Rules for Two-stage Proofs of Identifiability[1]

STA 2101F 2019

[1] See last slide for copyright information.

Overview

The two-stage model: $cov(\mathbf{D}_i) = \boldsymbol{\Sigma}$

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

- $\mathbf{X}_i$ is $p \times 1$, $\mathbf{Y}_i$ is $q \times 1$, $\mathbf{D}_i$ is $k \times 1$.
- $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_x$, $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $cov(\mathbf{F}_i) = cov \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) = \boldsymbol{\Phi} = \left(\begin{array}{cc} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^\top & \boldsymbol{\Phi}_{22} \end{array} \right)$
- $cov(\mathbf{e}_i) = \boldsymbol{\Omega}$

Identify parameter matrices in two steps

It does not really matter which one you do first.

- $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$
  $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_x$, $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$
  $cov(\mathbf{F}_i) = \boldsymbol{\Phi}$, $cov(\boldsymbol{e}_i) = \boldsymbol{\Omega}$

---

1. *Latent model*: Show $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Psi}$ can be recovered from $\boldsymbol{\Phi} = cov\left(\begin{array}{c}\mathbf{X}_i \\ \mathbf{Y}_i\end{array}\right)$.
2. *Measurement model*: Show $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$ can be recovered from $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$.

This means all the parameters can be recovered from $\boldsymbol{\Sigma}$.

Parameter count rule

A necessary condition overall and at each stage

If a model has more parameters than covariance structure equations, the parameter vector can be identifiable on at most a set of volume zero in the parameter space.

All the following rules

- Are sufficient conditions for identifiability from the covariance matrix; they are not necessary conditions.
- Assume that errors are independent of exogenous variables that are not errors.
- Assume all variables have expected value zero, so these models have been re-parameterized by centering — or by just ignoring intercepts and expected values.

Latent Model Rules

- $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$
- Here, identifiability means that the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Psi}$ are functions of $cov(\mathbf{F}_i) = \boldsymbol{\Phi}$.

Regression Rule

Someimes called the Null Beta Rule

Suppose

- No endogenous variables influence other endogenous variables.

- $\mathbf{Y}_i = \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$
- Of course $cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) = \mathbf{0}$, always.
- $\boldsymbol{\Psi} = cov(\boldsymbol{\epsilon}_i)$ need not be diagonal.

Then $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ are identifiable.

Acyclic Rule

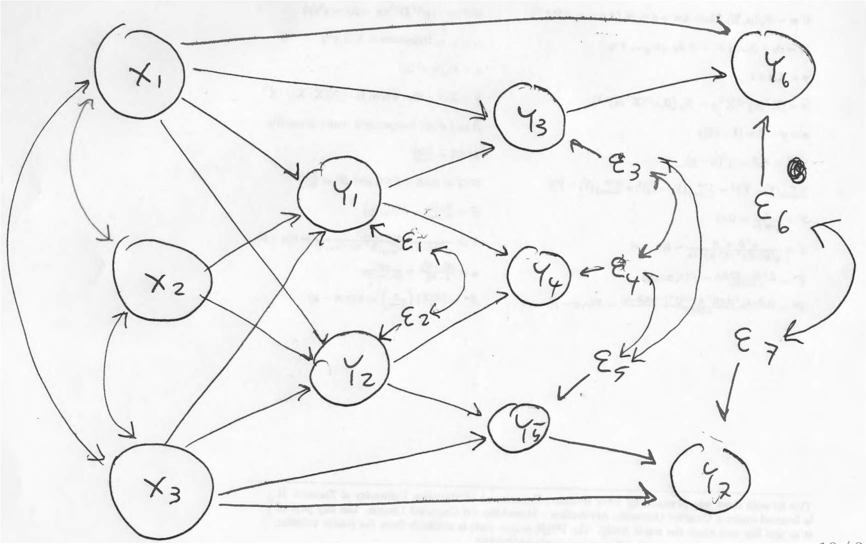
Note that each endogenous variable is influenced by exactly one error term, and by at least one other variable.

Parameters of the Latent Variable Model are identifiable if the model is acyclic (no feedback loops through straight arrows) and the following conditions hold.

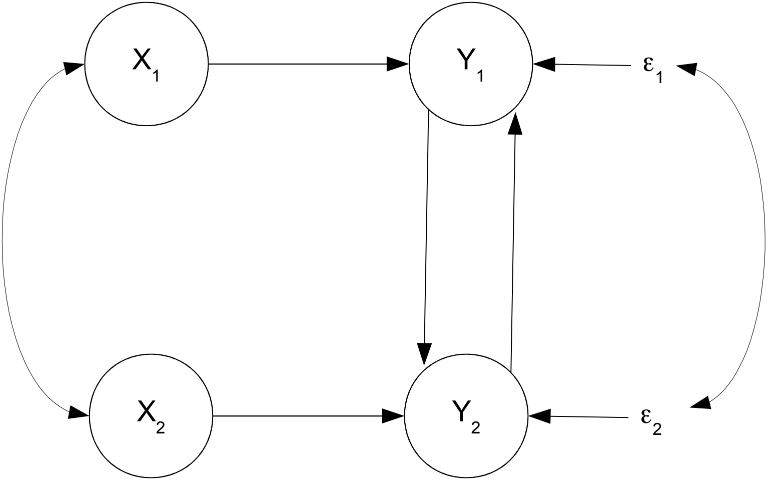
- Organize the variables that are not error terms into sets. Set 0 consists of all the exogenous variables.
- For $j = 1, \ldots, k$, each endogenous variable in set j is influenced by at least one variable in set $j-1$, and also possibly by variables in earlier sets.
- Error terms may be correlated within sets, but not between sets.

Proof: Repeated application of the Regression Rule.

An Acyclic model



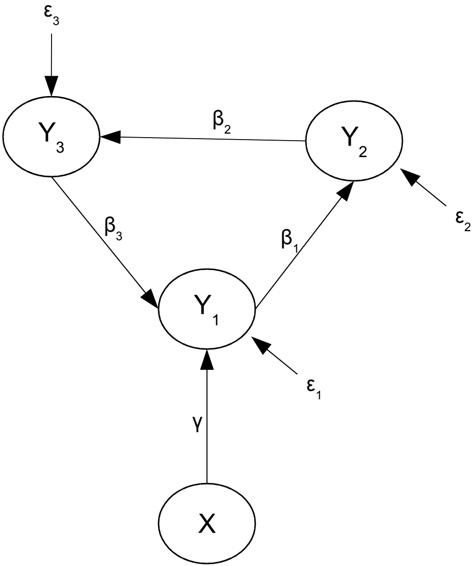
Parameters of this model are just identifiable



Shows that the acyclic rule is sufficient but not necessary.

The Pinwheel Model

Parameters are identifiable



Model equations for the 3-node Pinwheel Model

Assume all variances positive etc.

$$\begin{aligned} Y_1 &= \beta_3 Y_3 + \gamma X + \epsilon_1 \\ Y_2 &= \beta_1 Y_1 + \epsilon_2 \\ Y_3 &= \beta_2 Y_2 + \epsilon_3 \end{aligned}$$

In matrix terms:

$$\left(\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \end{array} \right) = \left(\begin{array}{ccc} 0 & 0 & \beta_3 \\ \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 \end{array} \right) \left(\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \end{array} \right) + \left(\begin{array}{c} \gamma \\ 0 \\ 0 \end{array} \right) X + \left(\begin{array}{c} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{array} \right)$$

To get $cov(\mathbf{Y})$

$$
\begin{aligned}
& \mathbf{Y} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{I}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}
\end{aligned}
$$

$(\mathbf{I} - \boldsymbol{\beta})^{-1}$ exists when $|\mathbf{I} - \boldsymbol{\beta}| \neq 0$

$$
\mathbf{I} - \boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & -\beta_3 \\ -\beta_1 & 1 & 0 \\ 0 & -\beta_2 & 1 \end{pmatrix}
$$

Calculate the determinant using Sage

```
sem = 'http://www.utstat.toronto.edu/~brunner/openSEM/sage/sem.sage'
load(sem)
B = ZeroMatrix(3,3)
B[0,2] = var('beta3'); B[1,0] = var('beta1'); B[2,1] = var('beta2')
ImB = IdentityMatrix(3)-B
show( ImB.determinant() )
```

$$-\beta_1\beta_2\beta_3 + 1$$

So the inverse will exist unless $\beta_1\beta_2\beta_3 = 1$.

Solve for Y_3

Starting with the model equations

$$\begin{array}{rcl} Y_1 & = & \beta_3 Y_3 + \gamma X + \epsilon_1 \\ Y_2 & = & \beta_1 Y_1 + \epsilon_2 \\ Y_3 & = & \beta_2 Y_2 + \epsilon_3 \end{array}$$

$$\begin{array}{rl} & Y_3 = \beta_1\beta_2\beta_3 Y_3 + \beta_1\beta_2\gamma X + \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3 \\ \Rightarrow & Y_3(1-\beta_1\beta_2\beta_3) = \beta_1\beta_2\gamma X + \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3 \end{array}$$

What happens if $(\mathbf{I}-\boldsymbol{\beta})^{-1}$ does not exist (and $\gamma \neq 0$)?

If $\beta_1\beta_2\beta_3 = 1$

Meaning that $(\mathbf{I}-\boldsymbol{\beta})^{-1}$ does not exist

$$
\begin{aligned}
& Y_3(1-\beta_1\beta_2\beta_3) = \beta_1\beta_2\gamma X + \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3 \\
\Rightarrow \quad & 0 = \beta_1\beta_2\gamma X + \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3 \\
\Rightarrow \quad & E(X\cdot 0) = E\left(X(\beta_1\beta_2\gamma X + \beta_1\beta_2\epsilon_1 + \beta_2\epsilon_2 + \epsilon_3)\right) \\
\Rightarrow \quad & 0 = \beta_1\beta_2\gamma E(X^2) + 0 \\
\Rightarrow \quad & \beta_1\beta_2\gamma\phi = 0
\end{aligned}
$$

with β_1, β_2, γ and ϕ all non-zero.

So $\beta_1\beta_2\beta_3 = 1$ contradicts the model.

Under the assumptions of the pinwheel model

- $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ exists.
- $\beta_1\beta_2\beta_3 \neq 1$.
- The surface $\beta_1\beta_2\beta_3 = 1$ forms a *hole* in the parameter space.

Covariance matrix of the factors: $\mathbf{\Phi}$

Factors are X, Y_1, Y_2, Y_3

$$\left(\begin{array}{cccc}
\phi & -\frac{\gamma\phi}{\beta_1\beta_2\beta_3-1} & -\frac{\beta_1\gamma\phi}{\beta_1\beta_2\beta_3-1} & -\frac{\beta_1\beta_2\gamma\phi}{\beta_1\beta_2\beta_3-1} \\
 & \frac{\beta_2^2\beta_3^2\psi_2+\beta_3^2\psi_3+\gamma^2\phi+\psi_1}{(\beta_1\beta_2\beta_3-1)^2} & \frac{\beta_1\beta_3^2\psi_3+\beta_1\gamma^2\phi+\beta_2\beta_3\psi_2+\beta_1\psi_1}{(\beta_1\beta_2\beta_3-1)^2} & \frac{\beta_1\beta_2\gamma^2\phi+\beta_2^2\beta_3\psi_2+\beta_1\beta_2\psi_1+\beta_3\psi_3}{(\beta_1\beta_2\beta_3-1)^2} \\
 & & \frac{\beta_1^2\beta_3^2\psi_3+\beta_1^2\gamma^2\phi+\beta_1^2\psi_1+\psi_2}{(\beta_1\beta_2\beta_3-1)^2} & \frac{\beta_1^2\beta_2\gamma^2\phi+\beta_1^2\beta_2\psi_1+\beta_1\beta_3\psi_3+\beta_2\psi_2}{(\beta_1\beta_2\beta_3-1)^2} \\
 & & & \frac{\beta_1^2\beta_2^2\gamma^2\phi+\beta_1^2\beta_2^2\psi_1+\beta_2^2\psi_2+\psi_3}{(\beta_1\beta_2\beta_3-1)^2}
\end{array}\right)$$

- $\phi = \phi_{11}$, $\beta_1 = \frac{\phi_{13}}{\phi_{12}}$ and $\beta_2 = \frac{\phi_{14}}{\phi_{13}}$ are easy.
- But then?

Solutions exist provided $\beta_1, \beta_2, \beta_3$ are all non-zero.

Using Sage …

$$\begin{aligned}
\beta_3 &= \frac{\phi_{12}\phi_{13}\phi_{23} - \phi_{13}^2\phi_{22}}{\phi_{12}\phi_{14}\phi_{33} - \phi_{13}\phi_{14}\phi_{23}} \\
\gamma &= \frac{\phi_{12}^2\phi_{33} - 2\,\phi_{12}\phi_{13}\phi_{23} + \phi_{13}^2\phi_{22}}{\phi_{11}\phi_{12}\phi_{33} - \phi_{11}\phi_{13}\phi_{23}} \\
\psi_3 &= \frac{(\phi_{13}\phi_{44} - \phi_{14}\phi_{34})(\phi_{12}^2\phi_{33} - 2\,\phi_{12}\phi_{13}\phi_{23} + \phi_{13}^2\phi_{22})}{(\phi_{12}\phi_{33} - \phi_{13}\phi_{23})\phi_{12}\phi_{13}} \\
\psi_2 &= \frac{\phi_{12}^2\phi_{33} - 2\,\phi_{12}\phi_{13}\phi_{23} + \phi_{13}^2\phi_{22}}{\phi_{12}^2} \\
\psi_1 &= \beta_1^2\beta_2^2\beta_3^2\phi_{22} - \beta_2^2\beta_3^2\psi_2 - 2\,\beta_1\beta_2\beta_3\phi_{22} - \beta_3^2\psi_3 - \gamma^2\phi + \phi_{22}
\end{aligned}$$

Parameters of this pinwheel model are identifiable

- Even though it does not fit any known rules
- And the proof is very difficult.

Measurement Model Rules

Factor Analysis

- In these rules, latent variables that are not error terms are called "factors."
- Unless otherwise noted, factors may have non-zero covariances with each other.
- All the models are surrogate models.

Double Measurement Rule

This has been proved

Model parameters are identifiable provided

- Each factor is measured twice.
- All factor loadings equal one.
- There are two sets of measurements, set one and set two.
- Correlated measurement errors are allowed within sets of measurements, but not between sets.

■

Three-Variable Rule for Standardized Factors

This has been proved

Model parameters are identifiable provided

- Errors are independent of one another.
- Each observed variable is influenced by only one factor.
- The variance of each factor equals one.
- There are at least 3 observed variables with non-zero loadings per factor.
- The sign of one non-zero loading is known for each factor.

■

Three-Variable Rule for Unstandardized Factors

This has been proved

Model parameters are identifiable provided

- Errors are independent of one another.
- Each observed variable is influenced by only one factor.
- For each factor, at least one factor loading equals one.
- There are at least 2 additional observed variables with non-zero loadings per factor.

■

Two-Variable Rule for Standardized Factors

A factor with just two observed variables may be added to a measurement model whose parameters are identifiable, and the parameters of the combined model will be identifiable provided

- The errors for the two additional observed variables are independent of one another and of those already in the model.
- The two new observed variables are influenced only by the new factor.
- The variance of the new factor equals one.
- Both new factor loadings are non-zero.
- The sign of one new loading is known.
- The new factor has a non-zero covariance with at least one factor already in the model.

■

Two-Variable Rule for Unstandardized Factors

A factor with just two observed variables may be added to a measurement model whose parameters are identifiable, and the parameters of the combined model will be identifiable provided

- The errors for the two additional observed variables are independent of one another and of those already in the model.
- The two new observed variables are influenced only by the new factor.
- At least one new factor loading equals one.
- The other new factor loading is non-zero.
- The new factor has a non-zero covariance with at least one factor already in the model.

■

Four-variable Two-factor Rule

The parameters of a measurement model with two factors and four observed variables will be identifiable provided

- All errors are independent of one another.
- Each observed variable is influenced by only one factor.
- Two observed variables are influenced by one factor, and two are influenced by the other.
- All factor loadings are non-zero.
- For each factor, either the variance of the factor equals one and the sign of one new loading is known, or at least one factor loading equals one.
- The covariance of the two factors does not equal zero.

■

Proof of the Four-variable Two-factor Rule

With standardized factors

The model equations are

$$
\begin{aligned}
D_1 &= \lambda_1 F_1 + e_1 \\
D_2 &= \lambda_2 F_1 + e_2 \\
D_3 &= \lambda_3 F_2 + e_3 \\
D_4 &= \lambda_4 F_2 + e_4,
\end{aligned}
$$

where all expected values are zero, $Var(e_j) = \omega_j$ for $j = 1, \ldots, 4$, and

$$
cov \left(\begin{array}{c} F_1 \\ F_2 \end{array} \right) = \left(\begin{array}{cc} 1 & \phi_{12} \\ \phi_{12} & 1 \end{array} \right)
$$

with $\phi_{12} \neq 0$. Also suppose $\lambda_1 > 0$, $\lambda_2 \neq 0$, $\lambda_3 > 0$ and $\lambda_4 \neq 0$.

Covariance matrix

For the 4-variable 2-factor problem

$$
\begin{aligned}
D_1 &= \lambda_1 F_1 + e_1 \\
D_2 &= \lambda_2 F_1 + e_2 \\
D_3 &= \lambda_3 F_2 + e_3 \\
D_4 &= \lambda_4 F_2 + e_4,
\end{aligned}
$$

$$
\boldsymbol{\Sigma} = \begin{array}{c|cccc}
 & D_1 & D_2 & D_3 & D_4 \\ \hline
D_1 & \lambda_1^2 + \omega_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3\phi_{12} & \lambda_1\lambda_4\phi_{12} \\
D_2 & & \lambda_2^2 + \omega_2 & \lambda_2\lambda_3\phi_{12} & \lambda_2\lambda_4\phi_{12} \\
D_3 & & & \lambda_3^2 + \omega_3 & \lambda_3\lambda_4 \\
D_4 & & & & \lambda_4^2 + \omega_4
\end{array}
$$

Using the assumption that $\lambda_1 > 0$ and $\lambda_3 > 0$

$$
\boldsymbol{\Sigma} = \begin{array}{c|cccc}
 & D_1 & D_2 & D_3 & D_4 \\ \hline
D_1 & \lambda_1^2 + \omega_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3\phi_{12} & \lambda_1\lambda_4\phi_{12} \\
D_2 & & \lambda_2^2 + \omega_2 & \lambda_2\lambda_3\phi_{12} & \lambda_2\lambda_4\phi_{12} \\
D_3 & & & \lambda_3^2 + \omega_3 & \lambda_3\lambda_4 \\
D_4 & & & & \lambda_4^2 + \omega_4
\end{array}
$$

$$
\frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} = \frac{\lambda_1^2\lambda_2\lambda_3\phi_{12}}{\lambda_2\lambda_3\phi_{12}} = \lambda_1^2
$$

$$
\Rightarrow \quad \lambda_1 = \sqrt{\frac{\sigma_{12}\sigma_{13}}{\sigma_{23}}}
$$

Similarly, $\lambda_3 = \sqrt{\frac{\sigma_{34}\sigma_{23}}{\sigma_{24}}}$, and the rest is easy.

Please don't do both!

Don't set the variance *and* a factor loading to one!

- Setting the variance of factors to one looks arbitrary, but it's really a smart re-parameterization.
- Setting one loading per factor to one also is a smart re-parameterization.
- It's smart because the resulting models impose the *same restrictions on the covariance that the original model does.*
- And, the *meanings* of the parameters have a clear connection to the meanings of the parameters of the original model.
- But if you do both, it's a mess. Most or all of the meaning is lost.
- And you put an *extra* restriction on $\boldsymbol{\Sigma}$ that is not implied by the original model.

Combination Rule

Suppose that the parameters of two measurement models are identifiable by any of the rules above. The two models may be combined into a single model provided that the error terms of the first model are independent of the error terms in the second model. The additional parameters of the combined model are the covariances between the two sets of factors, and these are all identifiable. ■

Cross-over Rule

This has been proved

Suppose that

- The parameters of a measurement models are identifiable, and
- For each factor there is at least one observed variable that is influenced only by that factor (with a non-zero factor loading).

Then any number of new observed variables may be added to the model and the result is a model whose parameters are all identifiable, provided that

- The error terms associated with the new observed variables are independent of the error terms in the existing model.

Each new observed variable may be influenced by any or all of the factors, potentially resulting in a cross-over pattern in the path diagram. The error terms associated with the new set of observed variables may be correlated with one another. Note that no new factors are added. ■

Error-Free Rule

This has been proved

A vector of observed variables may be added to the factors of a measurement model whose parameters are identifiable. Suppose that

- The new observed variables are independent of the errors in the measurement model, and
- For each factor in the measurement model there is at least one observed variable that is influenced only by that factor (with a non-zero factor loading).

Then the parameters of a new measurement model, where some of the observed variables are assumed to be measured without error, are identifiable. The practical consequence is that variables assumed to be measured without error may be included in the latent component of a structural equation model, provided that the measurement model for the other variables has identifiable parameters.

Copyright Information