the conditional distribution of the $\vec{\boldsymbol{\varepsilon}}$ vector conditional on $\mathbf{Z}^*$ as follows:

$$\begin{aligned} &\vec{\boldsymbol{\varepsilon}} \,|\, (\mathbf{Z}^* = \mathbf{z}^*) = \vec{\boldsymbol{\varepsilon}} \,|\, (\mathbf{G}^* = \mathbf{g}^*) \sim MVN_E(\vec{\boldsymbol{\mu}}_{\mathbf{g}^*}, \boldsymbol{\Theta}), \\ &\vec{\boldsymbol{\mu}}_{\mathbf{g}^*} = \boldsymbol{\Omega}_{G^*\vec{\varepsilon}}\boldsymbol{\Omega}_{G^*}^{-1}(\mathbf{g}^* - \boldsymbol{\mu}_{G^*}) + \mathbf{0}_E \text{ and } \boldsymbol{\Theta} = \boldsymbol{\Omega}_{\vec{\varepsilon}} - \boldsymbol{\Omega}_{G^*\vec{\varepsilon}}\boldsymbol{\Omega}_{G^*}^{-1}\boldsymbol{\Omega}'_{G^*\vec{\varepsilon}}, \text{ with } \boldsymbol{\mu}_{G^*} = (\boldsymbol{\alpha}_1'\mathbf{w}, \boldsymbol{\alpha}_2'\mathbf{w}, ..., \boldsymbol{\alpha}_L'\mathbf{w})'. \end{aligned} \tag{11}$$

The traditional mixed models (including the GHDM model) assume that all entries of the covariance matrix $\boldsymbol{\Omega}'_{G^*\vec{\varepsilon}}$ are identically zero, *a priori* imposing the restrictions that the latent construct vector (that is, factor vector) $\mathbf{Z}^*$ is independent of the kernel error terms of the dependent outcomes of interest. However, in our flexible model, we allow dependence between the $\mathbf{Z}^*$ vector and the kernel error terms that are not pure indicators of the $\mathbf{Z}^*$ vector.

3.3. Model Identification

Let $\boldsymbol{\delta}$ be the collection of parameters to be estimated: $\boldsymbol{\delta} = [\boldsymbol{\lambda}, \text{Vech}(\boldsymbol{\alpha}), \text{Vech}(\vec{\boldsymbol{\gamma}}), \text{Vech}(\vec{\boldsymbol{d}}), \tilde{\boldsymbol{\psi}}, \text{Vech}(\boldsymbol{\vartheta}), \text{Vech}(\boldsymbol{\Sigma}), \text{Vech}(\boldsymbol{\Omega})]$, where the operator "Vech(.)" vectorizes all the non-zero elements of the matrix/vector on which it operates. Sufficient conditions for identification of all parameters in $\boldsymbol{\delta}$ are identical to that discussed at length in Bhat (2015) based on O'Brien's exposition. However, the sufficiency condition for identification of $\vec{\boldsymbol{\gamma}}$ in Bhat (2015) is unnecessarily stringent; as long as the elements of $\boldsymbol{\alpha}$ in the SEM model component are identified solely from the ordinal indicator variables (for which $\tilde{\gamma}_n$ =0 in Equation (6)), there is no need for additional exclusion restrictions between the vectors $\boldsymbol{x}$ and $\mathbf{w}$ (see Bhat and Dubey, 2014)**.** In particular, identification is achieved if the following conditions hold:

(a) $\boldsymbol{\Omega}_{G^*}$ in the structural equation is specified to be a correlation matrix, with each latent variable correlated with at least one other latent variable (if $L \geq 2$),

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

(c) If $L \geq 2$, for each latent construct, there are at least two indicator variables that load only on that latent construct and no other latent construct (that is, there are at least two factor complexity one indicator variables for each latent variable) (see Reilly and O'Brien, 1996); If L=1, there are at least three indicator variables that load on that latent construct (that is, there are at least three factor complexity indcator variables for the latent variable).

(d) For the indicator variables that load only on a single latent construct, the dependence between the kernel error terms of those indicator variables and the corresponding latent construct is maintained to be zero.

(e) The dependence between the latent constructs and the kernel error term of the utility of one of the alternatives for the nominal variable (say the first alternative) is set to zero (this is, of course, because only utility differences matter in the nominal variable modeling).

3.4. Model Estimation

To estimate the model, the first order of business is to develop the conditional likelihood function of the dependent outcomes, given the latent construct values. To do so, using Equation (11), we begin by writing Equation (10) in conditional form as follows:

$$\vec{\boldsymbol{y}} \left| (\mathbf{Z}^* = \mathbf{z}^*) = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\vec{\boldsymbol{d}}}\mathbf{z}^* + \vec{\boldsymbol{\varepsilon}} \right| (\mathbf{Z}^* = \mathbf{z}^*) \sim MVN(\mathbf{B}, \boldsymbol{\Theta}), \text{ where } \mathbf{B} = \vec{\boldsymbol{\gamma}}\boldsymbol{x} + \vec{\vec{\boldsymbol{d}}}\mathbf{z}^* + \vec{\boldsymbol{\mu}}_{\mathbf{g}^*}. \tag{12}$$

Next, under the utility maximization paradigm, consider an individual who chose the alternative m. Then, $U_i - U_m$ must be less than zero for all $i \neq m$. Let $u_{im} = U_i - U_m (i \neq m)$, and stack the latent utility differentials into a vector $\mathbf{u} = \left[\left(u_{1m}, u_{2m}, ..., u_{Im} \right)' ; i \neq m \right]$. We now need to develop the

distribution of the vector $\boldsymbol{yu} = \left(\boldsymbol{y}', \left[\tilde{\boldsymbol{y}}^* \right]', \mathbf{u}' \right)'$ [$(E-1)\times 1$ vector] given $(\mathbf{Z}^* = \mathbf{z}^*)$ from that of $\vec{\boldsymbol{y}} \big| (\mathbf{Z}^* = \mathbf{z}^*)$. To do so, define a matrix **M** of size $\left[E-1\right] \times \left[E\right]$. Fill this matrix with values of zero. Then, insert an identity matrix of size *H*+*N* into the first *H*+*N* rows and *H*+*N* columns of the matrix **M**. Next, consider the rows from $H+N+1$ to $H+N+I-1$, and columns from $H+N+1$ to *E*. Insert an identity matrix of size $(I-1)$ after supplementing with a column of '-1' values in the column corresponding to the chosen alternative. With the matrix **M** as defined, we can write $\boldsymbol{yu} \big| (\mathbf{Z}^* = \mathbf{z}^*) \sim MVN_{\boldsymbol{E\text{-}I}}(\tilde{\mathbf{B}}, \tilde{\boldsymbol{\Theta}})$, where $\tilde{\mathbf{B}} = \mathbf{MB}$ and $\tilde{\boldsymbol{\Theta}} = \mathbf{M\Theta M}'$. Next, partition the vector $\tilde{\mathbf{B}}$ into components that correspond to the mean of the vector $\boldsymbol{y}$ (for the continuous variables), and $\tilde{\mathbf{u}} = \left(\tilde{\boldsymbol{y}}^{*\prime}, \mathbf{u}' \right)'$ (for the ordinal outcomes and the nominal outcomes). Correspondingly, partition the matrix $\tilde{\boldsymbol{\Theta}}$:

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\mathbf{B}}_{\boldsymbol{y}} \\ \tilde{\mathbf{B}}_{\tilde{\mathbf{u}}} \end{bmatrix} \; E \times 1 \text{ vector and } \tilde{\boldsymbol{\Theta}} = \begin{bmatrix} \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}} & \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}\tilde{\mathbf{u}}} \\ \tilde{\boldsymbol{\Theta}}'_{\boldsymbol{y}\tilde{\mathbf{u}}} & \tilde{\boldsymbol{\Theta}}_{\tilde{\mathbf{u}}} \end{bmatrix}$$

The conditional distribution of $\tilde{\mathbf{u}}$, given $\boldsymbol{y}$ and $(\mathbf{Z}^* = \mathbf{z}^*)$, is MVN with mean $\vec{\mathbf{B}}_{\tilde{\mathbf{u}}} = \tilde{\mathbf{B}}_{\tilde{\mathbf{u}}} + \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}\tilde{\mathbf{u}}} \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}}^{-1} \left(\boldsymbol{y} - \tilde{\mathbf{B}}_{\boldsymbol{y}} \right)$ and variance $\vec{\boldsymbol{\Theta}}_{\tilde{\mathbf{u}}} = \tilde{\boldsymbol{\Theta}}_{\tilde{\mathbf{u}}} - \tilde{\boldsymbol{\Theta}}'_{\boldsymbol{y}\tilde{\mathbf{u}}} \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}}^{-1} \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}\tilde{\mathbf{u}}}$. Next, define threshold vectors as follows:

$$\vec{\boldsymbol{\psi}}_{\boldsymbol{low}} = \left[\tilde{\boldsymbol{\psi}}'_{\boldsymbol{low}}, \left(-\boldsymbol{\infty}_{I-1} \right)', \right]' \; ([(N+I-1)\times 1] \text{ vector}) \text{ and } \vec{\boldsymbol{\psi}}_{\boldsymbol{up}} = \left[\tilde{\boldsymbol{\psi}}'_{\boldsymbol{up}}, \left(\mathbf{0}_{I-1} \right)' \right]' \; ([(N+I-1)\times 1] \text{ vector}),$$

where $-\boldsymbol{\infty}_{I-1}$ is a $(I-1)\times 1$-column vector of negative infinities, and $\mathbf{0}_{I-1}$ is another $(I-1)\times 1$-column vector of zeros. Also, let $\vec{\boldsymbol{\omega}}_{\tilde{\mathbf{u}}}$ be a diagonal matrix containing the square root of the variance elements of $\vec{\boldsymbol{\Theta}}_{\tilde{\mathbf{u}}}$, and let $\boldsymbol{\tau}_{\boldsymbol{low}} = \left(\vec{\boldsymbol{\omega}}_{\tilde{\mathbf{u}}} \right)^{-1} (\vec{\boldsymbol{\psi}}_{\boldsymbol{low}} - \vec{\mathbf{B}}_{\tilde{\mathbf{u}}})$, $\boldsymbol{\tau}_{\boldsymbol{up}} = \left(\vec{\boldsymbol{\omega}}_{\tilde{\mathbf{u}}} \right)^{-1} (\vec{\boldsymbol{\psi}}_{\boldsymbol{up}} - \vec{\mathbf{B}}_{\tilde{\mathbf{u}}})$, and $\vec{\boldsymbol{\Theta}}^*_{\tilde{\mathbf{u}}} = \left(\vec{\boldsymbol{\omega}}_{\tilde{\mathbf{u}}} \right)^{-1} \vec{\boldsymbol{\Theta}}_{\tilde{\mathbf{u}}} \left(\vec{\boldsymbol{\omega}}_{\tilde{\mathbf{u}}} \right)^{-1}$. Also, let *d* be an index that takes values from 1 to $(N+I-1)$, and let $\tau_{low,d}$ and $\tau_{up,d}$ represent the elements of the vectors $\boldsymbol{\tau}_{\boldsymbol{low}}$ and $\boldsymbol{\tau}_{\boldsymbol{up}}$, respectively. Then the likelihood function, conditional on $\mathbf{Z}^*$, may be written, in its most general form, as:

$$L(\boldsymbol{\delta}) \big| (\mathbf{Z}^* = \mathbf{z}^*) = f_H(\boldsymbol{y} \mid \tilde{\mathbf{B}}_{\boldsymbol{y}}, \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}}) \times \Pr\left[\vec{\boldsymbol{\psi}}_{\boldsymbol{low}} \le \tilde{\mathbf{u}} \le \vec{\boldsymbol{\psi}}_{\boldsymbol{up}} \right] . \tag{13}$$

Based on the inclusion-exclusion probability law, and for all Fretchet class of multivariate distribution functions with given univariate margins (of which the multivariate normal distribution is a part), the expression in Equation (13) can be re-written with the matrix definitions above as follows:

$$L(\boldsymbol{\delta}) \big| (\mathbf{Z}^* = \mathbf{z}^*) = f_H(\boldsymbol{y} \mid \tilde{\mathbf{B}}_{\boldsymbol{y}}, \tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}}) \times \sum_{S=1}^{S=2^{N+I-1}} (-1)^{A_S} \Phi_{N+I-1}(\tilde{\boldsymbol{\tau}}_S; \vec{\boldsymbol{\Theta}}^*_{\tilde{\mathbf{u}}}) \tag{14}$$

where S represents a specific combination of length $(N+I-1)$ of the $\tau_{low,d}$ and $\tau_{up,d}$ scalars across all possible d values such that both $\tau_{low,d}$ and $\tau_{up,d}$ are disallowed in the combination for any S (there are 2^{N+I-1} such combinations, each of length $(N+I-1)$, and we will represent the resulting vector of elements in combination S as $\tilde{\boldsymbol{\tau}}_S$), A_S is a count of the number of lower threshold elements $\tau_{low,d}$ appearing in the vector $\tilde{\boldsymbol{\tau}}_S$, and $\Phi_{N+I-1}(.,.)$ is the standard multivariate normal cumulative distribution (MVNCD) function of dimension $(N+I-1)$. In the case when all the ordinal variables are pure indicator variables of the latent constructs (that is, there is no dependence between the error terms underlying these ordinal variables and the latent constructs), the conditional likelihood of Equation (14) reduces to:

$$L(\boldsymbol{\delta})\left|(\mathbf{Z}^*=\mathbf{z}^*)=f_H(\boldsymbol{y}\,|\,\tilde{\mathbf{B}}_{\boldsymbol{y}},\tilde{\boldsymbol{\Theta}}_{\boldsymbol{y}})\times\left[\prod_{d=1}^{N}\left\{\Phi(\tau_{up,d})-\Phi(\tau_{low,d})\right\}\right]\times\Phi_{I-1}(\mathbf{0}_{\mathbf{I\text{-}1}};\ddot{\boldsymbol{\Theta}}^*_{\mathbf{u}}),\right. \tag{15}$$

where $\ddot{\boldsymbol{\Theta}}^*_{\mathbf{u}}$ refers to the sub-matrix of $\ddot{\boldsymbol{\Theta}}^*_{\ddot{\mathbf{u}}}$ corresponding only to the utility elements of the nominal variable. The MVNCD functions can be evaluated using the very accurate analytic approximation methods developed by Bhat (2018) and embedded within the MACML inference technique. In particular, in the case that N is very large or I is very large (>10 say), one may use a pairwise likelihood function for the conditional likelihood function of Equation (14), exactly similar to the traditional GHDM (see Section 3.3 of Bhat, 2015). Finally, the unconditional likelihood function may be obtained as follows:

$$L(\boldsymbol{\delta})=\int_{z_1^*=-\infty}^{z_1^*=+\infty}\int_{z_2^*=-\infty}^{z_2^*=+\infty}\ldots\int_{z_L^*=-\infty}^{z_L^*=+\infty}\left[L(\boldsymbol{\delta})\,|\,(\mathbf{Z}^*=\mathbf{z}^*)\right]h_L(\mathbf{z}^*)dz_1^*dz_2^*\ldots dz_L^* \tag{16}$$

The above equation can be re-written based on transforming the density of the random variable vector $\mathbf{Z}^*$ to the corresponding multivariate normal vector $\mathbf{G}^*$, using the usual Jacobian transformation result:

$$h_{\mathbf{Z}^*}(\mathbf{z}^*)=\text{ Prob}(\mathbf{Z}^*=\mathbf{z}^*)=f_{\mathbf{G}^*}(\mathbf{g}^*)\times abs\left|\frac{d\mathbf{g}^*}{d\mathbf{z}^*}\right|=f_{\mathbf{G}^*}(\mathbf{g}^*)\times\frac{dg_1^*}{dz_1^*}\times\frac{dg_2^*}{dz_2^*}\ldots\times\frac{dg_L^*}{dz_L^*},\text{ from which}$$
$$h_{\mathbf{Z}^*}(\mathbf{z}^*)dz_1^*dz_2^*\ldots dz_L^*=f_{\mathbf{G}^*}(\mathbf{g}^*)dg_1^*dg_2^*\ldots dg_L^*. \tag{17}$$

Substituting the above equality in Equation (16), and defining a diagonal matrix $\boldsymbol{\omega}_{G^*}$ whose entries correspond to the square root of the diagonal entries of the matrix $\boldsymbol{\Omega}_{G^*}$, the result is:

$$\begin{aligned}L(\boldsymbol{\delta})&=\int_{g_1^*=-\infty}^{g_1^*=+\infty}\int_{g_2^*=-\infty}^{g_2^*=+\infty}\ldots\int_{g_L^*=-\infty}^{g_L^*=+\infty}L(\boldsymbol{\delta})\,|\left(\mathbf{Z}^*=\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\mathbf{g}^*)\right)f_{\mathbf{G}^*}(\mathbf{g}^*)dg_1^*dg_2^*\ldots dg_L^*=\int_{\mathbf{g}^*=-\infty}^{\mathbf{g}^*=+\infty}L(\boldsymbol{\delta})\left|\left(\mathbf{Z}^*=\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\mathbf{g}^*)\right)\right.f_{\mathbf{G}^*}(\mathbf{g}^*)d\mathbf{g}^*\\&=\left(\prod_{l=1}^{L}\omega_{G_l^*}\right)^{-1}\times\int_{\tilde{\mathbf{g}}^*=-\infty}^{\tilde{\mathbf{g}}^*=+\infty}L(\boldsymbol{\delta})\left|\left(\mathbf{Z}^*=\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\omega}_{G^*}\tilde{\mathbf{g}}^*+\boldsymbol{\mu}_{\mathbf{G}^*})\right)\right.\phi_L\left(\tilde{\mathbf{g}}^*;\boldsymbol{\Omega}_{G^*}\right)d\tilde{\mathbf{g}}^*,\text{ where }\tilde{\mathbf{g}}^*=\boldsymbol{\omega}_{G^*}^{-1}[\mathbf{g}^*-\boldsymbol{\mu}_{\mathbf{G}^*}],\end{aligned} \tag{18}$$

$\omega_{G_l^*}$ is the *l*th diagonal element of the matrix $\boldsymbol{\omega}_{G^*}$, and $\phi_L(.;.)$ represents the standard multivariate normal density function. The integration in Equation (18) can be undertaken using usual simulation-based procedures, thus combining the maximum simulated likelihood approach with an analytic method to evaluate the MVNCD function.

There are two final issues. The first is that the covariance matrices $\boldsymbol{\Omega}$ has to be positive definite. The simplest 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 (rather than the original covariance matrices). Also, except for the diagonal elements corresponding to the continuous outcome variables, all other diagonal entries of this matrix are normalized to one (primarily for identification). To achieve this, we parameterize $\boldsymbol{\Omega}$ as $\boldsymbol{\omega}_{\boldsymbol{\Omega}}\boldsymbol{\Omega}^*\boldsymbol{\omega}_{\boldsymbol{\Omega}}$, where $\boldsymbol{\omega}_{\boldsymbol{\Omega}}$ is the diagonal matrix of the square root of the elements of $\boldsymbol{\Omega}$ (with the elements corresponding to the non-continuous outcomes/indicators fixed to one) and $\boldsymbol{\Omega}^*$ being a correlation matrix. Then, the parameters actually estimated correspond to the non-fixed elements of $\boldsymbol{\omega}_{\boldsymbol{\Omega}}$ and the Cholesky elements of $\boldsymbol{\Omega}^*$. We use a spherical parameterization of the Cholesky of the correlation matrix to ensure a positive-definite correlation matrix in estimation (see Bhat and Mondal, 2021). The parameters to be estimated in the model may be written as: $\boldsymbol{\delta} = [\boldsymbol{\lambda}, \text{Vech}(\boldsymbol{\alpha}), \text{Vech}(\vec{\boldsymbol{\gamma}}), \text{Vech}(\vec{\boldsymbol{d}}), \tilde{\boldsymbol{\psi}}, \text{Vech}(\boldsymbol{\Sigma}), \text{Vech}(\boldsymbol{\vartheta}), \text{Vech}(\boldsymbol{\omega}_{\boldsymbol{\Omega}}), \text{Vech}(\mathbf{L}_{\boldsymbol{\Omega}^*})]$, where $\mathbf{L}_{\boldsymbol{\Omega}^*}$ is the spherically-parameterized lower Cholesky matrix of the matrix $\boldsymbol{\Omega}^*$.

The second issue relates to the starting parameters. In our experimentation of alternative procedures to arrive at good starting values, the following procedure worked well: (a) First estimate the SEM model alone using only the indicators of the latent constructs (this is basically equivalent to the estimation of a multiple-indicator multiple-cause (MIMIC) model; see O'Brien, 1994 and Reilly and O'Brien, 1996 (that is, estimate $\boldsymbol{\lambda}, \text{Vech}(\boldsymbol{\alpha}), \tilde{\boldsymbol{d}}, \tilde{\boldsymbol{\psi}}$, and $\text{Vech}(\boldsymbol{\Omega}_{G^*})$), (b) Estimate the parameters of the measurement equation model for the actual dependent outcomes of interest, assuming zero correlations between the latent construct errors and the errors of the dependent outcomes of interest (that is $\boldsymbol{\gamma}, \boldsymbol{b}, \boldsymbol{d}, \text{Vech}(\boldsymbol{\vartheta})$, and $\text{Vech}(\boldsymbol{\Sigma})$, fixing the SEM coefficients to that estimated at step (a) and setting all the elements of $\boldsymbol{\Omega}_{G^*\tilde{\varepsilon}}$ to be zero) (c) Estimate the correlation parameters between the latent construct errors and the dependent outcome errors (that is, the elements of $\boldsymbol{\Omega}_{G^*\tilde{\varepsilon}}$) fixing all other coefficients from steps (a) and (b), and (d) Use the coefficient vector from the estimation results in step (c) to begin the iterations for the overall estimation of the model system.

4. MODEL APPLICATION

4.1. Background

In this section, we demonstrate an application of our proposed flexible GHDM framework and compare its effectiveness with the traditional GHDM structure and its variants. To do so, we undertake an empirical study to investigate individuals' bicycling propensity as an ordered

outcome and residential choice (high-density neighborhood living vs. low/medium density neighborhood living) as a binary outcome. Promoting bicycling and walking modes of transportation has been of substantial interest to transportation planners, as these non-motorized travel modes are associated with zero mobile-source emissions, lead to lower levels of traffic congestion, and foster an active and healthy lifestyle (Manaugh et al., 2017, Park and Akar, 2019, Bhat et al., 2017). In this regard, one thread of scientific enquiry has focused on evaluating the extent of benefits in investing in non-motorized-friendly built environment (BE) policies and infrastructure (for example, construction of exclusive bike lanes, traffic safety measures prioritizing bicycle users, building walk/bike friendly neighborhood with ample bicycle docking facilities and protected lanes). Bhat and Guo (2007) discuss many possible methodological directions that such an enquiry may take, each with its own advantages and disadvantages. But the data typically available for such an analysis is cross-sectional, and so much effort has gone into disentangling associative effects (that is, residential self-selection effects wherein individuals may choose a residential location based on their bicycling preferences) from "true" causal effects of residential BE attributes. This entails recognizing that residential choice may be endogenous to non-motorized mode use decisions, typically by modeling both choices as a "bundle" or a package choice using cross-sectional data (see, for example, Pinjari et al., 2008, Pinjari et al., 2011, Paleti et al., 2013, Bhat et al., 2013, Jarass and Scheiner, 2018, Wolday et al., 2019, Guan and Wang, 2019, Kroesen 2019, Deng and Zhao, 2022, and Millard-Ball et al., 2022). In this paper, along these lines, we consider the two dimensions of high-density neighborhood (HDN) living and bicycling frequency as our outcomes of interest in our proposed flexible GHDM methodological framework.

The express focus of this empirical study is primarily to demonstrate an application of our proposed flexible GHDM model. However, within the context of the data available, we tested several functional forms of the exogenous variables to arrive at the final specification. Subsequently, we provide a substantive interpretation of the variable effects. Moreover, we also provide a comparative analysis between our proposed model and the traditional GHDM variants (which are restricted versions of our proposed model), and underscore the potential pitfalls in terms of data fit as well as policy misinformation when non-normality of the latent construct error terms and/or the correlation between the latent constructs and the main outcomes are ignored.

4.2. Data and Sample Used

Our empirical application employs a sample drawn from the 2019 wave of the Puget Sound Regional Council Household Travel Study. The survey elicited, from a representative sample of the regional population, information on general lifestyle and travel-specific attitudes/preferences, residential location, and current travel patterns. For the current analysis, individuals below 18 years of age and whose survey responses were recorded through proxy-reporting were excluded. The final sample used in our study comprised 3645 individuals.

The residential locations of respondents was obtained in the survey in terms of the household Census tract of living. Each household's residential location was assigned to either a

low/medium density tract (≤10,000 individuals per square mile) or a high density tract (>10,000 individuals per square mile). Those residing in a high density tract are designated as living in a high-density neighborhood (HDN), which forms the binary outcome of interest ("HDN living" versus not). The use of such a binary residential density metric for residential choice provides a convenient way to capture land-use/BE effects on activity-travel behaviors, particularly because of the strong association between density and other BE elements. Indeed, there is a long and strong precedent for using such a binary residential density variable as a proxy for land-use/BE elements in the transportation literature (see, for example, Chen et al., 2008, Kim and Brownstone, 2013, Paleti et al., 2013, Cao and Fan, 2012, Bhat et al., 2016a, Falk and Katz-Gerro, 2017, and Gallo, 2020).

The second outcome of interest in our analysis corresponds to the bicycling use frequency, elicited from respondents through the following question: *"In the past 30 days, how often have you ridden a bicycle (for 15 minutes or more)?"*. The responses were recorded in six ordinal categories, which were collapsed in the current analysis into four categories because two of the six ordinal categories had very few responses. The four ordinal categories are as follows:

-**Never:** *"I never do this"*

-**Rarely:** *"I do this but not in the past 30 days" or "1-3 times in the past 30 days"*

-**Regularly:** *"A few days a week, but less than four days a week"*

-**Habitually:** *"5 days a week or 6-7 days a week"*

Table 1 provides the descriptive statistics of the two outcome variables in terms of the shares of each combination of the outcome variables. The sample is skewed toward the low use of bicycling as a travel mode, with about 70% of the individuals reporting that they have never bicycled (see last column of the first numeric row of the table) and only 12.4% reporting regular or habitual use of bicycling (the sum of the percentage entries in the third and fourth numeric rows of the last column of the table). However, among those living in an HDN, only 67.2% report never having bicycled relative to 71.9% not living in an HDN. Also, while only 9.9% of individuals not living in an HDN bicycle regularly/habitually, this percentage rises to 15.7% among those living in an HDN. These statistics suggest a balanced sample in terms of HDN (versus non-HDN) living, a sample skewed toward the low end of bicycling use, and a packaged or bundled choice of residential living and bicycle use.

4.2.1. Latent Constructs

For our empirical demonstration in the context of high density-living and bicycle use, we consider a single latent construct: Environmental consciousness or EC. Environmental consciousness (sometimes, also termed as green-lifestyle propensity) refers to the general concern in individuals regarding their own footprint effect on environmental quality, thereby predisposing these individuals (who have higher levels of EC) to adopt a more sustainable "pro-environmental" lifestyle (such as the frequent use of walking or bicycling as a means of transportation) to reduce negative anthropogenic impacts on the surrounding environment. Indeed, EC is very often used in

transportation or land-use studies in the study of the use and frequency of non-motorized modes of transportation (see for example, Sener et al., 2009, Heinen and Handy, 2012, Verma et al., 2016, Zhu et al., 2020, and Blazanin et al., 2022). But, because EC itself is not directly observed, it is considered as a latent stochastic construct, based on three ordinal indicators from the survey (all collected on a five-point Likert scale from "very unimportant" to "very important"): (a) Importance of being within a reasonably short commute to work, (b) Importance of a walkable neighborhood and being close to local activities when choosing residence location, and (c) Importance of being close to public transit when choosing residence location.

4.3. Model Results

4.3.1. Latent Construct Results

Table 2 provides the latent construct indicator loading results. All the indicator variables are found to be significant with expected directions of loadings (signs) on the latent construct. This forms one component of the measurement equation in the GHDM framework (note that all the components of the GHDM framework are jointly estimated, but are presented in separate sections for presentation ease). Table 3 presents the structural equation results that relate the environmental consciousness (EC) latent construct to observed demographic variables. Our results indicate that there is a significant gender difference in "pro-environmental" attitude; in particular, women are found to be generally more environmentally conscious compared to men. Social-psychological studies attribute this gender difference to the generally more cooperative and interdependent nature of socialization of women (relative to the typically competitive and independent nature of socialization of men), as well as to a more archetypically caring/altruistic behavior exhibited by women relative to men, which then manifests itself in the notion that the environment is a shared asset whose quality has to be preserved for the benefit of all through cooperative efforts and responsibility (see Gifford and Nilsson, 2014 and Desrochers et al., 2019).

Age is another key determinant of EC attitude. In particular, older individuals have lower levels of EC than their younger peers, a result that is again corroborated by earlier studies in environmental sociology (see, for example, Liu et al., 2014 and Clements, 2012). The younger generation has grown up in an era of rapid information technology development, through which to absorb and reflect upon environmental issues and sustainability challenges. This encourages them to adopt environmentally friendly lifestyles (Hassim, 2021). Similarly, the finding from the table that highly educated and employed individuals tend to be more concerned about the environment compared to individuals with lower levels of education and not-employed individuals, respectively, has been found in earlier literature (see, for example, Fisher et al., 2012; Franzen and Vogl, 2013, and Blazanin et al., 2022).

Our results also suggest that individuals with children in their household have a lower EC than other individuals. This finding has support in a recent study by Thomas et al. (2018), who find that, while the legacy hypothesis would suggest a higher EC because parents would be concerned about the legacy they leave for their offspring in regard to environmental quality, the time poverty parents face appears to discourage them from time-consuming sustainable practices.

This then gets manifested in the form of playing down the effects of sustainable practices in the parents' minds as a way of managing dissonance. Further, the child's immediate well-being appears to dominate over concerns over future environmental threats. On the other hand, the results also indicate that individuals from high income households (annual income greater than $100,000) have higher EC levels than those from lower income households. The latter result may be explained based on Maslow's theory of the hierarchy of human needs, which states that humans first focus on the survival-based instinct of meeting their immediate and basic material needs, and consider "higher level" needs such as the need for environmental quality only after the basic needs are satisfied. As income increases, so is the ability to consider the "higher level" needs, consistent with the findings of earlier studies (see for example, Awan and Abbasi, 2013, Sun and Wang, 2020 and Bülbül et al., 2020).

Finally, our results indicate a significant departure from normality for the distribution of the latent construct conditional on the exogenous sociodemographic variable (the transformation parameter in the last row of Table 3 is estimated to be λ =0.466, with a t-statistic of 3.68 for the with respect to the value of 1). This implies a statistically significant right-skew for the EC latent construct error term, suggesting that, conditional on observed characteristics, a higher fraction of individuals are closer to the low end of the EC spectrum; only a small fraction of the respondents are at the high EC range.

4.3.2. Main Outcome Results

Table 4 provides the main outcome results. We discuss, in turn, the latent construct effects, the individual characteristics effects, the household characteristics impacts, and the estimated correlation results. The coefficients in Table 4 represent the impacts of variables on the HDN living propensity latent variable and the bicycling latent propensity variable. The threshold effects toward the bottom of Table 4 do not have any substantive interpretations and simply provide the mechanism to translate the underlying latent propensities to the observed binary/ordinal categories.

Latent Construct Effects

As expected, EC positively impacts the propensity of high-density neighborhood (HDN) living as well as bicycling propensity. Environmentally concerned individuals generally have a higher preference for urban, high-density living because of, among other reasons, the increased opportunities for the use of pollution-free modes of transportation. In addition, the results reveal that the effect of EC on bicycling frequency is particularly elevated for workers relative to non-workers. This may simply be the result of habitual behavior (wherein the elevated bicycling frequency of a worker relative to a non-worker with a comparable EC is because of the regular commute rhythm), and/or may be the result of a conscious choice (wherein the worker realizes that taking up bicycling during the commute is a particularly impactful way of lowering one's travel carbon footprint). In any case, the net result is unobserved non-normal heterogeneity of the worker effect, through the stochasticity of the EC latent construct.

Individual Characteristics Effects

The individual effects in Table 4 represent direct effects after accommodating the moderating effects through the environmental consciousness (EC) latent construct effect. The results suggest that women are less likely to use the bicycle mode, a result that has consistently been reported in earlier studies (Ferdous et al., 2011, Sallis et al., 2013, Ma et al., 2014, Bhat et. al, 2017) and attributed, among other reasons, to the intrinsically heightened safety consciousness among women and the relative difficulty of riding a bicycle with the types of clothing sometimes worn by women. This negative direct effect (-0.295) dominates over the moderating EC effect (0.148×0.142=+0.021 for non-working women, and (0.148+0.114)×0.142=+0.037 for working women). Age is another key determinant of an individual's residential living choices and bicycling propensity. Older adults tend to have a lower inclination to reside in HDN relative to their younger peers, with the youngest of adults (less than 35 years of age) most predisposed toward HDN living, possibly a reflection of the desire for an urban, fast-paced, entertainment-accessible, and physically active lifestyle (De Vos and Alemi, 2020). The results also reveal a generally lower bicycling propensity among senior adults (age 65 years of greater), presumably due to their lower physical dexterity compared to those in the young or middle-aged categories. These direct age effects reinforce the moderating age effects through the EC latent construct. A similar reasoning as for young age may explain the general result that those with a higher education level and students have a higher HDN living propensity, while the higher bicycling propensity among those highly educated (even after conditioning for EC effects) may be associated with higher health consciousness in these population segments (see Jaafar et al., 2017).

Household Characteristics Effects

Among the household effects, individuals from households with three or more adults and with children (of any age) have a lower preference for HDN living relative to individuals from households with fewer than three adults and without children, respectively. These lower HDN living preferences may be attributed to the desire among those with many household members and with children for large living spaces (De Vos and Alemi, 2020), which are typically not in ample supply in HDN neighborhoods. Interestingly, individuals having small children (of age below 5 years) in the household tend to have a lower propensity to bicycle, though the effect gets reversed for individuals with older children in the household. While small children may constrain parents and other adults in the household from partaking in the use of active travel modes (due to time constraints that encourage the use of faster travel modes, as well as the difficulty taking a small child along in a bicycle), older children may promote active modes of travel through joint participation in recreational outdoor activities (see Ferdous et al., 2011 for a similar result).

Finally, household income is also found to influence an individual's residential and bicycling choices. Individuals from lower income households have a generally higher preference for HDN living relative to their higher household income peers, and those from the lowest income

group have a higher propensity to bicycle. These are consistent with earlier studies in the literature (see, for example, Kroesen and Handy, 2014).[6]

Endogenous Outcome Effect

The direct endogenous outcome effect (that is, the "true" causal effect of HDN living on bicycling propensity, as labeled by ϑ in Table 4) suggests that the built environment in high density neighborhoods has a positive causal effect on the propensity to use the bicycle mode. This is quite intuitive as HDNs are often characterized by exclusive bicycle lanes, protected lanes, as well as bicycle parking docks (see Zahabi et al., 2016, Bhat et al., 2017, and Arellana et al., 2020).

Correlations between Main Outcomes and Latent Construct

One of the salient features of the proposed flexible GHDM model (or FLEX-GHDM for short) is that it allows the error terms of the latent constructs to be correlated with the main outcome error terms (that is, the HDN living propensity error term and the bicycling propensity error term). In our model results, the correlation between the EC latent construct error term (in grade form or after transformation to a normally distributed error term) and the underlying HDN living propensity function normal error term is 0.149, while the correlation between the EC latent construct error term (again in grade form) and the bicycling propensity error term is 0.354. These correlations refer to the Spearman correlation in the copula literature (see Bhat and Eluru, 2009). In other words, there are common unobserved factors that simultaneously increase the EC level in individuals and their propensity to reside in an HDN or to bicycle. As discussed in Section 1.2, ignoring this positive EC endogeneity effect will overestimate the EC latent construct effect on the two main outcomes (that is, overestimate the self-selection effect of high-density living based on bicycling propensity), thereby underestimating the "true" HDN effect on bicycling frequency, as further discussed in the next section.

4.4. Estimating "True" Effects of Residential Densification on Bicycling Propensity

The FLEX-GHDM model provides estimates of the joint probability of an individual living in an HDN versus a non-HDN, and monthly bicycling use frequency. For presentation ease, we simplify the notation of Equation (12) and partition it into a component specific to HDN living propensity and another specific to bicycling propensity (underlying the ordinal bicycling frequency variable) as follows (we replace $\mathbf{z}^*$ by the single environmental consciousness latent construct ec^*, and introduce the index q for individuals):

$$\begin{aligned}
\vec{y}^*_{q,HDN}\left|ec^*_q\right. &= \vec{y}^*_{q,HDN}\left|\left[t_\lambda^{-1}\left(g^*_q\right)\right] = \vec{\gamma}_{HDN}\boldsymbol{x}_q + \vec{d}_{HDN}\left[t_\lambda^{-1}\left(g^*_q\right)\right] + \vec{\mu}_{\mathbf{g}^*_q,HDN} + \vec{\varepsilon}_{q,HDN}\right. \\
\vec{y}^*_{q,BF}\left|ec^*_q\right. &= \vec{y}^*_{q,BF}\left|\left[t_\lambda^{-1}\left(g^*_q\right)\right] = \vec{\gamma}_{BF}\boldsymbol{x}_{q,-HDN} + \vartheta y_{q,HDN} + \vec{d}_{BF}\left[t_\lambda^{-1}\left(g^*_q\right)\right] + \vec{\mu}_{\mathbf{g}^*_q,BF} + \vec{\varepsilon}_{q,BF}\right..
\end{aligned} \tag{19}$$

[6] Higher levels of car-ownership have also been consistently negatively associated with HDN living in the transportation and land-use literature (for example, see Shen et al., 2016, Ding et al., 2017, Zhang et al., 2014, De Vos and Alemi, 2020). However, we do not use household vehicle ownership as an exogenous variable in the model, because it is likely to be endogenous to HDN living and bicycle ownership decisions (see, for example, Pinjari et al. 2011).

For compatibility with earlier notation after introducing the index q for individuals, $\vec{\boldsymbol{\mu}}_{\mathbf{g}_q^*} = (\vec{\mu}_{\mathbf{g}_q^*,HDN}, \vec{\mu}_{\mathbf{g}_q^*,BF})'$ and $(\vec{\varepsilon}_{q,HDN}, \vec{\varepsilon}_{q,BF}) \sim BVN(\mathbf{0}_2, \boldsymbol{\Theta})$, where BVN is the bivariate normal distribution, $\mathbf{0}_2$ is a zero vector of length 2, and $\boldsymbol{\Theta}$ is a covariance matrix as defined earlier. $\boldsymbol{x}_{q,-HDN}$ is a vector of variables that does not include the high density living variable. For presentation ease, define $(\upsilon_{q,HDN} \mid g_q^*) = \vec{\boldsymbol{\gamma}}_{HDN}\boldsymbol{x}_q + \vec{d}_{HDN}\left[t_\lambda^{-1}\left(g_q^*\right)\right] + \vec{\mu}_{\mathbf{g}_q^*,HDN}$. Also, define the following:

$$\begin{aligned}&(\upsilon_{q,BF,y_{HDN}=0} \mid g_q^*) = \vec{\boldsymbol{\gamma}}_{BF}\boldsymbol{x}_{q,-HDN} + \vec{d}_{BF}\left[t_\lambda^{-1}\left(g_q^*\right)\right] + \vec{\mu}_{\mathbf{g}_q^*,BF}, \text{ and}\\ &(\upsilon_{q,BF,y_{HDN}=1} \mid g_q^*) = (\upsilon_{q,BF,y_{HDN}=0} \mid g_q^*) + \vartheta.\end{aligned} \tag{20}$$

With the notations above, we may write the equation system in (19) as:

$$\begin{aligned}&\vec{y}_{q,HDN}^* \,\big|\, g_q^* = (\upsilon_{q,HDN} \mid g_q^*) + \vec{\varepsilon}_{q,HDN},\\ &\vec{y}_{q,BF,y_{HDN}=0}^* \,\big|\, g_q^* = (\upsilon_{q,BF,y_{HDN}=0} \mid g_q^*) + \vec{\varepsilon}_{q,BF} \text{ observed if } \vec{y}_{q,HDN}^* < 0\\ &\vec{y}_{q,BF,y_{HDN}=1}^* \,\big|\, g_q^* = (\upsilon_{q,BF,y_{HDN}=0} \mid g_q^*) + \vartheta + \vec{\varepsilon}_{q,BF} \text{ observed if } \vec{y}_{q,HDN}^* > 0.\end{aligned} \tag{21}$$

There are three effects associated with HDN living propensity on bicycling propensity in the above equation system: (1) the *"true" causal effect* of HDN living on bicycling propensity, as captured by the ϑ shifter term for bicycling propensity based on HDN living versus non-HDN living, (2) a *sample selection effect* due to the unobserved correlation engendered between the HDN living and the bicycling frequency equations because of the conditioning of both these equations on the common g_q^* variable reflecting the transformed version of the EC stochastic latent construct (which then has to be unconditioned out in estimation as discussed in Section 3.4) and (3) the *EC endogeneity effect* that, based on Equation (11), leads to $\vec{\mu}_{\mathbf{g}_q^*,HDN} \neq 0$, $\vec{\mu}_{\mathbf{g}_q^*,BF} \neq 0$, and a covariation between the $\vec{\varepsilon}_{q,HDN}$ and $\vec{\varepsilon}_{q,BF}$ error terms $(\vec{\varepsilon}_{q,HDN}, \vec{\varepsilon}_{q,BF}) \sim BVN(\mathbf{0}_2, \boldsymbol{\Theta})$ (all generated by the non-zero correlation matrix $\boldsymbol{\Omega}_{G^*\vec{\varepsilon}}$ that contains the correlations between (a) the HDN living error and the EC construct error, and (b) the bicycling propensity error with the EC construct error). In the independent heteroscedastic data model (IHDM), which does not consider any jointness between the HDN living and bicycling frequency propensities due to unobserved factors (that is, the stochastic EC latent construct is not considered at all), the second and third effects above are completely ignored, and effectively get added on to the "true" causal effect. The net effect is an overestimation in the IHDM of the "true" causal effect of HDN living on bicycling propensity. In effect, then, the ϑ estimate from the IHDM model may be used as the overall cumulative of all three of the "true causal", the "sample selection", and the "EC endogeneity effect" of HDN living on bicycling propensity. In the traditional GHDM, the first and second effects above are considered, but the third "EC endogeneity effect" is ignored, which then gets added on to the self-selection effect and reduces the "true" HDN living effect on bicycling frequency, as discussed in Section 1.2 and Figure 1. We also estimate two additional models for comparison purposes. The

first, which we label as the not-dependent GHDM (or ND-GHDM), ignores the EC endogeneity effect, while retaining a non-normal distribution for the EC construct (that is, $\mathbf{\Omega}_{G^*\tilde{\varepsilon}}$ is assumed to be a zero matrix). The second additional model, which we label as the normal GHDM (or N-GHDM), considers the EC endogeneity effect, but considers a symmetric normal distribution for the EC construct. This is the same as our FLEX-GHDM, and considers all the three effects of "true causal", "sample selection", and "EC endogeneity" effects, but holds the EC construct to have a normal distribution.

To associate actual statistics to the three HDN effects on bicycling frequency, from equation system (19), and following the same approach as in Section 3.4, one can estimate, for each individual, the following four bivariate probabilities: $P(y_{q,HDN}=1, y_{q,BF,y_{HDN}=1}=j), P(y_{q,HDN}=1, y_{q,BF,y_{HDN}=0}=j), P(y_{q,HDN}=0, y_{q,BF,y_{HDN}=0}=j)$, and $P(y_{q,HDN}=0, y_{q,BF,y_{HDN}=1}=j), j=1,2,3,4.$

One can also readily compute the following univariate probabilities for bicycling frequency level: $P(y_{q,BF,y_{HDN}=0}=j)$ and $P(y_{q,BF,y_{HDN}=1}=j)$. For easy interpretability and presentation, we convert the ordinal bicycle use frequency categories into cardinal values c_j as follows: j=1 (never use = 0 instances per month); j=2 (rarely use = 2 instances per month), j=3 (regularly use = 12 instances per month), and j=4 (habitually use = 24 instances per month). With these assignments, one may compute two quantities as discussed in Bhat and Eluru (2009). The first is the average treatment effect (ATE) of residential living on bicycling frequency, which refers to the average (expected) increase in bicycling frequency for a random household from the population if it were to reside in an HDN as opposed to a non-HDN. This ATE constitutes the "true" effect of residential densification, and is given by:

$$ATE = \frac{1}{Q}\left[\sum_{j=1}^{4} c_j \left(P(y_{q,BF,y_{HDN}=1}=j) - P(y_{q,BF,y_{HDN}=0}=j)\right)\right]. \quad (22)$$

The second is the "Effect of Treatment on the Treated and Non-treated (TTNT)", which is the total of all the three effects of HDN living on bicycling frequency. This TTNT effect represents the average impact of treatment on the (currently) treated (that is, individuals currently residing in HDNs) *and* (currently) non-treated (TTNT) (that is, individuals currently residing in non-HDNs). In the current empirical context, it is the expected bicycling frequency change for a randomly picked individual who is relocated from the current residential neighborhood type of living to the other neighborhood type, measured in the *common direction* of change from a non-HDN to a HDN. The TTNT measure, in effect, provides the average expected change in bicycling frequency if all households were relocated to a HDN assuming that all households *currently* choose to locate in a non-HDN. It is given by:

$$TNTT = \frac{1}{Q}\left[\begin{array}{l} 1(y_{q,HDN}=1)\times\sum_{j=1}^{4} c_j\left(\frac{P(y_{q,HDN}=1, y_{q,BF,y_{HDN}=1}=j)}{P(y_{q,HDN}=1)} - \frac{P(y_{q,HDN}=1, y_{q,BF,y_{HDN}=0}=j)}{P(y_{q,HDN}=1)}\right)+ \\ 1(y_{q,HDN}=0)\times\sum_{j=1}^{4} c_j\left(\frac{P(y_{q,HDN}=0, y_{q,BF,y_{HDN}=1}=j)}{P(y_{q,HDN}=0)} - \frac{P(y_{q,HDN}=0, y_{q,BF,y_{HDN}=0}=j)}{P(y_{q,HDN}=0)}\right) \end{array}\right]. \quad (23)$$

The closer TTNT is to ATE, the lesser are the self-selection and EC endogeneity effects. Of course, in the limit that there is no self-selection and no EC endogeneity, TTNT collapses to the ATE, as the bivariate probabilities in the TNTT equation collapse to the product of univariate probabilities. This is the case for the IHDM model, and so the ATE for the IHDM is used as the common base as the TNTT for all the models for comparison of the substantive results across the many different models. The ATE as a percentage of TTNT provides the percentage "true" HDN densification effect. In the traditional GHDM, the difference between the ATE from the IHDM (=TNTT from the IHDM) and the ATE from the traditional GHDM provides the consolidated effect of self-selection and EC endogeneity (with the assumption of zero EC endogeneity effects). A similar situation holds for the ND-GHDM. For the FLEX-GHDM, the ATE is computed again as in Equation (22) based on the FLEX-GHDM estimates. To partition the remainder HDN (say, the RHDN) effect into a sample selection effect and an EC endogeneity effect, we compute the TNTT of the FLEX-GHDM and of the ND-GHDM. The difference of the TNTT between the ND-GHDM and the FLEX-GHDM (the subtraction applied in that order) as a proportion of the ND-GHDM TNTT provides the sample selection proportion of HDN in the Flex-GHDM model. Finally, to obtain a similar partitioning for the N-GHDM model, we take the difference between the TNTT values of the traditional GHDM and the N-GHDM, and designate this amount as a proportion of the traditional GHDM as the sample selection effect.

Table 5 presents the estimated "true effect" (ATE), the sample selection effect, and the EC endogeneity effect, both in absolute terms as well as a percentage. In doing so, for the IHDM model, we introduce the exogenous variables (sociodemographic variables) used to explain the latent constructs as exogenous variables in the choice dimension equations. This way, the contribution to the observed part of the utility due to sociodemographic variables is still maintained. The first row under the "IHDM model" heading indicates that a random individual shifted from a non-HDN to an HDN is, on an average, likely to increase monthly bicycling frequency by 0.321 instances (this represents about a 14% increase relative to the current sample average of 2.33 monthly bicycling frequency). Equivalently, if 100 random individuals are relocated from a non-HDN to an HDN, the point estimate from the IHDM indicates an increase in monthly bicycling frequency by about 32 instances. Because the IHDM completely ignores any self-selection effects, the ATE is the same as the TTNT, as mentioned earlier. However, the ATE for all the four GHDM versions are lower than the ATE for the IHDM, because all these GHDM models do acknowledge that some of the HDN effect as manifested in the ATE of the IHDM is not causal, but due to spurious self-selection and/or EC endogeneity effects. Across the four GHDM models, the estimated "true" HDN effect is lowest (and the self-selection effect is highest) in the case of the traditional-GHDM (TR-GHDM), while the estimated "true" HDN effect is

highest (and the self-selection effect is lowest) for the proposed flexible-GHDM (FLEX-GHDM). As explained in Section 1.2, the traditional GHDM ignores EC endogeneity, which then overestimates the EC effect on the two main outcomes and therefore the self-selection effect. This leads to an underestimation (quite substantially in our empirical context) of the "true" HDN living effect on the bicycling frequency (as evident from the first two numeric rows of Table 5). In fact, the "true" causal effect, as estimated by our proposed FLEX-GHDM model is almost twice that estimated by the TR-GHDM model, with the self-selection effect contribution dropping from 59.2% in the TR-GHDM model to 6.5% in our proposed model). Our results from Table 5 also suggest that, in our empirical context, the difference between the TR-GHDM and ND-GHDM, and between the N-GHDM and FLEX-GHDM, is rather marginal, suggesting that non-normality or normality of the EC construct does not play much of a role in estimating "true" HDN effects on bicycling frequency. Of course, this is specific to our empirical context, and can vary in other contexts. Besides, the predictions from a data fit stand point are still better for the non-normal EC models than the normal EC models, as discussed in the next section. Finally, it should be noted that the only way to test whether normal distributions suffice in an empirical context is by estimating a GHDM model allowing for non-normal latent constructs, and then testing with normally distributed latent constructs.

Overall, our results strongly emphasize the importance of allowing an unobserved dependency structure between attitudinal constructs and dependent outcomes in cross-sectional analyses of residential living and travel behavior, and demonstrate the pitfalls of using the traditional GHDM for such multivariate data analysis. In fact, as should be clear from the results, our analysis suggests that earlier cross-sectional studies that may have attempted to control for residential self-selection effects without considering attitude and lifestyle endogeneity may be even worse than simpler models that completely ignore self-selection effects (note that the IHDM ATE is closer to the FLEX-GHDM than is the TR-GHDM ATE in Table 5). Importantly, ignoring the package nature of the attitude-built environment-travel behavior (A-BE-TB) connection when investigating built environment effects on travel behavior can lead to mis-informed policies and mis-directed infrastructure investment decisions. In our case, the results suggest an underestimation of the "true" HDN living effect on bicycling propensity if EC endogeneity is ignored (such as in the traditional GHDM model). This may erroneously lead policy-makers to refrain from implementing neo-urbanist investments, or considering unnecessarily high investments in bicycle-friendly infrastructure to achieve a set goal of bicycle use. On the other hand, our results also suggest an overestimation of the "true" HDN living effect on bicycling propensity if both sample selection and EC endogeneity is ignored (as in the IHDM model), which can erroneously provide a much more "rosier" picture of the benefits accruing from densification to promote bicycling. Additionally, our results suggest an overestimation of EC effects on both HDN living and bicycling propensity if the package nature of the A-BE-TB connection is ignored. For example, in the case of bicycling frequency, the TR-GHDM indicates that the lower environmental consciousness of older adults (>65 years) contributes about 31% to the lower bicycling propensity of such individuals, while the proposed FLEX-GHDM indicates that this

lower environmental consciousness in older adults contributes only 20%. That is, our results suggest that campaigns to increase bicycling frequency among older adults through raising environmental consciousness is likely to have a much more limited effect than would be estimated by traditional models.

4.5. Data Fit Comparison

As discussed in the previous section, the substantive implications from the IHDM and the many GHDM models are quite different in regard to the effect of HDN living. While not presented here to conserve on space, the magnitude of the effects of other exogenous variables also vary across the models. In this section, to determine which of the models provides the best data fit, we present a comparative analysis of the five models; the IHDM, the TR-GHDM, the ND-GHDM, the N-GHDM, and the FLEX-GHDM. This analysis is undertaken using both likelihood and non-likelihood based measures. The likelihood-based measures are based on predictive likelihood at convergence for the two main outcome variables (since our primary goal is to understand the bicycling frequency and residential choice decisions, we focus on the predictive capability of our models for only these two main outcomes). The models may then be compared using a predictive Bayesian Information Criterion (BIC) statistic [$= -\mathcal{L}(\hat{\theta}) + 0.5$ (# of model parameters) log (sample size)] ($\mathcal{L}(\hat{\theta})$ is the predictive log-likelihood at convergence). The model with a lower BIC statistic is the preferred model. In addition, the adjusted likelihood ratio index of each of the four models is computed as follows with respect to the log-likelihood with only the constants in the two outcomes:

$$\bar{\rho}^2 = 1 - \frac{L(\hat{\boldsymbol{\theta}}) - M}{L(c)} \tag{24}$$

where $L(\hat{\boldsymbol{\theta}})$ and $L(c)$ are the predictive log-likelihood functions at convergence and at constants, respectively, and M is the number of parameters (excluding the constants) estimated in the model.

We also evaluate the data fit of the models intuitively and informally at both the disaggregate and aggregate levels. At the disaggregate level, we first compute the multivariate predictions for each of the two outcomes (this entails a total of $4 \times 2 = 8$ combinations). Then, we compute an average (across individuals) probability of correct prediction at this full combination level. The model with the highest average probability of correct prediction is preferred. At the aggregate level, we design a heuristic diagnostic check of model fit by computing the predicted aggregate number of individuals in each of the 8 combinations for all the GHDM frameworks. These are then compared with the actual shares and the absolute percentage error (APE) statistic is computed.

The results of the disaggregate data fit evaluations are provided in Table 6. The predictive log-likelihood at convergence, the BIC values, predictive adjusted likelihood ratio indices, and the average probability of correct prediction from the models indicate the better fit of the proposed FLEX-GHDM relative to all the other models. In terms of aggregate data fit too (see Table 7), the APE from the FLEX-GHDM framework is lower for almost all the combinations than those from

the restricted GHDM versions. As importantly, even though the "true" causal effect does not vary much between the N-GHDM and FLEX-GHDM models, the FLEX-GHDM model shows better data fit relative to the N-GHDM model, with just the addition of a single parameter. Overall, the weighted average (weighted by the share of each outcome combination) of the absolute percentage error (weighted MAPE) is the lowest for the FLEX-GHDM model and the highest for the IHDM model.

5. CONCLUSION

There is growing interest in multivariate dependent outcome models that include a mixture of different kinds of discrete and continuous variables. This may be attributed to at least two reasons. The first is the ability to generate multivariate distributions through the use of relatively flexible copula-based methods and/or the use of effective factorization techniques for the covariance matrices that reduce the number of covariance parameters to be estimated. The second is the development of computationally efficient ways to estimate models based on variational methods for Bayesian inference or maximum approximate composite marginal likelihood methods for frequentist inference However, there are two important assumptions in most earlier mixed data models, including the GHDM framework proposed by Bhat (2015): (i) marginal normality of unobserved factors that generate jointness, and (ii) independence between the unobserved factors and the propensity equations underlying the main outcomes of interest.

In the current paper, we simultaneously relax both these assumptions of earlier mixed data modeling efforts and develop a flexible GHDM model for mixed data modeling. We then propose a hybrid MSL-MACML inference approach for estimation. To our knowledge, this is the first study to propose such a flexible methodological structure for multiple mixed outcomes modeling. We demonstrate an application of our proposed model in the context of individuals' high-density residential neighborhood living choice and monthly bicycling frequency. The sample used is derived from the 2019 Puget Sound Regional Council Household Travel Study that collected socio-demographic, residential, and activity-travel data, as well as elicited information on attitudes/preferences through Likert scale indicator questions. Environmental consciousness is used as a non-normally distributed latent construct (that is, an unobserved stochastic factor), and is considered endogenous to the main ordinal outcomes of individuals' residential neighborhood living choice and bicycling frequency.

Overall, our study proposes a new methodological approach to introduce non-normality in mixed data models, while also recognizing factor endogeneity effects. Our empirical results demonstrate the importance of considering both these methodological issues, with normality being strongly rejected (based on the λ parameter being statistically significantly different from one), and clear evidence of statistically significant endogeneity effects of the latent constructs. Ignoring these effects, in our specific empirical context, leads to an underestimation of the true impact of high-density neighborhood living on bicycling frequency.

ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Human Resource Development (MHRD) of the Government of India through its Scheme for Promotion of Academic and Research Collaboration (SPARC) program. The authors are grateful to Lisa Macias for her help in formatting this document and also appreciate the anonymous reviewer comments on an earlier paper version.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

Arellana, J., Saltarín, M., Larrañaga, A., González, V. and Henao, C. (2020). Developing an urban bikeability index for different types of cyclists as a tool to prioritise bicycle infrastructure investments. *Transportation Research Part A*, *139*, 310-334.

Asmussen, K. E., Mondal, A., and Bhat, C. R. (2022). Adoption of partially automated vehicle technology features and impacts on vehicle miles of travel (VMT). *Transportation Research Part A*, *158*, 156-179.

Awan, U., and Abbasi, A. S. (2013). Environmental sustainability through determinism the level of environmental awareness, knowledge and behavior among business graduates. *Research Journal of Environmental and Earth Science*, *5*(9), 505-515.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C. R., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., de Palma, A., Gopinath, D., Karlstrom, A., and Muniziga, M. A. (2002). Hybrid choice models: Progress and challenges. *Marketing Letters*, *13*(3), 163-175.

Bhat, C. R. (1996). A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity. *Transportation Research Part B*, *30*(3), 189-207.

Bhat, C. R. (1998). Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transportation Research Part A*, *32*(7), 495-507.

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. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, *79*, 50-77.

Bhat, C. R. (2018). New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. *Transportation Research Part B*, *109*, 238-256.

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 Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, *43*(7), 749-765.

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 Lavieri, P. S. (2018). A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions. *Theory and Decision*, *84*(2), 239-275.

Bhat, C. R., and Mondal, A. (2021). On the almost exact-equivalence of the radial and spherical unconstrained Cholesky-based parameterization methods for correlation matrices. Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.

Bhat, C. R., and Sidharthan, R. (2012). A new approach to specify and estimate non-normally mixed multinomial probit models. *Transportation Research Part B*, *46*(7), 817-833.

Bhat, C. R., Sener, I. N., Eluru, N., (2010). A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B*, *44*(8-9), 903-921.

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.

Bhat, C. R., Born, K., Sidharthan, R., and Bhat, P. C. (2014). A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research*, *1*, 53-71.

Bhat, C. R., Dubey, S. K., and Nagel, K. (2015). Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice. *Transportation Research Part B*, *78*, 341-363.

Bhat, C. R., Astroza, S., Bhat, A. C., and Nagel, K. (2016a). Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. *Transportation Research Part B*, *91*, 52-76.

Bhat, C. R., Pinjari, A. R., Dubey, S. K., and Hamdi, A. S. (2016b). On accommodating spatial interactions in a generalized heterogeneous data model (GHDM) of mixed types of dependent variables. *Transportation Research Part B*, *94*, 240-263.

Bhat, C. R., Astroza, S., and Hamdi, A. S. (2017). A spatial generalized ordered-response model with skew normal kernel error terms with an application to bicycling frequency. *Transportation Research Part B*, *95*, 126-148.

Blazanin, G., Mondal, A., Asmussen, K. E., and Bhat, C. R. (2022). E-scooter sharing and bikesharing systems: An individual-level analysis of factors affecting first-use and use frequency. *Transportation Research Part C*, *135*, 103515.

Bolduc, D., Ben-Akiva, M., Walker, J., and 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.

Bülbül, H., Büyükkeklik, A., Topal, A., and Özoğlu, B. (2020). The relationship between environmental awareness, environmental behaviors, and carbon footprint in Turkish households. *Environmental Science and Pollution Research*, *27*(20), 25009-25028.

Cao, X., and Fan, Y. (2012). Exploring the influences of density on travel behavior using propensity score matching. *Environment and Planning B*, *39*(3), 459-470.

Chen, C., Gong, H., and Paaswell, R. (2008). Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation*, *35*(3), 285-299.

Chin, W. C., Lee, M. C., and Yap, G. L. C. (2016). Modelling financial market volatility using asymmetric-skewed-ARFIMAX and-HARX models. *Engineering Economics*, *27*(4), 373-381.

Clements, B. (2012). Exploring public opinion on the issue of climate change in Britain. *British Politics*, *7*(2), 183-202.

Dannemiller, K. A., Mondal, A., Asmussen, K. E., and Bhat, C. R. (2021). Investigating autonomous vehicle impacts on individual activity-travel behavior. *Transportation Research Part A*, *148*, 402-422.

de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics and Probability Letters*, *75*(1), 49-57.

de Leon, A. R., and Carriégre, K. C. (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: An overview. In de Leon, A.R., and Chough, K.C. (Eds), *Analysis of Mixed Data: Methods & Applications*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

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

De Vos, J., and Alemi, F. (2020). Are young adults car-loving urbanites? Comparing young and older adults' residential location choice, travel behavior and attitudes. *Transportation Research Part A*, *132*, 986-998.

De Vos, J., and Singleton, P. A. (2020). Travel and cognitive dissonance. *Transportation Research Part A*, *138*, 525-536.

De Vos, J., Ettema, D., and Witlox, F. (2018). Changing travel behaviour and attitudes following a residential relocation. *Journal of Transport Geography*, *73*, 131-147.

Deng, Y., and Zhao, P. (2022). Quantifying residential self-selection effects on commuting mode choice: A natural experiment. *Transportation Research Part D*, *104*, 103197.

Desrochers, J. E., Albert, G., Milfont, T. L., Kelly, B., and Arnocky, S. (2019). Does personality mediate the relationship between sex and environmentalism? *Personality and Individual Differences*, *147*, 204-213.

Dias, F. F., Lavieri, P. S., Sharda, S., Khoeini, S., Bhat, C. R., Pendyala, R. M., Pinjari, A. R., Ramadurai, G., and Srinivasan, K. K. (2020). A comparison of online and in-person activity engagement: The case of shopping and eating meals. *Transportation Research Part C*, *114*, 643-656.

Ding, C., Wang, D., Liu, C., Zhang, Y., and Yang, J. (2017). Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. *Transportation Research Part A*, *100*, 65-80.

Falk, M., and Katz-Gerro, T. (2017). Modeling travel decisions: Urban exploration, cultural immersion, or both?. *Journal of Travel and Tourism Marketing*, *34*(3), 369-382.

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

Ferdous, N., Pendyala, R.M., Bhat, C.R., and Konduri, K.C. (2011). Modeling the influence of family, social context, and spatial proximity on use of non-motorized transport mode. *Transportation Research Record: Journal of the Transportation Research* Board, *2230*, 111-120.

Festinger, L. (1957). *A Theory of Cognitive Dissonance* . Stanford University Press, Stanford, CA.

Fisher, C., Bashyal, S., and Bachman, B. (2012). Demographic impacts on environmentally friendly purchase behaviors. *Journal of Targeting, Measurement and Analysis for Marketing*, *20*(3), 172-184.

Franzen, A. and Meyer, R., (2010). Environmental attitudes in cross-national perspective: a multilevel analysis of the ISSP 1993 and 2000. *European Sociological Review*, *26* (2), 19-34.

Franzen, A., and Vogl, D. (2013). Two decades of measuring environmental attitudes: A comparative analysis of 33 countries. *Global Environmental Change*, *23*(5), 1001-1008.

Gallo, M. (2020). Assessing the equality of external benefits in public transport investments: The impact of urban railways on real estate values. *Case Studies on Transport Policy*, *8*(3), 758-769.

Gifford, R., and Nilsson, A. (2014). Personal and social factors that influence pro-environmental concern and behaviour: A review. *International Journal of Psychology*, *49*(3), 141-157.

Goerg, G. M. (2015). The lambert way to gaussianize heavy-tailed data with the inverse of Tukey's h transformation as a special case. *The Scientific World Journal*, *2015*, 909231.

Guan, X., and Wang, D. (2019). Residential self-selection in the built environment-travel behavior connection: Whose self-selection?. *Transportation Research Part D*, *67*, 16-32.

Hamed, M. M., and Mannering, F. L. (1993). Modeling travelers' postwork activity involvement: toward a new methodology. *Transportation Science*, *27*(4), 381-394.

Hassim, A. (2021). Why younger generations are more willing to change in the name of sustainability. GreenBiz.com. Available at: https://www.greenbiz.com/article/why-younger-generations-are-more-willing-change-name-sustainability

Heinen, E., and Handy, S. (2012). Similarities in attitudes and norms and the effect on bicycle commuting: Evidence from the bicycle cities Davis and Delft. *International Journal Of Sustainable Transportation*, *6*(5), 257-281.

Heydari, S., Miranda-Moreno, L., and Hickford, A. J. (2020). On the causal effect of proximity to school on pedestrian safety at signalized intersections: A heterogeneous endogenous econometric model. *Analytic Methods in Accident Research*, *26*, 100115.

Hoshino, T., and Bentler, P.M. (2013). Bias in factor score regression and a simple solution. In: De Leon, A.R., and Chough, K.C. (eds.), *Analysis of Mixed Data: Methods and Applications*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 43-61.

Jaafar, N., Ainin, S., and Yeong, M. (2017). Why bother about health? A study on the factors that influence health information seeking behaviour among Malaysian healthcare consumers. *International Journal of Medical Informatics*, *104*, 38-44.

Jarass, J., and Scheiner, J. (2018). Residential self-selection and travel mode use in a new inner-city development neighbourhood in Berlin. *Journal of Transport Geography*, *70*, 68-77.

Jiryaie, F., and Khodadadi, A. (2019). Simultaneous optimization of multiple responses that involve correlated continuous and ordinal responses according to the Gaussian copula models. *Journal of Statistical Theory and Applications*, *18*(3), 212-221.

Jiryaie, F., Withanage, N., Wu, B., and De Leon, A. R. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation*, *86*(9), 1643-1659.

Kang, S., Mondal, A., Bhat, A. C., and Bhat, C. R. (2021). Pooled versus private ride-hailing: A joint revealed and stated preference analysis recognizing psycho-social factors. *Transportation Research Part C*, *124*, 102906.

Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business and 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.

Kroesen, M. (2019). Residential self-selection and the reverse causation hypothesis: Assessing the endogeneity of stated reasons for residential choice. *Travel Behaviour and Society*, *16*, 108-117.

Kroesen, M., and Handy, S. (2014). The relation between bicycle commuting and non-work cycling: results from a mobility panel. *Transportation*, *41*(3), 507-527.

Kwon, S., Ha, I. D., Shih, J. H., and Emura, T. (2022). Flexible parametric copula modeling approaches for clustered survival data. *Pharmaceutical Statistics*, *21*(1), 69-88.

Lee, L. F. (1983). Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society*, 507-512.

Leung, K. Y., Astroza, S., Loo, B. P., and Bhat, C. R. (2019). An environment-people interactions framework for analysing children's extra-curricular activities and active transport. *Journal of Transport Geography*, *74*, 341-358.

Liu, X., Vedlitz, A., and Shi, L. (2014). Examining the determinants of public environmental concern: Evidence from national public surveys. *Environmental Science and Policy*, *39*, 77-94.

Loaiza-Maya, R., and Smith, M. S. (2019). Variational Bayes estimation of discrete-margined copula models with application to time series. *Journal of Computational and Graphical Statistics*, *28*(3), 523-539.

Lu, Y., Chen, L., Yang, Y., and Gou, Z. (2018). The association of built environment and physical activity in older adults: Using a citywide public housing scheme to reduce residential self-selection bias. *International Journal of Environmental Research and Public Health*, *15*(9), 1973.

Ma, L., Dill, J., Mohr, C. (2014). The objective versus the perceived environment: What matters for bicycling? *Transportation*, *41*(6), 1135-1152.

Manaugh, K., Boisjoly, G., and El-Geneidy, A. (2017). Overcoming barriers to cycling: Understanding frequency of cycling in a University setting and the factors preventing commuters from cycling on a regular basis. *Transportation*, *44*(4), 871-884.

Mannering, F. L., and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1-22.

Millard-Ball, A., West, J., Rezaei, N., and Desai, G. (2022). What do residential lotteries show us about transportation choices? *Urban Studies*, *59*(2), 434-452.

Moore, M. A., Lavieri, P. S., Dias, F. F., and Bhat, C. R. (2020). On investigating the potential effects of private autonomous vehicle use on home/work relocations and commute times. *Transportation Research Part C*, *110*, 166-185.

Müller, D., and Czado, C. (2018). Representing sparse Gaussian DAGs as sparse R-vines allowing for non-Gaussian dependence. *Journal of Computational and Graphical Statistics*, *27*(2), 334-344.

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.

O'Brien, R.M. (1994). Identification of simple measurement models with multiple latent variables and correlated errors. *Sociological Methodology*, *24*, 137-170.

Oh, D. H., and Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business and Economic Statistics*, *35*(1), 139-154.

Ong, V. M. H., Nott, D. J., and Smith, M. S. (2018). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, *27*(3), 465-478.

Paleti, R., Bhat, C. R. (2013). The composite marginal likelihood (CML) estimation of panel ordered-response models. *Journal of Choice Modelling*, *7*, 24-43.

Paleti, R., Bhat, C.R., Pendyala, R. (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.

Park, Y., and Akar, G. (2019). Understanding the effects of individual attitudes, perceptions, and residential neighborhood types on university commuters' bicycling decisions. *Journal of Transport and Land Use*, *12*(1), 419-441.

Patil, P. N., Dubey, S. K., Pinjari, A. R., Cherchi, E., Daziano, R., and Bhat, C. R. (2017). Simulation evaluation of emerging estimation techniques for multinomial probit models. *Journal of Choice Modelling*, *23*, 9-20.

Pinjari, A. R., Bhat, C. R., and Hensher, D. A. (2009). Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B*, *43*(7), 729-748.

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*, 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.

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

Sallis, J.F., Conway, T.L., Dillon, L.I., Frank, L.D., Adams, M.A., Cain, K.L., Saelens, B.E. (2013). Environmental and demographic correlates of bicycling. *Preventive Medicin*e, *57*(5), 456-460.

Sener, I. N., Eluru, N., and Bhat, C. R. (2009). Who are bicyclists? Why and how much are they bicycling? *Transportation Research Record: Journal of the Transportation Research Board*, *2134*(1), 63-72.

Shen, Q., Chen, P., and Pan, H. (2016). Factors affecting car ownership and mode choice in rail transit-supported suburbs of a large Chinese city. *Transportation Research Part A*, *94*, 31-44.

Smith, M. S., Loaiza-Maya, R., and Nott, D. J. (2020). High-dimensional copula variational approximation through transformation. *Journal of Computational and Graphical Statistics*, *29*(4), 729-743.

Spissu, E., Pinjari, A. R., Pendyala, R. M., and Bhat, C. R. (2009). A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel. *Transportation*, *36*(4), 403-422.

Stapleton, D. C. (1978). Analyzing political participation data with a MIMIC Model. *Sociological Methodology*, *9*, 52-74.

Sun, Y., and Wang, S. (2020). Understanding consumers' intentions to purchase green products in the social media marketing context. *Asia Pacific Journal of Marketing and Logistics*, *32*(4), 860-878

Teixeira-Pinto, A., and Harezlak, J. (2013). Factorization and latent variable models for joint analysis of binary and continuous outcomes. In: De Leon, A.R., and Chough, K.C. (eds.), *Analysis of Mixed Data: Methods and Applications*, CRC Press, Taylor and Francis Group, Boca Raton, FL, 81-91.

Thomas, G. O., Fisher, R., Whitmarsh, L., Milfont, T. L., and Poortinga, W. (2018). The impact of parenthood on environmental attitudes and behaviour: A longitudinal investigation of the legacy hypothesis. *Population and Environment*, *39*(3), 261-276.

van de Coevering, P., Maat, K., and van Wee, B. (2018). Residential self-selection, reverse causality and residential dissonance. A latent class transition model of interactions between the built environment, travel attitudes and travel behavior. *Transportation Research Part A*, *118*, 466-479.

van de Coevering, P., Maat, K., Kroesen, M., and van Wee, B. (2016). Causal effects of built environment characteristics on travel behaviour: A longitudinal approach. *European Journal of Transport and Infrastructure Research*, *16*(4), 674-697.

Verma, M., Rahul, T., Reddy, P., and Verma, A. (2016). The factors influencing bicycling in the Bangalore city. *Transportation Research Part A*, *89*, 29-40.

Vij, A., and Walker, J. L. (2014). Hybrid choice models: The identification problem. In *Handbook of Choice Modelling*. Edward Elgar Publishing.

Vinayak, P., Dias, F. F., Astroza, S., Bhat, C. R., Pendyala, R. M., and Garikapati, V. M. (2018). Accounting for multi-dimensional dependencies among decision-makers within a generalized model framework: An application to understanding shared mobility service usage levels. *Transport Policy*, *72*, 129-137.

Wang, D. and Lin, T. (2019). Built environment, travel behavior, and residential self-selection: A study based on panel data from Beijing, China. *Transportation*, *46*(1), 51-74.

Wei, Y., Tang, Y., and McNicholas, P. D. (2019). Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. *Computational Statistics and Data Analysis*, *130*, 18-41.

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41-78.

Wolday, F., Næss, P., and Cao, X. J. (2019). Travel-based residential self-selection: A qualitatively improved understanding from Norway. *Cities*, *87*, 87-102.

Wu, B., de Leon, A. R., and Withanage, N. (2013). Joint analysis of mixed discrete and continuous outcomes via copula models. *Analysis of Mixed Data: Methods and Applications*, 139-156.

Yeo, I. K., and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954-959.

Zahabi, S., Chang, A., Miranda-Moreno, L., and Patterson, Z. (2016). Exploring the link between the neighborhood typologies, bicycle infrastructure and commuting cycling over time and the potential impact on commuter GHG emissions. *Transportation Research Part D*, *47*, 89-103.

Zang, P., Lu, Y., Ma, J., Xie, B., Wang, R., and Liu, Y. (2019). Disentangling residential self-selection from impacts of built environment characteristics on travel behaviors for older adults. *Social Science and Medicine*, *238*, 112515.

Zhang, J., Yu, B., and Chikaraishi, M. (2014). Interdependences between household residential and car ownership behavior: A life history analysis. *Journal of Transport Geography*, *34*, 165-174.

Zhu, M., Hu, X., Lin, Z., Li, J., Wang, S., and Wang, C. (2020). Intention to adopt bicycle-sharing in China: introducing environmental concern into the theory of planned behavior model. *Environmental Science and Pollution Research*, *27*(33), 41740-41750.

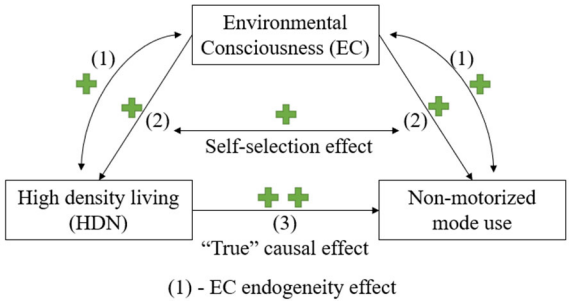

Figure 1a: True case

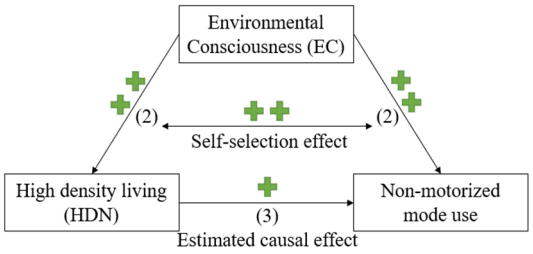

Figure 1b: Incorrect/biased case

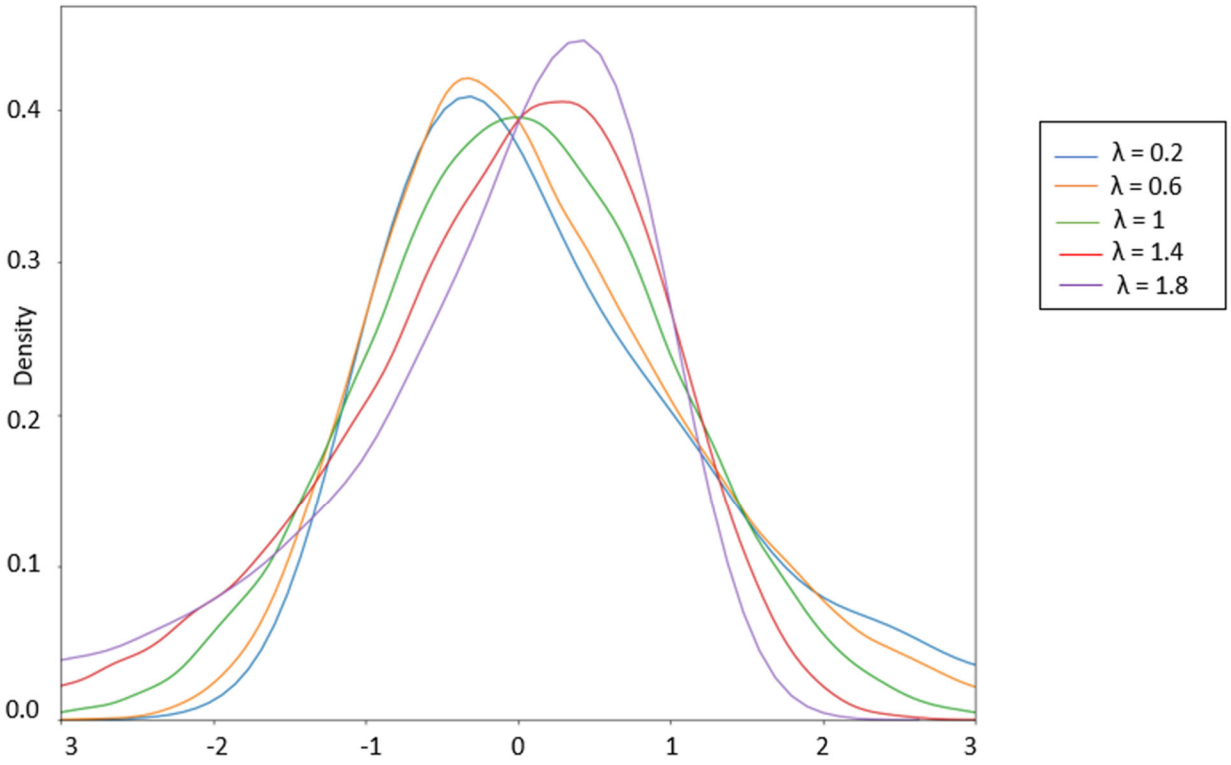

Figure 2: Density of transformed variable for different lambda values

Table 1: Dependent outcomes description

		High-Density Neighborhood (HDN) Living		Total
		No	**Yes**	
Bicycling Frequency	**Never**	1483 (58.2%)* (71.9%)+	1064 (41.8%) (67.2%)	2547 (69.8%)
	Rarely	376 (58.1%) (18.2%)	271 (41.9%) (17.1%)	647 (17.8%)
	Regularly	142 (47.0%) (6.9%)	160 (53.0.%) (10.1%)	302 (8.3%)
	Habitually	61 (40.9%) (3.0%)	88 (59.1%) (5.6%)	149 (4.1%)
Total		2062 (56.6%)	1583 (43.4%)	**3645 (100%)**

* Top percentage in parenthesis refers to row percentage. Thus, 58.2% of individuals who never bicycle do not live in HDNs, while 41.8% of individuals who never bicycle do live in HDNs.

+ Bottom percentage in parenthesis refers to column percentage. Thus, 71.9% of individuals who do not live in HDNs never bicycle, compared to 67.2% of individuals living in HDNs who never bicycle.

Table 2: Latent construct indicator loadings

	Environmental Consciousness (EC)	
	Coeff.	t-stat
Importance of being within a reasonably short commute to work.	0.611	33.13
Importance of having a walkable neighborhood and being near local activities when choosing residence location.	1.208	25.45
Importance of being close to public transit when choosing residence location.	1.107	27.23

Table 3: Latent construct SEM component

Variables (base category)	Environmental Consciousness (EC)	
	Coeff.	t-stat
Individual-Level Characteristics		
Female (base: male)	0.142	2.77
Age (base: below 35 years)		
35 to 44	-	-
45 to 64	-0.390	-8.77
65+	-0.658	-9.12
Education level (base: less than Bachelor's)		
Bachelor's degree or higher	0.529	5.17
Employed (base: not employed)	0.166	2.56
Household-Level Characteristics		
Presence of child below 18 years (base: not present)	-0.527	-9.17
Annual household income (base: less than $25,000)		
$25,000 - $49,999	-	-
$50,000 - $99,999	-	-
Greater $100,000	0.103	2.28
Transformation parameter (λ)	0.466	3.68*

*The standard error of λ is 0.145, and the t-statistic entry here is for the test of $\lambda = 1$ (that is, for the test of normality of the EC latent construct)

Table 4: Main outcome results (coefficients represent effects on underlying latent propensity)

Variables	High-Density Neighborhood Living		Bicycling Frequency	
	Coeff.	t-stat	Coeff.	t-stat
Latent Construct Effects				
Environmental consciousness (EC)	0.351	2.55	0.148	2.55
EC * Worker	-	-	0.114	2.05
Individual Characteristics				
Female (base: Male)	-	-	-0.295	-5.22
Age (base: below 35 years)				
35 to 44	-0.295	-4.02	-	-
45 to 64	-0.507	-2.99	-	-
65+	-0.678	-4.67	-0.275	-3.88
Education level (base: less than Bachelor's)				
Bachelor's degree	0.547	3.78	-	-
Graduate or higher degree	0.678	5.05	0.154	2.98
Student (base: non-student)	0.157	2.44	-	-
Household Characteristics				
Number of adults is 3 or more (base: 2 or less)	-0.725	-6.06	-	-
Child(ren) below 5 years of age present (base: not present)	-0.762	-5.13	-0.143	-3.11
Child(ren) 5-17 years of age present (base: not present)	-0.762	-5.13	0.248	3.51
Annual household income (base: less than $25,000)				
$25,000 - $49,999	-0.317	-3.12	-0.112	-2.45
$50,000 - $99,999	-0.582	-4.05	-0.112	-2.45
Greater $100,000	-0.428	-3.03	-0.112	-2.45
Endogenous Outcome Effect (ϑ parameter)				
High-density neighborhood (HDN) living	NA	NA	0.111	3.66
Thresholds				
Threshold 1	-0.023	-1.29	0.549	9.12
Threshold 2	NA	NA	1.212	12.56
Threshold 3	NA	NA	1.827	17.77
Correlations with Environmental Consciousness	0.149	2.08	0.354	2.77

Table 5: HDN living effect on monthly bicycling frequency

Metric	IHDM	TR-GHDM	ND-GHDM	N-GHDM	FLEX-GHDM
"True" causal HDN Effect (ATE)	0.321 (100.0%)*	0.131 (40.8%)	0.133 (41.4%)	0.261 (81.3%)	0.263 (81.9%)
Estimated self-selection effect	0.000 (0.0%)	0.190 (59.2%)	0.188 (58.6%)	0.023 (7.1%)	0.021 (6.5%)
Estimated EC endogeneity effect	0.000 (0.0%)	0.000 (0.0%)	0.000 (0.0%)	0.037 (11.6%)	0.037 (11.6%)

*Values in parenthesis provide the contribution of each effect as a percentage of the total HDN effect from the IHDM model (=0.321)

Table 6: Data fit measures summary

Summary Statistics	Model				
	IHDM	TR-GHDM	ND-GHDM	N-GHDM	FLEX-GHDM
Predictive log-likelihood at convergence	-5322.04	-5215.707	-5207.126	-5196.409	-5183.269
Number of parameters	36	51	52	53	54
Bayesian Information Criterion (BIC)	5469.66	5424.835	5420.355	5413.738	5404.699
Constants-only predictive log-likelihood	-5754.98				
Predictive adjusted likelihood ratio index	0.081	0.085	0.086	0.088	0.090
Average probability of correct prediction	0.289	0.310	0.317	0.323	0.334

Table 7: Aggregate shares prediction

Bicycling frequency (monthly)	HDN living?	Observed shares	Predicted shares				
			IHDM	TR-GHDM	ND-GHDM	N-GHDM	FLEX-GHDM
Never	No	40.69%	41.21%	41.10%	40.85%	40.91%	40.80%
Rarely	No	10.32%	9.49%	9.66%	9.71%	9.74%	9.85%
Regularly	No	3.90%	4.39%	4.25%	4.36%	4.31%	4.25%
Habitually	No	1.67%	2.00%	2.00%	2.00%	2.00%	1.95%
Never	Yes	29.19%	28.53%	28.75%	29.03%	29.00%	29.11%
Rarely	Yes	7.43%	8.20%	8.18%	8.09%	8.07%	8.01%
Regularly	Yes	4.39%	4.09%	4.01%	3.98%	3.95%	3.98%
Habitually	Yes	2.41%	2.09%	2.05%	1.98%	2.02%	2.05%
Weighted MAPE			**4.22%**	**3.68%**	**3.24%**	**3.18%**	**2.63%**