Structural Equation Models:
Path Analysis with Latent Variables

Latent Trait Measurement and
Structural Equation Models
Lecture #11

April 3, 2013

Today's Class

- Putting it all together:
  - Path Analysis
    - Observed variables
  - Confirmatory Factor Analysis / Measurement Models
    - Latent variables

- Concerns in building structural equation models
  - Model-predicted covariance matrices for path analysis with observed and latent variables

- Examples of SEM uses

UNDERLYING THEORY OF STRUCTURAL EQUATION MODELS

Structural Equation Models

- Although the term SEM can be applied to many settings, I view the label as being used to describe analyses with observed and latent variables

- A structural equation model consists of two "parts":
  - Measurement model(s) for each latent variable
  - Path analysis between the latent and observed variables

- Up to this point, we have covered both in isolation – today we put them together to show how the process works
  - You will see this extra step is pretty straight forward…
  - …but that added complexity becomes an issue when it comes to model fit

REVIEW OF PATH ANALYSIS

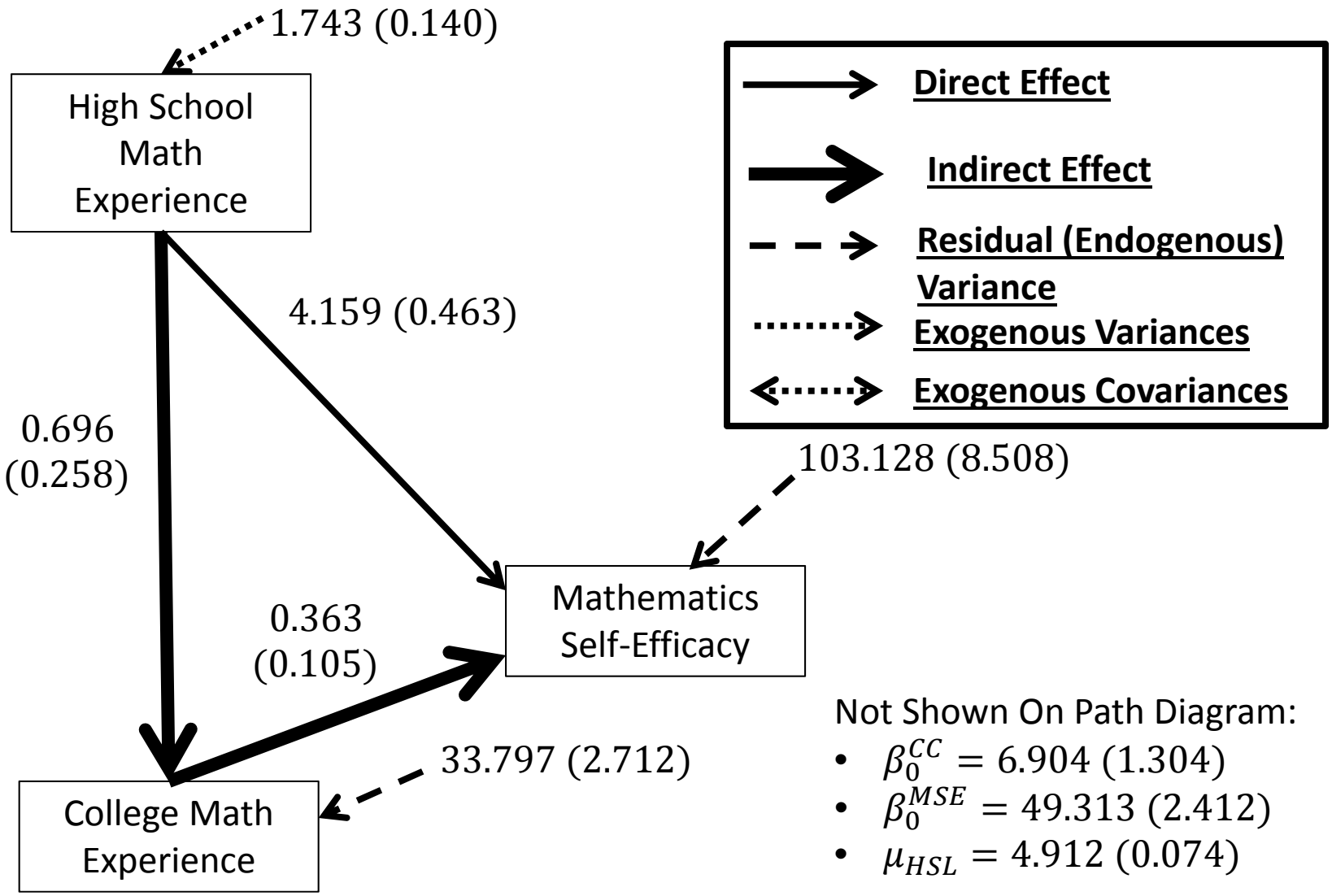
Types of Variables in the Analysis

- An important distinction in path analysis and SEM is between endogenous and exogenous variables

- Endogenous variable(s): variables whose variability ***is explained*** by one or more variables in a model
  - In linear regression, the **dependent variable** is the only endogenous variable in an analysis

- Exogenous variable(s): variables whose variability ***is not explained*** by any variables in a model
  - In linear regression, the **independent variable**(s) are the exogenous variables in the analysis

Direct and Indirect Effects of HSL on MSE



Not Shown On Path Diagram:

- $\beta_0^{CC} = 6.904\ (1.304)$
- $\beta_0^{MSE} = 49.313\ (2.412)$
- $\mu_{HSL} = 4.912\ (0.074)$

Path Analysis in Matrix Form

- Our path model simultaneous equations were:

$$CC_i = \beta_0^{CC} + \beta_{HSL}^{CC} HSL_i + e_i^{CC}$$
$$MSE_i = \beta_0^{MSE} + \beta_{CC}^{MSE} CC_i + \beta_{HSL}^{MSE} HSL_i + e_i^{MSE}$$

  - p =2 endogenous variables
  - q =1 exogenous variable

- Alternatively, we could rephrase this in matrix form:

$$\boldsymbol{y}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{y}_i + \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\zeta}_i$$

Where:

$\boldsymbol{x}_i = [HSL_i]$ (matrix of size $q\ x\ 1$ containing observed exogenous variables)

$\boldsymbol{y}_i = \begin{bmatrix} CC_i \\ MSE_i \end{bmatrix}$ (matrix of size $p\ x\ 1$ containing observed endogenous variables)

Then:

$\boldsymbol{\alpha} = \begin{bmatrix} \beta_0^{CC} \\ \beta_0^{MSE} \end{bmatrix}$ (matrix of size $p\ x\ 1$ containing intercepts for endogenous variables)

$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{CC}^{MSE} & 0 \end{bmatrix}$ (a $p\ x\ p$ matrix of coefficients relating the endogenous variables to themselves)

$\boldsymbol{\Gamma} = \begin{bmatrix} \beta_{HSL}^{CC} \\ \beta_{HSL}^{MSE} \end{bmatrix}$ (matrix of size $p\ x\ q$ relating exogenous variables to endogenous variable(s))

$\boldsymbol{\zeta}_i = \begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix} \sim N_2(\mathbf{0}, \boldsymbol{\Psi})$ (where $\boldsymbol{\Psi}$ is the $p\ x\ p$ residual covariance matrix)

Here, $\boldsymbol{\Psi}$ will be diagonal (no covariance) as we do not have any more degrees of freedom

Path Analysis in Matrix Form

- The equations from the previous slide are called the **structural form** of the path model

- Another form that exists in literature is the **reduced form**, where all endogenous variables are on the left-hand side

$$\boldsymbol{y}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{y}_i + \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\zeta}_i \leftrightarrow$$
$$\boldsymbol{y}_i - \mathbf{B}\boldsymbol{y}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\zeta}_i \leftrightarrow$$
$$(\mathbf{I} - \mathbf{B})\boldsymbol{y}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\zeta}_i \leftrightarrow$$
$$\boldsymbol{y}_i = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{x}_i + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}_i \leftrightarrow$$
$$\boldsymbol{y}_i = \boldsymbol{\Pi}_0 + \boldsymbol{\Pi}_1\boldsymbol{x}_i + \boldsymbol{\zeta}_i^*$$

Where $\boldsymbol{\zeta}_i^* \sim N_p(\mathbf{0}, \boldsymbol{\Psi}^*)$

- The reduced form is not as frequently used in practice, but does arise in some research areas and in identification

Path Analysis with Matrices

- Although not explained by our model, we could state that the mean vector of exogenous variables was:

$$\boldsymbol{\mu}_x = [\mu_{HSL}]$$

- Likewise, we can state that the covariance matrix of the exogenous variables is

$$\boldsymbol{\Phi} = [\sigma^2_{HSL}]$$

- We will use these terms in our matrix-version of the model predicted mean and covariance matrix

Model Predicted Mean Vector and Covariance Matrix

- The unconditional mean of the endogenous variables is:

$$\hat{\boldsymbol{\mu}}_y = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\mu}_x$$

- The covariance matrix of the exogenous and endogenous variables is then:

$$\boldsymbol{\Sigma}_{y,x} = \begin{bmatrix} \text{Y only} & \text{Y with X} \\ \text{X with Y} & \text{X only} \end{bmatrix} = \begin{bmatrix} (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi})(\mathbf{I} - \mathbf{B})^{T^{-1}} & (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}^T(\mathbf{I} - \mathbf{B})^{T^{-1}} & \boldsymbol{\Phi} \end{bmatrix}$$

- The point: that model specifications have direct implications for the parameters of the multivariate normal distribution

Matching Matrices with Results

- To more specifically link our results to the matrices from the previous page:

Name	Matrix	Model Estimates
Residual Covariance Matrix	$\boldsymbol{\Psi}$	$\begin{bmatrix} 33.797 & 0 \\ 0 & 103.128 \end{bmatrix}$
Regression Weights of Exogenous onto Endogenous	$\boldsymbol{\Gamma}$	$\begin{bmatrix} 0.696 \\ 4.159 \end{bmatrix}$
Covariance Matrix of Exogenous Variables	$\boldsymbol{\Phi}$	$[1.743]$
Mean Vector of Exogenous Variables	$\boldsymbol{\mu}_x$	$[4.912]$
Vector of Endogenous Variable Intercepts	$\boldsymbol{\alpha}$	$\begin{bmatrix} 6.904 \\ 49.313 \end{bmatrix}$
Matrix of Endogenous Regression Weights	$\mathbf{B}$	$\begin{bmatrix} 0 & 0 \\ 0.363 & 0 \end{bmatrix}$
Inverse matrix used in calculations	$(\mathbf{I} - \mathbf{B})^{-1}$	$\begin{bmatrix} 1 & 0 \\ -0.363 & 1 \end{bmatrix}$

Model Predicted Mean Vector and Covariance Matrix

- The estimated conditional mean of the endogenous variables is:

Model Estimated Means/Intercepts/Thresholds

	CC	MSE	HSL
1	10.322	73.495	4.912

Residuals for Means/Intercepts/Thresholds

	CC	MSE	HSL
1	0.000	0.000	0.000

  - These values correspond exactly (saturated model)
- The estimated covariance matrix of the exogenous and endogenous variables is:

Model Estimated Covariances/Correlations/Residual Correlations

	CC	MSE	HSL
CC	34.641		
MSE	17.629	141.526	
HSL	1.213	7.692	1.743

- These are mostly exact – small differences

Residuals for Covariances/Correlations/Residual Correlations

	CC	MSE	HSL
CC	0.000		
MSE	-0.002	-0.018	
HSL	0.000	-0.001	0.000

REVIEW OF CONFIRMATORY FACTOR ANALYSIS

One-Factor Model of Five GRI Items

- The CFA model for the five GRI items:

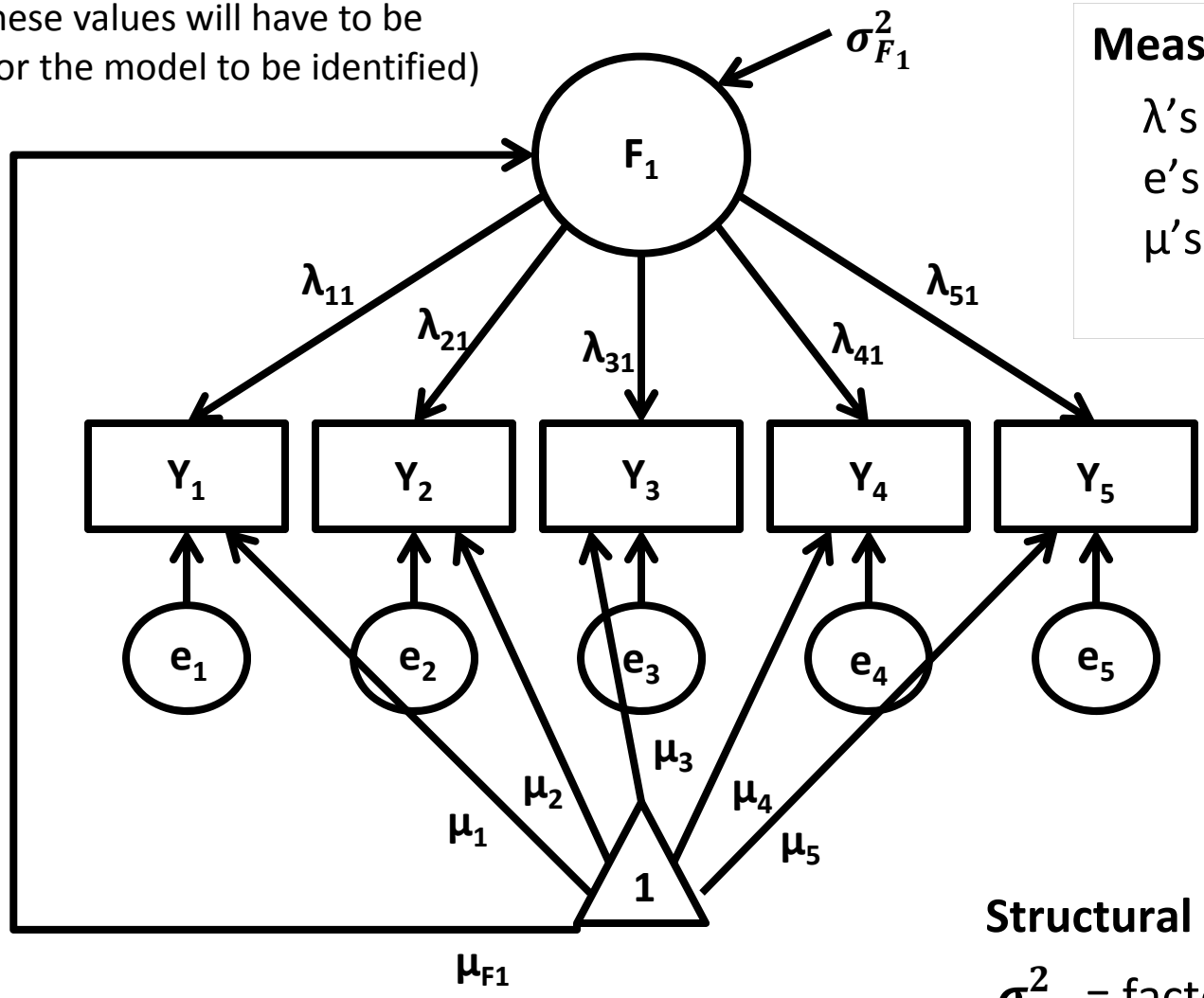
$$Y_{s1} = \mu_1 + \lambda_{11}F_{s1} + e_{s1}$$
$$Y_{s2} = \mu_2 + \lambda_{21}F_{s1} + e_{s2}$$
$$Y_{s3} = \mu_3 + \lambda_{31}F_{s1} + e_{s3}$$
$$Y_{s4} = \mu_4 + \lambda_{41}F_{s1} + e_{s4}$$
$$Y_{s5} = \mu_5 + \lambda_{51}F_{s1} + e_{s5}$$

- Here:
  - Y_{si} - response of subject s on item i
  - μ_i - intercept of item i (listed as a mean as this is typically what it becomes)
  - λ_{i1} - factor loading of item i on factor 1 (only one factor today)
  - F_{s1} - latent "factor score" for subject s (same for all items) to factor 1 (only one today)
  - e_{si} - regression-like residual for subject s on item i
    - We assume $e_{si} \sim N(0, \psi_i^2)$; ψ_i^2 is called the **unique variance** of item i
    - We also assume e_{si} and F_{s1} are independent

- Also, we will assume $F_{s1} \sim N(\mu_{F_1}, \sigma_{F_1}^2)$
  - Typically $\mu_{F_1} = 0$ (but not always)
  - Factor variance can be estimated or fixed (more on both in identification)

Our CFA Model Path Diagram

(Some of these values will have to be restricted for the model to be identified)



Measurement Model:

λ's = factor loadings
e's = error variances
μ's = item intercepts

Structural Model:

$\sigma^2_{F_1}$ = factor variance
μ_{F1} = factor mean

Model Predicted Mean Vector

- Combining across all items, the mean vector for the items is given by:

$$\boldsymbol{\mu}_Y = \boldsymbol{\mu}_I + \boldsymbol{\Lambda}\boldsymbol{\mu}_{\boldsymbol{F}}$$

$$\begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \\ \mu_{Y_3} \\ \mu_{Y_4} \\ \mu_{Y_5} \end{bmatrix} = \begin{bmatrix} \mu_{I_1} \\ \mu_{I_2} \\ \mu_{I_3} \\ \mu_{I_4} \\ \mu_{I_5} \end{bmatrix} + \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \\ \lambda_{51} \end{bmatrix} \begin{bmatrix} \mu_{F_1} \end{bmatrix} = \begin{bmatrix} \mu_{I_1} + \lambda_{11}\mu_{F_1} \\ \mu_{I_2} + \lambda_{21}\mu_{F_1} \\ \mu_{I_3} + \lambda_{31}\mu_{F_1} \\ \mu_{I_4} + \lambda_{41}\mu_{F_1} \\ \mu_{I_5} + \lambda_{51}\mu_{F_1} \end{bmatrix}$$

Model Implied Covariance Matrix

- Combining across all items, the covariance matrix for the items is given by:

$$\boldsymbol{\Sigma}_Y = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$$

  - Get used to seeing this – although you already have (see the regression slides)

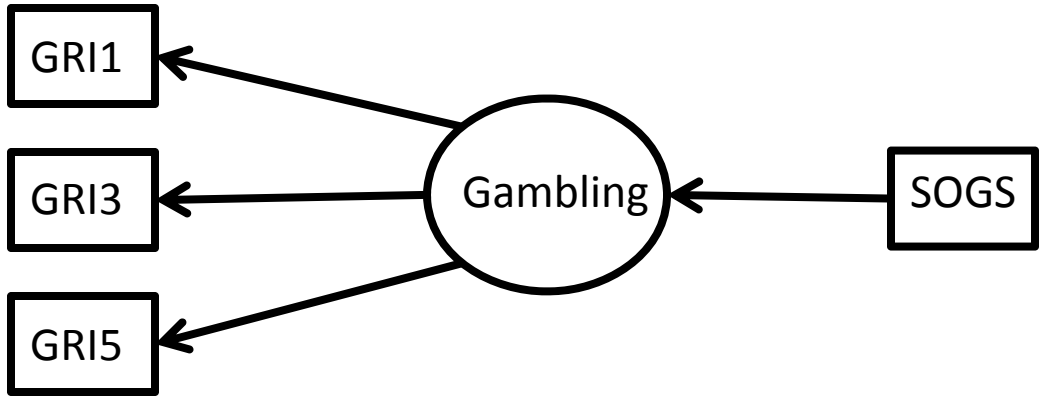
$$\begin{bmatrix} \sigma^2_{Y_1} & \sigma_{Y_1,Y_2} & \sigma_{Y_1,Y_3} & \sigma_{Y_1,Y_4} & \sigma_{Y_1,Y_5} \\ \sigma_{Y_1,Y_2} & \sigma^2_{Y_2} & \sigma_{Y_2,Y_3} & \sigma_{Y_2,Y_4} & \sigma_{Y_2,Y_5} \\ \sigma_{Y_1,Y_3} & \sigma_{Y_2,Y_3} & \sigma^2_{Y_3} & \sigma_{Y_3,Y_4} & \sigma_{Y_3,Y_5} \\ \sigma_{Y_1,Y_4} & \sigma_{Y_2,Y_4} & \sigma_{Y_3,Y_4} & \sigma^2_{Y_4} & \sigma_{Y_4,Y_5} \\ \sigma_{Y_1,Y_5} & \sigma_{Y_2,Y_5} & \sigma_{Y_3,Y_5} & \sigma_{Y_4,Y_5} & \sigma^2_{Y_5} \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \\ \lambda_{51} \end{bmatrix} \left[\sigma^2_{F_1}\right] [\lambda_{11} \quad \lambda_{21} \quad \lambda_{31} \quad \lambda_{41} \quad \lambda_{51}] + \begin{bmatrix} \psi^2_1 & 0 & 0 & 0 & 0 \\ 0 & \psi^2_2 & 0 & 0 & 0 \\ 0 & 0 & \psi^2_3 & 0 & 0 \\ 0 & 0 & 0 & \psi^2_4 & 0 \\ 0 & 0 & 0 & 0 & \psi^2_5 \end{bmatrix} =$$

PUTTING IT TOGETHER: PATH ANALYSIS WITH LATENT VARIABLES

A Small SEM Example

- To demonstrate how SEM works, we will use a very small example:
  - Measurement model: three GRI items forming one latent construct ("gambling")
    - Note: with three items, the measurement model is just-identified (meaning perfect fit)
  - Path model: The prediction of "gambling" by the SOG score
    - Note: here we treat SOGS score as being observed without error
      - The reason: the SOGS items are all binary indicators (0/1)…they won't work with an assumption of MVN
      - A better solution: model the SOGS items with a logit link function (called IRT/IFA) – covered if we have time



Step #1: Building the Measurement Model

- The first step in a structural equation model is to build the measurement model
  - Here, the measurement model is simplified so as to show how SEM works

```
TITLE:
    Gambling Research Instrument Items
    Data from 1192 College Students/144 Gamblers
    41 Likert Items (1-6): GRI1-GRI41
    12 SOGS items (SOGS4-SOGS15), mostly dichotomous
    ============================================================
    Identification: Marker Item Factor Variance, Zero Factor Mean
    ============================================================
    One-Factor GAMBLING tendencies model with 3 GRI items
    SEM with SOGS score predicting GAMBLING

DATA:
    FILE = gamblingdata.csv;

VARIABLE:
    NAMES = GRI1-GRI41 SOGS4-SOGS15 Student ID;
    USEVARIABLES = GRI1 GRI3 GRI5;
    IDVARIABLE = ID;
    MISSING = ALL(99);

DEFINE:
    SOGSsum = MEAN(SOGS4-SOGS15);

ANALYSIS:
    ESTIMATOR = MLR;

MODEL:
    GAMBLING by GRI1 GRI3 GRI5;

OUTPUT:
    STANDARDIZED MODINDICES(ALL 0) RESIDUAL;
```

Measurement Model Fit Assessment

- Our three-item measurement model fits perfectly

```
MODEL FIT INFORMATION

Number of Free Parameters                        9

Loglikelihood

          H0 Value                       -5254.609
          H0 Scaling Correction Factor       2.236
            for MLR
          H1 Value                       -5254.609
          H1 Scaling Correction Factor       2.236
            for MLR
```

```
Chi-Square Test of Model Fit

          Value                             0.000
          Degrees of Freedom                    0
          P-Value                           0.0000

RMSEA (Root Mean Square Error Of Approximation)

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

CFI/TLI

          CFI                               1.000
          TLI                               1.000
```

```
SRMR (Standardized Root Mean Square Residual)

          Value                             0.000
```

Measurement Model Parameter Estimates

MODEL RESULTS

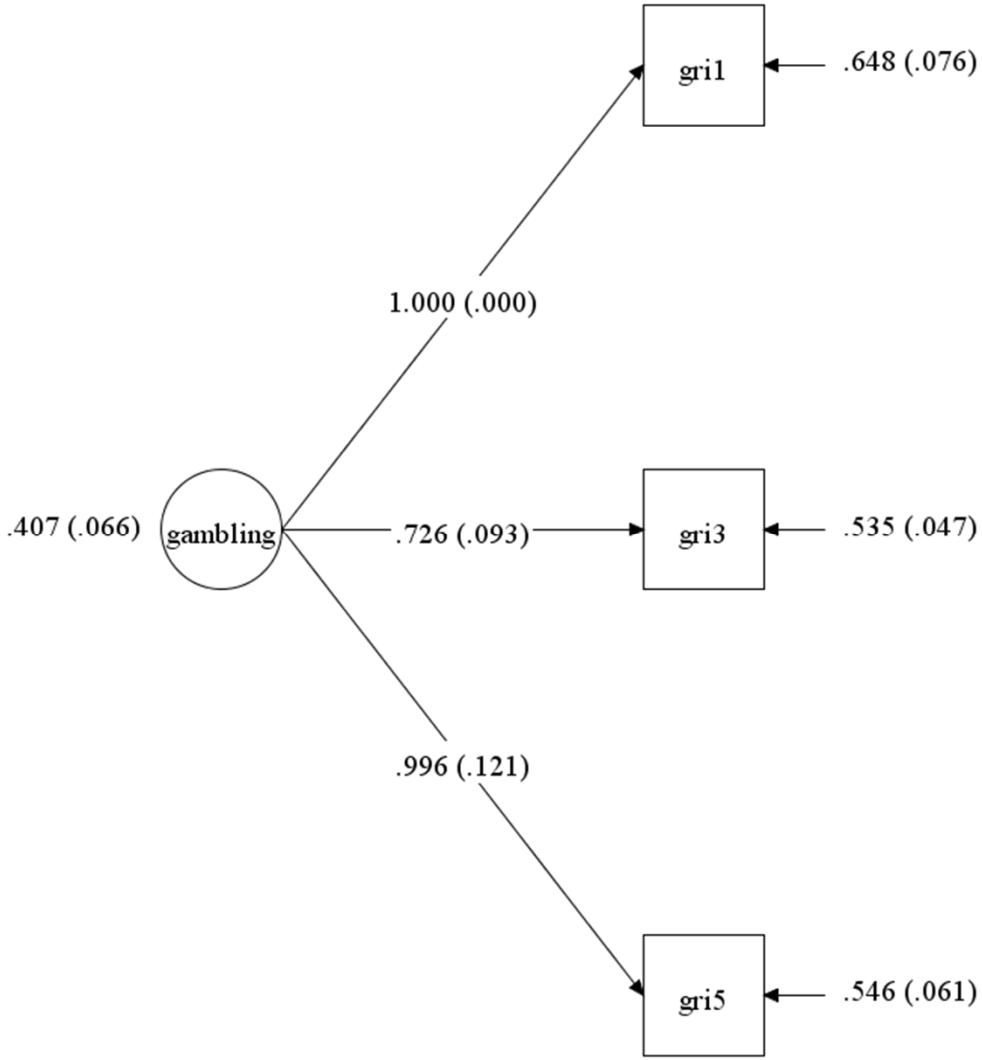
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GAMBLING BY				
GRI1	1.000	0.000	999.000	999.000
GRI3	0.726	0.062	11.792	0.000
GRI5	0.996	0.087	11.397	0.000
Intercepts				
GRI1	1.823	0.028	64.873	0.000
GRI3	1.548	0.024	65.364	0.000
GRI5	1.593	0.027	59.747	0.000
Variances				
GAMBLING	0.407	0.047	8.647	0.000
Residual Variances				
GRI1	0.648	0.042	15.246	0.000
GRI3	0.535	0.027	19.459	0.000
GRI5	0.546	0.040	13.631	0.000

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GAMBLING BY				
GRI1	0.621	0.031	19.963	0.000
GRI3	0.535	0.030	17.942	0.000
GRI5	0.652	0.032	20.596	0.000
Intercepts				
GRI1	1.775	0.044	40.426	0.000
GRI3	1.788	0.044	40.545	0.000
GRI5	1.635	0.042	39.090	0.000
Variances				
GAMBLING	1.000	0.000	999.000	999.000
Residual Variances				
GRI1	0.614	0.039	15.888	0.000
GRI3	0.714	0.032	22.354	0.000
GRI5	0.575	0.041	13.925	0.000

Measurement Model Path Diagram



Measurement Model: Implied Covariance Matrix

- The measurement model implied covariance matrix is:

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}_y \boldsymbol{\Phi} \boldsymbol{\Lambda}_y^T + \boldsymbol{\Psi}_y$$

$$= \begin{bmatrix} 1.000 \\ 0.726 \\ 0.996 \end{bmatrix} [0.407] [1.000 \quad 0.726 \quad 0.996]$$

$$+ \begin{bmatrix} 0.648 & 0 & 0 \\ 0 & 0.535 & 0 \\ 0 & 0 & 0.546 \end{bmatrix} \begin{bmatrix} 1.055 & 0.295 & 0.405 \\ 0.295 & 0.749 & 0.294 \\ 0.405 & 0.297 & 0.950 \end{bmatrix}$$

Model Estimated Covariances/Correlations/Residual Correlations

	GRI1	GRI3	GRI5
GRI1	1.055		
GRI3	0.295	0.749	
GRI5	0.405	0.294	0.950

Residuals for Covariances/Correlations/Residual Correlations

	GRI1	GRI3	GRI5
GRI1	0.000		
GRI3	0.000	0.000	
GRI5	0.000	0.000	0.000

Step 2: Estimating the Structural Equation Model

- Once the measurement model is found to fit, the next step is to estimate the full structural equation model

```
DEFINE:
    SOGSsum = MEAN(SOGS4-SOGS15);

MODEL:
    GAMBLING by GRI1 GRI3 GRI5;

    GAMBLING on SOGSsum;
```

- SOGSsum is treated as an **exogenous variable**
  - Also called an independent variable

- GAMBLING (and the items measuring it) are treated as **endogenous variables**
  - Also called dependent variables

SEM: Model Identification

- As SEM integrates both measurement and path models, the identification rules for SEM borrow from both
  - The measurement model (for all latent variables) must be locally identified
    - Including rules for setting scale of latent factor(s)
  - The path model must be identified

- A necessary but not sufficient way of ensuring identification is the t-rule (counting rule)
  - The number of parameters must be less than the total number of means + variances/covariances of **all** observed variables in the analysis

- Number of observed variables in our analysis: 4
  - Number of variances/covariances: 4*(4+1)/2 = 10
  - Number of means: 4
  - Total: 14

- Number of parameters in our analysis
  - 2 factor loadings + 1 factor variance + 3 unique variances + 1 direct effect + 3 item intercepts + 1 exogenous variance = 12

SEM: Mplus Syntax

- The Mplus syntax is a combination of path and measurement models

```
VARIABLE:
    NAMES = GRI1-GRI41 SOGS4-SOGS15 Student ID;
    USEVARIABLES = GRI1 GRI3 GRI5 SOGSsum;
    IDVARIABLE = ID;
    MISSING = ALL(99);

DEFINE:
    SOGSsum = MEAN(SOGS4-SOGS15);

ANALYSIS:
    ESTIMATOR = MLR;

MODEL:
    GAMBLING by GRI1 GRI3 GRI5;

    GAMBLING on SOGSsum;
    SOGSsum;
```

SEM: Model Fit Assessment

- We have fewer parameters than the total possible → we must now assess our model fit

```
MODEL FIT INFORMATION

Number of Free Parameters                       12

Loglikelihood

          H0 Value                          -4578.581
          H0 Scaling Correction Factor          2.383
            for MLR
          H1 Value                          -4577.428
          H1 Scaling Correction Factor          2.271
            for MLR
```

```
Chi-Square Test of Model Fit

          Value                              1.444*
          Degrees of Freedom                     2
          P-Value                           0.4859
          Scaling Correction Factor          1.597
            for MLR
```

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

          Estimate                           0.000
          90 Percent C.I.                    0.000  0.049
          Probability RMSEA <= .05           0.954

CFI/TLI

          CFI                                1.000
          TLI                                1.005
```

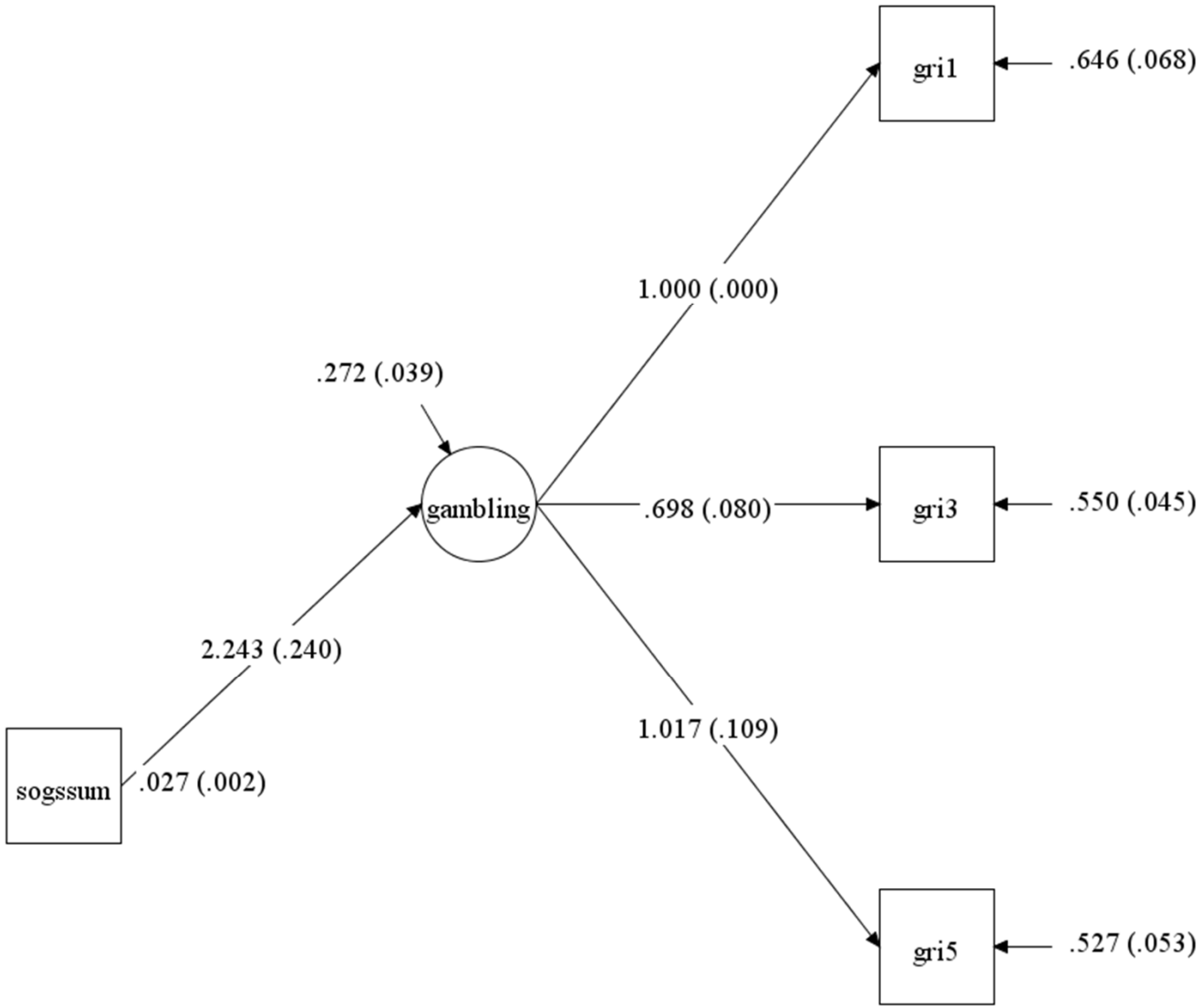
SEM: Model Parameter Output

- Note: our measurement model parameters have changed slightly
  - More on why in a moment

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GAMBLING BY				
GRI1	1.000	0.000	999.000	999.000
GRI3	0.698	0.080	8.763	0.000
GRI5	1.017	0.109	9.320	0.000
GAMBLING ON				
SOGSSUM	2.243	0.240	9.354	0.000
Means				
SOGSSUM	0.416	0.005	92.474	0.000
Intercepts				
GRI1	0.889	0.096	9.278	0.000
GRI3	0.896	0.066	13.586	0.000
GRI5	0.643	0.092	7.007	0.000
Variances				
SOGSSUM	0.027	0.002	11.749	0.000
Residual Variances				
GRI1	0.646	0.068	9.519	0.000
GRI3	0.550	0.045	12.117	0.000
GRI5	0.527	0.053	9.897	0.000
GAMBLING	0.272	0.039	7.046	0.000

SEM Model Path Diagram



SEM: Standardized Model Parameters

- Here, we see that the SOGS has a correlation of .577 with the GAMBLING latent variable

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GAMBLING BY				
GRI1	0.622	0.041	15.032	0.000
GRI3	0.516	0.042	12.401	0.000
GRI5	0.667	0.039	17.298	0.000
GAMBLING ON				
SOGSSUM	0.577	0.040	14.323	0.000
Means				
SOGSSUM	2.530	0.086	29.372	0.000
Intercepts				
GRI1	0.865	0.104	8.351	0.000
GRI3	1.035	0.095	10.917	0.000
GRI5	0.660	0.106	6.217	0.000
Variances				
SOGSSUM	1.000	0.000	999.000	999.000
Residual Variances				
GRI1	0.613	0.052	11.885	0.000
GRI3	0.734	0.043	17.109	0.000
GRI5	0.555	0.051	10.794	0.000
GAMBLING	0.667	0.047	14.316	0.000

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GRI1	0.387	0.052	7.516	0.000
GRI3	0.266	0.043	6.200	0.000
GRI5	0.445	0.051	8.649	0.000

Latent Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GAMBLING	0.333	0.047	7.161	0.000

SEM: Model Implied Covariance Matrices

- Notice our model-implied covariance matrix:

Model Estimated Covariances/Correlations/Residual Correlations

	GRI1	GRI3	GRI5	SOGSSUM
GRI1	1.055			
GRI3	0.285	0.749		
GRI5	0.415	0.290	0.949	
SOGSSUM	0.061	0.042	0.062	0.027

- And the residuals from the saturated model:

Residuals for Covariances/Correlations/Residual Correlations

	GRI1	GRI3	GRI5	SOGSSUM
GRI1	0.000			
GRI3	0.010	0.000		
GRI5	-0.010	0.004	0.000	
SOGSSUM	0.001	-0.003	0.001	0.000

- NEW WRINKLES:
  - MEASUREMENT MODEL DOES NOT FIT SATURATED MODEL PERFECTLY
  - OFF-DIAGONAL OF BLOCK CAN CAUSE MODEL MISFIT

Equation Form of Overall Structural Equation Model

- Our structural equation model simultaneous equations were

For the "measurement" portion:

$$GRI1_s = \mu_{I_1} + \lambda_{11} GAMBLING_{s1} + e_{s1}$$
$$GRI3_s = \mu_{I_3} + \lambda_{31} GAMBLING_{s1} + e_{s3}$$
$$GRI5_s = \mu_{I_5} + \lambda_{51} GAMBLING_{s1} + e_{s5}$$

For the "structural" portion:

$$GAMBLING_{s1} = \beta_0^{\text{GAMBLING}} + \beta_1^{SOGS} SOGS_s + \delta_{gs}$$

- $p = 3$ endogenous variables
- $q = 1$ exogenous variable

Matrix Form of Structural Equation Model

- In matrices, the measurement portion of the model is given by:

$$\mathbf{Y}_s = \boldsymbol{\mu}_I + \boldsymbol{\Lambda}\boldsymbol{\eta}_s + \mathbf{K}\mathbf{X}_s + \boldsymbol{e}_s$$

with $\boldsymbol{e}_s \sim N(0, \boldsymbol{\Psi}_y)$

- Further, the structural portion of the model is given by:

$$\boldsymbol{\eta}_s = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_s + \boldsymbol{\Gamma}\mathbf{X}_s + \boldsymbol{\zeta}_s$$

with $\boldsymbol{\zeta}_s \sim N(0, \boldsymbol{\Theta})$

- In terms of our model:
  - There are no direct exogenous predictors of our endogenous measurement model parameters (so $\mathbf{K} = \mathbf{0}$)
  - There are no direct predictors of our endogenous latent variables by other latent variables (so $\mathbf{B} = \mathbf{0}$)
  - We standardized our factor mean to zero (so $\boldsymbol{\alpha} = \mathbf{0}$)

Putting Values into Matrices: Measurement Model

- $\boldsymbol{\mu}_i = \begin{bmatrix} 0.889 \\ 0.896 \\ 0.643 \end{bmatrix}$ - item intercepts
- $\boldsymbol{\Lambda} = \begin{bmatrix} 1.000 \\ 0.698 \\ 1.017 \end{bmatrix}$ - factor loadings for endogenous variables
- $\boldsymbol{\Psi}_y = \begin{bmatrix} 0.646 & 0 & 0 \\ 0 & 0.550 & 0 \\ 0 & 0 & 0.527 \end{bmatrix}$ - unique variances of endogenous variables

Putting Values into Matrices: Structural Model

- $\boldsymbol{\Gamma} = [2.243]$ - direct regression coefficient of exogenous variable onto endogenous factor
- $\boldsymbol{\Theta} = [0.272]$ - residual variance of endogenous factor
- $\boldsymbol{\Phi} = [0.027]$ - variance/covariance matrix for the exogenous variables

Model Implied Covariance Matrix

- The covariance matrix of the exogenous and endogenous variables is then:

$$\boldsymbol{\Sigma}_{y,x} = \begin{bmatrix} \text{Y only} & \text{Y with X} \\ \text{X with Y} & \text{X only} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_y(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Theta})\boldsymbol{\Lambda}_y^T + \boldsymbol{\Psi}_y & \boldsymbol{\Lambda}_y\boldsymbol{\Gamma}\boldsymbol{\Theta} \\ \boldsymbol{\Theta}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}_y^T & \boldsymbol{\Phi} \end{bmatrix}$$

- The point: the structural equation model can have significant model misfit due to both the measurement model and the structural model

Issues in Building Structural Equation Models

- Because of the multiple ways SEMs can exhibit model misfit, the process of building SEMs can be difficult

- In general, current practice states that measurement models should be built first – then the full SEM

- Some researchers offer questionable advice:
  - Use only just-identified measurement models
    - Why: fewer degrees of freedom where misfit can happen
    - Bad idea: poor reliability for latent constructs
  - Build measurement models with SEMs simultaneously
    - Why: full calibration can lead to better overall model fit
    - Bad idea: measurement should happen in absence of exogenous variables
  - Use two-stage analyses for SEMs
    - Why: measurement model then cannot change
    - Bad idea: propagation of measurement error for some factor score methods

SEM IN PRACTICE: EXAMPLES FROM REAL WORLD ANALYSES

SEM in Practice

- To demonstrate the practical side of building structural equation models, I will go over a couple examples from real data analyses

- In these examples, the model-building process will be discussed, along with varying methods for analysis

- The data for these examples is not available – but the practice should show how decisions are made about how SEMs are constructed and interpreted

Example #1: Evaluation of Academic Progress

- This example comes from data from a large southeastern university

- Data include:
  - PRE: scores on a pretest of mathematics ability, administered to students when they arrive at the university
    - Scores are from total number correct – alpha reliability of .81
  - POST: scores on a posttest of mathematics ability (using the same items), administered to students after two years at the university
    - Scores are from total number correct – alpha reliability of .81
  - Course Enrollments:
    - If a student had enrolled in one of 29 courses related to math and science education at the university
      - Data are binary – 0 = did not enroll; 1 = enrolled

Example #1: Research Questions

- The evaluation sought to answer the following questions:
  - Did scores improve on the posttest when compared with the pretest?
  - Did coursework significantly affect the posttest scores?
  - Did the score on the pretest predict the coursework students took?
  - Did coursework mediate the relationship between pretest and posttest?

Building the SEM: Modeling Issues

- Because of the nature of the data, several modeling issues must be considered when using SEM to answer the research questions

- Because pretest and posttest are sum-scores (with a known reliability), each can be used as a single indicator
  - In this case, the posttest single indicator will be problematic because of the residual variance (after prediction) is less than the overall variance
    - So must put single indicator model in last

- Each of the courses is binary (dichotomous), so including them in the model directly is not an option
  - Model would treat them as normally distributed if not categorical
    - Software won't allow categorical mediators
  - Could use them as:
    - Counts for specific categories (then treat count as approximately normal)
      - What we did
    - Indicators of a coursework factor
      - Hard to envision

Modeling Strategy

- Courses:
  - Create counts of each course category (3 categories total)
  - Treat counts as approximately normal (and use MLR)
  - Use all variables in a path model where:
    - Pretest predicts course counts and posttest score
    - Course counts predict posttest score
  - Treat pretest and posttest as single indicators where variance of each is weighted by the .81 reliability of each
    - Final step in the analysis

Initial Syntax: For Descriptive Statistics

```
VARIABLE:
    NAMES = ID Pre Post PostEff PostImp G1_M103 G1_M105
            G1_M107 G1_M205 G1_M220 G1_M231 G1_M235
            G2_C120 G2_C131 G2_G112 G2_G101 G2_G121
            G2_P140 G2_P215 G2_P240 G3_B114 G3_B270 G3_G196
            G3_G103 G3_G110 G3_G200 G3_G211 G3_G102
            G3_G113 G3_G122 G3_G115 G3_A120 G3_A121
            G4_G104;

    USEVARIABLE =  Pre Post G1_SUM G2_SUM G3_SUM;

    IDVARIABLE = ID;
    MISSING = .;

DEFINE:
    G1_SUM = SUM(G1_M103 G1_M105 G1_M107 G1_M205 G1_M220 G1_M231 G1_M235);
    G2_SUM = SUM(G2_C120 G2_C131 G2_G112 G2_G101 G2_G121 G2_P140 G2_P215 G2_P240);
    G3_SUM = SUM(G3_B114 G3_B270 G3_G196 G3_G103 G3_G110 G3_G200 G3_G211 G3_G102
            G3_G113 G3_G122 G3_G115 G3_A120 G3_A121 G4_G104);

ANALYSIS:
    ESTIMATOR = MLR;
```

Initial Output: Descriptive Statistics

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Means				
PRE	46.226	0.278	166.563	0.000
POST	49.264	0.307	160.283	0.000
G1_SUM	1.073	0.021	52.167	0.000
G2_SUM	0.385	0.026	14.998	0.000
G3_SUM	0.513	0.026	20.065	0.000
Variances				
PRE	40.206	2.788	14.421	0.000
POST	49.313	4.199	11.744	0.000
G1_SUM	0.221	0.018	12.484	0.000
G2_SUM	0.344	0.024	14.331	0.000
G3_SUM	0.342	0.018	19.030	0.000

Model #1: Path Model w/o Posttest Single Indicator

- The Mplus syntax:

```
MODEL:
    PRETEST BY PRE@1;
    PRE (varPRE);

    G1_SUM ON PRETEST;
    G2_SUM ON PRETEST;
    G3_SUM ON PRETEST;
    POST ON G1_SUM G2_SUM G3_SUM PRETEST;
```

- Model fit:

```
Chi-Square Test of Model Fit

          Value                              6.026*
          Degrees of Freedom                     3
          P-Value                           0.1104
          Scaling Correction Factor          1.063
            for MLR

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.044
          90 Percent C.I.                    0.000  0.095
          Probability RMSEA <= .05           0.497

CFI/TLI

          CFI                                0.982
          TLI                                0.939

SRMR (Standardized Root Mean Square Residual)

          Value                              0.025
```

Model #1: Relevant Output

- For building a single indicator out of posttest:

Residual Variances				
PRE	7.639	0.000	999.000	999.000
POST	30.586	3.847	7.950	0.000
G1_SUM	0.218	0.017	12.574	0.000
G2_SUM	0.344	0.024	14.390	0.000
G3_SUM	0.342	0.018	19.116	0.000

Model #2: Pre/Post Single Indicators

- Mplus Syntax:

```
MODEL:
    PRETEST BY PRE@1;
    POSTTEST BY POST@1;

    PRE (varPRE);
    POST (varPOST);

    G1_SUM ON PRETEST;
    G2_SUM ON PRETEST;
    G3_SUM ON PRETEST;
    POSTTEST ON G1_SUM G2_SUM G3_SUM PRETEST;

MODEL CONSTRAINT:
    varPRE = (1-.81)*40.206;
    varPOST = (1-.81)*30.586;

MODEL INDIRECT:
    POSTTEST IND PRETEST;
```

Model #2: Model Fit Assessment

- Mplus Output:

Chi-Square Test of Model Fit

Value	6.026*
Degrees of Freedom	3
P-Value	0.1104
Scaling Correction Factor for MLR	1.063

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.044	
90 Percent C.I.	0.000	0.095
Probability RMSEA <= .05	0.497	

CFI/TLI

CFI	0.982
TLI	0.939

SRMR (Standardized Root Mean Square Residual)

Value	0.025

- Need for residual covariances between coursework sums

Normalized Residuals for Covariances/Correlations/Residual Correlations

	PRE	POST	G1_SUM	G2_SUM	G3_SUM
PRE	0.000				
POST	0.000	0.000			
G1_SUM	0.005	-0.009	0.000		
G2_SUM	0.020	-0.019	0.426	0.000	
G3_SUM	0.019	0.019	1.370	1.977	0.000

Model #3: Single Indicators with Residual Covariances

- Mplus syntax:

```
MODEL:
    PRETEST BY PRE@1;
    POSTTEST BY POST@1;

    PRE (varPRE);
    POST (varPOST);

    G1_SUM ON PRETEST;
    G2_SUM ON PRETEST;
    G3_SUM ON PRETEST;
    POSTTEST ON G1_SUM G2_SUM G3_SUM PRETEST;

    G1_SUM G2_SUM G3_SUM WITH G1_SUM G2_SUM G3_SUM;

MODEL CONSTRAINT:
    varPRE = (1-.81)*40.206;
    varPOST = (1-.81)*30.586;

MODEL INDIRECT:
    POSTTEST IND PRETEST;
```

- Note: this model has no degrees of freedom left – it is just-identified
  - Therefore model fit is perfect

Model #3: Results

MODEL RESULTS

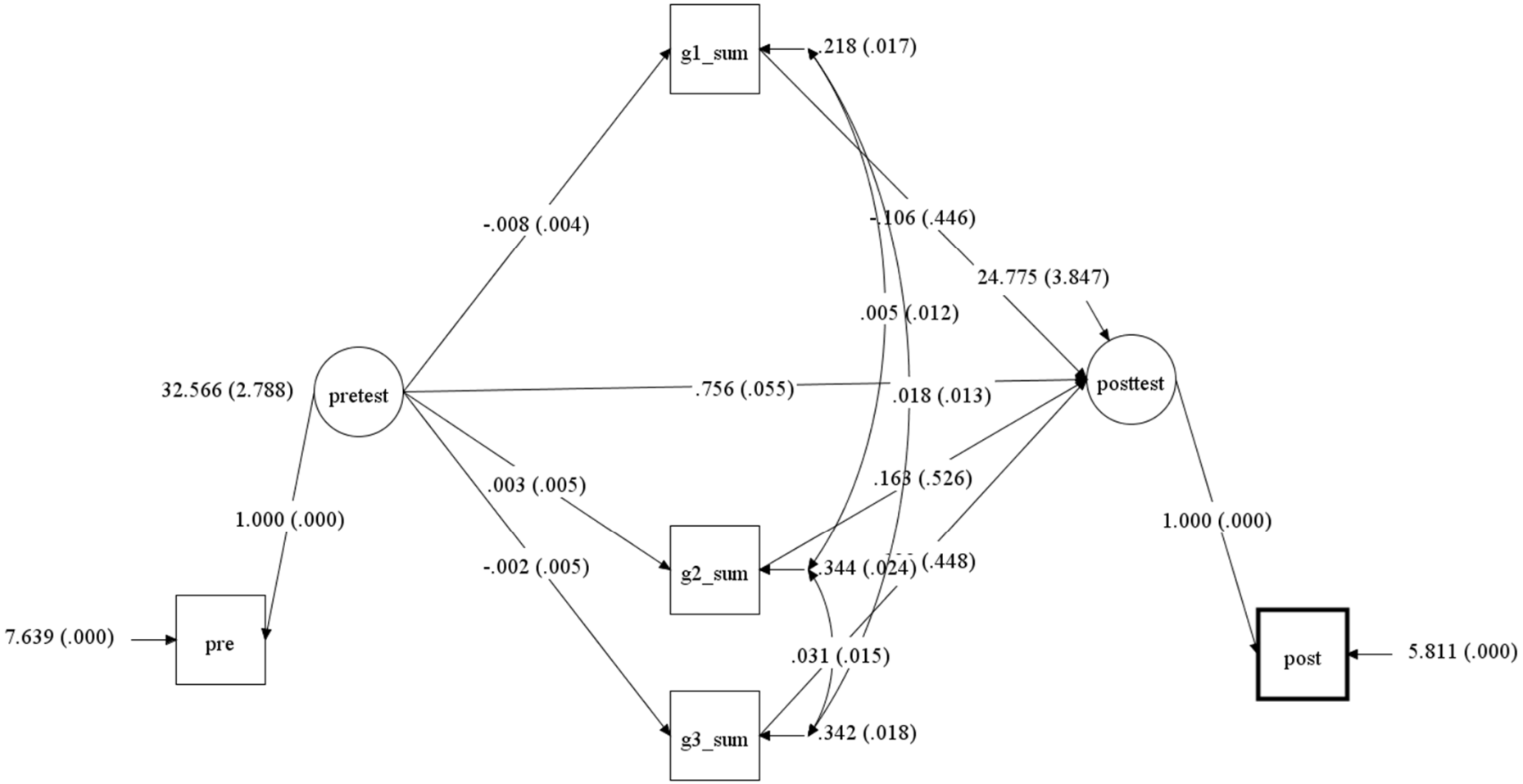
	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PRETEST BY				
PRE	1.000	0.000	999.000	999.000
POSTTEST BY				
POST	1.000	0.000	999.000	999.000
POSTTEST ON				
PRETEST	0.756	0.055	13.839	0.000
POSTTEST ON				
G1_SUM	-0.106	0.446	-0.237	0.813
G2_SUM	0.163	0.526	0.310	0.756
G3_SUM	-0.123	0.448	-0.275	0.784
Residual Variances				
PRE	7.639	0.000	999.000	999.000
POST	5.811	0.000	999.000	999.000
G1_SUM	0.218	0.017	12.571	0.000
G2_SUM	0.344	0.024	14.392	0.000
G3_SUM	0.342	0.018	19.111	0.000
POSTTEST	24.775	3.847	6.440	0.000
G1_SUM ON				
PRETEST	-0.008	0.004	-2.139	0.032
G2_SUM ON				
PRETEST	0.003	0.005	0.597	0.551
G3_SUM ON				
PRETEST	-0.002	0.005	-0.498	0.619
G1_SUM WITH				
G2_SUM	0.005	0.012	0.432	0.665
G3_SUM	0.018	0.013	1.384	0.166
G2_SUM WITH				
G3_SUM	0.031	0.015	1.984	0.047
Intercepts				
PRE	46.226	0.278	166.563	0.000
POST	49.378	0.615	80.270	0.000
G1_SUM	1.073	0.021	52.167	0.000
G2_SUM	0.385	0.026	14.998	0.000
G3_SUM	0.513	0.026	20.065	0.000
Variances				
PRETEST	32.566	2.788	11.681	0.000

Model #3 Results

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PRETEST BY				
PRE	0.900	0.007	122.956	0.000
POSTTEST BY				
POST	0.939	0.005	175.828	0.000
POSTTEST ON				
PRETEST	0.654	0.044	14.932	0.000
POSTTEST ON				
G1_SUM	-0.008	0.032	-0.237	0.813
G2_SUM	0.015	0.047	0.309	0.757
G3_SUM	-0.011	0.040	-0.275	0.783
G1_SUM ON				
PRETEST	-0.101	0.047	-2.171	0.030
G2_SUM ON				
PRETEST	0.030	0.050	0.599	0.549
G3_SUM ON				
PRETEST	-0.024	0.048	-0.498	0.618
G1_SUM WITH				
G2_SUM	0.020	0.045	0.432	0.666
G3_SUM	0.064	0.046	1.396	0.163
G2_SUM WITH				
G3_SUM	0.089	0.045	1.984	0.047

Model #3 Path Diagram



Model #3 Results

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PRE	0.810	0.013	61.478	0.000
POST	0.882	0.010	87.914	0.000
G1_SUM	0.010	0.009	1.085	0.278
G2_SUM	0.001	0.003	0.300	0.764
G3_SUM	0.001	0.002	0.249	0.803

Latent Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
POSTTEST	0.430	0.058	7.408	0.000

Model #3 Results

TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from PRETEST to POSTTEST				
Total	0.758	0.054	13.941	0.000
Total indirect	0.002	0.004	0.403	0.687
Specific indirect				
POSTTEST G1_SUM PRETEST	0.001	0.004	0.238	0.812
POSTTEST G2_SUM PRETEST	0.001	0.002	0.285	0.775
POSTTEST G3_SUM PRETEST	0.000	0.001	0.258	0.797
Direct				
POSTTEST PRETEST	0.756	0.055	13.839	0.000

Model #3 Results

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from PRETEST to POSTTEST				
Total	0.656	0.044	14.895	0.000
Total indirect	0.001	0.004	0.402	0.688
Specific indirect				
POSTTEST G1_SUM PRETEST	0.001	0.003	0.238	0.812
POSTTEST G2_SUM PRETEST	0.000	0.002	0.285	0.776
POSTTEST G3_SUM PRETEST	0.000	0.001	0.258	0.796
Direct				
POSTTEST PRETEST	0.654	0.044	14.932	0.000

Example #1: Research Questions…Answered

- The evaluation sought to answer the following questions:
  - Did scores improve on the posttest when compared with the pretest?
    - Yes, posttest scores improved by .654 SD for every one SD increase in the pretest score ($p < .001$), holding coursework constant
  - Did coursework significantly affect the posttest scores?
    - No, no coursework was significantly related to the posttest
  - Did the score on the pretest predict the coursework students took?
    - The G1 coursework was significantly reduced, with -.101 SD in number of courses taken for every SD increase in the pretest score ($p = .030$)
  - Did coursework mediate the relationship between pretest and posttest?
    - No, there was no indirect effect of pretest on posttest as mediated by coursework ($p = .687$)

CONCLUDING REMARKS

Wrapping Up…

- Today was about putting it all together: path analysis and measurement models

- The SEM framework allows for powerful inferential analyses to be conducted in a statistically rigorous manner
  - But with the power comes a lot of frustration – data do not always cooperate

- You will find that people take great liberties with how they conduct SEM analyses