

# Introduction to Structural Equation Models




















































April 2013





















































Paul D. Allison, Ph.D.  
Instructor

[www.StatisticalHorizons.com](http://www.StatisticalHorizons.com)
















Copyright © 2013 by Paul D. Allison

1	 Introduction to Structural Equation Models
2	 Structural Equation Models
3	 SEM
4	 Preview: A Latent Variable Model
5	 Latent Variable Model (cont.)
6	 Cautions
7	 Outline
8	 Software for SEMs
9	 Path Analysis of Observed Variables
10	 Some Rules and Definitions
11	 Three Predictor Variables
12	 Two-Equation System
13	 Why combine the two equations?
14	 Calculation of Indirect Effect
15	 A More Complex Model
16	 Decomposition of Direct & Indirect Effects
17	 Standardized Coefficients
18	 Numerical Examples
19	 More Complex Example
20	 Decomposition of Effects
21	 Examples Using Mplus
22	 Example 1. Data
23	 Mplus Program
24	 Selected Output
25	 Output (cont.)
26	 SAS Code for Example 1
27	 Path Diagram with Standardized Coefficients
28	 Indirect Effects
29	 Indirect Results
30	 Indirect Effects in SAS
31	 Example 2: From Rex Kline (2010)
32	 Sample Moment Matrix
33	 Example 2: Mplus Program
34	 Covariance Matrix
35	 Example 2: Results
36	 Standardized Results
37	 SAS Code for Example 2
38	 Partial Correlations
39	 Partial Correlations (cont.)
40	 Partial Correlations in Mplus
41	 Results with Partial Correlation
42	 Results (cont.)
43	 Partial Correlations in SAS
44	 Causal Ordering
45	 How to Decide
46	 Nonrecursive Systems
47	 Identification Problem in Nonrecursive Models
48	 Identification Problem (cont.)
49	 A Just-Identified Model
50	 Reduced Form Equations
51	 Solutions for Structural Parameters

52	 Sufficient Condition for Identification
53	 Varieties of Identification
54	 Problems with Instrumental Variables
55	 Example 3. Nonrecursive Model with Mplus
56	 Example 3 (cont.)
57	 Example 3. Mplus Program
58	 Example 3. Results
59	 Example 3. Results (cont.)
60	 SAS Code for Nonrecursive Model
61	 Classical Test Theory
62	 Random Measurement Error
63	 Reliability
64	 Parallel Measures
65	 Tau-Equivalent Measures
66	 Tau-Equivalence: Example
67	 Tau-Equivalence in Mplus
68	 Tau-Equivalence in SAS
69	 Congeneric Tests
70	 Three Congeneric Tests
71	 Three Congeneric Measures (cont.)
72	 Standardized Version
73	 Digression: Tracing Rule for Correlations
74	 Tracing Rule (cont.)
75	 Tracing Rule (cont.)
76	 Standardized Version (cont.)
77	 Three Congenerics: Example
78	 Three Congenerics: Mplus
79	 Three Congenerics: SAS
80	 Four Congeneric Measures
81	 Overidentification with 4 Congeneric Measures
82	 Four Congeneric Measures with Mplus
83	 Four Congeneric Measures with SAS
84	 Results for Four Congenerics
85	 Alternative Model for 4 Congeneric Measures
86	 Mplus for Alternative Model
87	 Results for Alternative Model
88	 Results (cont.)
89	 SAS Code for Alternative Model
90	 Heywood Case
91	 Factor Models
92	 Factor Models (cont.)
93	 Identification (Standardized)
94	 Identification (cont.)
95	 Two Approaches to Identification Problem
96	 Identification (Unstandardized)
97	 Determining Identification
98	 Normalizing Constraints
99	 Normalizing Constraints
100	 ML Estimation of CFA Models
101	 Multivariate Normality
102	 ML Details
103	 Chi-Square Test

104	<b>Example 4. Self-Concept Measurement</b>
105	<b>Self Concept Path Diagram</b>
106	<b>SAS Code for Example 4</b>
107	<b>Example 4. Mplus Results</b>
108	<b>Example 4. Results</b>
109	<b>Global Goodness of Fit Measures</b>
110	<b>Other Global Measures</b>
111	<b>Other Global Measures (cont.)</b>
112	<b>Specific Goodness of Fit Measures</b>
113	<b>Standardized Residuals for Self-Concept Model</b>
114	<b>Residuals in SAS</b>
115	<b>Modification Indices</b>
116	<b>Mod Indices for Self-Concept</b>
117	<b>Freeing Up Parameters</b>
118	<b>Results from Freeing 1 Parameter</b>
119	<b>Selected Results (cont.)</b>
120	<b>Correlated Errors</b>
121	<b>Two Correlated Errors</b>
122	<b>A Five-Indicator Model</b>
123	<b>A Two-Factor Model</b>
124	<b>Example: Self-Concept Data</b>
125	<b>Selected Results</b>
126	<b>Structural Relations Among Latent Variables</b>
127	<b>Example 5. 98 Farm Managers (Rock et al. 1977)</b>
128	<b>Farm Managers Path Diagram</b>
129	<b>Data and Mplus Code</b>
130	<b>Example 5: SAS Code</b>
131	<b>Example 5. Selected Results</b>
132	<b>Example 5. Selected Results (cont.)</b>
133	<b>A Tau-Equivalent Model</b>
134	<b>Parallel Model</b>
135	<b>Parallel Model in SAS</b>
136	<b>Identification in SEM Models</b>
137	<b>An Identified SEM Model</b>
138	<b>What to Do When a Model Doesn't Fit</b>
139	<b>Alternative Estimation Methods</b>
140	<b>GLS Example</b>
141	<b>GLS Results</b>
142	<b>Weighted Least Squares</b>
143	<b>WLS Example</b>
144	<b>WLS Output</b>
145	<b>Other ESTIMATOR Options</b>
146	<b>Multiple Group Analysis</b>
147	<b>Example 6. Multiple Groups</b>
148	<b>Example 6. Data</b>
149	<b>Example 6. Models</b>
150	<b>Example 6. Mplus Code for Model 1</b>
151	<b>Example 6. Mplus Code (cont.)</b>
152	<b>Example 6. Model 4 Code</b>
153	<b>Tests for Comparing the Groups</b>
154	<b>Interactions and Non-Linearities</b>
155	<b>Interactions with Latent Variables</b>

156		<b>Interactions with Latent Variables</b>
157		<b>Ordinal and Binary Data</b>
158		<b>Special Correlations</b>
159		<b>Special Correlations</b>
160		<b>Specialized Models</b>
161		<b>Mplus with Binary Data</b>
162		<b>Probit Results</b>
163		<b>Probit Results (cont.)</b>
164		<b>Other Features of Mplus</b>
165		<b>Cautions About SEMs</b>
166		<b>Misuse</b>
167		<b>Good Examples</b>
168		<b>SEMs and Causality</b>

# Introduction to Structural Equation Models

Paul D. Allison, Instructor

April 2013

[www.StatisticalHorizons.com](http://www.StatisticalHorizons.com)

1

## Structural Equation Models

The classic SEM model includes many common linear models used in the behavioral sciences:

- Multiple regression
- ANOVA
- Path Analysis
- Multivariate ANOVA and regression
- Factor Analysis
- Canonical Correlation
- Non-recursive simultaneous equations (econometrics)
- Seemingly unrelated regressions

2

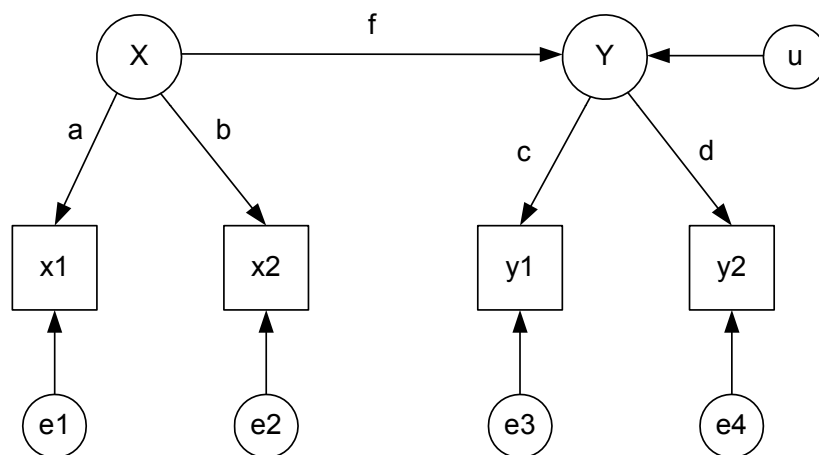
# SEM

## Convergence of psychometrics and econometrics

- Simultaneous equation models, possibly with reciprocal (nonrecursive) relationships
- Latent (unobserved) variables with multiple indicators.
- This course emphasizes models with latent variables. For example:

3

### Preview: A Latent Variable Model



$X$  and  $Y$  are unobserved variables,  $x_1$ ,  $x_2$ ,  $y_1$ , and  $y_2$  are observed indicators,  $e_1$ - $e_4$  and  $u$  are random errors.  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $f$  are correlation coefficients.

4

## Latent Variable Model (cont.)

- If we know the six correlations among the observed variables, simple hand calculations can produce estimates of  $\alpha$  through  $f$ . We can also test the fit of the model.
- Why is it desirable to estimate models like this?
  - Most variables are measured with at least some error.
  - In a regression model, measurement error in independent variables can produce severe bias in coefficient estimates.
  - We can correct this bias if we have multiple indicators for variables with measurement error.
  - Multiple indicators can also yield more powerful hypothesis tests.

5

## Cautions

- Although SEM's can be very useful, the methodology is often used badly and indiscriminately.
  - Often applied to data where it's inappropriate.
  - Can sometimes obscure rather than illuminate.
  - Easy to get sucked into overly complex modeling.

6



# Outline

1. Introduction to SEM
2. Path analysis of observed variables
3. Direct and indirect effects
4. Identification problem in nonrecursive models
5. Reliability: parallel and tau-equivalent measures
6. Multiple indicators of latent variables
7. Exploratory factor analysis
8. Confirmatory factor analysis
9. Goodness of fit measures
10. Structural relations among latent variables
11. Alternative estimation methods.
12. Multiple group analysis
13. Models for ordinal and nominal data

7

## Software for SEMs

**LISREL** – Karl Jöreskog and Dag Sörbom

**EQS** – Peter Bentler

**PROC CALIS (SAS)** – Wolfgang Hartmann

**MX** (freeware) – M.C. Neale

**Amos** – James Arbuckle

**Mplus** – Bengt Muthén

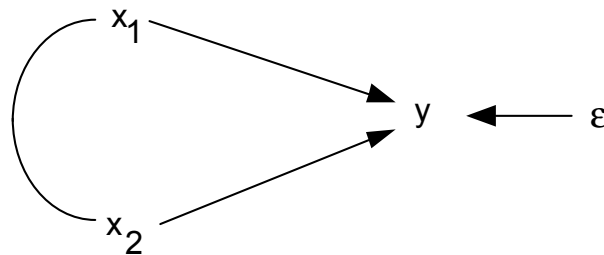
**sem (Stata)**

8

# Path Analysis of Observed Variables

In the SEM literature, it's common to represent a linear model by a path diagram.

- A diagrammatic method for representing a system of linear equations. There are precise rules so that you can write down equations from looking at the diagram.
- Single equation:  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



9

## Some Rules and Definitions



Direct causal effect



Correlation  
(no causal assumptions)

Why the curved line in the diagram?

Because omitting it implies that  $\rho_{12} = 0$ .

Endogenous variables: Variables caused by other variables in the system. These variables have straight arrows leading into them.

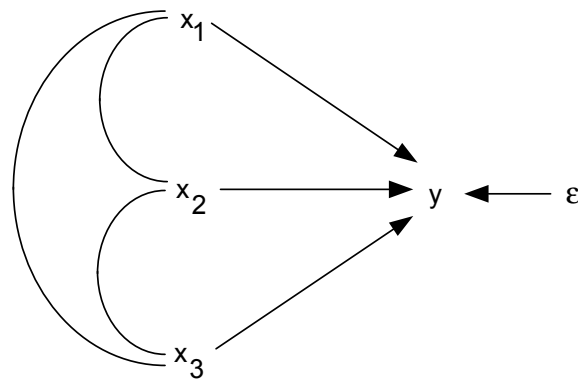
Exogenous variables: Variables not caused by others in the system. No straight arrows leading into them.

Not the same as dependent and independent because a variable that is dependent in one equation and independent in another equation is still endogenous.

Curved lines can only link *exogenous* variables.

10

## Three Predictor Variables



The fact that there are no curved arrows between  $\varepsilon$  and the  $x$ 's implies that  $\rho_{1\varepsilon} = 0$ ,  $\rho_{2\varepsilon} = 0$ , and  $\rho_{3\varepsilon} = 0$ . We make this assumption in the usual linear regression model.

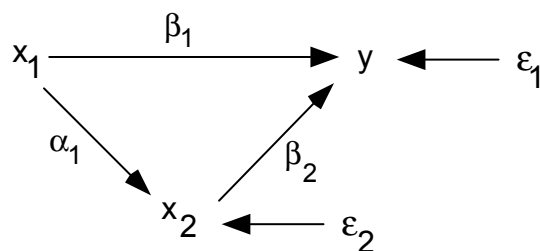
11

## Two-Equation System

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1$$

$$x_2 = \alpha_0 + \alpha_1 x_1 + \varepsilon_2$$

The diagram is now



Note: The diagram goes further than the equations by asserting that

$$\rho_{\varepsilon_1 \varepsilon_2} = 0, \rho_{\varepsilon_1 x_1} = 0, \rho_{\varepsilon_1 x_2} = 0, \rho_{x_1 \varepsilon_2} = 0$$

12

# Why combine the two equations?

Answer: to get further insight into the causal process.

To make this more concrete, let's suppose that

$y$  = income

$x_1$  = father's income

$x_2$  = years of schooling

What happens when you increase  $x_1$  by one unit? Then  $y$  changes by  $\beta_1$  units, holding  $x_2$  constant.

This can be misleading, however, because a one-unit increase in  $x_1$  *also* produces a change of  $\alpha_1$  units in  $x_2$ , which in turn produces a change in  $y$ .

Thus  $x_1$  has both a *direct* and an *indirect* effect on  $y$ . You wouldn't notice this with a single equation.

13

## Calculation of Indirect Effect

Substitute one equation into the other.

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 (\alpha_0 + \alpha_1 x_1 + \varepsilon_2) + \varepsilon_1 \\ &= (\beta_0 + \alpha_0 \beta_2) + (\beta_1 + \alpha_1 \beta_2) x_1 + (\varepsilon_1 + \beta_2 \varepsilon_2) \end{aligned}$$

The direct effect of  $x_1$  is  $\beta_1$ .

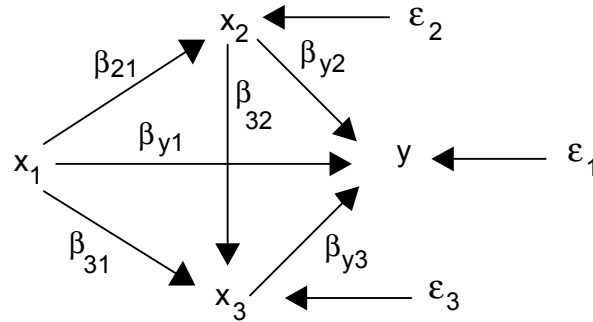
The indirect effect of  $x_1$  is  $\alpha_1 \beta_2$ .

The total effect of  $x_1$  is  $\beta_1 + \alpha_1 \beta_2$

**For recursive systems, indirect effects may be calculated by taking the product of coefficients on a particular path.**

14

## A More Complex Model



This diagram corresponds to three equations:

$$y = \beta_{y0} + \beta_{y1}x_1 + \beta_{y2}x_2 + \beta_{y3}x_3 + \varepsilon_1$$

$$x_3 = \beta_{30} + \beta_{31}x_1 + \beta_{32}x_2 + \varepsilon_3$$

$$x_2 = \beta_{20} + \beta_{21}x_1 + \varepsilon_2$$

15

## Decomposition of Direct & Indirect Effects

Decomposition of the total effect of  $x_1$  on  $y$ :

$\beta_{y1}$	Direct effect
$\beta_{31}\beta_{y3}$	Indirect effect through $x_3$
$\beta_{21}\beta_{y2}$	Indirect effect through $x_2$
$\beta_{21}\beta_{32}\beta_{y3}$	Indirect effect through $x_2$ and $x_3$

Decomposition of total effect of  $x_2$  on  $y$ :

$\beta_{y2}$	Direct effect
$\beta_{32}\beta_{y3}$	Indirect effect through $x_3$

As we'll see, you can use these results to compute percentages of the total effect. This works with either standardized or unstandardized coefficients, and you get the same results.

16

## Standardized Coefficients

If  $\beta_j$  is the *unstandardized* coefficient for  $x_j$ , the *standardized* coefficient is  $\beta_j^* = \beta_j s_{x_j} / s_y$ .

Interpretation: for each 1-standard deviation increase in  $x_j$ ,  $y$  increases by  $\beta_j^*$  standard deviations.

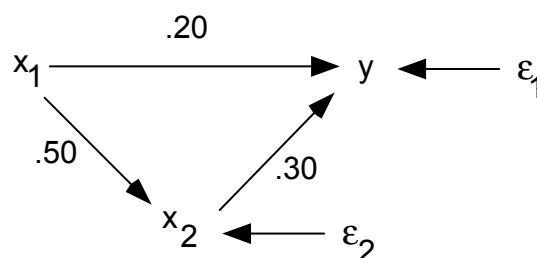
Standardized coefficients are appealing because they are *metric free*.

Facts:

- In a bivariate regression, the standardized coefficient is the same as the correlation coefficient.
- Standardized coefficients are usually between -1 and +1, but occasionally may be outside those bounds.
- The rank order of the standardized coefficients is usually the same as that of the t-statistics (but not always).

17

## Numerical Examples



Coefficients are standardized.

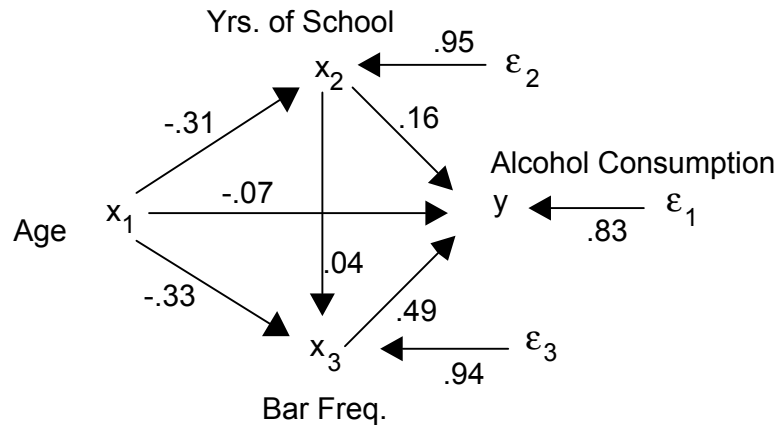
Total effect of  $x_1$  is  $.20 + .50 (.30) = .35$ .

The percentage of the total effect that is direct is

$$.20 / .35 = 57\%$$

18

## More Complex Example



The numbers on the arrows joining measured variables are standardized coefficients obtained by using OLS for each equation.

The numbers on the disturbance paths are found by the formula  $\sqrt{1 - R_k^2}$  where  $R_k^2$  is the R-squared for that particular equation.

19

## Decomposition of Effects

### Effect of Age on Drinking

	-.07	Direct	24%
-.33 (.49)	-.16	Indirect through bars	55%
-.31 (.16)	-.05	Indirect thru schooling	17%
-.31 (.04)(.49)	-.01	Thru schooling & bars	3%
Total Effect	-.29		99%

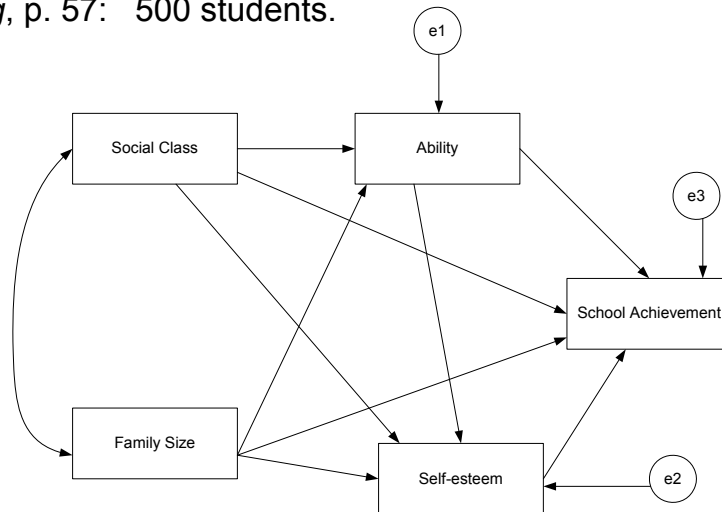
### Effect of Schooling on Drinking

Direct: .16	89%
Indirect: (.04)(.49) = .02	11%

20

# Examples Using Mplus

Example 1: Taken from Maruyama (1998) *Basics of Structural Equation Modeling*, p. 57: 500 students.



- Convention: Boxes are observed (manifest) variables and circles (or ellipses) are unobserved (latent) variables.
- This is a just-identified model: All possible relationships are present.

21

## Example 1. Data

Here's the correlation matrix for these variables:					
	Class	FamSize	Ability	Esteem	Achieve
Class	1.00				
FamSize	-.33	1.00			
Ability	.39	-.33	1.00		
Esteem	.14	-.14	.19	1.00	
Achieve	.43	-.28	.67	.22	1.00

This correlation matrix is contained in a text file called maruyama.txt. The contents look like this

```
1.00
-.33 1.00
.39 -.33 1.00
.14 -.14 .19 1.00
.43 -.28 .67 .22 1.00
```

Now we'll read the file into Mplus and specify the model:

22



# Mplus Program

## DATA:

```
FILE IS c:\data\maruyama.txt;  
TYPE IS CORR;  
NOBS IS 500;
```

## VARIABLE:

```
NAMES ARE class famsize ability esteem achieve;
```

## MODEL:

```
achieve ON esteem ability class famsize ;  
esteem ON ability class famsize;  
ability ON class famsize;
```

*My convention:* All upper case words are Mplus key words; all lower case words are variable names, parameter names, or data set names that you choose. (Mplus is not case sensitive).

23

# Selected Output

## TESTS OF MODEL FIT

### Chi-Square Test of Model Fit

Value	0.000
Degrees of Freedom	0
P-Value	0.0000

### Chi-Square Test of Model Fit for the Baseline Model

Value	469.444
Degrees of Freedom	9
P-Value	0.0000

### Loglikelihood

H0 Value	-3281.297
H1 Value	-3281.297

### Information Criteria

Number of Free Parameters	12
Akaike (AIC)	6586.594
Bayesian (BIC)	6637.169
Sample-Size Adjusted BIC	6599.080
(n* = (n + 2) / 24)	

24

## Output (cont.)

### MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ACHIEVE ON				
ESTEEM	0.082	0.033	2.505	0.012
ABILITY	0.575	0.036	15.993	0.000
CLASS	0.189	0.036	5.281	0.000
FAMSIZE	-0.016	0.035	-0.469	0.639
ESTEEM ON				
ABILITY	0.142	0.049	2.918	0.004
CLASS	0.060	0.049	1.238	0.216
FAMSIZE	-0.073	0.048	-1.537	0.124
ABILITY ON				
CLASS	0.315	0.042	7.433	0.000
FAMSIZE	-0.226	0.042	-5.323	0.000
Residual Variances				
ABILITY	0.801	0.051	15.811	0.000
ESTEEM	0.952	0.060	15.811	0.000
ACHIEVE	0.510	0.032	15.811	0.000

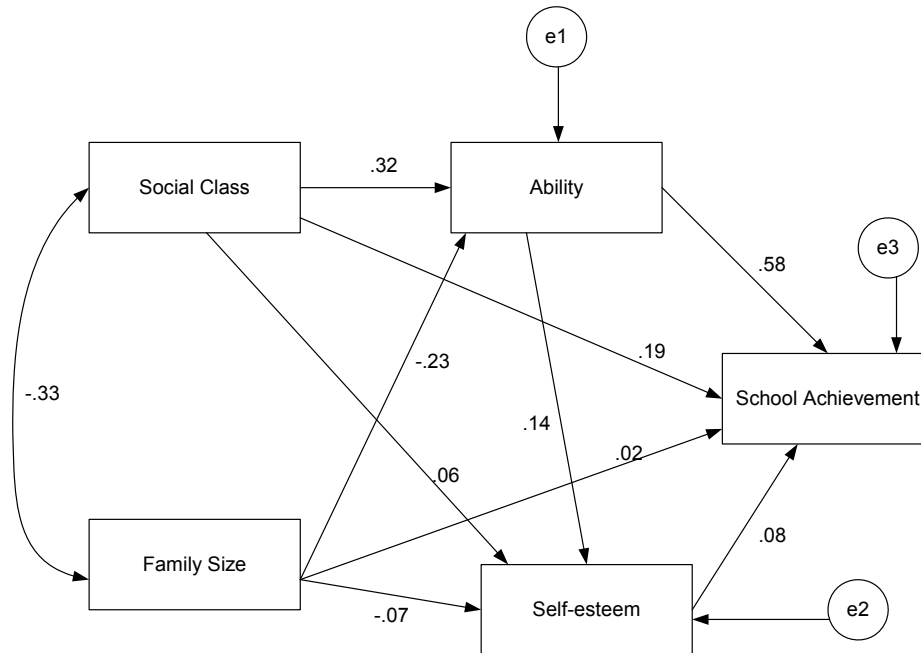
25

## SAS Code for Example 1

```
DATA maruyama(TYPE=CORR);  
  INPUT class famsize ability esteem achieve;  
  DATALINES;  
1.00 . . . .  
-.33 1.00 . . .  
.39 -.33 1.00 . .  
.14 -.14 .19 1.00 .  
.43 -.28 .67 .22 1.00  
PROC CALIS DATA=maruyama NOBS=500;  
PATH  
  achieve <- esteem ability class famsize,  
  esteem <- ability class famsize,  
  ability <- class famsize;  
RUN;
```

26

## Path Diagram with Standardized Coefficients



Results are the same as doing OLS for each dependent variable.

27

## Indirect Effects

Mplus can also calculate direct and indirect effects for specified variables. For example, to get all the direct and indirect effects of FAMSIZE on ACHIEVE:

MODEL:

```
achieve ON esteem ability class
famsize ;
esteem ON ability class famsize;
ability ON class famsize;
```

MODEL INDIRECT: achieve IND famsize;

28

## Indirect Results

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from FAMSIZE to ACHIEVE				
Total	-0.155	0.042	-3.672	0.000
Total indirect	-0.139	0.027	-5.218	0.000
Specific indirect				
ACHIEVE				
ABILITY				
FAMSIZE	-0.130	0.026	-5.051	0.000
ACHIEVE				
ESTEEM				
FAMSIZE	-0.006	0.005	-1.310	0.190
ACHIEVE				
ESTEEM				
ABILITY				
FAMSIZE	-0.003	0.001	-1.790	0.073
Direct				
ACHIEVE				
FAMSIZE	-0.016	0.035	-0.469	0.639

29

## Indirect Effects in SAS

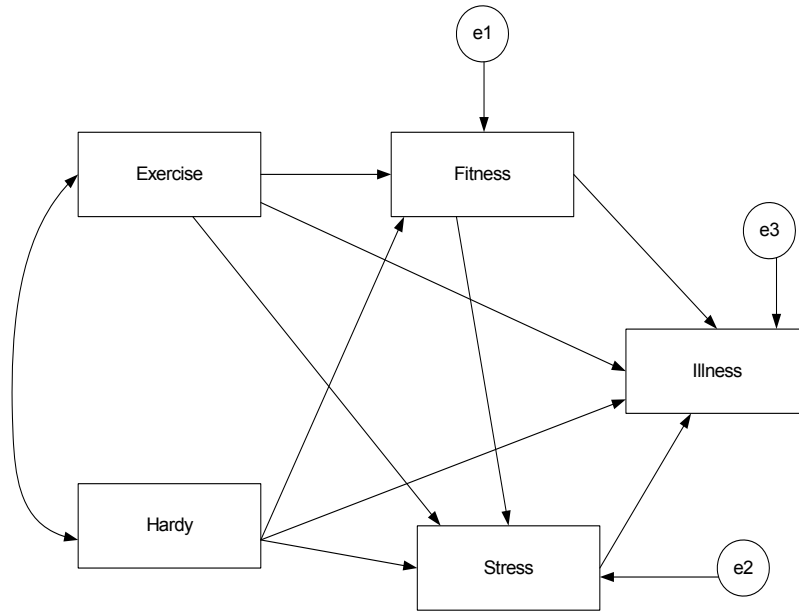
Get total and indirect effects with EFFPART option on the PROC statement:

Standardized Indirect Effects				
Effect / Std Error / t Value / p Value				
	ability	esteem	class	famsize
ability	0	0	0	0
achieve	0.0117 0.006118 1.9067 0.0566	0	0.1901 0.0259 7.3428 <.0001	-0.1386 0.0261 -5.3195 <.0001
esteem	0	0	0.0449 0.0164 2.7342 0.006253	-0.0322 0.0125 -2.5740 0.0101

Specific indirect effects can be gotten with the EFFPART *statement*.

30

## Example 2: From Rex Kline (2010)



Sample: 373 university students

31

## Sample Moment Matrix

For these data, we have means, standard deviations, and the correlation matrix in a data set called **illness.txt**. Here's what's in the data set:

```
40.9 0.0 67.1 4.8 716.7
66.5 3.8 18.4 6.7 624.8
1.00
-.03 1.00
.39 .07 1.00
-.05 -.23 -.13 1.00
-.08 -.16 -.29 .34 1.00
```

The means are in the first row, the standard deviations in the second, and the correlation matrix follows.

32

## Example 2: Mplus Program

DATA:

FILE IS c:\data\illness.txt;

TYPE IS CORR MEANS STD;

NOBS IS 373;

VARIABLE:

NAMES ARE exercise hardy fitness stress illness;

MODEL:

illness ON fitness exercise hardy stress ;

stress ON fitness exercise hardy;

fitness ON exercise hardy;

OUTPUT: STDYX;

By default, if Mplus is given the both correlations and standard deviations, it constructs and analyzes the **covariance matrix**, and reports **unstandardized** coefficients. But I've also requested the standardized coefficients, with the STDYX option.

Because this is a "recursive" model, ML estimates are identical to OLS estimates for each equation separately.

33

## Covariance Matrix

The sample covariance is defined as

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

It's related to the sample correlation by  $s_{xy} = r_{xy}s_x s_y$  where  $s_x$  and  $s_y$  are the standard deviations.

The sample covariance matrix has variances on the main diagonal and covariances off the diagonal.

Covariances always have the same sign as the corresponding correlations. They are hard to interpret because they depend on the units of measurement for each variable. But they are intermediate building blocks for many other statistics.

34

## Example 2: Results

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ILLNESS ON				
FITNESS	-8.835	1.744	-5.067	0.000
EXERCISE	0.318	0.479	0.663	0.507
HARDY	-12.146	7.939	-1.530	0.126
STRESS	27.125	4.521	6.000	0.000
STRESS ON				
FITNESS	-0.040	0.020	-1.993	0.046
EXERCISE	-0.001	0.005	-0.261	0.794
HARDY	-0.393	0.089	-4.427	0.000
FITNESS ON				
EXERCISE	0.109	0.013	8.249	0.000
HARDY	0.396	0.230	1.719	0.086
Intercepts				
FITNESS	62.659	1.026	61.048	0.000
STRESS	7.518	1.307	5.752	0.000
ILLNESS	1166.345	118.939	9.806	0.000

35

## Standardized Results

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ILLNESS ON				
FITNESS	-0.260	0.050	-5.204	0.000
EXERCISE	0.034	0.051	0.664	0.507
HARDY	-0.074	0.048	-1.535	0.125
STRESS	0.291	0.047	6.238	0.000
STRESS ON				
FITNESS	-0.109	0.054	-2.005	0.045
EXERCISE	-0.014	0.054	-0.261	0.794
HARDY	-0.223	0.049	-4.545	0.000
FITNESS ON				
EXERCISE	0.392	0.044	8.973	0.000
HARDY	0.082	0.047	1.726	0.084
R-SQUARE				
Observed				
Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FITNESS	0.159	0.035	4.574	0.000
STRESS	0.066	0.025	2.659	0.008
ILLNESS	0.183	0.036	5.065	0.000

36

## SAS Code for Example 2

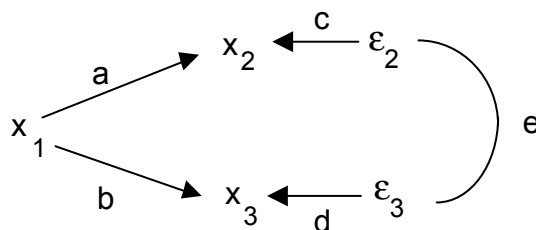
```
DATA illness (TYPE=CORR);  
  INPUT _TYPE_ $ exercise hardy fitness  
        stress illness;  
  DATALINES;  
  CORR  1.00 . . . .  
  CORR  -.03 1.00 . . .  
  CORR   .39 .07 1.00 . .  
  CORR  -.05 -.23 -.13 1.00 .  
  CORR  -.08 -.16 -.29 .34 1.00  
  STD   66.5 3.8 18.4 6.7 624.8  
  MEAN  40.9 0.0 67.1 4.8 716.7  
PROC CALIS DATA=illness NOBS=373 COV;  
PATH  
  illness <- fitness exercise hardy stress,  
  stress <- fitness exercise hardy,  
  fitness <- exercise hardy;  
RUN;
```

37

## Partial Correlations

If we have two exogenous variables, and we want to allow for a noncausal association between them, we join them with a curved line to represent a correlation.

But what about two endogenous variables? We don't allow a curved line between them. But we *can* put a curved line between their disturbance terms.



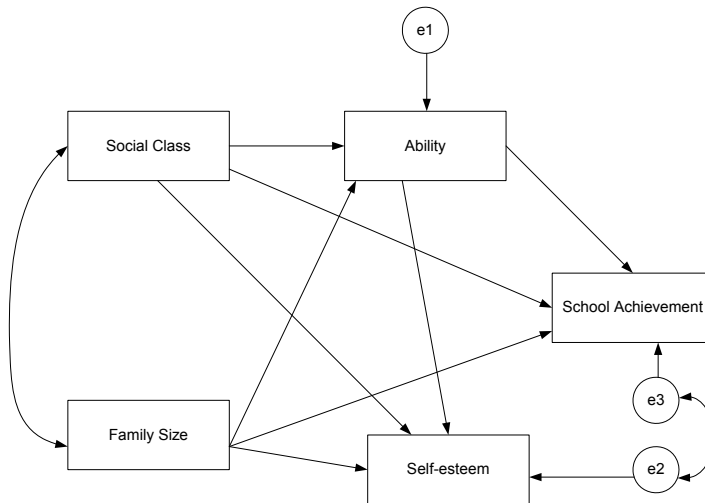
This model says that the unobserved causes of  $x_2$  and  $x_3$  are correlated. The quantity  $e$  is the **partial correlation** between  $x_2$  and  $x_3$ , controlling for  $x_1$ . Note that we cannot also have a direct effect of  $x_2$  on  $x_3$ , or vice versa.

38



## Partial Correlations (cont.)

Partial correlations can also be embedded in larger models.



This would be appropriate if you are unwilling to specify a causal direction between school achievement and self-esteem.

39

## Partial Correlations in Mplus

```
DATA: FILE IS c:\data\maruyama.txt;
      TYPE IS CORR;
      NOBS IS 500;
      VARIABLE: NAMES ARE class famsize
                  ability esteem achieve;
      MODEL:
        achieve ON ability class famsize ;
        esteem ON ability class famsize;
        ability ON class famsize;
        achieve WITH esteem;
      OUTPUT: STDYX;
```

The last line in the MODEL command says to allow a correlation between the residuals of **achieve** and **esteem**.

40

## Results with Partial Correlation

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	
ACHIEVE ON					
ABILITY	0.587	0.036	16.353	0.000	
CLASS	0.194	0.036	5.395	0.000	
FAMSIZE	-0.022	0.035	-0.639	0.523	
ESTEEM ON					
ABILITY	0.142	0.049	2.917	0.004	
CLASS	0.060	0.049	1.238	0.216	
FAMSIZE	-0.073	0.048	-1.537	0.124	
ABILITY ON					
CLASS	0.315	0.042	7.433	0.000	
FAMSIZE	-0.226	0.042	-5.323	0.000	
ACHIEVE WITH ESTEEM	0.078	0.032	2.474	0.013	(covariance)
Residual Variances					
ABILITY	0.801	0.051	15.811	0.000	
ESTEEM	0.952	0.060	15.811	0.000	
ACHIEVE	0.516	0.033	15.811	0.000	

41

## Results (cont.)

### STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	
ACHIEVE ON					
ABILITY	0.587	0.031	18.990	0.000	
CLASS	0.194	0.036	5.429	0.000	
FAMSIZE	-0.022	0.035	-0.639	0.523	
ESTEEM ON					
ABILITY	0.142	0.048	2.941	0.003	
CLASS	0.060	0.049	1.240	0.215	
FAMSIZE	-0.073	0.047	-1.540	0.123	
ABILITY ON					
CLASS	0.315	0.041	7.765	0.000	
FAMSIZE	-0.226	0.042	-5.434	0.000	
ACHIEVE WITH ESTEEM	<b>0.111</b>	<b>0.044</b>	<b>2.521</b>	<b>0.012</b>	

42

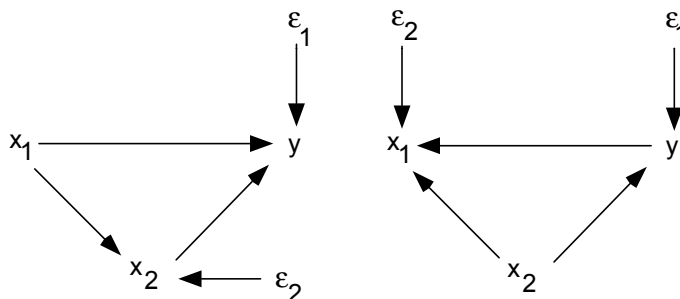
# Partial Correlations in SAS

```
PROC CALIS DATA=maruyama NOBS=500;  
PATH  
    achieve <- ability class famsize,  
    esteem <- ability class famsize,  
    ability <- class famsize,  
    esteem <-> achieve;  
RUN;
```

43

## Causal Ordering

Consider the following two path diagrams:



How do we know which one is right? Can't tell from the data.  
Have to decide the causal ordering in advance.

- Nothing new. We do this every time we estimate a regression model.
- Gets more complicated with multiple equations, however.

44

# How to Decide

Sometimes temporal ordering provides a good rationale, e.g.,  
Father's Education  $\Rightarrow$  Child's education  $\Rightarrow$  Child's income at age 30.

Sometimes theory or common sense is a good basis, e.g.

Crime rate in neighborhood  $\Rightarrow$  Fear of going out at night.

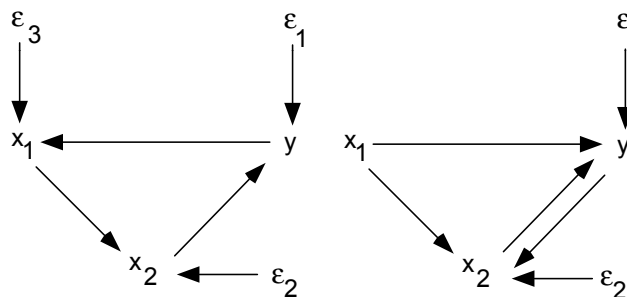
Can often make plausible, but not definitive arguments.

All the path diagrams considered so far have been *recursive* systems. In such systems, the causal ordering only goes in one direction. If you follow the arrows, you can never return to your starting point.

45

## Nonrecursive Systems

In a *nonrecursive* system, you *can* return to one or more starting points



- Recursive systems can be estimated by applying OLS regression to each equation separately.
- Nonrecursive systems require special estimation methods. Often they can't be estimated at all.
- Nonrecursive systems require special rules for tracing direct and indirect effects. See David Heise, *Causal Analysis*, 1975.

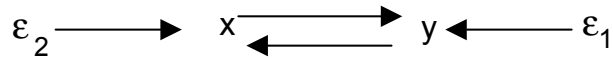
46

## Identification Problem in Nonrecursive Models

Here is the simplest possible nonrecursive model

$$y = \beta_0 + \beta_1 x + \varepsilon_1$$

$$x = \alpha_0 + \alpha_1 y + \varepsilon_2$$



Suppose we use OLS to regress  $y$  on  $x$  to for the first equation, and then regress  $x$  on  $y$  to estimate the second equation. We then have

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \hat{\alpha}_1 = \frac{s_{xy}}{s_y^2}$$

47

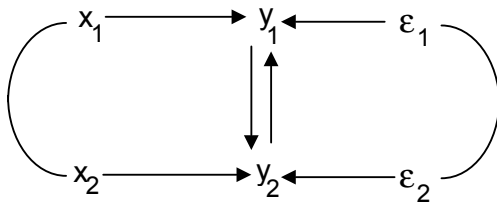
## Identification Problem (cont.)

- Since they have the same numerator and the denominator is positive, the signs must be the same, If one is zero, the other must be zero.
- This should warn us that there's something wrong with this method. For the method to be useful, we should be able to get different signs, or to have one be zero and the other be nonzero.
- The problem stems from the fact that  $x$  is necessarily correlated with  $\varepsilon_1$ , and  $y$  is necessarily correlated with  $\varepsilon_2$  (Why?).
- This situation is described by saying that the system is *underidentified*. This means that there is too little information to estimate the model. Most nonrecursive models are underidentified, either in whole or in part.

48

## A Just-Identified Model

Here's a nonrecursive model that *is* identified



$$y_1 = \beta_1 x_1 + \beta_2 y_2 + \varepsilon_1$$

$$y_2 = \alpha_1 y_1 + \alpha_2 x_2 + \varepsilon_2$$

These equations are called the *structural form* of the system. They cannot be directly estimated by ordinary least squares because  $y_2$  is necessarily correlated with  $\varepsilon_1$ , and  $y_1$  is necessarily correlated with  $\varepsilon_2$ .

We can solve these equations so that the endogenous variables appear only on the left hand side, and the exogenous variables appear only on the right-hand side. This is called the *reduced form* of the system:

49

## Reduced Form Equations

$$y_1 = \left( \frac{\beta_1}{1 - \alpha_1 \beta_2} \right) x_1 + \left( \frac{\alpha_2 \beta_2}{1 - \alpha_1 \beta_2} \right) x_2 + \left( \frac{1}{1 - \alpha_1 \beta_2} \right) \varepsilon_1 + \left( \frac{\beta_2}{1 - \alpha_1 \beta_2} \right) \varepsilon_2$$

$$y_2 = \left( \frac{\alpha_1 \beta_1}{1 - \alpha_1 \beta_2} \right) x_1 + \left( \frac{\alpha_2}{1 - \alpha_1 \beta_2} \right) x_2 + \left( \frac{\alpha_1}{1 - \alpha_1 \beta_2} \right) \varepsilon_1 + \left( \frac{1}{1 - \alpha_1 \beta_2} \right) \varepsilon_2$$

We can express these equations more simply as

$$y_1 = \gamma_{11} x_1 + \gamma_{12} x_2 + \varepsilon_1^*$$

$$y_2 = \gamma_{21} x_1 + \gamma_{22} x_2 + \varepsilon_2^*$$

where the  $\gamma$ 's correspond to the expressions in parentheses. There are four  $\gamma$ 's and four  $\alpha$ 's and  $\beta$ 's. They are functions of each other. If we knew the  $\gamma$ 's, we could solve for the  $\alpha$ 's and  $\beta$ 's.

50

## Solutions for Structural Parameters

$$\alpha_1 = \frac{\gamma_{21}}{\gamma_{11}} \quad \beta_2 = \frac{\gamma_{12}}{\gamma_{22}} \quad \beta_1 = \gamma_{11} \left( 1 - \frac{\gamma_{21}\gamma_{12}}{\gamma_{11}\gamma_{22}} \right) \quad \alpha_2 = \gamma_{22} \left( 1 - \frac{\gamma_{21}\gamma_{12}}{\gamma_{11}\gamma_{22}} \right)$$

We don't know the  $\gamma$ 's, but we can estimate them by applying OLS to the reduced form equations. The reduced form equations *do* satisfy the usual assumptions of the linear regression model.

Once we get estimates of the  $\gamma$ 's, we can substitute them into the formulas above to get estimates of the  $\alpha$ 's and  $\beta$ 's. These estimates are *approximately* unbiased. This method is called *indirect least squares*.

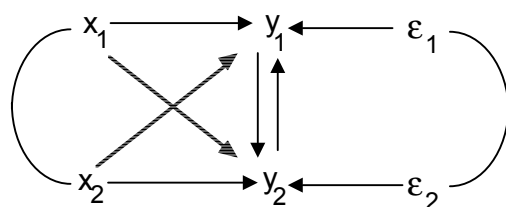
The identification problem can now be described as follows: Can the structural coefficients be obtained from the reduced form coefficients?

- If yes, then the system is identified.
- If no, then the system is underidentified.
- For identification, there must be at least as many reduced form coefficients as structural coefficients. A necessary but not sufficient condition.

51

## Sufficient Condition for Identification

What made this model identified? To answer, consider what would make the model underidentified. Note that two possible paths were excluded:



If these paths are included, the model becomes underidentified. There are 6 structural coefficients but only 4 reduced form coefficients. This suggests a *sufficient* condition for identification:

A nonrecursive model is identified if each endogenous variable in the feedback loop has at least one exogenous variable that affects it, but not the other endogenous variables. These exogenous variables are called "instrumental variables."

52

## Varieties of Identification

1. Just identified – the model is identified, and the number of reduced form coefficients is the same as the number of structural coefficients.
2. Partially identified – some structural coefficients can be estimated, but not others.
3. Over identified – the model is identified, and there are *more* reduced form coefficients than structural coefficients. This leads to multiple solutions for the structural coefficients—an embarrassment of riches.

Need some procedure to reconcile and combine these estimates:

- 2-stage least squares
- 3-stage least squares
- maximum likelihood.

We can do ML with Mplus, or any other SEM packages.

53

## Problems with Instrumental Variables

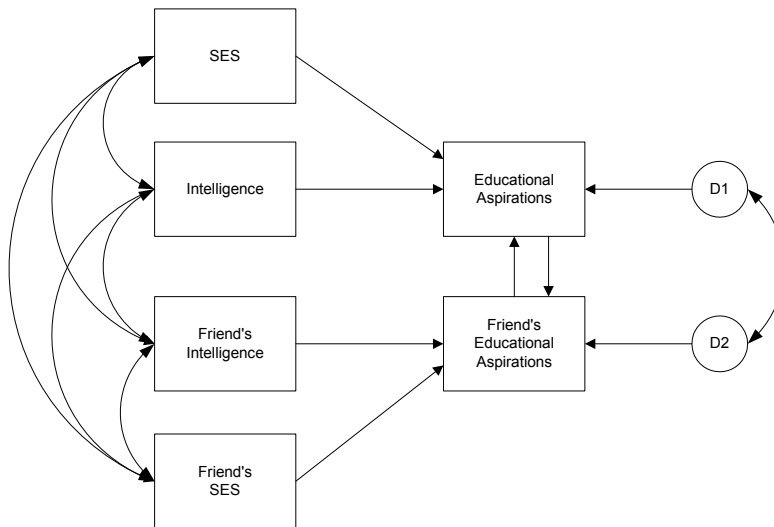
1. Must assume equilibrium.
2. To get instrumental variables, you must be able to exclude paths on theoretical grounds.
  - Rarely able to do this with any confidence in the social sciences.
  - No amount of data can tell you whether the paths can be excluded.
  - People frequently ignore this problem, but results can be very sensitive to the choice of instruments.
3. If instrumental variables are “weak” (have low correlations with their endogenous variables) standard errors will be high.

54



## Example 3. Nonrecursive Model with Mplus

Career aspirations of 329 seventeen year-old boys (Duncan, Haller and Portes, *American Journal of Sociology*, 1968). Each boy named his best friend, who was also interviewed.



55

## Example 3 (cont.)

This is an over-identified model. There are 2 fewer parameters than there are variances and covariances. What are the over-identifying restrictions? Equivalently, what additional parameters could be added without making the model underidentified?

For these data, we only have correlations and the sample size. Here are the correlations, contained in a data set called **aspirations.txt**:

```
1.00
.22 1.00
.40 .40 1.00
.34 .23 .29 1.00
.19 .27 .24 .30 1.00
.29 .31 .37 .52 .41 1.00
```

56

## Example 3. Mplus Program

```
DATA:
  FILE=c:\data\aspirations.txt;
  TYPE=CORR;
  NOBS=329;
VARIABLE:
  NAMES=intell ses edasp fintell
    fses fedasp;
MODEL:
  edasp ON fedasp intell ses;
  fedasp ON edasp fintell fses;
  edasp WITH fedasp;
OUTPUT: STDYX;
```

57

## Example 3. Results

### TESTS OF MODEL FIT

#### Chi-Square Test of Model Fit

Value	0.566
Degrees of Freedom	2
P-Value	0.7536

#### Chi-Square Test of Model Fit for the Baseline Model

Value	267.519
Degrees of Freedom	9
P-Value	0.0000

#### Loglikelihood

H0 Value	-2607.207
H1 Value	-2606.925

#### Information Criteria

Number of Free Parameters	9
Akaike (AIC)	5232.415
Bayesian (BIC)	5266.579
Sample-Size Adjusted BIC	5238.031
$(n^* = (n + 2) / 24)$	

58

## Example 3. Results (cont.)

### MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
EDASP	ON				
	FEDASP	0.282	0.099	2.843	0.004
	INTELL	0.249	0.051	4.921	0.000
	SES	0.268	0.050	5.324	0.000
FEDASP	ON				
	EDASP	0.383	0.108	3.538	0.000
	FINTELL	0.342	0.051	6.666	0.000
	FSES	0.218	0.046	4.769	0.000

*Because we used a correlation matrix, the above are all standardized coefficients.*

### EDASP WITH

FEDASP	-0.330	0.111	-2.963	0.003
--------	--------	-------	--------	-------

*But this is not a partial correlation. The correlation (from the standardized output) is  $-.487$ .*

59

## SAS Code for Nonrecursive Model

```
DATA nonrec(TYPE=CORR);
  INPUT intell ses edasp fintell fses fedasp;
DATALINES;
1.00 . . . . .
.22 1.00 . . . .
.40 .40 1.00 . . .
.34 .23 .29 1.00 . .
.19 .27 .24 .30 1.00 .
.29 .31 .37 .52 .41 1.00
PROC CALIS DATA=nonrec NOBS=329;
  PATH
    edasp <- fedasp intell ses,
    fedasp <- edasp fintell fses,
    edasp <-> fedasp;
RUN;
```

60

# Classical Test Theory

Developed by psychometricians beginning around 1930 to formalize notions of reliability, validity and measurement error.

Classic statement:

Lord and Novick (1968) *Statistical Theories of Mental Test Scores*.

Some notation:

$\rho_{XY}$  population correlation between X and Y

$r_{XY}$  sample correlation between X and Y

Population covariance:

$$\text{cov}(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

Sample covariance: 
$$s_{XY} = \frac{1}{N} \sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Population correlation: 
$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

61

## Random Measurement Error

We now formalize notions of random measurement error

T = true score

X = observed score

e = measurement error

Assume:  $X = T + e$

$$E(X|T) = T$$

This implies a lot of things:

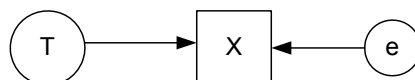
$$E(e|T) = E(e) = 0$$

$$E(X) = E(T)$$

$$\text{cov}(T, e) = 0 \quad (\text{no correlation between the true score and the error})$$

$$\text{var}(X) = \text{var}(T) + \text{var}(e)$$

The model can be represented by a path diagram:



62

# Reliability

**Reliability of X** — squared correlation between true score T and observed score X. Equivalently, the  $R^2$  for regressing X on T.

$$\rho_{TX}^2 = \left[ \frac{\text{cov}(T, X)}{\sqrt{\text{var}(T) \text{var}(X)}} \right]^2$$

Based on our assumptions about T and X, it can be shown that

$$\begin{aligned} \rho_{TX}^2 &= \frac{\text{var}(T)}{\text{var}(X)} = \frac{\text{true score variance}}{\text{observed score variance}} \\ &= \frac{\text{var}(T)}{\text{var}(T) + \text{var}(e)} \end{aligned}$$

But since we don't observe T, how can we ever estimate the reliability of X?

63

## Parallel Measures

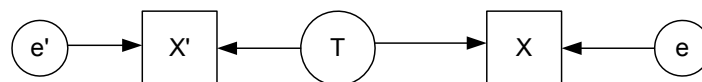
Introduce a **parallel** measure  $X'$ , satisfying the following assumptions:

$$X' = T + e'$$

$$E(X'|T) = T$$

$$\text{cov}(e', e) = 0 \quad (\text{errors are uncorrelated})$$

$$\text{var}(e') = \text{var}(e) \quad (\text{error variances are equal})$$



These assumptions imply that  $\rho_{TX}^2 = \rho_{TX'}^2$

Furthermore, it can be proved that

$$\rho_{XX'} = \rho_{XT}^2$$

Thus, reliability = the correlation between two parallel measures.

64

# Tau-Equivalent Measures

But parallel measures are hard to come by. Consider the weaker condition of **tau-equivalent** measures:

$$X = T + e$$

$$X' = T + e'$$

$$E(X|T) = T$$

$$E(X'|T) = T$$

$$\text{Cov}(e', e) = 0$$

$$\text{var}(e') \neq \text{var}(e)$$

These assumptions imply that  $\text{var}(X) \neq \text{var}(X')$  and  $\rho_{TX}^2 \neq \rho_{TX'}^2$

But we can still calculate each reliability based on observed data:

$$\rho_{XT}^2 = \rho_{XX'} \frac{\sigma_{X'}}{\sigma_X} \text{ and } \rho_{X'T}^2 = \rho_{XX'} \frac{\sigma_X}{\sigma_{X'}}$$

65

## Tau-Equivalence: Example

In the morning, 250 people are weighed on a bathroom scale ( $X_1$ ). The next morning, they are weighed on a doctor's scale ( $X_2$ ).

The correlation between the two measurements is .85. The standard deviation of  $X_1$  is 25. The standard deviation of  $X_2$  is 22.

Assuming that the two measures are tau-equivalent, the reliabilities are

$$\rho_{X_1T}^2 = .85 \times \frac{22}{25} = .75$$

$$\rho_{X_2T}^2 = .85 \times \frac{25}{22} = .97$$

In Mplus:

The data set looks like this:

25	22
1.0	
.85	1.0

66

## Tau-Equivalence in Mplus

DATA:

```
FILE IS c:\data\tauequiv.txt;
TYPE IS CORR STD;
NOBS IS 250;
VARIABLE: NAMES ARE bath doctor;
MODEL: truewt BY bath@1 doctor@1;
OUTPUT: STDYX;
```

“t BY x1 x2” says that the latent variable t is measured by x1 and x2.  
@1 constrains the coefficients to be 1.

Here is the key section of the output:

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
BATH	0.748	0.029	25.513	0.000
DOCTOR	0.966	0.038	25.513	0.000

67

## Tau-Equivalence in SAS

```
DATA tau(TYPE=COV);
  INPUT _TYPE_ $ bath doctor;
  DATALINES;
    CORR 1.0 .
    CORR .85 1.0
    STD 25 22
  ;
PROC CALIS DATA=tau COV NOBS=250;
  PATH
    bath doctor <- truewt = 1 1;
RUN;
```

Two 1's after the = sign constrain the coefficients to be 1.

68

## Congeneric Tests

$$X = T + e$$

$$X' = aT + e'$$

$$E(X|T) = T$$

$$E(X'|T) = aT$$

$$\text{Cov}(e', e) = 0$$

$$\text{Var}(e') \neq \text{var}(e)$$

$X$  and  $X'$  are both measures of  $T$ , but they have different units of measurement. Can't calculate reliability with only two congeneric tests. (The model is underidentified).

69

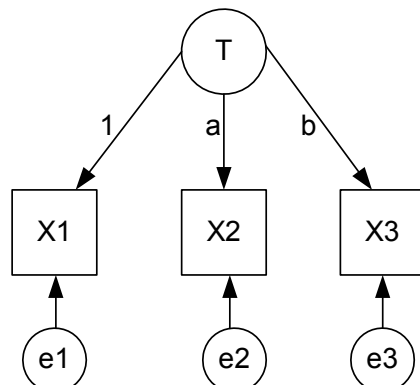
## Three Congeneric Tests

$$X_1 = T + e_1$$

$$X_2 = aT + e_2$$

$$X_3 = bT + e_3$$

$T$  is uncorrelated with all the  $e$ 's. The  $e$ 's are uncorrelated with each other.



70



## Three Congeneric Measures (cont.)

It can be shown that

$$\text{cov}(X_1, X_2) = \sigma_{12} = a \text{ var}(T)$$

$$\text{cov}(X_1, X_3) = \sigma_{13} = b \text{ var}(T)$$

$$\text{cov}(X_2, X_3) = \sigma_{23} = ab \text{ var}(T)$$

We have 3 equations and 3 unknowns. Solutions:

$$\text{var}(T) = \frac{\sigma_{13}\sigma_{12}}{\sigma_{23}}$$

$$a = \frac{\sigma_{23}}{\sigma_{13}}$$

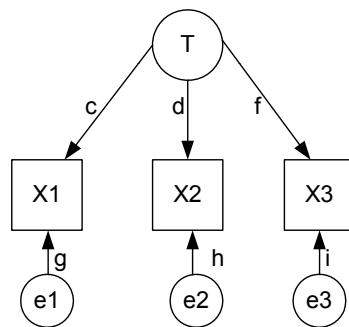
$$b = \frac{\sigma_{23}}{\sigma_{12}}$$

There are also solutions for the variances of  $e_1$ ,  $e_2$  and  $e_3$ .

71

## Standardized Version

Now let's redo the whole thing in terms of a standardized model, i.e., all variables are assumed to have a standard deviation of 1.



The equations are

$$X_1 = cT + ge_1$$

$$X_2 = dT + he_2$$

$$X_3 = fT + ie_3$$

$c$ ,  $d$  and  $f$  are just bivariate correlations. Their squares are the reliabilities.

72

## Digression: Tracing Rule for Correlations

As we will see later, it's often very useful to be able to express the correlation between two variables as functions of the parameters in a standardized path diagram.

Here is a rule for doing that in recursive models (Loehlin, *Latent Variable Models*, 1987):

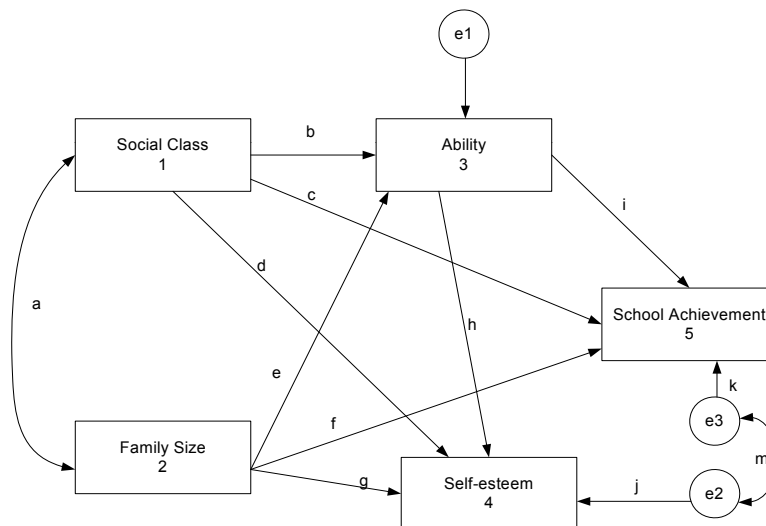
“The correlation between any two variables in the diagram can be expressed as the *sum of the compound paths connecting these two points*, where a compound path is a path along arrows that follows these rules:

- no loops;
- no going forward, then backward
- a maximum of one curved arrow per path.”

Let's apply this to the diagram on the next slide:

73

### Tracing Rule (cont.)



All the letters represent standardized coefficients or correlations. We now express each correlation between two observed variables as a function of those parameters.

74

## Tracing Rule (cont.)

$$\rho_{12} = a$$

$$\rho_{13} = b + ea$$

$$\rho_{23} = e + ab$$

$$\rho_{14} = d + ag + bh + aeh$$

$$\rho_{24} = g + ad + eh + abh$$

$$\rho_{34} = h + db + ge + gab + dae$$

$$\rho_{15} = c + bi + af + iea$$

$$\rho_{25} = f + ie + ac + iba$$

$$\rho_{35} = i + bc + fe + cae + fab$$

$$\rho_{45} = fg + cd + ih + ibd + ieg + ibag + cag + cbh + feh + jkm$$

These equations can be solved for the unknown parameters.

75

## Standardized Version (cont.)

By the tracing rules of path analysis

$$\rho_{12} = cd$$

$$\rho_{13} = cf$$

$$\rho_{23} = df$$

Again we have three equations and three unknowns. Solutions:

$$c^2 = \rho_{12}\rho_{13}/\rho_{23}$$

$$d^2 = \rho_{12}\rho_{23}/\rho_{13}$$

$$f^2 = \rho_{13}\rho_{23}/\rho_{12}$$

All these formulas are for the population parameters. But for sample data, we can substitute the corresponding sample correlations.

When the model is just-identified (same number of equations and unknowns), the resulting estimates will be maximum likelihood estimates.

76

## Three Congenerics: Example

Suppose T is “intelligence” and each person (in a sample of 200) takes three different IQ tests. The correlations are  $r_{12} = .78$ ,  $r_{13} = .84$ ,  $r_{23} = .80$ .

The three reliabilities are

$$(.78)(.84)/.80 = .82$$

$$(.78)(.80)/.84 = .74$$

$$(.84)(.80)/.78 = .86$$

Now let's do it in Mplus. The data set is:

```
1.0
.78 1.0
.84 .80 1.0
```

77

## Three Congenerics: Mplus

```
DATA:
  FILE IS c:\data\intelligence.txt;
  TYPE IS CORR;
  NOBS IS 200;
VARIABLE: NAMES ARE x1 x2 x3;
MODEL: intell BY x1 x2 x3;
OUTPUT: STDYX;
```

### R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
X1	0.819	0.033	24.703	0.000
X2	0.743	0.038	19.541	0.000
X3	0.862	0.031	27.913	0.000

78

# Three Congenerics: SAS

```
DATA intelligence (TYPE=CORR);
  INPUT x1 x2 x3;
  DATALINES;
1.0 . .
.78 1.0 .
.84 .80 1.0
;
PROC CALIS DATA=intelligence NOBS=200;
PATH
  intell -> x1 x2 x3 = 1;
RUN;
```

The 1 after the = sign constrains the coefficient for the effect of INTELL on X1 to be 1. Necessary for identification.

79

## Four Congeneric Measures

By tracing rules

$$\rho_{12} = ab$$

$$\rho_{13} = ac$$

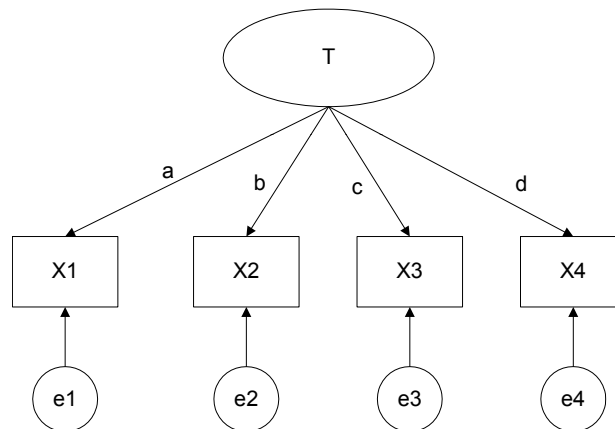
$$\rho_{23} = bc$$

-----

$$\rho_{14} = ad$$

$$\rho_{24} = bd$$

$$\rho_{34} = cd$$



Using only the first three equations, we get the same solutions as for the three-variable case. But we have three more equations, and only one more parameter. For each parameter, there are three alternative solutions, e.g.:

$$a = \sqrt{\frac{\rho_{12}\rho_{13}}{\rho_{23}}} = \sqrt{\frac{\rho_{12}\rho_{14}}{\rho_{24}}} = \sqrt{\frac{\rho_{13}\rho_{14}}{\rho_{34}}}$$

80

## Overidentification with 4 Congeneric Measures

These solutions may be mutually inconsistent

4 parameters, 6 correlations

implies 2 over-identifying restrictions.

The model implies two restrictions on the correlations:

$$\rho_{12}\rho_{34} = \rho_{13}\rho_{24} = \rho_{14}\rho_{23}$$

In any sample of real data, these equations won't exactly hold.

Why?

- Model may be wrong
- Sampling error

How to distinguish between these two possibilities?

- Likelihood ratio test

How to get a single estimate for each parameter?

- Take a simple, unweighted average.
- ML methods are much better.

81

## Four Congeneric Measures with Mplus

```
DATA: FILE IS c:\data\illness.txt;
      TYPE IS CORR MEANS STD;
      NOBS IS 373;
VARIABLE:
      NAMES = exercise hardy fitness stress illness;
      USEVARIABLES = hardy fitness stress illness;
MODEL:
      latvar BY illness hardy fitness stress ;
OUTPUT: STDYX;
```

82

## Four Congeneric Measures with SAS

```
DATA illness (TYPE=CORR);
  INPUT _NAME_ $ _TYPE_ $ exercise hardy fitness stress
    illness;
  DATALINES;
exercise CORR 1.00 . . . .
hardy CORR -.03 1.00 . . .
fitness CORR .39 .07 1.00 . .
stress CORR -.05 -.23 -.13 1.00 .
illness CORR -.08 -.16 -.29 .34 1.00
. STD 66.5 3.8 18.4 6.7 624.8
. MEAN 40.9 0.0 67.1 4.8 716.7
PROC CALIS DATA=illness NOBS=373;
PATH
  latvar -> illness stress fitness hardy = 1;
RUN;
```

83

## Results for Four Congenerics

### Chi-Square Test of Model Fit

Value	8.227
Degrees of Freedom	2
P-Value	0.0164

### MODEL RESULTS

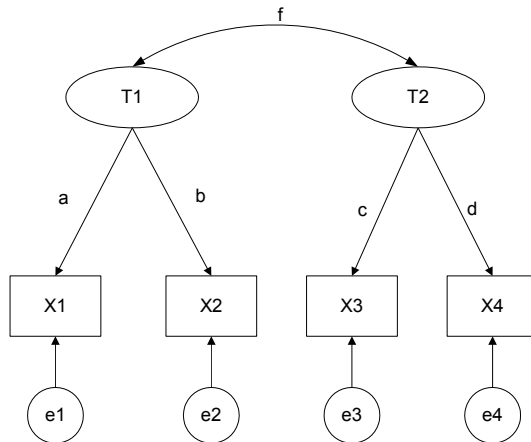
LATVAR	BY	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ILLNESS		1.000	0.000	999.000	999.000
HARDY		-0.002	0.001	-3.419	0.001
FITNESS		-0.015	0.004	-4.007	0.000
STRESS		0.007	0.002	4.233	0.000

### R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HARDY	0.076	0.036	2.121	0.034
FITNESS	0.136	0.048	2.839	0.005
STRESS	0.241	0.068	3.529	0.000
ILLNESS	0.493	0.118	4.186	0.000

84

## Alternative Model for 4 Congeneric Measures



$$\begin{aligned}\rho_{12} &= ab \\ \rho_{34} &= cd \\ \rho_{13} &= afc \\ \rho_{14} &= afd \\ \rho_{23} &= bfc \\ \rho_{24} &= bfd\end{aligned}$$

6 correlations, 5 parameters,  
1 over-identifying restriction

Some solutions:

$$a^2 = \frac{\rho_{12}\rho_{13}}{\rho_{23}} = \frac{\rho_{12}\rho_{14}}{\rho_{24}}$$

$$\begin{aligned}f^2 &= \frac{\rho_{14}\rho_{23}}{\rho_{12}\rho_{34}} = \frac{(afd)(bfc)}{(ab)(cd)} \\ &= \frac{\rho_{13}\rho_{24}}{\rho_{12}\rho_{34}}\end{aligned}$$

85

## Mplus for Alternative Model

Two solutions for every parameter.

Over-identifying restriction:  $\rho_{14}\rho_{23} - \rho_{13}\rho_{24} = 0$ .  
(This is called a vanishing tetrad).

```

DATA: FILE IS c:\data\illness.txt;
      TYPE = CORR MEANS STD; NOBS IS 373;
VARIABLE:
      NAMES = exercise hardy fitness stress illness;
      USEVARIABLES = hardy fitness stress illness;
MODEL:
      latvar1 BY illness stress;
      latvar2 BY fitness hardy;
OUTPUT: STDYX;
  
```

By default, latvar1 and latvar2 are allowed to be correlated.

86



# Results for Alternative Model

THE MODEL ESTIMATION TERMINATED NORMALLY

WARNING: THE LATENT VARIABLE COVARIANCE MATRIX (PSI) IS NOT POSITIVE DEFINITE. THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR A LATENT VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO LATENT VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO LATENT VARIABLES. CHECK THE TECH4 OUTPUT FOR MORE INFORMATION.  
PROBLEM INVOLVING VARIABLE LATVAR2.

Chi-Square Test of Model Fit

Value	7.590
Degrees of Freedom	1
P-Value	0.0059

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
LATVAR1 BY				
ILLNESS	1.000	0.000	999.000	999.000
STRESS	0.008	0.002	4.431	0.000
LATVAR2 BY				
FITNESS	1.000	0.000	999.000	999.000
HARDY	0.163	0.047	3.459	0.001
LATVAR2 WITH LATVAR1	-2952.922	598.620	-4.933	0.000

87

## Results (cont.)

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
LATVAR1 BY				
ILLNESS	0.678	0.080	8.450	0.000
STRESS	0.502	0.068	7.397	0.000
LATVAR2 BY				
FITNESS	0.298	0.114	2.620	0.009
HARDY	0.235	0.095	2.468	0.014
LATVAR2 WITH LATVAR1	-1.274	0.450	-2.834	0.005

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HARDY	0.055	0.045	1.234	0.217
FITNESS	0.089	0.068	1.310	0.190
STRESS	0.252	0.068	3.699	0.000
ILLNESS	0.459	0.109	4.225	0.000

88

# SAS Code for Alternative Model

```
PROC CALIS DATA=illness NOBS=373;  
PATH  
    latvar1 -> illness stress = 1,  
    latvar2 -> fitness hardy = 1;  
RUN;
```

89

## Heywood Case

When estimated correlations are greater than 1, or variances are less than 0, it's called a *Heywood* case. This can occur even in very simple models.

For example, suppose we have 3 observed variables with correlations .62, -.41, and -.24, and we postulate a single latent variable. Applying the formula for squared factor loadings in the three indicator case, we get

$$(.62)(-.41)/-.24 = 1.06$$

This could occur just by sampling variation, especially in small samples. But it also suggests that something fundamental is wrong with the model.

Note, however, that the chi-square test is insensitive to such problems. (E.g., the 3 indicator case is just identified and fits perfectly).

90

# Factor Models

“Factor” is another name for a latent variable. A general factor model for  $p$  observed variables may be written as

$$\begin{aligned}x_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \dots + \lambda_{1k}f_k + e_1 \\&\vdots \\x_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \dots + \lambda_{pk}f_k + e_p\end{aligned}$$

where  $k \leq p$ . We assume that

$$\begin{aligned}E(e_j) &= 0 && \text{for all } j \\ \text{cov}(f_i, e_j) &= 0 && \text{for all } i \text{ and } j \\ \text{cov}(e_i, e_j) &= 0 && \text{for all } i \neq j\end{aligned}$$

An *orthogonal* factor model assumes  $\text{cov}(f_i, f_j) = 0$  for all  $i \neq j$ .

An *oblique* model allows for correlations among the factors

91

## Factor Models (cont.)

In a factor model, all the correlations among the observed variables are due to their mutual dependence on a set of unobserved factors. Thus, factor models assume that all partial correlations among the observed variables (controlling for the factors) are 0:

$$\rho(x_i, x_j | f_1, \dots, f_k) = 0 \quad \text{for all } i \text{ and } j.$$

In matrix notation, the factor model can be expressed as

$$\underset{p \times 1}{\mathbf{x}} = \underset{p \times k}{\mathbf{\Lambda}} \underset{k \times 1}{\mathbf{f}} + \underset{p \times 1}{\mathbf{e}}$$

Assuming that  $\mathbf{f}$  and  $\mathbf{e}$  are uncorrelated, we have

$$V(\mathbf{x}) = \mathbf{\Lambda} V(\mathbf{f}) \mathbf{\Lambda}' + V(\mathbf{e})$$

$$\underset{p \times p}{\Sigma} = \underset{p \times k}{\mathbf{\Lambda}} \underset{k \times k}{\Phi} \underset{k \times p}{\mathbf{\Lambda}'} + \underset{p \times p}{\Psi}$$

In the classic version of the factor model,  $\Psi$  is a diagonal matrix (i.e., the error terms are uncorrelated with each other). In the orthogonal model  $\Phi$  is also diagonal (factors are uncorrelated).

92

## Identification (Standardized)

Are there more knowns than unknowns? There are  $p(p-1)/2$  terms in the observed correlation matrix.

Unknowns:

$pk$  factor loadings ( $\lambda$ 's)

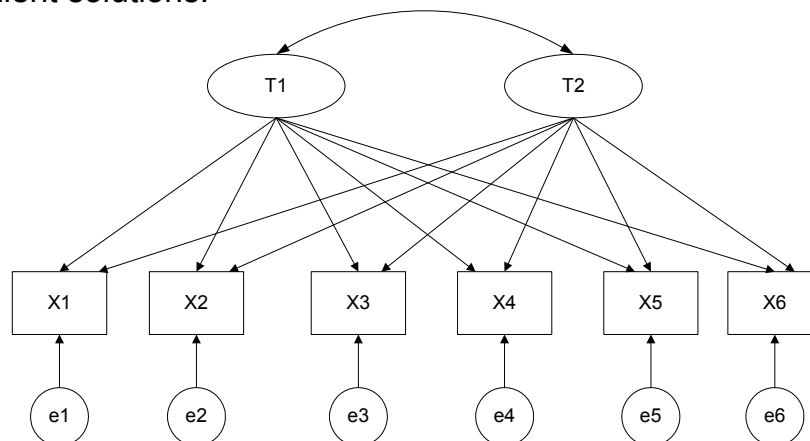
In an oblique model there are also  $k(k-1)/2$  factor correlations.

In many cases, there will be more knowns than unknowns. But the model is identified only if  $k=1$ . If there are more knowns than unknowns, there will be restrictions on the covariance matrix, but if  $k>1$  we still can't solve for the parameters.

93

## Identification (cont.)

Here's an intuitive justification. Consider the following oblique factor model. This model has more knowns than unknowns. But T1 and T2 are symmetric in the sense that they play exactly the same role in the model. Hence, there's no way to distinguish the loadings on one variable from the other. There are infinitely many different but equivalent solutions.



94

## Two Approaches to Identification Problem

*Exploratory factor analysis:* Leave the model underidentified. Then use a method called “rotation” to explore the infinite number of different possible solutions, until one is found that is conceptually appealing.

Many different methods of rotation, giving different results. Mplus offers 11 different rotation methods.

*Confirmatory factor analysis:* Impose theoretically based restrictions on the parameters until the model is identified, e.g.

Set some factor loadings equal to 0. For example, set loadings of  $x_1$  and  $x_2$  on  $T_2$  equal to 0 and set loadings of  $x_5$  and  $x_6$  on  $T_1$  equal to 0.

Constrain some parameters to be equal.

All the models we have seen for parallel, tau-equivalent, or congeneric tests were confirmatory factor models.

95

## Identification (Unstandardized)

In unstandardized form, there are more observed quantities and more parameters.

Observed quantities: Covariance matrix ( $p(p+1)/2$ ), plus  $p$  means.

Parameters:  $pk$  factor loadings,  $k(k+1)/2$  variances and covariances for the latent factors,  $p$  residual variances, and  $p$  intercepts.

The  $p$  means usually balance out the  $p$  intercepts, and the  $p$  residual variances usually balance out the  $p$  observed variances. And the  $k$  factor variances are usually balanced by  $k$  normalizing constraints (more shortly).

96

# Determining Identification

A sufficient condition for identification of confirmatory factor models:

- There are no correlated errors
- All factors are correlated
- Every factor has at least two exclusive indicators

How to determine whether a model is identified

- Solve the equations for every parameter (difficult)
- Try to estimate it and see if it works (not infallible)
- Break the model into submodels that are known to be identified.

97

## Normalizing Constraints

Because factors are not observed, they have no units of measurement. For a CFA model to have any meaning, you must arbitrarily choose a scale of measurement for the factors.

Two equivalent ways to do this:

1. Set the variances of the factors to be 1.0. Most useful when you're interested in correlations and standardized coefficients.
  - This method is usually problematic if the factors are endogenous variables. That's because the variance of an endogenous variable is not a parameter in the model.

98

## Normalizing Constraints

2. Set one factor loading for each factor equal to 1. Thus, for each latent variable, we choose one of its indicators to be a *reference* indicator. The units of measurement for the latent variable are the same as for the reference indicator.
  - This approach is most useful when you're interested in unstandardized coefficients and/or comparisons across groups.
  - Works best when you analyze a covariance matrix rather than a correlation matrix.
  - By default, Mplus sets the factor loading equal to 1 for the *first* observed variable in  
f BY x1 x2 x3

99

## ML Estimation of CFA Models

Choose as estimates, those values of the parameters that maximize the probability of getting the data in your sample.

ML solves several problems:

- Optimally combines multiple estimates for overidentified models
- Gives tests of overidentifying restrictions.
- Produces standard errors of estimates.

Properties: Consistency, asymptotic efficiency, asymptotic normality.

These are all large-sample approximations. But the approximations may not be great in small samples.

100

# Multivariate Normality

Conventional ML estimation of CF models assumes *multivariate normality* for endogenous variables That implies the following:

- Each variable has a normal distribution.
- All conditional distributions are normal, e.g.,  $f(X|Y=y)$  is normal.
- All conditional expectation functions are linear, e.g.,
$$E(Y| x, z, w) = \alpha + \beta x + \delta z + \gamma w$$
- All conditional variance functions are constant (homoscedasticity)

$$V(Y| x, z, w) = \sigma_y^2$$

101

## ML Details

Data consists of observed covariance matrix  $\mathbf{S}$ , which is an estimate of the population covariance matrix  $\mathbf{\Sigma}$ . According to the factor model

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$$

$\mathbf{\Lambda}$  is the matrix of factor loadings.  $\mathbf{\Phi}$  is the covariance matrix for the factors (latent variables).  $\mathbf{\Psi}$  is the covariance matrix for the errors.

ML estimation is accomplished by choosing  $\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Phi}}$  and  $\hat{\mathbf{\Psi}}$  to minimize

$$F(\hat{\mathbf{\Sigma}}) = \log|\hat{\mathbf{\Sigma}}| + tr(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}) - \log|\mathbf{S}| - p$$

where  $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Phi}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$

Note: When the model is just identified  $\hat{\mathbf{\Sigma}} = \mathbf{S}$  and  $F(\hat{\mathbf{\Sigma}}) = 0$ .

How is the function minimized?

- Many different numerical techniques for optimizing (maximizing or minimizing) a function.
- Most algorithms are iterative, i.e., they reach the solutions by successive approximations.

102



# Chi-Square Test

- If the specified model is correct,  $(n-1)F(\hat{\Sigma})$  has a chi-square distribution. The df is equal to the number of overidentifying restrictions.
- This statistic is a likelihood ratio chi-square comparing the fitted model with a saturated (just-identified) model that perfectly fits the data. If the chi-square is large and the  $p$ -value is small, it's an indication that the model should be rejected.
- Although this statistic is properly regarded as a test of the model, note that it is only testing the overidentifying restrictions.
- This test is sensitive to sample size. With a large sample, it may be difficult to find any parsimonious model that passes this test.

103

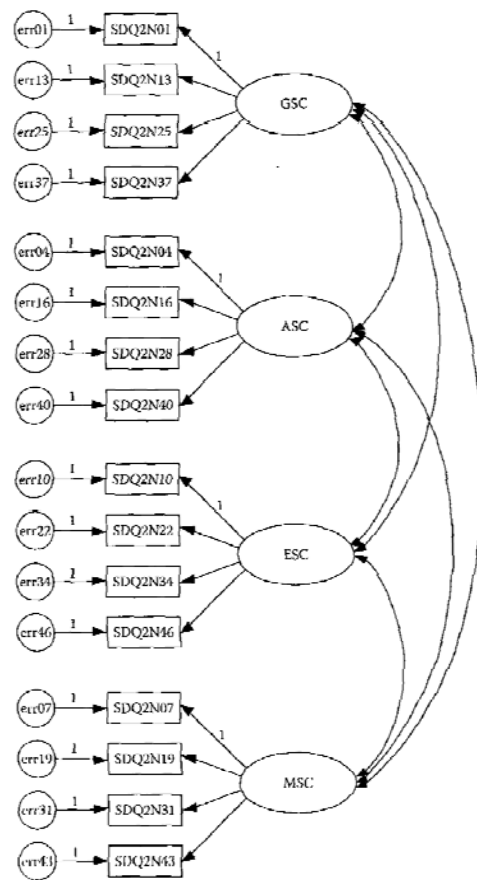
## Example 4. Self-Concept Measurement

From Byrne, *Structural Equation Modeling with AMOS* (2010). 265 seventh graders were asked a battery of questions. "Self concept" was hypothesized to have four dimensions, each with four indicators:

GSC – general self concept  
ASC – academic self concept  
ESC – English self concept  
MSC – mathematics self concept

```
DATA: FILE IS c:\data\asc7indm.txt;
VARIABLE: NAMES ARE v01-v24 sdq2n01 sdq2n13 sdq2n25 sdq2n37
sdq2n04 sdq2n16 sdq2n28 sdq2n40 sdq2n10 sdq2n22 sdq2n34 sdq2n46
sdq2n07 sdq2n19 sdq2n31 sdq2n43 masteng1 mastmat1 teng1 tmat1
seng1 smat1; MISSING=.;
USEVAR ARE sdq2n01 sdq2n13 sdq2n25 sdq2n37 sdq2n04 sdq2n16 sdq2n28
sdq2n40 sdq2n10 sdq2n22 sdq2n34 sdq2n46 sdq2n07 sdq2n19 sdq2n31
sdq2n43;
MODEL:
gsc BY sdq2n01 sdq2n13 sdq2n25 sdq2n37;
asc BY sdq2n04 sdq2n16 sdq2n28 sdq2n40;
esc BY sdq2n10 sdq2n22 sdq2n34 sdq2n46;
msc BY sdq2n07 sdq2n19 sdq2n31 sdq2n43;
OUTPUT: STDYX;
```

104



105

## SAS Code for Example 4

```

DATA asc7indm;
INFILE 'c:\data\asc7indmsp.txt';
INPUT v01-v24 sdq2n01 sdq2n13 sdq2n25 sdq2n37
      sdq2n04 sdq2n16 sdq2n28 sdq2n40 sdq2n10 sdq2n22 sdq2n34
      sdq2n46 sdq2n07 sdq2n19 sdq2n31 sdq2n43 masteng1
      mastmat1 teng1 tmat1 seng1 smat1;

PROC CALIS DATA=asc7indm;
  PATH
    gsc -> sdq2n01 sdq2n13 sdq2n25 sdq2n37 = 1,
    asc -> sdq2n04 sdq2n16 sdq2n28 sdq2n40 = 1,
    esc -> sdq2n10 sdq2n22 sdq2n34 sdq2n46 = 1,
    msc -> sdq2n07 sdq2n19 sdq2n31 sdq2n43 = 1,
RUN;

```

106

## Example 4. Mplus Results

### TESTS OF MODEL FIT

#### Chi-Square Test of Model Fit

Value	159.112
Degrees of Freedom	98
P-Value	0.0001

#### RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.049
90 Percent C.I.	0.034 0.062
Probability RMSEA <= .05	0.556

#### CFI/TLI

CFI	0.961
TLI	0.953

Chi-square test indicates that model should be rejected. But all the other fit indices indicate a good fit to the data.

107

## Example 4. Results

STANDARDIZED MODEL RESULTS		Estimate	S.E.	Est./S.E.	Two-Tailed
					P-Value
GSC	BY				
	SDQ2N01	0.582	0.055	10.613	0.000
	SDQ2N13	0.626	0.050	12.426	0.000
	SDQ2N25	0.544	0.056	9.644	0.000
	SDQ2N37	0.640	0.051	12.608	0.000
ASC	BY				
	SDQ2N04	0.536	0.048	11.143	0.000
	SDQ2N16	0.774	0.031	24.983	0.000
	SDQ2N28	0.703	0.037	19.071	0.000
	SDQ2N40	0.695	0.036	19.043	0.000
ESC	BY				
	SDQ2N10	0.711	0.044	16.167	0.000
	SDQ2N22	0.668	0.046	14.447	0.000
	SDQ2N34	0.322	0.064	5.002	0.000
	SDQ2N46	0.532	0.054	9.873	0.000
MSC	BY				
	SDQ2N07	0.854	0.020	41.953	0.000
	SDQ2N19	0.755	0.030	24.825	0.000
	SDQ2N31	0.923	0.016	59.292	0.000
	SDQ2N43	0.712	0.033	21.287	0.000
ASC	WITH				
	GSC	0.707	0.057	12.421	0.000
ESC	WITH				
	GSC	0.555	0.072	7.733	0.000
	ASC	0.758	0.052	14.652	0.000
MSC	WITH				
	GSC	0.534	0.061	8.756	0.000
	ASC	0.767	0.038	20.291	0.000
	ESC	0.266	0.073	3.652	0.000

108

## Global Goodness of Fit Measures

We want a single number that measures the similarity of  $\hat{\Sigma}$  and  $S$ , the predicted covariance matrix (based on the model) and the observed covariance matrix.

As a general approach to model evaluation, LR chi-square may be too sensitive to sample size. Many alternative statistics have been proposed. Here are the ones reported by Mplus.

### Tucker Lewis Index (TLI)

Also known as Bentler & Bonnet's NonNormed Fit Index

Let  $\chi_1^2$  be the chi-square for the fitted model and let  $\chi_0^2$  be the chi-square for some baseline model, usually the "independence" model which says all the observed covariances are really 0. (Mplus only considers covariances among endogenous variables and between endogenous and exogenous variables).

$$TLI = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_1^2}{df_1}}{\frac{\chi_0^2}{df_0} - 1}$$

Adjusts for relative complexity of the two models. Can sometimes be greater than 1, a possible indication of "overfitting".

109

## Other Global Measures

### Comparative fit index

$$CFI = \frac{(\chi_0^2 - df_0) - (\chi_1^2 - df_1)}{\chi_0^2 - df_0}$$

As with the TLI, models pay a penalty for more parameters. The formula can be greater than 1 or less than 0, in which case the CFI is simply set to 1 or 0.

### Root mean squared error of approximation:

$$RMSEA = \sqrt{\frac{\frac{\chi_1^2}{df_1} - 1}{N - 1}}$$

Good models have an RMSEA of .05 or less. Models whose RMSEA is .10 or more have poor fit. One nice thing about this statistic is that you can get a confidence interval.

110

## Other Global Measures (cont.)

### SRMR (Standardized Root Mean Square Residual)

The square root of the average of the squared differences between predicted correlations and observed correlations:

$$\sqrt{\frac{1}{q(q-1)/2} \sum_i \sum_j (\hat{\rho}_{ij} - r_{ij})^2}$$

where q is the number of variables.

### Akaike's Information Criterion

$$AIC = \chi^2 - 2df$$

Useful in comparing non-nested models. Choose the one with the lowest AIC.

### Schwarz's Bayesian Information Criterion

$$BIC = \chi^2 - \ln(N)df$$

Similar to AIC, but models are more strongly penalized for complexity.

111

## Specific Goodness of Fit Measures

Statistics that indicate poor fit for particular parameters or portions of the covariance matrix.

### Residuals

$s_{ij} - \hat{\sigma}_{ij}$ : observed covariance – covariance based on model.

These can also be calculated for variances, means, and other parameters.

Mplus also reports *standardized residuals*, which can be interpreted as z-scores

$$\frac{s_{ij} - \hat{\sigma}_{ij}}{se(s_{ij} - \hat{\sigma}_{ij})}$$

where *se* means standard error. If these are large (greater than 3, say), it suggests that the model is not fitting this particular parameter very well.

Mplus also reports *normalized residuals*:  $\frac{s_{ij} - \hat{\sigma}_{ij}}{se(s_{ij})}$

To get all three, just put RESIDUAL on the OUTPUT command.

OUTPUT: RESIDUAL;

112

## Standardized Residuals for Self-Concept Model

Standardized Residuals (z-scores) for Covariances/Correlations/Residual Corr					
	SDQ2N01	SDQ2N13	SDQ2N25	SDQ2N37	SDQ2N04
SDQ2N01	0.000				
SDQ2N13	0.357	0.000			
SDQ2N25	4.371	-1.310	0.000		
SDQ2N37	-2.331	0.928	-1.434	0.000	
SDQ2N04	0.036	2.354	-0.647	1.147	999.000
SDQ2N16	-1.093	-0.393	-1.019	1.878	0.636
SDQ2N28	-1.995	-1.239	-1.379	1.028	-0.170
SDQ2N40	-0.965	-0.368	-0.169	2.939	-3.319
SDQ2N10	-0.077	0.390	-0.807	-0.162	1.873
SDQ2N22	-0.788	0.001	-0.184	0.091	-0.671
SDQ2N34	0.990	1.631	0.695	0.880	-0.650
SDQ2N46	0.900	0.574	-1.016	-0.800	-0.511
SDQ2N07	-0.791	-1.987	0.017	-0.183	0.102
SDQ2N19	-0.381	-0.147	-1.157	-0.297	-0.097
SDQ2N31	-0.407	-0.071	0.833	2.550	0.668
SDQ2N43	-1.669	0.244	-0.293	-0.669	-0.985

A large value of the residual for two variables suggests a need to modify the model by including

- A direct effect of one variable on the other.
- Effects of other variable(s) on both variables.
- Correlated error terms.

113

## Residuals in SAS

You can get residuals in SAS by using the RESIDUAL option on the PROC statement. Here's what's available:

**RESIDUAL= NORM**

These are the normalized residuals of Mplus.

**RESIDUAL= ASYSTAND**

These are the standardized residuals of Mplus, which can be interpreted as test statistics.

**RESIDUAL= VARSTAND**

These are residuals based on correlations rather than covariances.

All three options also give the raw residuals. All the residuals can be written to a data set. There are also options for graphical displays.

114

## Modification Indices

Also known as Langrange multiplier tests, these statistics give the approximate change in chi-square that would occur if a constrained parameter were allowed to become a free parameter.

In Mplus, use MODINDICES option.

OUTPUT: MODINDICES;

The default is to report only modification index values that are greater than 10. If you want to see all them, use MODINDICES(0) .

The default is also to not report mod indices for effects of exogenous variables on endogenous variables. To see all of them use MODINDICES(ALL) .

In SAS, put the MOD option on the PROC statement..

115

## Mod Indices for Self-Concept

### MODEL MODIFICATION INDICES

NOTE: Modification indices for direct effects of observed dependent variables regressed on covariates may not be included. To include these, request MODINDICES (ALL).

Minimum M.I. value for printing the modification index 10.000

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
BY Statements					
ASC	BY SDQ2N07	11.251	-0.563	-0.422	-0.237
WITH Statements					
SDQ2N25	WITH SDQ2N01	17.054	0.359	0.359	0.319
SDQ2N31	WITH SDQ2N07	10.696	0.305	0.305	0.546
SDQ2N31	WITH SDQ2N19	17.819	-0.331	-0.331	-0.495

116

## Freeing Up Parameters

In deciding what parameters to “free up”, consider that direct effects of latent variables on observed variables are generally more conceptually meaningful than correlated errors.

It’s usually recommended to free up parameters one at a time, re-estimating the model in each case. Let’s allow SDQ2N07 to depend on ASC.

MODEL :

```
gsc BY sdq2n01 sdq2n13 sdq2n25 sdq2n37 ;
asc BY sdq2n04 sdq2n16 sdq2n28 sdq2n40 sdq2n07 ;
esc BY sdq2n10 sdq2n22 sdq2n34 sdq2n46;
msc BY sdq2n07 sdq2n19 sdq2n31 sdq2n43;
```

OUTPUT: MODINDICES;

117

## Results from Freeing 1 Parameter

Chi-Square Test of Model Fit

Value	147.328
Degrees of Freedom	97
P-Value	0.0008

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.044
90 Percent C.I.	0.029 0.058
Probability RMSEA <= .05	0.737

CFI/TLI

CFI	0.968
TLI	0.961

The chi-square has declined from 159 to 147. The RMSEA declined from .049 to .044. TLI improved from .953 to .961.

118



## Selected Results (cont.)

### STDYX Standardization

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ASC	BY				
	SDQ2N04	0.532	0.048	10.999	0.000
	SDQ2N16	0.775	0.030	25.418	0.000
	SDQ2N28	0.708	0.036	19.483	0.000
	SDQ2N40	0.699	0.036	19.265	0.000
	SDQ2N07	-0.257	0.080	-3.196	0.001
MSC	BY				
	SDQ2N07	1.070	0.070	15.281	0.000
	SDQ2N19	0.758	0.029	25.858	0.000
	SDQ2N31	0.915	0.015	60.123	0.000
	SDQ2N43	0.707	0.034	21.001	0.000

### MODEL MODIFICATION INDICES

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
WITH	Statements				
SDQ2N25	WITH SDQ2N01	17.228	0.361	0.361	0.320
SDQ2N31	WITH SDQ2N19	11.868	-0.265	-0.265	-0.381

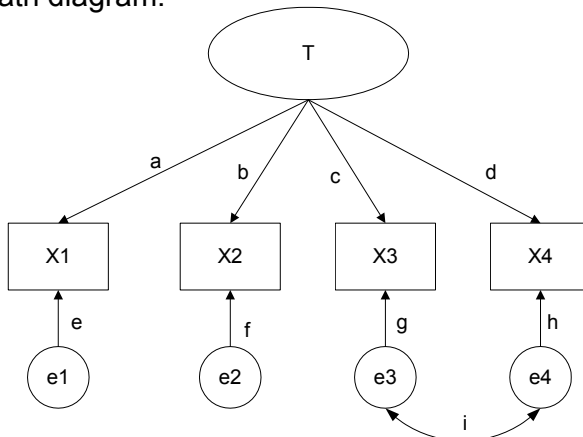
119

## Correlated Errors

We have only considered models with measurement errors that are uncorrelated. What kinds of models with correlated errors are identified?

Consider first some single-factor models. A model with one latent factor and three observed variables is just identified. Can't add any correlated errors.

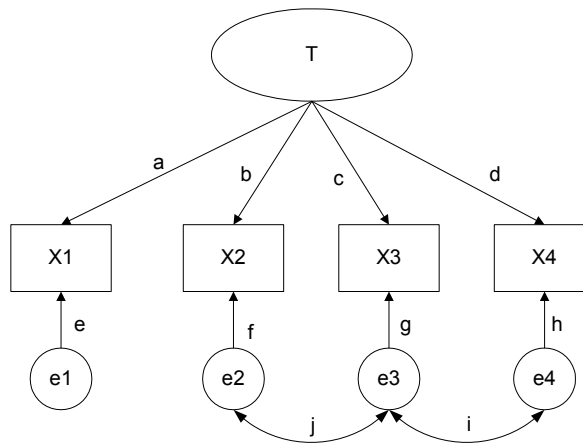
Consider one factor with four indicators. Ordinarily, this model has two overidentifying restrictions. Let's add one correlated error to a standardized path diagram:



This model can be estimated and it still has 1 overidentifying restriction. Can we add another error correlation?

120

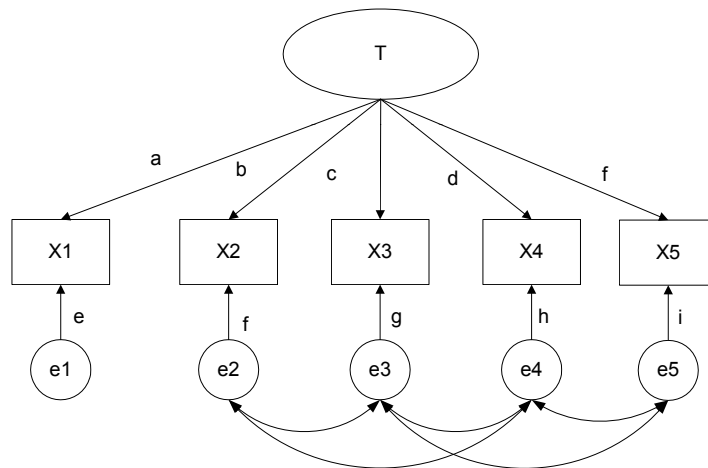
## Two Correlated Errors



Now the model is just identified, and we can't add any more error correlations. There are, however, several equivalent models.

121

## A Five-Indicator Model



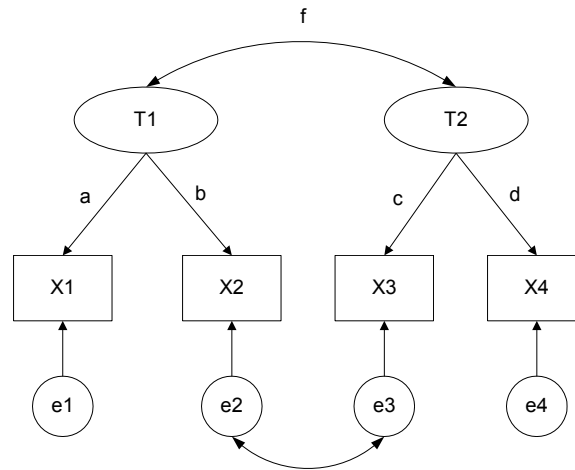
This model is just identified: 10 correlations,  
10 parameters (5 factor loadings, 5 error).

Note: No correlation between  $e_2$  and  $e_5$ .

Again, there are many equivalent models.

122

## A Two-Factor Model



Without the error correlation, this model has one over-identifying restriction. With the correlation it's just identified. How do we know?

If we delete  $X_2$ , we can still get c and d.

If we delete  $X_3$ , we can still get a and b.

Can't have a correlation between  $e_1$  and  $e_2$ , or between  $e_3$  and  $e_4$ .

123

## Example: Self-Concept Data

Let's add the two error correlations with mod indices > 10:

```
DATA: FILE IS c:\data\asc7indm.txt;
VARIABLE: NAMES ARE v01-v24 sdq2n01 sdq2n13 sdq2n25 sdq2n37
sdq2n04 sdq2n16 sdq2n28 sdq2n40 sdq2n10 sdq2n22 sdq2n34 sdq2n46
sdq2n07 sdq2n19 sdq2n31 sdq2n43 masteng1 mastmat1 teng1 tmat1
seng1 smat1;
USEVAR ARE sdq2n01 sdq2n13 sdq2n25 sdq2n37 sdq2n04 sdq2n16 sdq2n28
sdq2n40 sdq2n10 sdq2n22 sdq2n34 sdq2n46 sdq2n07 sdq2n19 sdq2n31
sdq2n43;
MODEL:
gsc BY sdq2n01 sdq2n13 sdq2n25 sdq2n37 ;
asc BY sdq2n04 sdq2n16 sdq2n28 sdq2n40 sdq2n07 ;
esc BY sdq2n10 sdq2n22 sdq2n34 sdq2n46;
msc BY sdq2n07 sdq2n19 sdq2n31 sdq2n43;
sdq2n25 WITH sdq2n01; sdq2n31 WITH sdq2n19 ;
OUTPUT: MODINDICES STDYX;
```

Note that if we use *WITH* to specify correlations among *endogenous* variables, we get correlations among their disturbance terms (partial correlations).

124

## Selected Results

### Chi-Square Test of Model Fit

Value	116.937
Degrees of Freedom	95
P-Value	0.0629

### RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.030	
90 Percent C.I.	0.000	0.046
Probability RMSEA <= .05	0.982	

### CFI/TLI

CFI	0.986
TLI	0.982

### SDQ2N25 WITH

SDQ2N01	0.283	0.063	4.489	0.000
---------	-------	-------	-------	-------

### SDQ2N31 WITH

SDQ2N19	-0.583	0.226	-2.586	0.010
---------	--------	-------	--------	-------

With enough correlated errors, any model will fit. But, beware of mindless addition of correlated errors as a way to make a model fit well.

125

## Structural Relations Among Latent Variables

We now consider a more general model that also allows

- Observed variables to affect latent variables
- Latent variables to affect other latent variables.

The general model can be written as follows:

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x}$$

where  $\mathbf{y}$  is a vector of endogenous variables (which may be either observed or latent) and  $\mathbf{x}$  is a vector of exogenous variables (which also includes the error terms and may be either observed or latent).

The matrix  $\mathbf{B}$  has zeros on the main diagonal, corresponding to the presumption that the  $y$ 's cannot affect themselves.

All models considered so far are special cases of this general model.

126

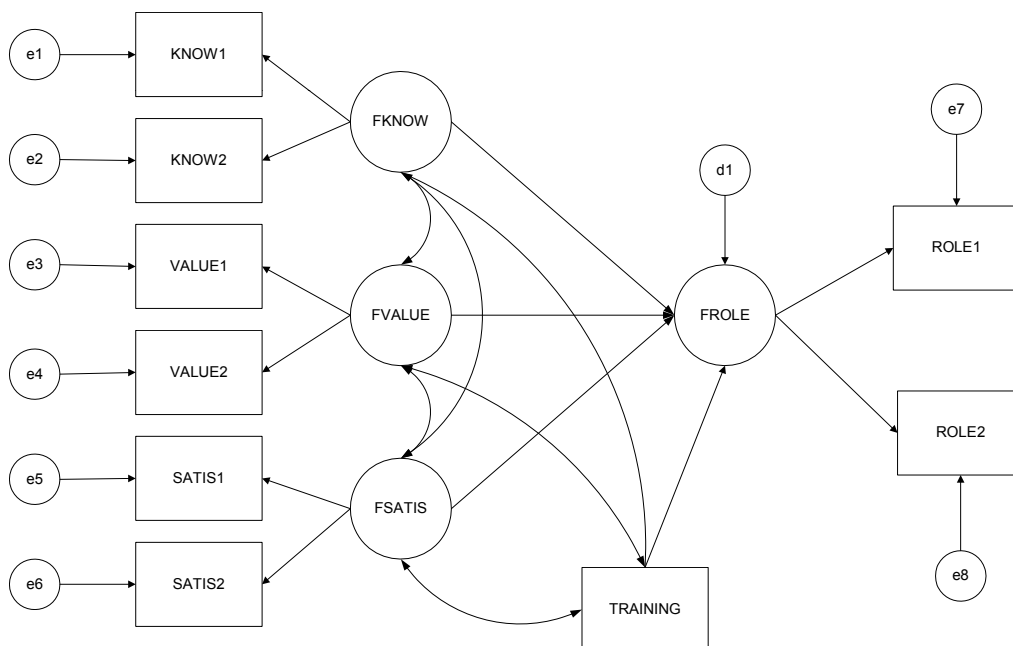
## Example 5. 98 Farm Managers (Rock et al. 1977)

Observed variables:

ROLE1	Split-half measure of role behavior
ROLE2	Split-half measure of role behavior
KNOW1	Split-half measure of knowledge
KNOW2	Split-half measure of knowledge
VALUE1	Split-half measure of value orientation
VALUE2	Split-half measure of value orientation
SATIS1	Split-half measure of role satisfaction
SATIS2	Split-half measure of role satisfaction
TRAINING	Measure of past training

127

### Farm Managers Path Diagram



128

## Data and Mplus Code

**Input data set (farmmgr.txt) in the form of a covariance matrix:**

```
.0271
.0172 .0222
.0219 .0193 .0876
.0164 .0130 .0317 .0568
.0284 .0294 .0383 .0151 .1826
.0217 .0185 .0356 .0230 .0774 .1473
.0083 .0011 -.0001 .0055 -.0087 -.0069 .1137
.0074 .0015 .0035 .0089 -.0007 -.0088 .0722 .1024
.0180 .0194 .0203 .0182 .0563 .0142 -.0056 -.0077 .0946
```

**Mplus program:**

```
DATA: FILE IS c:\data\farmmgr.txt; TYPE IS COVA; NOBS IS 98;
VARIABLE: NAMES = role1 role2 know1 know2 value1 value2 satis1 satis2
          training;
MODEL:
  knowledg BY know1 know2;
  roleperf BY role1 role2;
  values BY value1 value2;
  satisfac BY satis1 satis2;
  roleperf ON knowledg values satisfac training;
  training WITH knowledg values satisfac;
OUTPUT: STDYX;
```

129

## Example 5: SAS Code

```
DATA farmmgr(TYPE=COV);
INPUT role1 role2 know1 know2 value1 value2 satis1 satis2 training;
DATALINES;
.0271 . . . . .
.0172 .0222 . . . . .
.0219 .0193 .0876 . . . . .
.0164 .0130 .0317 .0568 . . . . .
.0284 .0294 .0383 .0151 .1826 . . . . .
.0217 .0185 .0356 .0230 .0774 .1473 . . .
.0083 .0011 -.0001 .0055 -.0087 -.0069 .1137 . .
.0074 .0015 .0035 .0089 -.0007 -.0088 .0722 .1024 .
.0180 .0194 .0203 .0182 .0563 .0142 -.0056 -.0077 .0946
PROC CALIS DATA=farmmgr COV NOBS=98;
PATH
  knowledg -> know1 know2 = 1,
  roleperf -> role1 role2 = 1,
  values -> value1 value2 = 1,
  satisfac -> satis1 satis2 = 1,
  roleperf <- knowledg values satisfac training;
RUN;
```

130

## Example 5. Selected Results

Chi-Square Test of Model Fit				
	Value			20.218
	Degrees of Freedom			18
	P-Value			0.3207
CFI/TLI				
	CFI			0.991
	TLI			0.982
RMSEA (Root Mean Square Error Of Approximation)				
	Estimate			0.035
	90 Percent C.I.			0.000 0.100
	Probability RMSEA <= .05			0.580
STDYX Standardization				
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
KNOWLEDG BY				
KNOW1	0.719	0.090	7.974	0.000
KNOW2	0.625	0.090	6.940	0.000
ROLEPERF BY				
ROLE1	0.836	0.050	16.610	0.000
ROLE2	0.839	0.050	16.737	0.000
VALUES BY				
VALUE1	0.855	0.100	8.531	0.000
VALUE2	0.552	0.094	5.892	0.000

131

## Example 5. Selected Results (cont.)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SATISFAC BY				
SATIS1	0.821	0.201	4.080	0.000
SATIS2	0.815	0.200	4.077	0.000
ROLEPERF ON				
KNOWLEDG	0.538	0.133	4.044	0.000
VALUES	0.312	0.142	2.197	0.028
SATISFAC	0.108	0.100	1.085	0.278
TRAINING	0.146	0.109	1.339	0.180
VALUES WITH				
KNOWLEDG	0.461	0.134	3.454	0.000
SATISFAC WITH				
KNOWLEDG	0.086	0.141	0.612	0.541
VALUES	-0.060	0.129	-0.465	0.642
TRAINING WITH				
KNOWLEDG	0.343	0.118	2.915	0.004
VALUES	0.462	0.104	4.420	0.000
SATISFAC	-0.081	0.112	-0.718	0.473

132

## A Tau-Equivalent Model

Since measures are split-halves of multi-item scales, it's reasonable to hypothesize that they are tau-equivalent and possibly parallel.

To impose tau-equivalence, set all loadings equal to 1:

MODEL:

```
knowledg BY know1 know2@1;  
roleperf BY role1 role2@1;  
values BY value1 value2@1;  
satisfac BY satis1 satis2@1;
```

Note: The first indicator on each line has its loading set to 1 by default (normalizing constraint)

This produces  $X^2=26.97$  with 22 d.f. Compare with initial model

$$\begin{array}{r} 26.97 \\ \underline{20.01} \\ 6.97 \end{array} \quad \begin{array}{r} 22 \\ \underline{18} \\ 4 \end{array} \quad p=.14$$

133

## Parallel Model

To impose parallelism, do tau-equivalence model, plus set error variances equal for each pair of measures.

MODEL:

```
knowledg BY know1 know2@1;  
roleperf BY role1 role2@1;  
values BY value1 value2@1;  
satisfac BY satis1 satis2@1;  
roleperf ON knowledg values satisfac training;  
training WITH knowledg values satisfac;  
know1 know2 (1)  
role1 role2 (2)  
value1 value2 (3)  
satis1 satis2 (4);
```

Compare with tau-equivalent model.

$$\begin{array}{r} 32.14 \\ \underline{26.97} \\ 5.17 \end{array} \quad \begin{array}{r} 26 \\ \underline{22} \\ 4 \end{array} \quad p=.27$$

134



# Parallel Model in SAS

```
PROC CALIS DATA=farmmgr NOBS=98;  
PATH  
  knowledg -> know1 know2 = 1 1,  
  roleperf -> role1 role2 = 1 1,  
  values -> value1 value2 = 1 1,  
  satisfac -> satis1 satis2 = 1 1,  
  roleperf <- knowledg values satisfac training,  
  <-> know1 know2 = a a,  
  <-> role1 role2 = b b,  
  <-> value1 value2 = c c,  
  <-> satis1 satis2 = d d;  
RUN;
```

135

## Identification in SEM Models

We can think of a general model as being composed of two parts:

- Measurement (a confirmatory factor model)
- Structural (causal relationships among latent and observed variables).

In most cases, we can approach identification for each part separately:

Is the measurement portion identified?

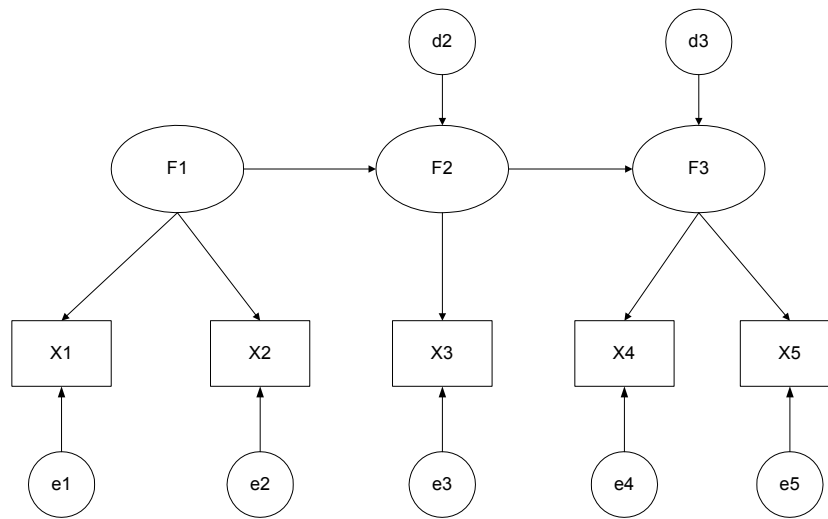
Is the structural portion identified?

If the answer to both questions is yes, then the model as a whole is identified.

This is not a NECESSARY condition for identification, however. Some models are identified even if the CFA portion is not.

136

## An Identified SEM Model



This model is identified even though the CFA model is not identified. What makes it so?

137

## What to Do When a Model Doesn't Fit

Possibly nothing, especially if the sample size is large and the *N*-invariant measures of fit look good.

Isolate the problem:

- Remove all restrictions on relationships among latent variables by fitting the corresponding CFA model. If it fits OK, then the problem is in the structural part. The chi-square for the structural part is the difference between the CFA chi-square and the overall chi-square.
- If the problem is in the structural part, try adding additional paths.
- If you have four or more indicators for some latent variables, try fitting measurement models for just those variables.
- Try fitting a model to subparts of the whole model, e.g., if you have five latent variables, try models for two at time.
- Problems in measurement part can sometimes be solved by deleting entire variables.

Modification indices and residuals can be useful, but use cautiously. Keep in mind the possibility that more latent variables may be needed (which won't be obvious from modification indices).

138

## Alternative Estimation Methods

Most SEM packages offer estimation methods that are alternatives to ML. We'll begin with a method that can be implemented with data in the form of a covariance matrix or correlation matrix.

### **Generalized Least Squares**

Here the function being minimized is

$$tr[(S^{-1}(S - \hat{\Sigma}))^2]$$

Like ML, if the data are multivariate normal, GLS estimators are consistent, asymptotically normal and asymptotically efficient. However, these properties also hold under somewhat less restrictive assumptions.

To implement this method in Mplus, use the ESTIMATOR=GLS option on the ANALYSIS command.

139

## GLS Example

Farm Manager Data

DATA:

FILE IS c:\data\farmmgr.txt;

TYPE IS COVA;

NOBS IS 98;

VARIABLE: NAMES ARE role1 role2 know1 know2 value1  
value2 satis1 satis2 training;

ANALYSIS: ESTIMATOR=GLS;

MODEL:

knowledg BY know1 know2;

roleperf BY role1 role2 ;

values BY value1 value2;

satisfac BY satis1 satis2;

roleperf ON knowledg values satisfac training;

training WITH knowledg values satisfac;

140

## GLS Results

WARNING: THE RESIDUAL COVARIANCE MATRIX (THETA) IS NOT POSITIVE DEFINITE. THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR AN OBSERVED VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO OBSERVED VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO OBSERVED VARIABLES.

CHECK THE RESULTS SECTION FOR MORE INFORMATION.  
PROBLEM INVOLVING VARIABLE SATIS2.

### Chi-Square Test of Model Fit

Value	18.115
Degrees of Freedom	18
P-Value	0.4481

### Residual Variances

ROLE1	0.007	0.002	3.418	0.001
ROLE2	0.006	0.002	3.508	0.000
KNOW1	0.042	0.011	3.953	0.000
KNOW2	0.030	0.007	4.635	0.000
VALUE1	0.028	0.034	0.820	0.412
VALUE2	0.085	0.017	5.022	0.000
SATIS1	0.064	0.027	2.385	0.017
SATIS2	-0.010	0.066	-0.154	0.877
ROLEPERF	0.007	0.002	2.856	0.004

Another Heywood case, which we didn't get with ML.

141

## Weighted Least Squares

Also known as Browne's distribution-free method, WLS is more robust to departures from normality. But it requires raw data rather than the covariance matrix. And it may require larger samples to get good results.

Fitting function:

$$\text{vec}(s_{ij} - \hat{\sigma}_{ij})' \mathbf{W}^{-1} \text{vec}(s_{ij} - \hat{\sigma}_{ij})$$

where  $\text{vec}(s_{ij} - \hat{\sigma}_{ij})$  is a vector of all the observed and predicted elements in the covariance matrix, and  $\mathbf{W}$  is an estimate of the covariance matrix for the *parameter estimates*.

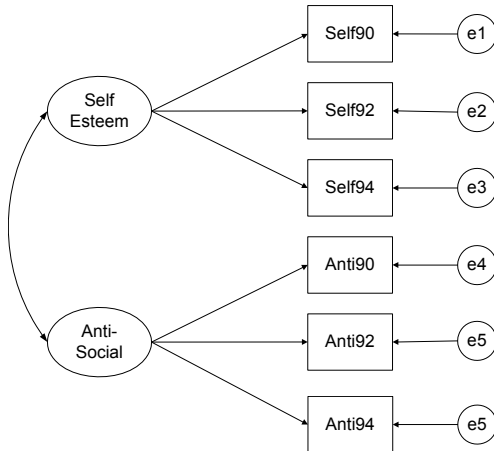
If the number of variables is large,  $\mathbf{W}$  will be extremely large. E.g., if the number of variables is 20, then there will be 210 variance and covariances. So  $\mathbf{W}$  will be 210 x 210, containing 22,155 unique elements. These will be poorly estimated unless the sample is large.

Example:

A sample of 581 children was surveyed in each of three years 1990, 1992, 1994. In each year, there were measures of anti-social behavior and self esteem. We'll estimate a confirmatory factor model:

142

## WLS Example



```

DATA: FILE IS c:\data\nlsy.dat;
VARIABLE: NAMES ARE anti90 anti92
         anti94 black childage gender
         hispanic married momage momwork
         pov90 pov92 pov94 self90 self92
         self94;
USEVARIABLES ARE anti90 anti92
         anti94 self90 self92 self94;
ANALYSIS: ESTIMATOR=WLS;
MODEL:
         anti BY anti90 anti92 anti94;
         self BY self90 self92 self94;
OUTPUT: STDYX;
  
```

143

## WLS Output

Chi-Square Test of Model Fit  
 Value  
 Degrees of Freedom  
 P-Value

25.958  
 8  
 0.0011

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ANTI BY				
ANTI90	0.759	0.027	28.465	0.000
ANTI92	0.849	0.026	33.273	0.000
ANTI94	0.737	0.032	22.838	0.000
SELF BY				
SELF90	0.430	0.047	9.098	0.000
SELF92	0.770	0.066	11.651	0.000
SELF94	0.555	0.060	9.310	0.000
SELF WITH ANTI	-0.197	0.057	-3.486	0.000

ML and GLS give very similar results.

144

# Other ESTIMATOR Options

Other ESTIMATOR= options in Mplus for linear models with raw data:

- MLR ML parameter estimates with robust (sandwich) standard errors. Robust to non-normality, heteroskedasticity, and dependence (if used with COMPLEX option)
- MLM ML parameter estimates with “mean adjusted” standard errors and chi-square that are robust to non-normality.
- MLMV ML parameter estimates with “mean and variance adjusted” standard errors and chi-square that are robust to non-normality.

145

## Multiple Group Analysis

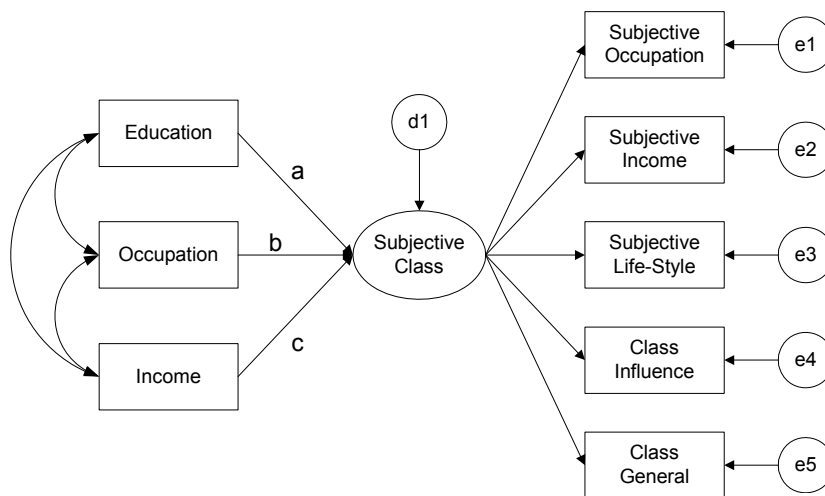
- Often we want to divide a sample into subgroups, estimate a model for each subgroup, then compare results across subgroups. Most SEM programs can fit models for multiple groups simultaneously.
- This makes it possible to constrain some parameters to be the same across groups while allowing others to differ. As we’ll see, this is one approach to testing for interaction.

Example: Kluegel et al. (1977) “Subjective class identification: A multiple indicator approach,” *American Sociological Review* 42: 599-611.

I used their data (432 Whites, 368 Blacks) but fit a different model.

146

## Example 6. Multiple Groups



I estimated models for blacks and for whites simultaneously using Mplus. The next slide shows the data, with standard deviations and correlations for each group, whites preceding blacks:

147

## Example 6. Data

```

1.203 21.277 2.198 .640 .670 .623 .647 .627 .668 .670 .736 .624
1
.495 1
.398 .292 1
.218 .282 .184 1
.299 .166 .383 .386 1
.272 .161 .321 .396 .553 1
.269 .169 .191 .382 .456 .534 1
.330 .231 .277 .431 .537 .608 .590 1
-.133 -.116 -.182 -.095 -.195 -.148 -.162 -.158 1
-.235 -.175 -.203 -.159 -.231 -.155 -.235 -.192 .401 1
-.248 -.225 -.140 -.152 -.137 -.137 -.194 -.172 .385 .508 1
-.277 -.171 -.239 -.138 -.290 -.154 -.189 -.170 .439 .504 .481 1
1.106 16.224 2.097 .747 .814 .786 .859 .784 .642 .729 .778 .726
1
.404 1
.268 .220 1
.216 .277 .268 1
.233 .183 .424 .550 1
.211 .270 .325 .574 .647 1
.207 .157 .282 .482 .517 .647 1
.202 .142 .238 .482 .472 .622 .632 1
-.143 -.185 -.133 -.213 -.095 -.139 -.182 -.142 1
-.283 -.105 -.159 -.123 -.188 -.196 -.259 -.117 .351 1
-.233 -.137 -.113 -.133 -.160 -.154 -.186 -.173 .316 .372 1
-.270 -.137 -.169 -.166 -.211 -.231 -.260 -.191 .327 .423 .523 1

```

148

## Example 6. Models

It's important to analyze the covariance matrix (not the correlation matrix) in order to compare unstandardized coefficients.

I fit four different models with varying degrees of sameness and difference:

	X <sup>2</sup>	df
1. No constraints across groups	113.7	34
2. Factor loadings the same	122.9	38
3. Factor loadings the same, plus $a^B=a^W$ , $b^B=b^W$ , $c^B=c^W$	132.9	41
4. All parameters the same (excluding variances and covariances of ed, occ, & inc).	208.1	47

149

## Example 6. Mplus Code for Model 1

```
DATA:
  FILE IS c:\data\kluegelcorr.txt;
  TYPE IS CORR STD;
  NOBS IS 432 368;
  NGROUPS=2;
VARIABLE:
  NAMES = ed occ inc scocc scinc sclife scinf scgen
    pol1-pol4;
  USEVARIABLES = ed occ inc scocc scinc sclife scinf
    scgen;
MODEL:
  subclass BY scocc scinc sclife scinf scgen;
  subclass ON ed occ inc;
MODEL g2:
  subclass BY scocc@1 scinc sclife scinf scgen;
```

The second MODEL command is necessary because, otherwise, the factor loadings would be constrained to be the same across groups by default. All the others parameters are allowed to vary across groups.

150



## Example 6. Mplus Code (cont.)

For Model 2, just delete the second MODEL command. For Model 3, the code is

MODEL:

```
    subclass BY scocc scinc sclife
      scinf scgen;
    subclass ON ed (1)
      occ (2)
      inc (3);
```

Numbered constraints must  
be on separate lines.

MODEL g2:

```
    subclass ON ed (1)
      occ (2)
      inc (3);
```

151

## Example 6. Model 4 Code

MODEL:

```
    subclass BY scocc scinc sclife scinf scgen;
    subclass ON ed (1)
      occ (2)
      inc (3);
    scocc (4)
    scinc (5)
    sclife (6)
    scinf (7)
    scgen (8)
    subclass (9);
```

MODEL g2:

```
    subclass ON ed (1)
      occ (2)
      inc (3);
    scocc (4)
    scinc (5)
    sclife (6)
    scinf (7)
    scgen (8)
    subclass (9);
```

These 3 lines constrain the  
regression coefficients to be  
the same in the 2 groups

These 6 lines  
constrain the residual  
variances to be the  
same in the 2 groups.

152

## Tests for Comparing the Groups

We now construct likelihood ratio tests to answer several questions:

	X <sup>2</sup>	df	
1. Are the two groups the same? (Model 1 vs. Model 4)	208.1 <u>-113.7</u> 94.4	47 <u>-34</u> 13	p<.001
2. Are the factor loadings the same? (Model 1 vs. Model 2)	122.9 <u>-113.7</u> 9.2	38 <u>-34</u> 4	p=.056
3. Are a, b and c the same? (Model 2 vs. Model 3)	132.9 <u>-122.9</u> 10.0	41 <u>-38</u> 3	p=.016

Choose model 2.

153

## Interactions and Non-Linearities

How to build interactions (product terms) and nonlinearities into SEMs?

If the two variables are both observed, it's easy. Just create a product variable (using the DEFINE command) and treat it as another variable in the analysis. Clearly, this requires raw data input, not a correlation or covariance matrix.

Any variables created in the DEFINE command must be included at the end of a USEVARIABLES command.

Example:

```
DATA: FILE IS c:\data\nlsy.dat;
VARIABLE: NAMES ARE anti90 anti92 anti94 black childage
gender hispanic married momage momwork pov90 pov92 pov94
self90 self92 self94;
USEVARIABLES=anti92 black hispanic childage gender
married momage momwork pov92 self92 marwork;
DEFINE: marwork=married*momwork;
MODEL:
anti92 ON black hispanic childage gender
married momage momwork pov92 self92 marwork;
OUTPUT: STDYX;
```

154

## Interactions with Latent Variables

If one variable is latent and the other is observed categorical, you can do multiple group analysis. (Model 2 for subjective social class had an interaction between race and occupation in their effect on subjective social class).

But it can be difficult if both variables are latent, or if one is latent and the observed variable is *not* categorical.

Mplus has good methods for doing this. You have to use raw data.

155

## Interactions with Latent Variables

```
DATA: FILE IS c:\data\nlsy.dat;
VARIABLE: NAMES ARE anti90 anti92 anti94 black childage
           gender hispanic married momage momwork pov90 pov92 pov94
           self90 self92 self94;
USEVARIABLES ARE anti94 black childage gender
                 hispanic married momage momwork self90 self92
                 self94;
ANALYSIS: TYPE = RANDOM; ALGORITHM = INTEGRATION;
MODEL: selfflat BY self90 self92 self94;
       selfXmom | selfflat XWITH momage;
       anti94 ON black hispanic childage gender
                 married momage momwork selfflat selfXmom;
       selfflat WITH black hispanic childage gender
                 married momage momwork;
```

In the ANALYSIS command, you must specify TYPE=RANDOM and ALGORITHM=INTEGRATION. The interaction is defined in the second line of the MODEL command.

156

# Ordinal and Binary Data

- Dichotomy (0,1)
- Polytomy (1,2,3,4) etc. Assume that it's ordered.

Clearly such variables cannot be normally distributed, and unlikely to be linearly related to other variables.

Available methods:

## 1. Treat as interval.

- Widely done
- Some studies show not too bad, e.g., Johnson & Creech (1983) "Ordinal measures in multiple indicator models." *American Sociological Review*.
- Some bias and inefficiency, especially in small samples. Standard errors may be underestimated. Robust standard errors (MLR option) may be helpful.

157

## Special Correlations

### 2. Compute a special correlation matrix and use as input to conventional model fitting software.

For two dichotomous variables, compute a *tetrachoric correlation*.

Assume latent continuous variables  $X^*$  and  $Y^*$ , that are normally distributed with correlation  $\rho$ . Observed dummy variables  $X$  and  $Y$  depend on  $X^*$  and  $Y^*$  by

$X=1$  if  $X^* > \phi$ , otherwise 0

$Y=1$  if  $Y^* > \theta$ , otherwise 0

You can get a ML estimate of  $\rho$  based on contingency table for  $X$  and  $Y$ .

158

## Special Correlations

For two ordinal polytomous variables, compute a *polychoric correlation* (generalization of tetrachoric). Again, the data is in the form of a contingency table.

Two other kinds of correlations may also be needed:

- Biserial: One dichotomous and one continuous.
- Polyserial: One polytomous and one continuous.

These can be computed in PRELIS, a companion program to LISREL.

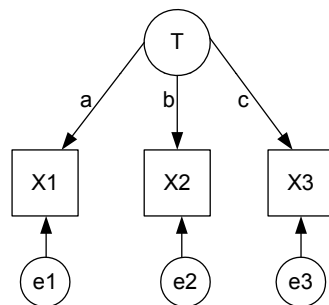
If you take this approach, you can only analyze a correlation matrix.

Conventional standard errors and test statistics may be inaccurate. There are methods for correcting these, but they require large samples.

159

## Specialized Models

**3. Develop a model especially for dichotomous or ordinal data and estimate it with efficient methods.**



Assume  $T$ ,  $X_1$ ,  $X_2$ , and  $X_3$  are multivariate normal.

But we don't directly observe  $X_1$ - $X_3$ . What we observe is  $Z_1$ - $Z_3$ , defined as

$Z_1=1$  if  $X_1 > \theta_1$ , otherwise 0

$Z_2=1$  if  $X_2 > \theta_2$ , otherwise 0

$Z_3=1$  if  $X_3 > \theta_3$ , otherwise 0

This leads to simultaneous probit models. Mplus is one of the few commercial packages I know that will estimate models like this.

160

## Mplus with Binary Data

```
DATA: FILE IS c:\data\nlsy.dat;
VARIABLE: NAMES ARE anti90 anti92 anti94 black childage
           gender hispanic married momage momwork pov90 pov92 pov94
           self90 self92 self94;
USEVARIABLES ARE pov90 pov92 pov94 black hispanic childage
                 gender married momage momwork;
CATEGORICAL ARE pov90 pov92 pov94;
MODEL:
    povlat BY pov90 pov92 pov94;
    povlat ON black hispanic childage gender married momage
             momwork;
OUTPUT: STDYX;
```

The default estimation method is weighted least squares with robust standard errors. Alternatively, one can do maximum likelihood with robust standard errors. In that case, the default “link” function is logit rather than probit. Implement this with the following statement, before the MODEL statement.

```
ANALYSIS: ESTIMATOR=MLR;
```

161

## Probit Results

```
Estimator      WLSMV
Chi-Square Test of Model Fit
Value          21.019*
Degrees of Freedom      14
P-Value        0.1011
```

- \* The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used for chi-square difference testing in the regular way. MLM, MLR and WLSM chi-square difference testing is described on the Mplus website. MLMV, WLSMV, and ULSMV difference testing is done using the DIFFTEST option.

```
RMSEA (Root Mean Square Error Of Approximation)
Estimate          0.029
90 Percent C.I.   0.000  0.054
Probability RMSEA <= .05      0.913

CFI/TLI
CFI              0.992
TLI              0.987
```

If you do logit rather than probit, you get none of this stuff—just a log-likelihood with AIC and BIC statistics.

162

## Probit Results (cont.)

### MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
POVLAT BY				
POV90	1.000	0.000	999.000	999.000
POV92	1.183	0.071	16.622	0.000
POV94	0.886	0.057	15.602	0.000
POVLAT ON				
BLACK	0.981	0.114	8.575	0.000
HISPANIC	0.441	0.128	3.443	0.001
CHILDAGE	0.027	0.076	0.351	0.725
GENDER	0.121	0.092	1.311	0.190
MARRIED	0.586	0.102	5.748	0.000
MOMAGE	-0.080	0.021	-3.733	0.000
MOMWORK	0.774	0.107	7.265	0.000
Thresholds				
POV90\$1	-0.231	1.186	-0.195	0.846
POV92\$1	-0.014	1.208	-0.012	0.991
POV94\$1	0.373	1.173	0.318	0.750
Residual Variances				
POVLAT	0.666	0.057	11.588	0.000

163

## Other Features of Mplus

Endogenous variables can be

- Ordered or unordered categorical
- Count data (Poisson or negative binomial regression)
- Censored (tobit models)
- Survival (Cox regression or piecewise exponential regression)

Latent variables can have indicators that are any combination of the above.

Mplus can handle missing data using full information maximum likelihood, for any of the data types listed above.

It can do multiple imputation based on most of the models that it can estimate.

With add-ons, Mplus can do multi-level modeling and latent class modeling.

Growth curve modeling (latent trajectories).

Monte Carlo simulations.

164

## Cautions About SEMs

Extremely useful method when appropriate.

But often applied inappropriately—one of the most overused and misused methods in social science.

My main complaint: Treating two or more variables as indicators of a single latent variable just because they seem vaguely related or fall within some very general theoretical concept.

Examples of “misuse”:

1. “Political participation.” Indicators:

- Voted in last election
- Contributed to a political party
- Worked in a political campaign
- Watched election debates

165

## Misuse

2. “Criminal behavior”

- Number of times arrested for crimes against property
- Number of times arrested for crimes against person
- Number of times convicted for felony
- Average seriousness of previous convictions

Is there some well-defined phenomenon that “causes” these indicators?

Especially annoying when a nicely measured variable like “number of offenses” is turned into some vague concept.

166



# Good Examples

Latent variable: Number of arrests

Indicators: self report, police records

Latent variable: Frequency of quarrels in a marriage

Indicators: husband's report, wife's report

Latent variable: Number of articles per year published by scientists

Indicators: counts from Chemical Abstracts, counts from Web of Science

Latent variable: Level of depression

Indicators: Beck depression inventory, CESD scale.

Split-half measures from multiple item scales.

167

## SEMs and Causality

Is it appropriate to describe SEMs as causal models?

- Much criticism on this point in recent years.
- There are really two questions, that often get confused:
  - Can non-experimental data “prove” causality?
    - Certainly not, although it can definitely provide evidence that supports a causal interpretation.
  - If the assumptions of the model are correct, can the parameters of the model be interpreted as causal effects?
    - Many statisticians in the “potential-outcome” school say no.
    - But for a vigorous and persuasive defense of a causal interpretation of SEM parameters, see:

Judea Pearl, *Causality* (2009)

\_\_\_\_\_, “The Causal Foundations of Structural Equation Modeling,” Chapter for R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2011). New York: Guilford Press.

168

## Exercise 1

Here's a sample covariance matrix for a study of 932 people

11.834					
6.947	9.364				
6.819	5.091	12.532			
4.783	5.028	7.495	9.986		
-3.839	-3.889	-3.841	-3.625	9.610	
-21.899	-18.831	-21.748	-18.775	35.522	450.288

The variables are

1. Anomie 1967
2. Powerlessness 1967
3. Anomie 1971
4. Powerlessness 1971
5. Education
6. Occupational Status

This covariance matrix is in a text file called asg1sem.txt. To read this file into Mplus, specify TYPE IS COVA on the DATA command. In SAS, you'll need a data step that looks like this:

```
data anomie(type=cov);  
  infile 'c:\data\Asg1SEM.txt' missover;  
  _type_='cov';  
  input  anomie67 power67 anomie71 power71 ed occ;  
run;
```

Use Mplus or SAS to estimate a path model with two endogenous variables, Anomie and Powerlessness, both in 1971. All the other variables are exogenous. Also include a direct effect of Anomie 1971 on Powerlessness 1971. Remember that, in Mplus, variable names should be limited to 8 characters.

Estimate the indirect effect of Anomie 1967 on Powerlessness 1971. Is it statistically significant?

*If you're using the demo version of Mplus, it has a limit of 2 independent variables. To get around this, add two more equations with Anomie 1967 and Powerlessness 1967 as dependent variables and Education and Occupational Status as independent variables. Also allow a partial correlation between Anomie 1967 and Powerlessness 1967. The model should be just identified (df=0, chi-square=0).*

## Exercise 2

1. Using the data in assignment 1, use Mplus or SAS to estimate a model with the same two endogenous variables, Anomie and Powerlessness in 1971, but now allow them to have effects on each other. To make the model identified, take Anomie67 out of the equations for Powerlessness, and take Powerless67 out of the equation for Anomie. Include a residual correlation. How would you interpret the results and how do they compare with the model of assignment 1.

*If you're using the demo version of Mplus, you will have to use the same trick as in assignment 1 to get around the limitation on number of independent variables.*

2. Re-estimate the model after setting the residual correlation to 0. Does it make a difference?

Keep the residual correlation at 0, and do the following variations:

- a) Put Powerless67 in the equation for Anomie, and re-estimate the model.
- b) Undo step (a), put Anomie67 in the equation for Powerlessness, and re-estimate the model.
- c) Let all 1967 variables affect the endogenous variables, and re-estimate the model.

What does all this tell you, if anything?

### Exercise 3. Confirmatory Factor Analysis

Jaccard, Weber and Lundmark (1975) measured attitudes toward cigarette smoking (C) and capital punishment (P) by four different methods: semantic differential (1), Likert (2), Thurstone (3) and Guilford (4). The correlation matrix for C1-C4 P1-P4 is given below. Specify a sample size of 350.

```
1.0
.78 1.0
.81 .77 1.0
.76 .71 .81 1.0
.29 .23 .19 .10 1.0
.28 .29 .18 .09 .84 1.0
.26 .31 .24 .08 .81 .89 1.0
.27 .24 .23 .15 .84 .91 .85 1.0
```

The correlation matrix is in a text file called asg3sem.txt. The data step for SAS will be

```
data cig_pun(type=corr);
  infile 'c:\data\Asg3SEM.txt' missover;
  _type_='corr';
  input c1-c4 p1-p4;
run;
```

Estimate a confirmatory factor model in which all four C variables are indicators of a single latent variable and all four P variables are indicators of another latent variable. The two latent variables should be allowed to correlate, but there should be no correlations among the error variables. Can you reject the hypothesis that the correlation between the two latent variables is 0?

*If you're using the demo version of Mplus, you'll to exclude C4 and P4 to get around the variable limits.*

Re-estimate the model, allowing for correlations among the error terms for P1 and C1, P2 and C2, P3 and C3, P4 and C4. Give an interpretation for this second model. Test the difference in chi-squares for the two models, and interpret the results of this test.

#### **Exercise 4**

For the data you analyzed in Assignments 1 and 2, estimate an SEM model with the following measurement relationships:

1. A latent variable Alienation67 is measured by Anomie67 and Powerlessness67.
2. A latent variable Alienation71 is measured by Anomie71 and Powerlessness71.
3. A latent variable SES is measured by education and occupational status.

The structural part of the model should be fully recursive, with Alienation71 depending on both Alienation67 and SES, and Alienation67 depending on SES. If the model doesn't fit (according to the chi-square), modify it until it does fit. Interpret the results.

## **Mplus Code for Exercises**

### **Exercise 1**

DATA:

FILE IS c:\data\asg1SEM.txt;

TYPE IS COVA;

NOBS IS 932;

VARIABLE:

NAMES ARE ano67 pow67 ano71 pow71 ed occ;

MODEL:

ano71 ON ano67 pow67 ed occ;

pow71 ON ano71 ano67 pow67 ed occ;

MODEL INDIRECT: pow71 IND ano67;

OUTPUT: STDYX;

## Exercise 2

DATA:

FILE IS c:\data\asg1SEM.txt;

TYPE IS COVA;

NOBS IS 932;

VARIABLE:

NAMES ARE ano67 pow67 ano71 pow71 ed occ;

MODEL:

ano71 ON ano67 pow71 ed occ;

pow71 ON ano71 pow67 ed occ;

pow71 WITH ano71;

OUTPUT: STDYX;

DATA:

FILE IS c:\data\asg1SEM.txt;

TYPE IS COVA;

NOBS IS 932;

VARIABLE:

NAMES ARE ano67 pow67 ano71 pow71 ed occ;

MODEL:

ano71 ON ano67 pow71 ed occ;

pow71 ON ano71 pow67 ed occ;

OUTPUT: STDYX;

To this model, you can add pow67 to the equation for ano71, or you can add ano67 to the equation for pow71. But you can't do both because then the model is underidentified.

### Exercise 3

```
DATA: FILE = c:\data\asg3SEM.txt;
      NOBS=35; TYPE=CORR;
VARIABLE:
      NAMES =c1-c4 p1-p4;
MODEL:
      cig BY c1-c4;
      pun BY p1-p4;
OUTPUT: STDYX;
```

```
DATA: FILE = c:\data\asg3SEM.txt;
      NOBS=35; TYPE=CORR;
VARIABLE:
      NAMES =c1-c4 p1-p4;
MODEL:
      cig by c1-c4;
      pun by p1-p4;
      c1-c4 PWITH p1-p4;
OUTPUT: STDYX;
```

The PWITH operator allows a correlation between the error terms for C1 and P1, between C2 and P2, etc. This is plausible because the pairs share a common method. None of the correlations is statistically significant, however, probably because of the small sample size.



## Exercise 4

DATA:

```
FILE IS c:\data\asg1SEM.txt;  
TYPE IS COVA;  
NOBS IS 932;
```

VARIABLE:

```
NAMES ARE ano67 pow67 ano71 pow71 ed occ;
```

MODEL:

```
alien67 by ano67 pow67;  
alien71 by ano71 pow71;  
ses by ed occ;  
alien71 on alien67 ses;  
alien67 on ses;  
ano71 WITH ano67;
```

OUTPUT: STDYX MODINDICES;

This model has a p-value less than .0001. However, based on the modification indices, I tried the following model which yields a p-value above .25. This model is plausible, for the same argument for the correlated error model in assignment 3 was plausible.

DATA:

```
FILE IS c:\data\asg1SEM.txt;  
TYPE IS COVA;  
NOBS IS 932;
```

VARIABLE:

```
NAMES ARE ano67 pow67 ano71 pow71 ed occ;
```

MODEL:

```
alien67 by ano67 pow67;  
alien71 by ano71 pow71;  
ses by ed occ;  
alien71 on alien67 ses;  
alien67 on ses;  
ano71 WITH ano67;
```

OUTPUT: STDYX;

## SAS Code for Exercises

### Exercise 1

```
data anomie(type=cov);  
    infile 'c:\data\Asg1SEM.txt' missover;  
    _type_='cov';  
    input  anomie67 power67 anomie71 power71 ed occ;  
run;  
  
proc calis data=anomie cov nobs=932 effpart;  
    path  
        anomie71 <- anomie67 power67 ed occ,  
        power71  <- anomie71 anomie67 power67 ed occ;  
run;
```

## Exercise 2

```
proc calis data=anomie cov nobs=932;
  path
    anomie71 <- anomie67 power71 ed occ,
    power71 <- anomie71 power67 ed occ,
    power71 <-> anomie71;
run;
```

```
proc calis data=anomie cov nobs=932;
  path
    anomie71 <- anomie67 power71 ed occ,
    power71 <- anomie71 power67 ed occ;
run;
```

```
proc calis data=anomie cov nobs=932;
  path
    anomie71 <- anomie67 power71 power67 ed occ,
    power71 <- anomie71 power67 ed occ;
run;
```

```
proc calis data=anomie cov nobs=932;
  path
    anomie71 <- anomie67 power71 ed occ,
    power71 <- anomie71 power67 anomie67 ed occ;
run;
```

```
proc calis data=anomie cov nobs=932;
  path
    anomie71 <- anomie67 power71 power67 ed occ,
    power71 <- anomie71 power67 anomie67 ed occ;
run;
```

### Exercise 3

```
data cig_pun(type=corr);  
  infile 'c:\data\Asg3SEM.txt' missover;  
  _type_='corr';  
  input  c1-c4 p1-p4;  
run;  
  
proc calis data=cig_pun nobs=350;  
  path  
    cig -> c1-c4 = 1,  
    pun -> p1-p4 = 1;  
run;  
  
proc calis data=cig_pun nobs=350;  
  path  
    cig -> c1-c4 = 1,  
    pun -> p1-p4 = 1,  
    c1 <-> p1, c2 <-> p2, c3 <-> p3, c4 <-> p4;  
run;
```

## Exercise 4

```
proc calis data=anomie cov nobs=932;
  path
    alien67 -> anomie67 power67 = 1,
    alien71 -> anomie71 power71 = 1,
    ses -> ed occ = 1,
    alien71 <- alien67 ses,
    alien67 <- ses;
run;
```

```
proc calis data=anomie cov nobs=932;
  path
    alien67 -> anomie67 power67 = 1,
    alien71 -> anomie71 power71 = 1,
    ses -> ed occ = 1,
    alien71 <- alien67 ses,
    alien67 <- ses,
    anomie71 <-> anomie67;
run;
```