ENDOGENEITY OF INDICATOR VARIABLES IN HYBRID CHOICE MODELS:
MONTE CARLO INVESTIGATION VS. STATED PREFERENCE STUDY

Wiktor Budziński, Mikołaj Czajkowski, Ewa Zawojska

University of Warsaw, Faculty of Economic Sciences

Supported by:

FNP Foundation for Polish Science

National Science Centre Poland

Outline

- Introduction
  - Choice modeling and unobserved/latent factors
  - Problems with indicator variables
  - Hybrid choice models
- Endogeneity and hybrid choice models
- Monte Carlo simulation
  - General setup
  - Results
- Empirical study
  - Stated preferences and consequentiality
  - Data
  - Results
- Conclusions

Budzinski and Czajkowski (2022). *Endogeneity and measurement bias of the indicator variables in hybrid choice models: A Monte Carlo investigation*

Budzinski, Czajkowski and Zawojska (2022). *Endogeneity of self-reported consequentiality in stated preference studies*

Introduction

Choice modeling and unobserved/latent factors

- In the choice modeling field, we are interested in estimating the preferences of the individual
- Usually, we parametrize the preferences using random utility function
  - For example, multinomial logit model or mixed logit

$$U_{ijt} = \boldsymbol{\beta}_i{}' X_{ijt} + e_{ijt}$$

- Marginal utilities are often allowed to be heterogeneous
  - Unobserved heterogeneity leads to the random parameters or latent class specifications

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda} Z_i + \boldsymbol{\beta}_i^*$$

- Simplified choice scenario:

	Alternative 1	Alternative 2	Alternative 3
Attribute 1	X_{11}	X_{12}	X_{13}
Attribute 2	X_{21}	X_{22}	X_{23}
Attribute 3	X_{31}	X_{32}	X_{33}
Cost	X_{41}	X_{42}	X_{43}
Your choice:	☐	☐	☐

Introduction
Choice modeling and unobserved/latent factors

- In many applications some of the independent variables are not directly observed by the researcher or not easily measurable
- This could be attributes of the choice alternatives
  - In marketing it could be quality of the good
  - In transportation it could be a comfort of the transportation mode
  - In environmental economics it could be water quality
- Or some characteristics of the respondent
  - In stated preference studies it could be survey consequentiality
  - In many applications, risk aversion could also be a relevant factor
- These factors can be necessary to correctly estimate/interpret the model, or they may be of interest by themselves from the policy or research perspective

Introduction
Problems with indicator variables

- If we collect the data with a survey, then we can try to measure relevant factors by asking some attitudinal questions
- For example, if we model individuals' wine choice and want to measure their perception of the quality of the given wine bottle we can ask:
  - "Using a scale from 1 to 5, where 1 means "*I strongly disagree*" and 5 means "*I strongly agree*", please indicate your level of the agreement with the following statement:

    I believe this wine is excellent."
- Analogously, when measuring survey consequentiality we can ask respondents to rate the following statement
  - *"To what extent do you agree with the statement that the results of the survey will influence future policy?"*

Introduction
Problems with indicator variables

- Having such indicator variables put directly into choice model is usually considered to be methodologically flawed
  - Indicators are not direct measures of latent constructs but rather their functions – measurement bias
  - There may be unobserved effects that influence both a respondent's choice and their responses to indicator questions – endogeneity bias
- To resolve these issues, it was proposed to use the so-called Hybrid Choice Models instead

Introduction
Hybrid choice models

- Initially proposed by Ben-Akiva et al. (1999) and Ben-Akiva et al. (2002)
- General framework for more complex choice models
  - Flexible disturbances
  - Explicit modelling of latent psychological factors
  - Latent segmentation for different decision protocols
- Integrated Choice and Latent Variables models focus mostly on the second point
  - Although currently the names are used interchangeably
- Useful when we are interested in the effect of '*soft*' (not objectively measureable) variables, such as perceptions and attitudes, on choices / preferences
  - More '*behavioral*' approach for explaining preference heterogeneity
- We conduct a Monte Carlo simulation to test the claims that this class of models mitigate measurement and endogeneity biases caused by indicator variables

Introduction

Hybrid choice models

Discrete choice model
(e.g. mixed logit model)

Latent variables influence preferences

$$U_{ijt} = \boldsymbol{\beta}_i' X_{ijt} + e_{ijt}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\Lambda} \mathbf{LV}_i + \boldsymbol{\Omega} \mathbf{SD}_i + \boldsymbol{\beta}_i^*$$

$$\mathbf{LV}_i = \boldsymbol{\Psi}' \mathbf{X}_i^{str} + \boldsymbol{\xi}_i$$

Latent variables
(structural equation)

Unobserved psychological factors

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

Measurement equations

Latent variables influence indicators
(e.g. "*To what extent do you agree with the statement that the results of the survey will influence future policy?*"
(from 1 – 'definitely disagree' to 5 – 'definitely agree')

Endogeneity and hybrid choice models

- Measurement error causes endogeneity in and by itself (Walker et al. 2010)
  - We treat it as a separate issue
- Chorus and Kroesen (2014) list possible reasons for endogeneity of latent variables:
  - Missing variables which influence both latent variable and choices of individuals
  - Learning effects
  - Individuals tend to align their attitudes with their actual choices in order to seem consistent
- Hybrid choice models have been used to solve endogeneity issue caused by some observed covariates
  - For example, effect of price when quality is unobserved
  - HCM can be used to impute missing variable
  - <u>We do not study this</u>. In our case indicators are the cause of endogeneity, rather than the solution for it

Endogeneity and hybrid choice models

- We consider two types of indicator variables' endogeneity:
  - LV-endogeneity
    - Latent variable is endogenous in itself
    - Correlated error terms in choice model and structural equations
  - M-endogeneity
    - Indicator variables are endogenous, but latent variable is not
    - Correlated error terms in choice model and measurement equations
- Simulation with 1'000 individuals, 6 choice tasks per person, 3 alternatives per choice task (including the Status Quo)
- 1000 repetitions

Monte Carlo simulation
General setup

	Choice model	
Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$ $SQ_{ijt} \in \{0,1\}$ $Quality_{ijt} \sim U(0,2)$ $Cost_{ijt} \sim U(0,2)$ $e_{ijt} \sim EV_I(0,1)$	$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$ $\beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}\right)$ $X_i^{Miss} \sim N(0,1)$
	Latent variable (structural component)	**Indicator variables (measurement component)**
	$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$ $\xi_i \sim N(0,1)$ $X_i^{SD} \sim N(0,1)$	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$ $\eta_{i1} \sim N(0,1)$ $\eta_{i2} \sim N(0,1)$

Monte Carlo simulation
General setup

	Choice model	
Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$ $SQ_{ijt} \in \{0,1\}$ **Choice situation is defined by 3 attributes** $e_{ijt} \sim EV_I(0,1)$	$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$ $\beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}\right.$ $X_i^{Miss} \sim N(0,1)$
	Latent variable (structural component)	**Indicator variables (measurement component)**
	$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$ $\xi_i \sim N(0,1)$ $X_i^{SD} \sim N(0,1)$	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$ $\eta_{i1} \sim N(0,1)$ $\eta_{i2} \sim N(0,1)$

Monte Carlo simulation
General setup

Choice model

Utility function

$$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$$

$$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$$

$$\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$$

$$\beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}\right)$$

Preference heterogeneity is driven by a latent variable and one observable variable

$$Cost_{ijt} \sim U(0,2)$$

$$X_i^{Miss} \sim N(0,1)$$

$$e_{ijt} \sim EV_I(0,1)$$

Latent variable (structural component)

$$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$$

$$\xi_i \sim N(0,1)$$

$$X_i^{SD} \sim N(0,1)$$

Indicator variables (measurement component)

$$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$$

$$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$$

$$\eta_{i1} \sim N(0,1)$$

$$\eta_{i2} \sim N(0,1)$$

Monte Carlo simulation
General setup

Choice model

Utility function

$$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$$
$$SQ_{ijt} \in \{0,1\}$$
$$Quality_{ijt} \sim U(0,2)$$
$$Cost \sim U(0,2)$$
$$e_{ijt}$$

$$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$$
$$\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$$
$$\beta_{3i} = \exp(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}$$
$$X_i^{Miss} \sim N(0,1)$$

Latent variable is correlated with some observable variables

L
(struct

$$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$$
$$\xi_i \sim N(0,1)$$
$$X_i^{SD} \sim N(0,1)$$

Indicator variables
(measurement component)

$$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$$
$$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$$
$$\eta_{i1} \sim N(0,1)$$
$$\eta_{i2} \sim N(0,1)$$

Monte Carlo simulation
General setup

	Choice model	
Utility function	$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$ $SQ_{ijt} \in \{0,1\}$ $Quality_{ijt} \sim U(0,2)$ $Cost_{ijt} \sim U(0,2)$ $e_{ijt} \sim EV_I(0,1)$	$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$ $\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$ $\beta_{3i} = \exp(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}$ $X_i^{Miss} \sim N(0,1)$
	Latent variable (structural component)	**(measurement component)**
	$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$ $\xi_i \sim N(0,1)$ $X_i^{SD} \sim N(0,1)$	$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$ $I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$ $\eta_{i1} \sim N(0,1)$ $\eta_{i2} \sim N(0,1)$

We assume that instead of LV we observe two indicator variables

Monte Carlo simulation

General setup

We induce endogeneity in the simulation by having a missing variable which affects choices both directly and indirectly (through latent variable)

Choice model

U

$$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$$
$$SQ_{ijt} \in \{0,1\}$$
$$Quality_{ijt} \sim U(0,2)$$
$$Cost_{ijt} \sim U(0,2)$$
$$e_{ijt} \sim EV_I(0,1)$$

$$\beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss}$$
$$\beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss}$$
$$\beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}\right.$$
$$X_i^{Miss} \sim N(0,1)$$

Latent variable (structural component)

$$LV_i^* = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i$$
$$\xi_i \sim N(0,1)$$
$$X_i^{SD} \sim N(0,1)$$

Indicator variables (measurement component)

$$I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1}$$
$$I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2}$$
$$\eta_{i1} \sim N(0,1)$$
$$\eta_{i2} \sim N(0,1)$$

Monte Carlo simulation
General setup

- Estimated models:
  - Base models allow us to check whether simulation works properly, and the extent of a measurement bias:

	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

Monte Carlo simulation
General setup

- Next we analyze the extent of error arising due to:
  - Endogeneity and measurement bias jointly
  - Endogeneity bias and ignoring the preference heterogeneity
  - Endogeneity bias

	Model type	**Measurement error**	**Endogeneity**	**Description**
Model 3	MXL	Yes	Yes	missing variable, random parameters indicator variables entering directly
Model 4	Hybrid MNL	Controlled	Yes	missing variable
Model 5	Hybrid MXL	Controlled	Yes	missing variable, random parameters

- These are models which are most likely to be used by researchers

Monte Carlo simulation
General setup

- In Model 5 we add correlated random parameters to account for unobserved preference heterogeneity caused by omitted variable

$$\begin{cases} V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt} \\ \beta_{1i} = \alpha_{11} + \alpha_{12} LV + \beta_{1i}^{*} \\ \beta_{2i} = \alpha_{21} + \alpha_{22} LV + \beta_{2i}^{*} \\ \beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV + \beta_{3i}^{*}\right) \\ LV_i = \alpha_{61} X_i^{SD} + \xi_i^{**} \end{cases},$$

Monte Carlo simulation
General setup

- Lastly, we propose two different methods to mitigate endogeneity bias:
  - Directly modeling the correlation between latent factor and random parameters
  - Incorporating additional latent variable to account for residual correlation between error terms

	Model type	**Measurement error**	**Endogeneity**	**Description**
Model 6	Hybrid MXL	Controlled	Controlled	missing variable, random parameters, additional correlation
Model 7	Hybrid MNL	Controlled	Controlled	missing variable, additional LV in model specification

Monte Carlo simulation
General setup

- Model 7 utilizes the fact that the two formulations below are equivalent

$$(A)$$

$$\begin{cases} V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt} \\ \qquad \beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13} X_i^{Miss} \\ \qquad \beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23} X_i^{Miss} \\ \qquad \beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33} X_i^{Miss}\right) \\ \qquad LV_i = \alpha_{61} X_i^{SD} + \alpha_{62} X_i^{Miss} + \xi_i \\ I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{43} \eta_{i1i1} \\ I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{53} \eta_{i2} \end{cases}$$

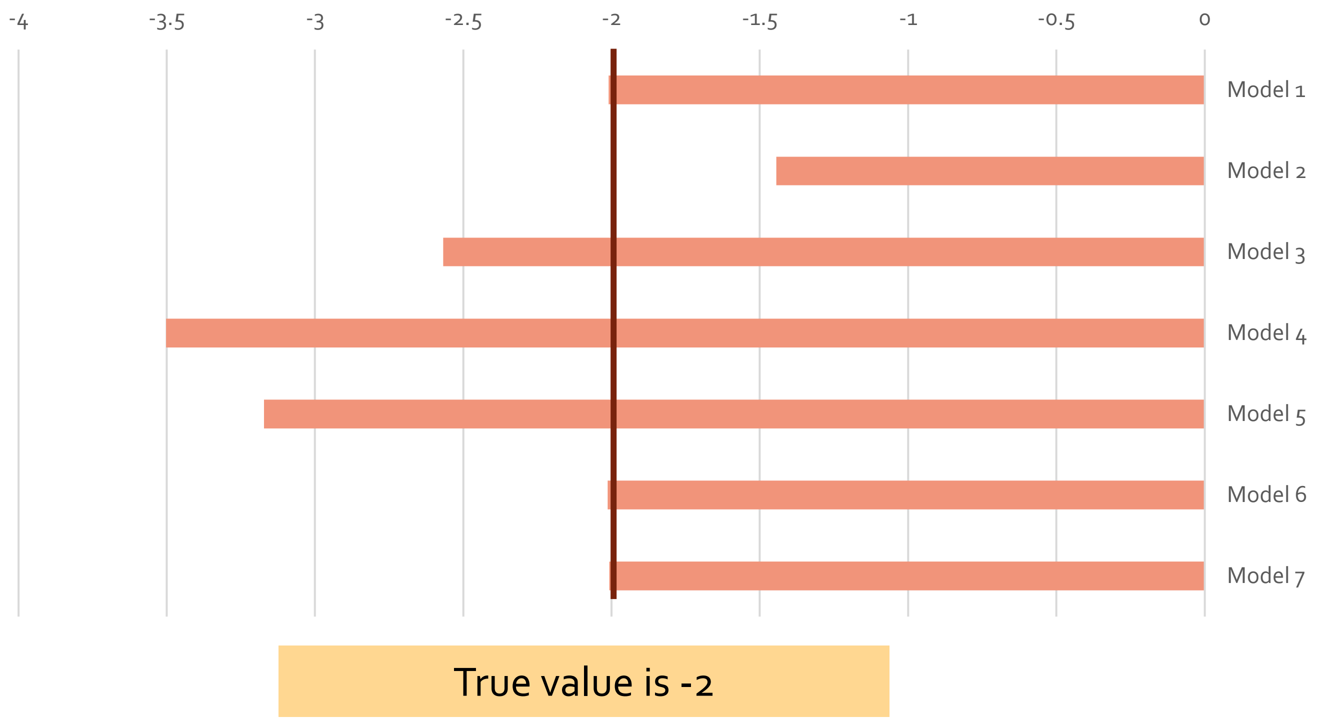
$$(B)$$

$$\begin{cases} V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt} \\ \qquad \beta_{1i} = \alpha_{11} + \alpha_{12} LV_i + \alpha_{13}^{*} X_i^{Miss} \\ \qquad \beta_{2i} = \alpha_{21} + \alpha_{22} LV_i + \alpha_{23}^{*} X_i^{Miss} \\ \qquad \beta_{3i} = \exp\left(\alpha_{31} + \alpha_{32} LV_i + \alpha_{33}^{*} X_i^{Miss}\right) \\ \qquad LV_i = \alpha_{61} X_i^{SD} + \xi_i \\ I_{i1} = \alpha_{41} + \alpha_{42} LV_i + \alpha_{44}^{*} X_i^{Miss} + \alpha_{43} \eta_{i1} \\ I_{i2} = \alpha_{51} + \alpha_{52} LV_i + \alpha_{54}^{*} X_i^{Miss} + \alpha_{53} \eta_{i2} \end{cases}$$

- And then impute additional LV for the missing variable

Monte Carlo simulation
Results

- Interaction with ASC

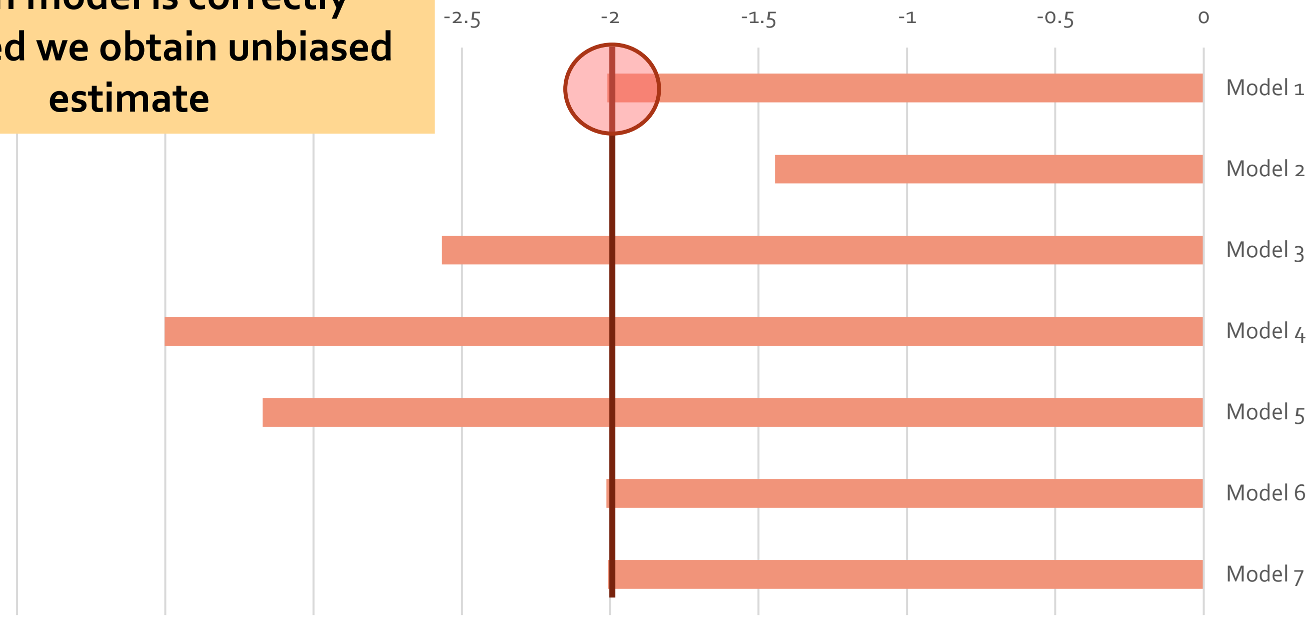


Monte Carlo simulation

Results

- Interaction with ASC

When model is correctly specified we obtain unbiased estimate

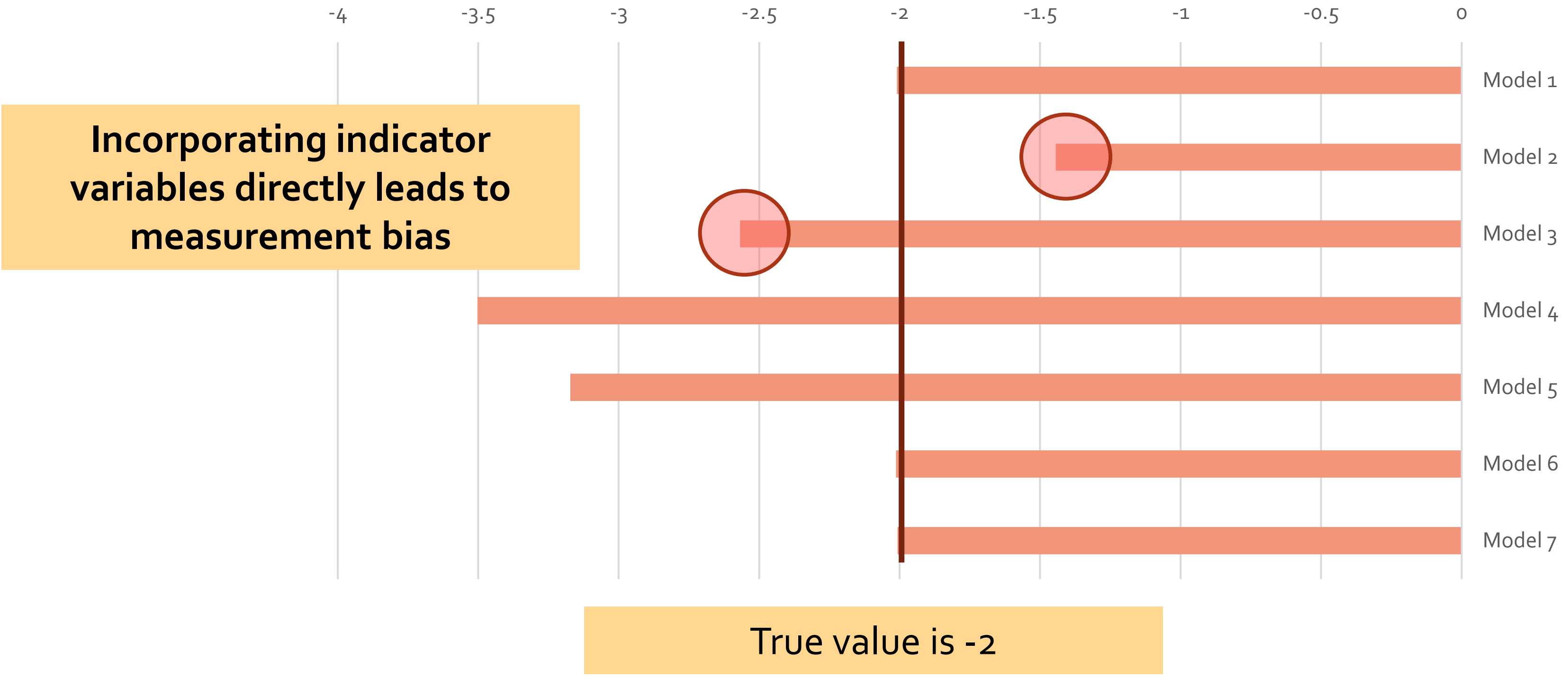


True value is -2

Monte Carlo simulation

Results

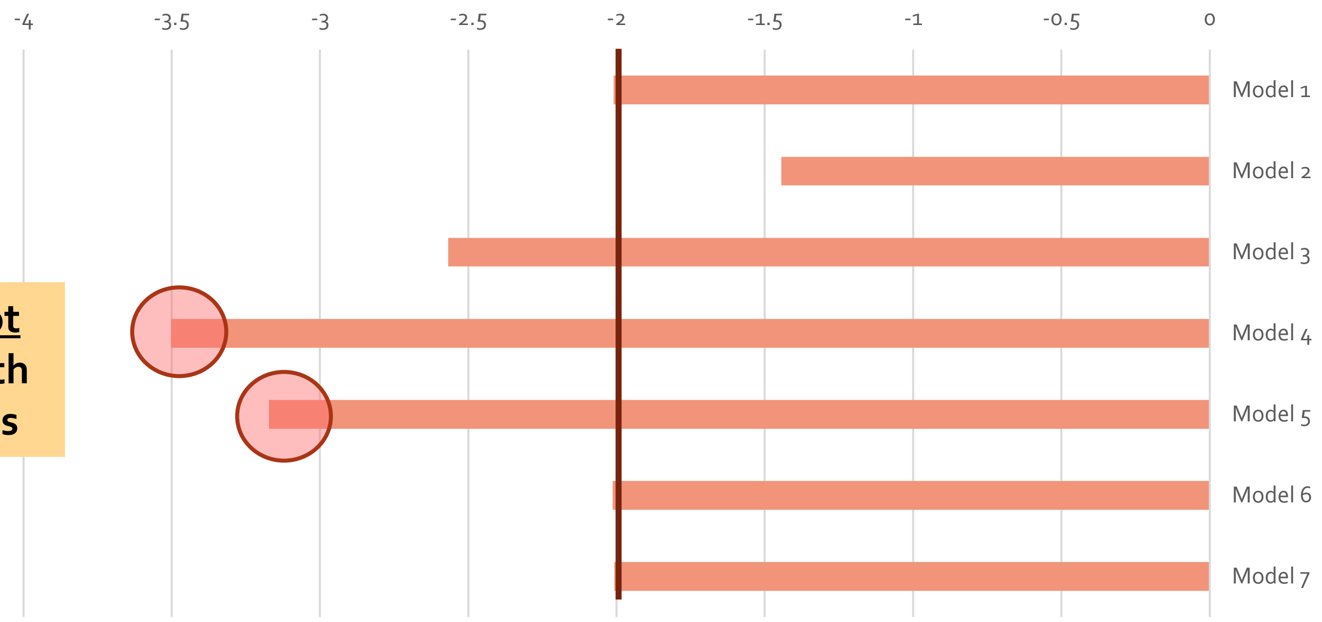
- Interaction with ASC



Monte Carlo simulation

Results

- Interaction with ASC



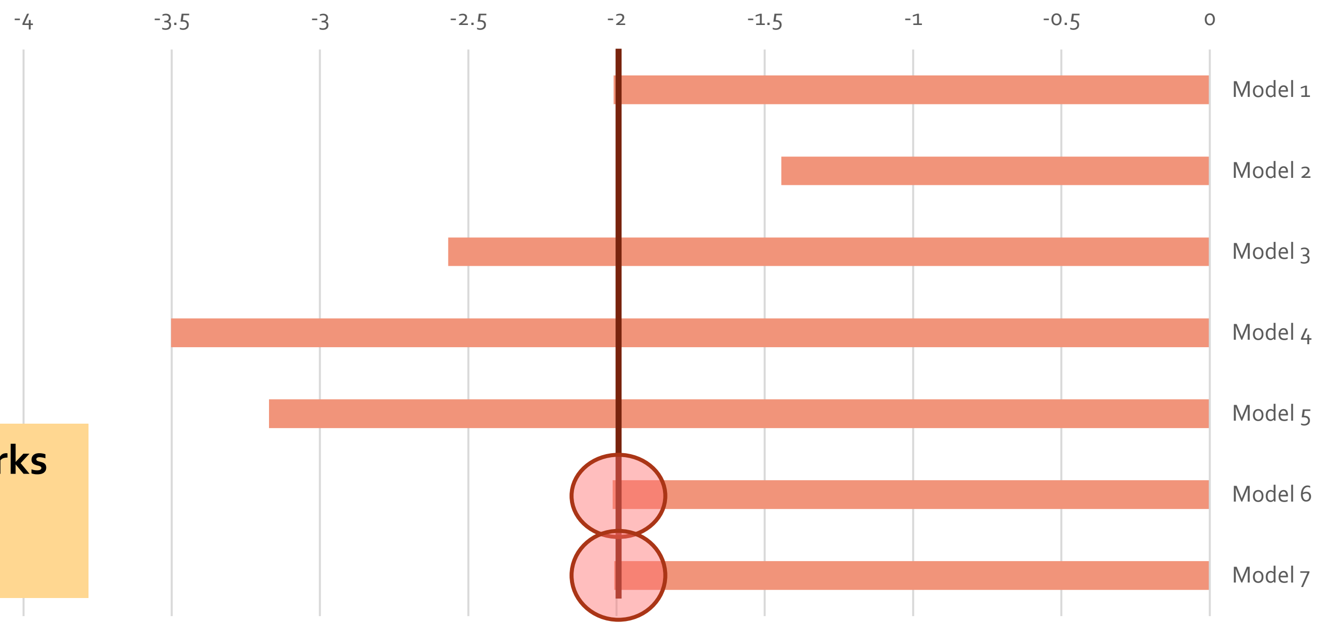
Hybrid choice models <u>do not account for endogeneity</u> with their standard specifications

True value is -2

Monte Carlo simulation

Results

- Interaction with ASC



The proposed solution works as expected and lead to unbiased estimates

True value is -2

Monte Carlo simulation
Results

- We obtain the same results for other coefficients measuring effect of the LV on choices
  - In many cases, for models which do not control for endogeneity, other coefficients are biased as well
- We observe analogous results for the M-endogeneity
  - The exception is that Model 6 no longer recovers true values of the parameters

Monte Carlo simulation
Results

- In stated preference research the main variable of interest is willingness to pay (WTP)
- If we have a following utility function:

$$V_{ijt} = \beta_{1i} SQ_{ijt} + \beta_{2i} Quality_{ijt} - \beta_{3i} Cost_{ijt} + e_{ijt}$$

- Then the WTP for the unit increase of *Quality* will be given by

$$WTP_{2i} = \frac{\beta_{2i}}{\beta_{3i}}$$

- As WTP follows certain distribution, we usually estimate its mean

Monte Carlo simulation

Results

- We can check how the analyzed models perform in terms of recovering mean WTP

	True mean WTP	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Model		**Hybrid MNL**	**MNL**	**MXL**	**Hybrid MNL**	**Hybrid MXL**	**Hybrid MXL**	**Hybrid MNL, 2LV**
Endogeneity		**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes, controlled**	**Yes, controlled**
Measurement error		**No**	**Yes**	**Yes**	**No**	**No**	**No**	**No**
SQ	-7.33	-7.668*	-9.184	-8.34	-3.616	-7.874*	-7.445**	-7.743*
Quality	10.324	10.697**	14.171	11.506	6.601	10.790*	9.979**	10.732**

Monte Carlo simulation
Results

- We can check how the analyzed models perform in terms of recovering mean WTP

	True mean WTP	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Model		Hybrid MNL	MNL				brid XL	Hybrid MNL, 2LV
Endogeneity		No	No				es, controlled	Yes, controlled
Measurement error		No	Yes	Yes	No	No	No	No
SQ	-7.33	-7.668*	-9.184	-8.34	-3.616	-7.874*	-7.445**	-7.743*
Quality	10.324	10.697**	14.171	11.506	6.601	10.790*	9.979**	10.732**

As expected Models 1, 6 and 7 lead to correct estimates of mean WTP

Monte Carlo simulation
Results

- We can check how the analyzed models perform in terms of recovering mean WTP

	True mean WTP	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Model		**Hybrid MNL**	**MNL**	[illegible]	[illegible]	[illegible]	**brid XL**	**Hybrid MNL, 2LV**
Endogeneity		**No**	**No**	**Yes**	**Yes**	**Yes**	**Yes, controlled**	**Yes, controlled**
Measurement error		**No**	**Yes**	**Yes**	**No**	**No**	**No**	**No**
SQ	-7.33	-7.668*	-9.184	-8.34	-3.616	-7.874*	-7.445**	-7.743*
Quality	10.324	10.697**	14.171	11.506	6.601	10.790*	9.979**	10.732**

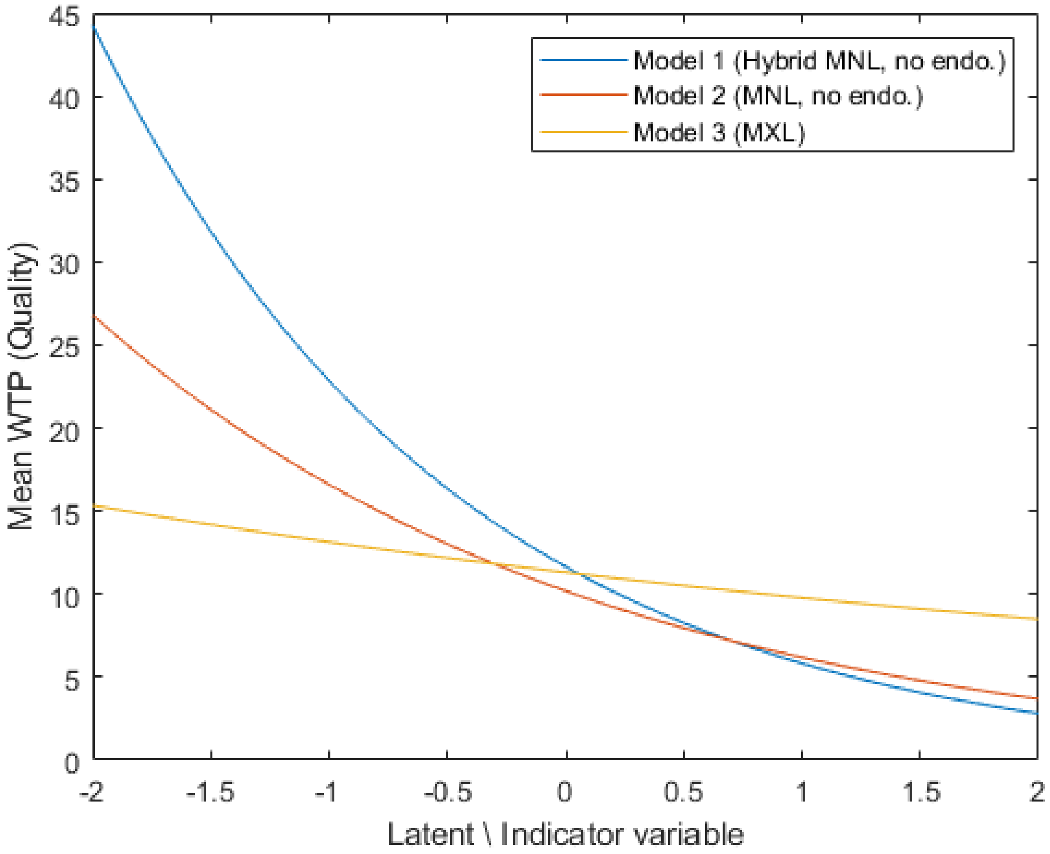
Model 5 is also not far off from true values

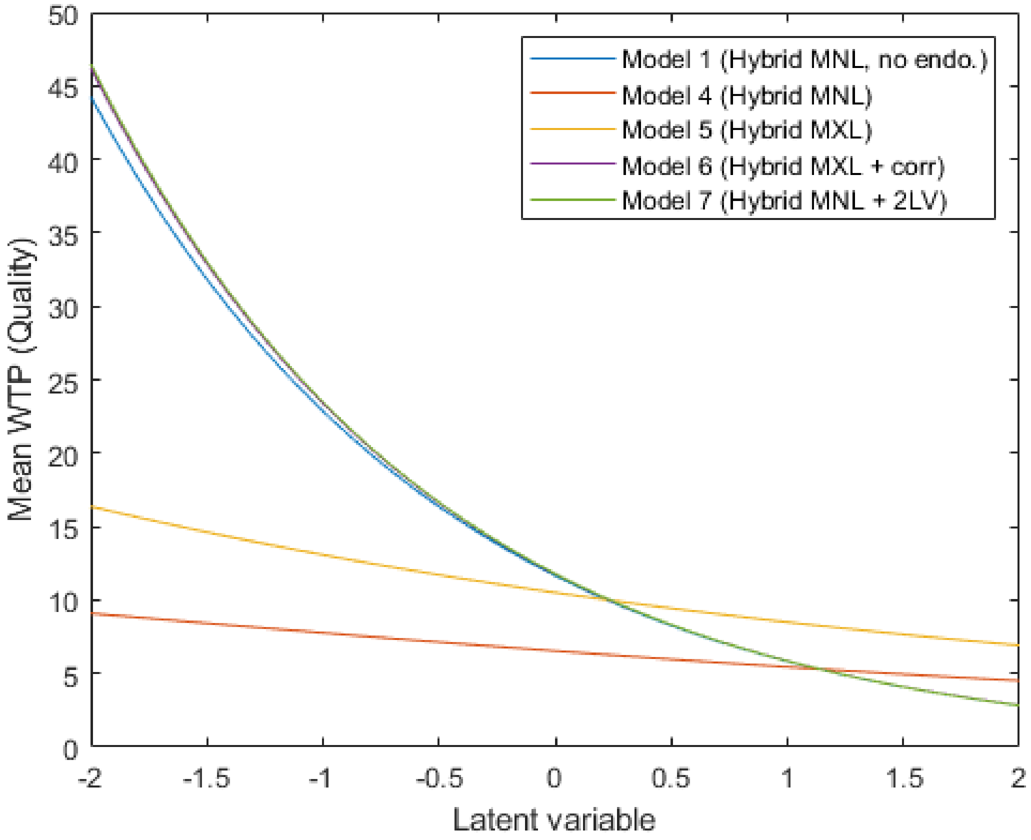
Monte Carlo simulation
Results

- When working with HC models the main goal is usually to find how WTP is affected by a given LV
  - One can actually estimate regular choice model without LVs and indicator variables to obtain mean WTP estimate
  - It would constitute a reduced-form model of HCM
- We can test how well different specifications recover this relationship

Monte Carlo simulation

Results





Simulation results - summary

- Usually used specifications of HCMs do not account for the endogeneity of indicator variables
  - May be sufficient to estimate mean WTP
  - Nonetheless, they do not properly recover the true relationship between WTP and LV
- Measurement bias can be substantial
  - Even with continuous indicator variables
- Possible solutions
  - Allowing for correlation between error terms in structural equations and choice model may help
  - Additional Latent Variables to capture residual correlation
    - Identification may be impossible, particularly with the two-step estimation procedure

Endogeneity of Self-Reported Consequentiality in Stated Preference Studies

Wiktor Budziński, Mikołaj Czajkowski, Ewa Zawojska

Stated preference methods

- Widely used to measure the value of non-market goods, especially public goods
- In transportation, marketing, health, culture, environmental economics, …
- Based on surveys
- Many advantages:
  - Capture use and passive-use values
  - Go beyond the scope of the existing data
- But also important disadvantages:
  - Not based on market behavior
  - Might be viewed as not related to direct consequences
  - Incentive properties insufficiently understood

Conditions for truthful preference disclosure
(Carson and Groves 2007; Carson et al. 2014; Vossler et al. 2012)

One of the conditions requires the survey consequentiality

A necessary condition for truthful preference disclosure:

Consequentiality

- "a survey's results are seen by the agent as potentially influencing an agency's actions and the agent cares about the outcomes of those actions" (Carson and Groves 2007)
- "an individual faces or perceives a nonzero probability that their responses will influence decisions related to the outcome in question and they will be required to pay for that outcome" (*Contemporary Guidance for Stated Preference Studies*, Johnston et al. 2017)

policy consequentiality

payment consequentiality

Other dimensions of consequentiality?

E.g., pivotality?

Endogeneity of consequentiality perceptions

explored in previous studies

- Herriges et al. (2010) – an exogenous information treatment and a Bayesian treatment-effect model; importance of controlling for endogeneity
- No significant problem of endogeneity especially in studies using socio-demographics as instruments:
  - Vossler et al. (2012) – a generalized method of moments over-identification test
  - Interis and Petrolia (2014) – a two-step instrumental variable probit model
- Groothuis et al. (2017) – a bivariate probit approach; perceived consequentiality found to be endogenous; unobserved factors strengthen the consequentiality and decrease the likelihood of voting for the program
- Lloyd-Smith et al. (2019) – a special multi-step estimator for a scaled probit model; importance of controlling for endogeneity; with no endogeneity control, perceived consequentiality affects voting behavior, but the effect disappears with the special regressor

Endogeneity of consequentiality perceptions

explored in previous studies

- Herriges et al. (2010) – an exogenous information treatment and a Bayesian treatment-effect model; importance of controlling for endogeneity
- No significant problem of endogenei demographics as instruments:
  - Vossler et al. (2012) – a generalized m
  - Interis and Petrolia (2014) – a two-ste
- Groothuis et al. (2017) – a bivariate p found to be endogenous; unobserve decrease the likelihood of voting for
- Lloyd-Smith et al. (2019) – a special multi-step estimator for a scaled probit model; importance of controlling for endogeneity; with no endogeneity control, perceived consequentiality affects voting behavior, but the effect disappears with the special regressor

Limitations:

- Little evidence – very few studies
- Mixed evidence
- Step-wise procedures
- Do not account for the measurement error
- Single indicator (measurement) questions for consequentiality

Endogeneity control in hybrid choice models

Budziński and Czajkowski (2022)

Model 1

- Standard hybrid choice models do not resolve endogeneity

Model 2

- Solutions:
  - Directly modeling the correlation between latent variables and random parameters

Model 3

  - Adding a latent variable to capture the correlation caused by missing covariates

- **These models correspond to Models 5, 6 and 7 from the simulation**

Measurement equations
(ordered probit)

Latent variables influence self-reports about beliefs in survey consequentiality

Latent variables

Unobserved beliefs about survey consequentiality

Discrete choice model
(interactions in the mixed logit model)

Latent variables influence stated preferences

Discrete choice experiment

- Public-good scenario: Extension of public theater offer in Poland (a number of shows)
- 4 choice tasks per person; CAWI; a representative sample of 2,863 residents of Poland

	Variant A	**Variant B** No changes	**Attribute levels**
Entertainment theaters	+ 25%	no change	+ 25%, + 50%, no change
Drama theaters	+ 50%	no change	
Children's theaters	no change	no change	
Experimental theaters	+ 50%	no change	
Annual cost for you (tax)	50 PLN	0 PLN	5, 10, 20, 50 PLN
Your choice	☐	☐	

Consequentiality elicitation

- Randomized statements assessed on a Likert scale with seven levels: from 'definitely disagree' to 'definitely agree' + don't know
- Used in the measurement $\rightarrow$ 9 ordered probit models as measurement equations

I think that …
[1] … by participating in this survey, I will have influence on the future theater offer.
[2] … the results of this survey will determine if to change the theater offer.
[3] … the results of this survey will be used to decide if to change the theater offer.
[4] … if the theater offer is decided to be changed, the results of this survey will be used to decide which type of shows will be played more and less.
[5] … the increase of the theater offer as described in this survey is possible to be implemented.
[6] … a decision to expand the theater offer will indeed result in more shows and premiers, as described in this survey.
[7] … a decision to expand the theater offer will indeed result in higher (tax) fees, which will increase my household expenditures, as described in this survey.
[8] … I am one of many people participating in this survey, so my responses do not have a chance to affect the survey final results.
[9] … a decision whether to change the theater offer will be taken independently of the survey results.

Results

Measurement equations
(ordered probit)

Latent variables influence self-reports about beliefs in survey consequentiality

Latent variables

Unobserved beliefs about survey consequentiality

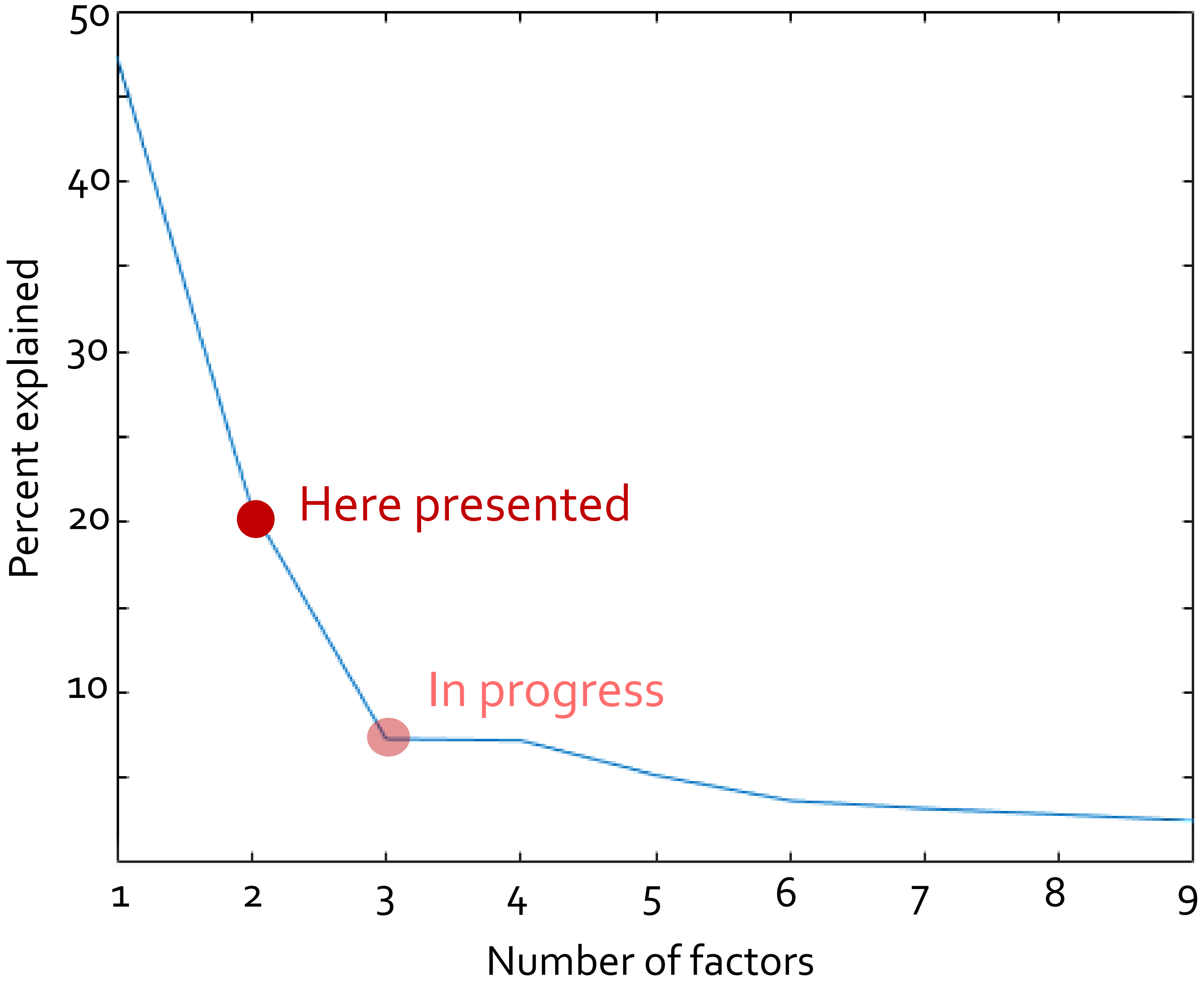
Discrete choice model
(interactions in the mixed logit model)

Latent variables influence stated preferences

	Model 1	Model 2	Model 3
	Standard	Corr. LVs and random parameters	+ 1 LV

How many latent variables to include?

How many dimensions of consequentiality do we have?



Results

Measurement equations
(ordered probit)

Latent variables influence self-reports about beliefs in survey consequentiality

Latent variables

Unobserved beliefs about survey consequentiality

Discrete choice model
(interactions in the mixed logit model)

Latent variables influence stated preferences

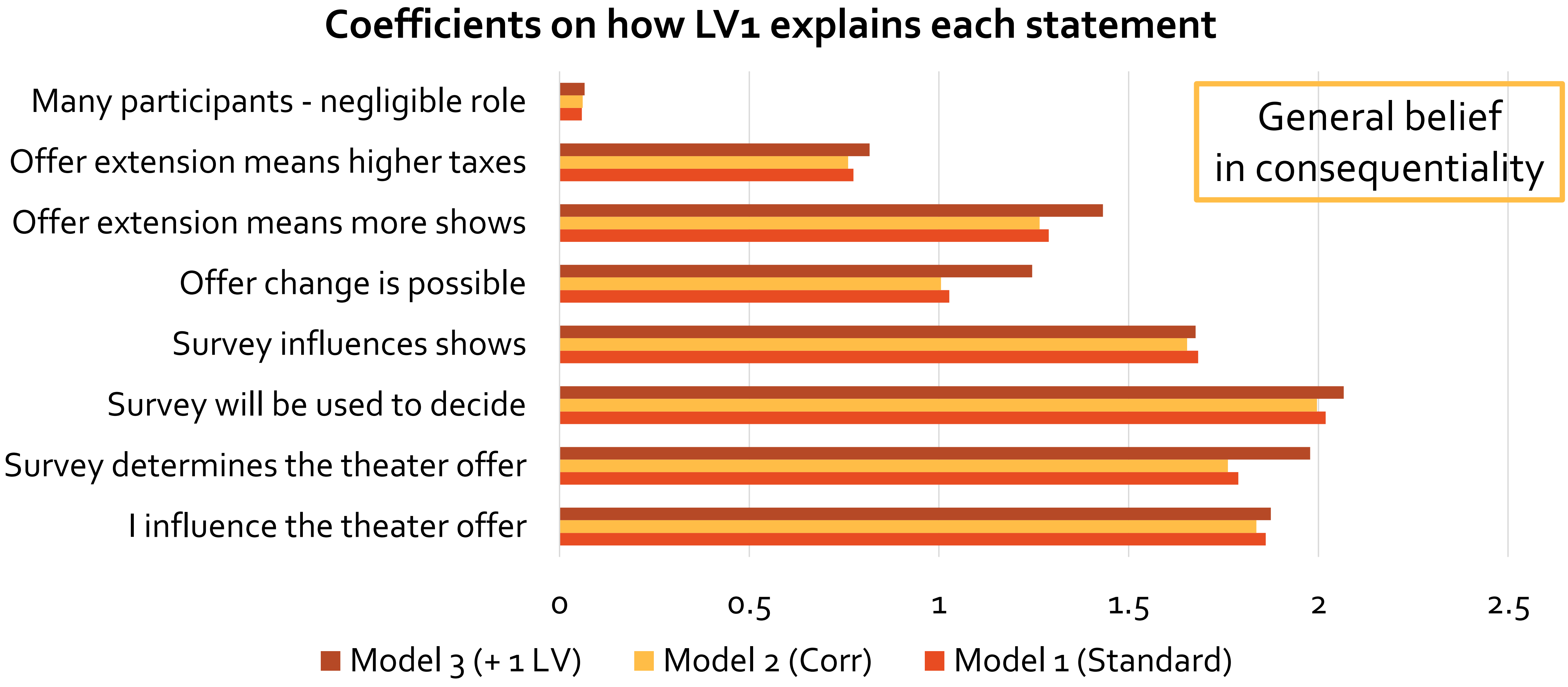
	Model 1	Model 2	Model 3
	Standard	**Corr. LVs and random parameters**	**+ 1 LV**
LL	-38,620.1	-38,564.6	-38,465.4
AIC/n	6.764	6.756	6.739

better → even better

- Responses to consequentiality statements are explained with latent variables
- Two latent variables (LVs) expressing perceived consequentiality:
  - General belief in consequentiality
  - Lack of belief in pivotality

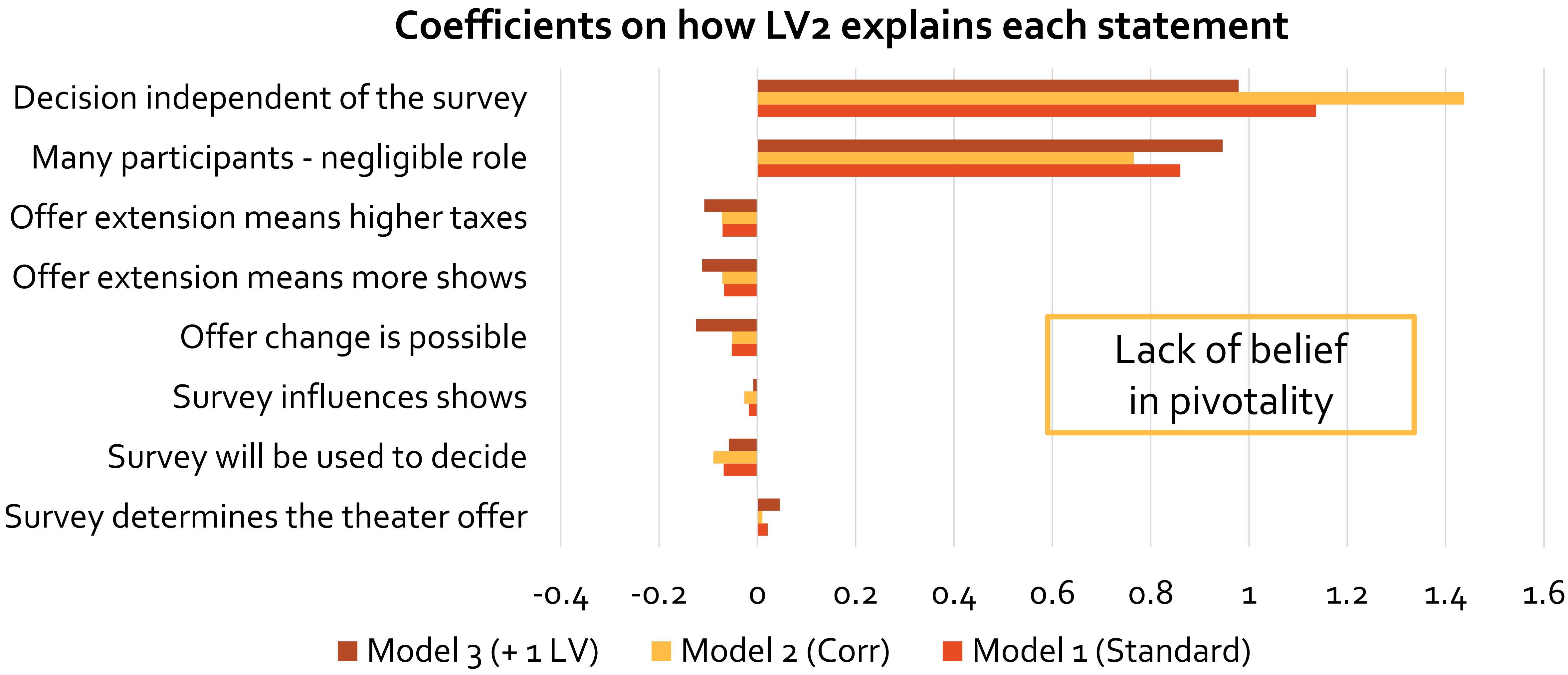
Results: Measurement equations

Ordered probits



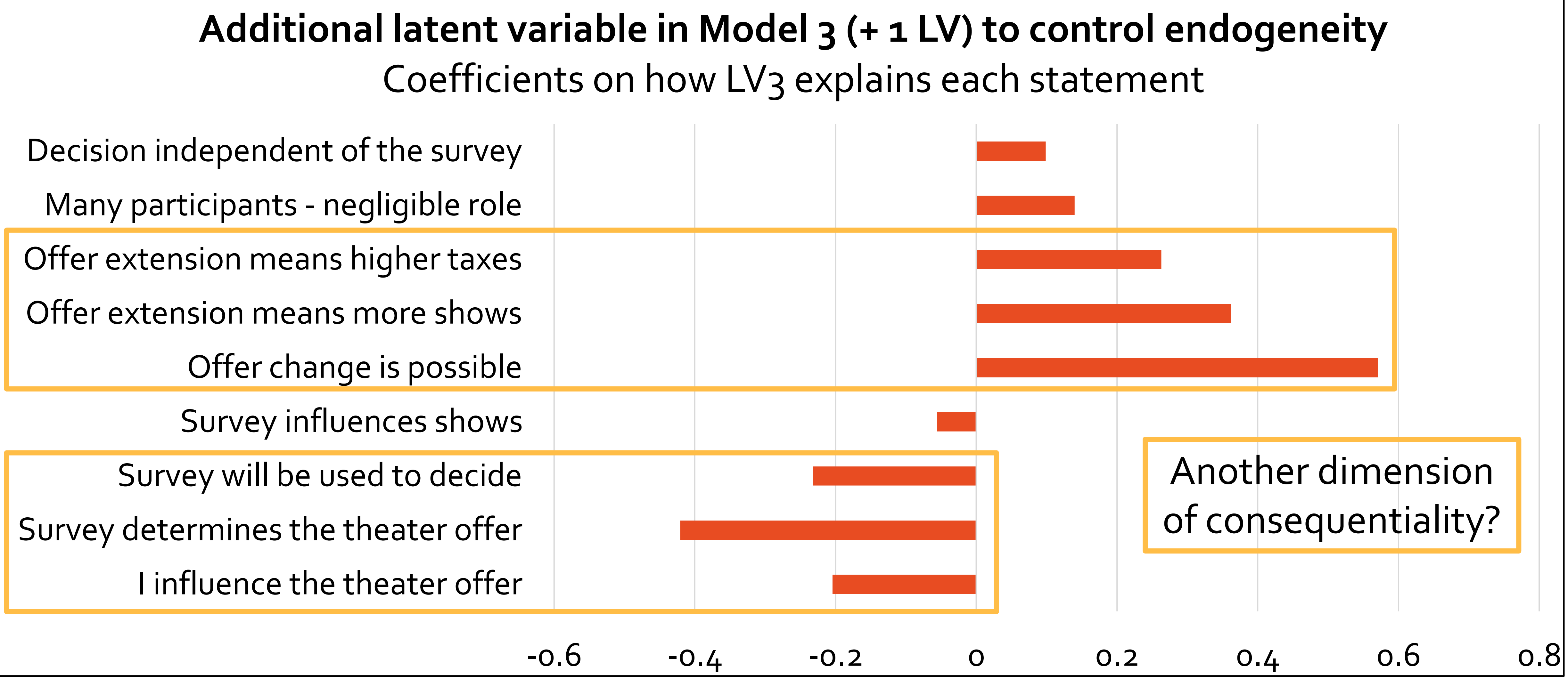
Results: Measurement equations

Ordered probits



Results: Measurement equations

Ordered probits



Results

Measurement equations
(ordered probit)

Latent variables influence self-reports about beliefs in survey consequentiality

Latent variables

Unobserved beliefs about survey consequentiality

Discrete choice model
(interactions in the mixed logit model)

Latent variables influence stated preferences

	Model 1	Model 2	Model 3
	Standard	**Corr. LVs and random parameters**	**+ 1 LV**
LL	-38,620.1	-38,564.6	-38,465.4
BIC/n	6.834	6.835	6.819

better

even better

- Two latent variables (LVs) expressing perceived consequentiality:
  - General belief in consequentiality
  - Lack of belief in pivotality

Results: Discrete choice component

Mixed logits with means interacted with LVs

Mean coefficient estimates

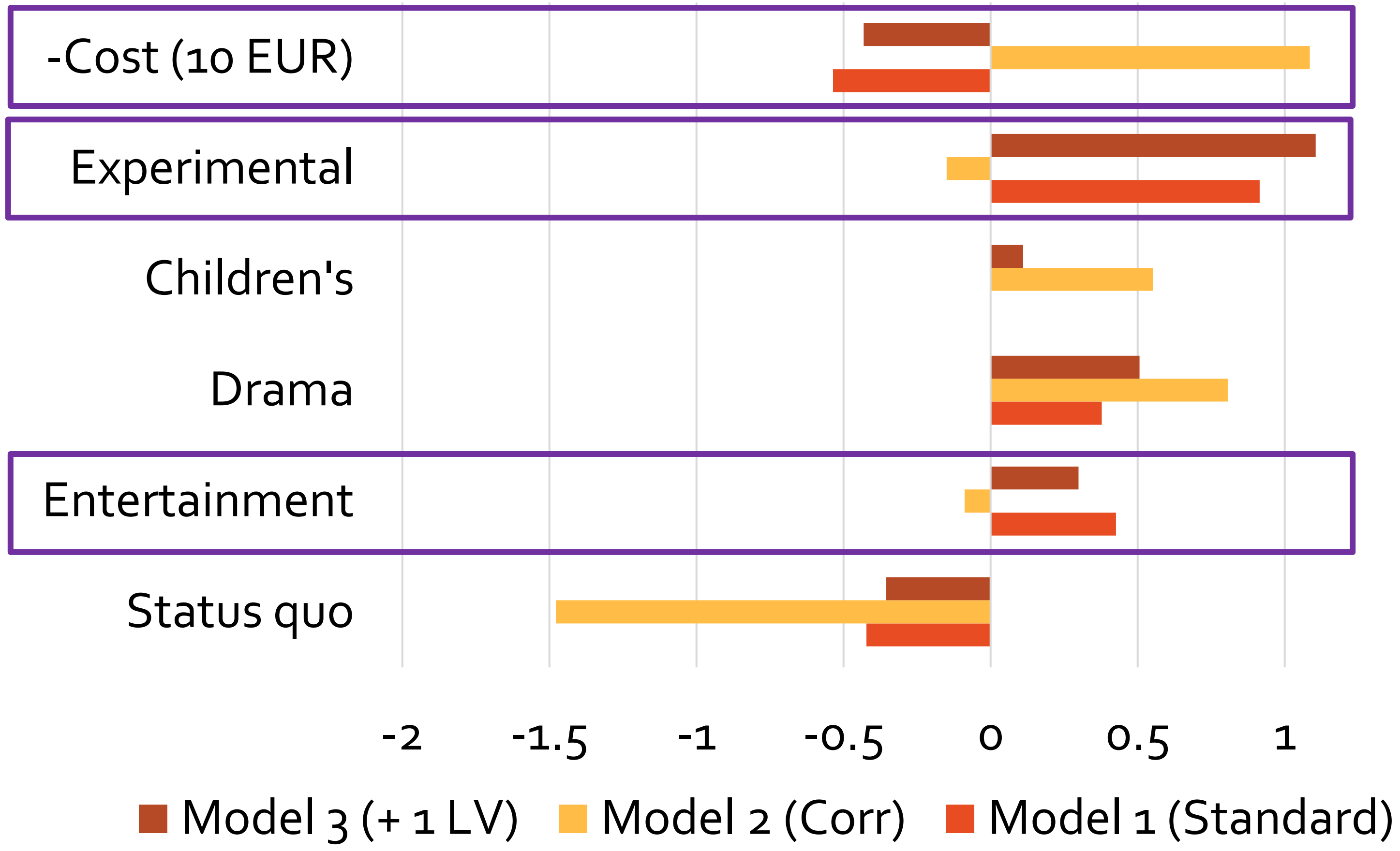
	Model 1	Model 2	Model 3
	Standard	Corr. LVs and random parameters	+ 1 LV
Status quo	0.4719***	0.4459***	0.4711***
Entertainment	0.8926***	0.999***	0.9151***
Drama	0.5769**	0.464*	0.4259
Children's	0.1364	0.1099	0.0443
Experimental	-0.4336	-0.502*	-0.409
– Cost (10 EUR)	3.7752***	3.8161***	3.6282***

- Preference parameters are random
- For all, standard deviations are (highly) significant
- Mean coefficient estimates are similar across models

Results: Discrete choice component

Mixed logits with means interacted with LVs

Coefficients of interactions of means with LV1 (general consequentiality)

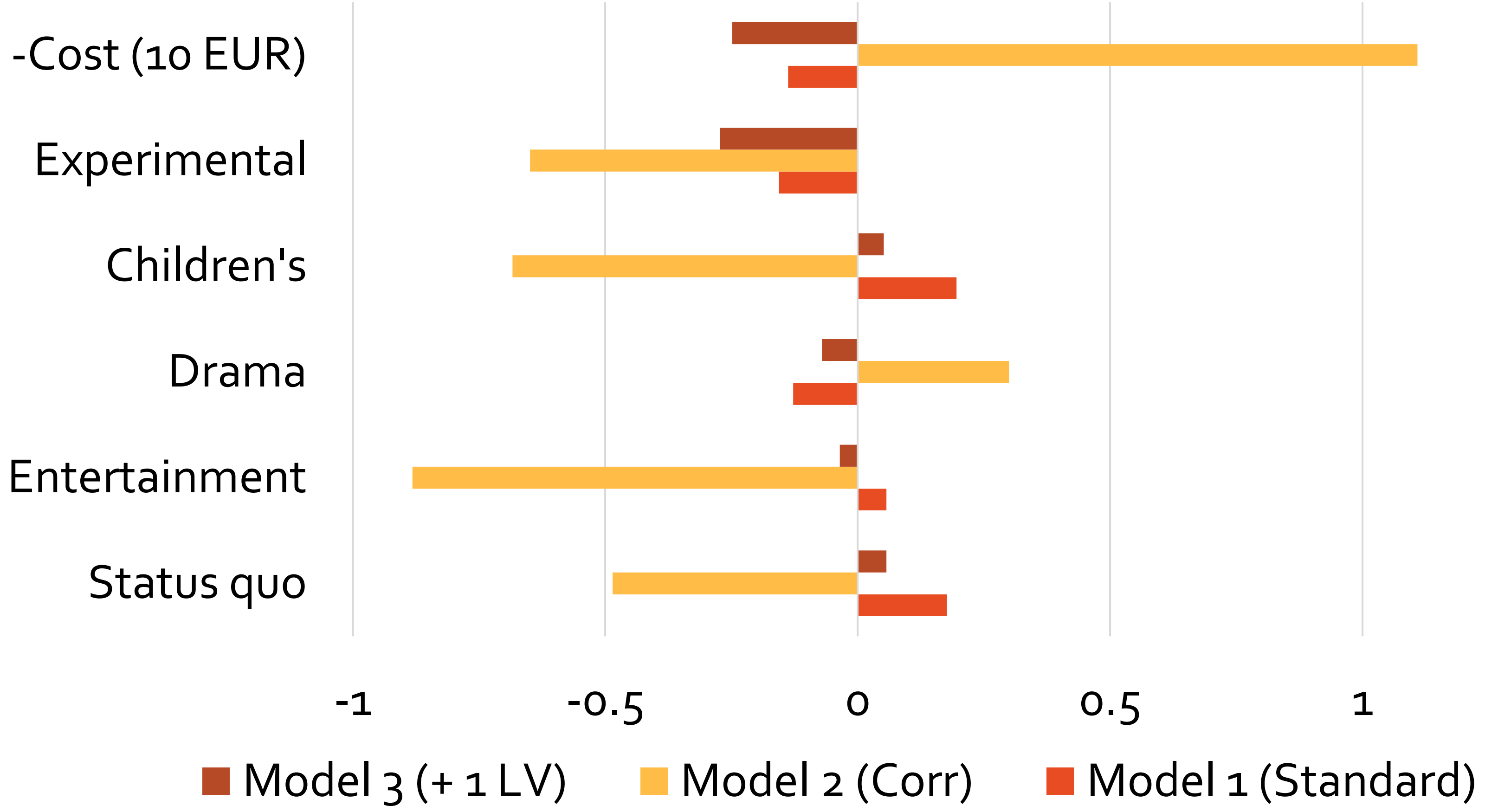


- Model 2 (Corr) leads to identification of quite different effects from Model 1
- Model 3 (+1 LV) seems to result in very similar results to Model 1
  - Maybe another consequentiality dimension? Does not fully account for residual correlation

Results: Discrete choice component

Mixed logits with means interacted with LVs

Coefficients of interactions of means with LV2 (pivotality)

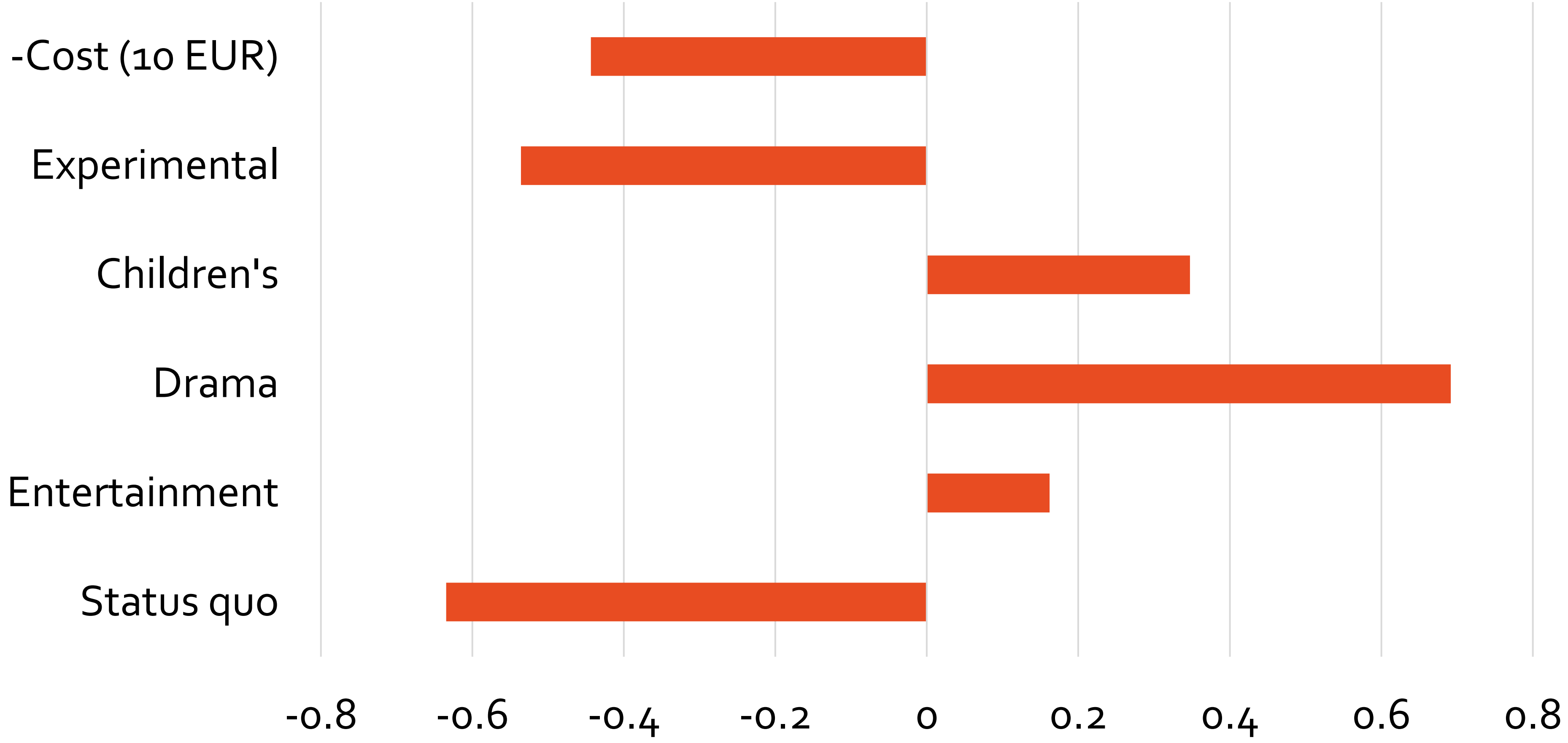


- Similar findings
- Endogeneity control in Model 2 matters for many attributes
- In Model 3, maybe another latent factor is needed?

Results: Discrete choice component

Mixed logits with means interacted with LVs

Coefficients of interactions of means with LV3



Closing thoughts

- Accounting for endogeneity matters
- The proposed solutions works well when we have well defined latent constructs
- No theory regarding dimensions of consequentiality (or other attitudes captured)
  - This could guide designing indicator questions to elicit respondents' perceptions
  - No construct validity
    - Theory does not predict what effect on preferences we should expect
- Maybe use some algorithm to find proper specification, similar to:
  - Paz, Alexander, Cristian Arteaga, and Carlos Cobos. "*Specification of mixed logit models assisted by an optimization framework.*" Journal of choice modelling 30 (2019): 50-60.
- Some problems with the interpretation of the additional LV
- Maybe the issue is in the lack of variables in the structural equations for the latent factors

THANK YOU!

Wiktor Budziński, Mikołaj Czajkowski, Ewa Zawojska

University of Warsaw, Faculty of Economic Sciences