University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 21/2018 (280)

Hybrid Choice Models vs. Endogeneity of Indicator Variables: A Monte Carlo Investigation

Wiktor Budziński
Mikołaj Czajkowski

Warsaw 2018

UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
WORKING PAPERS

Hybrid choice models vs. endogeneity of indicator variables: a Monte Carlo investigation

Wiktor Budziński*, Mikołaj Czajkowski

University of Warsaw, Department of Economics, Dluga 44/50, 00-241 Warsaw, Poland
** Corresponding authors: wbudzinski@wne.uw.edu.pl*

Abstract: We investigate the problem of endogeneity in the context of hybrid choice (integrated choice and latent variable) models. We first provide a thorough analysis of potential causes of endogeneity and propose a working taxonomy. We demonstrate that although it is widely believed that the hybrid choice framework is devoid of the endogeneity problem, there is no theoretical reason to expect that this is the case. We then demonstrate empirically that the problem exists in the hybrid choice framework too. By conducting a Monte Carlo experiment, we display the extent of the bias resulting from measurement and endogeneity biases. Finally, we propose two novel solutions to address the problem: by explicitly accounting for correlation between structural and discrete choice component error terms (or with random parameters in a utility function), or by introducing additional latent variables. Using simulated data, we demonstrate that these approaches work as expected, that is, they result in unbiased estimates of all model parameters.

Keywords: hybrid choice models, endogeneity, measurement bias, attitudinal variables, indicators

JEL codes: C35, C51, Q51, R41

http://doi.org.10.26405/WP/WNE/2018/280/021

Acknowledgments: The authors wish to thank the participants of the 5th International Choice Modelling Conference, Cape Town, 2017, and the participants of the 23rd Annual Conference of the European Association of Environmental and Resource Economists, Athens, 2017, who facilitated this study with helpful comments on earlier versions of this paper. WB gratefully acknowledges the support of the National Science Centre of Poland (project 2016/21/N/HS4/02094) and the support of the Foundation for Polish Science (FNP). MC gratefully acknowledges the support of the Polish Ministry of Science and Higher Education and the National Science Centre of Poland (project 2017/25/B/HS4/01076).

1. Introduction

A hybrid choice (HC) model is a flexible tool that incorporates perceptions and cognitive processes into a random utility framework commonly used to model individuals' choices. Behavioral factors, such as psychological or sociological constructs, are represented through latent variables that enter the choice model component, rather than entering the model directly. An HC model can therefore be viewed as a combination of a classical discrete choice model, such as the mixed logit (MXL) model (Revelt and Train 1998) with a Multiple Indicators, Multiple Causes (MIMIC) model (Jöreskog and Goldberger 1975).

Including behavioral variables directly in the choice model is considered methodologically flawed because of (among other factors) potential endogeneity of indicator variables.[1] It is commonly believed that the HC model resolves this problem. For example, Daly et al. (2011) state that "The advantages of the latent variable framework over deterministic attitude incorporation are clear; the model is not affected by endogeneity bias […]." Other papers that make similar statements include Hess and Stathopoulos (2013), Hess et al. (2013), Kløjgaard and Hess (2014), and Bello and Abdulai (2015). This is surprising as, to the best of our knowledge, in all applications of the HC framework to date, correlations between error terms in structural or measurement equations and error terms or random parameters in the choice component are not accounted for. As a result, the endogeneity problem extends to the HC framework.

In this study, we conduct a Monte Carlo simulation to demonstrate that unless correlation of the error terms is accounted for, the HC model does not solve the endogeneity problem and the resulting estimates are biased. We then propose two methods of accounting for endogeneity and demonstrate that they mitigate the bias.

Section 2 presents the general econometric framework of the hybrid choice. Section 3 describes the design of our Monte-Carlo experiment, the data generating process (DGP), the models we compare, and the methodology of comparisons. Next, the results are presented and interpreted in detail. The last section provides a summary, acknowledges the limitations of our study, and concludes with recommendations for future research. The paper is accompanied by

[1] Another potential problem is the measurement error, resulting from the fact that the indicator variables are usually not direct measures of latent constructs, but rather their functions. In this case, the HC models can help by incorporating various econometric models for recovering the latent constructs and explicitly accounting for the measurement error.

an online supplement presenting the results of a Monte Carlo experiment in a more complicated (and hence realistic) DGP setting, demonstrating that the endogeneity problems become even more pronounced in this case. In addition, we make our software codes available online to make the use of hybrid choice models for empirical studies easier and to facilitate future research.[2]

2. The hybrid choice model framework

HC models can consist of up to three parts: a discrete choice model, measurement component, and structural component. We describe each part in detail to set the scene for the empirical illustration that follows.

2.1. Discrete choice model

The discrete choice component of the HC model is based on random utility theory . Individual i's utility (V) from choosing alternative j in choice situation t depends on a vector of observed characteristics $\mathbf{X}$ and unobserved idiosyncrasies, represented by a stochastic component e:

$$V_{ijt} = \mathbf{X}_{ijt}\boldsymbol{\beta}_i + e_{ijt}, \tag{1}$$

where $\boldsymbol{\beta}_i$ denotes a vector of individual-specific parameters, thus allowing for heterogeneous preferences amongst respondents and leading to a mixed logit model (MXL).[3] The stochastic component of the utility function (e_{ijt}) is of unknown, possibly heteroskedastic variance $\left(\operatorname{var}\left(e_{ijt}\right) = s_i^2\right)$. Identification of the model typically relies on normalizing this variance, such that the error term $\varepsilon_{ijt} = e_{ijt}\dfrac{\pi}{\sqrt{6}s_i}$ is an i.i.d. type I extreme value with constant variance $\operatorname{var}\left(\varepsilon_{ijt}\right) = \pi^2/6$. This normalization is particularly convenient because it leads to a closed-form logit expression of the likelihood function.

---

[2] The models estimated herein used the DCE package developed in Matlab and are available at https://github.com/czaj/DCE. The code and data for estimating the specific models presented in this study are available from http://czaj.org/research/supplementary-materials.

[3] Is it typically assumed that individual parameters follow a particular distribution (possibly multivariate distribution allowing for non-zero correlation of model parameters), rather than being separately estimated for each parameter. The distributions can be continuous, leading to a so-called random parameter model, or discrete, resulting in a latent class model. Assuming instead that parameters are the same for all respondents implies homogenous preferences and leads to the multinomial logit model (MNL) as a special case.

The HC model allows random parameters of the utility function to be a function of individual-specific latent variables, denoted by $\mathbf{LV}_i$, and socio-demographic or other directly observable variables (such as different treatments in the survey) collected in the vector $\mathbf{SD}_i$.[4] For a normally distributed $\boldsymbol{\beta}_i$, this dependence can be specified in the following way:

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^* + \mathbf{LV}_i\boldsymbol{\gamma} + \mathbf{SD}_i\boldsymbol{\varphi}, \tag{2}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ are matrices of estimable coefficients and $\boldsymbol{\beta}_i^*$ has a multivariate normal distribution with a vector of means and a covariance matrix to be estimated.[5] As a result, the conditional probability of individual i's choices $\mathbf{y}_i$, for all T_i choice tasks, is given by:

$$P\left(\mathbf{y}_i \mid \mathbf{X}_i, \boldsymbol{\beta}_i^*, \mathbf{LV}_i, \boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\theta}\right) = \prod_{t=1}^{T_i} \frac{\exp\left(\mathbf{X}_{ijt}\boldsymbol{\beta}_i\right)}{\sum_{k=1}^{C}\exp\left(\mathbf{X}_{ikt}\boldsymbol{\beta}_i\right)}, \tag{3}$$

where $\boldsymbol{\theta}$ is a vector of parameters on which $\boldsymbol{\beta}_i^*$ depends.

2.2 Measurement component

The main purpose of including latent variables in an HC model is a belief that they are describing some behavioral or other factors, which cannot be measured in a direct way (unlike, e.g., age or gender). Instead, various indicators are used, which can be expected to be determined by the latent variables.

The model choices for the indicator equations depend on the particular application. The measurement equations could be linear, ordered, binary, multinomial, or count regressions, whatever best fits an interpretation of each indicator. Throughout the simulation that follows, we will use continuous indicator variables and therefore we assume a linear specification of the form:

$$\mathbf{I}_i = \mathbf{LV}_i\boldsymbol{\Gamma} + \mathbf{X}_i^{Mea}\boldsymbol{\Phi} + \boldsymbol{\eta}_i, \tag{4}$$

[4] There are other possible specifications, in which latent variables enter the choice model differently. For example, they can explain class probabilities in the latent class model or variance of error term (scale). We use this specification as we find it to be the most straightforward way of connecting attitudes with individuals' tastes.

[5] The number of columns in $\boldsymbol{\gamma}$ is equal to the number of latent variables, and the number of rows is equal to the number of attributes.

where $\mathbf{I}_i$ is a vector of indicator variables, $\mathbf{X}_i^{Mea}$ is a vector of additional variables that influence indicator variables, but not through the latent variable itself,[6] $\mathbf{\Gamma}$ and $\mathbf{\Phi}$ are matrices of coefficients and $\boldsymbol{\eta}_i$ denotes a vector of error terms assumed to come from a multivariate normal distribution with 0 means and an identity covariance matrix.[7] Essentially, we assume that indicators $\mathbf{I}_i$ are driven by (and hence they are used to measure) unobserved latent variables $\mathbf{LV}_i$ and potentially also by some other observed individual-specific characteristics $\mathbf{X}_i^{Mea}$ while allowing for measurement error, represented by the errors component $\boldsymbol{\eta}_i$.

2.3. Structural component

Latent variables can also directly depend on exogenous factors, such as socio-demographic variables, which are stacked in the vector $\mathbf{X}_i^{str}$. Vector $\mathbf{X}_i^{str}$ may, in principle, overlap with vector $\mathbf{SD}_i$. This relationship is described by the following structural equation:

$$\mathbf{LV}_i = \mathbf{X}_i^{str}\mathbf{\Psi} + \xi_i , \tag{5}$$

with a matrix of coefficients $\mathbf{\Psi}$ and error terms ξ_i, which are assumed to come from a multivariate normal distribution.[8] Generally, linking socio-demographic variables with latent variables through structural equations is not necessary. In the absence of such structural equations, latent variables become similar to random parameters; they capture the correlation between individuals' preferences and measurement variables.

2.4. Identification

In order to make identification of hybrid choice models possible, the scale of a latent variable needs to be normalized (Daly et al. 2012). This can be done by normalizing variances of the error terms in the structural equations or by normalizing some coefficients in the $\mathbf{\Gamma}$ matrix for each latent variable (Raveau et al. 2012). In this study we adopt the former approach. Contrary to most studies conducted till date, we do not normalize variance of ξ_i to one. Instead, we use

---

[6] For example, some individuals may have a tendency to overstate (or understate) their real attitudes.

[7] It is important to note that the number of measurement equations need not equal the number of latent variables. For instance, cases may arise where more than one indicator may be available for a latent variable (e.g., there may be two survey questions targeting the same underlying psychological construct). This framework can accommodate such a setting by specifying multiple measurement equations for a single latent variable.

[8] This is a common assumption, although Bhat et al. (2015) introduce a specification that allows for non-normal error terms.

normalization to assure that the variance of every latent variable in $\mathbf{LV}_i$ is equal to one. Although such an approach introduces additional nonlinearities into the model, it is very useful, as all latent variables now have the same scale (even with socio-demographic variables in structural equations) and, therefore, their relative importance (e.g., in measurement equations) can be easily assessed. We do not observe any additional issues with convergence due to this normalization.

We formally define $\mathbf{LV}_i^* = \mathbf{X}_i^{str}\mathbf{\Psi}^* + \xi_i^*$, with $\mathbf{\Psi}^*$ being a matrix of parameters to be estimated and ξ_i^* being a vector of independent normally distributed variables with mean zero and unit standard deviation. For $\mathbf{LV}_{\bullet k}^*$, representing a vector of values of the *k*-th non-normalized latent variable for all individuals and $\boldsymbol{\delta}_k = std\left(\mathbf{LV}_{\bullet k}^*\right)$, its standard deviations, we have $\mathbf{LV}_{\bullet k} = \mathbf{LV}_{\bullet k}^* / \boldsymbol{\delta}_k$, $\mathbf{\Psi}_k = \mathbf{\Psi}_k^* / \boldsymbol{\delta}_k$ and $\xi_{\bullet k} = \xi_{\bullet k}^* / \boldsymbol{\delta}_k$.[9]

Unfortunately, exact conditions for identification of the HC model are not yet known; they depend on the number of latent variables and measurement equations (Bahamonde-Birke et al. 2015). We follow Bollen and Davis (2009) to ensure the necessary condition for identification of structural equation models holds; our specifications satisfy the "2+ emitted paths rule" (we assume that each latent variable has two unique indicators in the measurement component and is linked with three preference parameters in the discrete choice component).[10] In the simulation that follows, we use several different specifications to compare their ability to recover true values of parameters. Because the final specification we use (with the highest number of parameters) encountered no problems in identification and produces stable results, we conclude that the model is identified.[11] Since other specifications are nested (have more restrictions on parameters), they also are identified.

2.5. Estimation

Finally, we combine the discrete choice model specified in (3), the measurement equations defined in (4), and structural equations described in (5) to obtain the full-information likelihood

---

[9] $\mathbf{\Psi}_k$ denotes *k*-th row of $\mathbf{\Psi}$ matrix, and $\xi_{\bullet k}$ denotes stacked values of the random term in the *k*-th structural equation for all individuals.

[10] One exception is the Model 9 (see below), which uses two latent variables and two indicators. In this case we have also used simulation to confirm that the model is identified and produces stable results.

[11] This approach for testing whether the HC model is identified was suggested by Ben-Akiva et al. (2002).

function for the HC model (for ease of exposition, we stack the parameters $\boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ into $\boldsymbol{\Omega}$):

$$L_i = \int P\left(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{X}_i^{str}, \boldsymbol{\beta}_i^*, \boldsymbol{\xi}_i^*, \boldsymbol{\Omega}\right) P\left(\mathbf{I}_i \mid \mathbf{X}_i^{str}, \boldsymbol{\xi}_i^*, \mathbf{X}_i^{Mea}, \boldsymbol{\Omega}\right) f\left(\boldsymbol{\beta}_i^*, \boldsymbol{\xi}_i^* \mid \boldsymbol{\theta}\right) d\left(\boldsymbol{\beta}_i^*, \boldsymbol{\xi}_i^*\right). \quad (6)$$

As random disturbances of $\boldsymbol{\beta}_i^*$ and (non-normalized) error terms in structural equations $\boldsymbol{\xi}_i^*$ are not directly observed, they must be integrated out of the conditional likelihood. This multidimensional integral can be approximated using a simulated maximum likelihood approach. As can be seen, we use one-step estimation. This approach has two main advantages over a two-step (or multi-step) method. First, it is more efficient, and second, it allows for identification of more flexible specifications because it has more degrees of freedom.

3. The setup of the Monte Carlo investigation of the effects of endogeneity

In mathematical terms, endogeneity boils down to the correlation of error terms.[12] Chorus and Kroesen (2014) list missing variables that influence both latent variables and an individual's choices as one of the most common causes of endogeneity.[13] We use missing variables to cause endogeneity and investigate its effects because it is a convenient way to generate such data. However, our results are not limited to this case only. If different reasons for correlation of error terms arise, the results would be qualitatively the same.

The literature is not always clear on what exactly is understood by endogeneity of indicator variables. Most papers consider indicator variables to be endogenous, but do not consider underlying latent factors as such. This is in stark contrast to Chorus and Kroesen (2014), who indicate the endogeneity of attitudes and perceptions themselves as one of the three most apparent reasons for endogeneity. To make this distinction clearer, we differentiate between the two in our data-generating process. In the case when a latent variable is endogenous (there is correlation of the error terms of the structural equation with the error terms in the discrete choice model), we label it as "LV-endogeneity." In the case when an indicator variable in the measurement component is endogenous (but not due to LV-

---

[12] That is, the correlation between error terms in structural ($\boldsymbol{\xi}_i$) or measurement ($\boldsymbol{\eta}_i$) equations with error terms in the discrete choice component (e_{ijt}) or with random parameters ($\boldsymbol{\beta}_i^*$).

[13] Other common causes include learning effects, which is a bi-directional dependence between experience with attributes and their perceptions, and aligning individual's attitudes with their actual choices in order to appear consistent (c.f. Ariely et al. 2003).

endogeneity), we refer to it as "M-endogeneity" (in this case the error terms in measurement equations are correlated with the error terms in the discrete choice model).[14]

To investigate the effects of endogeneity under various model specifications we designed and conducted a Monte Carlo simulation. Essentially, we assumed various DGPs that included different kinds of endogeneity, simulated data, and investigated how well different model specifications perform in terms of recovering the original parameters. The process was repeated multiple times to make sure the results are not coincidental. This exercise allows us to clearly demonstrate which specifications suffer from the endogeneity problem (and hence result in biased estimates) and how can the problem be controlled.

3.1. Data generating process

The DGP we selected is relatively simple and mimics the usual settings of studies dealing with discrete choice data. The discrete choice consists of three choice alternatives and six choice tasks per respondent. It includes three attributes: a binary variable SQ_{ijt} representing an alternative specific constant for the first (status quo) alternative, and two continuous attributes $Quality_{ijt}$ and $Cost_{ijt}$, assumed to always equal 0 for the status quo alternative, and distributed (independently) uniformly between 0 and 2. Each artificial sample consists of 1,000 individuals. The individual-specific explanatory variables X_i^{SD} and X_i^{Miss} were assumed to have standard normal distribution. Table 1 describes the details of the DGP.

[14] LV-endogeneity arises in all cases listed by Chorus and Kroesen (2014). M-endogeneity occurs when the same unobserved factor influences measurement error and individual choice. In stated preference studies, an example of the second type of factor would be the "yea-saying" latent factor, which can make individuals overstate their real attitudes (whether the indicator questions are framed positively or negatively) as well as more likely to choose costly improvement alternatives.

Table 1: Description of the data-generating process used for Monte Carlo simulations

	LV-endogeneity	M-endogeneity
Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $SQ_{ijt} \in \{0,1\}$ $Quality_{ijt} \sim U(0,2)$ $Cost_{ijt} \sim U(0,2)$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$ $\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$ $e_{ijt} \sim EV_I(0,1)$	
Individual-specific characteristics	$X_i^{SD} \sim N(0,1)$ $X_i^{Miss} \sim N(0,1)$	
Indicator variables (measurement component)	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$ $\eta_{i1} \sim N(0,1)$ $\eta_{i2} \sim N(0,1)$	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1} + \alpha_{44} X_i^{Miss}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2} + \alpha_{54} X_i^{Miss}$ $\eta_{i1} \sim N(0,1)$ $\eta_{i2} \sim N(0,1)$
Latent variables (structural component)	$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i + \alpha_{62} X_i^{Miss}$ $\xi_i \sim N(0,1)$	$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$ $\xi_i \sim N(0,1)$

We assume that the only sources of preference heterogeneity are the individual-specific X_i^{Miss} variable and the latent variable. Omitting X_i^{Miss} in model specification will make the error terms correlated and hence cause endogeneity. The difference between the two endogeneity types results from the way in which X_i^{Miss} enters the model. In the case of LV-endogeneity, X_i^{Miss} enters the structural component (the latent variable equation) and hence if it is not included in the model specification, it makes error terms of structural and discrete choice components correlated $\left(cor\left(\xi_i, e_{ijt}\right) \neq 0\right)$. In contrast, in the M-endogeneity case, X_i^{Miss} enters through the measurement component (indicator variable equation) and if omitted, it makes utility function and measurement equations error terms correlated $\left(cor\left(\eta_i, e_{ijt}\right) \neq 0\right)$.[15]

[15] Note that we assume that X_i^{Miss} enters the utility function through the parameter of the alternative specific constant only. This is useful because omitting X_i^{Miss} from the specification will linearly alter the error term of the status quo alternative only. If it entered through the parameters of all other attributes, the error terms of the alternatives would be affected in a non-linear manner, making interpretation of the results more difficult. The results of such a specification are available in an online supplement to this paper (Appendix B). It does not qualitatively change our results.

3.2. Comparison of the models

We use different model specifications to examine whether the presence of LV- or M-endogeneity affects the results and to what extent it is controlled. Table 2 provides an overview of model specifications.

The first model (Model 1) reflects the data-generating process with no missing variables. It is used to test if we are able to correctly recover the parameters when there is no endogeneity issue present. Models 2 and 3 are similar in terms of no missing variables, but they do not use the hybrid framework. Instead, the X_i^{Miss} and indicator variables enter the model directly as explanatory variables of random parameters ($\mathbf{SD}_i$ in equation (2)), rather than through latent variables. Additionally, Model 3 has a random parameter for the status quo alternative specific constant (ASC). These standard discrete choice models (multinomial logit, MNL, and mixed logit, MXL) allow us to illustrate the bias resulting from the measurement error only (as no variable is missing).

In the following specifications we omit X_i^{Miss} as if it was unobserved and hence the models suffer from LV- or M-endogeneity depending on the DGP. Model 4 and Model 5 are MNL and MXL models, respectively, which suffer from both endogeneity and measurement biases. The MXL model has a random parameter for status quo ASC, which captures unobserved preference heterogeneity caused by the omitted X_i^{Miss} variable.

Models 6 and 7 are hybrid specifications commonly used in empirical studies. They account for measurement bias by using a latent variables framework, although they are still affected by the endogeneity issue. In Model 7, as in Model 5, there is a normally distributed random parameter for status quo ASC, which captures unobserved preference heterogeneity.[16]

The last two specifications represent different ways to control for endogeneity and show that if the correlation of the error terms is accounted for, the HC model can correctly recover the parameters even in the case of unobserved X_i^{Miss}. Model 8 is the hybrid MXL that allows for estimable correlation between the error term in the structural component (ξ_i) and the

[16] In real-life situations, researchers should probably make parameters of all attributes random. We do not do this in order to make it easier to conduct the simulation and interpret the results (note that the DGP does not require random parameter specification for the other attributes). In the online supplement to this paper (Appendix B), we present the results of a model in which all parameters are random.

random parameter of the SQ attribute (β_{i1}). This is equivalent to directly modelling the correlation between ξ_i and e_{ijt}, but it is easier to operationalize. Model 9 is a hybrid MNL that uses a different approach to control for endogeneity and is applicable to both LV- and M-endogeneity cases. It does not make the SQ parameter random; instead, it assumes there is an additional latent variable that enters both measurement equations. By using this additional latent variable, we impute the unobserved X_i^{Miss} variable into the model. In doing so, it will recover both the unobserved preference heterogeneity of the SQ parameter and the effect of the unobserved X_i^{Miss} on indicator variables.

Table 2: Overview of model specifications used to investigate the endogeneity bias in the controlled and uncontrolled cases

	Model type	**Measurement error**	**Endogeneity**	**Description**
Model 1	Hybrid MNL	No	No	No missing variables
Model 2	MNL	Yes	No	No missing variables, indicator variables entering directly
Model 3	MXL	Yes	No	No missing variables, β_{i1} random, indicator variables entering directly
Model 4	MNL	Yes	Yes	X_i^{Miss} missing, indicator variables entering directly
Model 5	MXL	Yes	Yes	X_i^{Miss} missing, β_{i1} random, indicator variables entering directly
Model 6	Hybrid MNL	Controlled	Yes	X_i^{Miss} missing
Model 7	Hybrid MXL	Controlled	Yes	X_i^{Miss} missing, β_{i1} random
Model 8	Hybrid MXL	Controlled	Controlled[17]	X_i^{Miss} missing, β_{i1} random, correlation between ξ_i and β_{i1} allowed
Model 9	Hybrid MNL	Controlled	Controlled	X_i^{Miss} missing, additional LV in model specification

[17] LV-endogeneity only.

We assume that Model 1 will recover the DGP parameters correctly, while the estimates of Models 2–7 will be biased due to measurement and/or endogeneity bias. If Model 8 performs as expected, it should recover the parameters well in the LV-endogeneity case (i.e., it controls for endogeneity and measurement error in this case). Model 9 should work well in both the LV- and M-endogeneity cases. The reason why it works in the LV-endogeneity case is that even though X_i^{Miss} should be included in the structural equation of the latent variable, it is possible to instead use it as an interaction with all attributes in the utility function ($\mathbf{SD}_i$ in equation (2)) and as an additional explanatory variable in measurement equations ($\mathbf{X}_i^{Mea}$ in equation (4)), which leads to mathematically equivalent specification. That is, the additional latent variable in Model 9 can be used to control for the missing X_i^{Miss} variable. We provide full description of each model specification for the cases of both LV- and M-endogeneity in Appendix A.

3.3. Rescaling of the parameters

Identification of a hybrid choice model requires normalization of selected parameters in the structural component. In the case of missing variables, normalization leads to rescaling of the remaining parameters (e.g., variance of the error term). As a result, rescaled parameters may differ from the parameters assumed by the DGP, but they will not be biased.

To illustrate, consider the case of LV-endogeneity, in which we assume that the parameter for status quo ASC and the structural equation for a latent variable are given by:

$$\begin{aligned} \beta_{1i} &= \alpha_{11} + \frac{\alpha_{12}}{\delta} LV_i^* + \alpha_{13} X_i^{Miss} \\ LV_i^* &= \alpha_{61} X_i^{SD} + \xi_i + \alpha_{62} X_i^{Miss} \end{aligned}, \tag{7}$$

where $\delta = \sqrt{1+\alpha_{61}^2+\alpha_{62}^2}$ is the standard deviation of the latent variable before normalization as described in section 2.4. When X_i^{Miss} is missing (Model 8), we assume a different formulation, namely (for detailed specification see Appendix A):

$$\begin{aligned} \beta_{1i} &= \alpha_{11}^* + \frac{\alpha_{12}^*}{\delta^*} LV_i^* + \alpha_{13}^* \beta_{1i}^* \\ LV_i^* &= \alpha_{61}^* X_i^{SD} + \xi_i^* \end{aligned}, \tag{8}$$

where β_{1i}^{*} is a normally distributed parameter correlated with ξ_i^{*} (it accounts for unobserved heterogeneity caused by X_i^{Miss} as well as endogeneity) and $\delta^{*}=\sqrt{1+\left(\alpha_{61}^{*}\right)^{2}}$. Because X_i^{Miss} is no longer observed, it enters the error term of the structural equation, $\xi_i^{**}=\xi_i+\alpha_{62}X_i^{Miss}$. However, as the error term is normalized for unit variance, the structural equation becomes rescaled, leading to $\alpha_{61}^{*}=\frac{\alpha_{61}}{\sqrt{1+\alpha_{62}^{2}}}$ and $\xi_i^{*}=\frac{1}{\sqrt{1+\alpha_{62}^{2}}}\xi_i+\frac{\alpha_{62}}{\sqrt{1+\alpha_{62}^{2}}}X_i^{Miss}$. The mean value of individual-specific parameters will not change, therefore $\alpha_{11}^{*}=\alpha_{11}$. Lastly, the effect of X_i^{SD} on β_{1i} (through the latent variable) should be the same in both specifications, it therefore needs to hold that $\frac{\alpha_{12}^{*}}{\delta^{*}}\alpha_{61}^{*}=\frac{\alpha_{12}}{\delta}\alpha_{61}$, from which it follows that $\alpha_{12}^{*}=\alpha_{12}$. Therefore, in Model 8 only the coefficient for non-missing variables in the structural equation becomes rescaled, while all other coefficients should be the same as in the DGP.[18, 19]

On the other hand, when X_i^{Miss} becomes the missing variable and Model 9 is used, the following formulation is assumed:

$$\begin{array}{l}\beta_{1i}=\alpha_{11}^{**}+\frac{\alpha_{12}^{**}}{\delta^{**}}LV_i^{*}+\alpha_{13}^{**}LV_{2,i}\\ LV_{1,i}^{*}=\alpha_{61}^{**}X_i^{SD}+\xi_i^{**}\end{array}, \tag{9}$$

where $\delta^{**}=\sqrt{1+\left(\alpha_{61}^{**}\right)^{2}}$ and $LV_{2,i}$ is an additional latent variable, which we impute for X_i^{Miss}. In this case we remove X_i^{Miss} from $LV_{1,i}^{*}$, so it does not enter the error term and therefore $\alpha_{61}^{**}=\alpha_{61}$ and $\xi_i^{**}=\xi_i$. Once again, the effect of X_i^{SD} on β_{1i} (through the latent variable) should be the same in this specification as in true DGP, it therefore needs to hold that $\frac{\alpha_{12}^{**}}{\delta^{**}}\alpha_{61}^{**}=\frac{\alpha_{12}}{\delta}\alpha_{61}$, from which it follows that $\alpha_{12}^{*}=\alpha_{12}\frac{\sqrt{1+\alpha_{61}^{2}}}{\sqrt{1+\alpha_{61}^{2}+\alpha_{62}^{2}}}$. As a result, when using Model 9 in the LV-endogeneity case, coefficients in the structural equation will stay the same as in the DGP, but the effect of the latent variable on preferences will be rescaled.

---

[18] In the brief analysis above we considered only the equation for β_{1i}, but analogous analysis can be conducted for parameters of other attributes and coefficients in measurement equations.

[19] The rescaling considered here is not caused by the way in which the latent variable is normalized. If another typically used normalization is employed (ξ_i with variance equal one and $\delta=1$), not only the coefficient in the structural equation would be rescaled, but also the coefficients of the interactions of latent variables and preferences (α_{12}).

This brief discussion illustrates that, similar to the case of coefficients for utility function, where parameters do not have a quantitative interpretation, parameters associated with latent variables should also be interpreted with caution when different specifications are compared. The coefficients may change due to rescaling, rather than a true change in predicted effects. We take this effect into account when interpreting our results.

3.4. Methodology of the comparisons

To compare the estimates resulting from using the different models to their true values, we require a method that not only looks at their expected values, but also penalizes for variance. Consider the usual case when one tests if x_i is statistically equal to its true value x_{true} (e.g., using the standard t-test). The larger the variance associated with x_i, the more difficult it is to reject the equality hypothesis. As a result, models that would result in high variation of the estimates (high standard errors) would make it easier to falsely conclude that an estimate is not statistically significantly different from its true value.

To address this problem, we base our comparisons on equivalence tests (Hauck and Anderson 1984, Kristofersson and Navrud 2005). Equivalence tests reverse the null hypothesis and the alternative hypothesis; instead of testing if x_i is equal to x_{true}, we test if the absolute difference between them is higher than an *a priori* defined "acceptable" level. Czajkowski and Ščasný (2010) and Czajkowski et al. (2017a) operationalize equivalence tests by proposing to search for a Minimum Tolerance Level (MTL), that is, the minimum "acceptable" difference that allows us to conclude that two values are equivalent at the required level of statistical significance.

For a random variable ω, MTL is formally defined as the minimum $\theta \geq 0$ that satisfies:

$$P\left(\left|\omega - \omega_{true}\right| > \theta \cdot |\omega_{true}|\right) = \alpha, \tag{10}$$

where α is the required significance level (e.g., 0.05). In our case,[20] the probability can be evaluated using Two One-Sided T-Tests, while MTL can be found as:

$$MTL_{\alpha} = \underset{\theta \in [0,+\infty)}{\arg\min} \{\theta\} \text{ s.t.} P\left(\left|\omega - \omega_{true}\right| > \theta \cdot |\omega_{true}|\right) \leq \alpha. \tag{11}$$

[20] A collection of Matlab functions that are useful for calculating MTL is available at https://github.com/czaj/BTtools.

MTL has an intuitive interpretation. For example, $MTL_{0.05} = 0.01$ means that, with 95% probability, the deviation of the estimated coefficients from the true values will not be larger than 1%.

4. Results

We generated 1,000 datasets following LV- and M-endogeneity DGP. For each dataset, we estimated each of the models presented in section 3.2. The results are reported in Table 3 (the LV-endogeneity) and Table 4 (M-endogeneity).

In the case of both the LV- and M-endogeneity, the results of Model 1 indicate that our modelling framework works well. If there are no missing variables and a latent variable framework is used, the true parameter values are recovered with satisfactory precision.

This is not the case if indicator variables are used directly as interactions with choice attributes (Model 2 and Model 3).[21] In Model 2 and Model 3, although no variables are missing, parameter estimates are substantially different from the true values due to measurement bias. In most cases, estimated coefficients are smaller (in absolute values) than the true coefficients. In this case, not only interactions with indicator variables are biased, but sometimes other coefficients, such as interaction with X_i^{Miss} or even main effects for attributes are also biased. The same holds for Models 4 and 5, which suffer from both measurement and endogeneity bias, although, surprisingly, the distance from the true values appears to be smaller in some cases, as measurement bias and endogeneity bias can work in opposite directions. Overall, we consider this convincing evidence against using any attitudinal variables as direct interactions of the model parameters.[22] Hybrid choice models have an obvious advantage in this regard by directly accounting for the measurement error.

The next two models (Model 6 and Model 7) correspond to the case when the measurement bias is controlled by using the hybrid choice framework. However, it is apparently not enough to account for either LV- or M-endogeneity, as some parameter estimates remain biased. In the case of LV-endogeneity, the parameter that has the largest bias

---

[21] Indicator variables I_1 and I_2 enter Models 2 and 3 (MNL and MXL) with means normalized to 0 in order to ensure that mean estimates of the parameters associated with SQ, $Quality$, and $Cost$ are comparable with those of other models. We also normalized standard deviations to 1 so that those variables have the same range as latent variables used in other specifications.

[22] It is likely that the problem is even more significant in the case of ordinal or count, rather than linear indicators, and if the relationship between LV and indicator variables is not one-to-one, as assumed in the DGP used here.

is the interaction of latent variable with the status quo (ASC), as we would expect from the DGP.[23] Accounting for unobserved heterogeneity in Model 7 somewhat helps and takes other coefficients closer to the true values, but it does not help much with regard to the status quo-LV interaction. In the M-endogeneity case, the interactions of LV with all attributes are biased, and accounting for unobserved heterogeneity does not qualitatively change this result. What is more, M-endogeneity makes coefficients in structural and measurement equations biased as well, which we do not observe in the LV-endogeneity case.

Finally, the last two models represent attempts to control for endogeneity, either by explicitly allowing for correlation between error terms in the structural and discrete choice components (Model 8) or assuming the existence of an additional LV to compensate for the missing variable (Model 9). We find that in the LV-endogeneity case, both models perform well, recovering the expected coefficients values (although some of them are rescaled as described in Section 3.3). On average, both specifications have similar log-likelihood, although we note that these models are not nested, and Model 9 employs more coefficients, so some difference in terms of LL is expected. In the M-endogeneity case, we observe that Model 8 does not mitigate the bias, and its results are very close to Model 6 and Model 7, which is not surprising as log-likelihood does not on average substantially change between these specifications. On the other hand, including an additional latent factor (Model 9) works well and, as expected, recovers all model parameters with satisfactory precision.

Finally, we note that Appendix B presents the results of analogous simulation in which the missing variable is interacted not only with status quo ASC but also with other attributes. In such a case we obtain very similar results, but with more biased parameters. Nevertheless, for the LV-endogeneity case, both Model 8 and Model 9 recover DGP well, whereas for M-endogeneity case only Model 9 recovers DGP. In Appendix C we present results of analogous simulation in which we use four measurement equations instead of two. We find that, especially for the LV-endogeneity case, this setting allows for higher precision of estimates, with parameters of Models 8 and 9 being closer to the true values, as indicated by lower values of $MTL_{0.05}$.

[23] In the more realistic case when the missing variable interacts with all attributes in DGP, the results of which are presented in the online Appendix B, these hybrid choice models result in many more biased parameters, specifically, parameters for interactions of LV with all attributes are biased.

Table 3: Results of Monte Carlo simulation of the effects of LV-endogeneity for parameter estimates under different specifications – mean values of the estimates in 1,000 simulations

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0127***	-3.3037	-4.0170***	-2.227	-4.0178***	-3.6394*	-4.0121***	-3.9927***	-4.0152***
	SQ (std. dev.)		2.0000	-	-	1.2469	-	2.3309	-	1.8052	2.0156**	-
	SQ (LV)	α_{12}	-2.0000	-1.9947***	-	-	-	-	-3.0446	-3.1074	-1.9005*	-
	SQ (LV , with LV_2)		-1.6330	-	-	-	-	-	-	-	-	-1.6406**
	SQ (LV_2)		-3.1547	-	-	-	-	-	-	-	-	-3.1356**
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0072***	-2.0724**	-2.3467	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-0.717	-0.9133	-0.8964	-1.5797	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.7284	0.9313	0.9152	1.6142	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0093***	4.6848*	4.8977**	4.3441	4.8995**	4.9246**	5.0068***	5.0006***	5.0192***
	Quality (LV)	α_{22}	1.0000	1.0031***	-	-	-	-	1.1113	1.0051**	1.0082**	-
	Quality (LV , with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8117**
	Quality (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5863**
	Quality ($\cdot I_1$)		1.0000	-	0.5093	0.4554	0.5902	0.456	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.5102	-0.4574	-0.5906	-0.4573	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0073***	-2.7946*	-2.9554**	-2.5314	-2.9540**	-2.9254**	-3.0059***	-3.0041***	-3.0127***
	Cost (LV)	α_{32}	1.0000	1.0055**	-	-	-	-	0.9116*	1.0028***	1.0030***	-
	Cost (LV , with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8131**
	Cost (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5792**
	Cost ($\cdot I_1$)		1.0000	-	0.4249	0.4931	0.3135	0.4919	-	-	-	-
	Cost ($\cdot I_2$)		-1.0000	-	-0.4184	-0.4848	-0.3099	-0.4846	-	-	-	-
Me	I_1 (constant)	α_{41}	-1.0000	-1.0003***	-	-	-	-	-1.0006***	-1.0014***	-1.0006***	-1.0004***

	I_1 (LV)	α_{42}	1.0000	0.9999***	-	-	-	-	0.9940***	1.0041***	0.9996***	-
	I_1 (LV, with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8101**
	I_1 (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5762**
	I_1 (η_1)	α_{43}	0.5000	0.4990***	-	-	-	-	0.5126**	0.4932**	0.5014***	0.4986***
	I_2 (constant)	α_{51}	1.0000	1.0007***	-	-	-	-	1.0011***	1.0018***	1.0011***	1.0008***
	I_2 (LV)	α_{52}	-1.0000	-1.0008***	-	-	-	-	-0.9951***	-1.0052***	-1.0008***	-
	I_2 (LV, with LV_2)		-0.8165	-	-	-	-	-	-	-	-	-0.8107**
	I_2 (LV_2)		-0.5774	-	-	-	-	-	-	-	-	-0.5775**
	I_2 (η_2)	α_{53}	0.5000	0.4979***	-	-	-	-	0.5111**	0.4917**	0.4998***	0.4974***
Structural component	LV^* (X^{SD})		-0.7071	-	-	-	-	-	-0.6476*	-0.6693*	-0.7041***	-
	LV^* (X^{SD} with LV_2)		-1	-	-	-	-	-	-	-	-	-1.0769*
	LV^* (X^{SD} with X^{Miss})	α_{61}	-1	-0.9989***	-	-	-	-	-	-	-	-
	LV^* (X^{Miss})	α_{62}	1	0.9983***	-	-	-	-	-	-	-	-
	$cor(\xi, \beta_1)$		-0.7071	-	-	-	-	-	-	-	-0.7608*	-
	Mean LL			-4011.0178	-1960.126	-1932.4828	-2218.4631	-2067.3666	-4466.0403	-4443.5069	-4428.3391	-4425.0995
	Parameters			15	10	11	9	10	13	14	15	18

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

Table 4: Results of Monte Carlo simulation of the effects of M-endogeneity for parameter estimates under different specifications – mean values of the estimates in 1,000 simulations

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0104***	-3.2799	-3.9678**	-3.0848	-3.9685**	-3.8938**	-3.9389**	-3.9037**	-4.0120***
	SQ (std. dev.)		2.0000	-	-	1.2422	-	1.4475	-	0.855	0.4263	-
	SQ (LV)	α_{12}	-2.0000	-2.0112**	-	-	-	-	-2.6444	-2.6438	-2.7633	-
	SQ (LV , with LV_2)		-2.0000	-	-	-	-	-	-	-	-	-1.9912**
	SQ (LV_2)		-2.0000	-	-	-	-	-	-	-	-	-2.0327**
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0015***	1.2794	1.447	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-3.1585	-3.5645	-1.5931	-1.8609*	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.148	0.3346	0.7018	1.046	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0226***	4.5435*	4.7139*	4.4948	4.7136*	4.7878**	4.7930**	4.7871**	5.0183***
	Quality (LV)	α_{22}	1.0000	1.0011***	-	-	-	-	0.8223	0.8092	0.8166	-
	Quality (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	0.9988***
	Quality ($\cdot I_1$)		1.0000	-	-0.1436	-0.2174	-0.1234	-0.2178	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.9115*	-0.8623	-0.926*	-0.862	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0151***	-2.7486*	-2.8785**	-2.7101*	-2.8781**	-2.9022**	-2.9089**	-2.9017**	-3.0132***
	Cost (LV)	α_{32}	1.0000	1.0058**	-	-	-	-	0.8303	0.8378	0.8289	-
	Cost (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	1.0057**
	Cost ($\cdot I_1$)		1.0000	-	-0.2265	-0.1971	-0.2342	-0.1968	-	-	-	-
	Cost ($\cdot I_2$)		-1.0000	-	-0.8391	-0.9304*	-0.8113	-0.9302*	-	-	-	-
Measurement	I_1 (constant)	α_{41}	-1.0000	-1.0010***	-	-	-	-	-1.0010***	-1.0010***	-0.9999***	-1.0008***
	I_1 (LV)	α_{42}	1.0000	0.9998***	-	-	-	-	1.4972	1.4942	1.4938	0.9980***
	I_1 (LV_2)		1.5000	-	-	-	-	-	-	-	-	1.4923***
	X^{Miss} (in I_1)	α_{44}	1.5000	1.4981***	-	-	-	-	-	-	-	-

I_1 (η_1)	α_{43}	0.5000	0.4984***	-	-	-	-	1.1186	1.1224	1.1198	0.4932***
I_2 (constant)	α_{51}	1.0000	1.0005***	-	-	-	-	1.0004***	1.0004***	0.9996***	1.0002***
I_2 (LV)	α_{52}	-1.0000	-0.9998***	-	-	-	-	-1.1472	-1.1526	-1.1502	-0.9972***
I_2 (LV_2)		-0.5000	-	-	-	-	-	-	-	-	-0.5051***
X^{Miss} (in I_2)	α_{54}	-0.5000	-0.4983***	-	-	-	-	-	-	-	-
I_2 (η_2)	α_{53}	0.5000	0.4991***	-	-	-	-	0.4276	0.4123	0.4142	0.4959***
LV^* (X^{SD})	α_{61}	-1.0000	-1.0007***	-	-	-	-	-0.7556	-0.7557	-0.752	-1.0087**
$cor(\xi, \beta_1)$		-	-	-	-	-	-	-	-	0.274	-
Mean LL			-4153.49	-2000.84	-1976.97	-2037.89	-2000.59	-5037.74	-5037.04	-5035.3	-4960.62
Parameters			18	10	11	9	10	13	14	15	16

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

5. Discussion and conclusion

The hybrid choice framework is an approach that has quickly gained popularity. Vij and Walker (2016) analyze the possible advantages of employing an HC framework and identify a wide range of situations in which its use is justified. Most of the applications till date appear in the environmental economics literature (e.g., Dekker et al. 2012, Hess and Beharry-Borg 2012, Hoyos et al. 2015, Czajkowski et al. 2017b, Czajkowski et al. 2017c, Pakalniete et al. 2017) and transportation (e.g., Vredin Johansson et al. 2006, Daly et al. 2012, Daziano and Bolduc 2013). However, none of these papers explicitly account for the potential correlation between discrete choice and measurement (M-endogeneity) or structural components (LV-endogeneity).

We show that despite the fact that it is commonly assumed that the hybrid choice framework addresses the endogeneity problem, it does not. Using a Monte Carlo simulation, in which we can control the DGP and induce LV- or M-endogeneity, we are able to study the performance of different model specifications. We find that while hybrid choice models generally help with measurement error, this is not enough to address the endogeneity bias. We find that M-endogeneity is generally more severe, but both result in biased estimates.

The LV- and M-endogeneity can be controlled by explicitly allowing for correlation between structural and discrete choice component error terms (or with random variables in utility function) or introducing an additional latent variable. We demonstrate that these approaches work as expected; they result in unbiased estimates of all model parameters. Although the practical usefulness of these approaches is yet to be confirmed,[24] they exhibit that endogeneity should and can be controlled.

We acknowledge that there is a possible taxonomy confusion, which could, to some extent, explain the overall belief in the ability of a hybrid choice model to mitigate endogeneity bias. Namely, if one considers the M-endogeneity case in which there is a correlation between error terms in utility function and those in measurement equations, then switching from a standard discrete choice model (e.g., MXL) to the hybrid framework would resolve the problem of endogeneity bias. This is because in the hybrid framework, indicator variables do not enter the choice model directly, and there is therefore no endogeneity.[25] However, this does not mean

[24] For example, because of the possible non-identification of a model with more latent variables given available data.

[25] This explanation does not work well for LV-endogeneity, as latent variables still enter the utility function directly.

that standard hybrid choice model will work well. Indeed, as we have shown in Table 4 (Models 6-8), estimates are still biased in such cases. The only thing that changes is the type of the bias, which should probably be called misspecification rather than endogeneity. Nevertheless, we believe that for applied research, it is more important whether there is any bias, rather than its type, so for clarity we have chosen to keep referring to this problem as M-endogeneity throughout the paper.

It should be noted that there is another line of research devoted to endogeneity problems in hybrid choice models. However, the nature of the problem analyzed therein is different. In many revealed preference studies, endogeneity of a cost attribute can occur because it can be correlated with other attributes related to the quality of a good being chosen, which are not observed by the researcher. To mitigate this, one can impute these missing attributes as latent variables using several indicator variables as measurement equations. For a review of this literature, see, for example, Guevara and Ben-Akiva (2010). Other methods to address this type of endogeneity include the BLP method (Berry et al. 1995), using a Control Function (Rivers and Vuong 1988, Train 2009)) and Multiple Indicator Solution (Guevara and Polanco 2016). For the comparison of performance of different methods see Guevara (2015). For recent application of these methods in the transportation literature see Fernández-Antolín et al. (2016), and in the environmental literature see Mariel et al. (2018). Nevertheless, This area is outside the scope of our study as we are considering endogeneity of indicators themselves, rather than endogeneity due to omitted attributes.

Finally, we acknowledge that our investigation concerns the case of individual-specific latent variables, or as Bahamonde-Birke et al. (2015) refer to them, “*non-alternative related attitudes*,” in contrast to “*alternative related attitudes*” and “*perceptions*.” We believe that our results are general, although we note that addressing endogeneity in the case of alternative related attitudes or perceptions would likely be much more difficult to address from the modelling perspective.

In summary, our study shows that while hybrid choice models can mitigate the measurement error, they can still suffer from endogeneity bias. We highlight the potential problem, provide a thorough analysis of its potential causes, and propose a method of classification. We use a Monte Carlo experiment to demonstrate the existence and extent of endogeneity bias, propose two ways of addressing it, and confirm that they work. Overall, we hope that our study stimulates further research in this area and will be considered by applied researchers, who often seem to assume that hybrid choice models address the problem of endogeneity, which is generally acknowledged when indicator variables are directly included in the discrete choice model.

References

.

Ariely, D., Loewenstein, G., Prelec, D. 2003. "Coherent arbitrariness": Stable demand curves without stable preferences. Quarterly Journal of Economics 118, 73–105.

Bahamonde-Birke, F.J., Kunert, U., Link, H., Ortúzar, d. J.D. 2015. About attitudes and perceptions: finding the proper way to consider latent variables in discrete choice models. Transportation 44, 1–19.

Bello, M., Abdulai, A. 2015. Measuring heterogeneity, survey engagement and response quality in preferences for organic products in Nigeria. Applied Economics 48, 1–13.

Ben-Akiva, M., Walker, J., Bernardino, A.T., Gopinath, D.A., Morikawa, T., Polydoropoulou, A. 2002. Integration of choice and latent variable models. Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges, 431–470.

Berry, S., Levinsohn, J., Pakes. A. 1995. Automobile prices in market equilibrium. Econometrica 63, 841–890.

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

Bollen, K.A., Davis, W.R. 2009. Two rules of identification for structural equation models. Structural Equation Modeling: A Multidisciplinary Journal 16, 523–536.

Chorus, C.G., Kroesen, M. 2014. On the (im-)possibility of deriving transport policy implications from hybrid choice models. Transport Policy 36, 217–222.

Czajkowski, M., Ahtiainen, H., Artell, J., Meyerhoff, J. 2017a. Choosing a Functional Form for an International Benefit Transfer: Evidence from a Nine-country Valuation Experiment. Ecological Economics 134, 104–113.

Czajkowski, M., Hanley, N., Nyborg, K. 2017b. Social norms, morals and self-interest as determinants of pro-environment behaviours: The case of household recycling. Environmental and Resource Economics 66, 647–670.

Czajkowski, M., Ščasný, M. 2010. Study on benefit transfer in an international setting. How to improve welfare estimates in the case of the countries' income heterogeneity? Ecological Economics 69, 2409–2416.

Czajkowski, M., Vossler, C.A., Budziński, W., Wiśniewska, A., Zawojska, E. 2017c. Addressing empirical challenges related to the incentive compatibility of stated preference methods. Journal of Economic Behavior and Organization 142, 47–63.

Daly, A., Hess, S., Patruni, B., Potoglou, D., Rohr, C. 2011. Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. Transportation 39, 267–297.

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

Dekker, T., Hess, S., Brouwer, R., Hofkes, M. 2012. Implicitly or explicitly uncertain? International Choice Modelling Conference, Sydney, Australia.

Fernández-Antolín, A., Guevara, C.A., Lapparent de, M., Bierlaire, M. 2016. Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data. Journal of Choice Modelling 20, 1–15.

Guevara, C.A. 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transportation Research Part A: Policy and Practice 82, 240–254.

Guevara, C.A., Ben-Akiva, M. 2010. Addressing endogeneity in discrete choice models: Assessing control-function and latent-variable methods, in: Hess, S., Daly, A. (Eds.). Choice Modelling: The State-of-the Art and the State-of-practice, pp.353–371.

Guevara, C.A., Polanco, D. 2016. Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution. Transportmetrica A: Transport Science 12, 458–478.

Hauck, W.W., Anderson, S. 1984. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. Journal of Pharmacokinetics and Biopharmaceutics 12, 83–91.

Hess, S., Beharry-Borg, N. 2012. Accounting for latent attitudes in willingness-to-pay studies: The case of coastal water quality improvements in Tobago. Environmental and Resource Economics 52, 109–131.

Hess, S., Shires, J., Jopson, A. 2013. Accommodating underlying pro-environmental attitudes in a rail travel context: Application of a latent variable latent class specification. Transportation Research Part D: Transport and Environment 25, 42–48.

Hess, S., Stathopoulos, A. 2013. A mixed random utility – Random regret model linking the choice of decision rule to latent character traits. Journal of Choice Modelling 9, 27–38.

Hoyos, D., Mariel, P., Hess, S. 2015. Incorporating environmental attitudes in discrete choice models: An exploration of the utility of the awareness of consequences scale. Science of The Total Environment 505, 1100–1111.

Jöreskog, K.G., Goldberger, A.S. 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association 70, 631–639.

Kløjgaard, M.E., Hess, S. 2014. Understanding the formation and influence of attitudes in patients' treatment choices for lower back pain: Testing the benefits of a hybrid choice model approach. Social Science & Medicine 114, 138–150.

Kristofersson, D., Navrud, S. 2005. Validity tests of benefit transfer – Are we performing the wrong tests? Environmental and Resource Economics 30, 279–286.

Mariel, P., Hoyos, D., Artabe, A., Guevara, C.A. 2018. A multiple indicator solution approach to endogeneity in discrete-choice models for environmental valuation. Science of the Total Environment 633, 967–980.

Pakalniete, K., Aigars, J., Czajkowski, M., Strake, S., Zawojska, E., Hanley, N. 2017. Understanding the distribution of economic benefits from improving coastal and marine ecosystems. Science of the Total Environment 584–585, 29–40.

Raveau, S., Yáñez, M.F., Ortúzar, J. d. D. 2012. Practical and empirical identifiability of hybrid discrete choice models. Transportation Research Part B: Methodological 46, 1374–1383.

Revelt, D., Train, K. 1998. Mixed logit with repeated choices: Households' choices of appliance efficiency level. Review of Economics and Statistics 80, 647–657.

Rivers, D., Vuong, Q.H.. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. Journal of Econometrics 39, 347–366.

Train, K.E. 2009. Discrete Choice Methods with Simulation. 2 edition. Cambridge University Press, New York.

Vij, A., Walker, J.L. 2016. How, when and why integrated choice and latent variable models are latently useful. Transportation Research Part B: Methodological 90, 192–217.

Vredin Johansson, M., Heldt, T., Johansson, P. 2006. The effects of attitudes and personality traits on mode choice. Transportation Research Part A: Policy and Practice 40, 507–525.

Appendix A

Table A1: Details of model specifications used to investigate the endogeneity bias in the controlled and uncontrolled case

<table>
<tr><th></th><th></th><th>LV-endogeneity</th><th>M-endogeneity</th></tr>
<tr><td rowspan="3">Model 1 (Hybrid MNL)</td><td>Utility function</td><td colspan="2">$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$
$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$
$\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$
$\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$</td></tr>
<tr><td>Indicator variables (measurement component)</td><td>$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$
$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$</td><td>$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1} + \alpha_{44} X_i^{Miss}$
$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2} + \alpha_{54} X_i^{Miss}$</td></tr>
<tr><td>Latent variables (structural component)</td><td>$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i + \alpha_{62} X_i^{Miss}$</td><td>$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$</td></tr>
<tr><td>Model 2 (MNL)</td><td>Utility function</td><td colspan="2">$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$
$\beta_{1i} = \alpha_{11} + \alpha_{12} I_{i1} + \alpha_{13} I_{i2} + \alpha_{14} X_i^{Miss}$
$\beta_{2i} = \alpha_{21} + \alpha_{22} I_{i1} + \alpha_{23} I_{i2}$
$\beta_{3i} = \alpha_{31} + \alpha_{32} I_{i1} + \alpha_{33} I_{i2}$</td></tr>
<tr><td>Model 3 (MXL)</td><td>Utility function</td><td colspan="2">$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$
$\beta_{1i} = \alpha_{11} + \alpha_{12} I_{i1} + \alpha_{13} I_{i2} + \alpha_{14} X_i^{Miss} + \beta_{1i}^*$
$\beta_{2i} = \alpha_{21} + \alpha_{22} I_{i1} + \alpha_{23} I_{i2}$
$\beta_{3i} = \alpha_{31} + \alpha_{32} I_{i1} + \alpha_{33} I_{i2}$</td></tr>
<tr><td>Model 4 (MNL)</td><td>Utility function</td><td colspan="2">$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$
$\beta_{1i} = \alpha_{11} + \alpha_{12} I_{i1} + \alpha_{13} I_{i2}$
$\beta_{2i} = \alpha_{21} + \alpha_{22} I_{i1} + \alpha_{23} I_{i2}$
$\beta_{3i} = \alpha_{31} + \alpha_{32} I_{i1} + \alpha_{33} I_{i2}$</td></tr>
<tr><td>Model 5 (MXL)</td><td>Utility function</td><td colspan="2">$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$
$\beta_{1i} = \alpha_{11} + \alpha_{12} I_{i1} + \alpha_{13} I_{i2} + \beta_{1i}^*$
$\beta_{2i} = \alpha_{21} + \alpha_{22} I_{i1} + \alpha_{23} I_{i2}$
$\beta_{3i} = \alpha_{31} + \alpha_{32} I_{i1} + \alpha_{33} I_{i2}$</td></tr>
</table>

Model 6 (Hybrid MNL)	Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$ $\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$
	Indicator variables (measurement component)	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$
	Latent variables (structural component)	$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$
Model 7 (Hybrid MXL)	Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} \beta_{1i}^*$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$ $\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$
	Indicator variables (measurement component)	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$
	Latent variables (structural component)	$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$
Model 8 (Hybrid MXL, correlation controlled)	Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} \beta_{1i}^*$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i$ $\beta_{3i} = \alpha_{31} + \alpha_{32} LV_i$ $corr\left(\beta_{1i}^*, \xi_i\right) = \rho$
	Indicator variables (measurement component)	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$
	Latent variables (structural component)	$LV_i^* = \alpha_{61} X_i^{SD} + \xi_i$
Model 9 (Hybrid MNL, additional LV)	Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} + \beta_{3i} Cost_{ijt} + e_{ijt}$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_{1,i} + \alpha_{13} LV_{2,i}$ $\quad$ $\beta_{1i} = \alpha_{11} + \alpha_{12} LV_{1,i} + \alpha_{13} LV_{2,i}$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_{1,i} + \alpha_{23} LV_{2,i}$ $\quad$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_{1,i}$ $\beta_{3i} = \alpha_{31} + \alpha_{22} LV_{1,i} + \alpha_{23} LV_{2,i}$ $\quad$ $\beta_{3i} = \alpha_{31} + \alpha_{22} LV_{1,i}$
	Indicator variables (measurement component)	$I_{i1} = \alpha_{41} + \alpha_{42} LV_{1,i} + \alpha_{43} \eta_{i1} + \alpha_{44} LV_{2,i}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_{1,i} + \alpha_{53} \eta_{i2} + \alpha_{54} LV_{2,i}$
	Latent variables (structural component)	$LV_{1,i}^* = \alpha_{61} X_i^{SD} + \xi_{1,i}$ $LV_{2i}^* = \xi_{2,i}$

Appendix B

Table B1: The results of Monte Carlo simulation of the effects of LV-endogeneity for parameter estimates under different specifications - mean values of the estimates in 1,000 simulations. Missing variable enters data generating process as an interaction with all attributes.

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0181***	-3.2466	-4.0607**	-1.6242	-4.0507**	-3.9780***	-4.0843**	-4.0002***	-4.0264**
	SQ (std. dev.)		2.0000	-	-	0.7907	-	2.0155**	-	1.8393*	2.0070***	-
	SQ (LV)	α_{12}	-2.0000	-2.0148**	-	-	-	-	-3.6088	-3.1759	-2.0011***	-
	SQ (LV , with LV_2)		-1.6330	-	-	-	-	-	-	-	-	-1.6516**
	SQ (LV_2)		-3.1547	-	-	-	-	-	-	-	-	-3.1764**
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0051***	-1.8671*	-2.2217	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-0.7599	-0.967	-0.7692	-1.6048	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.7493	0.9532	0.7595	1.5885	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0154***	4.7049*	5.0786**	3.4589	5.0487**	4.6108*	5.0757**	5.0026***	5.0112***
	Quality (std. dev.)		2.0000	-	-	1.7716	-	2.0075**	-	1.8577*	1.9910***	-
	Quality (LV)	α_{22}	1.0000	1.0026**	-	-	-	-	2.3064	2.1267	1.0053**	-
	Quality (LV , with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8264**
	Quality (LV_2)		2.5774	-	-	-	-	-	-	-	-	2.5709***
	Quality (X^{Miss})	α_{23}	2.0000	2.0034***	1.9364**	2.1079*	-	-	-	-	-	-
	Quality ($\cdot I_1$)		1.0000	-	0.475	0.4766	0.7722	1.0761*	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.4807	-0.4824	-0.7712	-1.0819*	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0072***	-2.7984*	-3.0462**	-2.0469	-3.0226***	-2.7137*	-3.0456**	-3.0004***	-3.0046***
	Cost (std. dev.)		1.0000	-	-	5.2812	-	1.8457	-	1.2194	0.9936**	-
	Cost (LV)	α_{32}	1.0000	0.9999***	-	-	-	-	0.2706	0.4507	0.9965***	-
	Cost (LV , with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8079**

	$Cost\ (LV_2)$		-0.4226	-	-	-	-	-	-	-	-	-0.4120**
	$Cost\ (X^{Miss})$	α_{33}	-1.0000	-0.9985***	-0.861	-0.9127*	-	-	-	-	-	-
	$Cost\ (\cdot I_1)$		1.0000	-	0.3985	0.4709	0.0907	0.2087	-	-	-	-
	$Cost\ (\cdot I_2)$		-1.0000	-	-0.4016	-0.4748	-0.096	-0.214	-	-	-	-
Measurement component	I_1 (constant)	α_{41}	-1.0000	-1.0004***	-	-	-	-	-1.0005***	-1.0012***	-1.0004***	-1.0007***
	I_1 (LV)	α_{42}	1.0000	0.9989***	-	-	-	-	0.9739**	1.0037***	0.9989***	-
	I_1 (LV, with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8131***
	I_1 (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5778***
	I_1 (η_1)	α_{43}	0.5000	0.4986***	-	-	-	-	0.5468*	0.4903**	0.4992***	0.4984***
	I_2 (constant)	α_{51}	1.0000	1.0001***	-	-	-	-	1.0002***	1.0009***	1.0002***	1.0004***
	I_2 (LV)	α_{52}	-1.0000	-1.0009***	-	-	-	-	-0.9758**	-1.0057***	-1.0008***	-
	I_2 (LV, with LV_2)		-0.8165	-	-	-	-	-	-	-	-	-0.8147***
	I_2 (LV_2)		-0.5774	-	-	-	-	-	-	-	-	-0.5791***
	I_2 (η_2)	α_{53}	0.5000	0.4998***	-	-	-	-	0.5483*	0.4915**	0.5004***	0.4997***
Structural component	LV^* (X^{SD})		-0.7071	-	-	-	-	-	-0.5966	-0.6653*	-0.7060***	-
	LV^* (X^{SD} with LV_2)		-1	-	-	-	-	-	-	-	-	-1.0146**
	LV^* (X^{SD} with X^{Miss})	α_{61}	-1	-1.0009***	-	-	-	-	-	-	-	-
	LV^* (X^{Miss})	α_{62}	1	1.0018***	-	-	-	-	-	-	-	-
	$cor(\xi, \beta_1)$		-0.7071	-	-	-	-	-	-	-	-0.7068***	-
	$cor(\xi, \beta_2)$		0.7071	-	-	-	-	-	-	-	0.7068***	-
	$cor(\xi, \beta_3)$		-0.7071	-	-	-	-	-	-	-	-0.7068***	-
	Mean LL			-4068.46	-2024.79	-1980.27	-2747.35	-2255.36	-4731.95	-4641.86	-4612.5	-4609.77
	Parameters			17	12	18	9	15	13	19	22	18

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

Table B2: The results of Monte Carlo simulation of the effects of M-endogeneity for parameter estimates under different specifications - mean values of the estimates in 1,000 simulations. Missing variable enters data generating process as an interaction with all attributes.

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0176***	-3.2558	-4.0464**	-3.0402	-4.0244***	-3.604	-4.1216**	-4.0865**	-4.0160***
	SQ (std. dev.)		2.0000	-	-	0.785	-	0.7657	-	0.6254	0.4332	-
	SQ (LV)	α_{12}	-2.0000	-2.0070***	-	-	-	-	-2.5529	-2.8811	-2.6026	-
	SQ (LV , with LV_2)		-2.0000	-	-	-	-	-	-	-	-	-1.9939***
	SQ (LV_2)		-2.0000	-	-	-	-	-	-	-	-	-2.0409**
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0121**	-0.3245	-0.2447	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-1.318	-1.6711	-1.5941	-1.9451**	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.8793	1.1115	0.682	0.9895	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0156***	4.6915*	5.0719**	4.4514	5.0343***	4.3711	5.0619**	5.0430**	5.0175***
	Quality (std. dev.)		2.0000	-	-	1.8662	-	0.8817	-	1.3925	1.2463	-
	Quality (LV)	α_{22}	1.0000	1.0055**	-	-	-	-	1.7644	1.9835	1.2857	-
	Quality (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	0.9973**
	Quality (LV_2)		2.0000	-	-	-	-	-	-	-	-	2.0284**
	Quality (X^{Miss})	α_{23}	2.0000	2.0067***	0.9664	1.1185	-	-	-	-	-	-
	Quality ($\cdot I_1$)		1.0000	-	0.8341	0.8397	1.9173	2.1482	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.5513	-0.5578	-0.0416	-0.0032	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0105***	-2.8016*	-3.0455**	-2.6026	-3.0162***	-2.4928	-3.0305**	-3.0286**	-3.0122***
	Cost (std. dev.)		1.0000	-	-	5.2299	-	3.0591	-	1.8506	1.7777	-
	Cost (LV)	α_{32}	1.0000	0.9996***	-	-	-	-	0.2598	0.2732	1.2634	-
	Cost (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	1.0032***
	Cost (LV_2)		-1.0000	-	-	-	-	-	-	-	-	-1.0069**
	Cost (X^{Miss})	α_{33}	-1.0000	-1.0048***	-1.72	-1.9083	-	-	-	-	-	-

	$Cost\ (\cdot I_1)$		1.0000	-	0.7169	0.8403	-1.2779	-1.3972	-	-	-	-
	$Cost\ (\cdot I_2)$		-1.0000	-	-0.4657	-0.5446	-1.2736	-1.4755	-	-	-	-
Measurement component	I_1 (constant)	α_{41}	-1.0000	-1.0017***	-	-	-	-	-1.0025***	-1.0028***	-1.0033***	-1.0038***
	$I_1\ (LV)$	α_{42}	1.0000	1.0010***	-	-	-	-	1.5819	1.5796	1.6013	0.9980***
	$I_1\ (LV_2)$		1.5000	-	-	-	-	-	-	-	-	1.5061***
	X^{Miss} (in I_1)	α_{44}	1.5000	1.4996***	-	-	-	-	-	-	-	-
	$I_1\ (\eta_1)$	α_{43}	0.5000	0.5002***	-	-	-	-	1.0016	1.0048	0.9702	0.5122***
	I_2 (constant)	α_{51}	1.0000	1.0002***	-	-	-	-	1.0007***	1.0009***	1.0012***	1.0009***
	$I_2\ (LV)$	α_{52}	-1.0000	-1.0001***	-	-	-	-	-1.1165	-1.1198	-1.1098	-0.9964***
	$I_2\ (LV_2)$		-0.5000	-	-	-	-	-	-	-	-	-0.5100***
	X^{Miss} (in I_2)	α_{54}	-0.5000	-0.4998***	-	-	-	-	-	-	-	-
	$I_2\ (\eta_2)$	α_{53}	0.5000	0.4990***	-	-	-	-	0.5054**	0.4975***	0.5192**	0.4994***
	$LV^*\ (X^{SD})$	α_{61}	-1.0000	-1.0012***	-	-	-	-	-0.6983	-0.7001	-0.6827	-1.0070**
	$cor(\xi, \beta_1)$			-	-	-	-	-	-	-	-0.7075	-
	$cor(\xi, \beta_2)$			-	-	-	-	-	-	-	0.7208	-
	$cor(\xi, \beta_3)$			-	-	-	-	-	-	-	-0.7207	-
	Mean LL			-4297.05	-2092.36	-2048.92	-2184.68	-2127.8	-5303.05	-5263.64	-5232.34	-5121.25
	Parameters			18	12	18	9	15	13	19	22	18

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

Appendix C

Table C1: The results of Monte Carlo simulation of the effects of LV-endogeneity for parameter estimates under different specifications - mean values of the estimates in 1,000 simulations. Four measurement equations used in hybrid specifications.

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0148***	-3.4172	-4.0197***	-2.3268	-4.0246**	-3.564	-4.0082***	-3.9960***	-4.0185***
	SQ (std. dev.)		2.0000	-	-	1.1261	-	2.2225	-	1.774	2.0083***	-
	SQ (LV)	α_{12}	-2.0000	-2.0039***	-	-	-	-	-2.9829	-3.1354	-2.0004***	-
	SQ (LV , with LV_2)		-1.6330	-	-	-	-	-	-	-	-	-1.6607**
	SQ (LV_2)		-3.1547	-	-	-	-	-	-	-	-	-3.1521***
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0129**	-2.0674**	-2.2946	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-0.6301	-0.768	-0.7733	-1.3138	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.6381	0.778	0.7619	1.2967	-	-	-	-
	SQ ($\cdot I_3$)		-1.0000	-	-0.1957	-0.2406	-0.2552	-0.4237	-	-	-	-
	SQ ($\cdot I_4$)		1.0000	-	0.1947	0.2386	0.2507	0.4235	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0140***	4.7513*	4.9309**	4.4085	4.9298**	4.9105**	5.0096***	5.0036***	5.0254***
	Quality (LV)	α_{22}	1.0000	1.0041**	-	-	-	-	1.1237	1.0069**	1.0080**	-
	Quality (LV , with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8184**
	Quality (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5805**
	Quality ($\cdot I_1$)		1.0000	-	0.4154	0.3785	0.4844	0.3802	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.4158	-0.3788	-0.4857	-0.3879	-	-	-	-
	Quality ($\cdot I_3$)		0.5000	-	0.132	0.1183	0.1431	0.1092	-	-	-	-
	Quality ($\cdot I_4$)		-0.5000	-	-0.1355	-0.125	-0.1576	-0.1199	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0049***	-2.8368*	-2.9741**	-2.5668	-2.9660**	-2.9080**	-3.0036***	-2.9983***	-3.0113***
	Cost (LV)	α_{32}	1.0000	1.0060**	-	-	-	-	0.899	1.0039***	1.0049***	-

	$Cost$ (LV, with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8163***
	$Cost$ (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5818**
	$Cost$ ($\cdot I_1$)		1.0000	-	0.3538	0.4012	0.2691	0.4058	-	-	-	-
	$Cost$ ($\cdot I_2$)		-1.0000	-	-0.355	-0.402	-0.2626	-0.3949	-	-	-	-
	$Cost$ ($\cdot I_3$)		0.5000	-	0.1115	0.1272	0.0849	0.1281	-	-	-	-
	$Cost$ ($\cdot I_4$)		-0.5000	-	-0.1091	-0.123	-0.0824	-0.1291	-	-	-	-
Measurement component	I_1 (constant)	α_{41}	-1.0000	-1.0003***	-	-	-	-	-1.0008***	-1.0019***	-1.0009***	-1.0005***
	I_1 (LV)	α_{42}	1.0000	1.0001***	-	-	-	-	0.9961***	1.0041***	1.0025***	-
	I_1 (LV, with LV_2)		0.8165	-	-	-	-	-	-	-	-	0.8147***
	I_1 (LV_2)		0.5774	-	-	-	-	-	-	-	-	0.5740**
	I_1 (η_1)	α_{43}	0.5000	0.4993***	-	-	-	-	0.5095**	0.4950**	0.4993***	0.4983***
	I_2 (constant)	α_{51}	1.0000	1.0005***	-	-	-	-	1.0010***	1.0021***	1.0011***	1.0007***
	I_2 (LV)	α_{52}	-1.0000	-0.9999***	-	-	-	-	-0.9956***	-1.0035***	-1.0020***	-
	I_2 (LV, with LV_2)		-0.8165	-	-	-	-	-	-	-	-	-0.8143***
	I_2 (LV_2)		-0.5774	-	-	-	-	-	-	-	-	-0.5736**
	I_2 (η_2)	α_{53}	0.5000	0.4993***	-	-	-	-	0.5101**	0.4956**	0.4999***	0.4988***
	I_3 (constant)	α_{61}	0.0000	-0.0009	-	-	-	-	-0.0011	-0.0017	-0.0012	-0.0010
	I_3 (LV)	α_{62}	0.5000	0.4990***	-	-	-	-	0.4983***	0.5003***	0.5001***	-
	I_3 (LV, with LV_2)		0.4082	-	-	-	-	-	-	-	-	0.4057**
	I_3 (LV_2)		0.2887	-	-	-	-	-	-	-	-	0.2874**
	I_3 (η_3)	α_{63}	0.5000	0.4996***	-	-	-	-	0.5009***	0.4992***	0.4997***	0.4990***
	I_4 (constant)	α_{71}	-1.0000	-0.9995***	-	-	-	-	-0.9993***	-0.9987***	-0.9993***	-0.9995***
	I_4 (LV)	α_{72}	-0.5000	-0.4997***	-	-	-	-	-0.4991***	-0.5011***	-0.5010***	-
	I_4 (LV, with LV_2)		-0.4082	-	-	-	-	-	-	-	-	-0.4064***
	I_4 (LV_2)		-0.2887	-	-	-	-	-	-	-	-	-0.2876**

	$I_4(\eta_4)$	α_{73}	0.5000	0.4993***	-	-	-	-	0.5006***	0.4989***	0.4993***	0.4985***
Structural component	LV^* (X^{SD})		-0.7071	-	-	-	-	-	-0.6545*	-0.6738**	-0.7065***	-
	LV^* (X^{SD} with LV_2)		-1	-	-	-	-	-	-	-	-	-1.0357**
	LV^* (X^{SD} with X^{Miss})	α_{81}	-1	-1.0019***	-	-	-	-	-	-	-	-
	LV^* (X^{Miss})	α_{82}	1	1.0010***	-	-	-	-	-	-	-	-
	$cor(\xi, \beta_1)$		-0.7071	-	-	-	-	-	-	-	-0.7087***	-
	Mean LL			-5528.69	-1928.87	-1909.22	-2178.78	-2045.27	-5996.69	-5971.28	-5954.92	-5949.88
	Parameters			21	16	17	15	16	19	20	21	26

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

Table C2: The results of Monte Carlo simulation of the effects of M-endogeneity for parameter estimates under different specifications - mean values of the estimates in 1,000 simulations. Four measurement equations used in hybrid specifications.

	Variable	Parameter	True value of the parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model9
Utility function	*SQ* (constant)	α_{11}	-4.0000	-4.0161***	-3.4123	-4.0154***	-3.4064	-4.0153***	-3.8850**	-3.9424**	-3.9212**	-4.0141***
	SQ (std. dev.)		2.0000	-	-	1.1451	-	1.1516	-	0.8213	0.7378	-
	SQ (LV)	α_{12}	-2.0000	-2.0101**	-	-	-	-	-2.6202	-2.6212	-2.8252	-
	SQ (LV , with LV_2)		-2.0000	-	-	-	-	-	-	-	-	-2.0159**
	SQ (LV_2)		-2.0000	-	-	-	-	-	-	-	-	-2.0538**
	SQ (X^{Miss})	α_{13}	-2.0000	-2.0104***	-0.2207	-0.2414	-	-	-	-	-	-
	SQ ($\cdot I_1$)		-2.0000	-	-1.5978	-1.8561*	-1.7213	-1.9938**	-	-	-	-
	SQ ($\cdot I_2$)		2.0000	-	0.7674	0.9210	0.7691	0.9246	-	-	-	-
	SQ ($\cdot I_3$)		-1.0000	-	-0.0312	-0.0877	0.0658	0.0180	-	-	-	-
	SQ ($\cdot I_4$)		1.0000	-	0.1184	0.1662	0.0774	0.1218	-	-	-	-
	Quality (constant)	α_{21}	5.0000	5.0182***	4.7593**	4.9145**	4.7577**	4.9144**	4.7944**	4.7998**	4.7899**	5.0116***
	Quality (LV)	α_{22}	1.0000	1.0016***	-	-	-	-	0.837	0.8193	0.8147	-
	Quality (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	1.0040**
	Quality ($\cdot I_1$)		1.0000	-	0.4605	0.3891	0.4618	0.3894	-	-	-	-
	Quality ($\cdot I_2$)		-1.0000	-	-0.4491	-0.4171	-0.4489	-0.4168	-	-	-	-
	Quality ($\cdot I_3$)		0.5000	-	0.4045	0.403	0.4047	0.4031	-	-	-	-
	Quality ($\cdot I_4$)		-0.5000	-	-0.2404	-0.2348	-0.2405	-0.2348	-	-	-	-
	Cost (constant)	α_{31}	-3.0000	-3.0116***	-2.8470*	-2.9682**	-2.8457*	-2.9681**	-2.9013**	-2.9093**	-2.9020**	-3.0094***
	Cost (LV)	α_{32}	1.0000	1.0033***	-	-	-	-	0.8357	0.8443	0.834	-
	Cost (LV , with LV_2)		1.0000	-	-	-	-	-	-	-	-	1.0067**
	Cost ($\cdot I_1$)		1.0000	-	0.3501	0.4178	0.3491	0.4177	-	-	-	-
	Cost ($\cdot I_2$)		-1.0000	-	-0.3852	-0.4307	-0.3851	-0.4309	-	-	-	-

	$Cost\ (\cdot I_3)$		0.5000	-	0.3884	0.4149	0.388	0.4149	-	-	-	-
	$Cost\ (\cdot I_4)$		-0.5000	-	-0.2193	-0.2368	-0.2191	-0.2368	-	-	-	-
Measurement component	I_1 (constant)	α_{41}	-1.0000	-1.0019***	-	-	-	-	-1.0019***	-1.0019***	-1.0025***	-1.0055***
	I_1 (LV)	α_{42}	1.0000	0.9997***	-	-	-	-	1.4817	1.4789	1.4799	1.0014***
	I_1 (LV_2)		1.5000	-	-	-	-	-	-	-	-	1.5209***
	X^{Miss} (in I_1)	α_{44}	1.5000	1.4985***	-	-	-	-	-	-	-	-
	I_1 (η_1)	α_{43}	0.5000	0.4997***	-	-	-	-	1.1387	1.1431	1.1422	0.5084***
	I_2 (constant)	α_{51}	1.0000	1.0003***	-	-	-	-	1.0002***	1.0002***	1.0006***	1.0021***
	I_2 (LV)	α_{52}	-1.0000	-1.0006***	-	-	-	-	-1.1517	-1.1594	-1.1631	-1.0040***
	I_2 (LV_2)		-0.5000	-	-	-	-	-	-	-	-	-0.5108***
	X^{Miss} (in I_2)	α_{54}	-0.5000	-0.5000***	-	-	-	-	-	-	-	-
	I_2 (η_2)	α_{53}	0.5000	0.4988***	-	-	-	-	0.4174	0.3964	0.3852	0.4980***
	I_3 (constant)	α_{61}	0.0000	0.0009	-	-	-	-	0.0011	0.0011	0.0012	0.0022
	I_3 (LV)	α_{62}	0.50000	0.5002***	-	-	-	-	-0.029	-0.0263	-0.0286	0.5012***
	I_3 (LV_2)		-1.0000	-	-	-	-	-	-	-	-	-1.0117***
	X^{Miss} (in I_3)	α_{64}	-1.0000	-0.9990***	-	-	-	-	-	-	-	-
	I_3 (η_3)	α_{63}	0.5000	0.4992***	-	-	-	-	1.2226	1.2227	1.2226	0.5002***
	I_4 (constant)	α_{71}	-1.0000	-0.9996***	-	-	-	-	-0.9997***	-0.9997***	-0.9996***	-1.0000***
	I_4 (LV)	α_{72}	-0.5000	-0.5007***	-	-	-	-	-0.194	-0.1955	-0.194	-0.5022***
	I_4 (LV_2)		0.5000	-	-	-	-	-	-	-	-	0.5047***
	X^{Miss} (in I_4)	α_{74}	0.5000	0.4990***	-	-	-	-	-	-	-	-
	I_4 (η_4)	α_{73}	0.5000	0.4993***	-	-	-	-	0.8433	0.843	0.8433	0.4996***
	LV^* (X^{SD})	α_{81}	-1.0000	-1.0009***	-	-	-	-	-0.7665	-0.7636	-0.7551	-0.9939***
	$cor(\xi, \beta_1)$		-	-	-	-	-	-	-	-	0.3631	-
	Mean LL			-5680.03	-1918.55	-1900.52	-1919.56	-1901.2	-7909.17	-7908.24	-7906.33	-7006.48
	Parameters			26	16	17	15	16	19	20	21	24

*, **, *** indicate $MTL_{0.05}$ at the 0.10, 0.05, 0.01 level, respectively.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl