Model-Based synthesis of indicators

Statistical Composite Indicators to convey consistent policy messages

Carlo Cavicchia, Maurizio Vichi

Department of Statistical Sciences
Sapienza University of Rome

em: carlo.cavicchia@uniroma1.it maurizio.vichi@uniroma1.it

Workshop: 9-10 November 2017

Outline of the Presentation

FROM DATA to KNOWLEDGE by Dimensionality Reduction:
a model-based Composite Indicator (CI) is the result of the dimensional reduction of the observed multivariate data.

We will start from

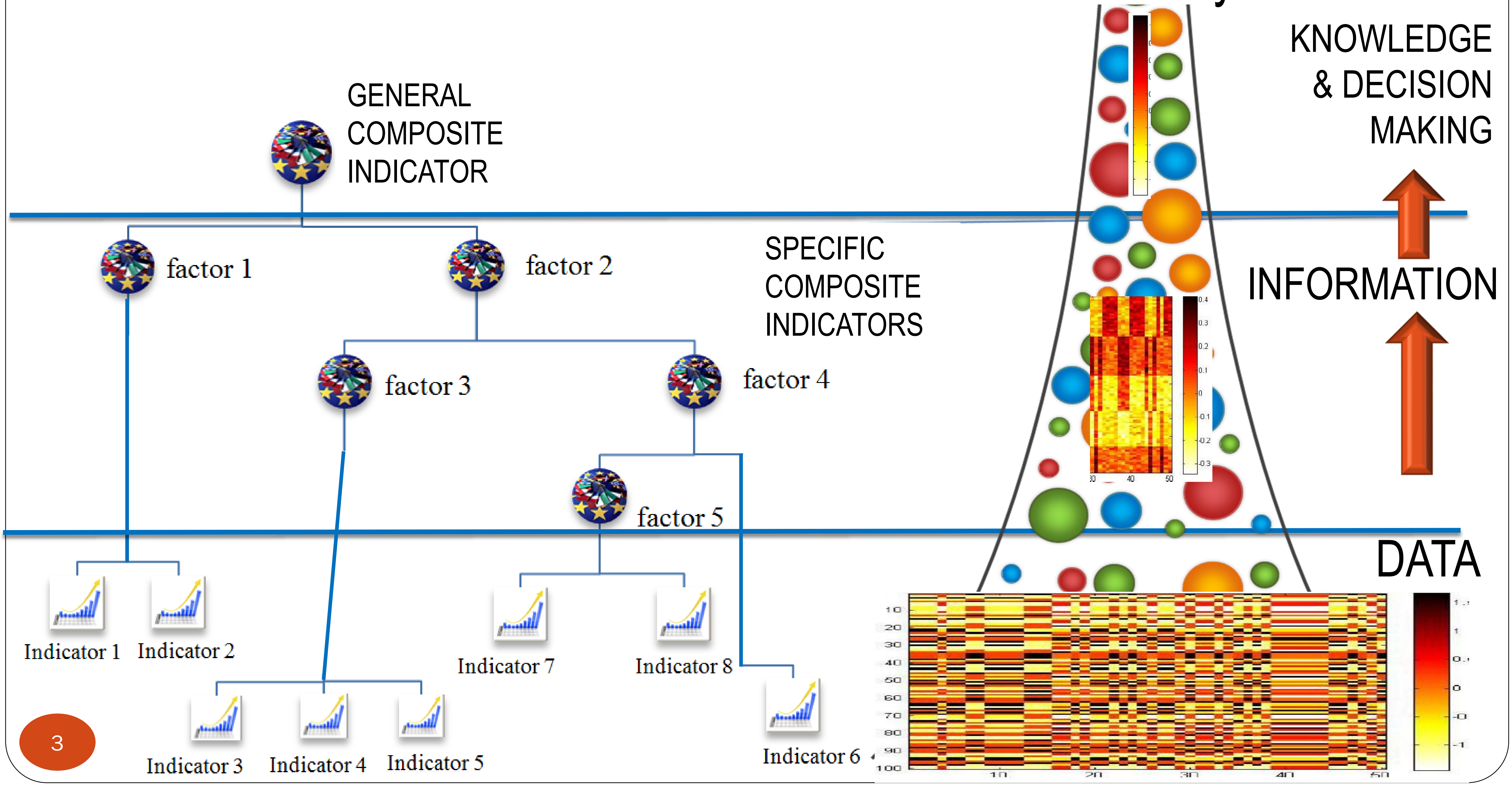
- **Properties on which a Composite Indicator should be based**

We will discuss

- **HDNFA: Hierarchical Disjoint Non-negative Factor Analysis Model**

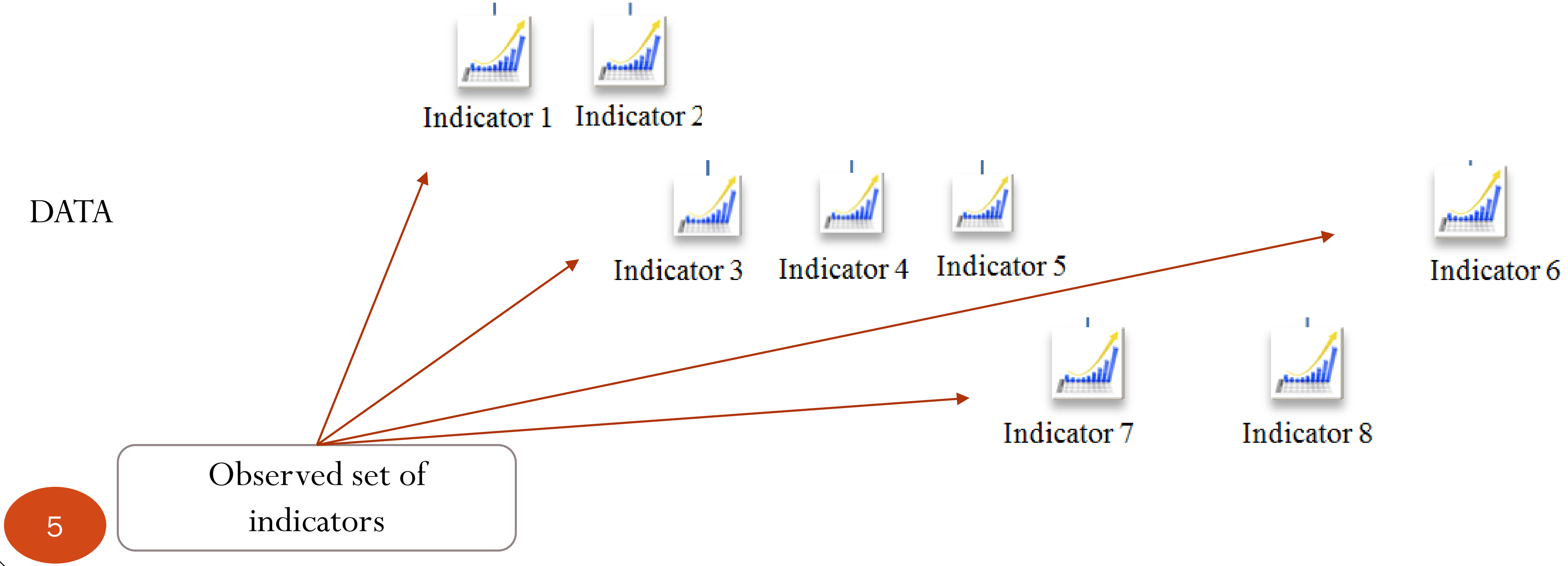
- **Application study: SDGs for Europe with data by EUROSTAT**

From DATA to KNOWLEDGE via Data Dimensionality Reduction



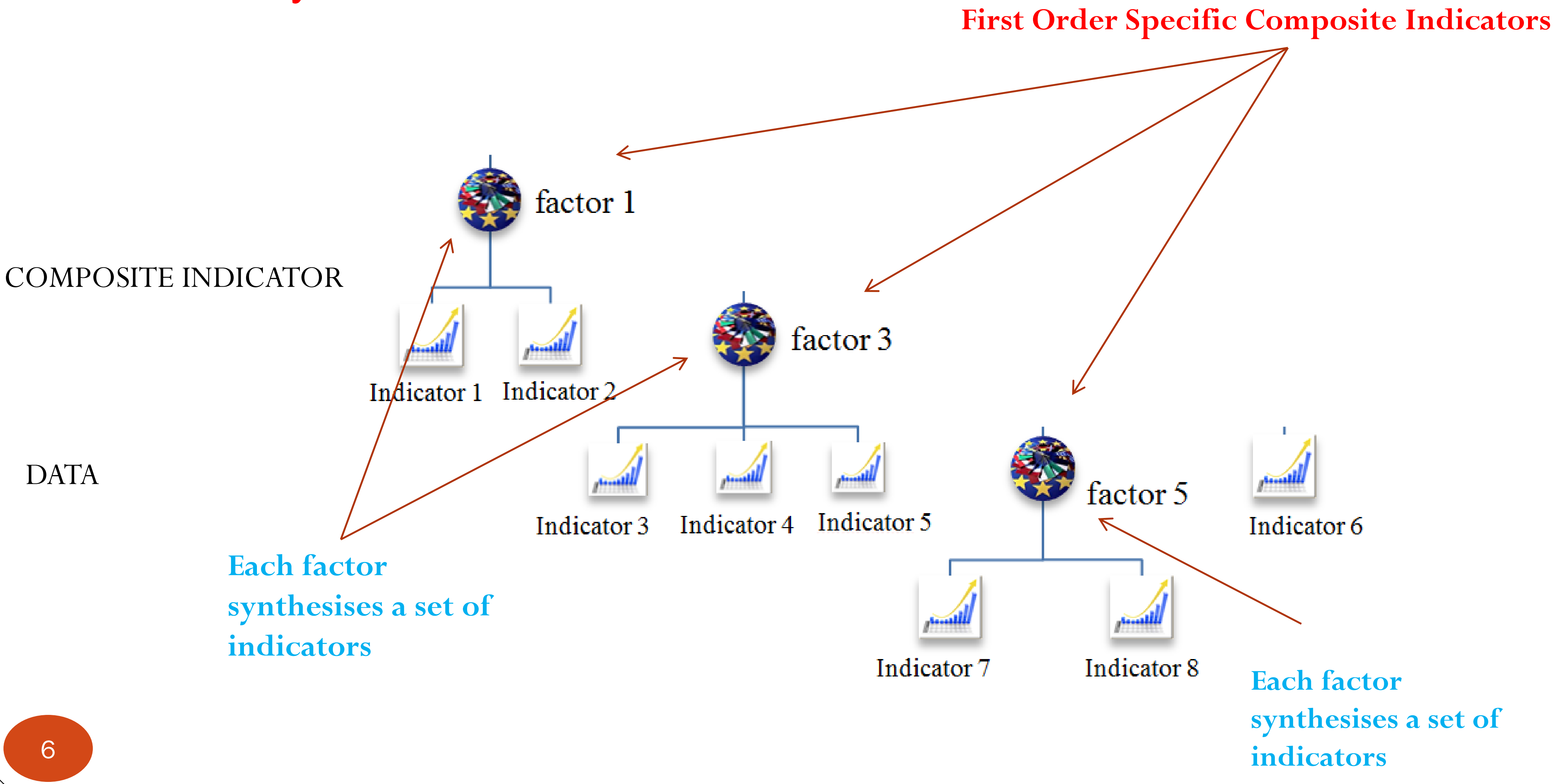
Moving from DATA to KNOWLEDGE by a Dimensionality Reduction

Let us start from the Observed Data (Manifest Indicators) on a phenomenon



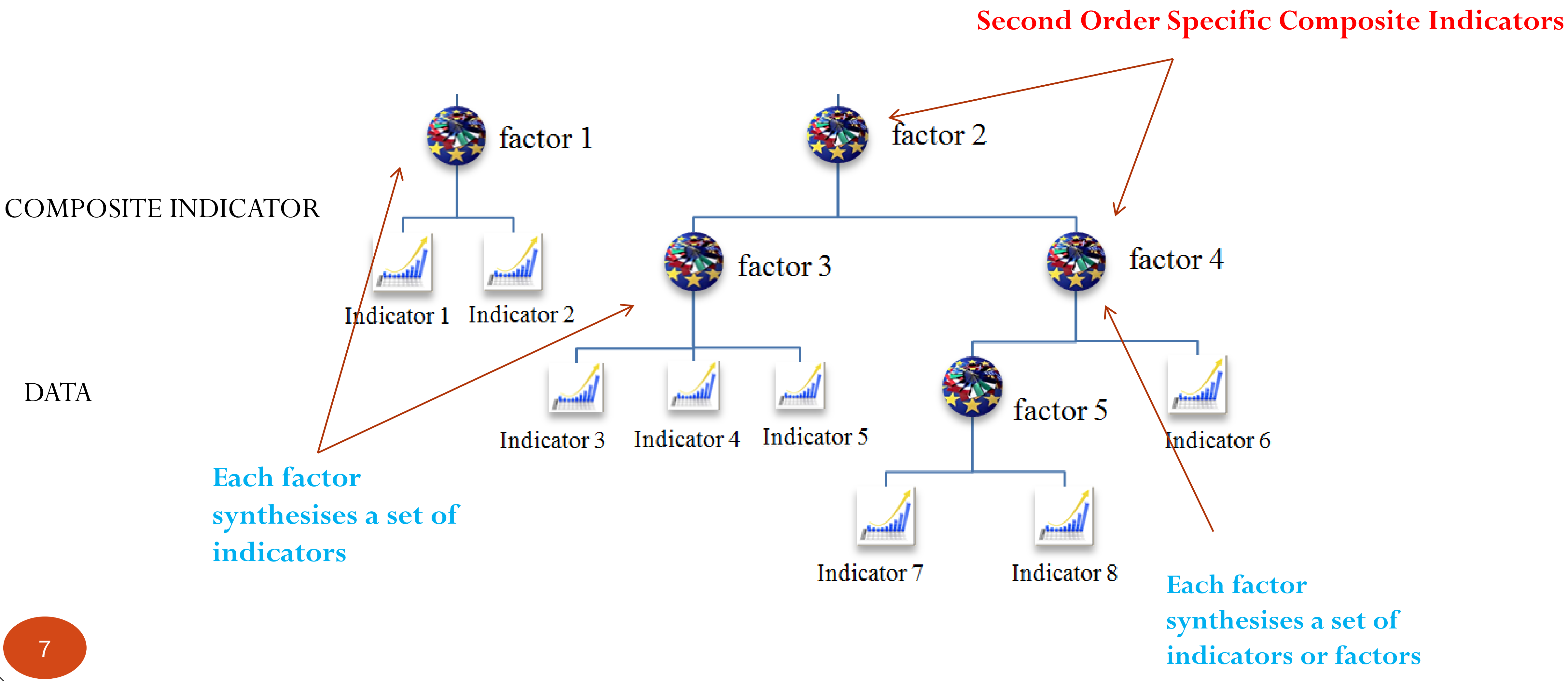
Specific Composite Indicators for detecting the “specific concepts” describing the phenomenon (1/2)

First Level Synthesis



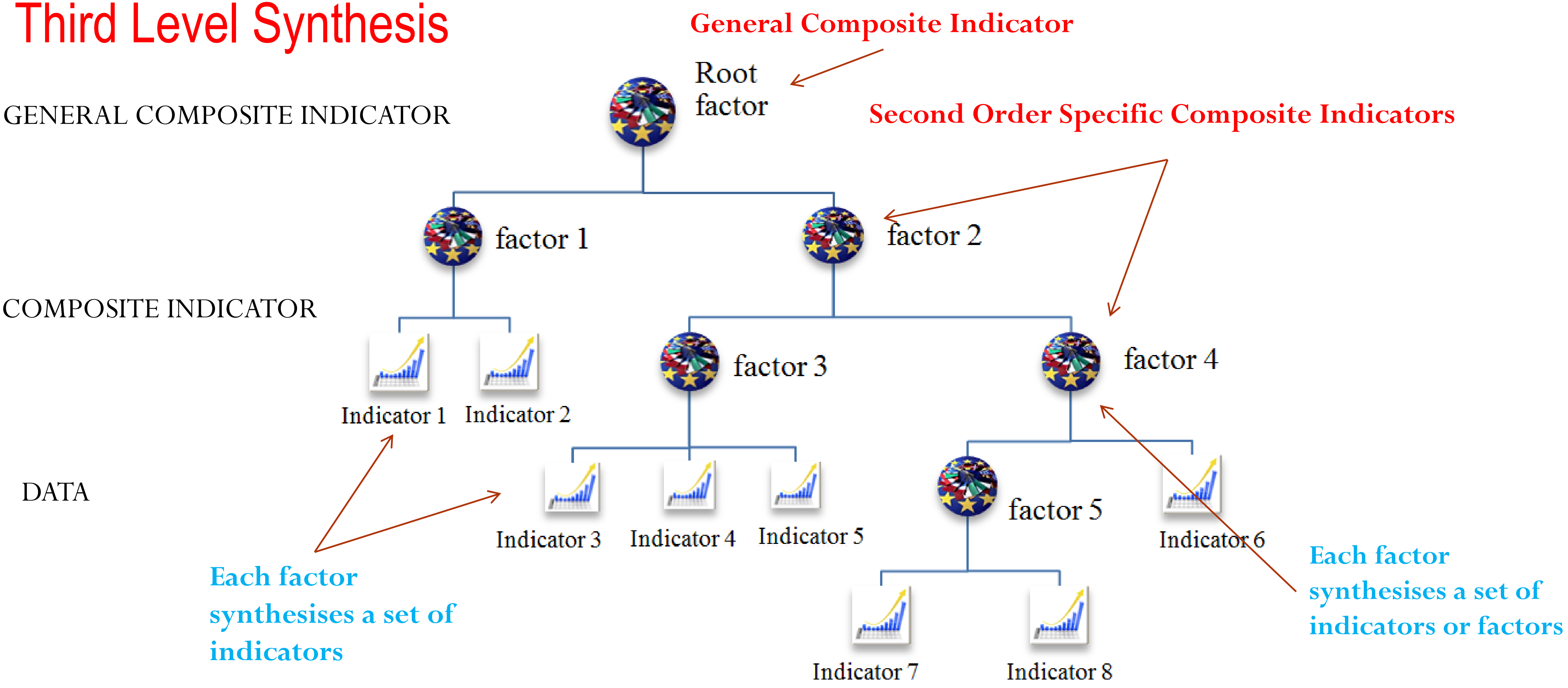
Specific Composite Indicators for detecting the “main concepts” describing the phenomenon (2/2)

Second Level Synthesis



General Composite Indicator (for decision making)

Third Level Synthesis



Hierarchical CI model to move from DATA to KNWOLEDGE

PROPERTIES of CI

Model-Based

Statistically estimated (non-normative)

Confirmatory, Exploratory or Mixed

Reflective and/or Formative

Scale-invariant CI is a latent variable that is not sensitive to linear transformations such as normalization methods (i.e., standardization, Min-Max method);

Non-Compensable & Non-Negative

Reliable

Unidimensional

Statistical Model for Hierarchical CI

Model-based CI and its statistical estimation (i.e., non-normative):

Data = Hierarchical CI model + error

Advantages
Statistical estimation (LS, MLE, …)
Goodness of Fit (to confirm the model)
Inference on the weights, GoF, …

Which typology of research approach:

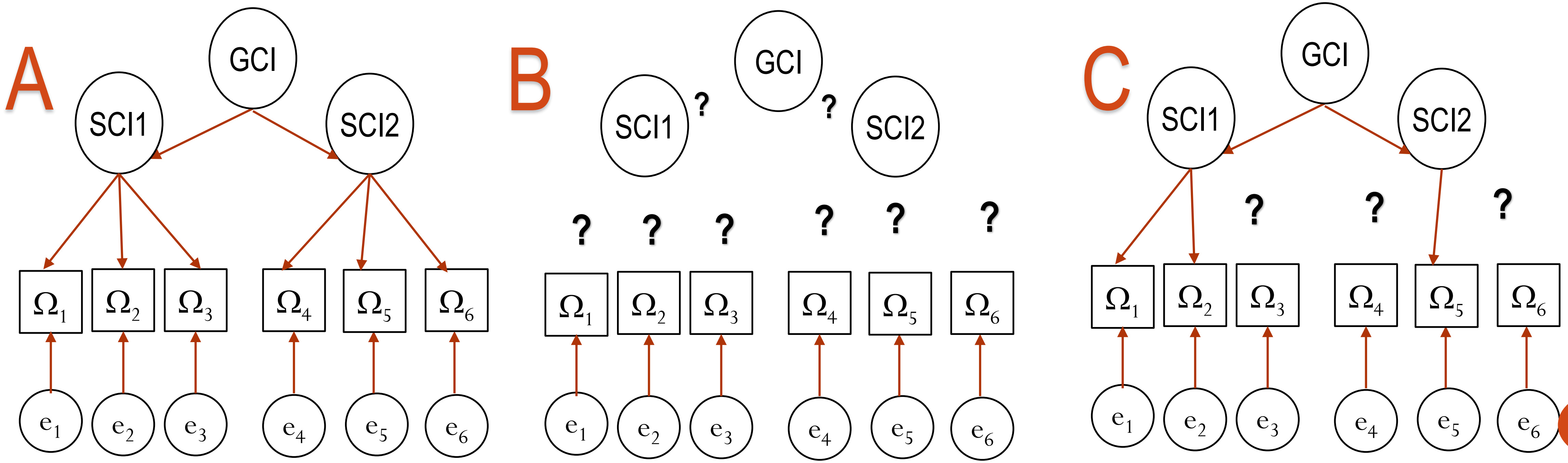
- **Confirmatory –** a Scientific Theory (ST) is assumed and has to be confirmed by the observed indicators;
- **Exploratory –** no clear ST is known, thus, regularities are searched in the data;
- **Mixed Confirmatory & Exploratory –** part of the ST is known, but it is not completely known

Which typology of relations between indicators:

- **Reflective**
- **Formative**

Confirmatory, Exploratory, Mixed-Confirmatory/Exploratory

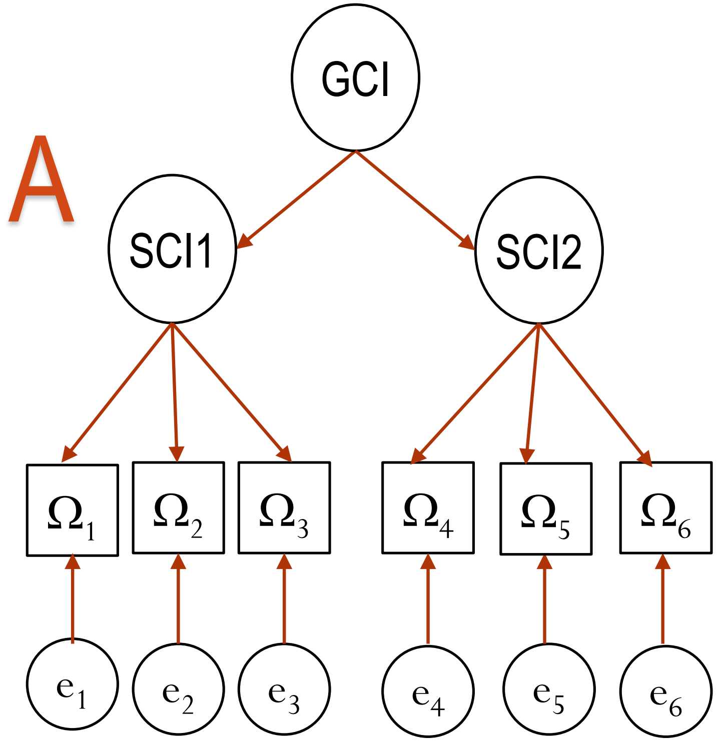
- **Confirmatory** model: if a theory on the model of the CI is available, i.e., **all relationships** between manifest variables and latent variables **are** and a **priori known**;
- **Exploratory** model: **all relationships** between manifest variables and latent variables **are not a priori known**;
- **Mixed-confirmatory/exploratory** : **some relationships** are **known** according to a theory and **some** are **unknown** and must be achieved by exploratory analysis.



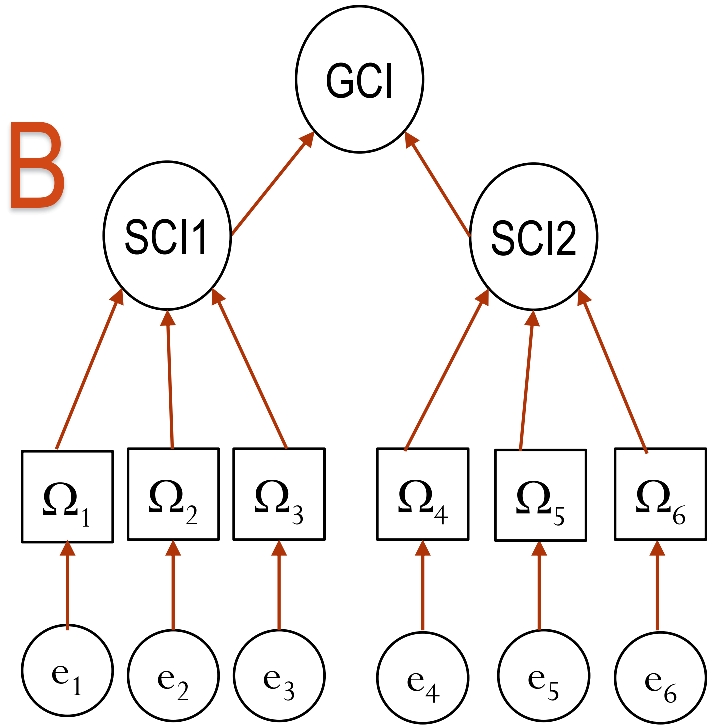
Relations between Composite Indicators (GCI & SCIs) and Manifest Indicators

A) Reflective

B) Formative



The General Composite Indicator is a determinant (causes) the Specific Composite Indicators & these last are determinant (causes) of the Manifest Indicators, i.e., The GCI reconstructs the SCIs that reconstruct the MI



Independent Manifest Indicators are determinant (cause, explain) of independent Specific Composite indicators that are determinant of the General Composite Indicator)

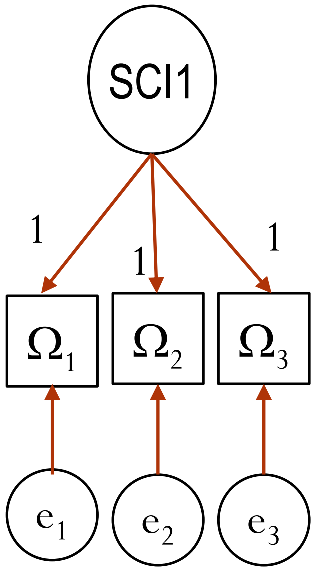
Why we use a hierarchical, non-negative and Reflective model? (1/3)

1. Why reflective and non-negative

Let us consider that loadings are positive correlations and let us suppose that the SCI1 exactly determines three MIs

Why Reflective

A



SCI1 determines 100% MIs

$$\Omega_1 = a_{11}SCI1 + e_1$$
$$\Omega_2 = a_{21}SCI1 + e_2$$
$$\Omega_3 = a_{31}SCI1 + e_3$$

The Specific Composite Indicator is the latent determinant of the Manifest Indicators, i.e.,
The SCI reconstructs the three MIs

Why non-negative

Loading Matrix	a_{11}	1			
A =	a_{21}	1			
	a_{31}	1			
	Ω_1	Ω_2	Ω_3		Var SCI
Var-Cov Matrix	1	1	1	PCA or FA	3
$\Sigma x = A*I*A' + \Psi x =$	1	1	1	eig(Σx) =	0
	1	1	1		0
				Var. explaided	100%

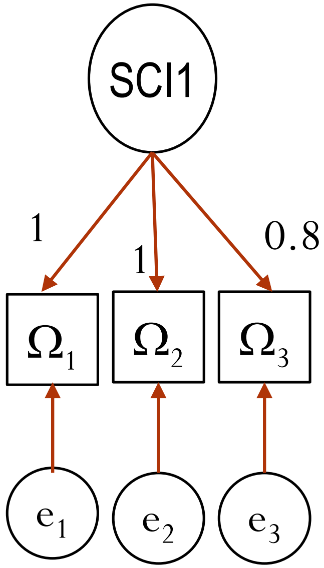
Weights are positive correlations (values between 0 and 1)
1 if SCI perfectly determines the MI,
0 if SCI dos not determine the MI

Why a hierarchical non-negative and Reflective model? (1/3)

Let us decrease the weight of SCI1 on Ω_3

Why Reflective

A



SCI1 determines 91% MIs

$$\Omega_1 = a_{11}\mathrm{SCI1} + e_1$$
$$\Omega_2 = a_{21}\mathrm{SCI1} + e_2$$
$$\Omega_3 = a_{31}\mathrm{SCI1} + e_3$$

The Specific Composite Indicator is the latent determinant of the Manifest Indicators, i.e.,
The SCI reconstructs the three MIs

Why non-negative

Loading Matrix	a_{11} 1.0			
A =	a_{21} 1.0			
	a_{31} 0.8			
	Ω_1 Ω_2 Ω_3			Var SCI
Var-Cov Matrix	1.0 1.0 0.8		PCA or FA	2.74
$\Sigma x = A*I*A'+\Psi x$ =	1.0 1.0 0.8		eig(Σx) =	0.26
	0.8 0.8 1.0			0
			Var explained	91%

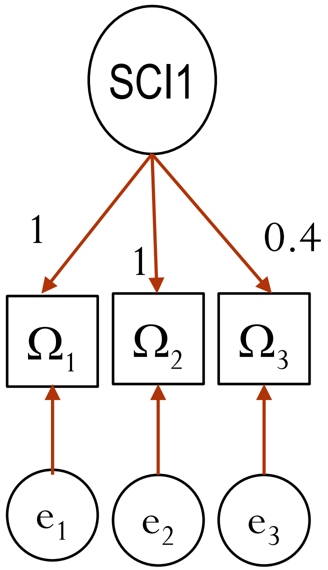
Weights are positive correlations (values between 0 and 1)
1 if SCI perfectly determines the MI,
0 if SCI dos not determine the MI

Why a hierarchical non-negative and Reflective model? (2/3)

Let us continue to decrease the weight of SCI1 on Ω_3

Why Reflective

A



SCI1 determines 75% MIs

$\Omega_1 = a_{11}SCI1 + e_1$

$\Omega_2 = a_{21}SCI1 + e_2$

$\Omega_3 = a_{31}SCI1 + e_3$

The Specific Composite Indicator is the latent determinant of the Manifest Indicators, i.e.,
The SCI reconstructs the three MIs

Why non-negative

Loading Matrix	a_{11}	1.0			
A =	a_{21}	1.0			
	a_{31}	0.4			
	Ω_1	Ω_2	Ω_3		Var SCI
Var-Cov Matrix	1.0	1.0	0.4	PCA or FA	2.26
$\Sigma x = A*I*A'+\Psi x =$	1.0	1.0	0.4	eig(Σx) =	0.74
	0.4	0.4	1.0		0
				Var expained	75.3%

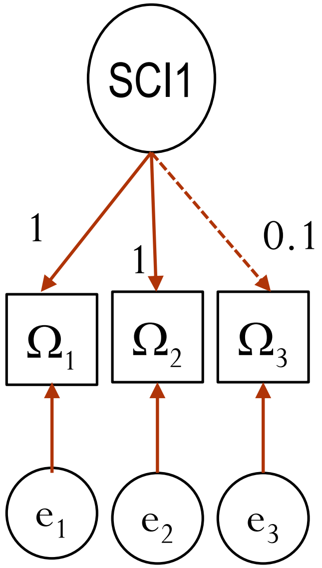
Weights are positive correlations (values between 0 and 1)
1 if SCI perfectly determines the MI,
0 if SCI dos not determine the MI

Why a hierarchical non-negative and Reflective model? (1/3)

Let us continue to decrease the weight of SCI1 on Ω_3

Why Reflective

A



SCI1 determines 67% MIs

$$\Omega_1 = a_{11}\text{SCI1} + e_1$$
$$\Omega_2 = a_{21}\text{SCI1} + e_2$$
$$\Omega_3 = a_{31}\text{SCI1} + e_3$$

The Specific Composite Indicator causes the Manifest Indicators,
i.e.,
The SCI reconstructs the three MIs

Why non-negative

Loading Matrix	a_{11} 1.0		
A =	a_{21} 1.0		
	a_{31} 0.1		
	Ω_1 Ω_2 Ω_3		Var SCI
Var-Cov Matrix	1.0 1.0 0.1	PCA or FA	2.02
$\Sigma x = A{*}I{*}A' + \Psi x =$	1.0 1.0 0.1	eig(Σx) =	0.98
	0.1 0.1 1.0		0
		Var expained	67.3%

Why a hierarchical non-negative and Reflexive model? (2/3)

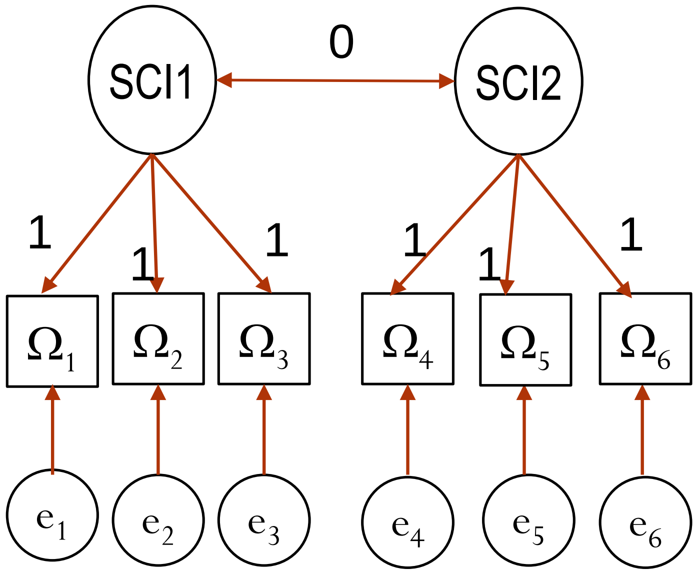
Let us consider that SCI1 and SCI2 exactly (100%) determine two groups of three MIs.
Let us first suppose the two SCIs are uncorrelated

Why hierarchical

SCI1 determine 50% MIs

SCI2 determine 50% MIs

A



Two Specific Composite Indicators determine two groups of Manifest Indicators, i.e.,
The two SCIs reconstruct the two groups of MIs

Loading Matrix	a_{11} a_{12}	1	0							
	a_{21} a_{22}	1	0						1.0	0.0
A =	a_{31} a_{32}	1	0					Σy =	0.0	1.0
	a_{41} a_{42}	0	1							
	a_{51} a_{52}	0	1							
	a_{61} a_{62}	0	1							
	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6				Var SCI
	1	1	1	0	0	0				3
Var-cov Matrix	1	1	1	0	0	0		PCA		3
$\Sigma x = A{*}I{*}A'+\Psi x$ =	1	1	1	0	0	0		eig(Σx) =		0
	0	0	0	1	1	1				0
	0	0	0	1	1	1				0
	0	0	0	1	1	1				0

$\Omega_1 = a_{11}SCI1 + e_1$

$\Omega_2 = a_{21}SCI1 + e_2$

$\Omega_3 = a_{31}SCI1 + e_3$

$\Omega_4 = a_{42}SCI1 + e_4$

$\Omega_5 = a_{52}SCI1 + e_5$

$\Omega_6 = a_{62}SCI1 + e_6$

Why a hierarchical non-negative and Reflexive model? (2/3)

Why do we need the Hierarchical model rather than a Factor Analysis (FA)?

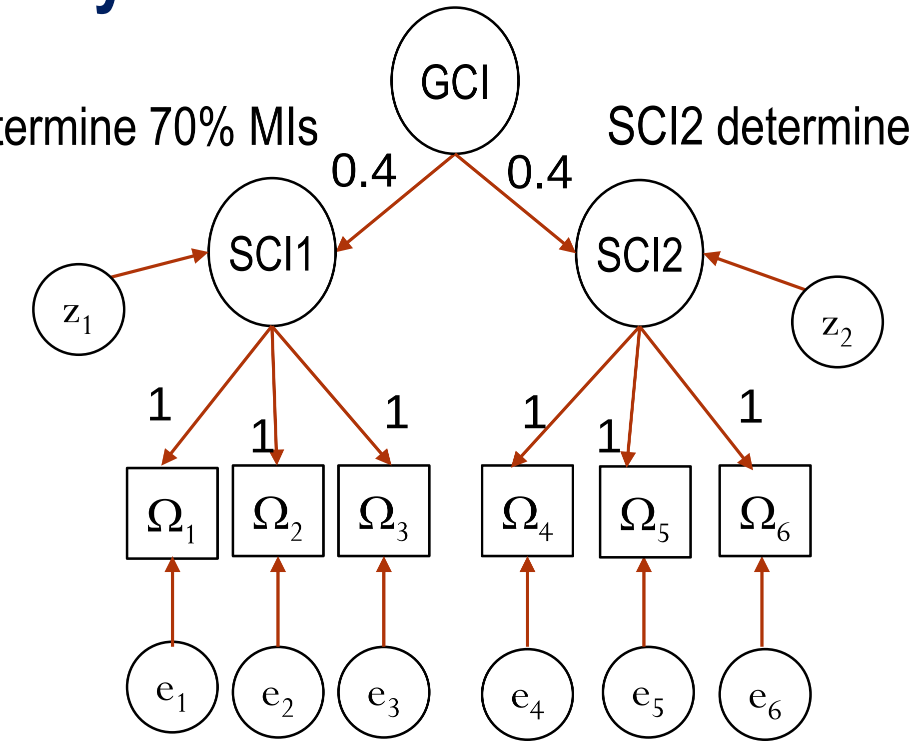
Because FA or PCA are not appropriate methodologies.

In fact they tend to identify a single factor, and the hierarchical structure of the GCI is masked

Why hierarchical

SCI1 determine 70% MIs

SCI2 determine 30% MIs



Loading matrix

$$A = \begin{matrix} a_{11}\ a_{12} \\ a_{21}\ a_{22} \\ a_{31}\ a_{32} \\ a_{41}\ a_{42} \\ a_{51}\ a_{52} \\ a_{61}\ a_{62} \end{matrix} \quad \begin{matrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{matrix} \qquad \Sigma y = \begin{matrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{matrix}$$

	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6		Var SCI
Observed var-cov	1.0	1.0	1.0	0.4	0.4	0.4		4.2
$\Sigma x = A^*\Sigma y^*A+\Psi x =$	1.0	1.0	1.0	0.4	0.4	0.4	PCA	1.8
	1.0	1.0	1.0	0.4	0.4	0.4	eig(Σx) =	0
	0.4	0.4	0.4	1.0	1.0	1.0		0
	0.4	0.4	0.4	1.0	1.0	1.0		0
	0.4	0.4	0.4	1.0	1.0	1.0		0

The General Composite Indicator causes the Specific Composite Indicators & these last cause the Manifest Indicators,
i.e.,
The GCI reconstructs the SCIs that reconstruct the MI

$\Omega_1 = a_{11}SCI1 + e_1$ $\quad$ $\Omega_4 = a_{42}SCI2 + e_4$ $\quad$ $SCI1 = c_1GCI + z_1$

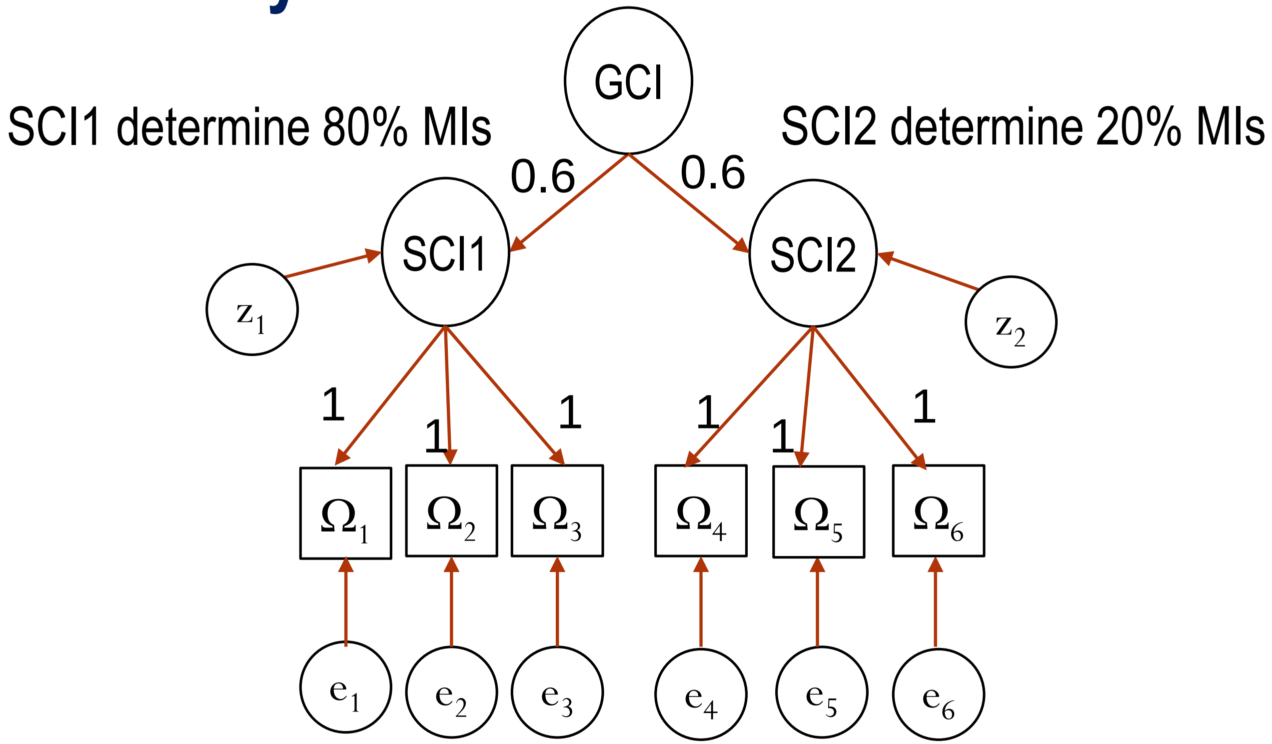
$\Omega_2 = a_{21}SCI1 + e_2$ $\quad$ $\Omega_5 = a_{52}SCI2 + e_5$ $\quad$ $SCI2 = c_2GCI + z_2$

$\Omega_3 = a_{31}SCI1 + e_3$ $\quad$ $\Omega_6 = a_{62}SCI2 + e_6$

Why a hierarchical Reflective model? (2/3)

Why do we need the Hierarchical model rather than a Factor Analysis (FA)?
Because FA or PCA are not appropriate methodologies.
In fact they tend to identify a single factor, and the hierarchical structure of the GCI is masked

Why hierarchical



Loading matrix	a_{11} a_{12}	1	0						
	a_{21} a_{22}	1	0					1.0	0.6
A =	a_{31} a_{32}	1	0				Σy =	0.6	1.0
	a_{41} a_{42}	0	1						
	a_{51} a_{52}	0	1						
	a_{61} a_{62}	0	1						
	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6			Var SCI
Observed var-cov	1.0	1.0	1.0	0.6	0.6	0.6			4.8
$\Sigma x = A*\Sigma y*A'+\Psi x$ =	1.0	1.0	1.0	0.6	0.6	0.6		PCA	1.2
	1.0	1.0	1.0	0.6	0.6	0.6		eig(Σx) =	0
	0.6	0.6	0.6	1.0	1.0	1.0			0
	0.6	0.6	0.6	1.0	1.0	1.0			0
	0.6	0.6	0.6	1.0	1.0	1.0			0

The General Composite Indicator causes the Specific Composite Indicators & these last cause the Manifest Indicators,
i.e.,
The GCI reconstructs the SCIs that reconstruct the MI

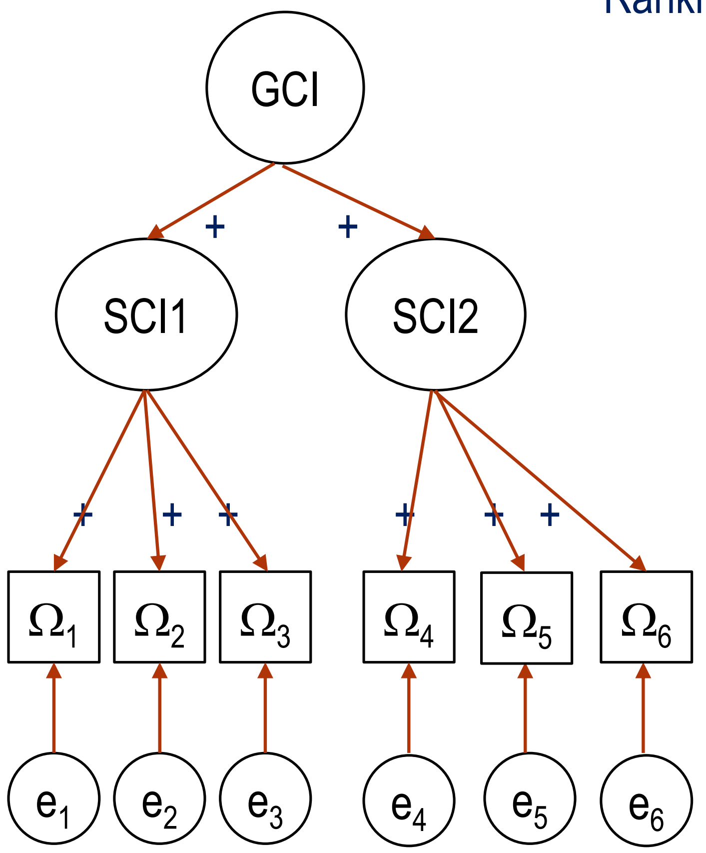
$\Omega_1 = a_{11}SCI1 + e_1$ $\quad \Omega_4 = a_{42}SCI2 + e_4$ $\quad SCI1 = c_1GCI + z_1$

$\Omega_2 = a_{21}SCI1 + e_2$ $\quad \Omega_5 = a_{52}SCI2 + e_5$ $\quad SCI2 = c_2GCI + z_2$

$\Omega_3 = a_{31}SCI1 + e_3$ $\quad \Omega_6 = a_{62}SCI2 + e_6$

Non-Compensability & Non-Negativity.

The CI satisfies the non-compensability property if its relationships with latent and/or manifest variables are all positives. Thus, the effect of the latent and/or manifest variables do not compensate each other.



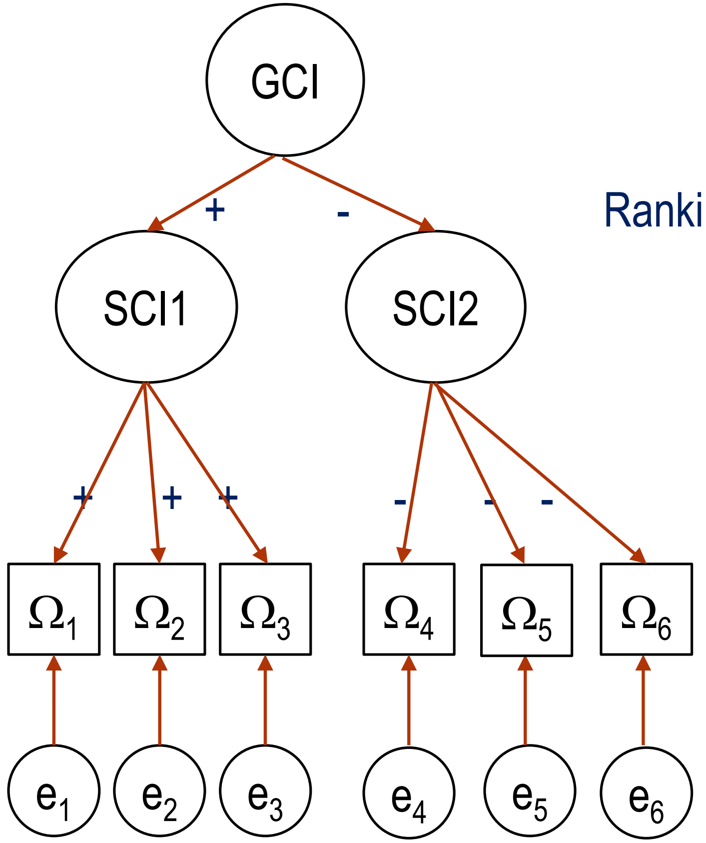
Ranking of the not compensated model:

3	10	1	7	5	9	4	2	8	6

So **non-negativity** and non-compensability are strictly connected.

Non-Compensability & Non-Negativity.

The CI satisfies the non-compensability property if its relationships with latent and/or manifest variables are all positives. Thus, the effect of the latent and/or manifest variables do not compensate each other.



Ranking of the not compensated model:

3	10	1	7	5	9	4	2	8	6

Ranking of the compensated model:

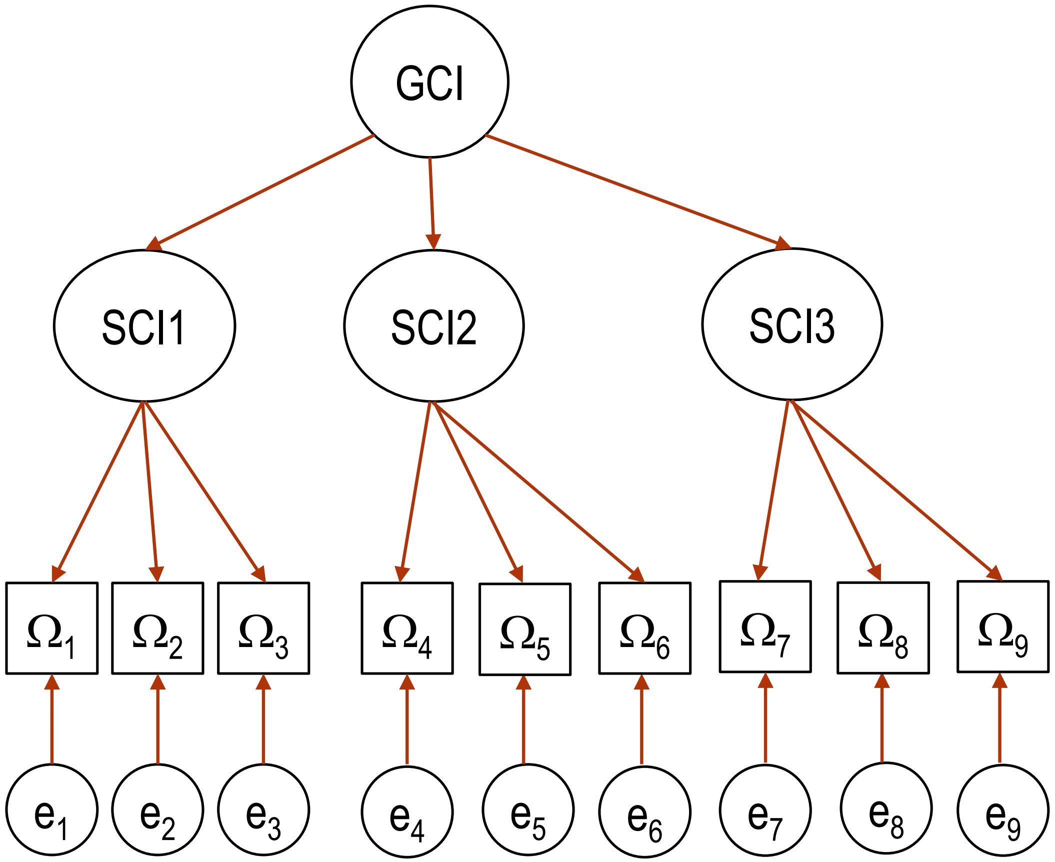
5	9	3	7	1	6	8	4	2	10

So **non-negativity** and non-compensability are strictly connected.

Reliability assesses internal consistency, that is, how well a set of manifest variables measures a single unidimensional latent variable. The Cronbach Coefficient Alpha (c-alpha) is the most common estimator of internal consistency. A latent factor is **reliable** if the Cronbach's alpha is larger than 0.7 (Nunnally, 1978).

Unidimensionality evaluates to which extent a latent variable includes a single concept.

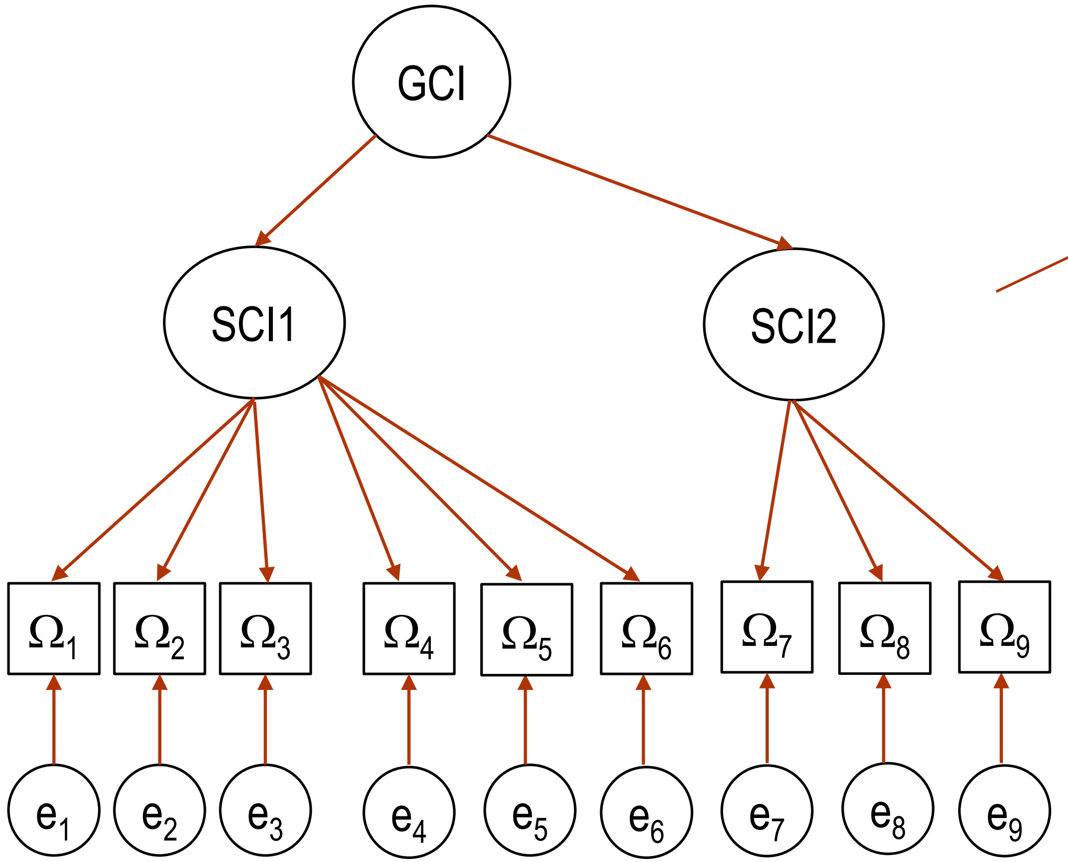
Let us suppose that DATA have this structure:



Reliability assesses internal consistency, that is, how well a set of manifest variables measures a single unidimensional latent variable. The Cronbach Coefficient Alpha (c-alpha) is the most common estimator of internal consistency. A latent factor is **reliable** if the Cronbach's alpha is larger than 0.7 (Nunnally, 1978).

Unidimensionality evaluates to which extent a latent variable includes a single concept.

<u>And let us try to estimate the model by using only two SCI</u>



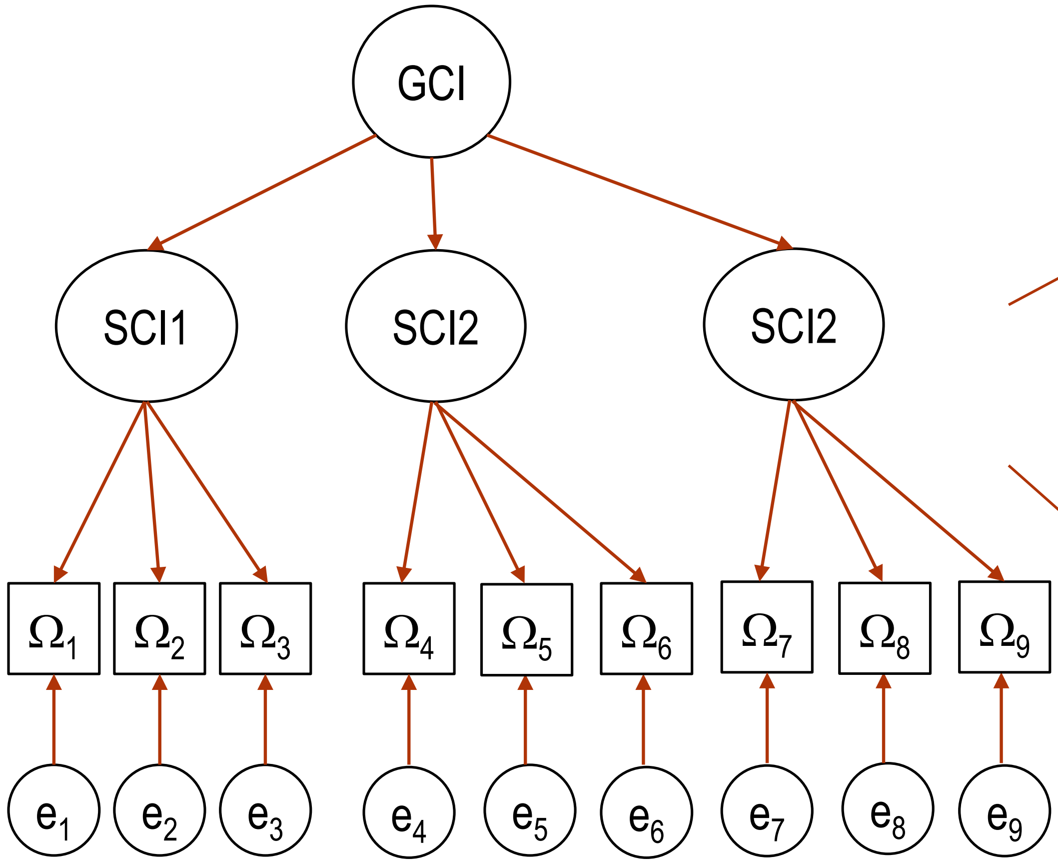
K=2

Unidimensionality: F1= **2.73766**
F2= 0.55638
variance of the second component of class

Reliability: F1= 0.526587
Cronbach's Alpha F2= 0.475812

Reliability assesses internal consistency, that is, how well a set of manifest variables measures a single unidimensional latent variable. The Cronbach Coefficient Alpha (c-alpha) is the most common estimator of internal consistency. A latent factor is **reliable** if the Cronbach's alpha is larger than 0.7 (Nunnally, 1978).
Unidimensionality evaluates to which extent a latent variable includes a single concept.

<u>Let us suppose that DATA have this structure:</u>



K=2

Unidimensionality: F1= **2.73766**
F2= 0.55638
variance of the second component of class

Reliability: F1= 0.526587
Cronbach's Alpha F2= 0.475812

K=3

Unidimensionality: F1= 0.40059
F2= 0.55649
F3= 0.61888
variance of the second component of class

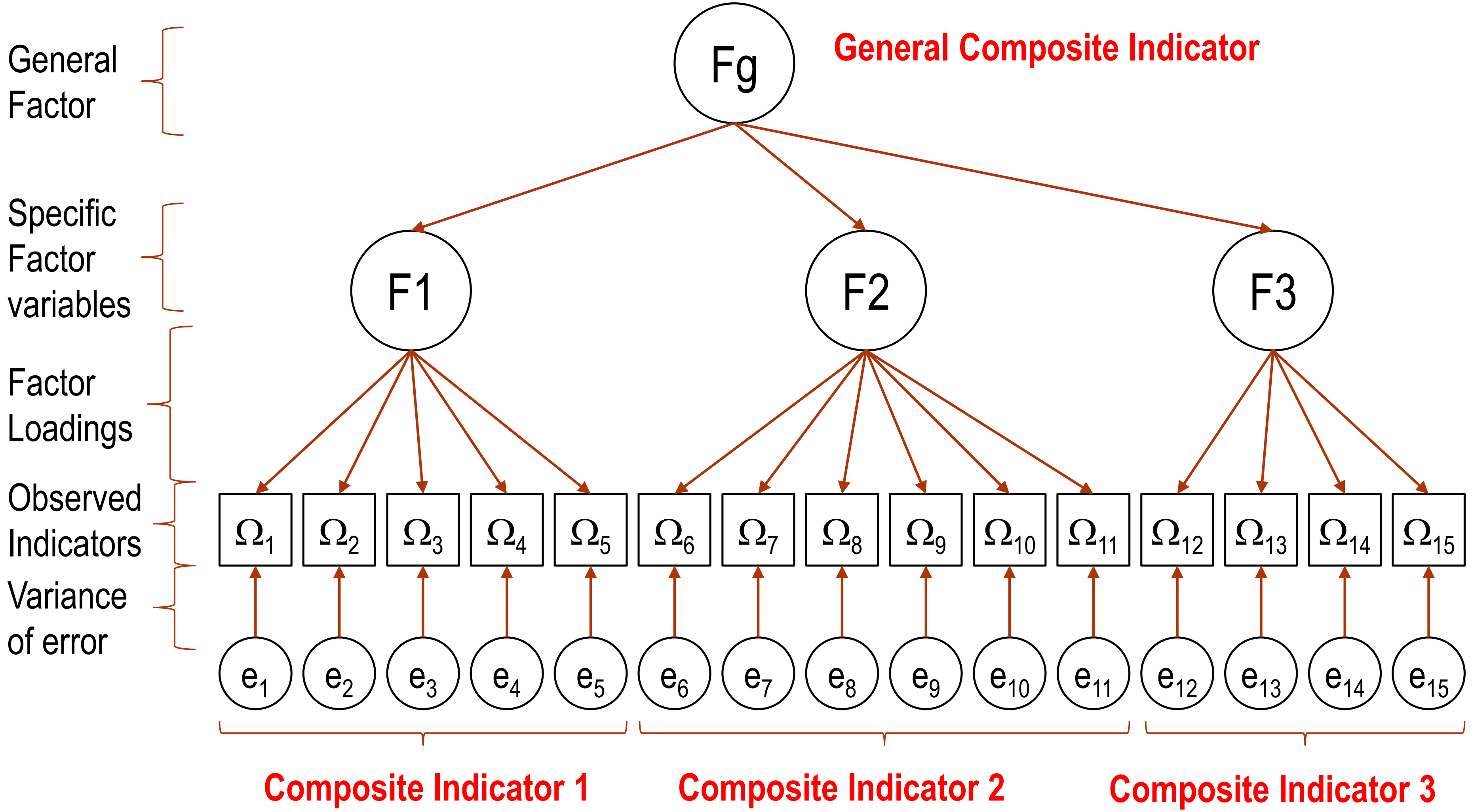
Reliability: F1= 0.781364
Cronbach's Alpha F2= 0.794888
F3= 0.781499

Our strategy for computing model-based composite indicators

Let us consider a two level hierarchical confirmatory factor model

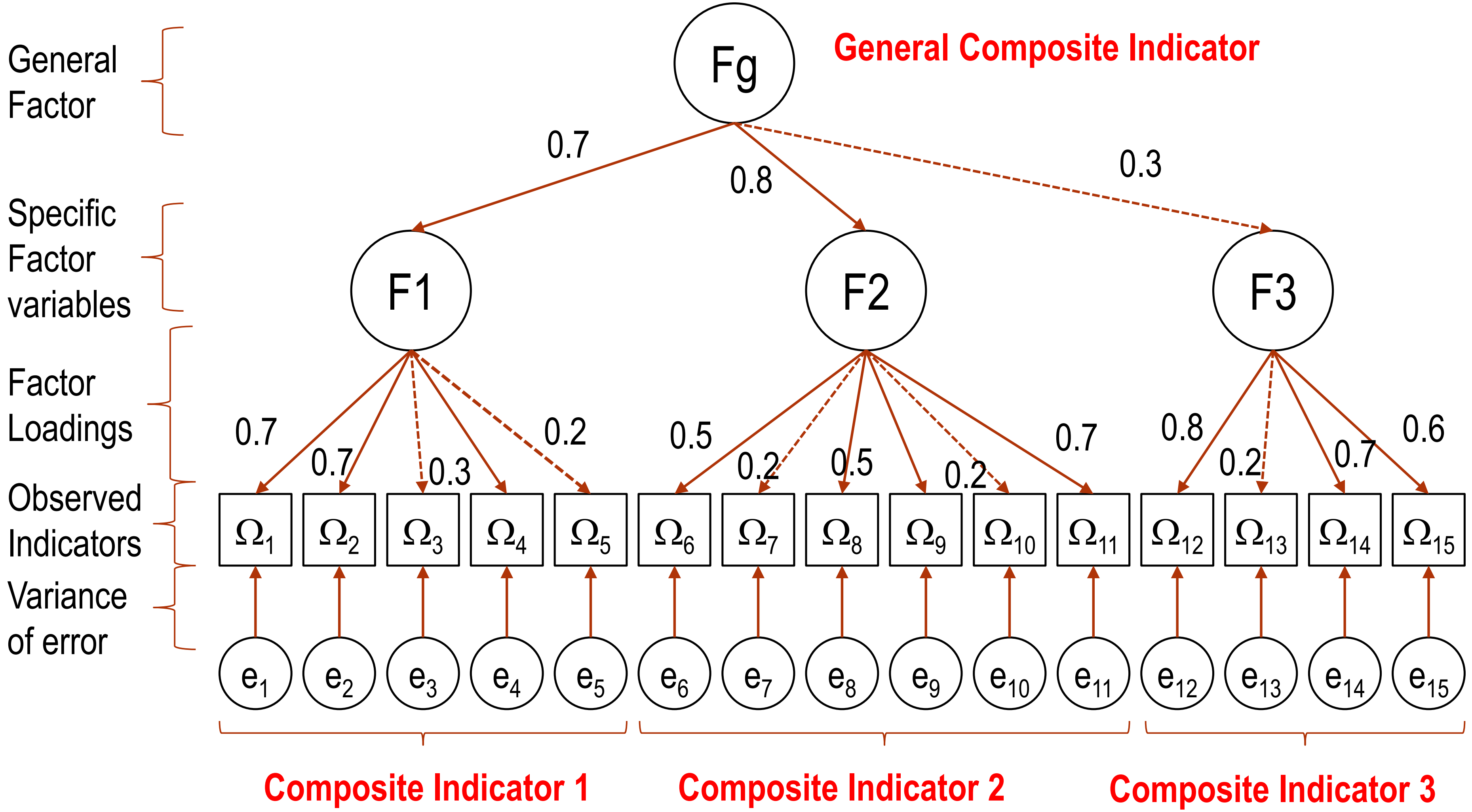
Two level Hierarchical Simple Structure Model

Confirmatory approach: both the number of specific factors and the association between observed indicators and specific composite indicators (represented by arrows) are supposed known while the level of correlation has to be estimated.



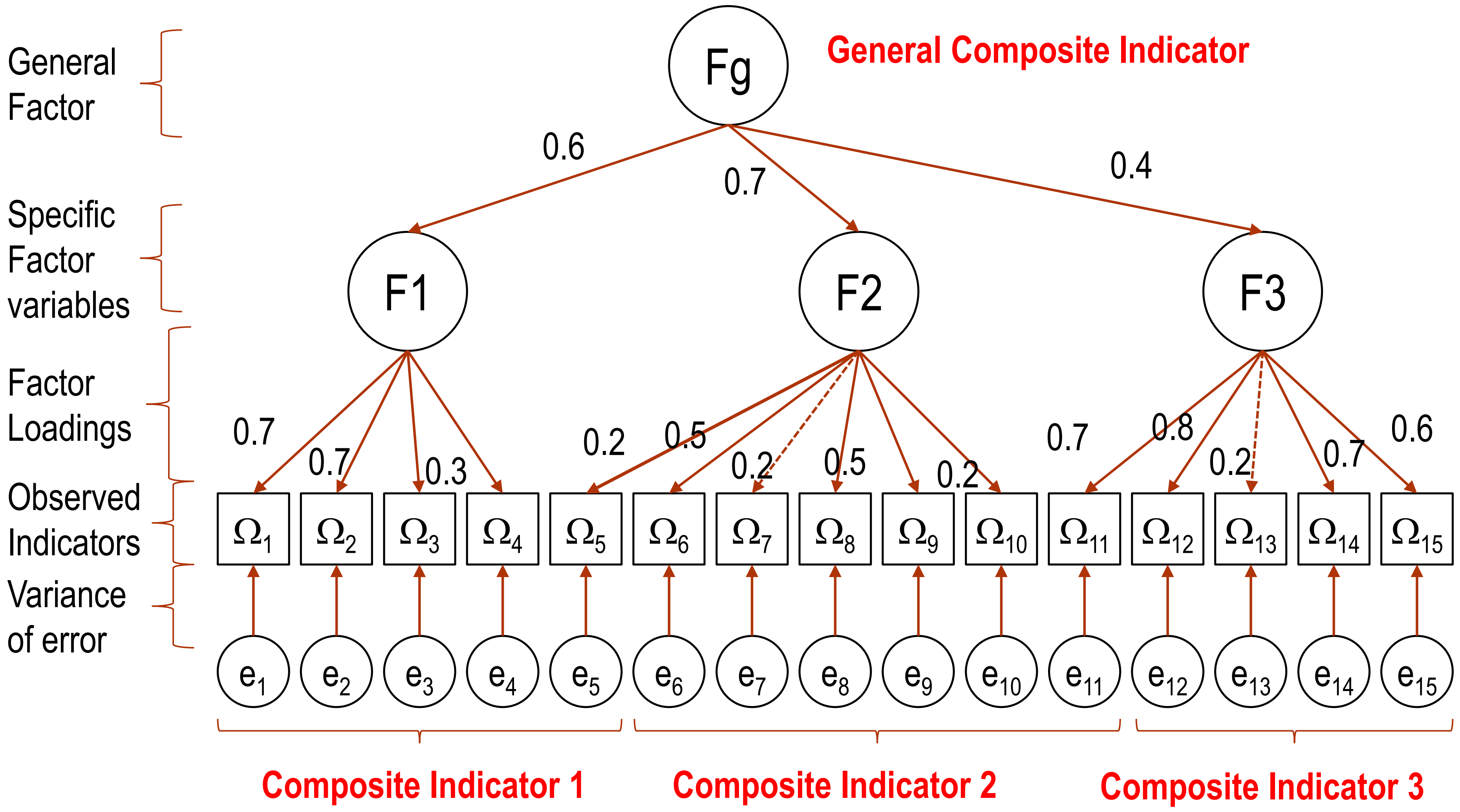
Two level Hierarchical Simple Structure Model

ESTIMATION: correlation between variables and factors, between general and specific factors. Some associations may be not statistically significant (correlations are substantially null) and FIT POOR



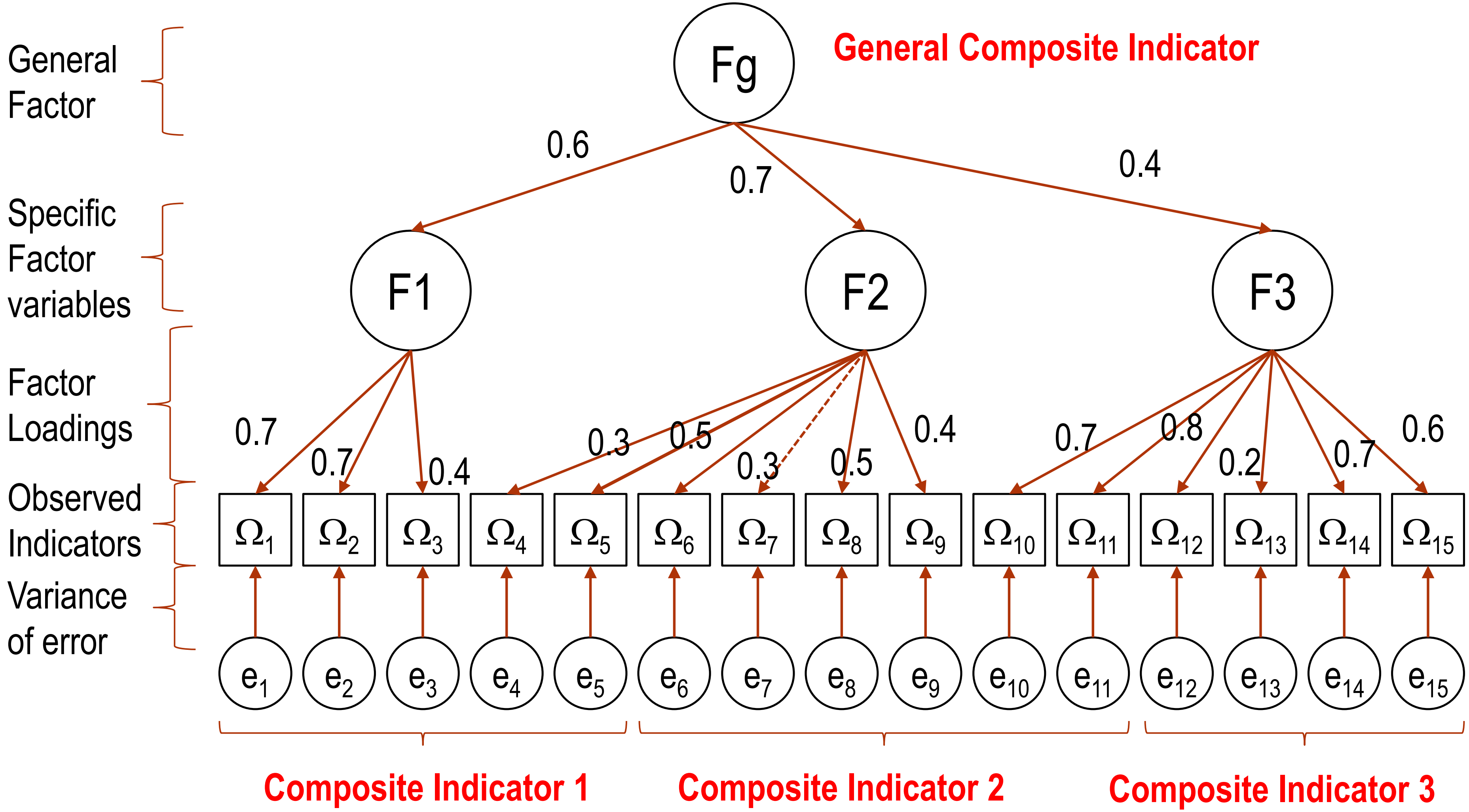
Two level Hierarchical Simple Structure Model

At this point the researcher starts to play with different models hypothesizing some changes that do not have a theory behind.



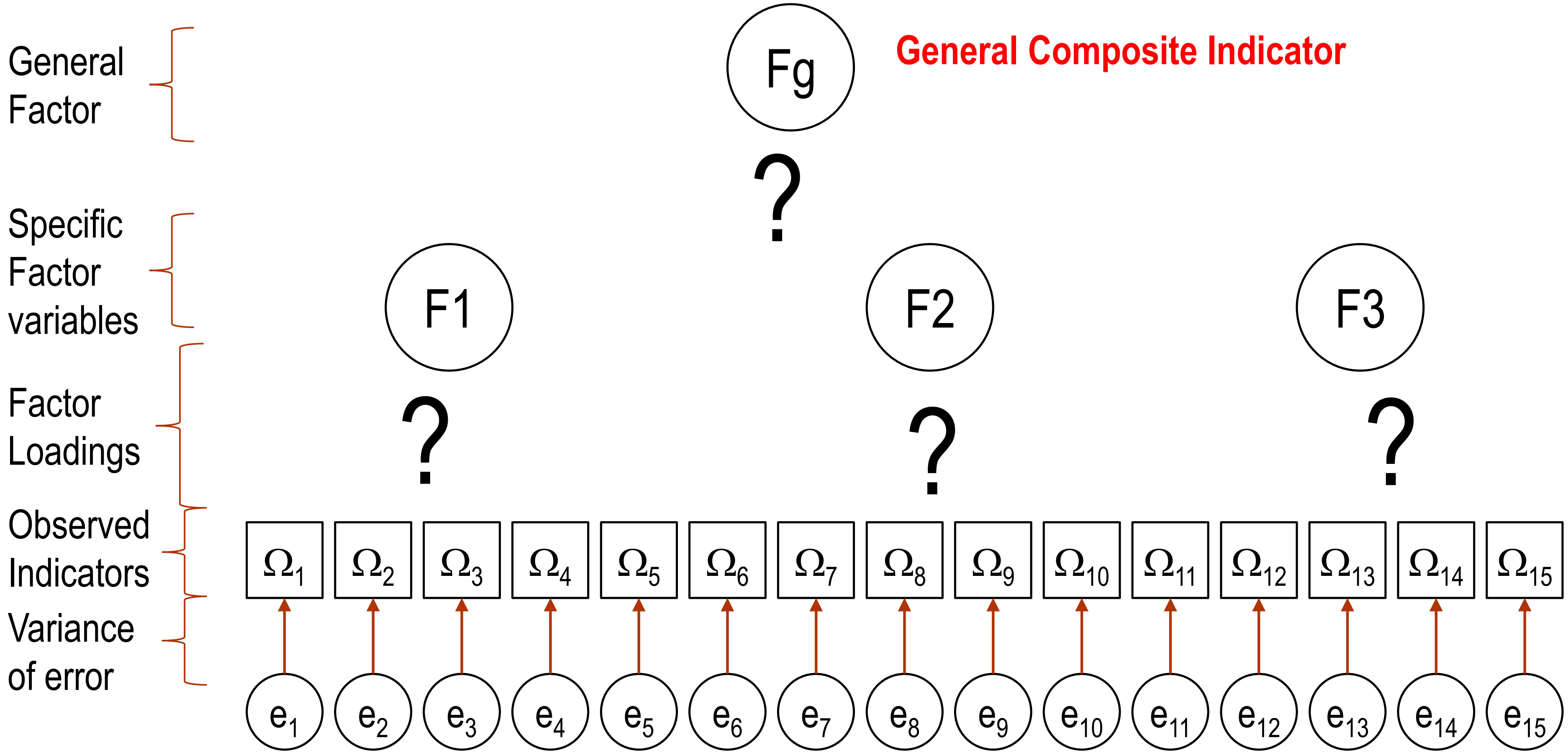
Two level Hierarchical Simple Structure Model

The **final model** obtained by the researcher is **only "partially" sustained by a theory**. The **modification** is **not 'optimised'** and thus the **model selection becomes** an **"artisanal skill"** of the researcher.



Two level Hierarchical Simple Structure Model

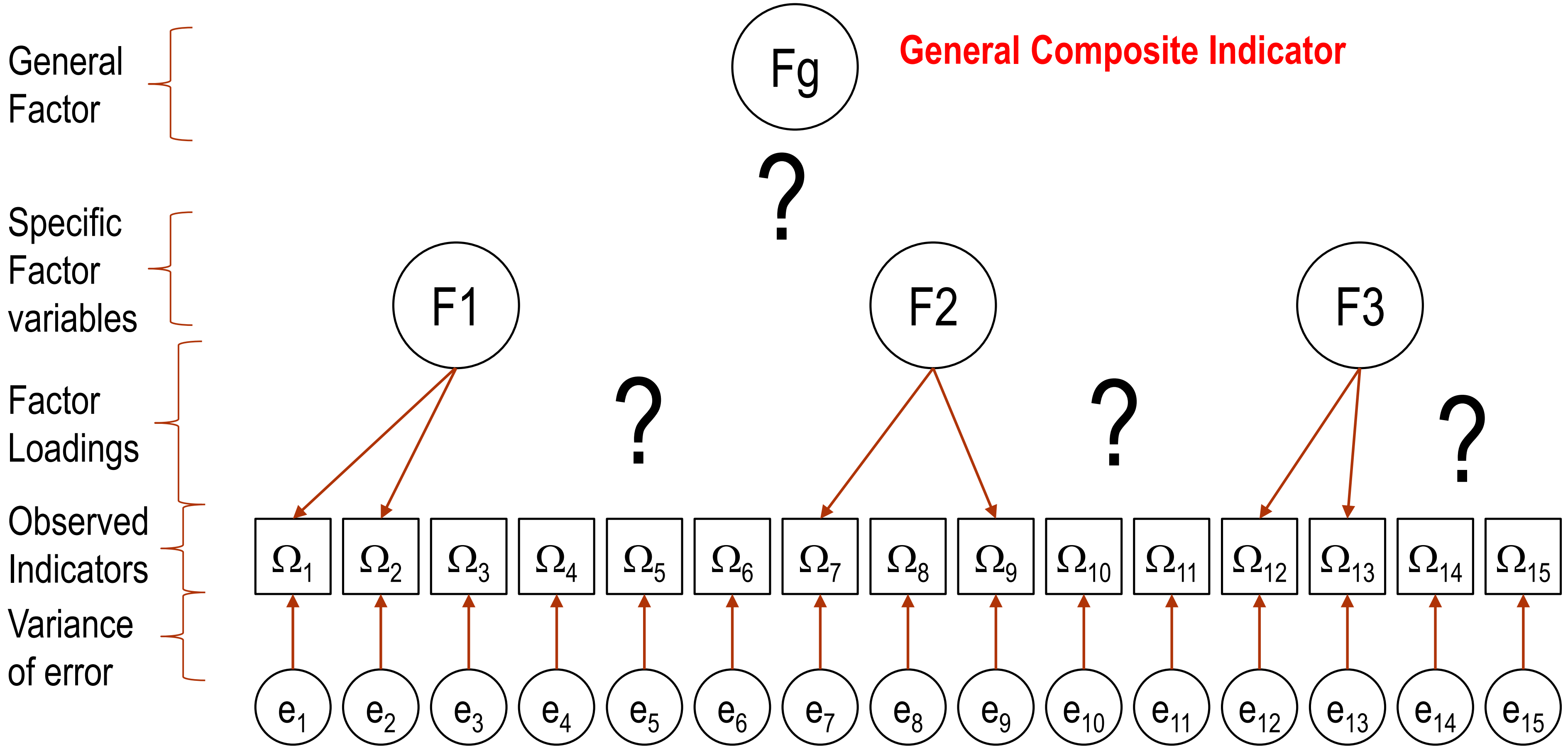
OUR PROPOSAL (1/3): Only the # of specific CIs is known



Association between CIs and between CIS and Indicators **are unknown**

Two level Hierarchical Simple Structure Model

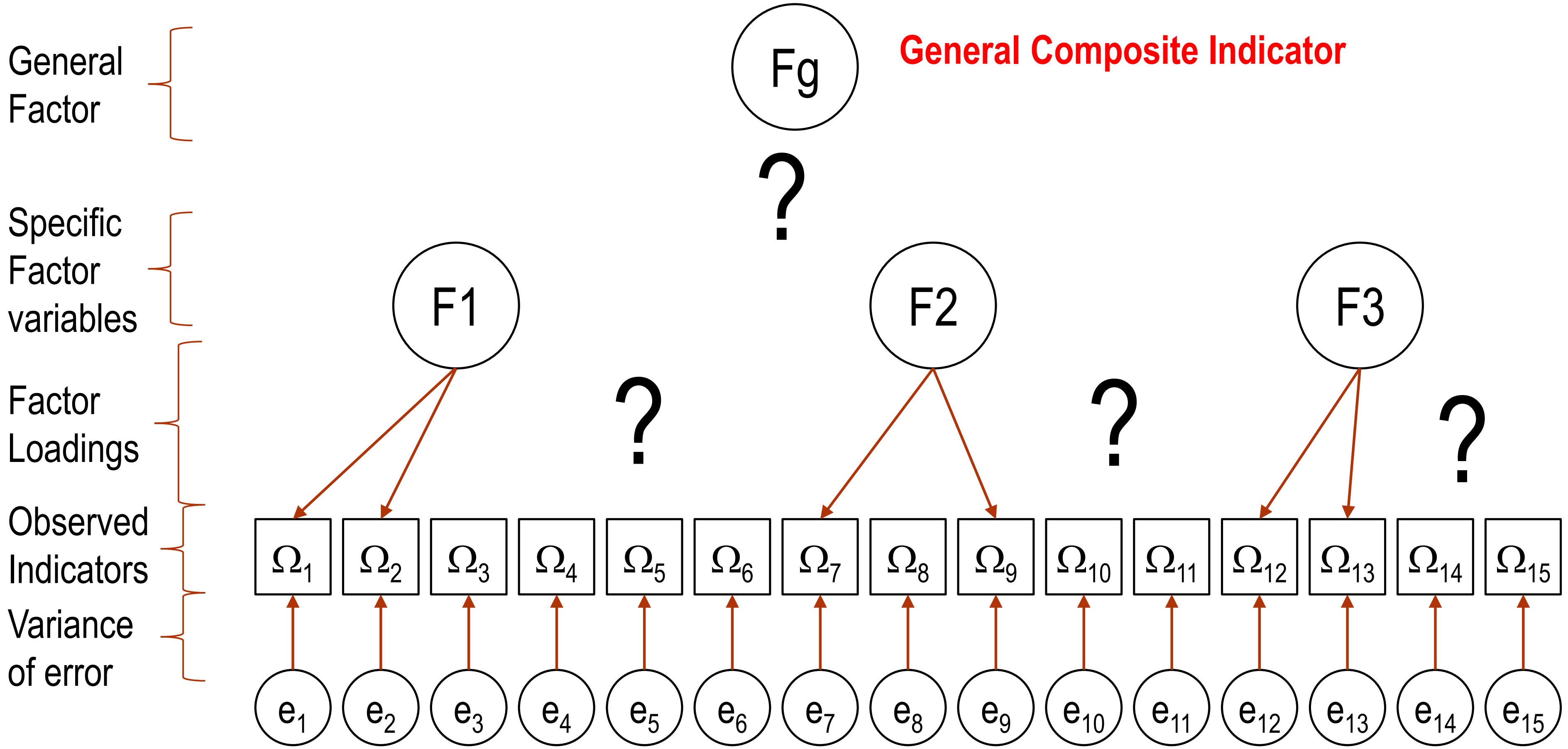
OUR PROPOSAL (2/3) SOME FLEXIBILITY: also part of associations are known, because these are sustained by a theory.



Associations between CIs and observed Indicators **are partially known**

Two level Hierarchical Simple Structure Model

OUR PROPOSAL (3/3) Statistical Coherence of correlations.
Weights must be non-negative, one-dimensionality and reliability of specific CIs



NON-NEGATIVE WEIGHTS (CORRELATIONS)

Hierarchical Disjoint Factor Analysis

a model to identify the latent Hierarchical Composite Indicator and set of specific Composite Indicators that best reconstruct the observed data

SOME METODOLOGICAL CONSIDERATIONS

Hierarchical Disjoint Factor Analysis

$$\mathbf{x} - \boldsymbol{\mu}_\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}_\mathbf{x}, \quad (\mathbf{y}\ \textbf{Specific factors}) \tag{1}$$

$$\mathbf{y} = \mathbf{c}g + \mathbf{e}_\mathbf{y}, \quad (g\ \textbf{General factor}) \tag{2}$$

Let include model (2) into model (1) and considering the loading matrix **A** is restricted to the product **A**=**BV** the **HDFA model** is defined

$$\mathbf{x} - \boldsymbol{\mu}_\mathbf{x} = \mathbf{BV}(\mathbf{c}g + \mathbf{e}_\mathbf{y}) + \mathbf{e}_\mathbf{x} = \mathbf{BV}\mathbf{c}g + \mathbf{BV}\mathbf{e}_\mathbf{y} + \mathbf{e}_\mathbf{x}. \tag{3}$$

Let rewrite the model in matrix form

$$\mathbf{X} = \mathbf{g}\mathbf{c}'\mathbf{V}'\mathbf{B} + \mathbf{E}_\mathbf{y}\mathbf{V}'\mathbf{B} + \mathbf{E}_\mathbf{x}. \tag{4}$$

with

$$\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{BV}(\mathbf{c}\frac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{c}' + \boldsymbol{\Psi}_\mathbf{y})\mathbf{V}'\mathbf{B} + \boldsymbol{\Psi}_\mathbf{x} = \mathbf{BV}\boldsymbol{\Sigma}_\mathbf{y}\mathbf{V}'\mathbf{B} + \boldsymbol{\Psi}_\mathbf{x}, \tag{5}$$

where $\boldsymbol{\Sigma}_\mathbf{y} = \mathbf{c}\frac{1}{n}(\mathbf{g}'\mathbf{g})\mathbf{c}' + \boldsymbol{\Psi}_\mathbf{y}.$ (6)

such that

$\mathbf{V} = [v_{jh} : \forall\ v_{jh} \in \{0,1\}]$ (binary) (7)

$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J$ (row stochastic) (8)

$\mathbf{B} = diag(b_1, \ldots, b_J)$ with $b_j^2 > 0$ (diagonal, non-null) (9)

$\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = diag(b_{.1}^2, \ldots, b_{.H}^2)$, with $b_{.h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0$ (orthogonal, non-empty) (10)

Estimation of HDFA

Minimization of the **discrepancy functions** w.r.t. **B, V** and **Ψ**

- **Least-Squares Estimation**

$$LSE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = \|\mathbf{S} - \mathbf{BV}(\mathbf{cc}' + \boldsymbol{\Psi}_\mathbf{y})\mathbf{V}'\mathbf{B} - \boldsymbol{\Psi}_\mathbf{x}\|^2 \rightarrow \min_{\mathbf{B},\mathbf{V},\mathbf{c},\boldsymbol{\Psi}\mathbf{x},\boldsymbol{\Psi}\mathbf{y}} \qquad (12)$$

- **Maximum likelihood Estimation**

$$MLE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = ln\left|\mathbf{BV}(\mathbf{cc}'+\boldsymbol{\Psi}_y)\mathbf{V}'\mathbf{B}+\boldsymbol{\Psi}_x\right| - ln\,|\mathbf{S}| + tr\left(\left(\mathbf{BV}(\mathbf{cc}'+\boldsymbol{\Psi}_y)\mathbf{V}'\mathbf{B}+\boldsymbol{\Psi}_x\right)^{-1}\mathbf{S}\right) - J \rightarrow \min_{\mathbf{B},\mathbf{V},\mathbf{c},\boldsymbol{\Psi}\mathbf{x},\boldsymbol{\Psi}\mathbf{y}} \qquad (12')$$

- **Generalised Least-Squares Estimation**

$$GLSE(\mathbf{B},\mathbf{V},\boldsymbol{\Psi}) = \|(\mathbf{S} - \mathbf{BV}(\mathbf{cc}' + \boldsymbol{\Psi}_\mathbf{y})\mathbf{V}'\mathbf{B} - \boldsymbol{\Psi}_\mathbf{x})\mathbf{S}^{-1/2}\|^2 \rightarrow \min_{\mathbf{B},\mathbf{V},\boldsymbol{\Psi}} \qquad (12'')$$

such that

$\mathbf{V} = [\, v_{jh} : \forall\, v_{jh} \in \{0,1\}]$	(binary)	(13)
$\mathbf{V}\mathbf{1}_H = \mathbf{1}_J$	(row stochastic)	(14)
$\mathbf{B} = diag(b_1, \ldots, b_J)$ with $b_j^2 > 0$	(diagonal, non-null)	(15)
$\mathbf{V}'\mathbf{BBV} = diag(b_{.1}^2, \ldots, b_{.H}^2)$, with $b_{.h}^2 = \sum_{j=1}^{J} b_{jh}^2 > 0$	(orthogonal, non-empty)	(16)

A coordinated descendent algorithm has been developed this problem.

NOTE: This is a discrete and continuous problem that cannot be solved by a quasi-Newton type algorithm

Hierarchical Disjoint Non-Negative Factor Analysis

The **General factor frequently corresponds** to a **Composite indicator** where **each subset of variable is consistent and reliable, thus the loadings must be positive**

Recall that the discrepancy function $D(\mathbf{B}, \hat{\mathbf{V}}, \hat{\Psi})$ is minimized with respect to $\mathbf{B}_h = diag(\mathbf{b}_h)$ by

$$\widehat{\boldsymbol{b}_h} = \hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{u}_{1h} (\lambda_{1h} - 1)^{\frac{1}{2}} \quad h=1,\ldots,H.$$

where λ_{1h} is the largest eigenvalue and $\mathbf{u}_{1h}$ is the corresponding eigenvector of the variance-covariance matrix $\hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}}$ corresponding to variables identified by $\mathbf{v}_{.h}$, that corresponds to h-th column of $\mathbf{V}$. Values λ_{1h} and $\mathbf{u}_{1h}$ minimize $\|\hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{S}_h \hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \lambda_{1h} \mathbf{u}_{1h} \mathbf{u}'_{1h}\|^2$, or equivalently

$$\| \mathbf{X}_h \hat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}'_{1h} \|^2, \tag{17}$$

where $\mathbf{X}_h$ is the centered data matrix formed by variables identified by $\mathbf{v}_{.h}$ and $\mathbf{y}_h$ is the factor score vector.

$$\| \mathbf{X}_h \, \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} - \sqrt{\lambda_{1h}} \mathbf{y}_h \mathbf{u}'_{1h} \|^2, \qquad (18)$$

SOLUTION OF a REGRESSION PROBLEM

can be solved by an **ALS algorithm that alternates two regression problems**.

Given $\widehat{\mathbf{u}}_{1h}$ compute $\mathbf{y}_h$ by

$$\mathbf{y}_h = \mathbf{X}_h \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \widehat{\mathbf{u}}_{1h} (\widehat{\mathbf{u}}'_h \widehat{\mathbf{u}}_{1h})^{-1}. \qquad (19)$$

Given $\widehat{\mathbf{y}}_h$ compute $\mathbf{u}_{1h}$ by

$$\mathbf{u}_{1h} = \widehat{\Psi}_h^{-\frac{1}{2}} \mathbf{X}'_h \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}'_h \widehat{\mathbf{y}}_h)^{-1}. \qquad (20)$$

At each reiteration of the two steps (19) and (20), the loss function (18) decreases or at least does not increase. The algorithm stops when function (18) decreases less than a positive arbitrary constant.

Now **vector $\mathbf{u}_{1h}$ must be non-negative**.

The **solution** can be **found** by the **Non-Negative LS Algorithm** (Lawson and Hanson, 1974).
This is an **active set algorithm**, where the H inequality constraints are active if $\mathbf{u}'_{1h}$ are negative (or zero) when estimated unconstrained, otherwise constraints are passive.
The non-negative **solution** of (18) with respect to $\mathbf{u}_{1h}$ is the **unconstrained least squares solution using only the variables of the passive set**, setting the regression coefficients of the active set to zero. Therefore

$$\mathbf{u}_{1h} = \begin{cases} \widehat{\Psi}_{\mathbf{x}h}^{-\frac{1}{2}} \mathbf{X}'_{h+} \widehat{\mathbf{y}}_h (\widehat{\mathbf{y}}'_h \widehat{\mathbf{y}}_h)^{-1} \\ 0 \quad \text{otherwise} \end{cases}. \qquad (21)$$

where $\mathbf{X}_{h+}$ is the set of passive variables.

Application to Sustainable Development Goals

SDGs Europe: 100 Indicators, 17 Goals

Goal1: 6 on 6
- People at risk of poverty or social exclusion 01.11
- People at risk of poverty after social transfers 01.12
- Severely materially deprived people 01.13
- People living in households with very low work intensity 01.14
- Housing cost overburden rate 01.21
- Share of total population living in a dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames or floor 01.22

Goal2: 4 on 6
- Obesity rate 02.11
- Agricultural factor income per annual work unit (AWU) 02.21
- Government support to agricultural research and development 02.26
- Area under organic farming 02.31
- Ammonia emissions from agriculture 02.52
- Gross nutrient balance on agricultural land 02.54

Goal3: 5 on 6
- Life expectancy at birth 03.11
- Self-perceived health 03.14
- Death rate due to chronic diseases 03.25
- Suicide death rate 03.31
- Smoking prevalence 03.36
- Self-reported unmet need for medical examination and care 03.41

Goal4: 4 on 6
- Early childhood education and care 04.10
- Early leavers from education and training 04.20
- Tertiary educational attainment 04.30
- Employment rate of recent graduates 04.31
- Adult participation in learning 04.40
- Underachievement in reading, maths and science 04.50

Goal5: 2 on 6
- Gender pay gap 05.10
- Gender employment gap 05.12
- Proportion of seats held by women in national parliaments and local government 05.20
- Proportion of women in senior management positions 05.21
- Physical and sexual violence by a partner or a non-partner 05.33
- Inactivity rates due to caring responsibilities 05.44

Goal6: 4 on 6
- Share of total population having neither a bath, nor a shower, nor indoor flushing toilet in their household 06.11
- Population connected to urban wastewater treatment with at least secondary treatment 06.13
- Biochemical oxygen demand in rivers 06.21
- Nitrate in groundwater 06.24
- Phosphate in rivers 06.26
- Water exploitation index (WEI) 06.41

Goal7: 4 on 6
- Percentage of people affected by fuel poverty (inability to keep home adequately warm) 07.10
- Share of renewable energy in gross final energy consumption 07.20
- Primary energy consumption; final energy consumption by sector 07.30
- Final energy consumption in households per capita 07.32
- Energy dependence 07.33
- Energy productivity 07.35

Goal8: 4 on 6
- Real GDP per capita - growth rate 08.10
- Young people neither in employment nor in education and training 08.20
- Total employment rate 08.30
- Long-term unemployment rate 08.31
- Involuntary temporary employment 08.35
- Fatal accidents at work by sex (NACE Rev. 2, A, C-N) - Unstandardised incidence rate 08.60

Goal9: 1 on 6

- Gross domestic expenditure on R&D 09.10
- Employment in high- and medium-high technology manufacturing sectors and knowledge-intensive service sectors 09.11
- Total R&D personnel 09.13
- Patent applications to the European Patent Office (EPO) 09.14
- Share of collective transport modes in total passenger land transport 09.40
- Share of rail and inland waterways activity in total freight transport 09.41

Goal10: 2 on 6

- GDP per capita in PPS 10.10
- Real adjusted gross disposable income of households per capita in PPS 10.11
- Relative median at-risk-of-poverty gap 10.22
- Gini coefficient of equivalised disposable income 10.24
- Income growth of the bottom 40 per cent of the population and the total population 10.25
- Number of first time asylum applications (total and accepted) per capita 10.31

Goal11: 6 on 6

- Overcrowding rate by degree of urbanisation 11.12
- Distribution of population by level of difficulty in accessing public transport 11.21
- People killed in road accidents 11.25
- Urban population exposure to air pollution by particulate matter 11.31
- Proportion of population living in households considering that they suffer from noise 11.36
- Recycling rate of municipal waste 11.52

Goal12: 3 on 6

- Generation of waste excluding major mineral wastes 12.10
- Recycling and landfill rate of waste excluding major mineral wastes 12.11
- Consumption of toxic chemicals 12.30
- Resource productivity 12.40
- Average CO2 emissions per km from new passenger cars 12.51
- Volume of freight transport relative to GDP 12.54

Goal13: 3 on 6

- Greenhouse gas emissions (indexed totals and per capita) 13.11
- Greenhouse gas emissions intensity of energy consumption 13.14
- Global (and European) near surface average temperature 13.21
- Economic losses caused by climate extremes (consider climatological, hydrological, meteorological) 13.45
- Contribution to the 100bn international commitment on climate related expending (public finance) 13.51
- Share of EU population covered by the new Covenant of Mayors for Climate and Energy (integrating mitigation, adaptation, and access to clean and affordable energy) 13.63

Goal14: 2 on 5

- Bathing water quality 14.13
- Sufficiency of marine sites designated under the EU habitats directive 14.21
- Ocean acidification (CLIM 043) 14.31
- Catches in major fishing areas 14.41
- Assessed fish stocks exceeding fishing mortality at maximum sustainable yield (Fmsy) 14.43

Goal15: 3 on 6

- Forest area as a proportion of total land area 15.11
- Artificial land cover per capita 15.11
- Change in artificial land cover per year 15.24
- Common bird index 15.31
- Sufficiency of terrestrial sites designated under the EU habitats directive 15.32
- Estimated soil erosion by water 15.41

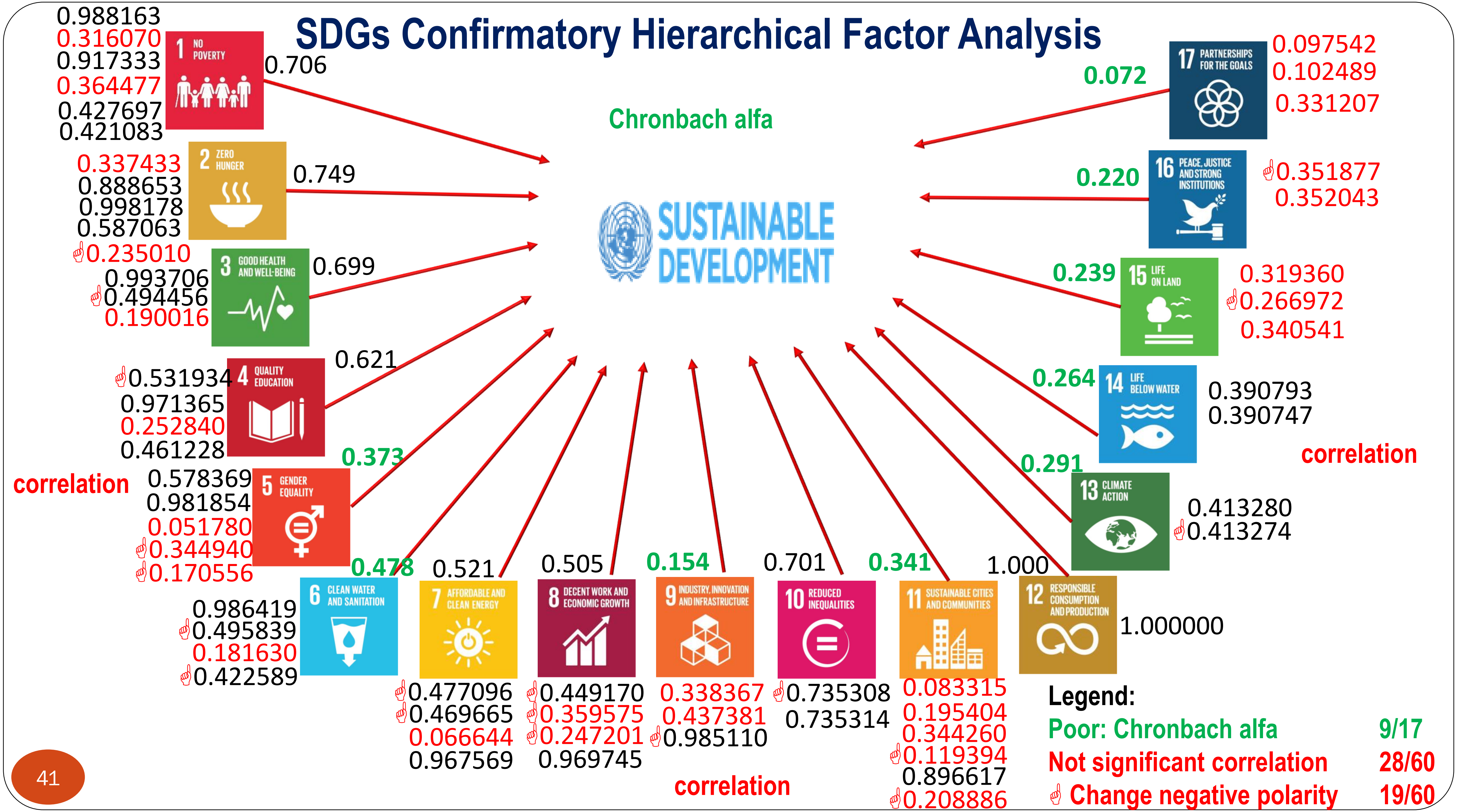
Goal16: 2 on 6

- Death due to homicide, assault, by sex 16.10 (tps00146)
- Share of population which reported occurrence of crime, violence or vandalism in their area 16.19
- General government total expenditure on law courts 16.32
- Corruption Perception Index 16.50
- Perceived independence of the justice system 16.61
- Level of citizens' confidence in EU institutions 16.62

Goal17: 5 on 5

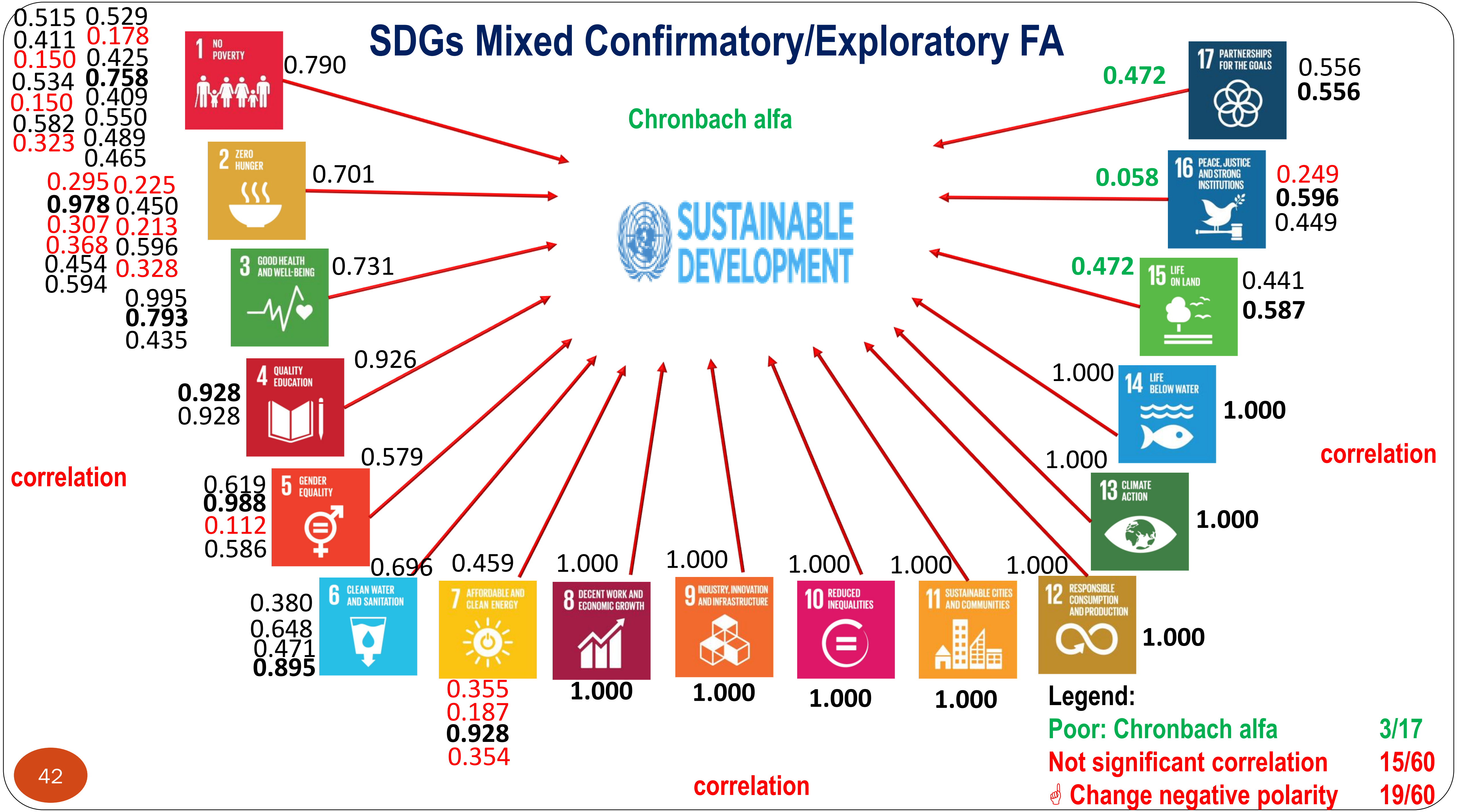
- Official development assistance as share of gross national income 17.10
- EU financing for developing countries 17.11
- EU Imports from developing countries 17.12
- General government gross debt 17.13
- Shares of environmental and labour taxes in total tax revenues 17.19

SDGs Confirmatory Hierarchical Factor Analysis



Legend:

Poor: Chronbach alfa	9/17
Not significant correlation	28/60
☝ Change negative polarity	19/60



SDGs Mixed Confirmatory/Exploratory FA
Chronbach alfa
SUSTAINABLE DEVELOPMENT
0.515 0.529
0.411 0.178
0.150 0.425
0.534 0.758
0.150 0.409
0.582 0.550
0.323 0.489
0.465
1 NO POVERTY
0.790
0.295 0.225
0.978 0.450
0.307 0.213
0.368 0.596
0.454 0.328
0.594
2 ZERO HUNGER
0.701
0.995
0.793
0.435
3 GOOD HEALTH AND WELL-BEING
0.731
0.928
0.928
4 QUALITY EDUCATION
0.926
correlation
0.619
0.988
0.112
0.586
5 GENDER EQUALITY
0.579
0.380
0.648
0.471
0.895
6 CLEAN WATER AND SANITATION
0.696
7 AFFORDABLE AND CLEAN ENERGY
0.459
0.355
0.187
0.928
0.354
8 DECENT WORK AND ECONOMIC GROWTH
1.000
1.000
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
1.000
1.000
10 REDUCED INEQUALITIES
1.000
1.000
11 SUSTAINABLE CITIES AND COMMUNITIES
1.000
1.000
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
1.000
1.000
correlation
13 CLIMATE ACTION
1.000
1.000
14 LIFE BELOW WATER
1.000
1.000
15 LIFE ON LAND
0.472
0.441
0.587
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
0.058
0.249
0.596
0.449
17 PARTNERSHIPS FOR THE GOALS
0.472
0.556
0.556
correlation
Legend:
Poor: Chronbach alfa 3/17
Not significant correlation 15/60
Change negative polarity 19/60

Can we integrate and reduce the number of Development Goals?

SDGs Indicators: Best Model

Living condition

Factor 1:
- No Poverty
- Zero Hunger
- Good Health and Well-Being
- Clean Water and Sanitation
- Reduce Inequalities
- Responsible Consumption and Production

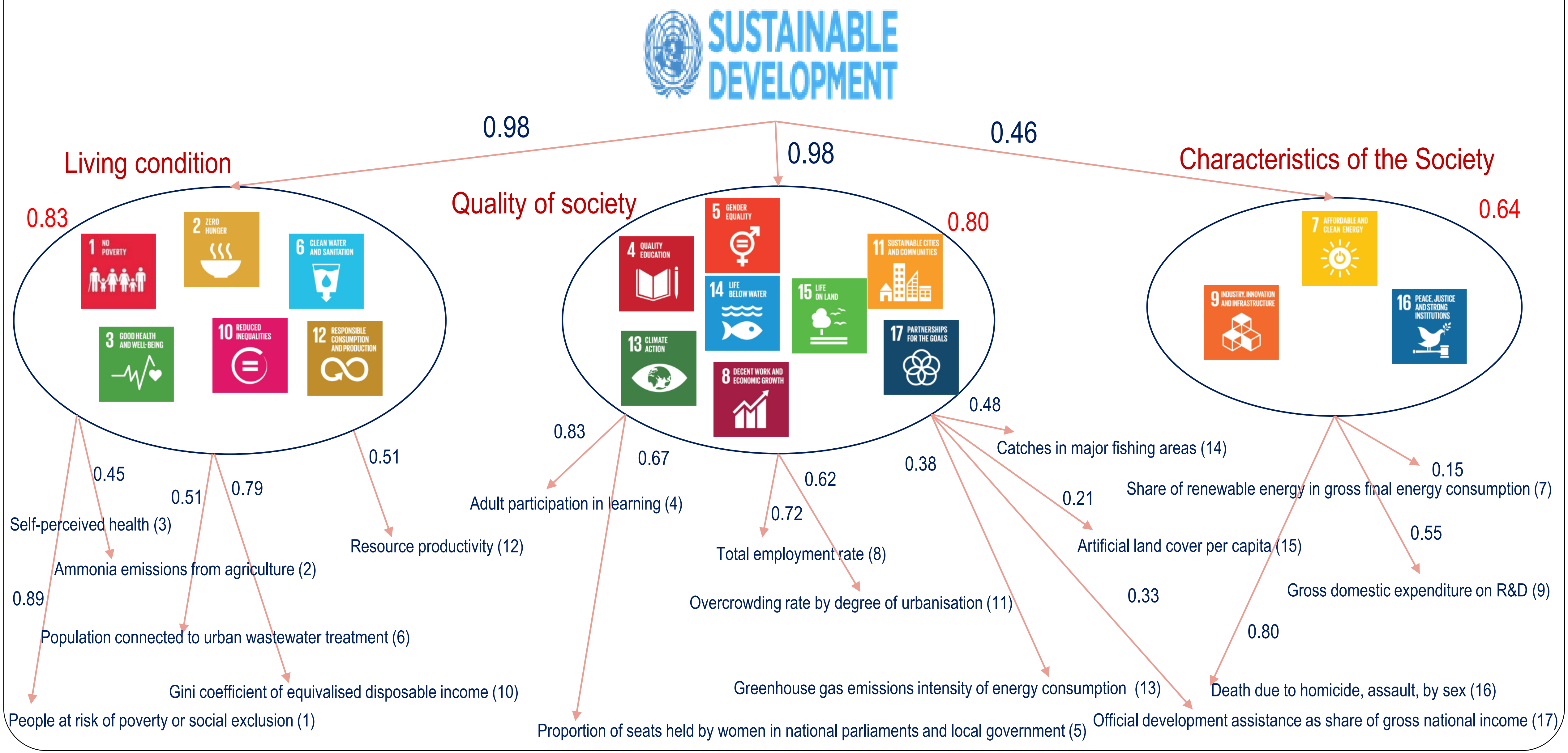
Quality of society

Factor 2:
- Quality Education
- Gender Equality
- Decent Work and Economic Growth
- Sustainable Cities and Communities
- Climate Action
- Life Below Water
- Life on land
- Partnership for the goals

Characteristics of the Society

Factor 3:
- Industry, Innovation and Infrastructure
- Affordable and Clear Energy
- Peace, Justice and Strong Institutions

SDGs Indicators: Best Model



SDGs Indicators: Best Model

Ranking of 28 countries of EU based on Sustainable Development:

Sweden	1.000
Netherlands	0.986
Luxembourg	0.985
Denmark	0.944
Germany	0.881
France	0.832
UK	0.803
Austria	0.798
Finland	0.791
Belgium	0.757
Ireland	0.756
Malta	0.657
Spain	0.616
Italy	0.585
Czech Republic	0.564
Slovenia	0.517
Portugal	0.453
Republic of Cyprus	0.447
Poland	0.396
Estonia	0.379
Slovakia	0.354
Hungary	0.268
Lithuania	0.221
Croatia	0.212
Greece	0.198
Latvia	0.128
Bulgaria	0.066
Romania	0.000

- The rank correlation* between Sustainable Development index and Better Life Index (by OECD) is 0.95

*computed only for the 22 common countries

SDGs Indicators: Best Model

Ranking of 28 countries of EU based on Sustainable Development divided in six clusters:

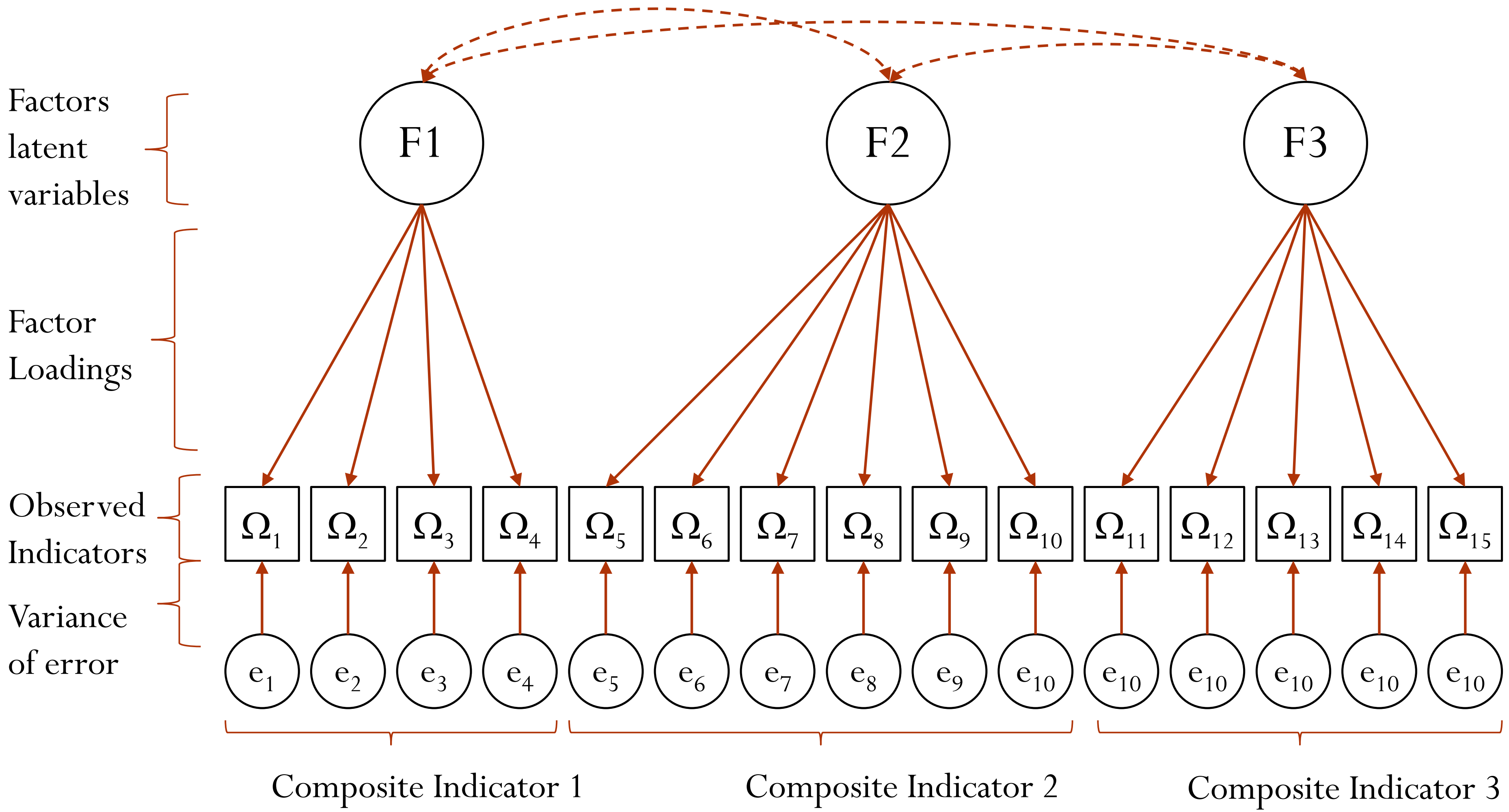
Sweden	1.000	1
Netherlands	0.986	
Luxembourg	0.985	
Denmark	0.944	
Germany	0.881	2
France	0.832	
UK	0.803	
Austria	0.798	
Finland	0.791	
Belgium	0.757	
Ireland	0.756	
Malta	0.657	3
Spain	0.616	
Italy	0.585	
Czech Republic	0.564	
Slovenia	0.517	
Portugal	0.453	4
Republic of Cyprus	0.447	
Poland	0.396	
Estonia	0.379	
Slovakia	0.354	
Hungary	0.268	5
Lithuania	0.221	
Croatia	0.212	
Greece	0.198	
Latvia	0.128	6
Bulgaria	0.066	
Romania	0.000	

Thank you for your kind attention.

Cross-Loadings

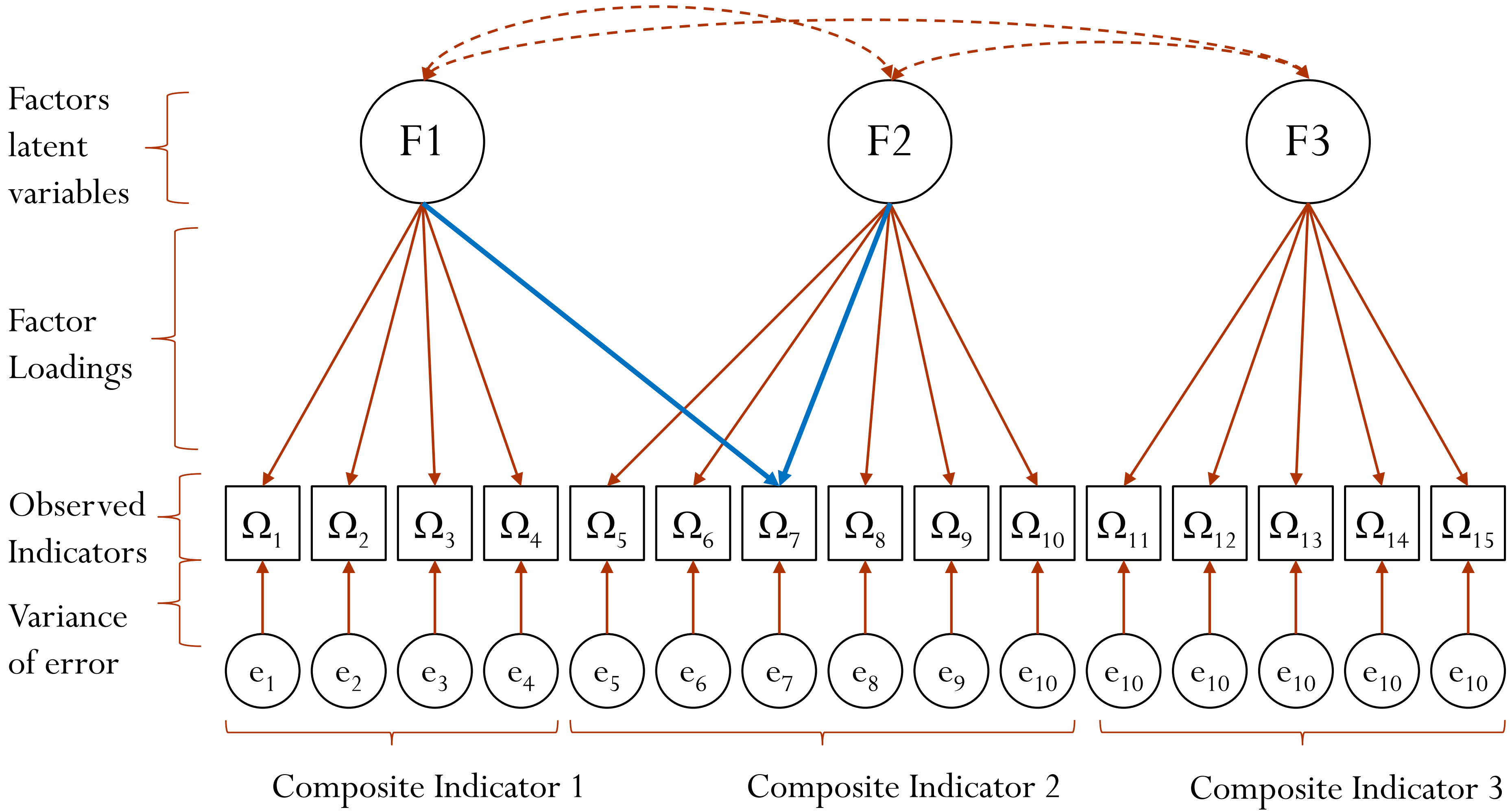
The fit of the SSM may be poor: for the uncorrect choice of the number of factors
for the presence of cross-loadings

PROCEDURE: FIRST ESTIMATE the best SSM



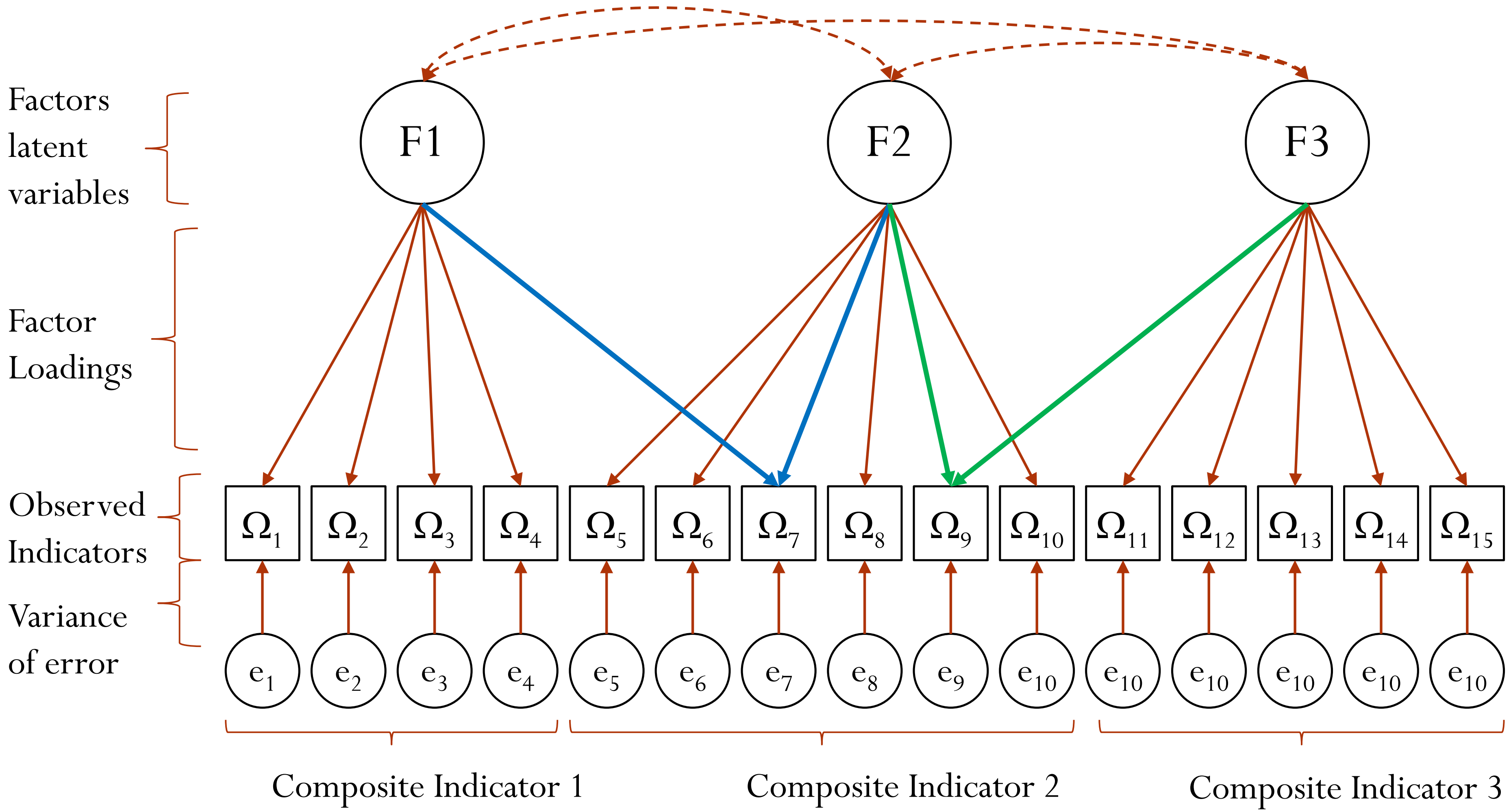
Cross-Loadings

- IDENTIFY the Cross-Loanding that most reduce BIC



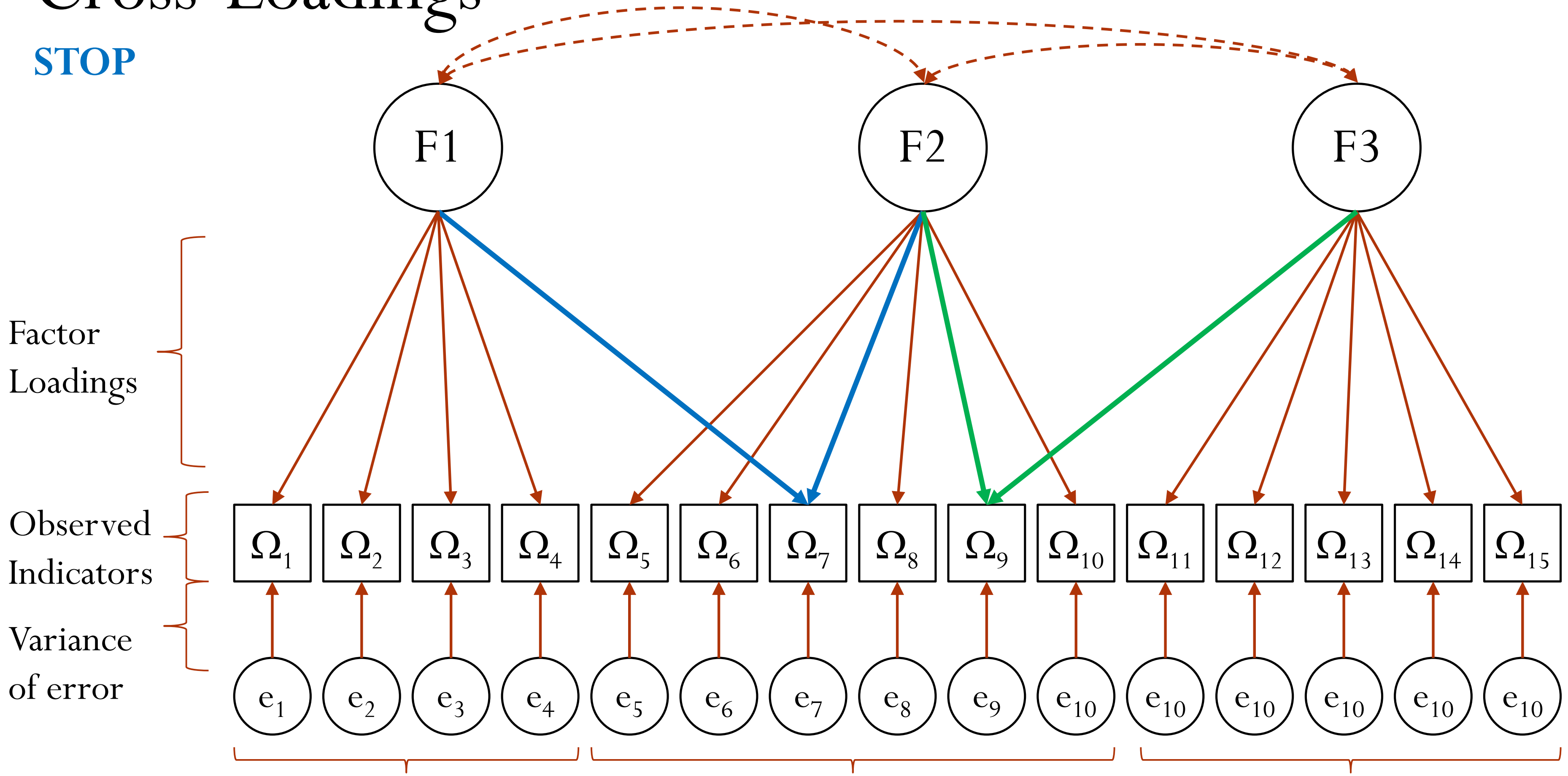
Cross-Loadings

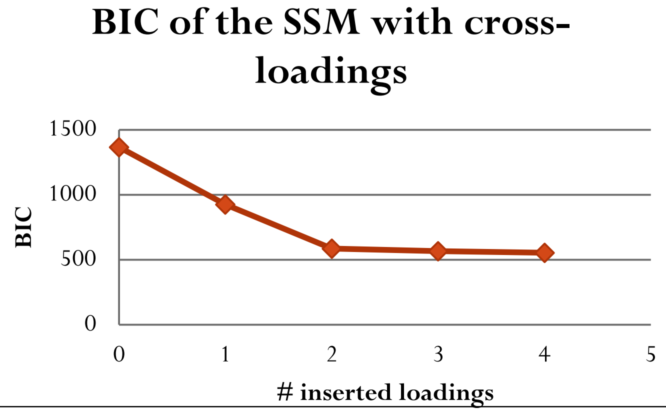
Continue to IDENTIFY Cross-Loadings that reduce BIC



Cross-Loadings

STOP

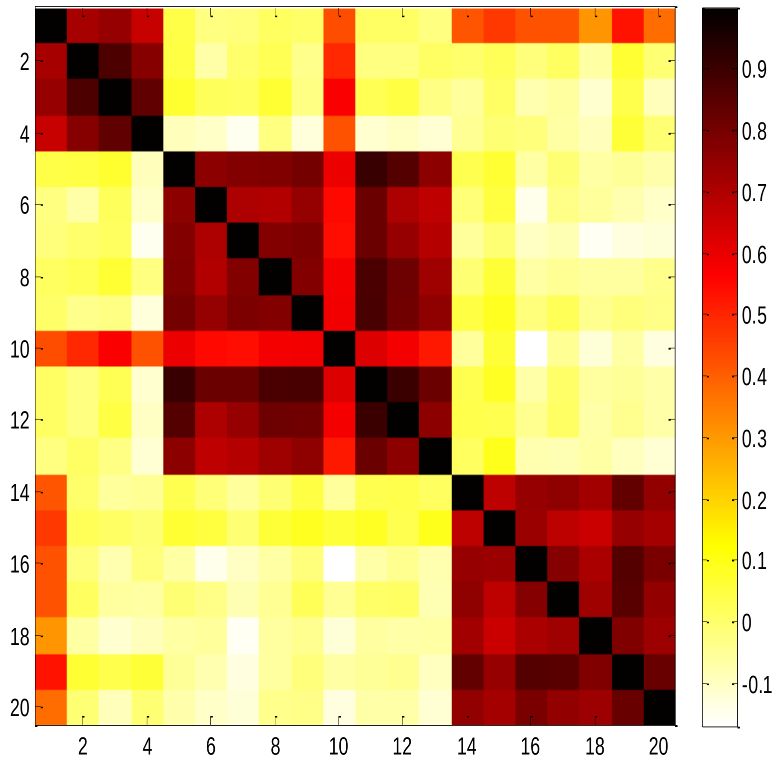




Cross-Loadings

Example: 20 x 20 correlation matrix

3 blocks generated according DFA + cross-loadings $a_{10,1}$, $a_{1,3}$



Estimation:

first estimate the best Simple Structure Model [3 blocks (V1-V4), (V5-13), (V14-V20)]

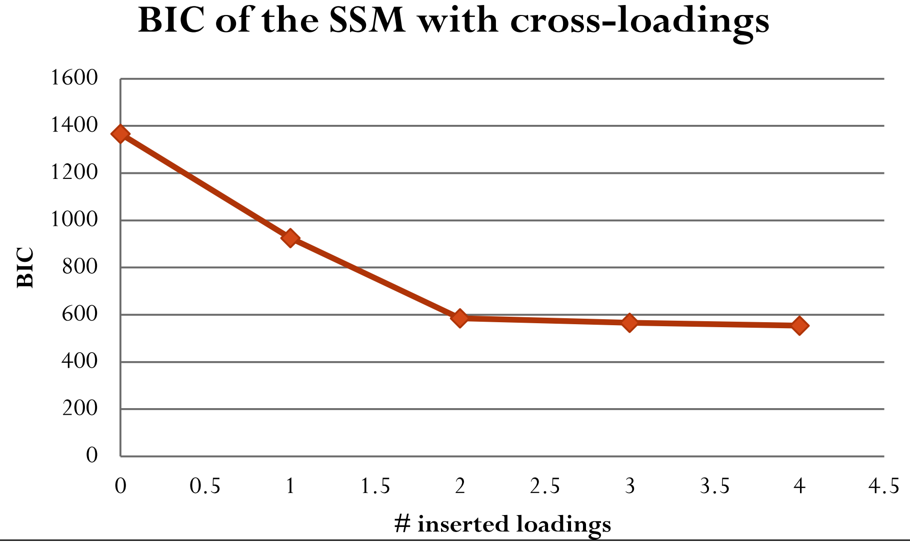
BIC = 1366

First cross-loading estimated $a_{10,1}$ =0.57;

BIC = 925

Second cross-loading estimated $a_{1,3}$ =0.52;

BIC = 585



SDGs Indicators: Best Model

Living condition

Factor 1:
- 1. No Poverty
- 3. Good Health and Well-Being
- 5. Gender Equality
- 6. Clean Water and Sanitation
- 12. Responsible Consumption and Production
- 13. Climate Action
- 14. Life Below Water

Quality of society

Factor 2:
- 2. Zero Hunger
- 4. Quality Education
- 8. Decent Work and Economic Growth
- 11. Sustainable Cities and Communities
- 15. Life on land
- 17. Partnership for the goals

Society

Factor 3:
- 7. Affordable and Clear Energy
- 9. Industry, Innovation and Infrastructure
- 10. Reduce Inequalities
- 16. Peace, Justice and Strong Institutions

SDGs Indicators: Best Model