Structural Equation Models: Path Analysis with Latent Variables

Structural Equation Modeling
Lecture #7

December 3, 2025

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

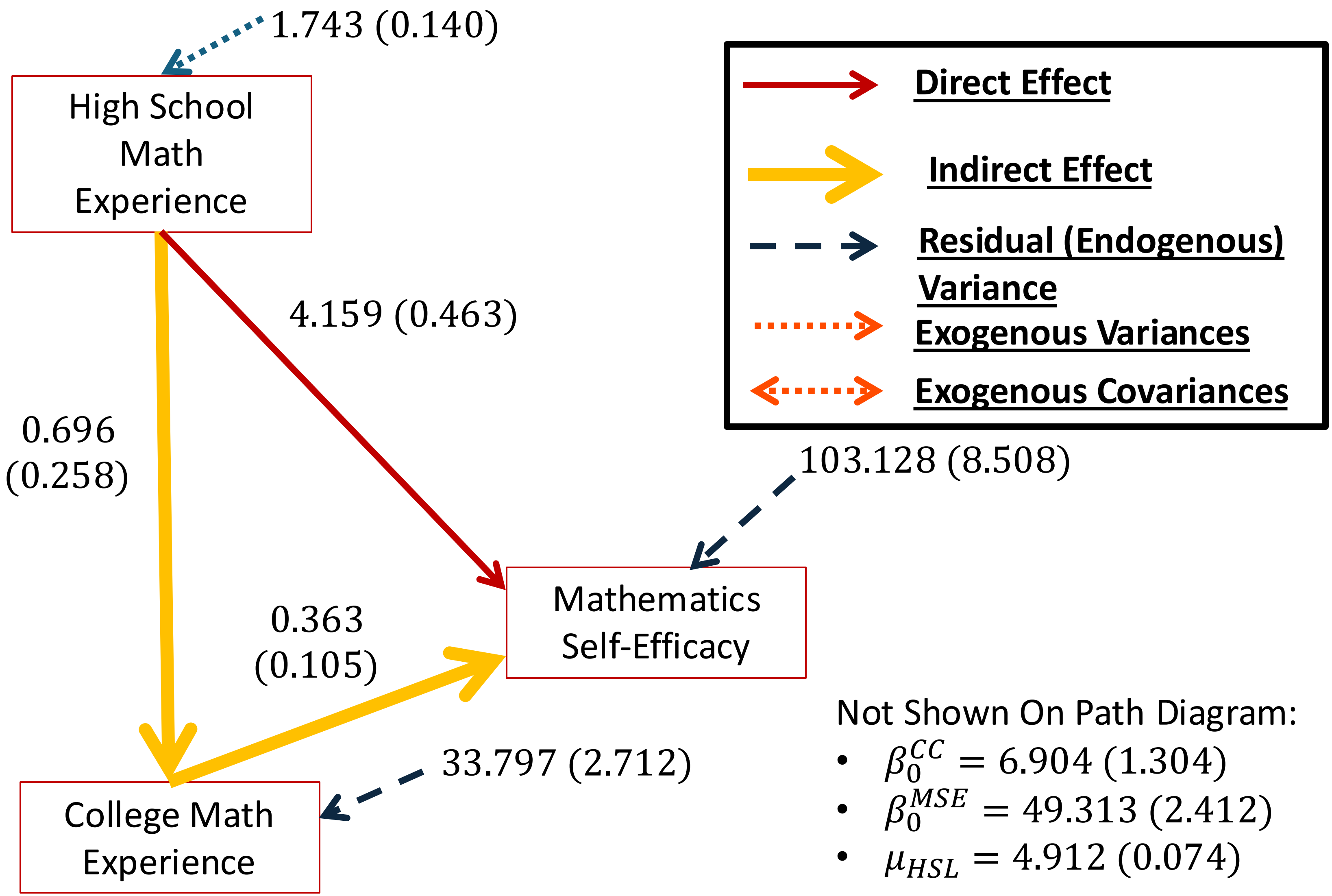
Types of Variables in the Analysis

- 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:

$$CC_p = \beta_0^{CC} + \beta_{HSL}^{CC} HSL_p + e_p^{CC}$$
$$MSE_p = \beta_0^{MSE} + \beta_{CC}^{MSE} CC_p + \beta_{HSL}^{MSE} HSL_p + e_p^{MSE}$$

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

- Alternatively, we could rephrase this in matrix form:

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

Where:

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

$\boldsymbol{y}_p = \begin{bmatrix} CC_p \\ MSE_p \end{bmatrix}$ (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))

$\boldsymbol{\zeta}_p = \begin{bmatrix} e_p^{CC} \\ e_p^{MSE} \end{bmatrix} \sim N_2(\mathbf{0}, \boldsymbol{\Psi})$ (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:

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

- 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	$\mathbf{\Psi}$	$\begin{bmatrix} 33.797 & 0 \\ 0 & 103.128 \end{bmatrix}$
Regression Weights of Exogenous onto Endogenous	$\mathbf{\Gamma}$	$\begin{bmatrix} 0.696 \\ 4.159 \end{bmatrix}$
Covariance Matrix of Exogenous Variables	$\mathbf{\Phi}$	$\begin{bmatrix} 1.743 \end{bmatrix}$
Mean Vector of Exogenous Variables	$\boldsymbol{\mu}_x$	$\begin{bmatrix} 4.912 \end{bmatrix}$
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:

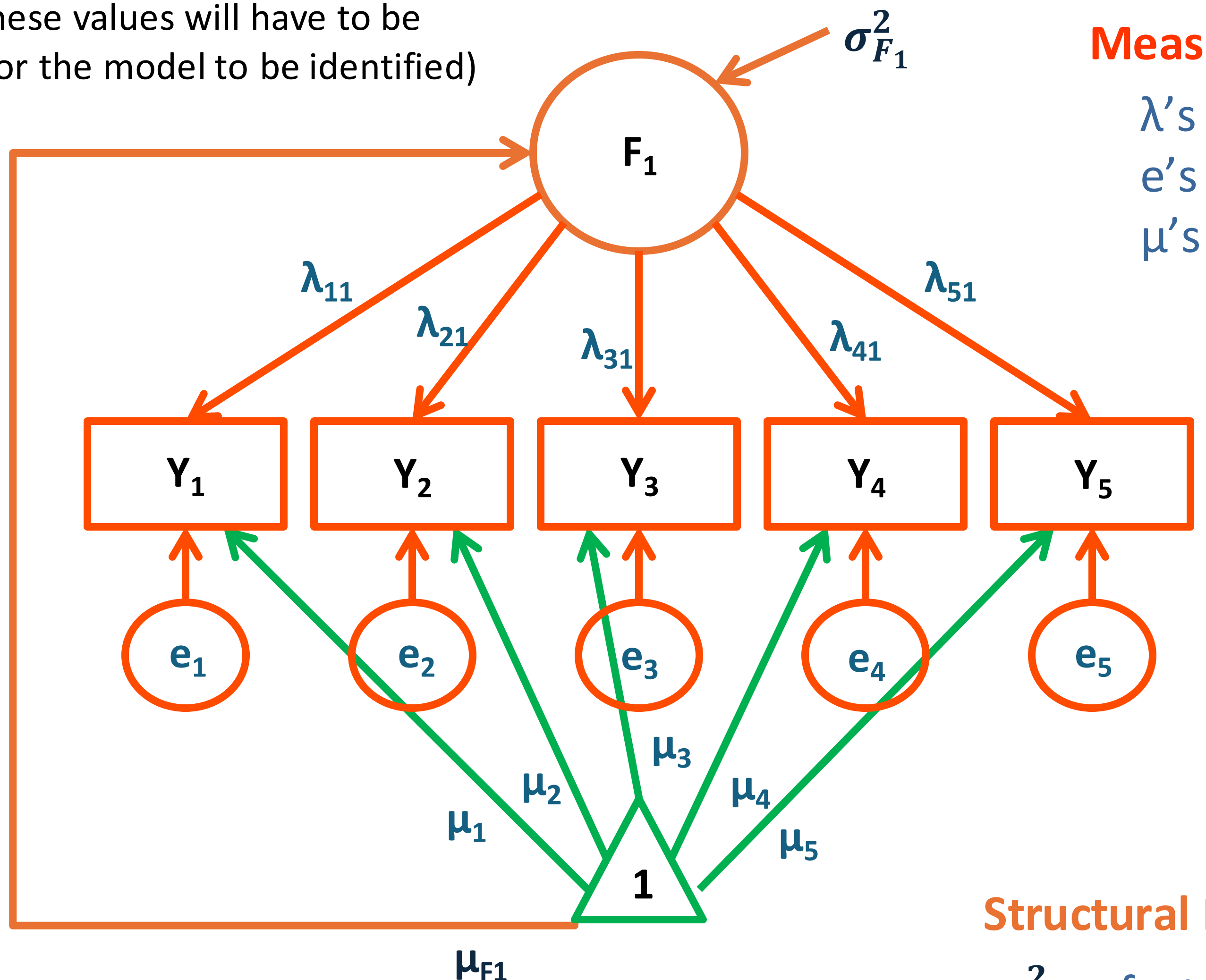
$$
\begin{aligned}
Y_{p1} &= \mu_1 + \lambda_{11} F_{p1} + e_{p1} \\
Y_{p2} &= \mu_2 + \lambda_{21} F_{p1} + e_{p2} \\
Y_{p3} &= \mu_3 + \lambda_{31} F_{p1} + e_{p3} \\
Y_{p4} &= \mu_4 + \lambda_{41} F_{p1} + e_{p4} \\
Y_{p5} &= \mu_5 + \lambda_{51} F_{p1} + e_{p5}
\end{aligned}
$$

- Here:
  - Y_{pi} - response of person p 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_{p1} - latent "factor score" for person p (same for all items) to factor 1 (only one today)
  - e_{pi} - regression-like residual for person p on item i
    - We assume $e_{pi} \sim N\left(0, \psi_i^2\right)$; ψ_i^2 is called the **unique variance** of item i
    - We also assume e_{pi} and F_{p1} are independent

- Also, we will assume $F_{p1} \sim N\left(\mu_{F_1}, \sigma_{F_1}^2\right)$
  - 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:

$$\boldsymbol{\mu}_Y = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\mu}_F$$

$$\begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \\ \mu_{Y_3} \\ \mu_{Y_4} \\ \mu_{Y_5} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \\ \lambda_{51} \end{bmatrix} \begin{bmatrix} \mu_{F_1} \end{bmatrix} = \begin{bmatrix} \mu_1 + \lambda_{11}\mu_{F_1} \\ \mu_2 + \lambda_{21}\mu_{F_1} \\ \mu_3 + \lambda_{31}\mu_{F_1} \\ \mu_4 + \lambda_{41}\mu_{F_1} \\ \mu_5 + \lambda_{51}\mu_{F_1} \end{bmatrix}$$

Model Implied Covariance Matrix

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

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

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

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

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

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 status of the person (student = 1 vs. non-student = 0) – status is observed directl
  - Note: We are assuming that the gambling construct is the same for both students and non-students (we must test this assumption: in two weeks)

Step #1: Building the Measurement Model(s)

- 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

```
#MODEL 01: Gambling GRI Single Factor Model ----------------------------------------------------
model01.syntax = "
  GAMBLING =~ GRI1 + GRI3 + GRI5
"

model01.fit = sem(model01.syntax, data=data01, estimator = "MLR", mimic="Mplus", fixed.x=FALSE)
summary(model01.fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

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

```
#MODEL 02: Full Structural Equation Model ------
model02.syntax = "
  GAMBLING =~ GRI1 + GRI3 + GRI5
  GAMBLING ~ Student
"
```

- The =~ defines the GAMBLING factor
- The ~ says the GAMBLING factor is predicted by the Student variable

- STUDENT 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

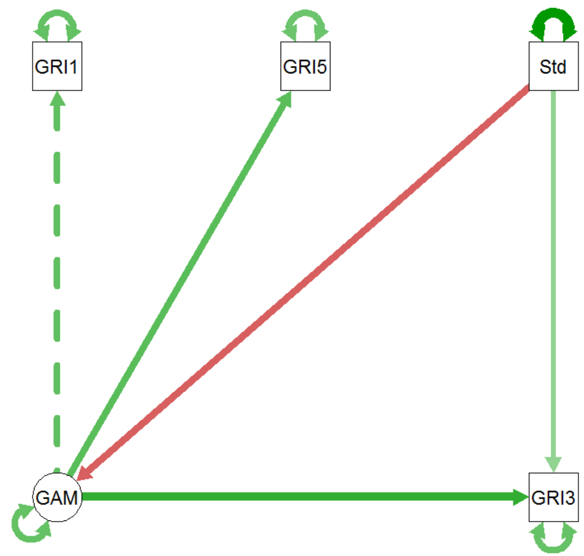
- A necessary but not sufficient way of ensuring identification is the t-rule (counting rule)
  - The number of parameters must be less than the total number of means + variances/covariances of **all** observed variables in the analysis

- 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

Model 03: Adding a Direct Effect of STUDENT on GRI3 (A MIMIC Model)

- As the biggest source of misfit came from the covariance between STUDENT and item GRI3, we will add a direct effect of STUDENT on GRI3
  - STUDENT is predicting GRI3
  - This type of model is called a Multiple Indicators/Multiple Causes (MIMIC) model



- Lavaan syntax:

```
#MODEL 03: Structural Equation Model #2 --
model03.syntax = "
  GRI3 ~ Student
  GAMBLING =~ GRI1 + GRI3 + GRI5
  GAMBLING ~ Student
"
```

- The equation for item GRI3 is:

$$Y_{p3} = \mu_3 + \lambda_3 GAMBLING_p + \beta_{Student}^{GRI3} Student_p + e_p$$

Item 3: If I lost a lot of money gambling one day, I would be more likely to want to play again the following day.

Equation Form of Overall Structural Equation Model

- The structural equation model simultaneous equations

For the "measurement" portion:

$$GRI1_p = \mu_{I_1} + \lambda_{11} GAMBLING_p + e_{p1}$$

$$GRI3_p = \mu_{I_3} + \lambda_{31} GAMBLING_p + \beta_{Student}^{GRI3} Student_p + e_{p3}$$

$$GRI5_s = \mu_{I_5} + \lambda_{51} GAMBLING_p + e_{p5}$$

For the "structural" portion:

$$GAMBLING_p = \beta_0^{\mathrm{GAMBLING}} + \beta_{Student}^{GAMBLING} Student_p + \delta_p$$

- 3 endogenous variables (the latent variable does not count)
- 1 exogenous variable

Regarding Marker Item vs. Std Factor Identification

- The R^2 reported at the end of the parameters is interpreted as any other R^2

```
R-Square:

   GRI1                0.389
   GRI3                0.399
   GRI5                0.422
   GAMBLING            0.392
```

- For instance, the R^2 for GAMBLING indicates that the student variable accounts for 39.2% of variation in the GAMBLING latent variable

- The same value would have occurred had we used the standardized factor identification method
  - But look at the differences in the other parameters of the model

Marker Item Identification: LL and Parameters

Without Prediction of Gambling

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)          -5570.865   -5570.865
  Scaling correction factor                              2.167
    for the MLR correction

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  GAMBLING =~
    GRI1              1.000                               0.639    0.622
    GRI3              0.857    0.097    8.800    0.000    0.548    0.611
    GRI5              0.993    0.106    9.357    0.000    0.635    0.651

Regressions:
  GRI3 ~
    Student           0.435    0.096    4.545    0.000    0.435    0.151
  GAMBLING ~
    Student           0.000                               0.000    0.000

Intercepts:
    GRI1              1.823    0.028   64.871    0.000    1.823    1.775
    GRI3              1.160    0.087   13.329    0.000    1.160    1.295
    GRI5              1.593    0.027   59.749    0.000    1.593    1.635
    Student           0.892    0.008  105.162    0.000    0.892    2.877
    GAMBLING          0.000                               0.000    0.000

Variances:
    GRI1              0.647    0.072    8.931    0.000    0.647    0.613
    GRI3              0.485    0.048   10.064    0.000    0.485    0.604
    GRI5              0.547    0.056    9.754    0.000    0.547    0.576
    GAMBLING          0.408    0.062    6.582    0.000    1.000    1.000
    Student           0.096    0.007   14.450    0.000    0.096    1.000

R-Square:

    GRI1              0.387
    GRI3              0.396
    GRI5              0.424
    GAMBLING          0.000
>
```

With Prediction of Gambling

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)          -5399.671   -5399.671
  Scaling correction factor                              2.186
    for the MLR correction

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  GAMBLING =~
    GRI1              1.000                               0.641    0.624
    GRI3              1.088    0.136    8.026    0.000    0.697    0.805
    GRI5              0.988    0.087   11.341    0.000    0.633    0.649

Regressions:
  GRI3 ~
    Student           1.203    0.176    6.827    0.000    1.203    0.431
  GAMBLING ~
    Student          -1.293    0.114  -11.387    0.000   -2.018   -0.626

Intercepts:
    GRI1              2.976    0.111   26.768    0.000    2.976    2.898
    GRI3              1.729    0.094   18.435    0.000    1.729    1.998
    GRI5              2.732    0.123   22.176    0.000    2.732    2.804
    Student           0.892    0.008  105.162    0.000    0.892    2.877
    GAMBLING          0.000                               0.000    0.000

Variances:
    GRI1              0.644    0.067    9.603    0.000    0.644    0.611
    GRI3              0.450    0.052    8.603    0.000    0.450    0.601
    GRI5              0.549    0.049   11.105    0.000    0.549    0.578
    GAMBLING          0.250    0.039    6.449    0.000    0.608    0.608
    Student           0.096    0.007   14.450    0.000    0.096    1.000

R-Square:

    GRI1              0.389
    GRI3              0.399
    GRI5              0.422
    GAMBLING          0.392
```

STD Factor Identification: LL and Parameters

Without Prediction of Gambling

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)           -5570.865    -5570.865
  Scaling correction factor                                2.167
    for the MLR correction

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  GAMBLING =~
    GRI1              0.639    0.049   13.164    0.000    0.639    0.622
    GRI3              0.548    0.047   11.544    0.000    0.548    0.611
    GRI5              0.635    0.049   12.869    0.000    0.635    0.651

Regressions:
  GRI3 ~
    Student           0.435    0.096    4.545    0.000    0.435    0.151
  GAMBLING ~
    Student           0.000                               0.000    0.000

Intercepts:
    GRI1              1.823    0.028   64.871    0.000    1.823    1.775
    GRI3              1.160    0.087   13.329    0.000    1.160    1.295
    GRI5              1.593    0.027   59.749    0.000    1.593    1.635
    Student           0.892    0.008  105.162    0.000    0.892    2.877
    GAMBLING          0.000                               0.000    0.000

Variances:
    GRI1              0.647    0.072    8.931    0.000    0.647    0.613
    GRI3              0.485    0.048   10.064    0.000    0.485    0.604
    GRI5              0.547    0.056    9.754    0.000    0.547    0.576
    GAMBLING          1.000                               1.000    1.000
    Student           0.096    0.007   14.450    0.000    0.096    1.000

R-Square:

    GRI1              0.387
    GRI3              0.396
    GRI5              0.424
    GAMBLING          0.000
```

With Prediction of Gambling

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)           -5399.671    -5399.671
  Scaling correction factor                                2.186
    for the MLR correction

                   Estimate  Std.err  Z-value  P(>|z|)   Std.lv  Std.all
Latent variables:
  GAMBLING =~
    GRI1              0.500    0.039   12.897    0.000    0.641    0.624
    GRI3              0.543    0.049   11.135    0.000    0.697    0.805
    GRI5              0.494    0.039   12.712    0.000    0.633    0.649

Regressions:
  GRI3 ~
    Student           1.203    0.176    6.827    0.000    1.203    0.431
  GAMBLING ~
    Student          -2.587    0.261   -9.909    0.000   -2.018   -0.626

Intercepts:
    GRI1              2.976    0.111   26.768    0.000    2.976    2.898
    GRI3              1.729    0.094   18.435    0.000    1.729    1.998
    GRI5              2.732    0.123   22.176    0.000    2.732    2.804
    Student           0.892    0.008  105.162    0.000    0.892    2.877
    GAMBLING          0.000                               0.000    0.000

Variances:
    GRI1              0.644    0.067    9.603    0.000    0.644    0.611
    GRI3              0.450    0.052    8.603    0.000    0.450    0.601
    GRI5              0.549    0.049   11.105    0.000    0.549    0.578
    GAMBLING          1.000                               0.608    0.608
    Student           0.096    0.007   14.450    0.000    0.096    1.000

R-Square:

    GRI1              0.389
    GRI3              0.399
    GRI5              0.422
    GAMBLING          0.392
```

Same model fit as marker item identification—but parameters go crazy when factor variance is fixed to 1

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

WHY SEM MATTERS: MEASUREMENT ERROR PROPAGATES THROUGH TO ESTIMATES

Previous Analysis…without SEM

- The previous analysis was built to be a demonstration of how structural equation models can be built and how results are interpreted
- Perhaps more important is why we are using SEMs in the first place: the GAMBLING variable does not exist
- Without SEM, a similar analysis using the sum score of the three gambling items could have been conducted
  - Likely that's more prevalent in educational and social sciences research
- However, such analyses will have biased estimates (regression slopes) and biased standard errors
  - Next, we compare and contrast such analyses

Analysis with a Sum Score: Creating the Sum Score

- To create a sum score with the GAMBLING 3-item scale:

```
#creating sum score for GAMBLING 3-item Survey
data02 = data01
data02$GRI135sum = data02$GRI1 + data02$GRI3 + data02$GRI5
```

- We can also calculate the reliability of that sum score using the Guttman-Chronbach alpha
  - The three-item sum-score GC reliability was .352

- Side note: I calculated this from a CFA model (we'll discuss reliability and how to do this next week)

```
#Model 06: Calculation of Alpha Reliablity with Tau-Equivalent CFA Model ------------------------------
model06.syntax = "
  GAMBLING =~ (loading)*GRI1 + (loading)*GRI3 + (loading)*GRI5
  GRI1 ~~ (U1)*GRI1
  GRI3 ~~ (U3)*GRI3
  GRI5 ~~ (U5)*GRI5

  GCalpha := (3*loading*loading)/( (3*loading*loading) + (U1 + U3 + U5))
"
model06.fit = sem(model06.syntax, data=data02, estimator = "MLR", mimic="Mplus", fixed.x=FALSE, std.lv = TRUE)
summary(model06.fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

Comparing Results Between Models

- Sum-score model:

```
model05.syntax = "
  GRI135sum ~ Student
"
model05.fit = sem(model05.syntax, data=data02, estimator = "MLR", mimic="Mplus", fixed.x=FALSE)
summary(model05.fit, fit.measures=TRUE, rsquare=TRUE, standardized=TRUE)
```

- Std.nox coefficient of interest:

```
> standardizedSolution(model05.fit, type="std.nox")[1,]
        lhs op     rhs est.std    se       z pvalue
1 GRI135sum  ~ Student  -1.273 0.119 -10.673      0
```

- Compared with model 3a (best model+MIMIC):

```
> standardizedSolution(model03a.fit, type="std.nox")[5,]
       lhs op     rhs est.std    se       z pvalue
1 GAMBLING  ~ Student  -2.018 0.177 -11.387      0
```

  - Not a valid comparison
  - Sum-score model has no way to indicate misfit or lack of true assumptions

- Compared with model 2 (no MIMIC):

```
> standardizedSolution(model02.fit, type="std.nox")[4,]
       lhs op     rhs est.std   se      z pvalue
1 GAMBLING  ~ Student  -1.745 0.21 -8.327      0
```

Side-by-Side Comparison

```
> standardizedSolution(model05.fit, type="std.nox")[1,]
        lhs op     rhs est.std    se       z pvalue
1 GRI135sum  ~ Student  -1.273 0.119 -10.673      0
```

```
> standardizedSolution(model02.fit, type="std.nox")[4,]
       lhs op     rhs est.std   se      z pvalue
1 GAMBLING  ~ Student  -1.745 0.21 -8.327      0
```

- Using sum scores results in:
  - The estimate of the standardized mean difference between students and non-students is lower (-1.273 for sum score vs. -1.745 in latent)
    - This could result in Type-II error issues
  - The standard error is much lower (.119 for sum score vs. .21 in latent)
    - The sum score standard error is about 57% the size of the latent version
    - This could result in Type-I error issues
- The reason: sum scores contain multiple sources of error:
  - Measurement error
  - Model fit error
- We will discuss sum scores next week…and the best practices of to do when you cannot use a full SEM

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 university in a state that is supposed to be for lovers
  - I no longer have the data, so these results come from Mplus

- 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          1.063
            for MLR

SRMR (Standardized Root Mean Square Residual)

          Value                              0.025
```

```
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
```

- Normalized residuals:

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

- Need for residual covariances between coursework sums

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

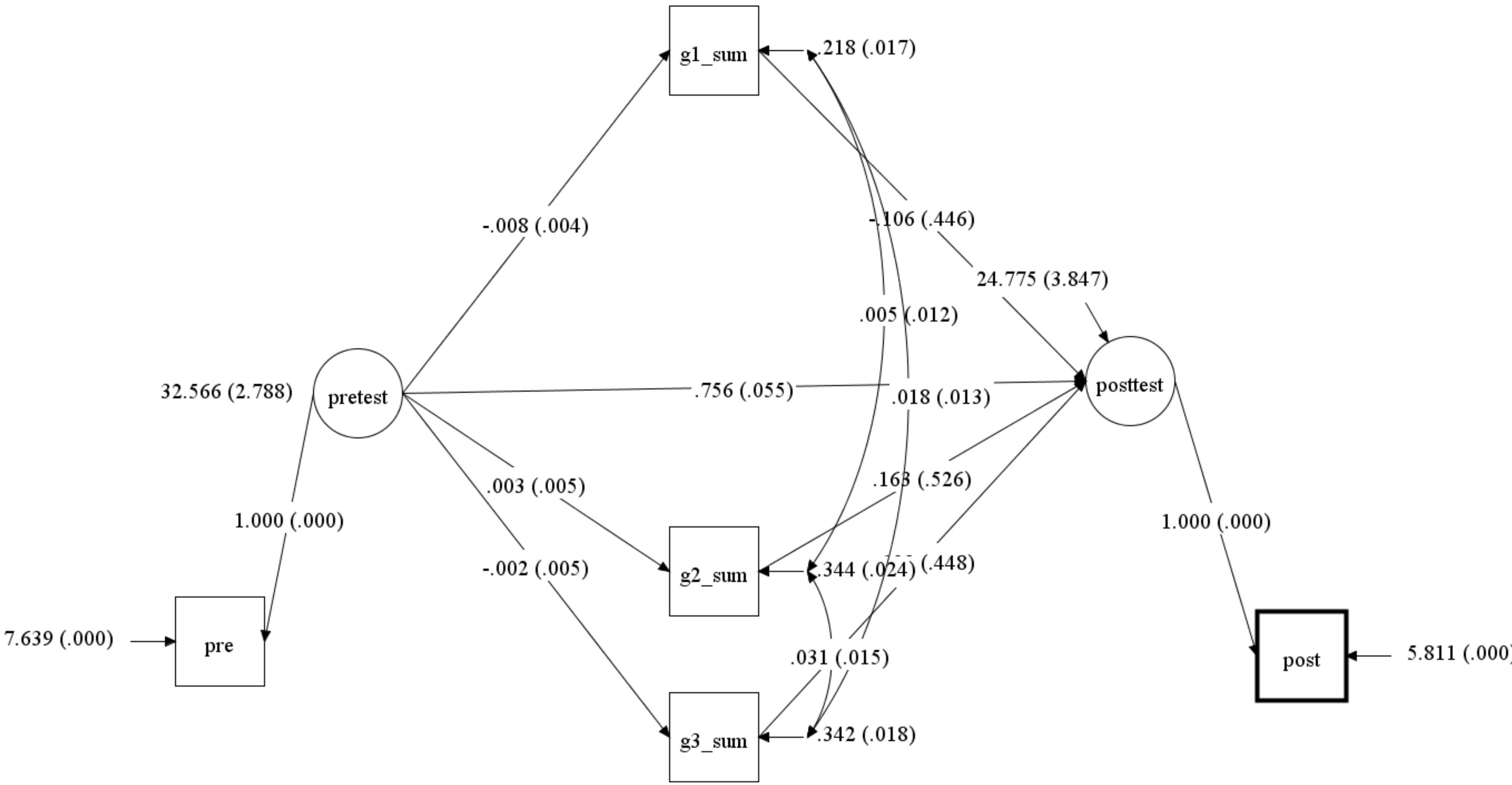
```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    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


                                                   Two-Tailed
                    Estimate       S.E.  Est./S.E.    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

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    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