Validity of capability estimates obtained through structural equation modelling in a multidimensional setting - an agent-based simulation exercise

Jaya Krishnakumar (University of Geneva)
Joint paper with Florian Chávez-Juárez (LNPP-CIDE, México)

Research Seminar
Madras School of Economics
Chennai, February 2020

UNIVERSITÉ DE GENÈVE
CIDE

Overview

- How do we evaluate the performance of structural equation models regarding capability 'measures' when we do not observe the true capabilities?
- In Krishnakumar and Chávez-Juárez (2016) we simulated an **artificial world** using **agent-based** modelling techniques, thus generating the true capabilities along with other variables needed for the evaluation exercise.
- However, this study only concerned a single dimension (education).
- In this study, we enhance this first attempt by:
  1. moving from 1 to 3 dimensions: education, health, social relations
  2. bringing the simulated world closer to reality by using real world data to feed the model
- Our results show that the SEM approach yields very good estimates of capabilities (correlation with true capabilities above 0.8), thus reconfirming our previous result in a multidimensional setting
- We also find that the performance of SEM is better than that of individual single-dimensional models.

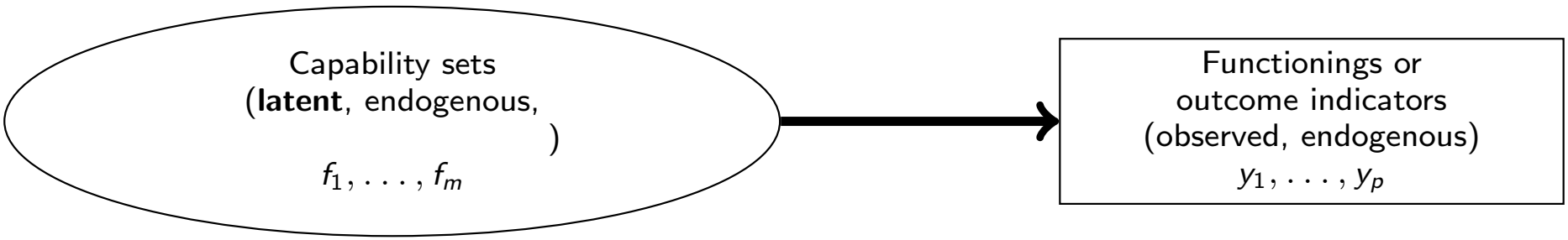
Outline

The key issue

- One of the key challenges of the Capability Approach is its operationalisation: **how can we get empirical measures of capabilities?**
- **Difficulty:** Capabilities are hard to observe, and hence there are no 'true' values to which any estimates/measures can be compared to test their validity.
- **Our approach:** We simulate a world with various agents and heterogeneous behaviours. This enables us to obtain data on both observable and unobservable variables (*viz.* capabilities). We can then evaluate any technique against the *true* (generated) capability values.
- **Our techniques:**
  - Agent-based models to simulate the world
  - Structural equations models to estimate capabilities

The latent variable approach and the CA

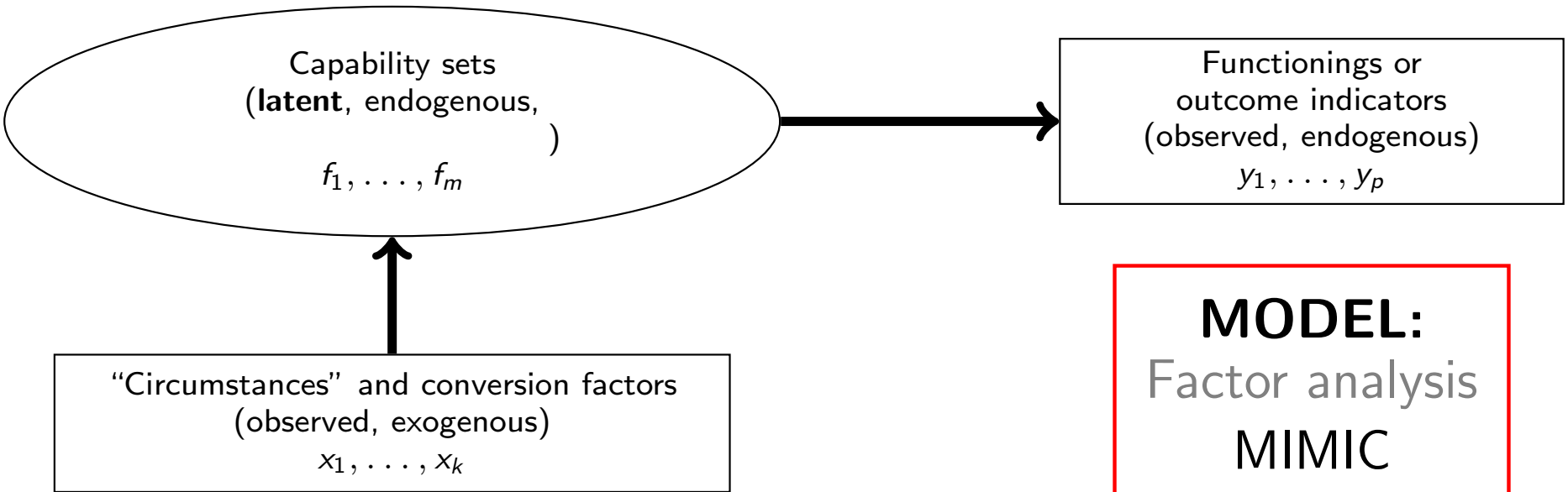
Figure: CA in a diagram



MODEL:
Factor analysis
MIMIC
SEM

The latent variable approach and the CA

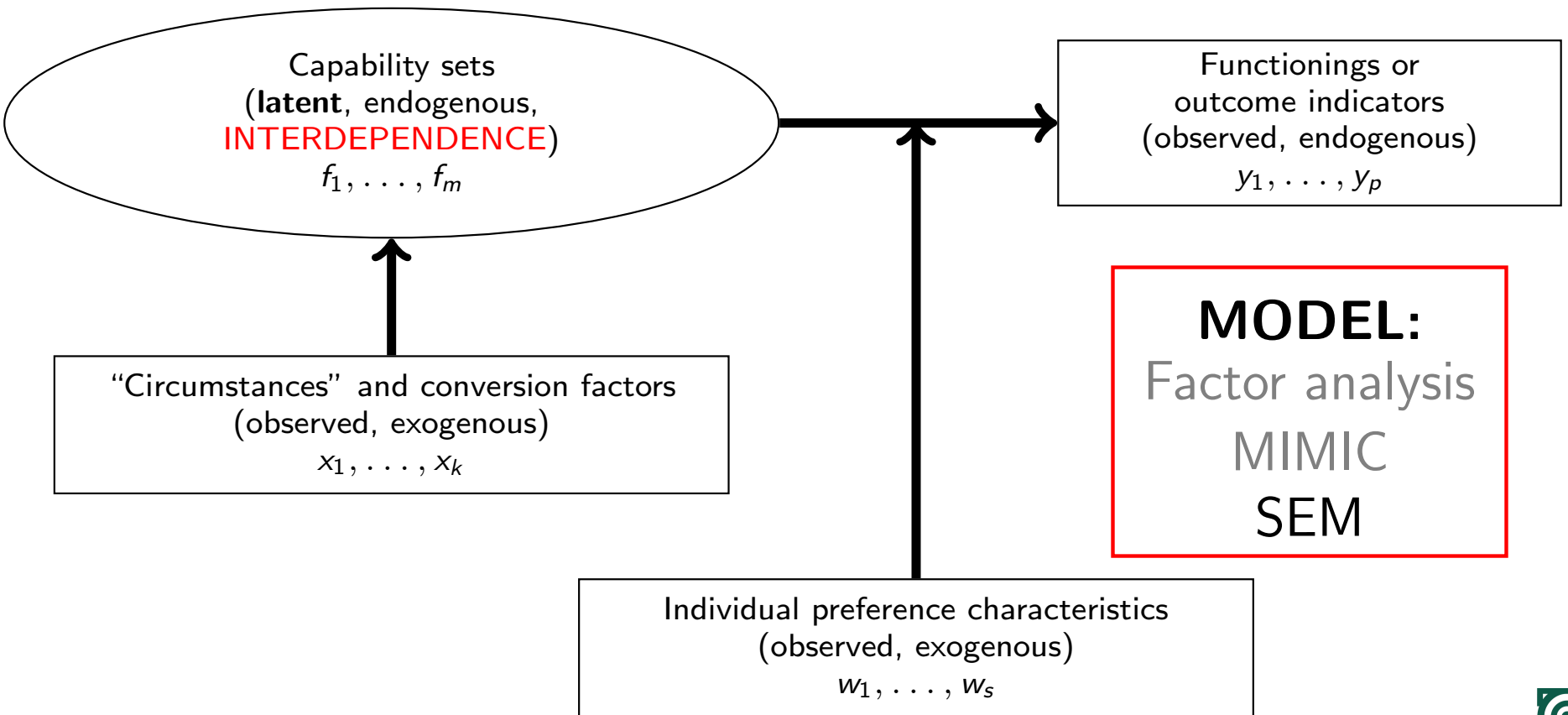
Figure: CA in a diagram



MODEL:
Factor analysis
MIMIC
SEM

The latent variable approach and the CA

Figure: CA in a diagram



Key features of our model

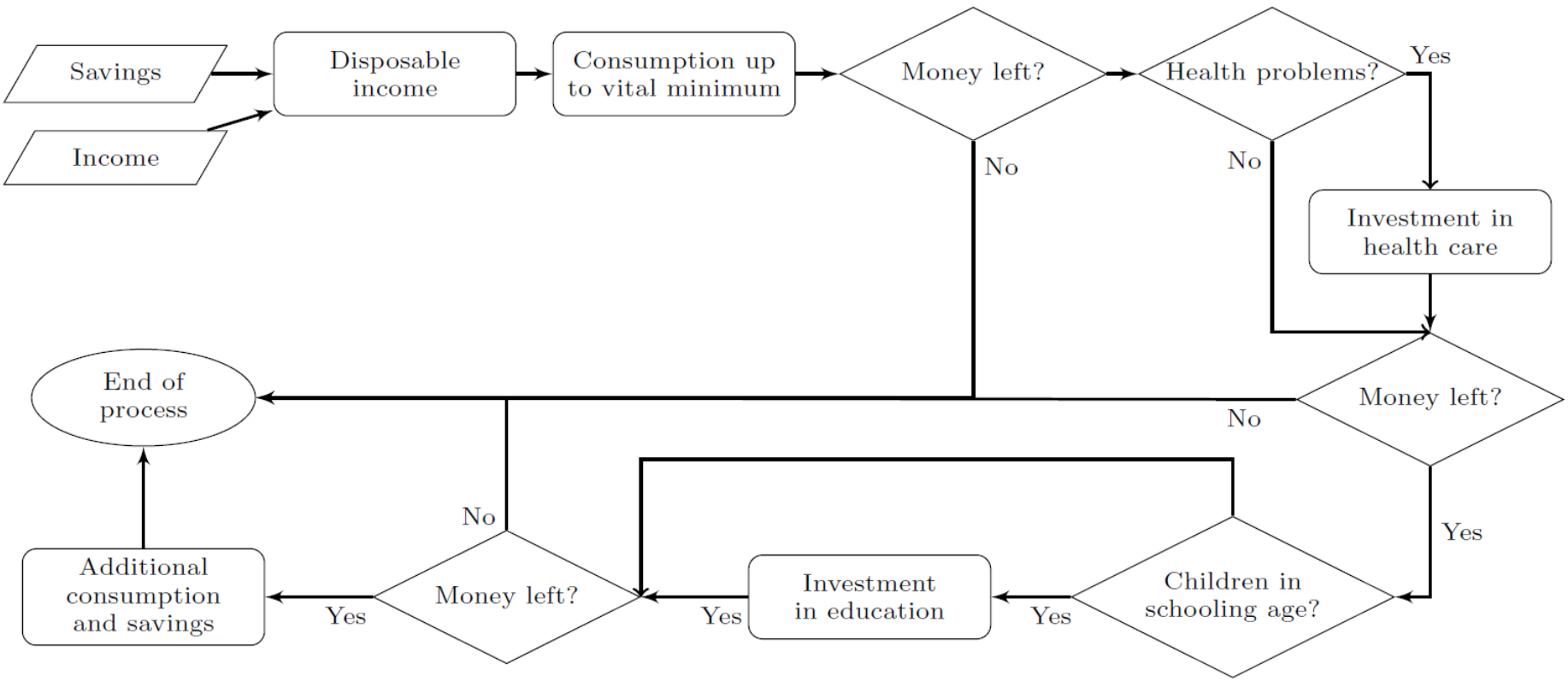
Agent Based Model

- Initialised and calibrated with survey data from Mexico
  - We use 4 types of localities (Capital city, Large city, Small city and Village) with characteristics estimated using data (e.g. density of schools)
  - Survey data on education, type of education, IQ, family structure etc.
- Focus on three dimensions: education, health and social networks
- Once initialised, in each period a series of processes take place, generating endogenously all values of the model.
- Several processes (next slide) are directly calibrated on data:
  - Fertility rates estimated based on age and education level of the mother
  - Health shocks: partially calibrated with survey data
  - Partner search: calibrated with survey data
  - Wage level estimated through modified Mincer equations and heterogeneous across locality types.

Main processes taking place each period

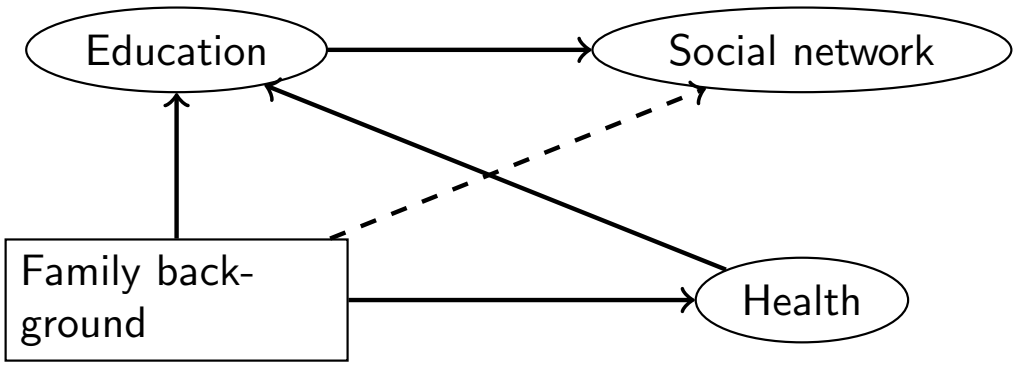
1	Procreation based on empirical evidence
2	Health shocks affecting individual with a calibrated probablity.
3	For each enrolled child we determine if he/she **passes the grade** or not.
4	Mating process for all eligible individuals
5	Families receive their labour income (based on data)
6	The **family decides how to use the resources**. They can spend it for health recovery, investment in schooling of children, consumption or save it for the next period. The decision making process is based on a priority approach rather than a utility maximisation.
7	The **social network** of each individual is built at school, where grade-mates might enter the social network and the longer they stay together at school, the more likely the relationship will last for their whole life.

Diagram of priority approach



Relationship among dimensions in our model

According to the processes implemented in the model, we might expect to have the following relationships across dimensions:



- Family background directly affects the ability to invest in health and education: family background $\rightarrow$ health and education
- Passing grades at school depends on the health status health $\rightarrow$ education
- The social network is formed at school: education $\rightarrow$ social network

Data export and simulation strategy

- We run our **simulation 100 times**, generating 100 different populations
- For each simulation, we consider the information of all individuals "born during the simulation" at the **age of 25 years**
- Our variables combine average values of their youth (e.g. average family income) and values observed at the age of 25 years (e.g. size of social network at that age).
- We also include background variables such as:
  - Gender
  - IQ
  - Parental education
  - Average income and average savings of the family
  - etc.

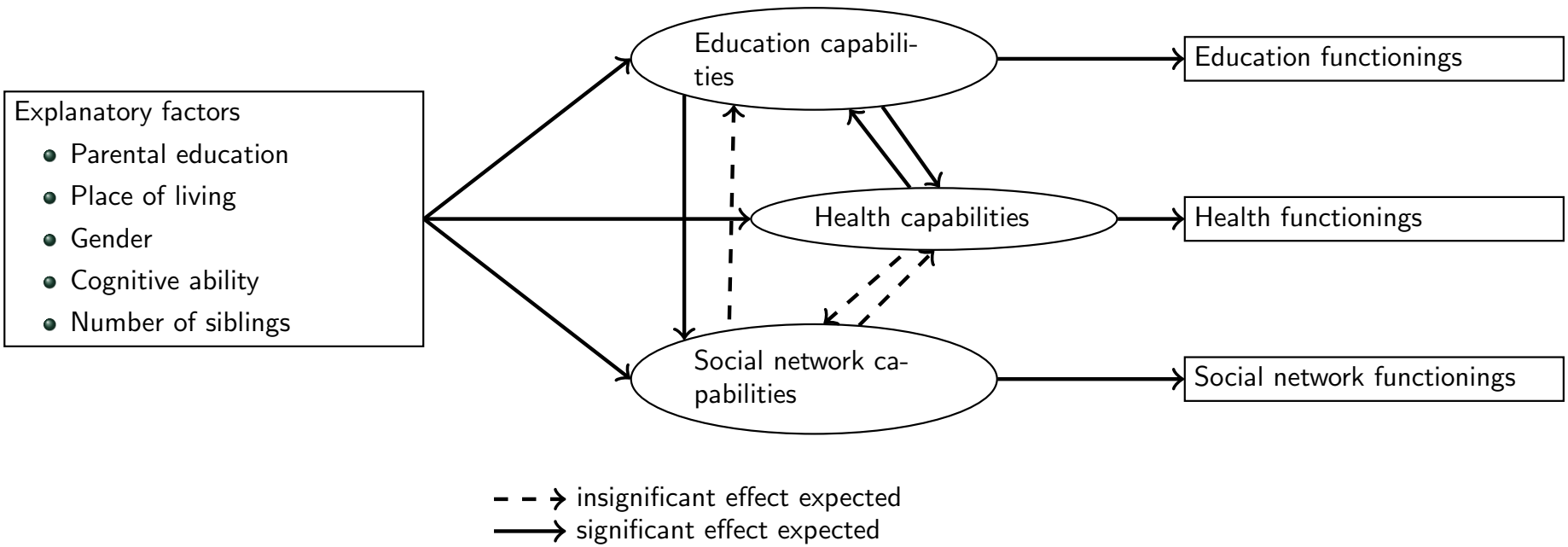
Data: functionings

Education	**Health**	**Social capital**
Years of education	Number of health shocks	Number of people in the social network (at age 25)
Acquired human capital	Average size of health shocks	Average years of education in social network
Percentage of schooling years in private schools	Average health status	Number of people with high school or more in social network
Average school quality	Health status at the age of 25	

Data: capabilities

Dim.	Potential outcome	Capabilities Choice approach
Education	Counter-factual number of years of education	Forced transfer to public school
	Counter-factual human capital	
Health	Counter-factual health stock	Full recovery possible?
Social network	Counter-factual social network (All in)	Number of possible additions
		Number of lost contacts
	Counter-factual social network (No out)	

Our econometric model



The same using equations

The simple model can be written as:

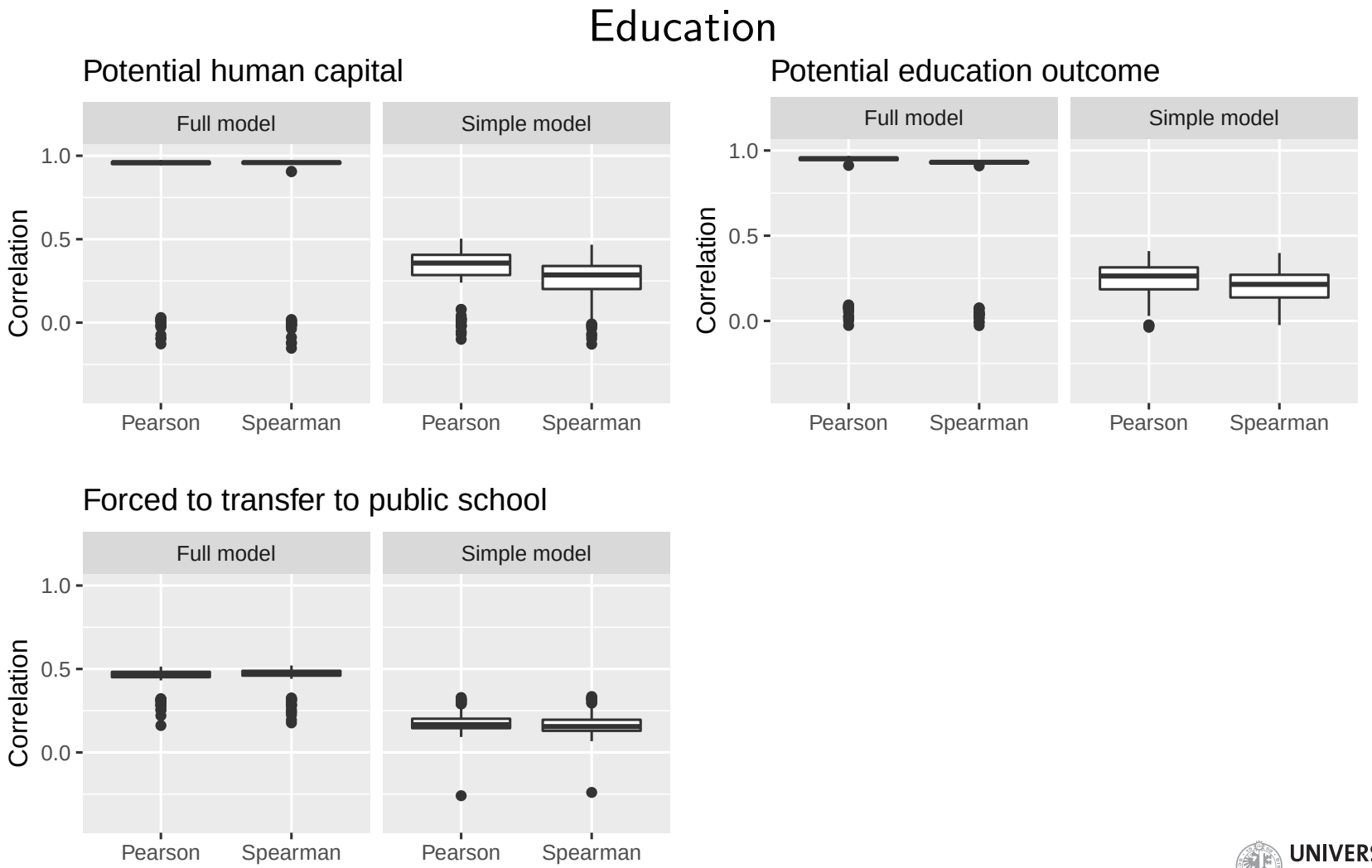
$$\begin{cases} F_d &= \Gamma c_d + v \\ c_d &= X_d \beta + u \end{cases}$$

The full model is given by:

$$\begin{cases} f_e &= \gamma_e' c_e + v_e \\ f_h &= \gamma_h' c_h + v_h \\ f_s &= \gamma_s' c_s + v_s \\ c_e &= w' \alpha_e + x_e' \beta_e + \theta_{he} c_h + \theta_{se} c_s + u_e \\ c_h &= w' \alpha_h + x_h' \beta_d + \theta_{eh} c_e + \theta_{sh} c_s + u_h \\ c_s &= w' \alpha_s + x_s' \beta_s + \theta_{hs} c_h + \theta_{es} c_e + u_s \end{cases}$$

To obtain the capabilities, we use the factor scores
$\hat{c_d} = (I + \Lambda' \Psi^{-1} \Lambda)^{-1} X_d \beta + (I + \Lambda' \Psi^{-1} \Lambda)^{-1}) \Lambda' \Psi^{-1} (F_d - \Gamma w)$
where $V(v) = \Psi$, $V(u) = \sigma^2 I$ and $\Omega = (\Psi + \Lambda \Lambda')$.

Rank correlation between true and estimated capabilities - Education

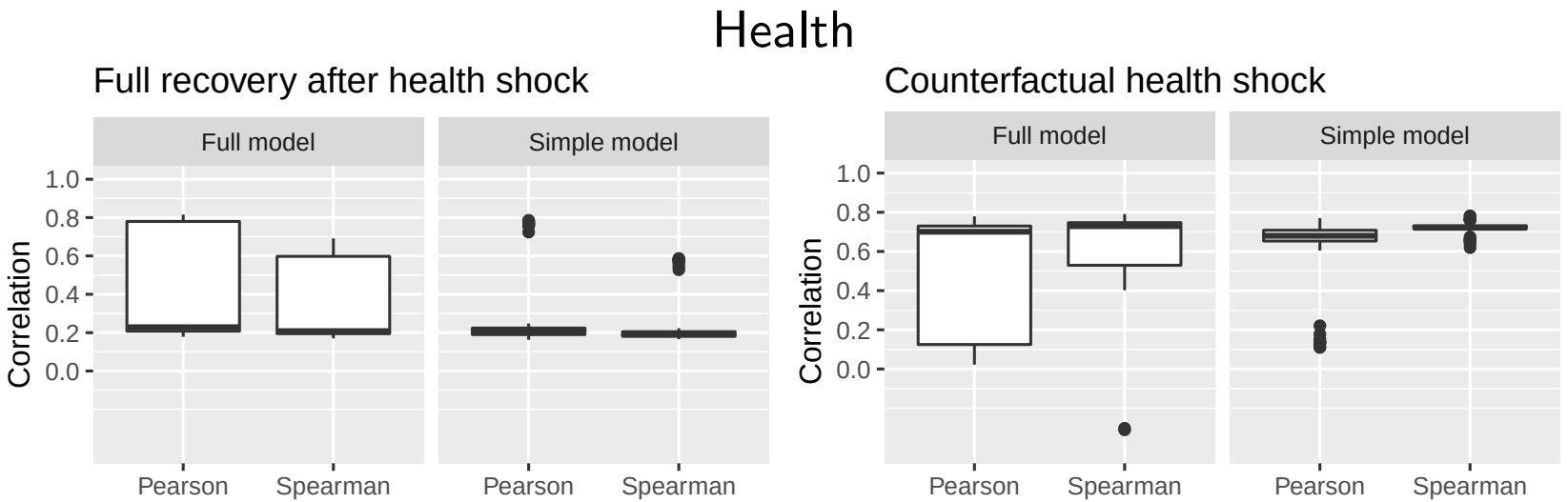


Education results

Education dimension results:

- Linear correlations and rank correlations are almost identical
- Simultaneous model results better than single-dimensional models
- Results better for potential outcome approach than choice approach

Rank correlation between true and estimated capabilities - Health



Health Results

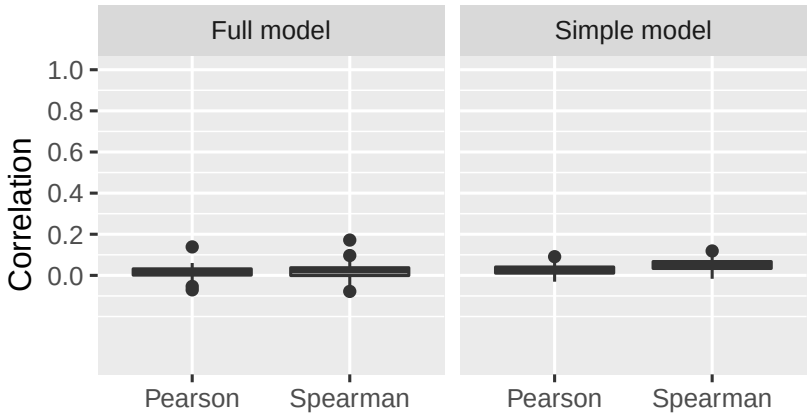
Health dimension results

- Correlations slightly lower than for education
- For choice approach, the full model results better than those of single-dimensional models, but it is the other way around for the potential outcome approach
- Correlation results better for potential outcome approach than choice approach

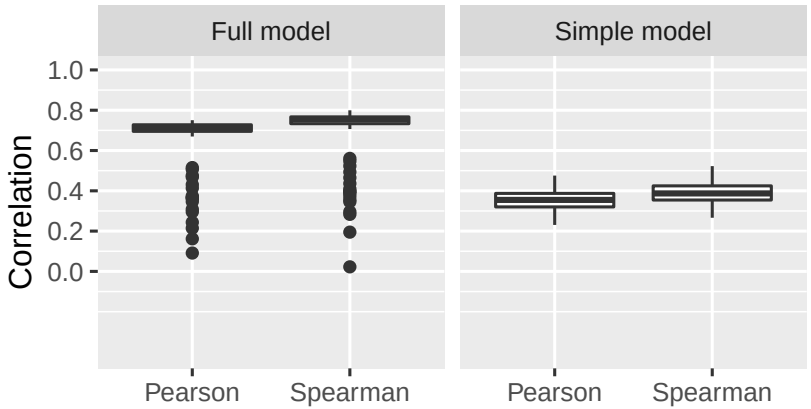
Rank correlation between true and estimated capabilities - Social Network

Social network / social capital

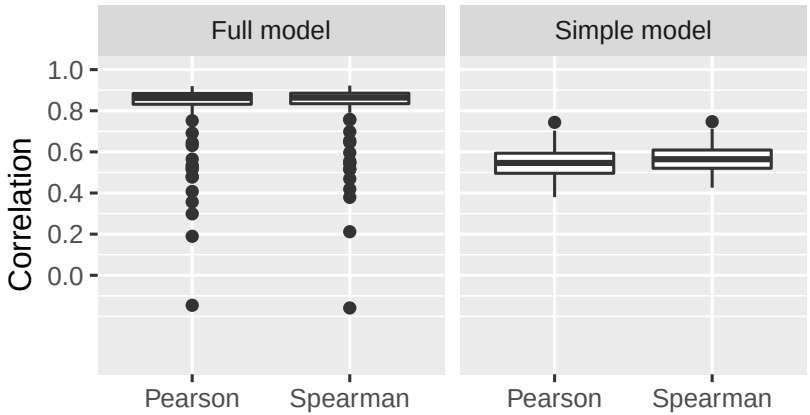
Number of lost links in SN



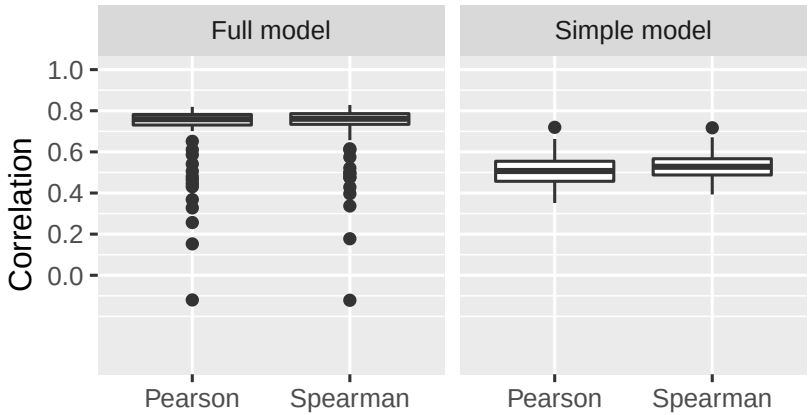
Number of possible additions to SN



Potential SN: all in



Potential SN: no out



SITÉ ÈVE CIDE

Social Network results

Social Network dimension results

- High correlations like for education
- Simultaneous model results better than single-dimensional models in most cases
- Results better for potential outcome approach than choice approach

Further Discussion of our research findings I

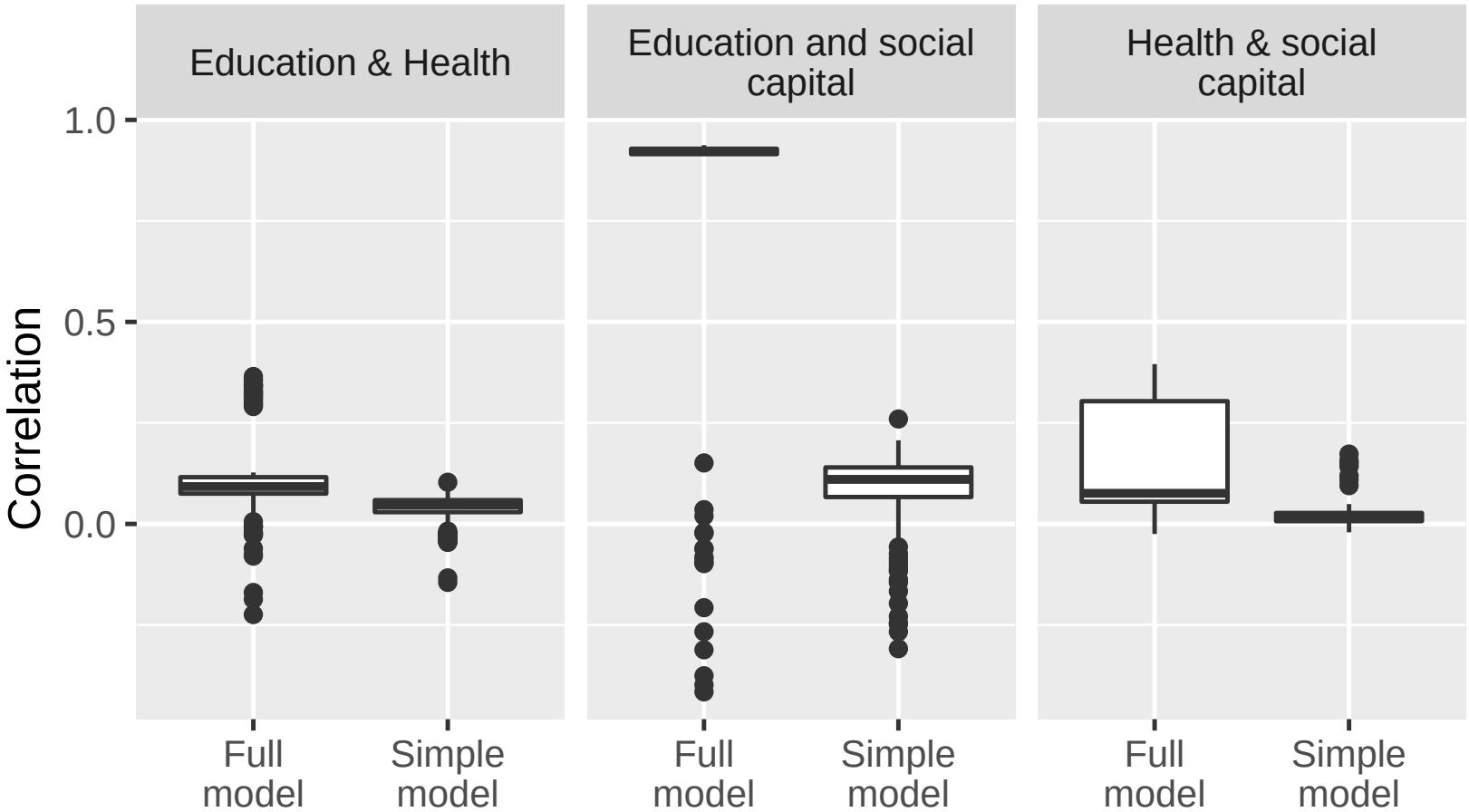
- Why does the full model generally performs better than the simple model?
- Obvious answer: It is able to take into account the links between dimensions.
- But then, are there always strong links between all dimensions in real-life? Not always!
- Hence, a nuanced answer: it works better for dimensions that are closely related like education and social network in our example.
- Is there a problem in applying SEM when the dependence between dimensions is weak? No, the results remain good.
- Our results show that when correlations between the 'true' capabilities are low for certain dimensions (say between health and social network in our case), the *estimated* capabilities reproduce these weak correlations.

Correlations among true capabilities

	1	2	3	4	5	6	7	8	9
1: Forced to move to public school	1.00								
2: Potential education outcome	-0.48	1.00							
3: Potential human capital	-0.45	0.96	1.00						
4: Full recovery after health shock	0.01	0.02	0.01	1.00					
5: Counterfactual health shocks	0.09	0.39	0.38	0.01	1.00				
6: Number of lost links in social network	0.51	-0.04	-0.04	0.02	0.09	1.00			
7: Number of possible additional to soc.net	-0.30	0.77	0.75	0.02	0.29	0.05	1.00		
8: Potential social network: all in	-0.38	0.82	0.79	0.02	0.33	0.14	0.77	1.00	
9: Potential social network: no out	-0.28	0.66	0.64	0.02	0.26	0.34	0.60	0.94	1.00

Note: For each simulation and related estimation, we compute the correlation at the individual level. In this table we present the average value of all these correlation estimates.

Distribution of correlations among capability estimates



Further Discussion of our research findings I contd.

- Use the full model when interdependence is strong among dimensions. Dimensions can be separated when the correlation is weak.
- Practically speaking, regroup capabilities into blocks that have high correlations within them and low correlations between them, so that a SEM can be formulated within each block and the blocks can be estimated separately without much loss of efficiency.
- But how do we know the correlations in the absence of capability measures?
- Implement a two-stage process by first estimating a full SEM and then re-estimating blocks after dividing into groups depending on the correlations for getting closer to reality.

Further Discussion of our research findings II

- Why is the SEM approach able to better capture capabilities inspired by potential outcomes rather than the choice approach?
- Our SEM approach focuses on continuous latent variables thus producing continuous capability measures. In contrast, choices are typically not continuous but rather categorical (e.g. to enrol child or not). Thus the type of latent variables that we postulated are better suited for potential outcome indicators.
- How to remedy? Use *latent class* models. The approach of latent class models is the same as that of SEM except that the underlying latent variable is discrete.

Further Discussion of our research findings III

- Why are our results poor (low correlations) for certain capability measures? This is the case for 2 out of our 9 'true capabilities' in our model (continuous or discrete).
- Possibly these two 'true capabilities' are not well-defined to start with! In the case of *full recovery after a health shock* we might actually capture an income effect rather than a true decision taken by the family. Similarly for the number of contacts lost from the network, it may not be the result of a deliberate choice on the part of the individual (rather of choices made by others) ...

Sensitivity analysis

Name	Combination	Runs per setting	Parameters	# sims
RiskyBehaviour	Yes	5	$\phi_{RF} = \{1.0, \mathbf{1.2}, 1.4\}$ $\phi_{RB} = \{0.4, \mathbf{0.5}, 0.6\}$	45
SnStart	Yes	5	$\delta_{out} = \{\mathbf{0.0\%}, 0.1\%\}$ $\delta_{inherit} = \{0.1, \mathbf{0.3}, 0.5\}$	30
SnEvol	Yes	5	$\delta^{+}_{SN} = \{0.1, \mathbf{0.2}, 0.3\}$ $\delta^{-}_{SN} = \{0.0, \mathbf{0.05}, 0.1\}$	45
HCacum	Yes	5	$\psi_s = \begin{bmatrix} 0.7; 0.7; 0.7; 0.8; 0.75; 0.75; 0.75; 0.75 \\ 0.75; 0.75; 0.75; 0.85; 0.8; 0.8; 0.8; 0.8 \\ 0.65; 0.65; 0.65; 0.75; 0.7; 0.7; 0.7; 0.7 \end{bmatrix}$	15
Misc	No	5	$\xi = \{\mathbf{0.0}, 1.0\}$ $\sigma_{pers} = \{0.0, \mathbf{0.1}, 0.2\}$ $\sigma_H = \{-1, 0.0, \mathbf{0.1}\}$	35
			TOTAL	**170**

Sensitivity analysis

Summary of main findings of our sensitivity analysis:

1. Overall, the performance of SEM is not substantially affected by modifications in values of model parameters. Put differently, the results presented in the main body of the paper also seem to hold true for other combinations of parameters.
2. Among the three dimensions, health is the most affected one (although the impact is small in absolute terms). This is the same dimension that already produced worse results in the baseline model. Hence, we might argue that dimensions where the model performs less well to start with are also prone to be more sensitive to small changes in the model itself.
3. The SEM results worsen in the scenarios involving changes in the social network dimension parameters, but only for this particular dimension (the other two dimensions are largely unaffected). Here we believe that the relative importance of initial conditions versus subsequent evolution of social relations should be further investigated.

Concluding remarks

Concluding Remarks

- We find very high correlations between our estimates and most 'true' capabilities.

Concluding remarks

Concluding Remarks

- We find very high correlations between our estimates and most 'true' capabilities.
- The estimates are better when using a multidimensional model taking into account the potential links across development dimensions.

Concluding remarks

Concluding Remarks

- We find very high correlations between our estimates and most 'true' capabilities.
- The estimates are better when using a multidimensional model taking into account the potential links across development dimensions.
- Uni-dimensional structural equation models risk producing unstable results when different types of functionings are used (e.g. health) [The results oscillate between the two functionings in the simple model whereas in the full model the simulations point to one of them being the 'appropriate measure']

Concluding remarks

Concluding Remarks

- We find very high correlations between our estimates and most 'true' capabilities.
- The estimates are better when using a multidimensional model taking into account the potential links across development dimensions.
- Uni-dimensional structural equation models risk producing unstable results when different types of functionings are used (e.g. health) [The results oscillate between the two functionings in the simple model whereas in the full model the simulations point to one of them being the 'appropriate measure']
- A sensitivity analysis shows that the results are relatively 'robust' to changes in the values of parameters with a few exceptions.

Next steps

Next steps

- Analyse and compare different operationalisation methods of the CA using our modelling tool
- Allow the capability community to access the generated data for use in research
- Investigate the use of ABM for testing model validity in a general setting (outside the CA)

Krishnakumar, Jaya and Florian Chávez-Juárez, "Estimating Capabilities with Structural Equation Models: How Well are We Doing in a 'Real' World?," *Social Indicators Research*, Nov 2016, *129* (2), 717–737.

Model slides

School Quality

$$SQ_s \sim \begin{cases} U[0.5, 1.0] & \text{if private primary or lower/upper secondary} \\ U[0.6, 0.8] & \text{if public primary or lower/upper secondary} \\ U[0.55, 1.0] & \text{if public or private tertiary school} \end{cases}$$

Hospital Coverage

$$\phi_H = \begin{cases} 0.002 & \text{for villages} \\ 0.003 & \text{for small and large towns} \\ 0.005 & \text{for the capital city} \end{cases}$$

Model slides

Social Link Intensity

$$I_{t+1}^{AB} = \begin{cases} I_t^{AB} + \delta_{SM}^{+} & \text{if A and B are school-mates in time t and } I_t^{AB} > 0 \\ I_t^{AB} - \delta_{SM}^{-} & \text{if A and B are NOT school-mates in time t and } I_t^{AB} < \delta_{SM}^{T} \\ I_t^{AB} & \text{if A and B are NOT school-mates in time t and } I_t^{AB} \geq \delta_{SM}^{T} \end{cases}$$

$$Probability of connection P(\text{connection}) = \theta_{SN} \times (1 - |P_A - P_B|)$$

where P_A and P_B refer to the personality index (bound between zero and one) of the two individuals.

Model slides

Wage equation

$$\tilde{w}_{it} = \theta_{inc} \times X\hat{\beta} + \varepsilon^{I}_{it} \qquad \text{with } \varepsilon^{I}_{it} \sim \mathcal{N}(0, \hat{\sigma} \times \theta_{incvar})$$

Health Effect on Wages

$$w_{it} = H^{\xi}_{it} \times \tilde{w}$$

Model slides

Initial Health *Stock*

$$HS_{it} = min\left[1, max\left(0.1, \frac{HS_{it}^{mother} + HS_{it}^{father}}{2} + \varepsilon_i^H\right)\right] \qquad \text{with } \varepsilon_i^H \sim \mathcal{N}(0, \sigma_H)$$

Health Stock evolution

$$HS_{it} = HS_{i,t-1} - hs_{it} + r_{it}$$

Model slides

Health *Shock*

$$hs_{it} = \begin{cases} \sim \exp(\phi_{age}) & \text{with probability } \pi_{it}^{hs} \\ 0 & \text{with probability } 1 - \pi_{it}^{hs} \end{cases}$$

Health Shock probability

$$\pi_{it}^{hs} = \left\{(0.217 - 0.002 age_{it} + 0.000063 age_{it}^2) + (1 - H_{it})\right\} \times \phi_{RF}$$

with $\phi_{RF} = \begin{cases} 1 & \text{if no risky behaviour} \\ > 1 & \text{if risky behaviour} \end{cases}$

Recovery

$$r_{it} = \phi_H \times \sqrt{hci_{it}}$$

Health care investment

$$hci_{ith} = \exp\left(\frac{r}{\phi_H}\right) + \tau_D D_{ih}$$

Model slides

Child enrolment decision (only at each level)

$$\text{enrol child } i \;=\; \begin{cases} \textit{true} & \textit{if}\,(age_i < 9 \wedge Asp_p > educ_i) \vee (age_i \geq 9 \wedge Asp_i > educ_i) \\ \textit{false} & \textit{otherwise} \end{cases}$$

Educational Aspiration of Child

$$Aspiration_i = round\left(\frac{Aspiration_{mother} + Aspiration_{father}}{2} + \varepsilon_i^A\right) \quad \text{with } \varepsilon_i^A \sim \mathcal{N}(0,2)$$

Human Capital Accumulation

$$HC_{it} = HC_{i,t-1} + \left(SQ_s \times \frac{IQ_i}{100} \times HS_{it} + \varepsilon_{it}^E\right)$$

Condition to pass a grade: $HC_{it} \geq \Psi_{gs} \equiv \psi_s g$

Education years accumulation

$$educ_{it} = educ_{i,t-1} + \mathbb{1}\left(HC_{it} \geq \Psi_{gs}\right)$$

Model slides

Mating probability

$$Prob(match_{ij}) = \theta_m \times \frac{1 - \Delta P_{ij}}{\sum_j (1 - \Delta P_{ij})}$$

Personality distance

$$\Delta P_{ij} = abs(P_i - P_j)$$

Personality of child

$$P_{child} = \min\left\{1, \max\left(0, \frac{1}{2}(P_{mother} \times P_{father}) + \varepsilon\right)\right\} \qquad \text{with } \varepsilon \sim \mathcal{N}(0, \sigma_{pers})$$

Model slides

Probability of risky behaviour at birth

$$P(RB_i) = \begin{cases} 0.75 & \text{if both parents have risky behaviour} \\ 0.5 & \text{if one of the parents has risky behaviour} \\ 0.25 & \text{if none of the parents have a risky behaviour} \end{cases}$$

Probability of risky behaviour

$$P(RB_{it}) = \phi_{RB} \overline{RB}^{SN}_{i,t-1} + (1 - \phi_{RB}) RB_{i,t-1}$$

Risky Behaviour

$$RB_{it} = \begin{cases} 1 & \text{with probability } P(RB_{it}) \\ 0 & \text{with probability } 1 - P(RB_{it}) \end{cases}$$