Causal Discovery with Latent Variables: the Measurement Problem

Richard Scheines
Carnegie Mellon University

Peter Spirtes, Clark Glymour, Ricardo Silva,
Steve Riese, Erich Kummerfeld, Renjie Yang, Max Mansolf

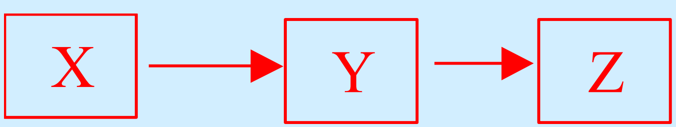
Outline

1. Measurement Error, Coarsening and
   Conditional Independence
2. Conventional Strategies
3. Latent Variable Models to the Rescue
4. The Problem of Impurity
5. Strategies for Handling Impurity (Rank Constraints)
6. Application to Psychometric Models

Causal Structure

Testable Statistical Predictions

Causal Graphs



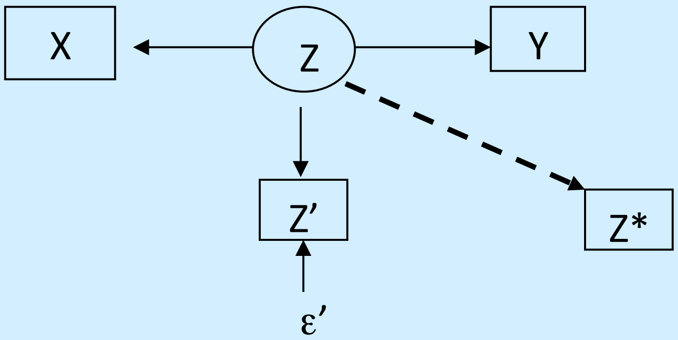
e.g., Conditional Independence

X _||_ Z | Y

$\forall x,y,z \;\; P(X = x, Z=z \mid Y=y) =$

$P(X = x \mid Y=y)\, P(Z=z \mid Y=y)$

Measurement Error and Coarsening Endanger conditional Independence



$X \perp\!\!\!\perp Y \mid Z$

$Z' = Z + \varepsilon$

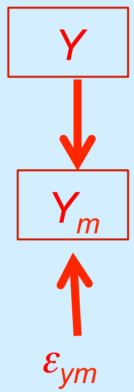
Measurement Error

Coarsening:
$-\infty < Z < 0 \rightarrow Z^* = 0$
$0 \leq Z < i \rightarrow Z^* = 1$
$i \leq Z < j \rightarrow Z^* = 2$
..
$k \leq Z < \infty \rightarrow Z^* = k$

$X \not\perp\!\!\!\perp Y \mid Z'$ (unless $Var(\varepsilon') = 0$)

$X \not\perp\!\!\!\perp Y \mid Z^*$ (almost always)

Parameterizing Measurement Error



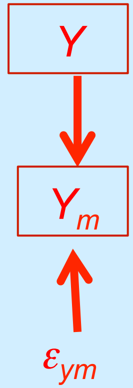
$Y_m := Y + \varepsilon_{ym}$ $\varepsilon_{ym} \sim N(0, \sigma_2{}^2)$

$Measurement\_Error = \varepsilon_{Ym}$

$Var(Y_m) = Var(Y) + Var(\varepsilon_{Ym})$

Amount of Measurement Error $= Var(eYm)/Var(Ym) = Var(Ym) - Var(Y)/Var(Ym)$

Unstandardized

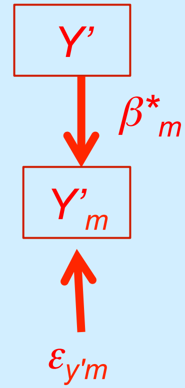


$Y_m := Y + \varepsilon_{ym}$

$\varepsilon_{ym} \sim N(0, \sigma^2)$

Measurement Error = $Var(eYm)/Var(Ym)$

Standardized



$Y'_m := \beta^*_m Y' + \varepsilon_{y'm}$

$\varepsilon_{y'm} \sim N(0, \sigma'^2)$

$Var(Y) = Var(Y'_m) = 1.0$

$Var(\varepsilon_{y'm}) = \sigma'^2 = 1 - \beta^{*2}_m$

$\beta^*_m = \rho(Y', Y'_m)$

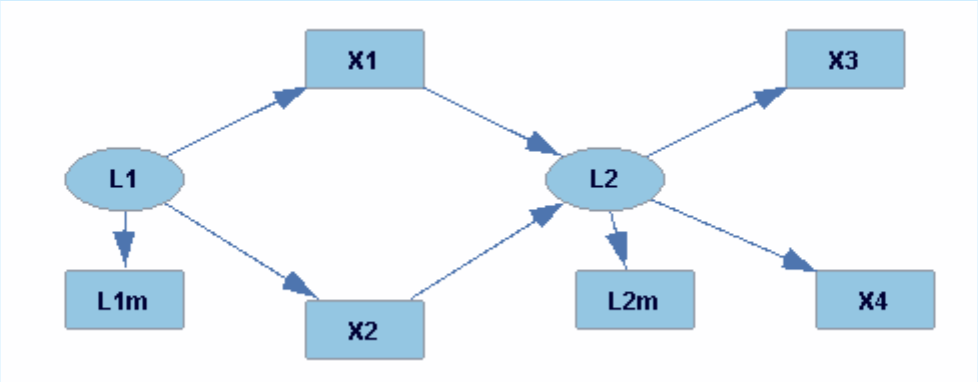
$Var(\varepsilon_{y'm}) = 1 - \rho(Y', Y'_m)^2$

Measurement Error = $Var(eY'm)/Var(Y'm)$

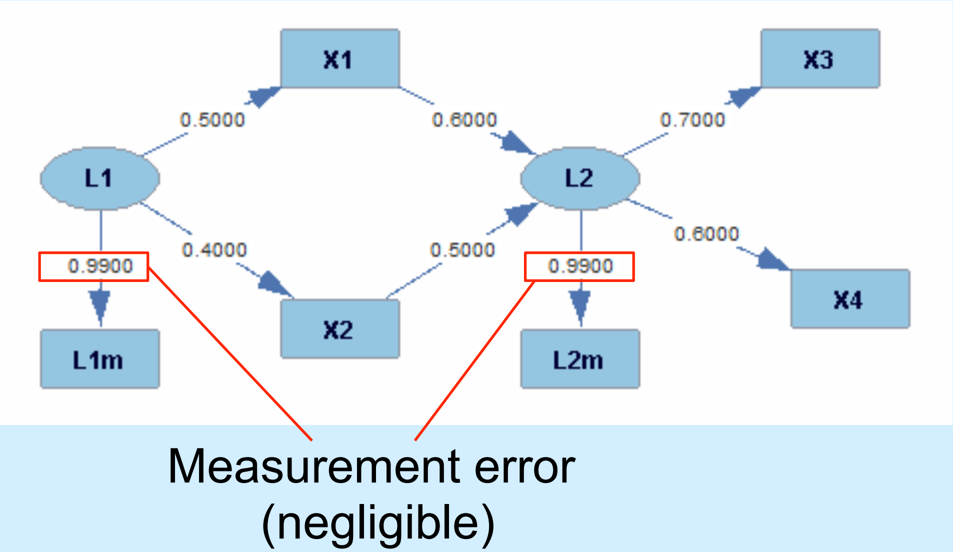
Measurement Error = $1 - r(Y', Y'm)2/1.0$

Measurement Error

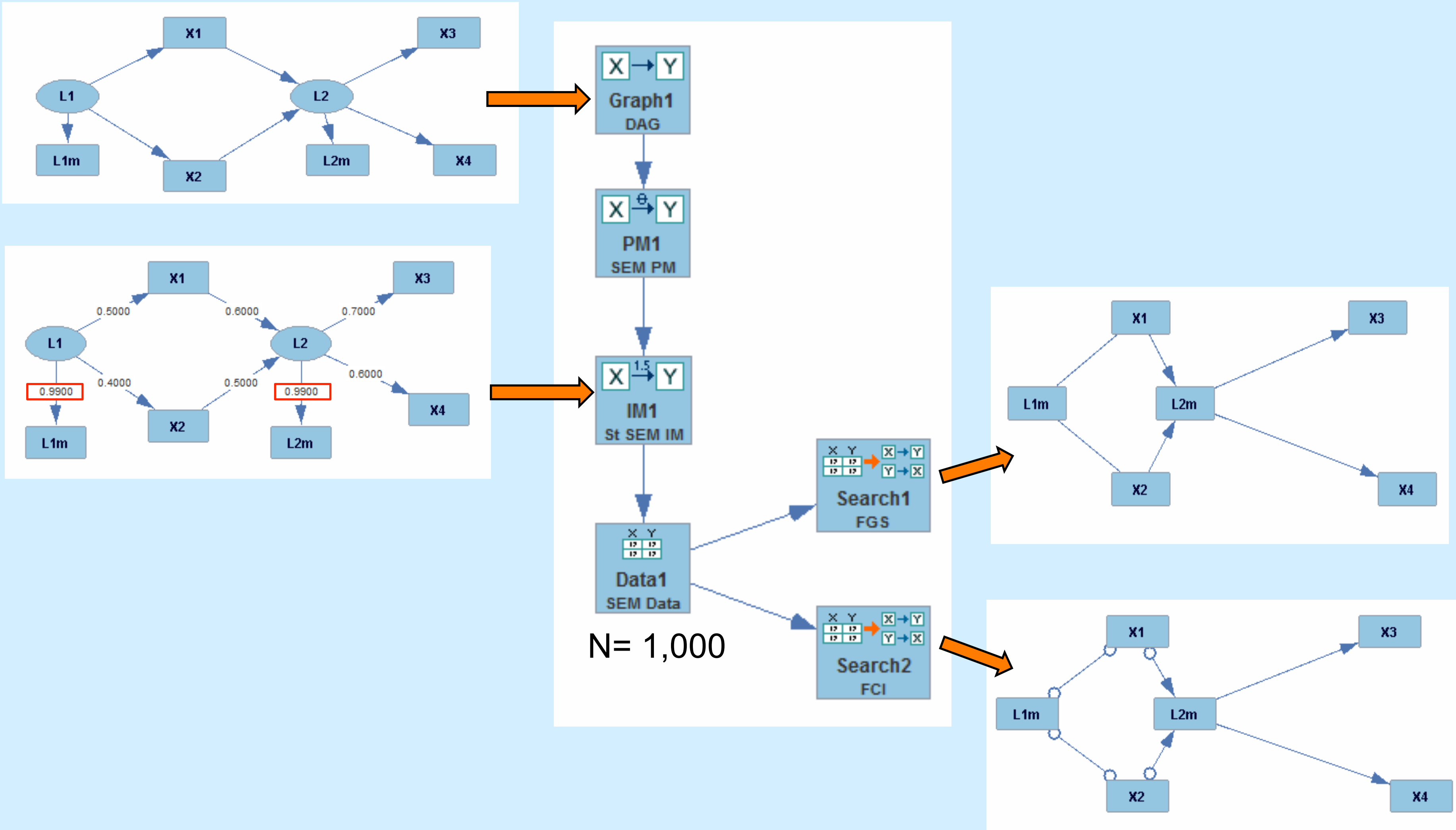
True Graph



Parameterization 1:
Standardized SEM IM

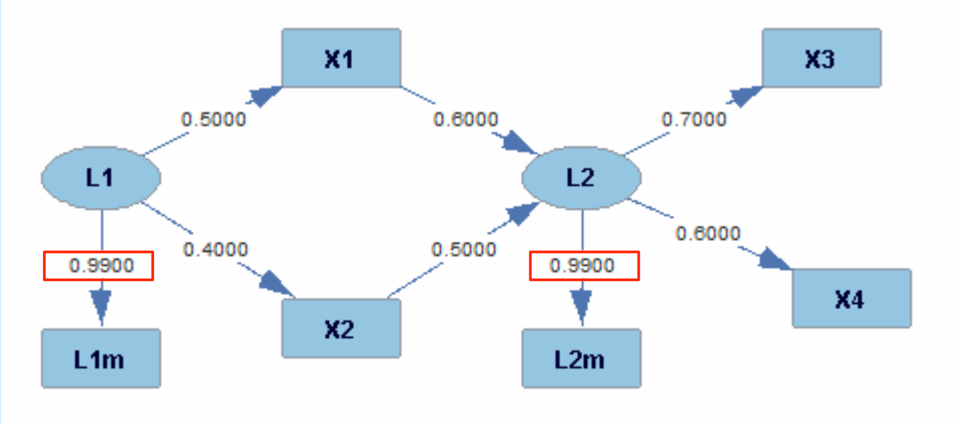


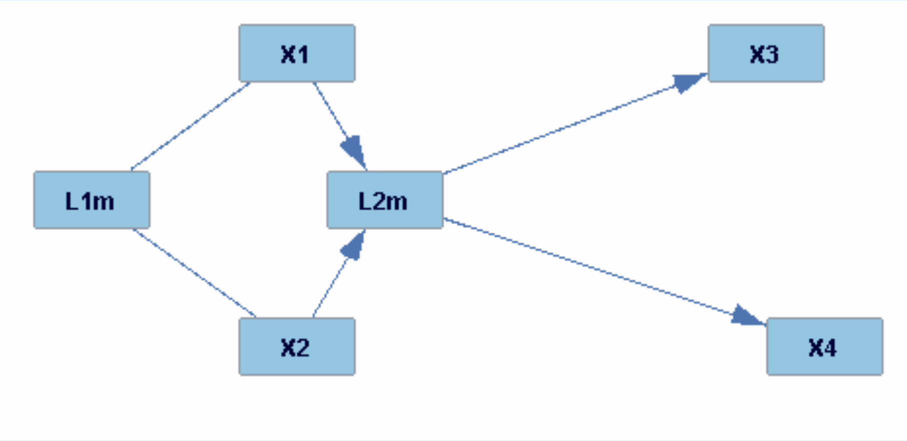
Measurement Error



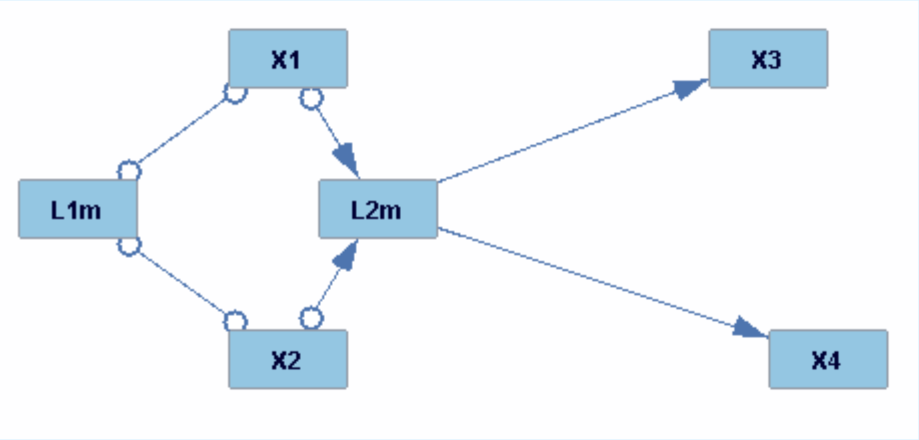
Measurement Error

Parameterization 1:
Negligible
Measurement Error





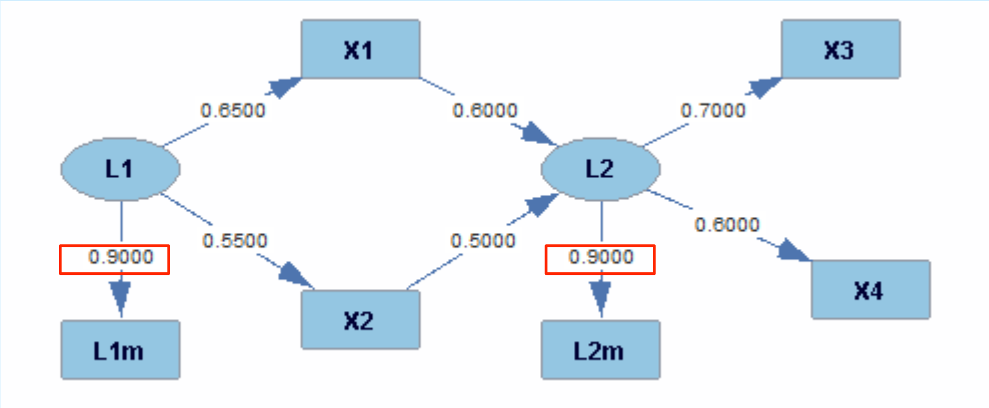
GES Output

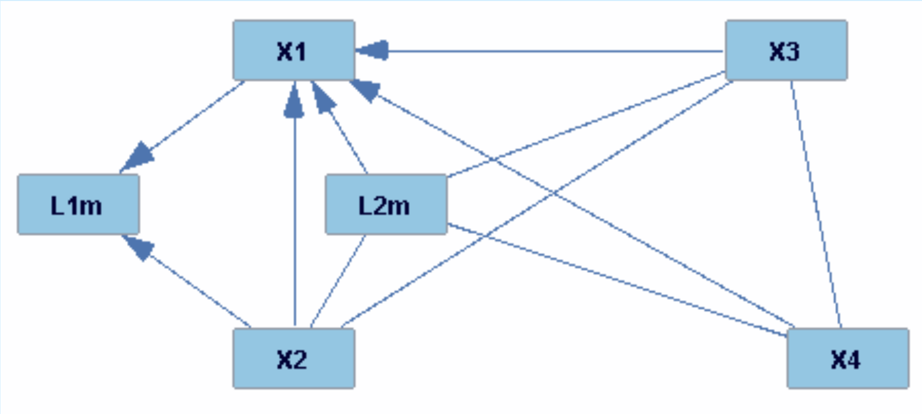


FCI Output

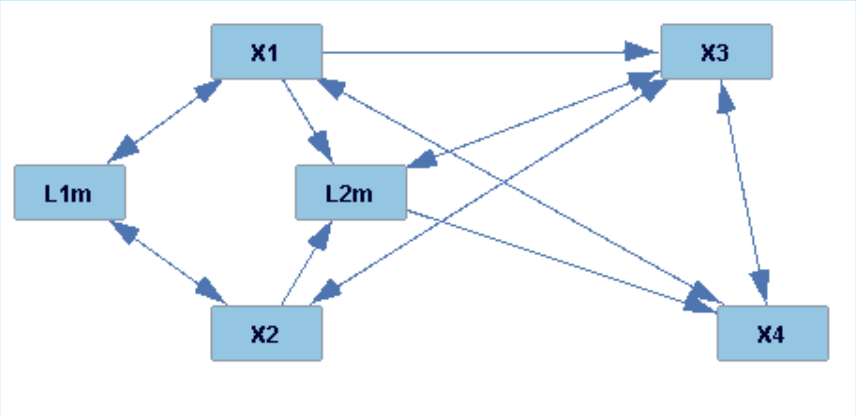
Measurement Error

Parameterization 2:
Small
Measurement Error





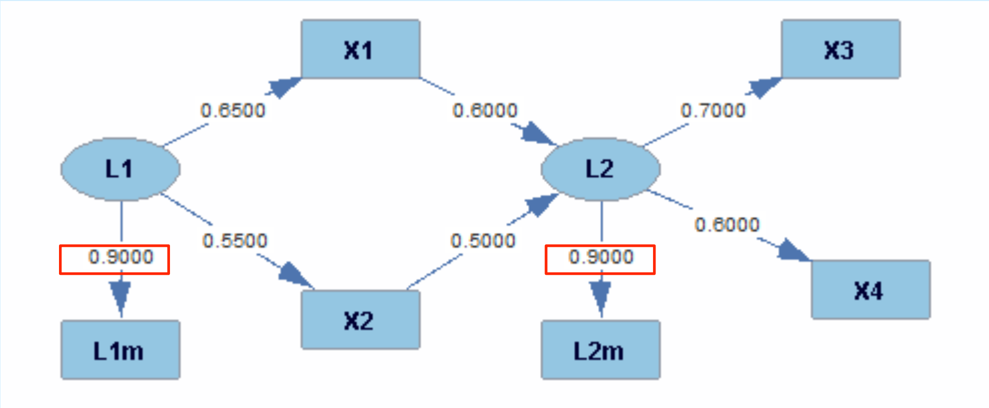
GES Output

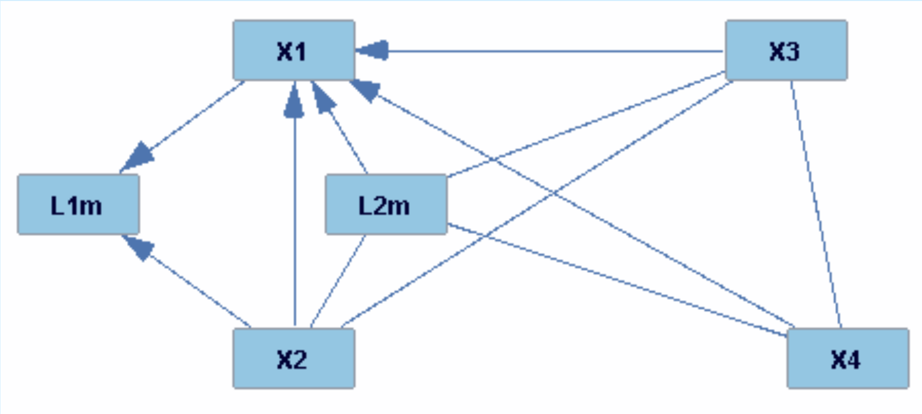


FCI Output
(α = .05)

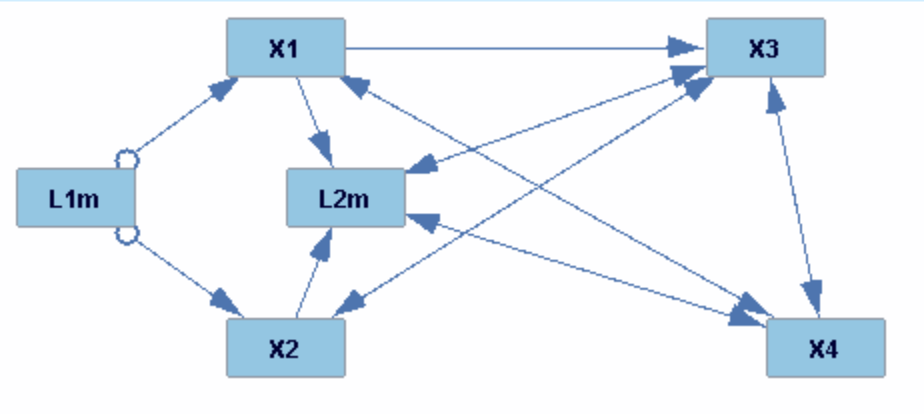
Measurement Error

Parameterization 2:
Small
Measurement Error





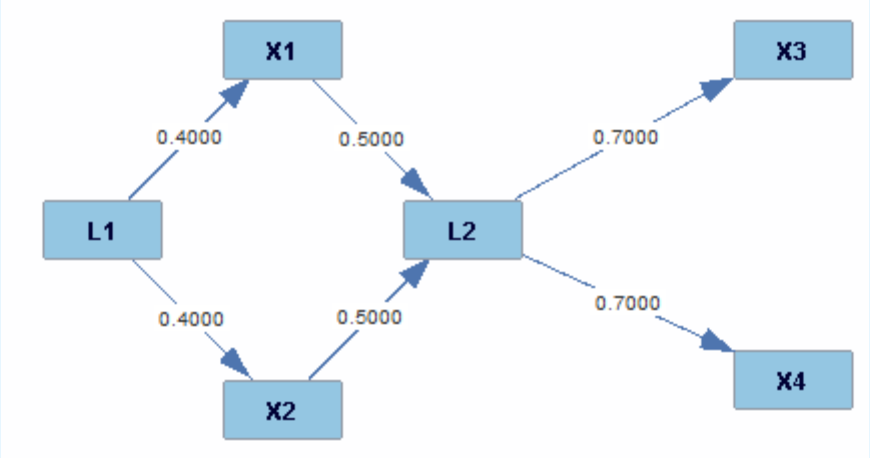
GES Output

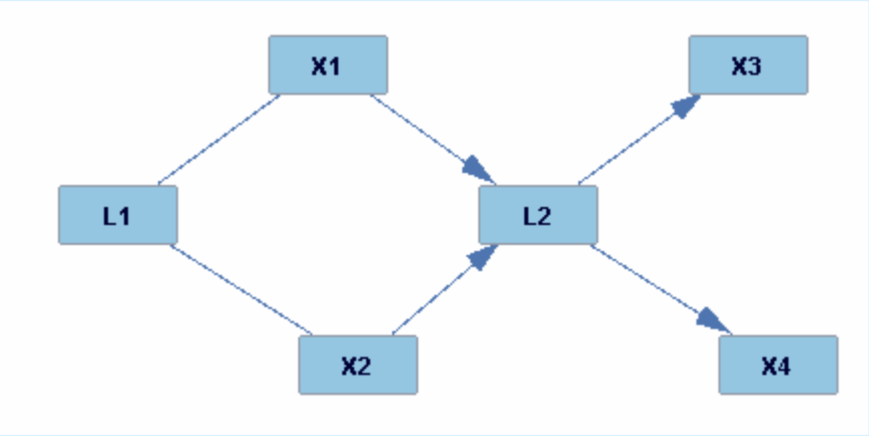


FCI Output
(α = .01)

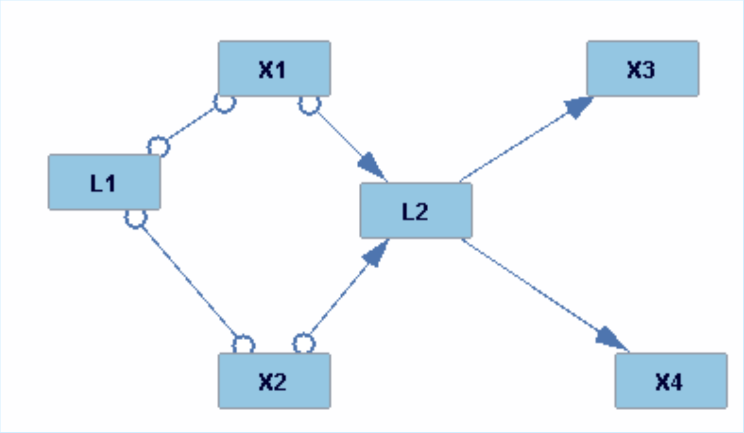
Coarsening

Parameterization 3:



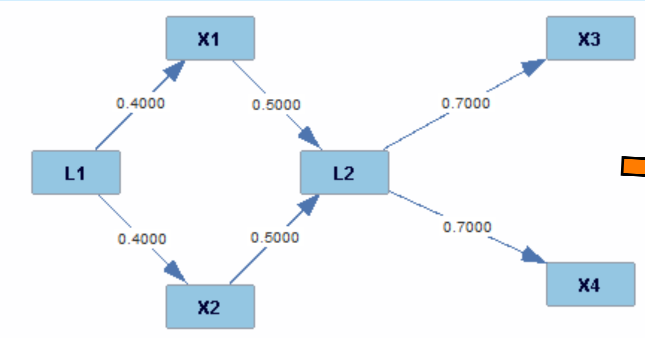


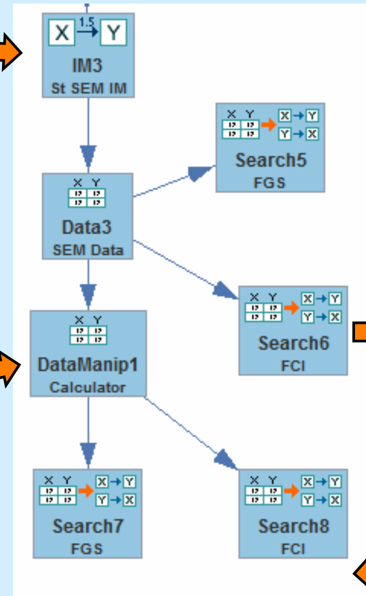
GES Output

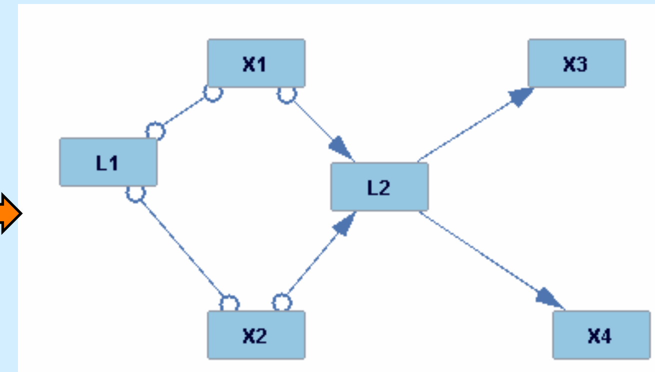


FCI Output
(α = .05)

Coarsening







L1c

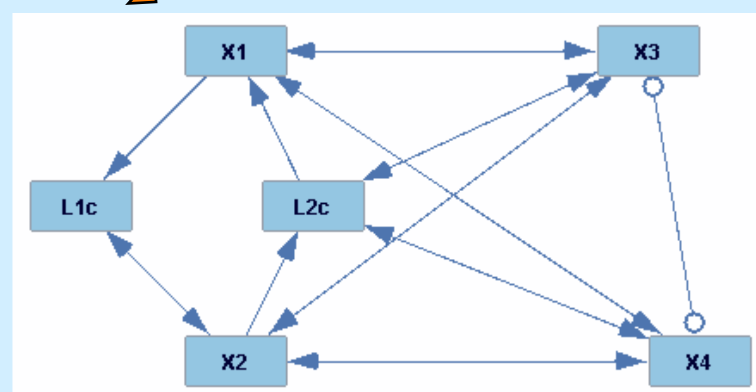
L1 > 0.0 → L1c = 1.0

L1 ≤ 0.0 → L1c = 0.0

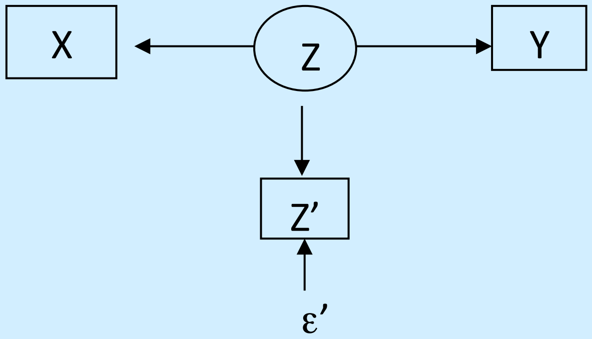
L2c

L1 > 0.0 → L1c = 1.0

L1 ≤ 0.0 → L1c = 0.0



Strategies



1. Guess (the amount of measurement error):

- Sensitivity Analysis
- Bayesian Analysis
- Bounds

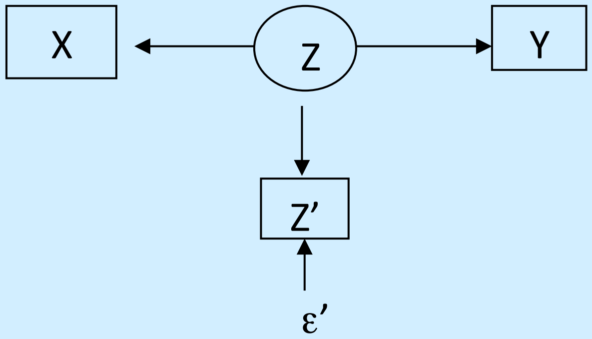
$$\textit{Measurement Error} = 1 - \rho(Z, Z')^2$$

Prior over

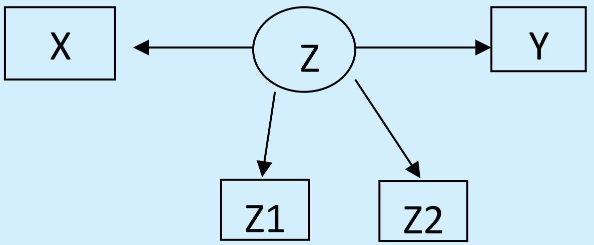
Strategies

1. Guess the measurement error:

- Sensitivity Analysis
- Bayesian Analysis
- Bounds

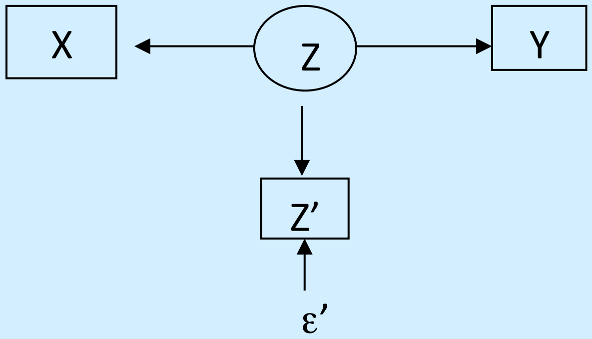


2. Multiple Indicators:

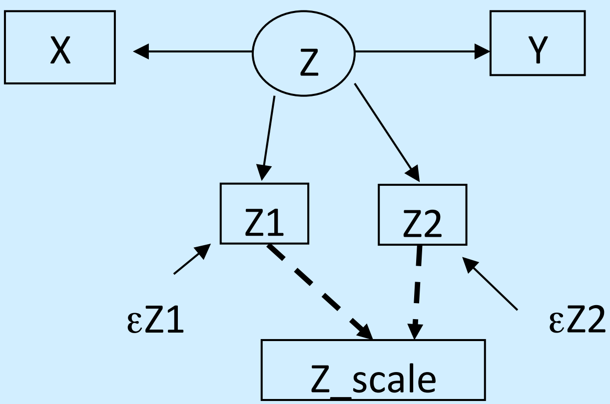


Strategies

1. Parameterize measurement error:
- Sensitivity Analysis
- Bayesian Analysis
- Bounds

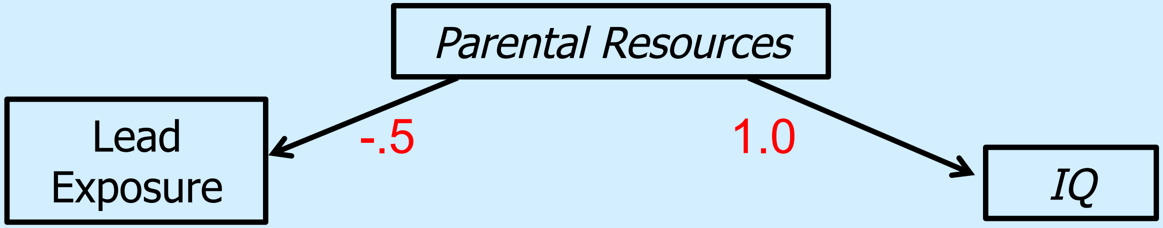


2. Multiple Indicators:
- Scales



X ~~_||_~~ Y | Z_scale

Simulated Example: Lead and IQ



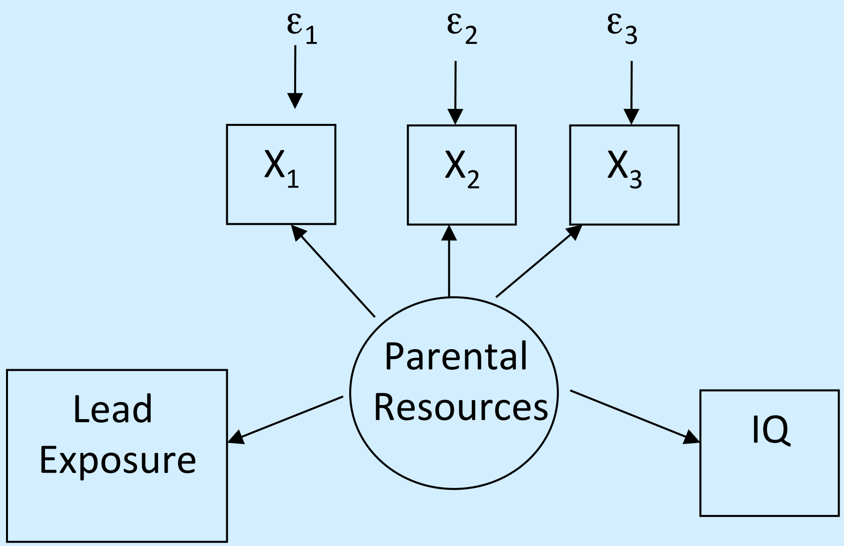
Psuedorandom sample: N = 2,000

Test of Lead _||_ IQ | Parental Resources:

Regression: Dependent Variable: IQ
Independent Variables: Lead, PR

Independent Variable	Coefficient Estimate	p-value
PR	0.98	0.000
Lead	-0.088	0.378

Multiple Measures of the Confounder

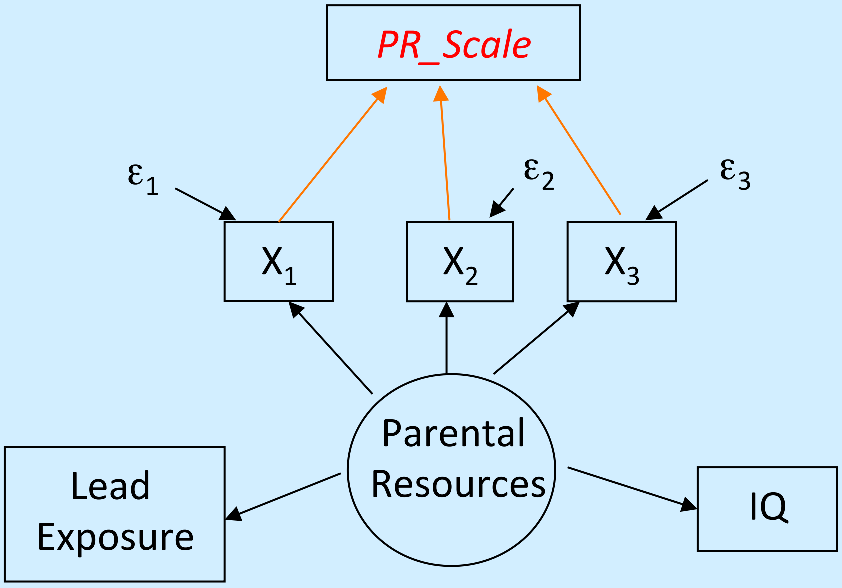


X_1 := Parental Resources + ε_1

X_2 := Parental Resources + ε_2

X_3 := Parental Resources + ε_3

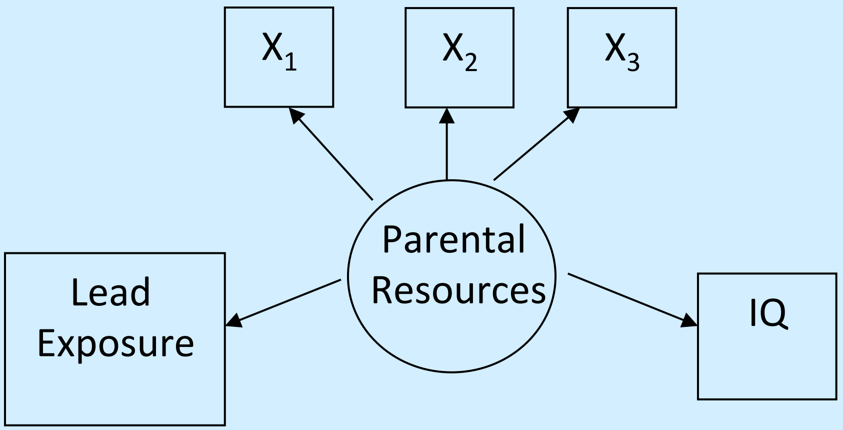
Scales don't preserve conditional independence



$$PR\_Scale = (X_1 + X_2 + X_3) / 3$$

Independent Variable	Coefficient Estimate	p-value
PR_scale	0.290	0.000
Lead	-0.423	0.000

Indicators Don't Preserve Conditional Independence

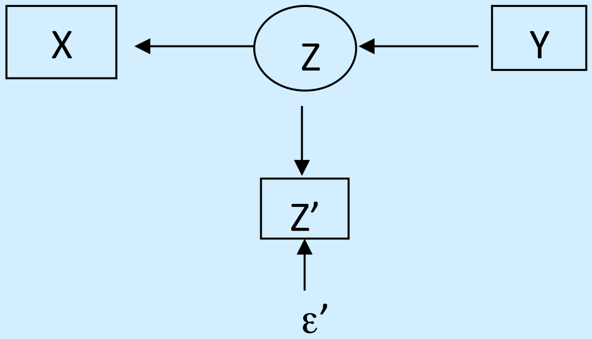


Regress IQ on: Lead, X_1, X_2, X_3

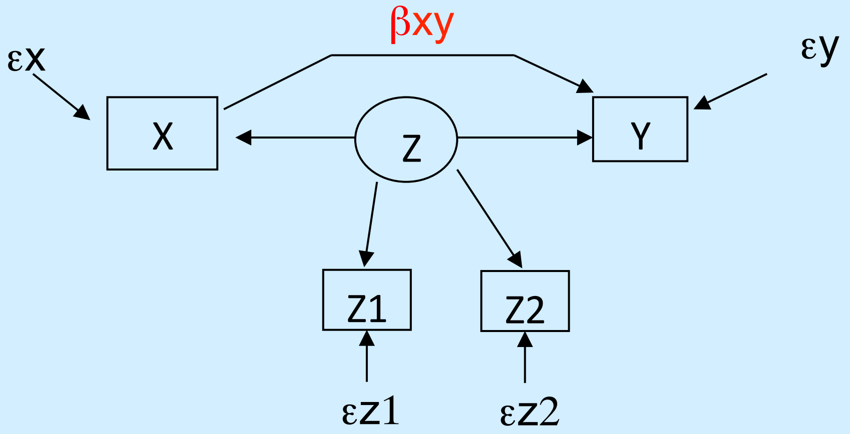
Independent Variable	Coefficient Estimate	p-value
X_1	0.22	0.002
X_2	0.45	0.000
X_3	0.18	0.013
Lead	-0.414	0.000

Strategies

1. Parameterize measurement error:
- Sensitivity Analysis
- Bayesian Analysis
- Bounds

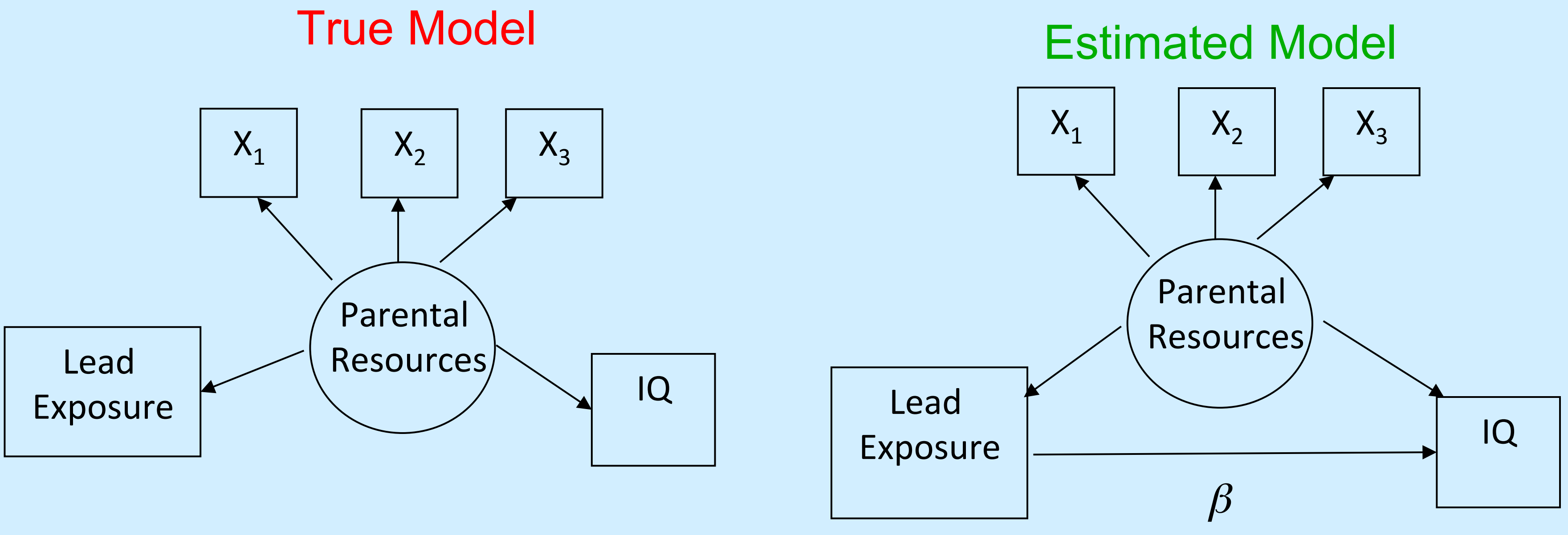


2. Multiple Indicators:
- Scales
- SEM



$$E(\hat{\beta}_{yx}) = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

Structural Equation Models Work!

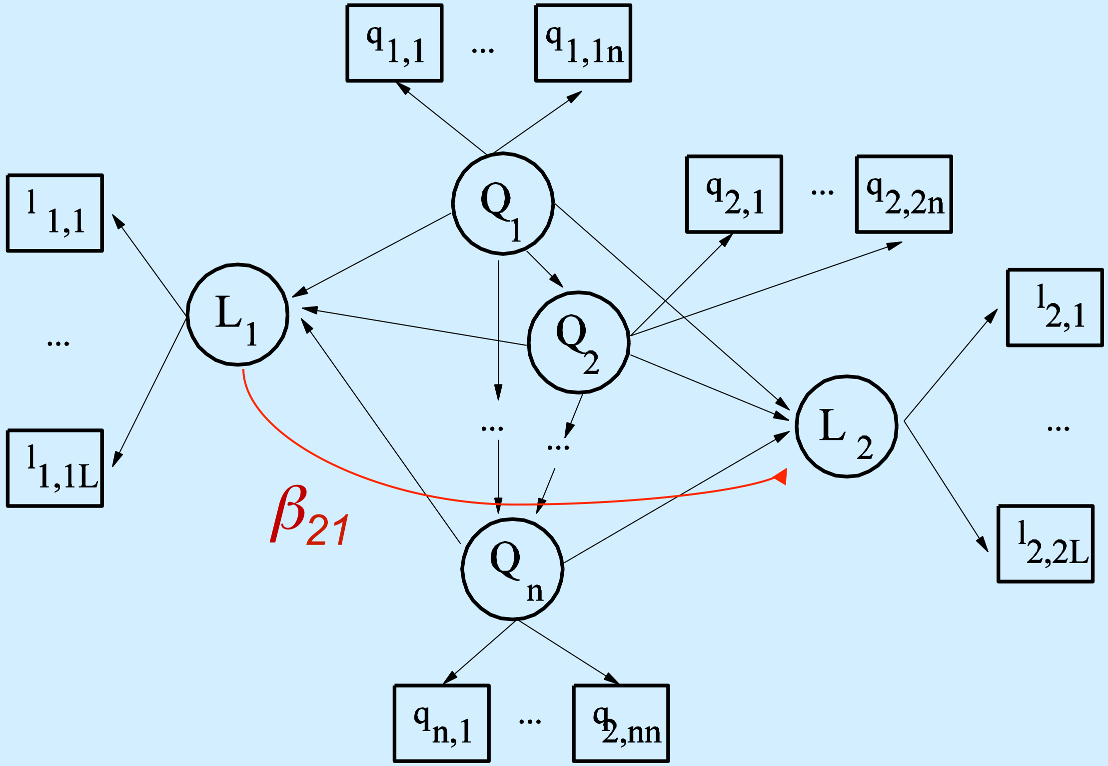


In the Estimated Model

- $E(\hat{\beta}) = 0$
- $\hat{\beta} = .07$ (p-value = .499)
- Lead and IQ detectibly "screened off" by PR

Test conditional independence relations among latents *generally* in a SEM

Question: $L_1 \perp\!\!\!\perp L_2 \mid \{Q_1, Q_2, \ldots, Q_n\}$



$$E(\hat{\beta}_{21}) = 0 \iff L_1 \perp\!\!\!\perp L_2 \mid \{Q_1, Q_2, \ldots, Q_n\}$$

Pure Measurement Models: Causal Discovery among latents *feasible*

Independence Questions: e.g.,

$L_1 \perp\!\!\!\perp L_2 \mid \{L_3, L_4, \ldots, L_n\}$

↓

Constraint Based Search over L

MIMBuild (PC)

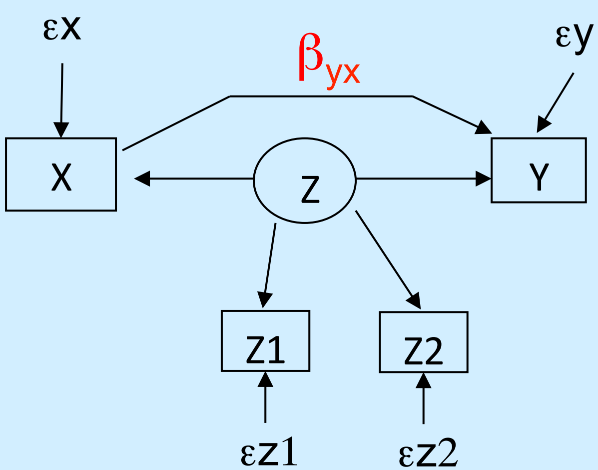
↓

*Pattern, PAG over **L***

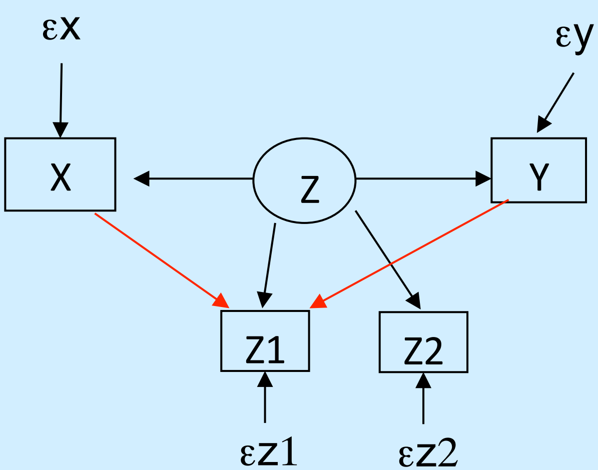
The Problem of Impurities

(“Unmodeled Complexity”)

Specified Model



True Model

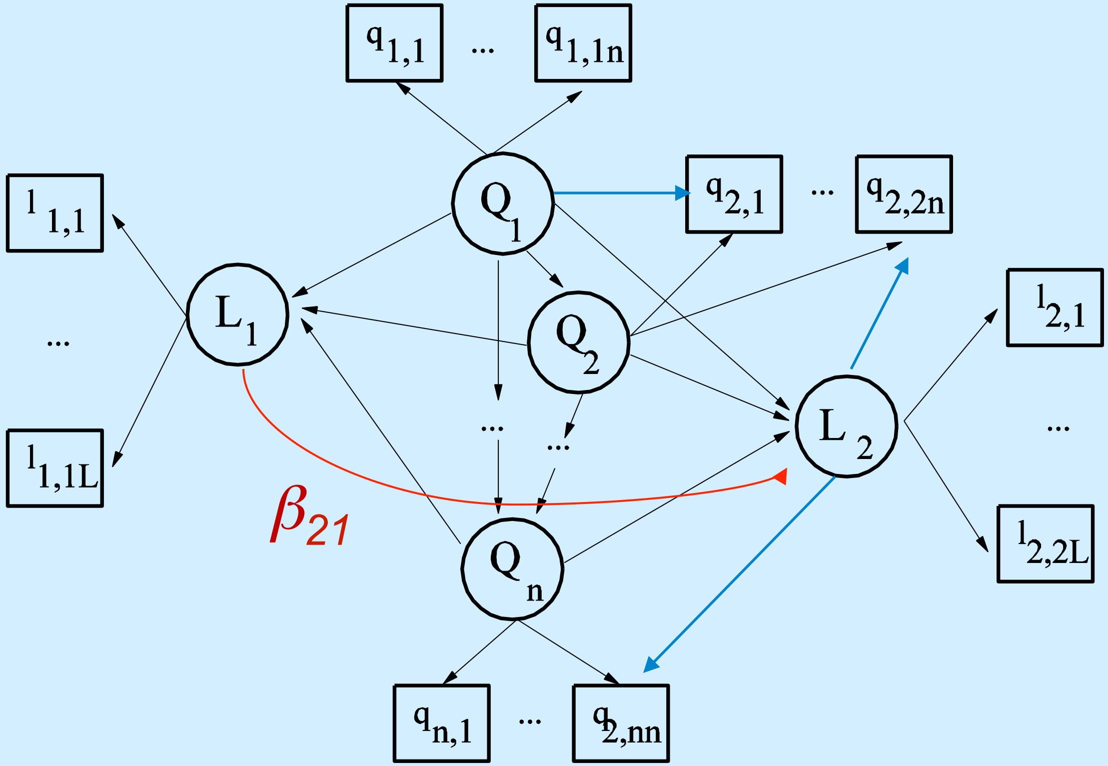


$E(\hat{\beta}_{yx}) = 0$ ~~⇔~~ X _||_ Y | Z

$E(\hat{\beta}_{yx}) \neq \beta_{xy}$

Test conditional independence relations among latents generally

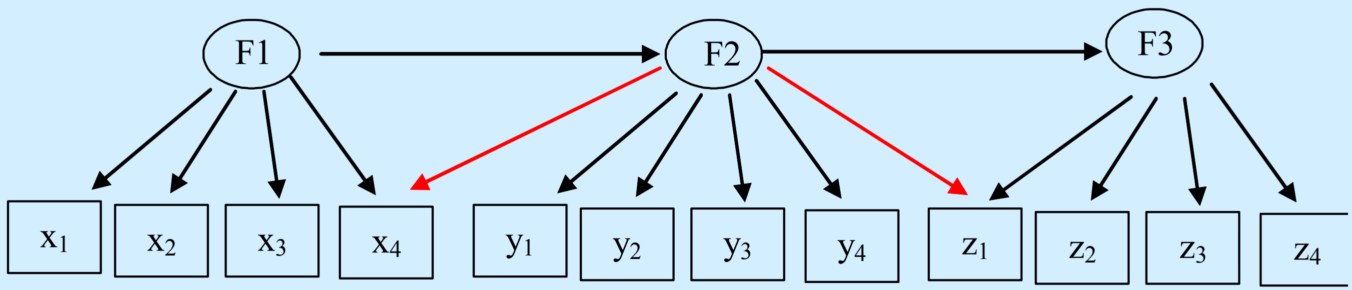
Question: $L_1 \_||\_ L_2 \mid \{Q_1, Q_2, \ldots, Q_n\}$



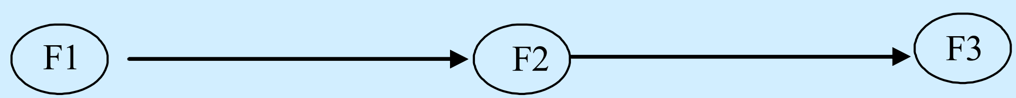
$$E(\hat{\beta}_{21}) = 0 \not\Leftrightarrow L_1 \_||\_ L_2 \mid \{Q_1, Q_2, \ldots, Q_n\}$$

Strategy: Purify the Measurement Model

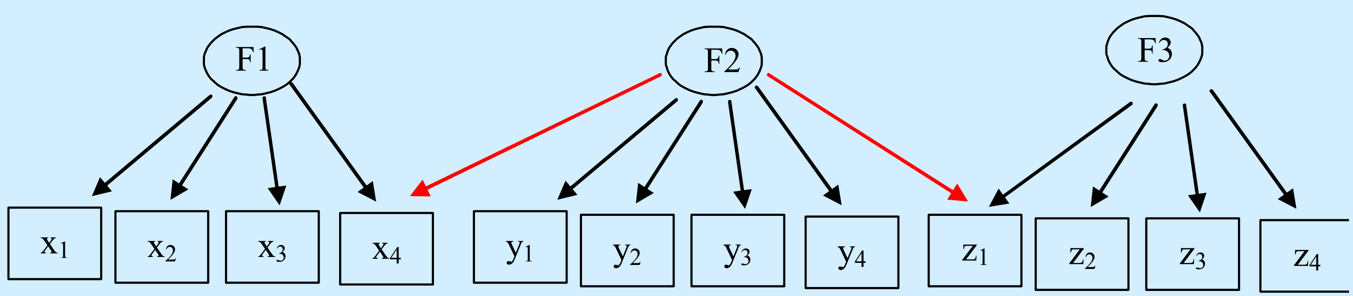
Full Model



Structural Model

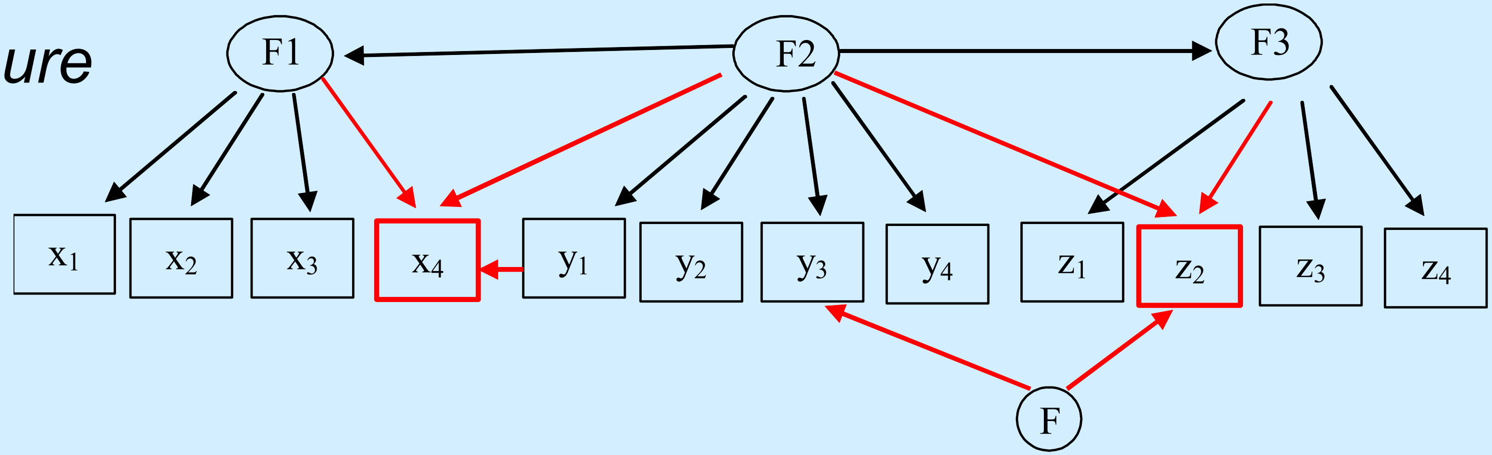


Measurement Model

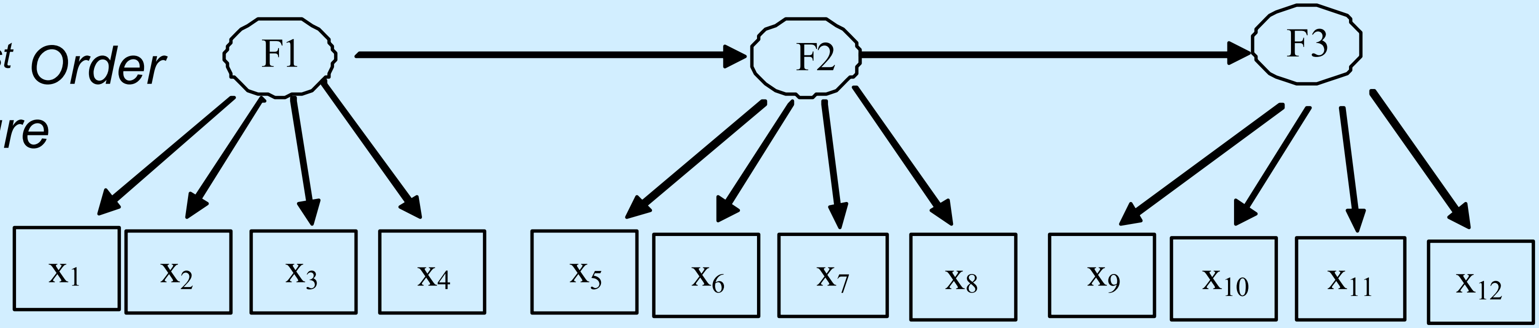


Local Independence / Pure Measurement Models

Impure

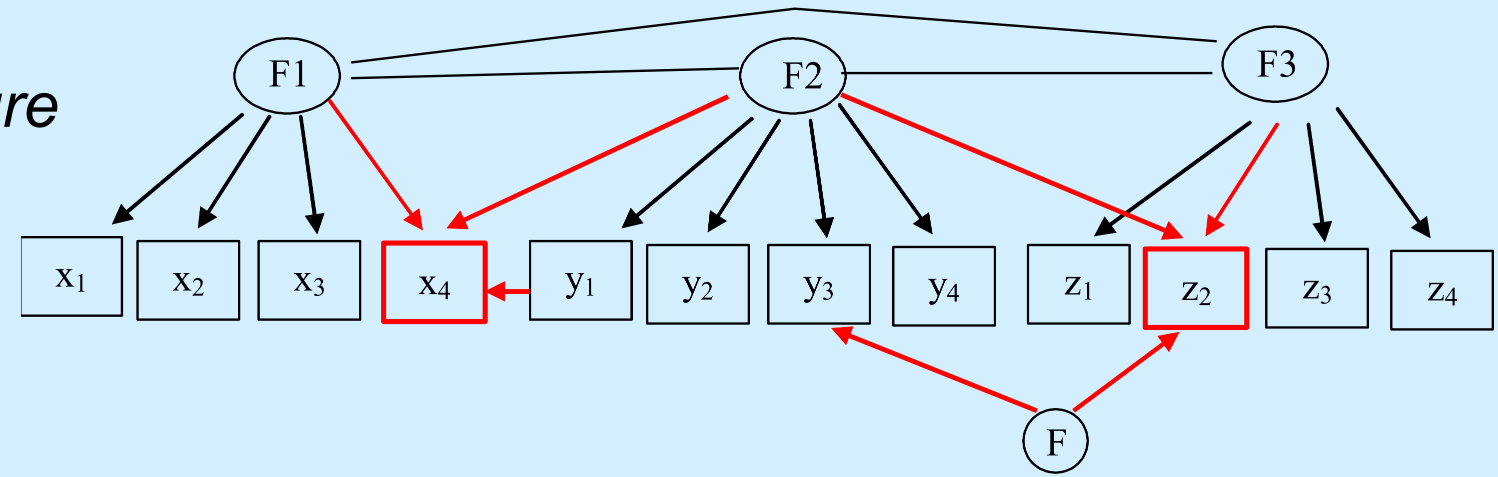


1st Order Pure

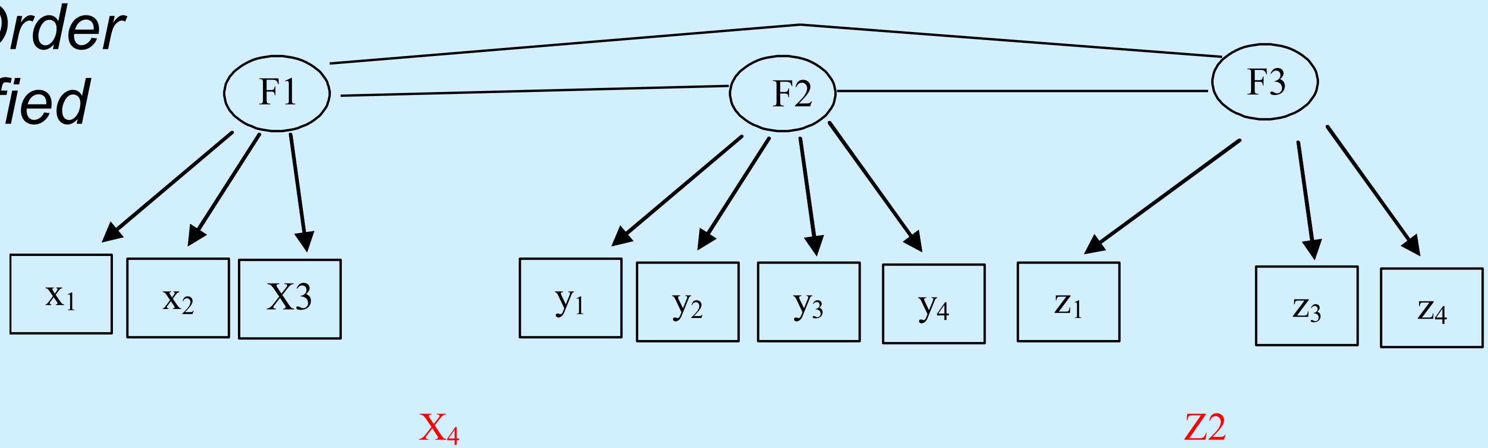


Purify

Impure



1st Order Purified



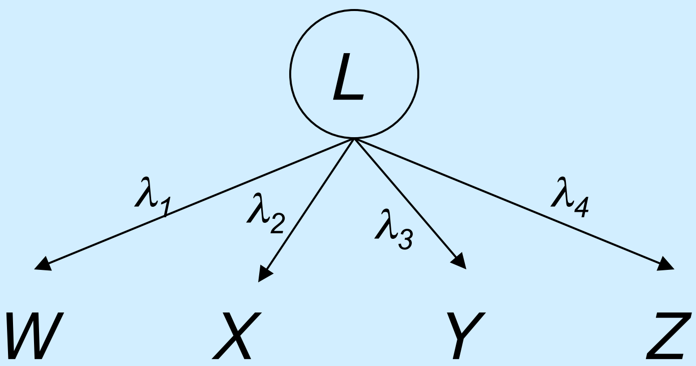
Testable/Observable Contraints

- Latent variable models that are linear below the latents entail testable rank constraints on the measured covariance matrix *regardless of the structural model*.

- *Impurities* selectively defeat such implications, and are thus in many circumstances *detectable and localizable*.

Rank 1 Constraints: Tetrad Equations

- Fact: given



$$W = \lambda_1 L + \varepsilon_1$$
$$X = \lambda_2 L + \varepsilon_2$$
$$Y = \lambda_3 L + \varepsilon_3$$
$$Z = \lambda_4 L + \varepsilon_4$$

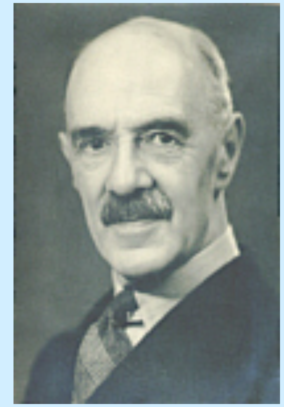
- it follows that

$$Cov_{WX}Cov_{YZ} = (\lambda_1\lambda_2\sigma^2_L)\,(\lambda_3\lambda_4\sigma^2_L) =$$
$$= (\lambda_1\lambda_3\sigma^2_L)\,(\lambda_2\lambda_4\sigma^2_L) = Cov_{WY}Cov_{XZ}$$

tetrad constraints

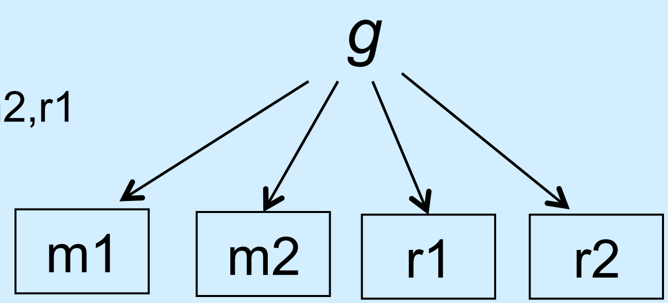
$$\sigma_{WX}\sigma_{YZ} = \sigma_{WY}\sigma_{XZ} = \sigma_{WZ}\sigma_{XY}$$

Charles Spearman (1904)



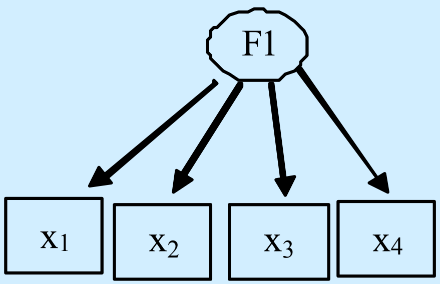
Statistical Constraints → Measurement Model Structure

$$\rho_{m1,m2} * \rho_{r1,r2} = \rho_{m1,r1} * \rho_{m2,r2} = \rho_{m1,r2} * \rho_{m2,r1}$$

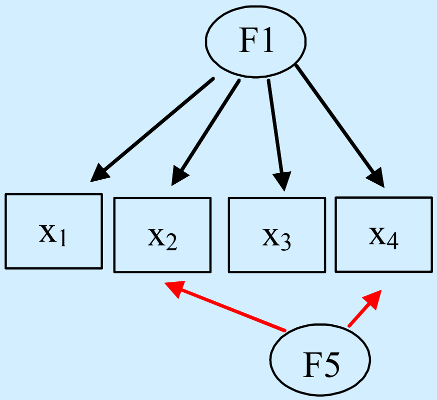


Impurities defeat rank (e.g., tetrad) constraints selectively

Truth



Truth



$$\rho_{x1,x2} * \rho_{x3,x4} = \rho_{x1,x3} * \rho_{x2,x4}$$

$$\rho_{x1,x2} * \rho_{x3,x4} = \rho_{x1,x4} * \rho_{x2,x3}$$

$$\rho_{x1,x3} * \rho_{x2,x4} = \rho_{x1,x4} * \rho_{x2,x3}$$

$$\rho_{x1,x2} * \rho_{x3,x4} \neq \rho_{x1,x3} * \boxed{\rho_{x2,x4}}$$

$$\rho_{x1,x2} * \rho_{x3,x4} = \rho_{x1,x4} * \rho_{x2,x3}$$

$$\rho_{x1,x3} * \boxed{\rho_{x2,x4}} \neq \rho_{x1,x4} * \rho_{x2,x3}$$

2006: Build (1st-Order) Pure Clusters (BPC)

Input:

- Covariance matrix over set of original items

↓

BPC (FOFC)

1) Cluster (complicated boolean combinations of tetrads)

2) Purify

↓

Output: Equivalence class of 1st order pure clusters

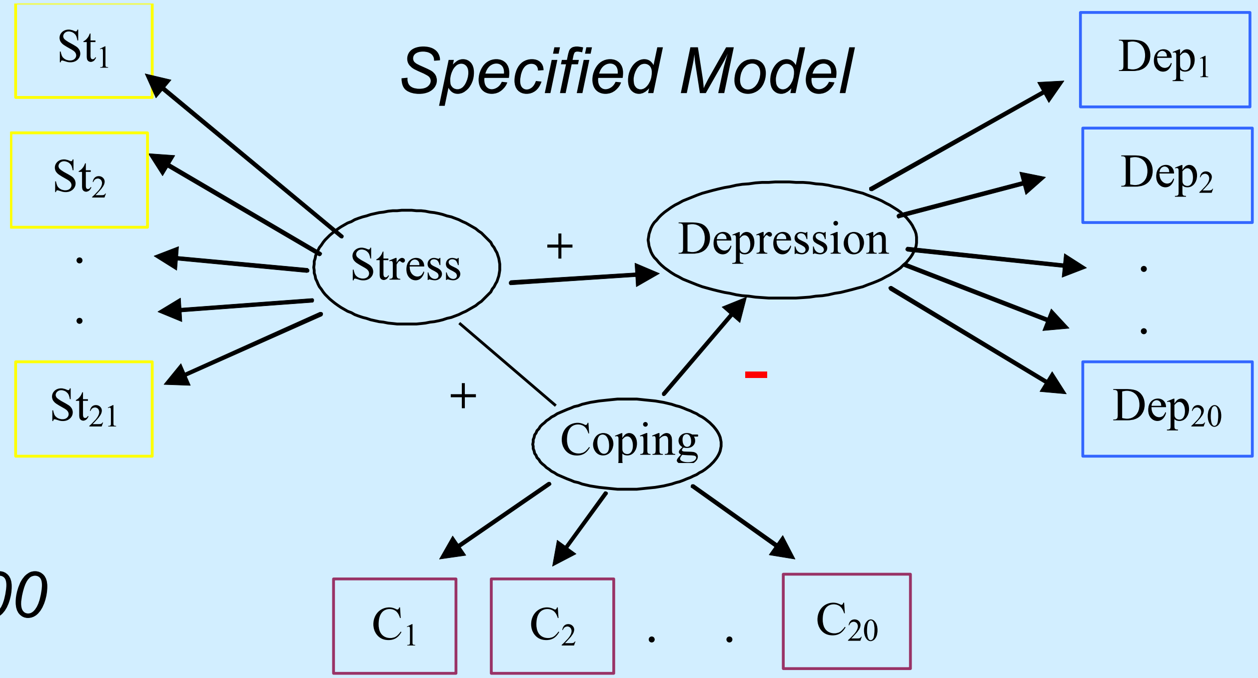
BPC: Pointwise consistent

Silva, R., Glymour, C., Scheines, R. and Spirtes, P. (2006) “Learning the Structure of Latent Linear Structure Models,” *Journal of Machine Learning Research*, 7, 191-246.

Case Study: Stress, Depression, and Religion

Masters Students (N = 127) 61 - item survey (Likert Scale)

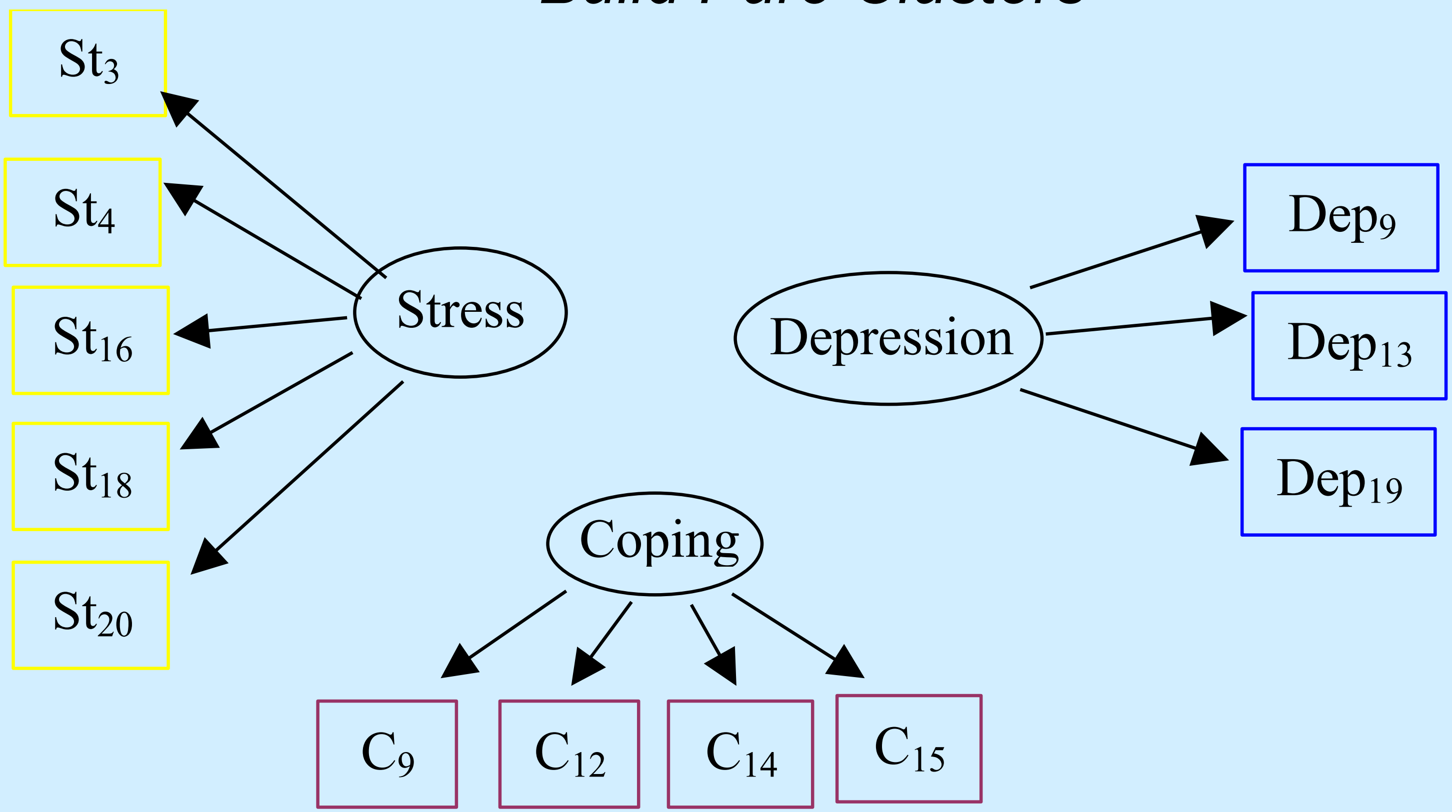
- *Stress: St_1 - St_{21}*
- *Depression: D_1 - D_{20}*
- *Religious Coping: C_1 - C_{20}*



P(χ2) = 0.00

Case Study: Stress, Depression, and Religion

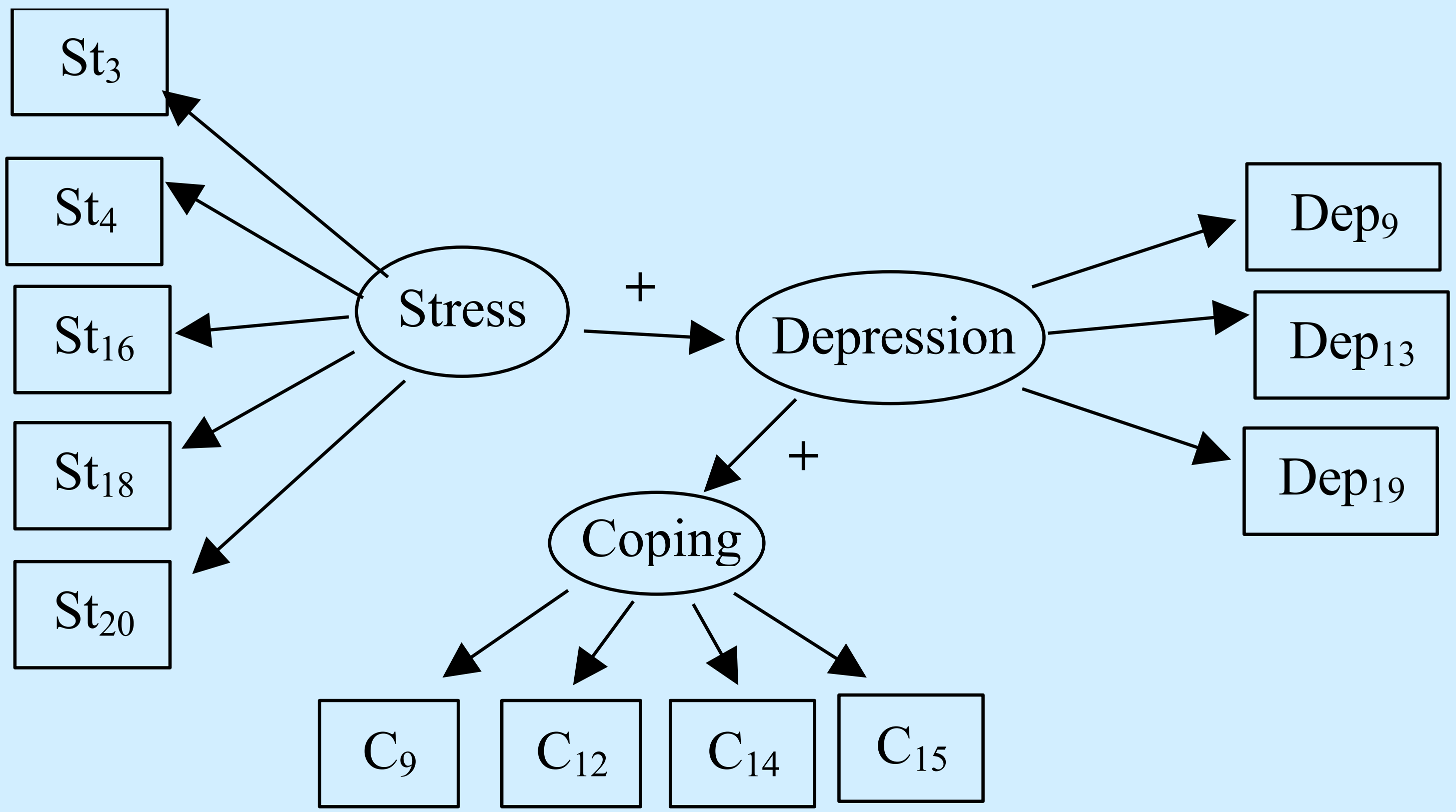
Build Pure Clusters



Case Study: Stress, Depression, and Religion

Assume : Stress causally prior to Depression

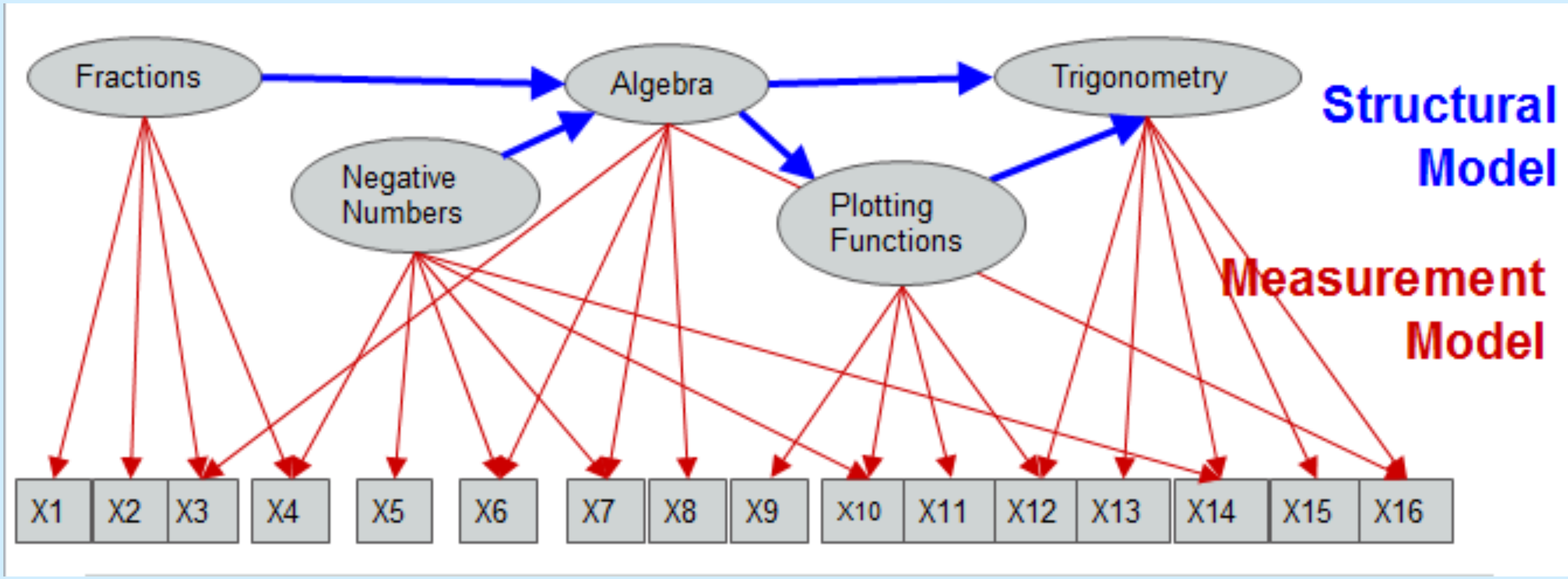
Find : Stress _||_ Coping | Depression



P(χ2) = 0.28

Psychometric Models in Practice:

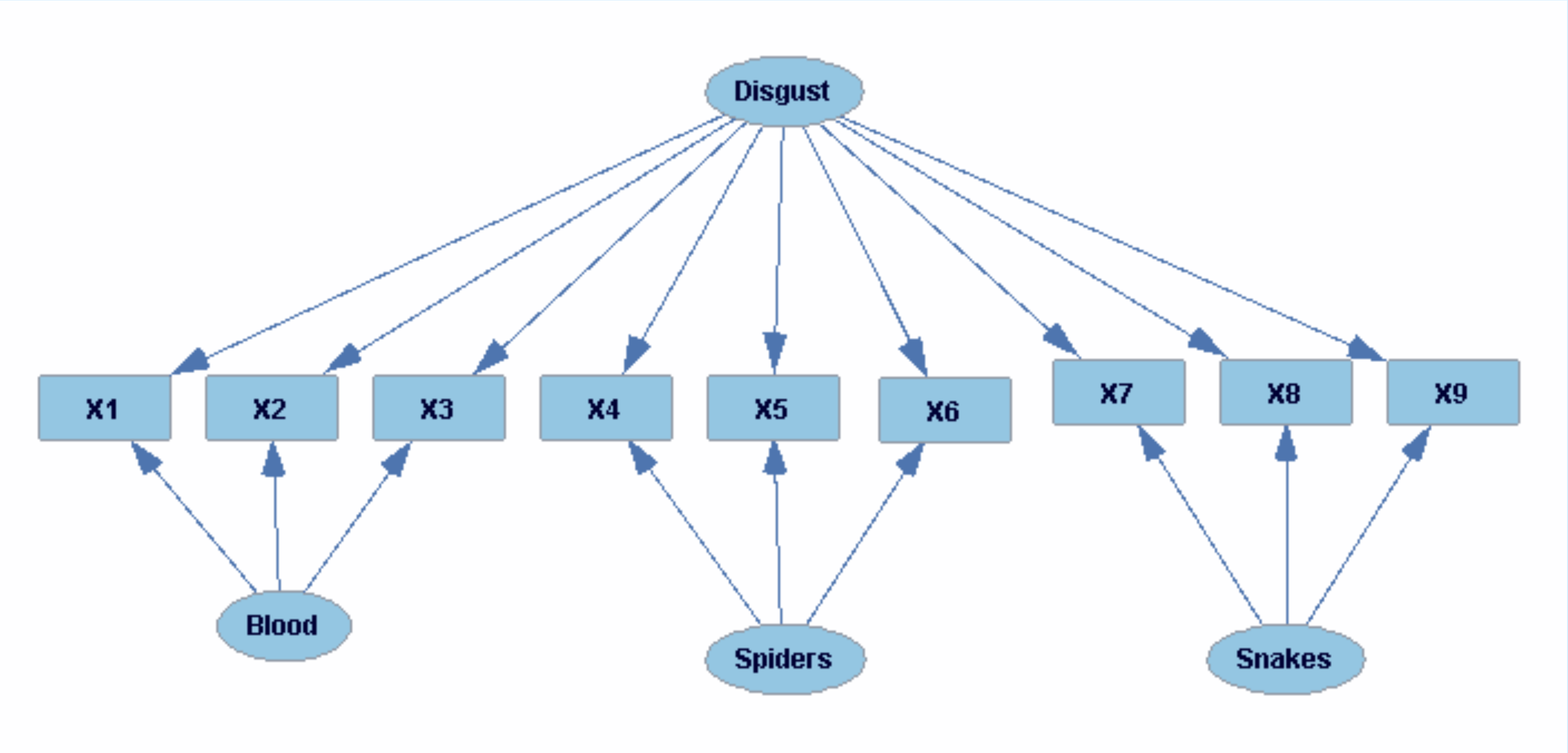
Almost Never 1st Order Pure – or 1st-order Purifiable



Psychometric Models in Practice:

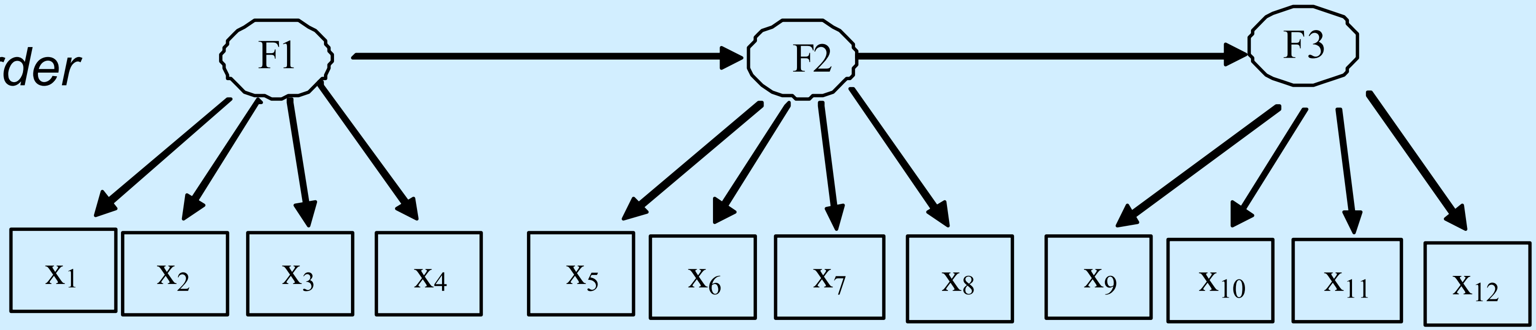
Almost Never 1st Order Pure – or 1st-order Purifiable

"Bi-factor" structure

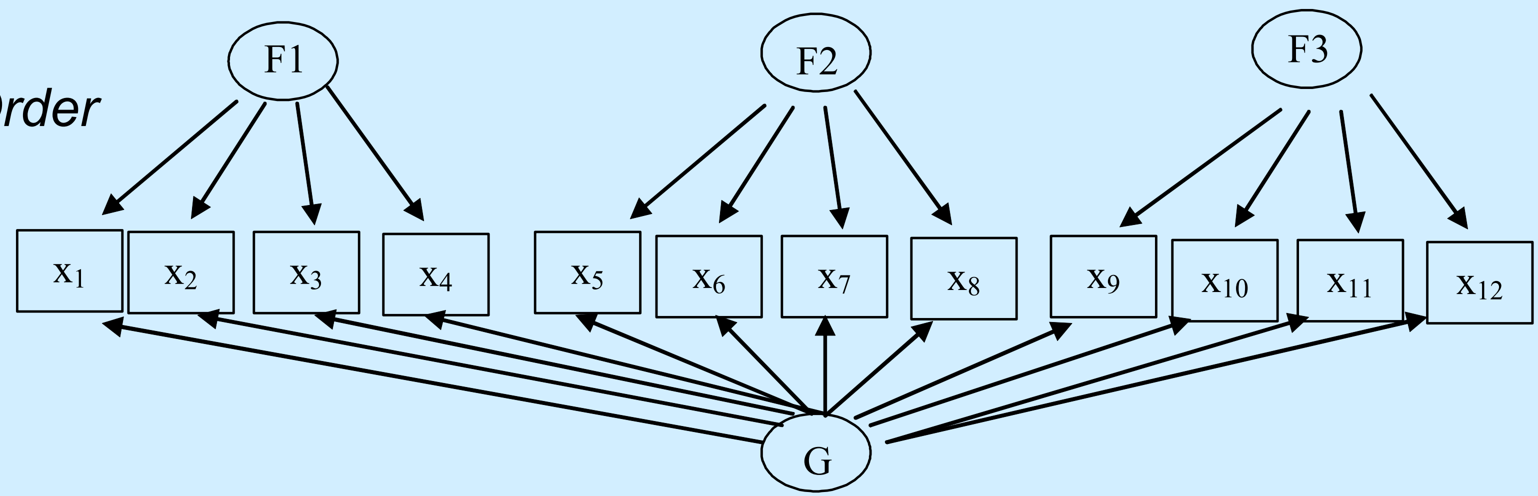


Local Independence / Pure Measurement Models

1st Order Pure



2nd Order Pure



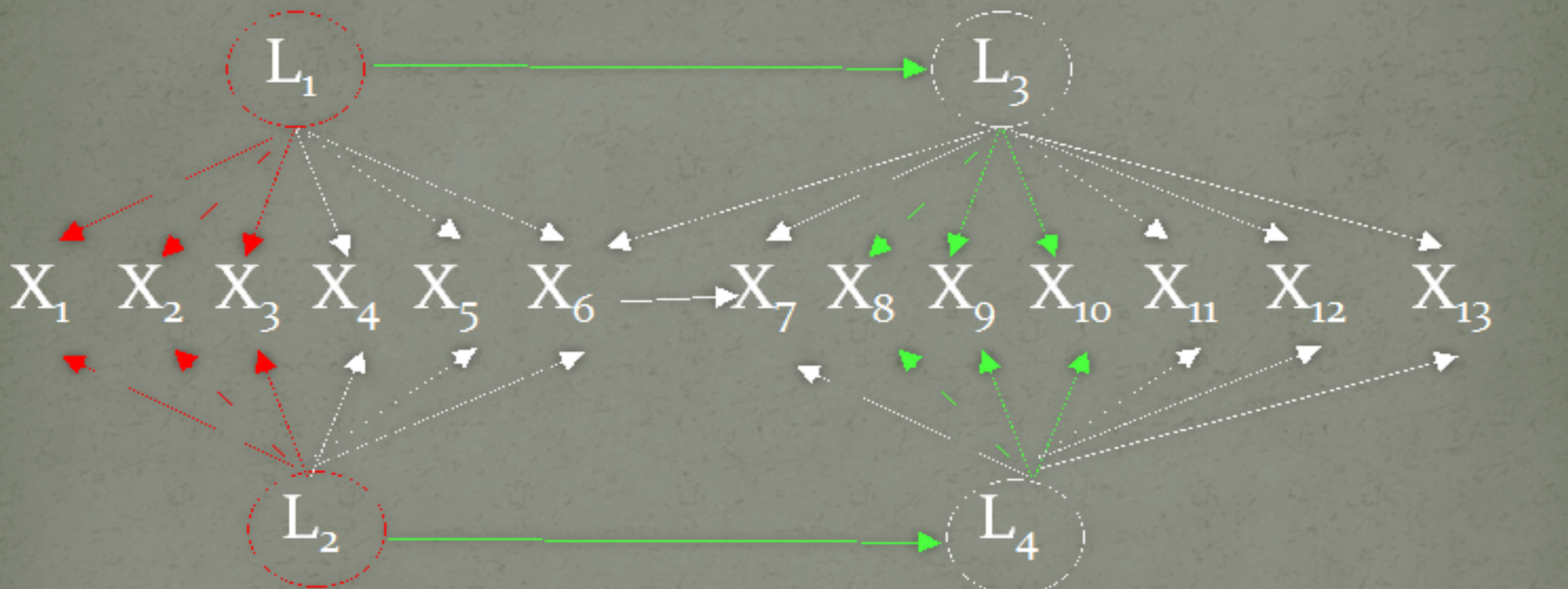
Beyond 1st-Order Purity and Tetrads

- Drton, M., Sturmfels, B., Sullivant, S. (2007) Algebraic factor analysis: tetrads, pentads and beyond, *Probability Theory and Related Fields*, 138, 3-4, 463-493

- Sullivant, S., Talaska, K., & Draisma, J. (2010). Trek Separation for Gaussian Graphical Models. *Annals of Statistics*, 38(3), 1665-1685

Trek-Separation

< $\{L_1, L_2\}, \varnothing$ > Trek-Separate {1,2,3}:{8,9,10}



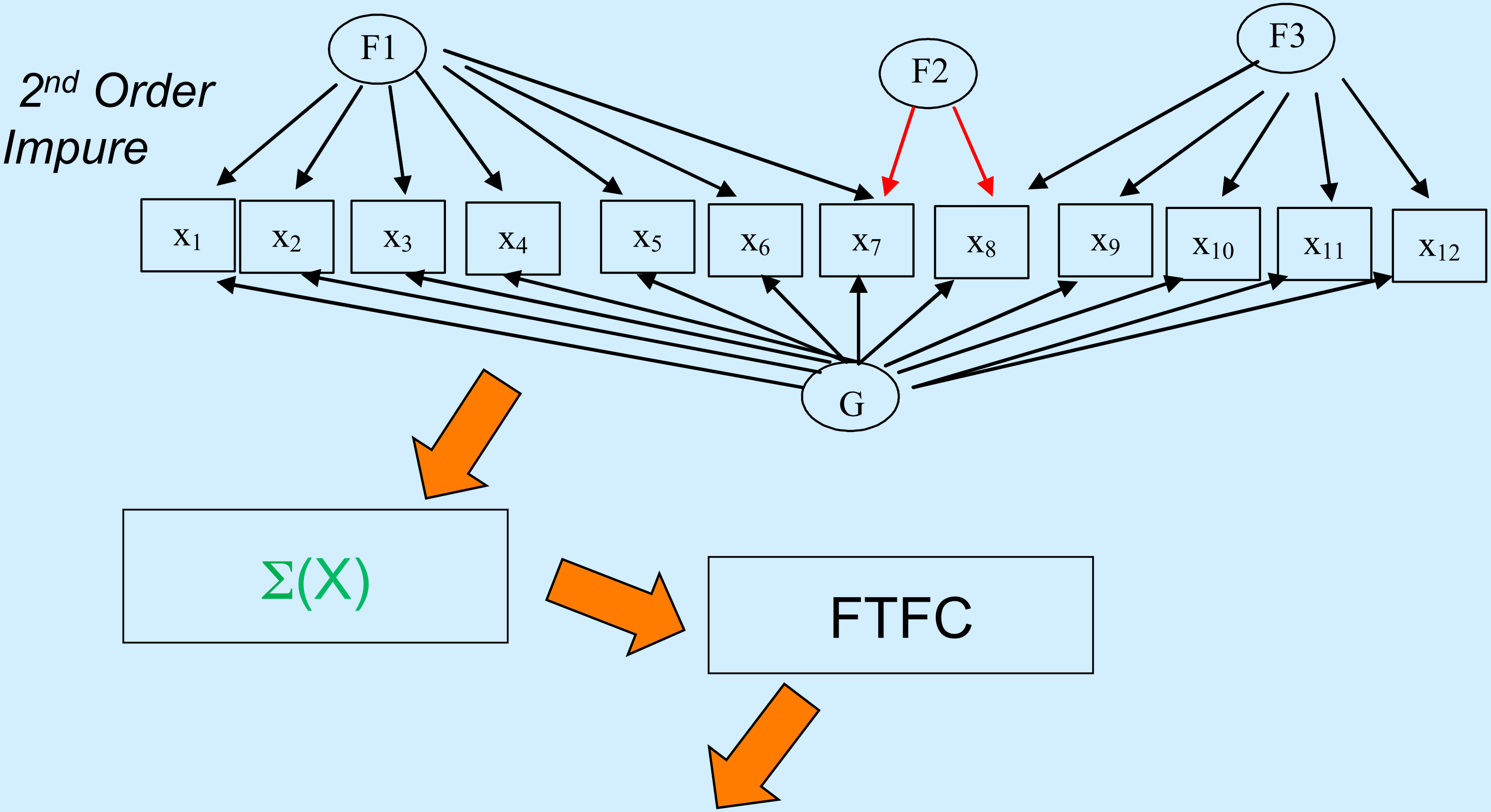
$$rank\begin{pmatrix} \rho(X_1,X_8) & \rho(X_1,X_9) & \rho(X_1,X_{10}) \\ \rho(X_2,X_8) & \rho(X_2,X_9) & \rho(X_2,X_{10}) \\ \rho(X_3,X_8) & \rho(X_3,X_9) & \rho(X_3,X_{10}) \end{pmatrix} \le \#\{L_1,L_2\} + \#\varnothing = 2$$

2013 FTFC
(2^{nd}-order pure measurement clusters)

Input: Covariance Matrix of measured items:

Output: Subset of items and clusters that are 2^{nd} Order Pure

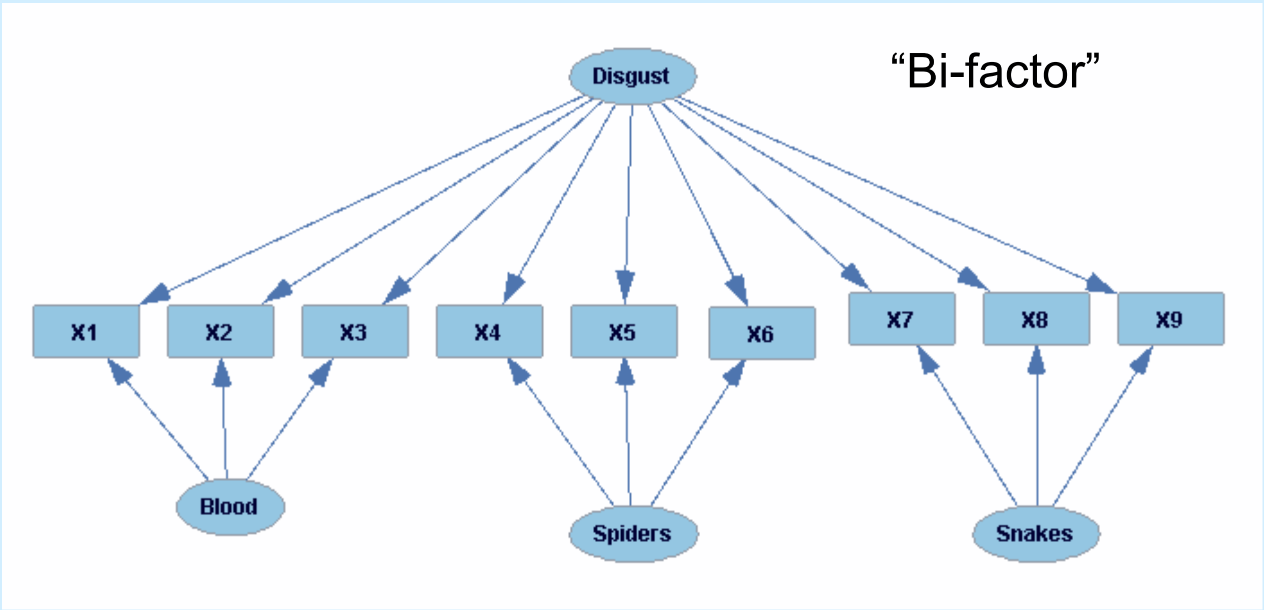
Kummerfeld and Ramsey

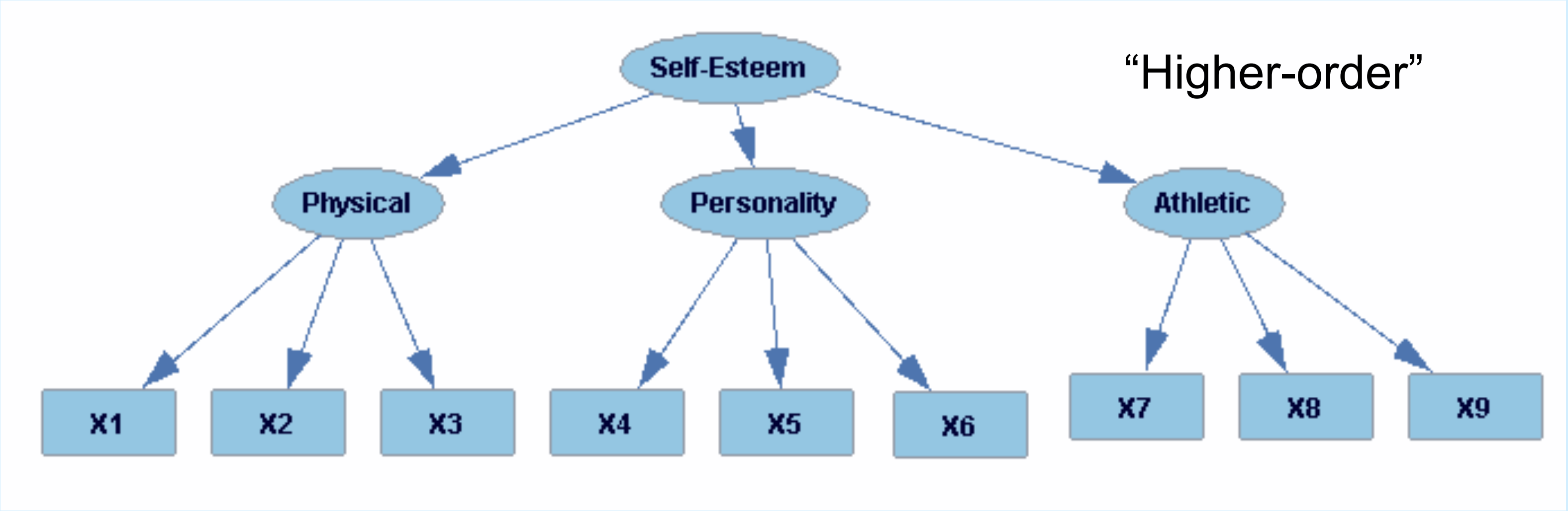


2^{nd}-Order Pure Clusters:

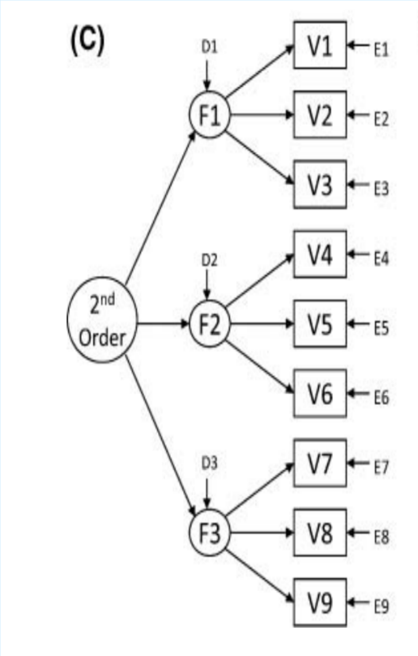
- $\{X_1, X_2, X_3, X_4, X_5, X_6\}$
- $\{X_8, X_9, X_{10}, X_{11}, X_{12}\}$

Psychometric Models in Practice:Bi-factor vs. Higher-Order

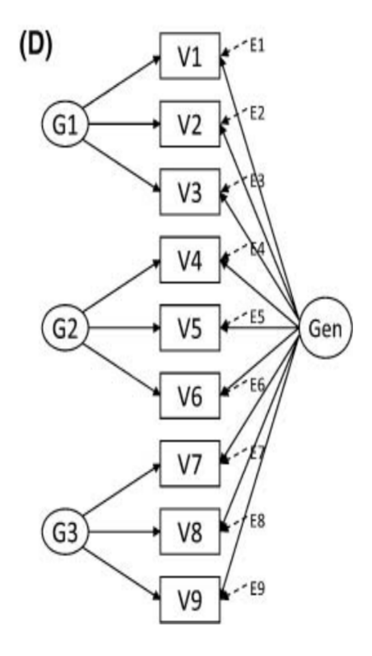




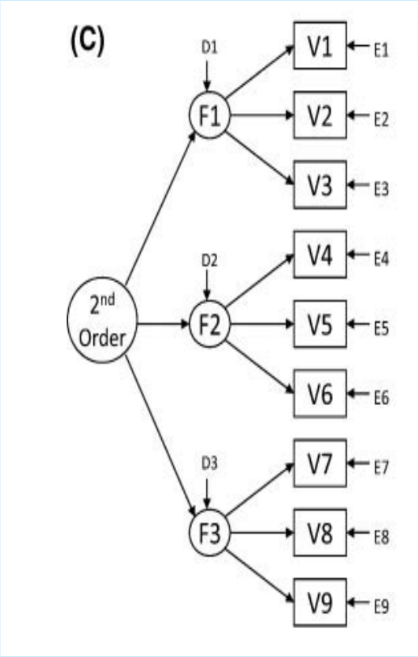
Model Comparison: 2^{nd}-Order vs. Bi-factor



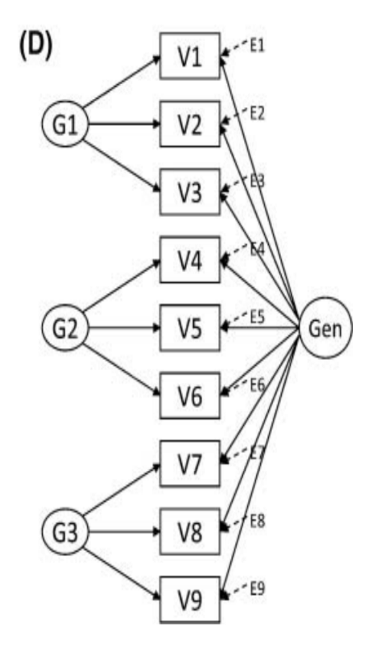
- 2^{nd}-Order entails more constraints (simpler, more degrees of freedom, nesting not simple)
- *Statistical comparison: fit - complexity*
  fit to data (good)
  complexity (bad)
- p-value(χ^2), BIC, AIC, etc.
- Theoretically – BIC desirable.



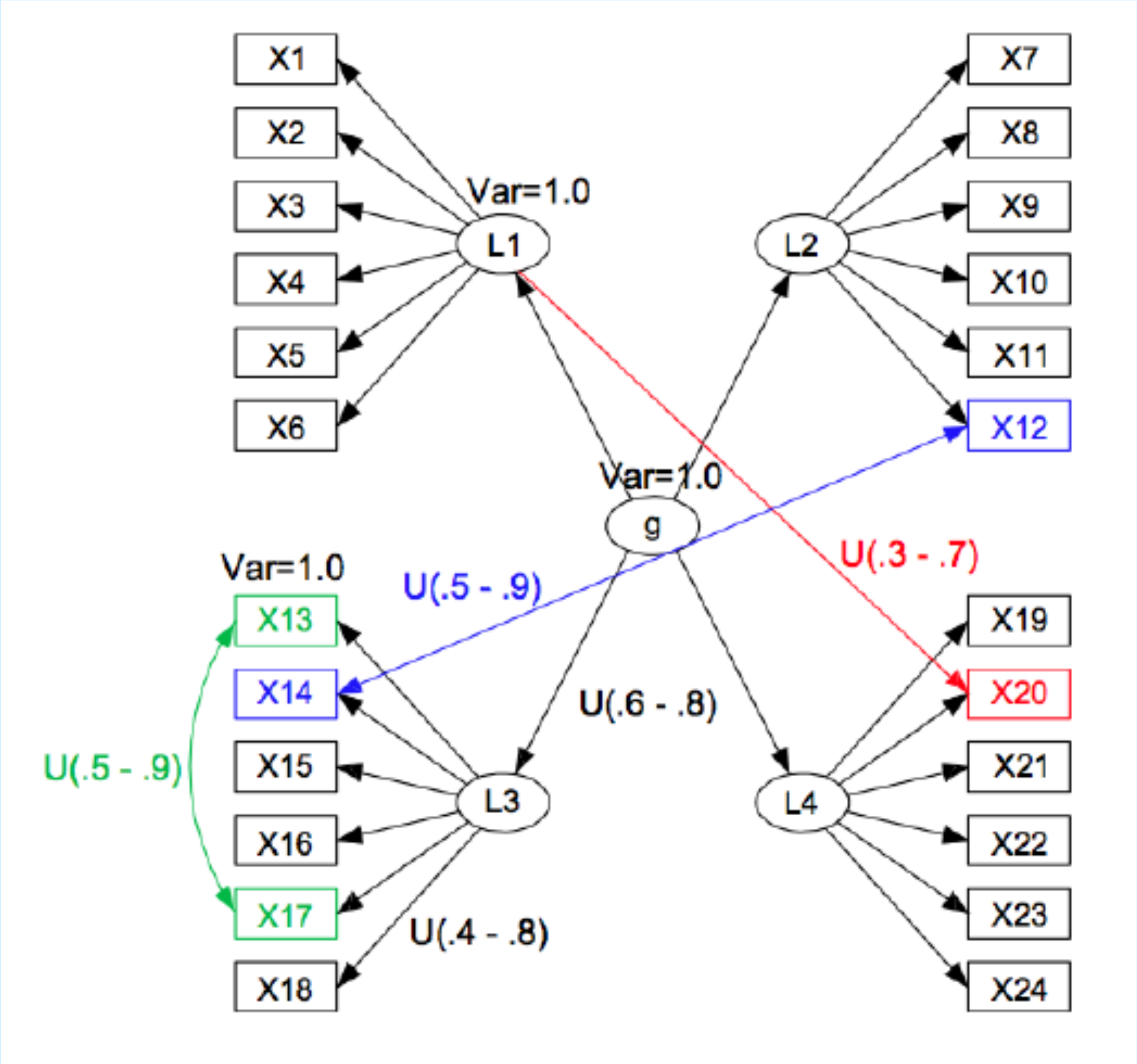
Model Comparison: 2nd-Order vs. Bi-factor



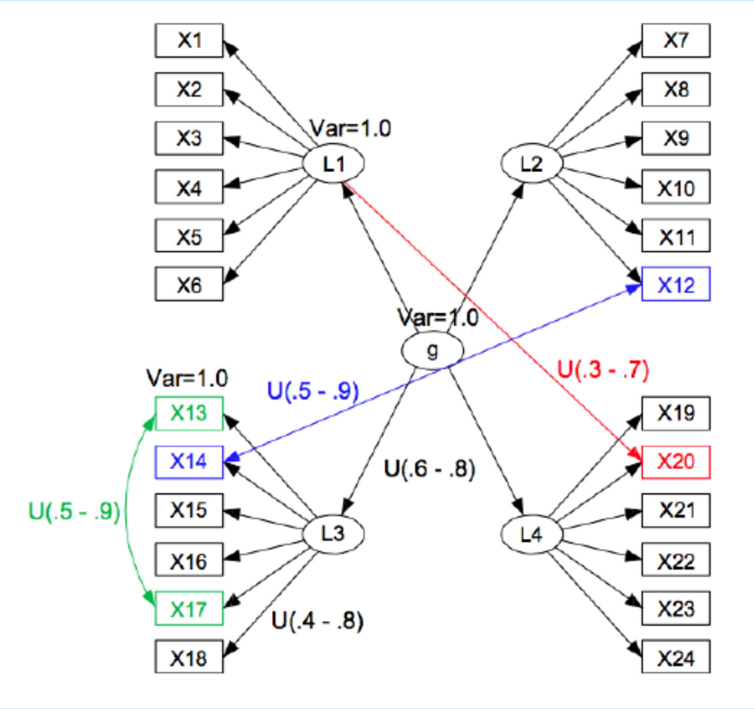
- Empirical finding: Bi-factor usually wins (65 published studies comparing: all chose bi-factors)
- What if data-generating process is neither 2nd-order nor bi-factor?
- What if there is "unmodeled complexity"?



Simulation Study: Truth: 2nd-Order



Simulation Study: 2 Strategies

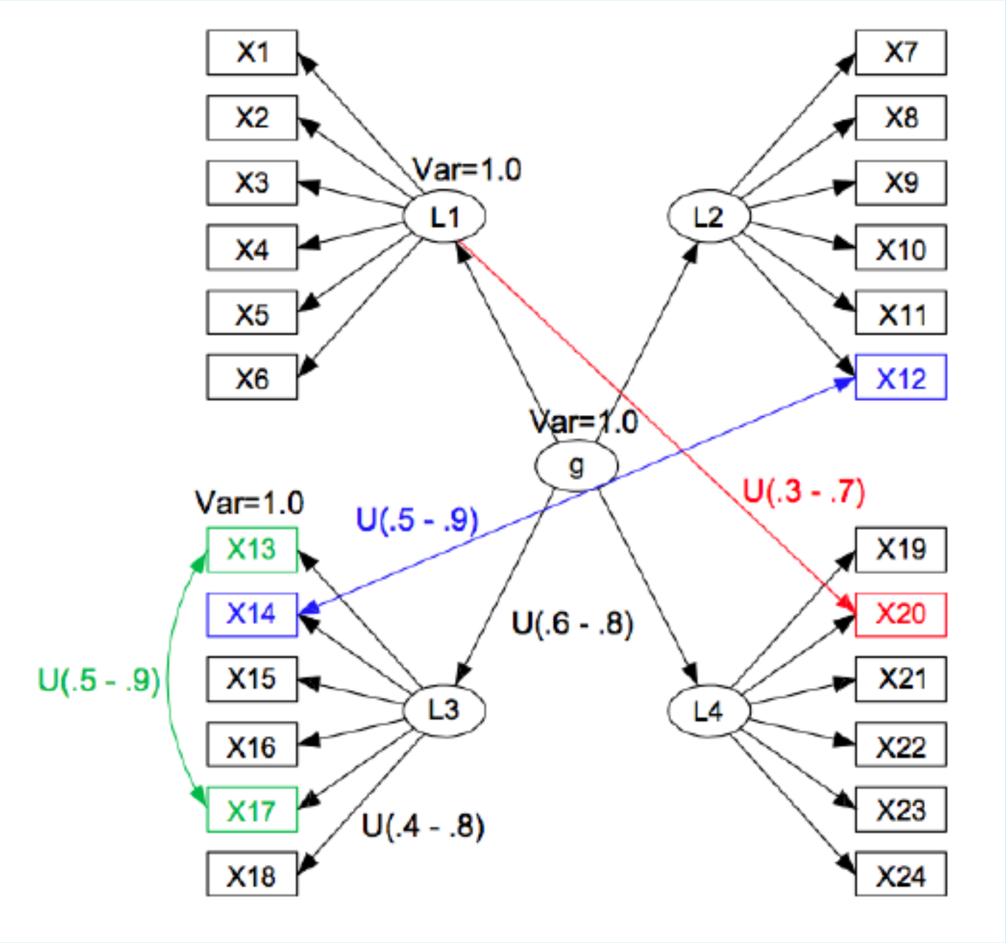


Data from *impure* 2nd-order model

1. **Before FOFC**:
   Fit and compare *pure* versions of
   2nd-order, bi-factor models
   with *correct* clustering – on *all* items

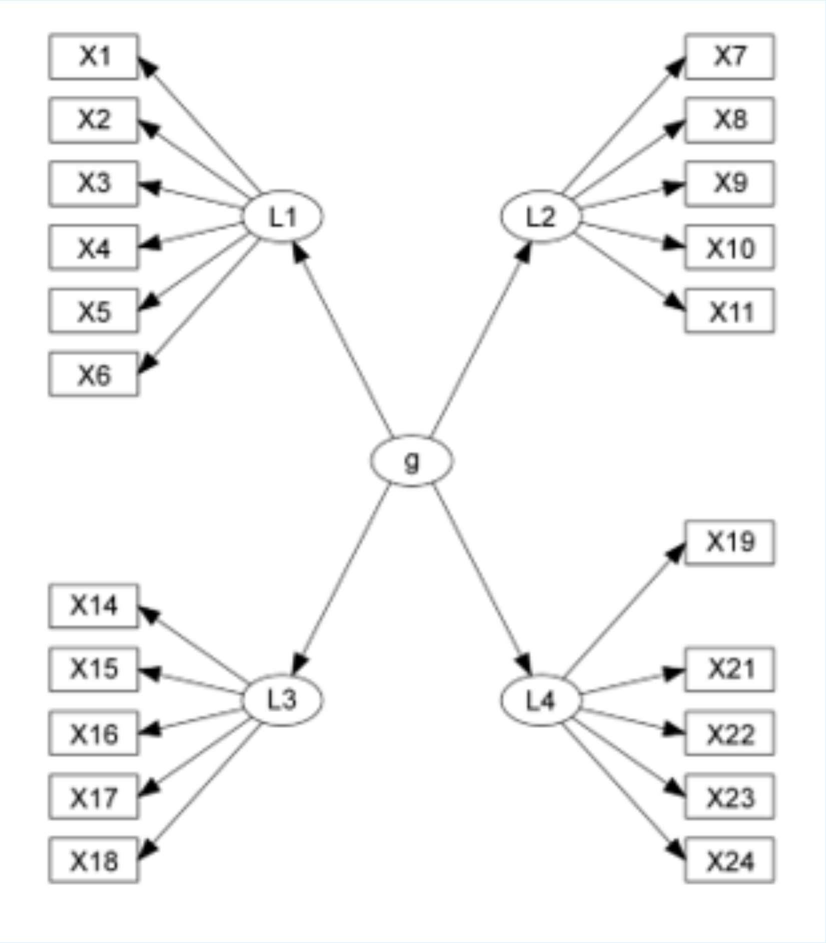
2. **After FOFC/FTFC**
   Locate and discard impure indicators
   Fit and compare *pure* versions of
   2nd-order, bi-factor models
   with *correct* clustering – on *retained* items

Truth



Purified Version of Truth

{X12, X13, X20} removed

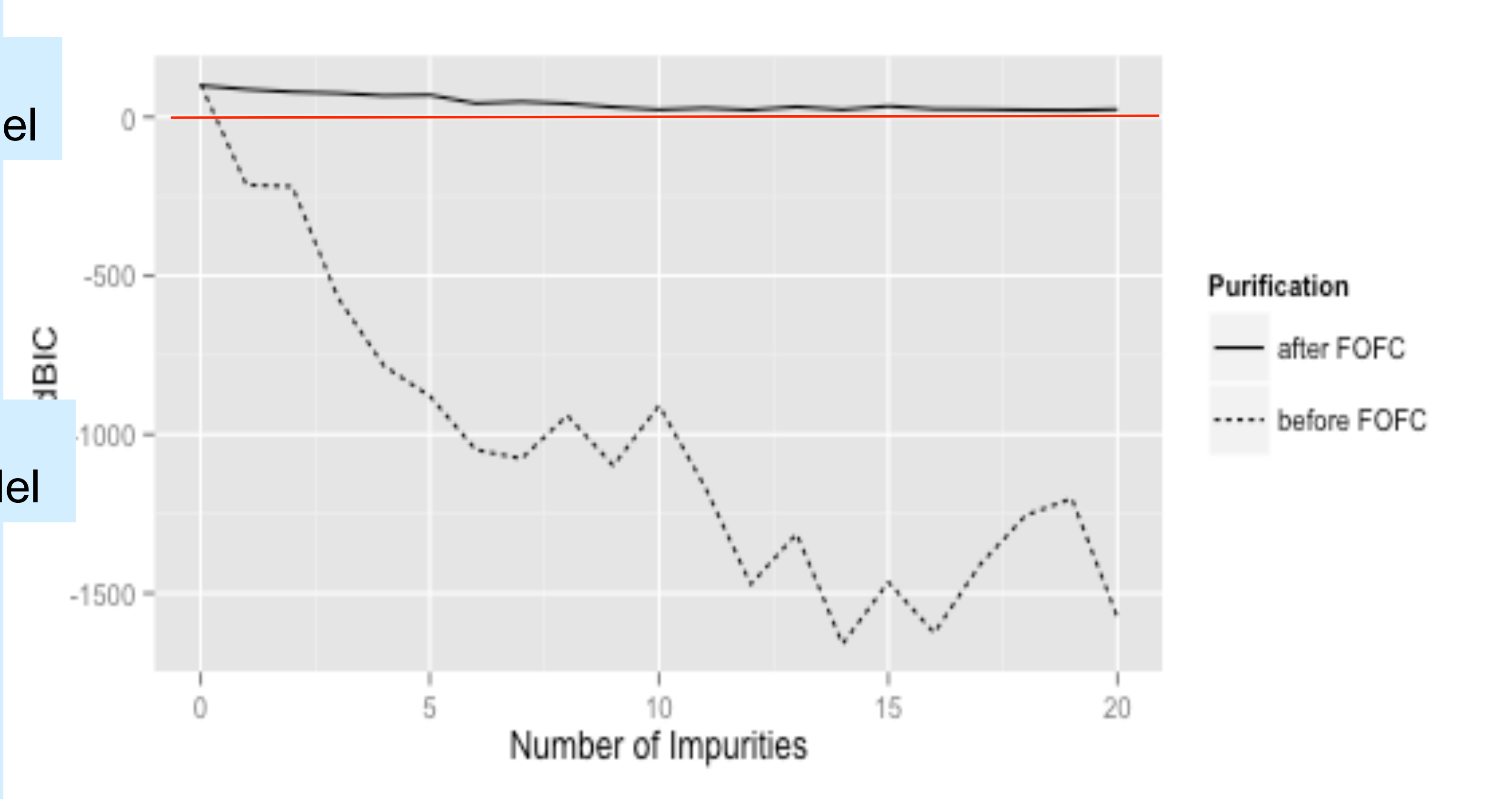


Simulation Study

Truth: 2nd-Order

dBIC > 0:
2nd-Order Model

dBIC < 0:
Bi-factor Model

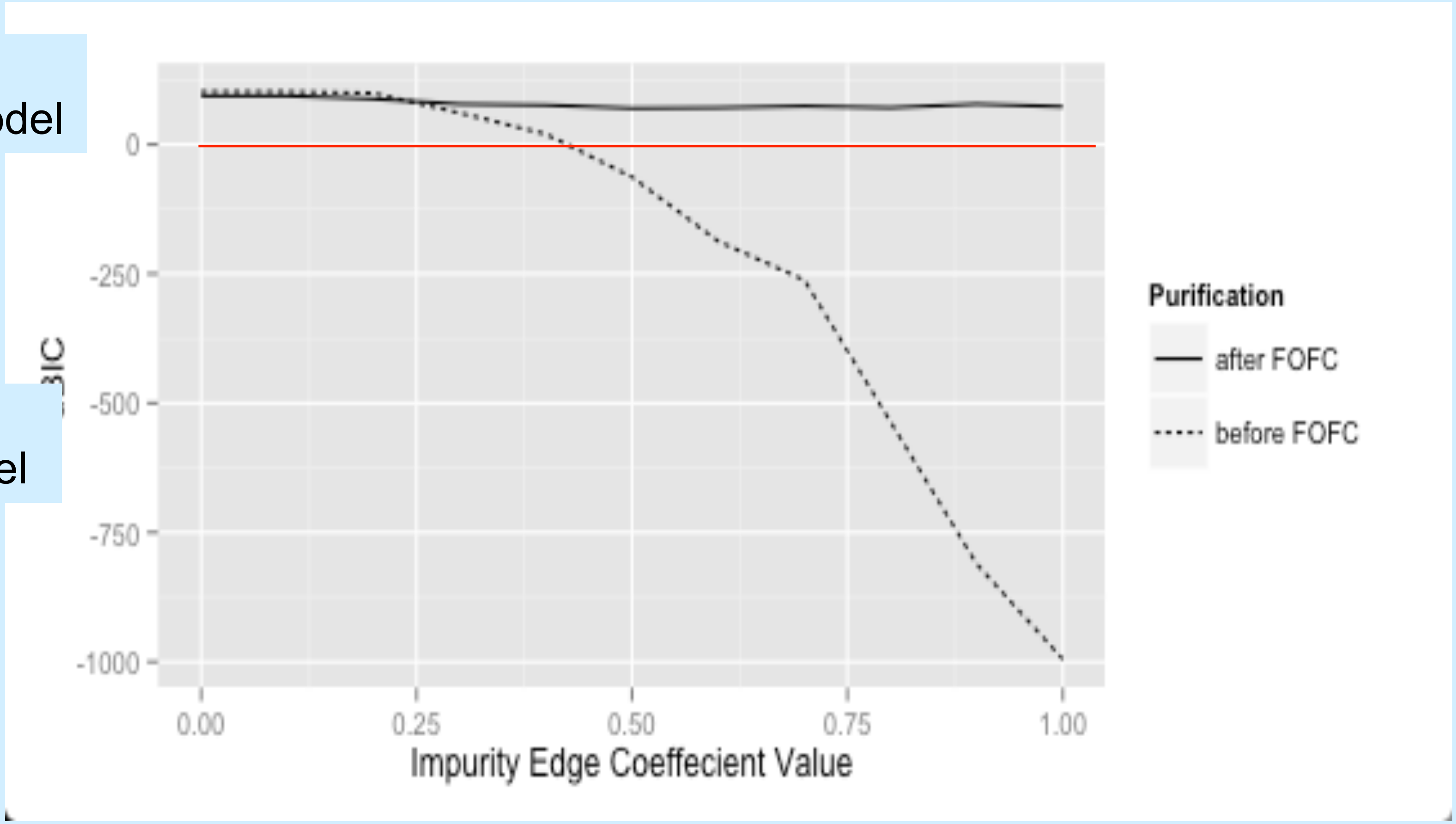


Size of the "impurities"

Truth: 2nd-Order

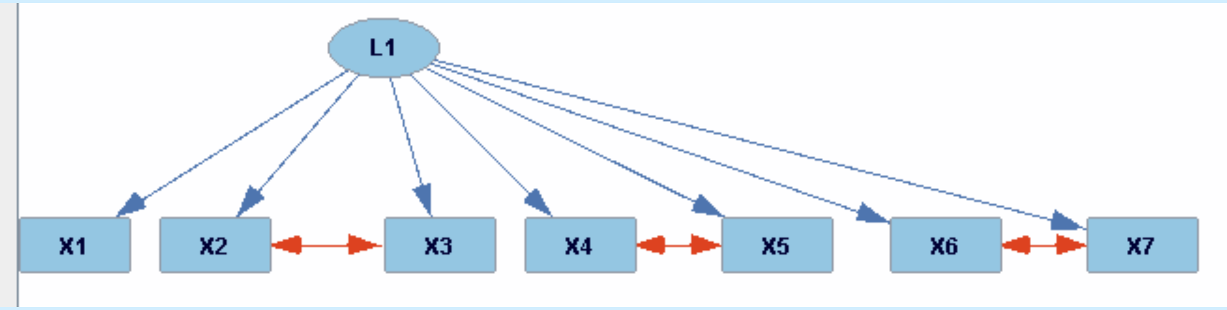
dBIC > 0:
2nd-Order Model

dBIC < 0:
Bi-factor Model



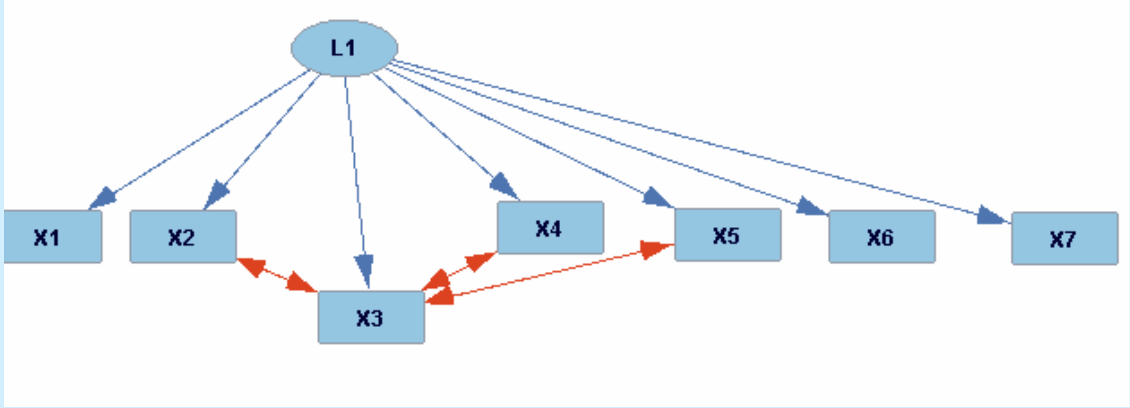
Level of Inhomogeneity

Inhomogeneiity: the degree to which the impurities are concentrated on a small number of indicators

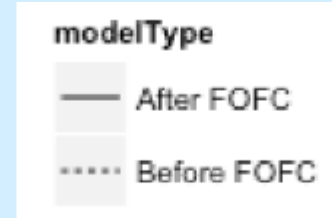


Low Inhomogeneiity

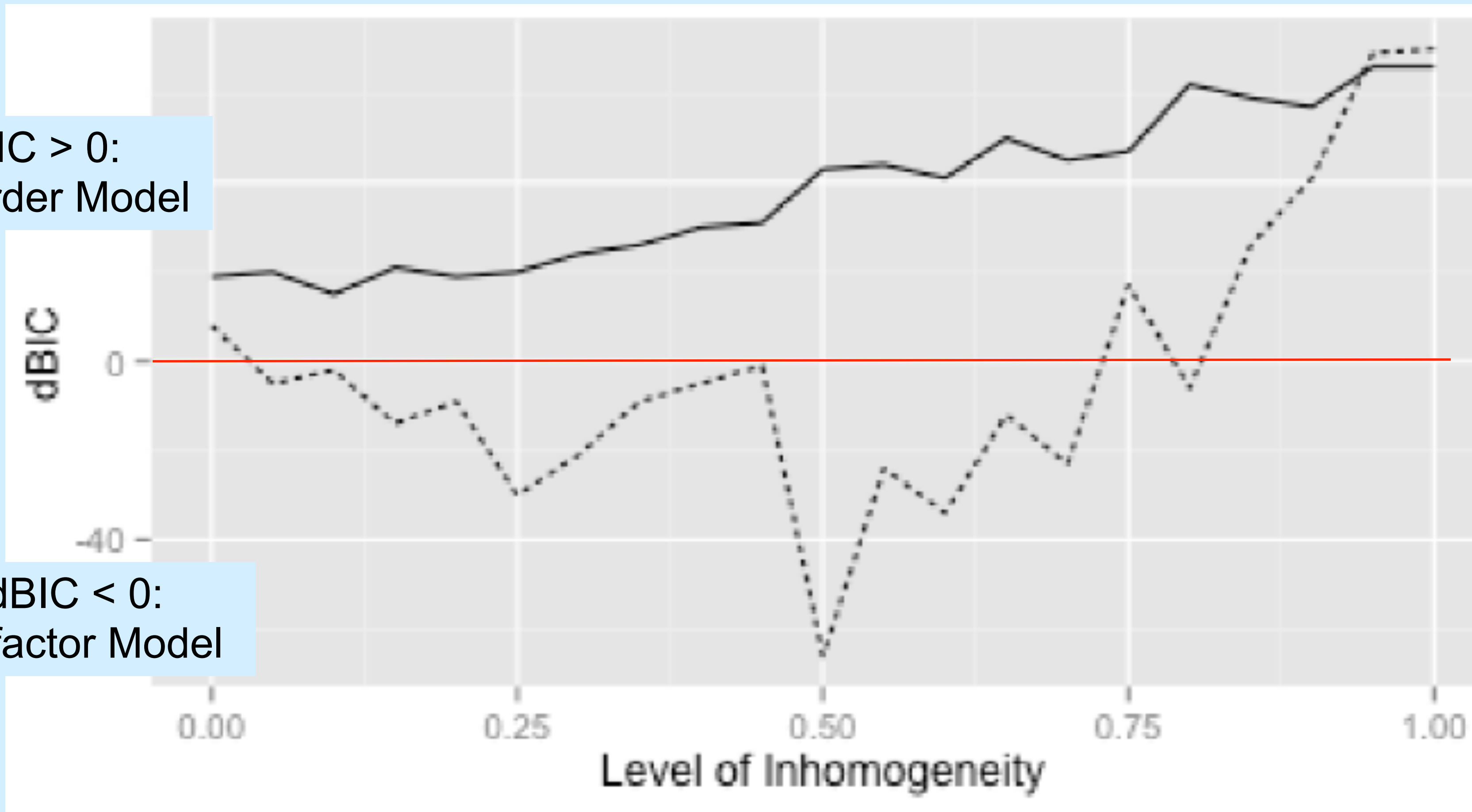
High Inhomogeneiity



Level of Inhomogeneity



Truth: 2nd-Order

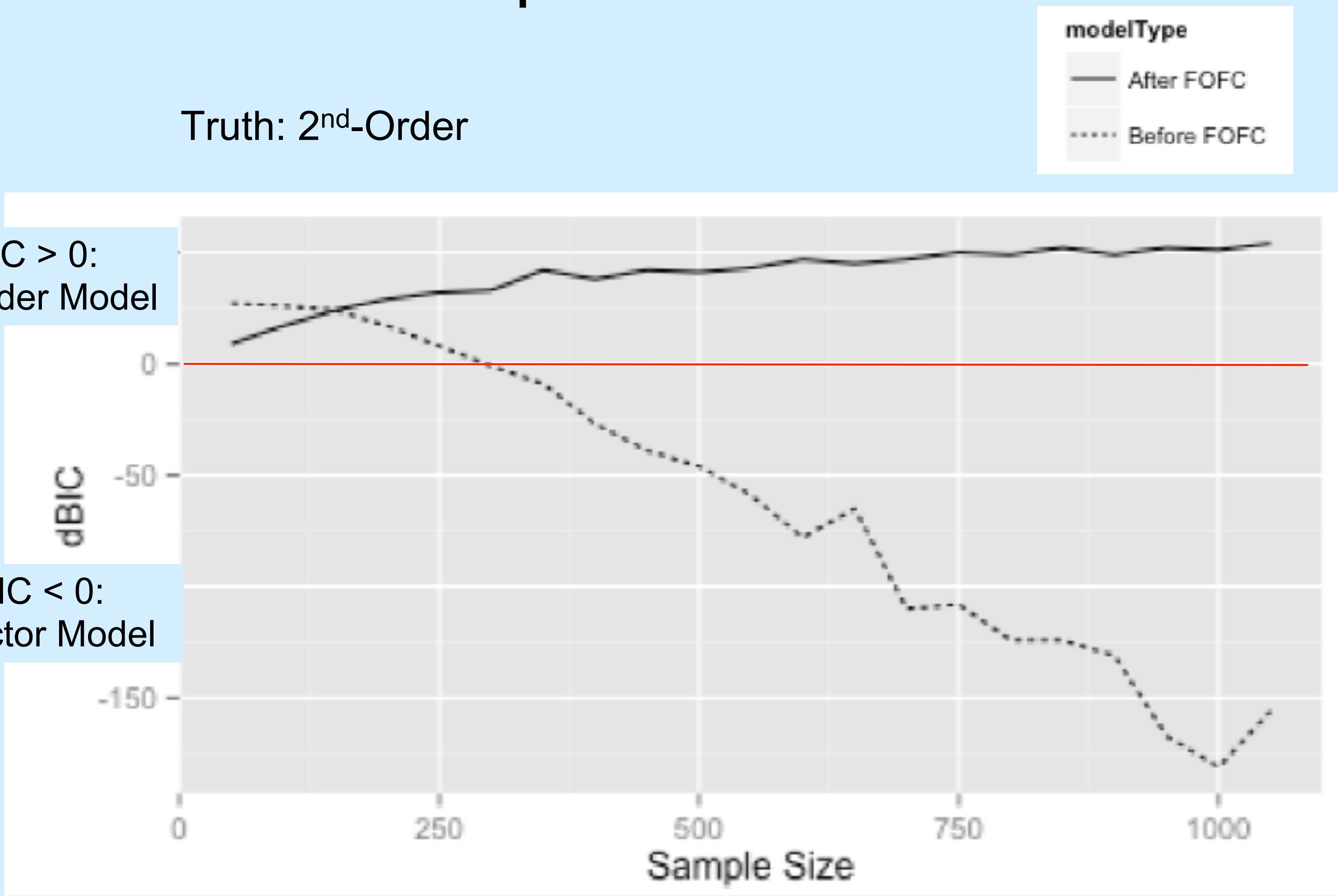


dBIC > 0:
2nd-Order Model

dBIC < 0:
Bi-factor Model

Sample Size

Truth: 2nd-Order



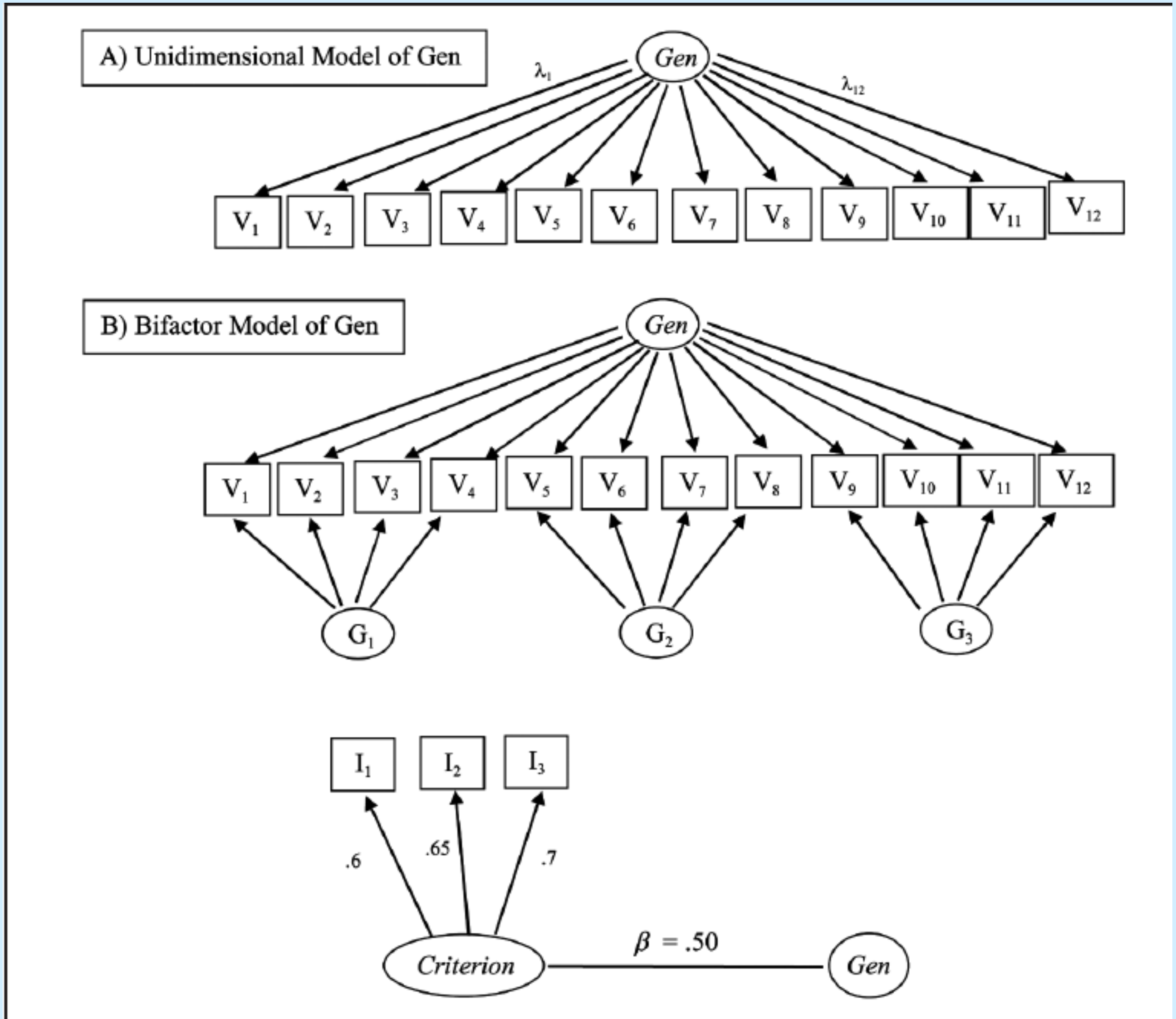
dBIC > 0:
2nd-Order Model

dBIC < 0:
Bi-factor Model

Thanks

Measurement Model Misspecification

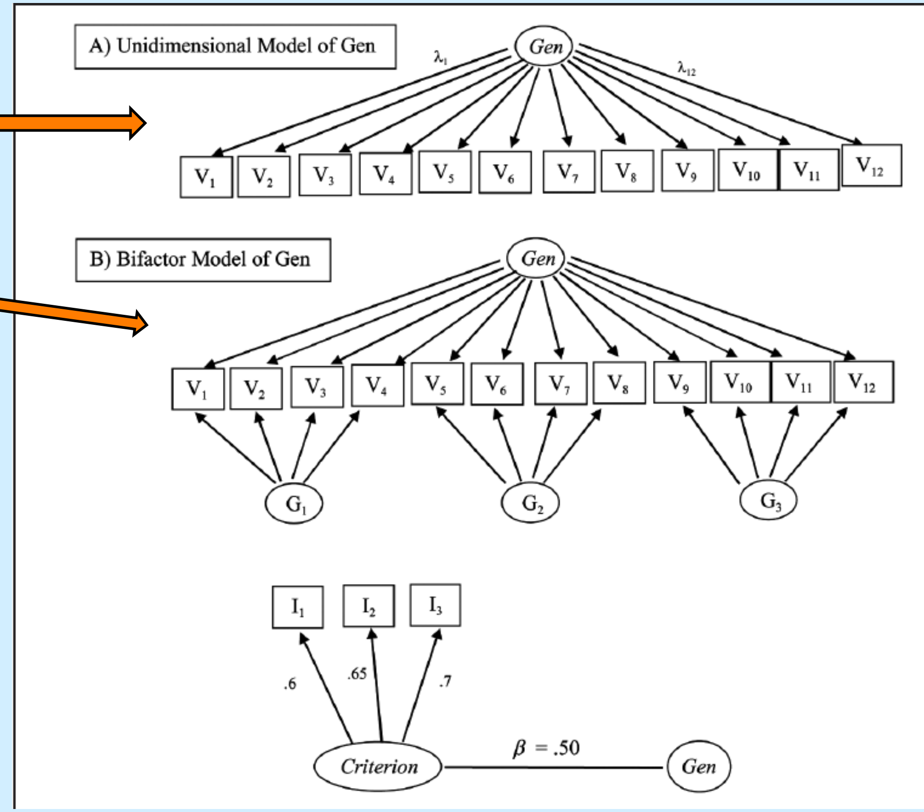
→ Structural Parameter Bias



Measurement Model Misspecification → Structural Parameter Bias

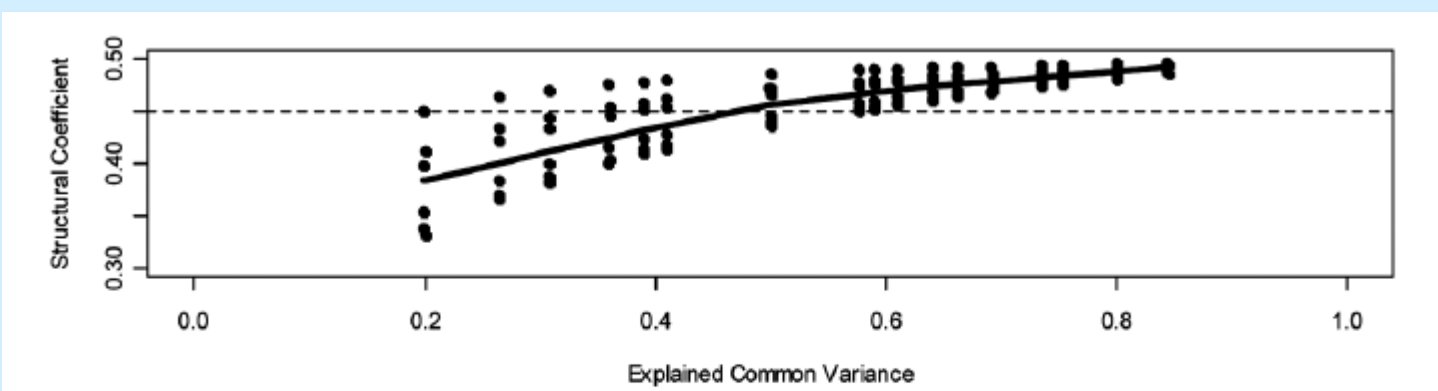
Specified Model

True Model

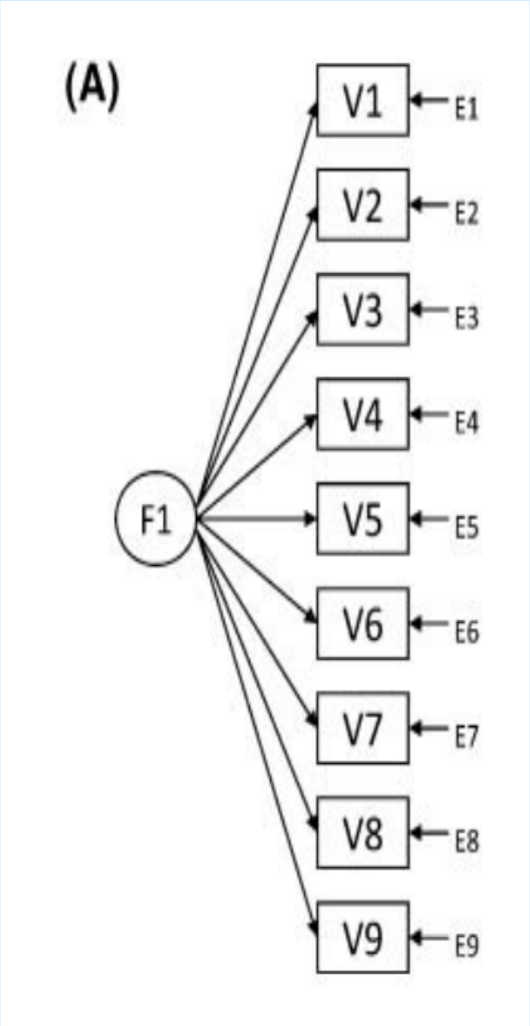


Parametric Measure of Difference:
Explained Common Variance (ECV)

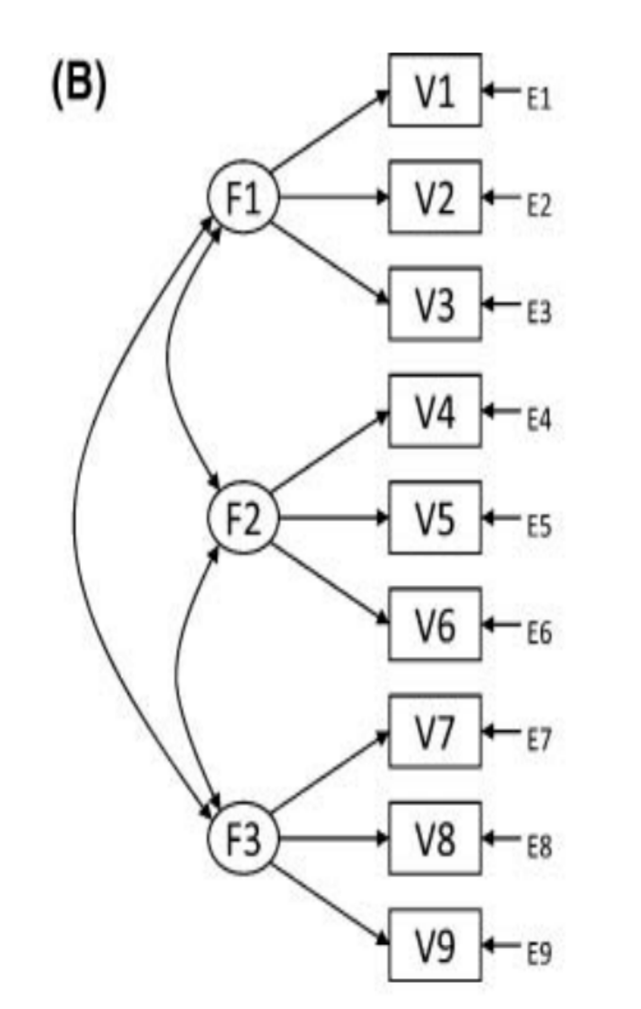
$$\text{ECV} = \frac{\sum \lambda_{Gen}^2}{\sum \lambda_{Gen}^2 + \sum \lambda_{GR1}^2 + \sum \lambda_{GR2}^2 + \sum \lambda_{GR3}^2}.$$



Measurement Models



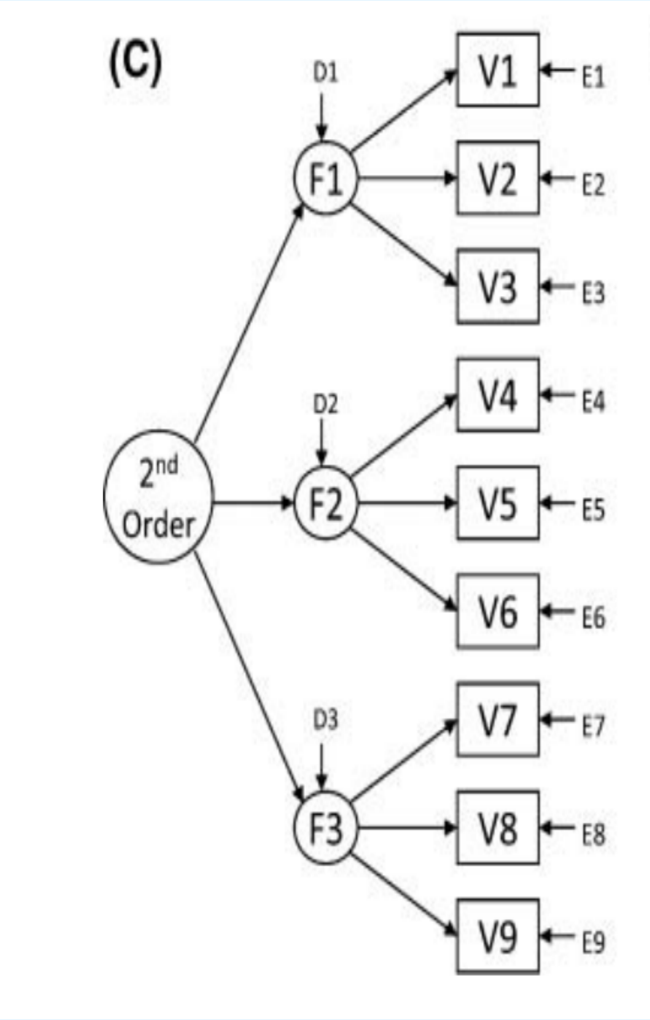
Unidimensional



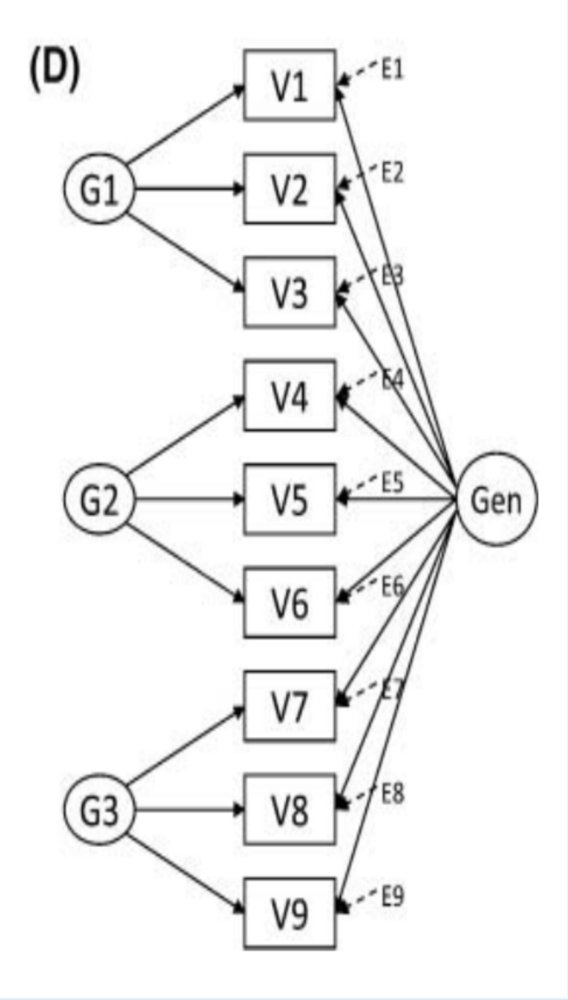
Correlated Trait

Psychometric Models in Practice:

Bi-factor vs. 2nd (Higher)-Order Models



2nd Order



Bi-factor