Path Analysis

Latent Trait Measurement and Structural Equation Models
Lectures #3 and #4

January 23 and 30, 2013

Today's Lecture

- Path analysis
  - Starting with linear regression…
  - …then moving to multivariate regression…
  - …then moving to a "small" path model…
  - …then arriving at our final destination

- Path analysis details:
  - Standardized coefficients (introduced in regression)
  - Model fit (introduced in multivariate regression)
  - Model modification (introduced in multivariate regression and path analysis)
  - Direct and indirect effects (introduced in path analysis)

- Additional issues in path analysis
  - Estimation types
  - Variable considerations

Today's Data Example

- Data are simulated based on the results reported in:

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: a path analysis. *Journal of Educational Psychology, 86*, 193-203.

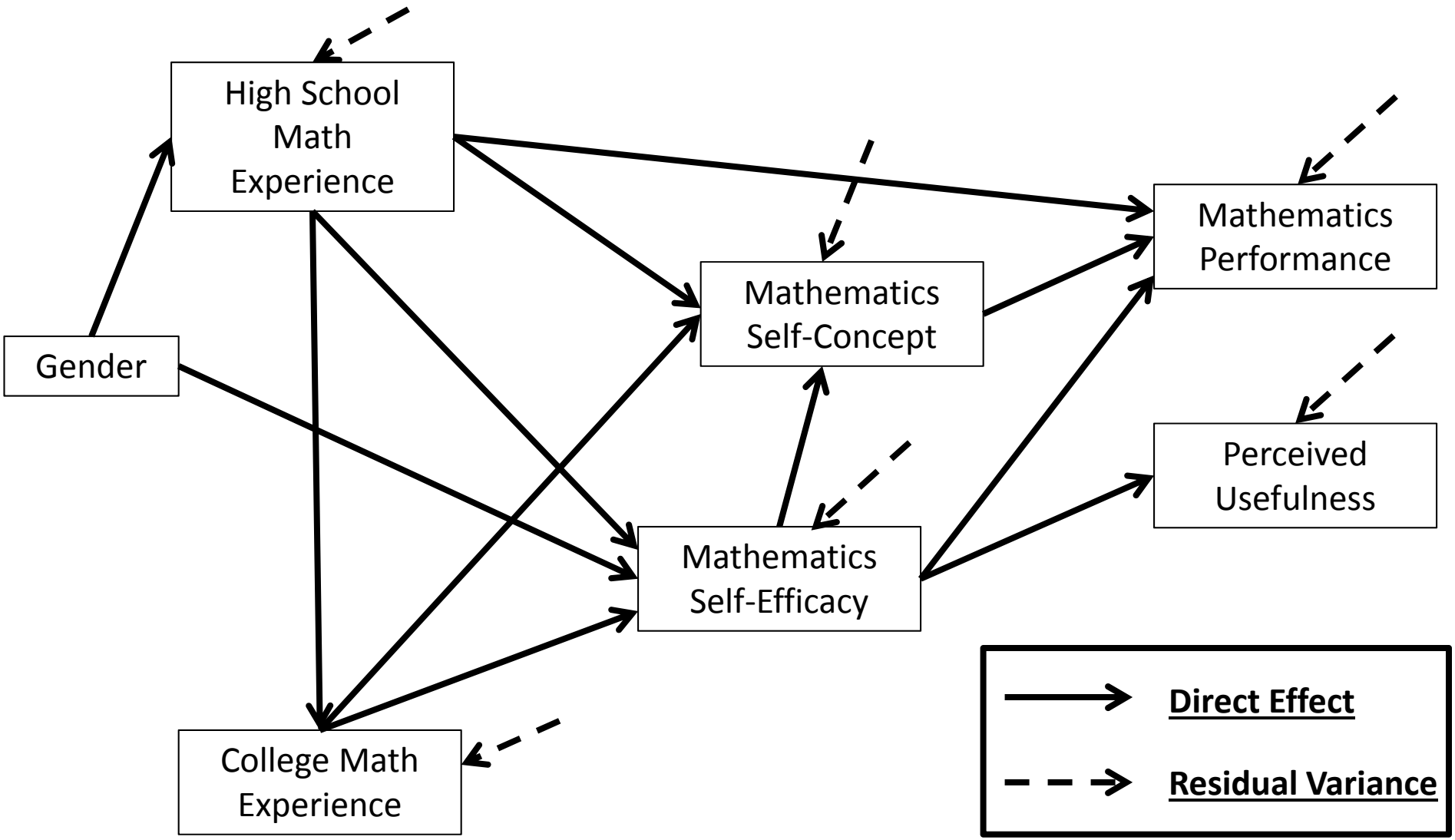
- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- Note: simulated data characteristics differ from actual data (some variables extend beyond their official range)
  - Simulated using Multivariate Normal Distribution
    - Some variables had boundaries that simulated data exceeded
  - Results will not match exactly due to missing data and boundaries

Variables of Data Example

- Gender (1 = male; 0 = female)
- Math Self-Efficacy (MSE)
  - Reported reliability of .91
  - Assesses math confidence of college students
- Perceived Usefulness of Mathematics (USE)
  - Reported reliability of .93
- Math Anxiety (MAS)
  - Reported reliability ranging from .86 to .90
- Math Self-Concept (MSC)
  - Reported reliability of .93 to .95
- Prior Experience at High School Level (HSL)
  - Self report of number of years of high school during which students took mathematics courses
- Prior Experience at College Level (CC)
  - Self report of courses taken at college level
- Math Performance (PERF)
  - Reported reliability of .788
  - 18-item multiple choice instrument (total of correct responses)

Our Destination: Overall Path Model



The Big Picture

- Path analysis is a multivariate statistical method that assumes the variables in an analysis are multivariate normally distributed
  - Mean vectors
  - Covariance matrices

- By specifying simultaneous regression equations (the core of path models), a very specific covariance matrix is implied
  - Similar to last week's homework with Models #1 (independent variables, common variance) and #2 (common covariance and common variance)

- Much like MANOVA and multilevel models, the key to path analysis is finding an effective approximation to the unstructured (saturated) covariance matrix
  - With fewer parameters, if possible

- The art to path analysis is in specifying models that blend theory and statistical evidence to produce valid, generalizable results

LINEAR REGRESSION: A BASIC PATH MODEL

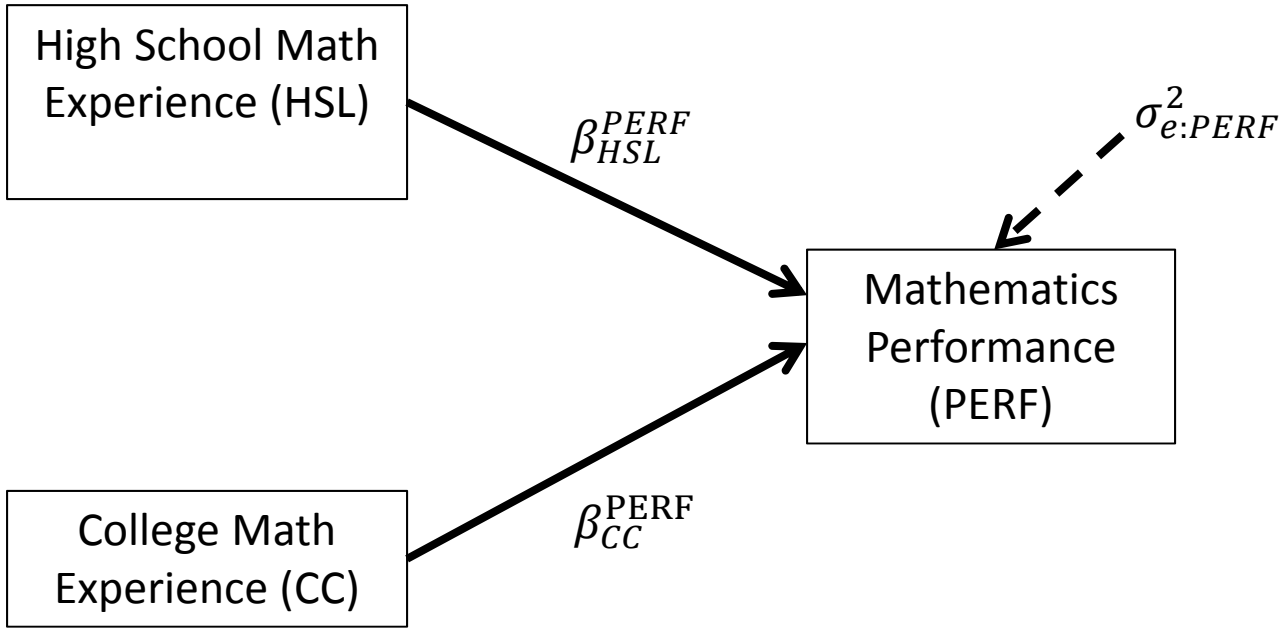
Linear Regression Framed As A Path Model

- We will begin our discussion by starting with linear regression: predicting mathematics performance (PERF) with high school (HSL) and college experience (CC)

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$

- As typical, we assume $e_i^{PERF} \sim N(0, \sigma_{e:PERF}^2)$
- A guide to my notation:
  - β_X^Y - the regression slope where variable Y is being predicted by variable X
  - β_0^Y - the intercept for the regression line predicting variable Y
  - e_i^Y - the residual for variable Y for observation i
  - $\sigma_{e:Y}^2$ - the **residual** variance (note the e: in the subscript) for the prediction of variable Y
  - σ_X^2 - the variance of variable X (not a residual – unexplained)

Linear Regression Path Diagram



Types of Variables in the Analysis

- An important distinction in path analysis 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
    - Mathematics Performance (PERF) in our example

- 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
    - High school (HSL) and college (CC) experience

Linear Regression in Mplus

- The basic code for linear regression in Mplus uses the ON statement:

```
VARIABLE:
    NAMES = id gender hsl cc use msc mas mse perf;
    USEVARIABLE = hsl perf cc;
    IDVARIABLE = id;
    MISSING = .;

ANALYSIS:
    ESTIMATOR = MLR;

MODEL:
    perf ON hsl cc;

OUTPUT:
    STANDARDIZED RESIDUAL;
```

- Mplus uses ML by default to estimate the parameters of the model
  - Listwise deletion happens for any independent variables (right of ON) with missing data

```
*** WARNING
  Data set contains cases with missing on x-variables.
  These cases were not included in the analysis.
  Number of cases with missing on x-variables:  68
```

- Sample should be 350 subjects
  - Mplus uses 237

```
SUMMARY OF ANALYSIS

Number of groups                                         1
Number of observations                                 237

Number of dependent variables                            1
Number of independent variables                          2
Number of continuous latent variables                    0
```

DETOUR #1: MISSING DATA WITH MAXIMUM LIKELIHOOD IN MPLUS

More Linear Regression in Mplus

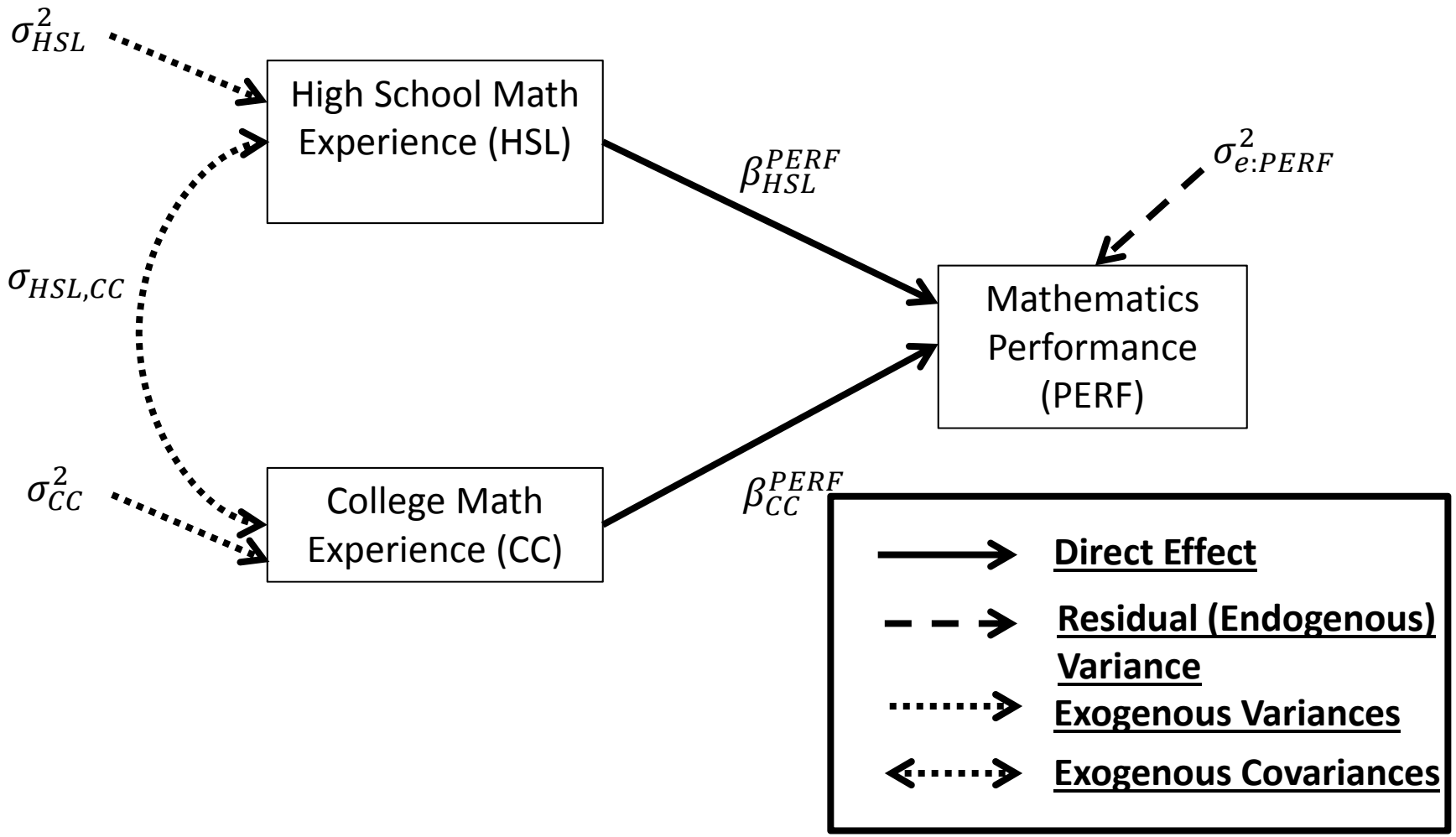
- One way to get around Mplus omitting independent (exogenous) variables is to add the command to estimate their covariance:

```
MODEL:
    perf ON hsl cc;
    hsl; cc;
    hsl WITH cc;
```

- Now, Mplus attempts to estimate the covariance matrix of all variables, using the multivariate normal distribution
  - There is one omitted case (missing on all three variables)

```
SUMMARY OF ANALYSIS

Number of groups                                         1
Number of observations                                 349

Number of dependent variables                            1
Number of independent variables                          2
Number of continuous latent variables                    0
```

Full Model Linear Regression Path Diagram



Full Model in Statistical Distributions

- The full model uses maximum likelihood for the multivariate normal (MVN) distribution for all variables
  - Including the exogenous (independent) variables does not affect the direct effects in the model
  - Assumes Missing At Random for all variables in the model

- The MVN likelihood function has its mean vector and covariance re-expressed as a function of:
  - The model parameters (β_0^{PERF}, β_{HSL}^{PERF}, β_{CC}^{PERF}, $\sigma_{e:PERF}^2$)
  - The exogenous variable means and covariances (μ_{HSL}, μ_{CC}, σ_{HSL}^2, σ_{CC}^2, $\sigma_{HSL,CC}$)

Full Model Likelihood Function

- The log-likelihood function (from the previous lecture):

$$L(\mathbf{X}_i|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}) = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log(|\boldsymbol{\Sigma}|) - \sum_{i=1}^{N}\frac{\left(\mathbf{x}_i^T - \boldsymbol{\mu}_i\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i^T - \boldsymbol{\mu}_i\right)}{2}$$

- The MVN distribution has two matrices of parameters:
  - A mean vector $\boldsymbol{\mu}_i$
    - In regression this is called the ***conditional mean*** – the predicted value of the dependent variable for an observation i
  - A covariance matrix $\boldsymbol{\Sigma}$
    - In regression, this is the same for all observations (so no subscript)

- These matrices result from the parameters in the full regression model

Model-Predicted Mean Vector

- The mean vector now becomes different for each observation i
- This is called a conditional mean vector
  - Conditional on the values of the independent variables:

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i,PERF} = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i \\ \mu_{HSL} \\ \mu_{CC} \end{bmatrix}$$

- The means for HSL and CC are not conditional as they are exogenous (not explained)

Model-Predicted Covariance Matrix

- The model predicted covariance matrix is not conditional as it does not depend on the values of the independent variables

$$\Sigma = \begin{bmatrix} \sigma^2_{PERF} & \sigma_{PERF,HSL} & \sigma_{PERF,CC} \\ \sigma_{PERF,HSL} & \sigma^2_{HSL} & \sigma_{HSL,CC} \\ \sigma_{PERF,CC} & \sigma_{HSL,CC} & \sigma^2_{CC} \end{bmatrix}$$

- Note that there is no subscript – assumed the same for all i

Where, by the regression model:

$$\sigma^2_{PERF} = \left(\beta^{PERF}_{HSL}\right)^2 \sigma^2_{HSL} + 2\beta^{PERF}_{HSL}\beta^{PERF}_{CC}\sigma_{HSL,CC} + \left(\beta^{PERF}_{CC}\right)^2 \sigma^2_{CC} + \sigma^2_{e:PERF}$$

- $\sigma^2_{e:PERF}$ is the residual variance; σ^2_{PERF} is the variance

- $\hat{\sigma}_{PERF,HSL} = \beta^{PERF}_{HSL}\sigma^2_{HSL} + \beta^{PERF}_{CC}\sigma_{HSL,CC}$
- $\hat{\sigma}_{PERF,CC} = \beta^{PERF}_{CC}\sigma^2_{CC} + \beta^{PERF}_{HSL}\sigma_{HSL,CC}$

Maximum Likelihood with Missing Data

- Because the MLR algorithm in Mplus assumes a MVN distribution, if one or more variables are missing, a reduced mean vector/covariance matrix is used

- For instance, if an observation was missing HSL:

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i,PERF} \\ \mu_{CC} \end{bmatrix}; \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2_{PERF} & \sigma_{PERF,CC} \\ \sigma_{PERF,CC} & \sigma^2_{CC} \end{bmatrix}$$

- I highlight these terms because the linear regression model (and path models) make specific predictions about a person's MVN distribution matrices
  - A regression model is a saturated (unstructured) model for the distributional parameters of the MVN
  - Number of distributional parameters (3 means, 3 variances, 3 covariances = 9) equals number of model parameters (1 intercept, 2 slopes, 1 residual variance, 2 means, 2 variances, 1 covariance = 9)

Saturated/Unstructured Model in Mplus

- The Model Fit Information section of Mplus output confirms that the number of parameters in our model is the same as the saturated model (the unstructured model from last week)

```
MODEL FIT INFORMATION

Number of Free Parameters                        9

Loglikelihood

          H0 Value                       -2215.144
          H0 Scaling Correction Factor       0.9520
            for MLR
          H1 Value                       -2215.144
          H1 Scaling Correction Factor       0.9520
            for MLR


Chi-Square Test of Model Fit

          Value                              0.000*
          Degrees of Freedom                     0
          P-Value                           0.0000
          Scaling Correction Factor          1.0000
            for MLR
```

- Here the log-likelihood for H1 (unstructured/saturated) model and H0 (the current model) are identical
  - Therefore, no tests of model fit are possible – no DF
  - Model predicted covariance is exact – no residual variances or covariances

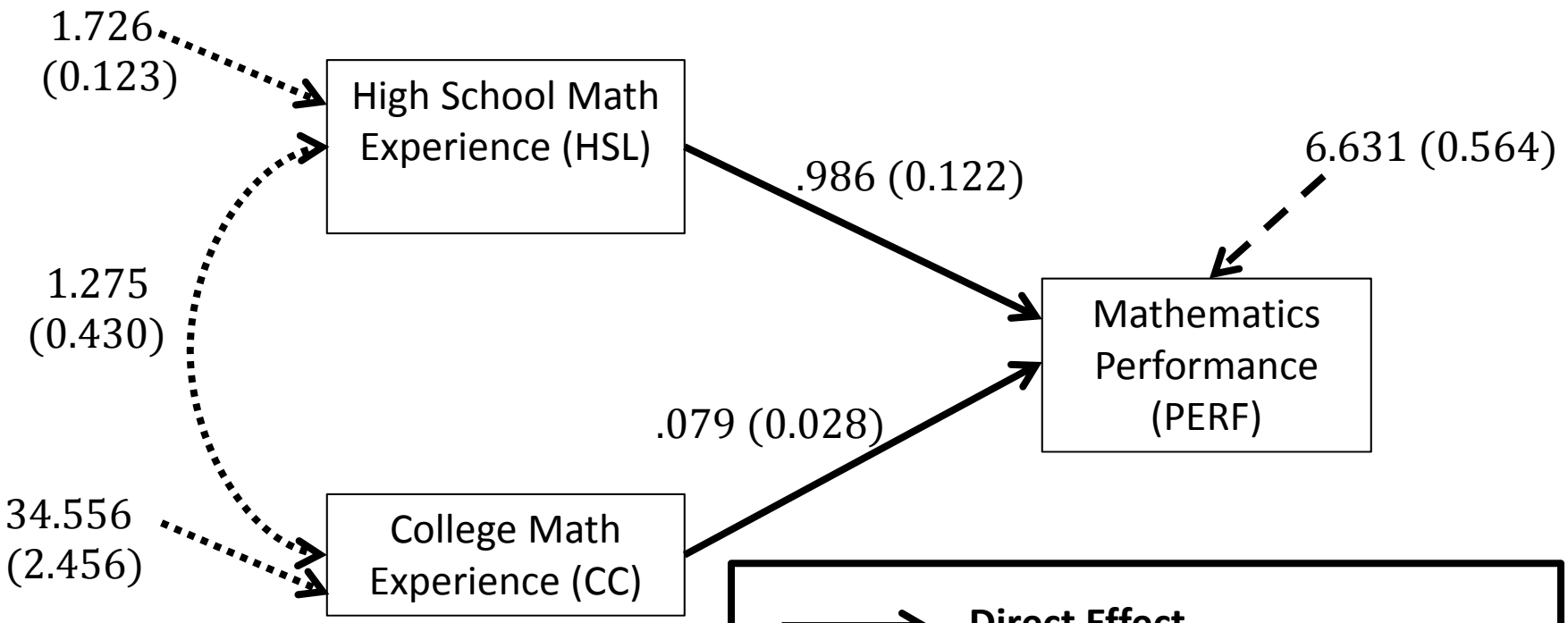
…AND NOW BACK TO OUR REGULARLY SCHEDULED REGRESSION ANALYSIS

Linear Regression Results in Mplus

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.986	0.122	8.103	0.000
CC	0.079	0.028	2.861	0.004
HSL WITH				
CC	1.275	0.430	2.967	0.003
Means				
HSL	4.925	0.073	67.051	0.000
CC	10.331	0.331	31.189	0.000
Intercepts				
PERF	8.253	0.598	13.807	0.000
Variances				
HSL	1.726	0.123	14.039	0.000
CC	34.556	2.456	14.069	0.000
Residual Variances				
PERF	6.631	0.564	11.759	0.000

Linear Regression Path Diagram with Results



Not Shown On Path Diagram:

- $\beta_0^{PERF} = 8.253\ (0.598)$
- $\mu_{HSL} = 4.925\ (0.073)$
- $\mu_{CC} = 10.331\ (0.331)$

Interpreting Linear Regression Results

- The linear regression results are interpreted as follows:
  - $\beta_0^{PERF} = 8.253$: the intercept for PERF – the value of PERF when all predictors are zero (HSL = 0 and CC = 0)
  - $\beta_{HSL}^{PERF} = 0.986$: the slope for HSL. Indicates that for every one-unit increase in HSL (holding CC constant), PERF increases by .986
  - $\beta_{CC}^{PERF} = 0.079$: the slope for CC. Indicates that for every one-unit increase in CC (holding HSL constant), PERF increases by .079
  - $\sigma_{e:PERF}^2 = 6.631$: the residual (or unexplained) variance in PERF
  - Note: the rest of the parameters are the descriptive statistics for the independent (exogenous) variables and are not explained by the regression model

Explained Variance

- To demonstrate the concept of explained variance, consider the dependent variable, math performance
  - "Empty Model" – estimate of its variance: $\sigma^2_{PERF} = 8.722$

ESTIMATED SAMPLE STATISTICS

Means

	PERF	HSL	CC
1	13.923	4.925	10.331

Covariances

	PERF	HSL	CC
PERF	8.722		
HSL	1.802	1.726	
CC	3.980	1.275	34.556

Note: the sample statistics are from the unstructured (saturated) model estimated with ML – if you have missing data or are using unbiased estimates, these will not match other programs

- The independent (exogenous) variables in the analysis seek to explain the variability in math performance
  - Adding significant IVs will reduce the variance, therefore "explaining" a portion of the DV

Regression Model Explained Variance

- After adding both independent variables HSL and CC, the residual variance of performance was $\sigma^2_{e:PERF} = 6.631$

- Therefore, the inclusion of these variables reduced the variance of PERF from 8.722 to 6.631, for an

$$R^2 = \frac{8.722 - 6.631}{8.722} = .24$$

- Mplus reports this value under the standardized coefficients output (explained next):

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF	0.240	0.043	5.566	0.000

Standardized Coefficients

- The scale of the (unstandardized) slope coefficients is given in terms of UNITS of Y (SD Y) per UNITS of X (SD X)
  - Y goes up β_X^Y UNITS of Y for every UNIT of X
    - HSL has SD of 1.31; CC has SD of 5.88
  - If the UNITS of X differ for the various IVs in a model, it can be hard to compare relative strengths of coefficients
    - $\beta_{HSL}^{PERF} = .986$ (but HSL has SD of 1.31)
    - $\beta_{CC}^{PERF} = .079$ (but CC has SD of 5.88)

- Standardized coefficients are the coefficients that would be obtained if Y and X were standardized:
  - Standardized = variance of 1 (i.e. z-scores used for analysis)

- Standardized coefficients are useful for comparing the relative effects of each IV in the model

Standardization in Mplus

- Under the output section, the word STANDARDIZED will produce standardized coefficients in Mplus output

- Three types of standardizations are given:
  - **STDYX**: These are the standardized regression coefficients; use these for continuous IVs (used for our current analysis)
  - **STDY**: These only standardize based on variance of Y (the DV). Use when binary variables are IVs (like gender dummy coding) as unit of X has no meaning
  - **STD:** Discussed when we get to models with latent variables

Standardized Coefficients Output

- Standardized Coef:

$$b_{effect} = \beta_{effect} \frac{SD(X_{effect})}{SD(Y)}$$

- For HSL:

$$b_{HSL}^{PERF} = .986 \frac{1.726}{8.722} = .439$$

  - PERF increases .439 SD when HSL increases 1 SD (holding CC constant)

- For CC:

$$b_{CC}^{PERF} = .079 \frac{34.556}{8.722} = .157$$

  - PERF increases .157 SD when CC increases 1 SD (holding HSL constant)

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.439	0.047	9.415	0.000
CC	0.157	0.055	2.875	0.004

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.986	0.122	8.103	0.000
CC	0.079	0.028	2.861	0.004
Variances				
HSL	1.726	0.123	14.039	0.000
CC	34.556	2.456	14.069	0.000

RESIDUAL OUTPUT

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	HSL	CC
PERF	8.722		
HSL	1.802	1.726	
CC	3.980	1.275	34.556

REGRESSION IN MATRIX FORM

Regression in Matrices

- Many path modeling texts use matrix algebra to denote complicated path models
  - The easiest way to dissect these texts is to start with a linear regression model – all regression models can be phrased as path models

- Matrix algebra helps to understand which models are identified (able to be estimated)

- Because I believe one should know more than just the path diagram approach, I will re-express our regression using matrix algebra (that will come back in path analysis)
  - I will borrow the notation used by Kaplan (2009)

Linear Regression in Matrices

- Our linear regression equation was given by:

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$

  - Here we have:
    - $p = 1$ endogenous variable
    - $q = 2$ exogenous variables

- Alternatively, we could rephrase this:

$$y_i = \beta_0 + \mathbf{\Gamma} \boldsymbol{x}_i + e_i^y$$

Where:

- $e_i^y \sim N(0, \sigma_{e:y}^2)$
- $\mathbf{\Gamma} = [\beta_{HSL}^{PERF} \quad \beta_{CC}^{PERF}]$ (matrix of size $p\ x\ q$ relating exogenous variables to endogenous variable(s))
- $\boldsymbol{x}_i = \begin{bmatrix} HSL_i \\ CC_i \end{bmatrix}$ (matrix of size $q\ x\ 1$ containing observed exogenous variables)

More Regression with Matrices

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

$$\boldsymbol{\mu}_x = \begin{bmatrix} \mu_{HSL} \\ \mu_{CC} \end{bmatrix}$$

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

$$\boldsymbol{\Phi} = \begin{bmatrix} \sigma^2_{HSL} & \sigma_{HSL,CC} \\ \sigma_{HSL,CC} & \sigma^2_{CC} \end{bmatrix}$$

- 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 conditional mean of the endogenous variables is:

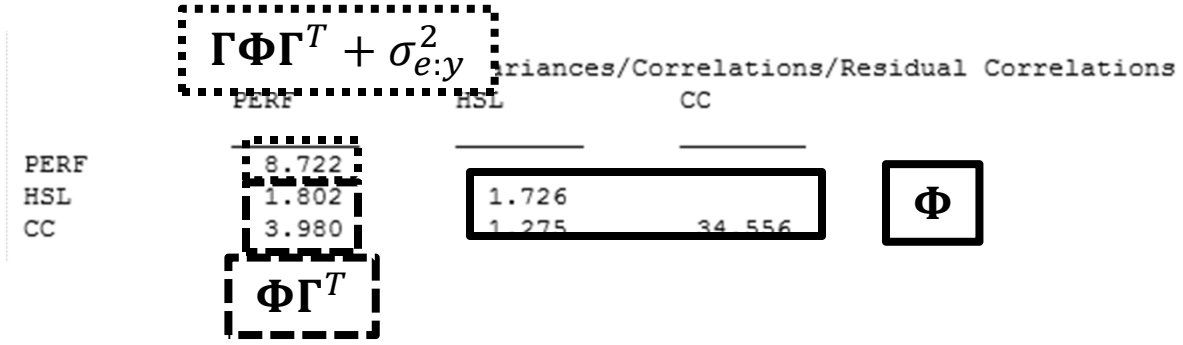
$$\hat{\mu}_y = \beta_0^y + \mathbf{\Gamma}\boldsymbol{\mu}_x$$

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

This covariance matrix has a very specific structure

$$\mathbf{\Sigma}_{y,x} = \begin{bmatrix} \text{Y only} & \text{Y with X} \\ \text{X with Y} & \text{X only} \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma\Phi\Gamma}^T + \sigma^2_{e:y} & \mathbf{\Gamma\Phi} \\ \mathbf{\Phi\Gamma}^T & \mathbf{\Phi} \end{bmatrix}$$

- This is given in Mplus output:



Matching Matrices with Results

- To more specifically link our results (for both the full and reduced model) to the matrices from the previous page:

Name	Matrix	Full Model Estimates
Residual Variance	$\sigma^2_{e:PERF}$	6.631
Regression Weights of Exogenous onto Endogenous	$\boldsymbol{\Gamma}$	$[0.986 \quad 0.079]$
Regression Intercept	β_0^{PERF}	8.253
Covariance Matrix of Exogenous Variables	$\boldsymbol{\Phi}$	$\begin{bmatrix} 1.726 & 1.275 \\ 1.275 & 34.556 \end{bmatrix}$
Mean Vector of Exogenous Variables	$\boldsymbol{\mu}_x$	$\begin{bmatrix} 4.925 \\ 10.331 \end{bmatrix}$

Predicted Model Mean and Covariance Matrix

- The conditional mean of the endogenous variables is:

$$\hat{\mu}_y = \beta_0^y + \boldsymbol{\Gamma}\boldsymbol{\mu}_x = 8.253 + [0.986 \quad 0.079]\begin{bmatrix} 4.925 \\ 10.331 \end{bmatrix} = 13.923$$

Model Estimated Means/Intercepts/Thresholds

	PERF	HSL	CC
1	13.923	4.925	10.331

Residuals for Means/Intercepts/Thresholds

	PERF	HSL	CC
1	0.000	0.000	0.000

The Model Perfectly Reproduces "Saturated" Mean Vector

$$\boldsymbol{\Sigma}_{y,x} = \begin{bmatrix} \text{Y only} & \text{Y with X} \\ \text{X with Y} & \text{X only} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \sigma_{e:y}^2 & \boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}^T & \boldsymbol{\Phi} \end{bmatrix}$$

$$= \begin{bmatrix} [0.986 \quad 0.079]\begin{bmatrix} 1.726 & 1.275 \\ 1.275 & 34.556 \end{bmatrix}\begin{bmatrix} 0.986 \\ 0.079 \end{bmatrix} + 6.631 & [0.986 \quad 0.079]\begin{bmatrix} 1.726 & 1.275 \\ 1.275 & 34.556 \end{bmatrix} \\ \begin{bmatrix} 1.726 & 1.275 \\ 1.275 & 34.556 \end{bmatrix}\begin{bmatrix} 0.986 \\ 0.079 \end{bmatrix} & \begin{bmatrix} 1.726 & 1.275 \\ 1.275 & 34.556 \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 8.722 & 1.802 & 3.980 \\ 1.802 & 1.726 & 1.275 \\ 3.980 & 1.275 & 34.556 \end{bmatrix}$$

The Model Perfectly Reproduces "Saturated" Covariance Matrix

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	HSL	CC
PERF	8.722		
HSL	1.802	1.726	
CC	3.980	1.275	34.556

Residuals for Covariances/Correlations/Re

	PERF	HSL	CC
PERF	0.000		
HSL	0.000	0.000	
CC	0.000	0.000	0.000

MULTIVARIATE REGRESSION

Multivariate Regression

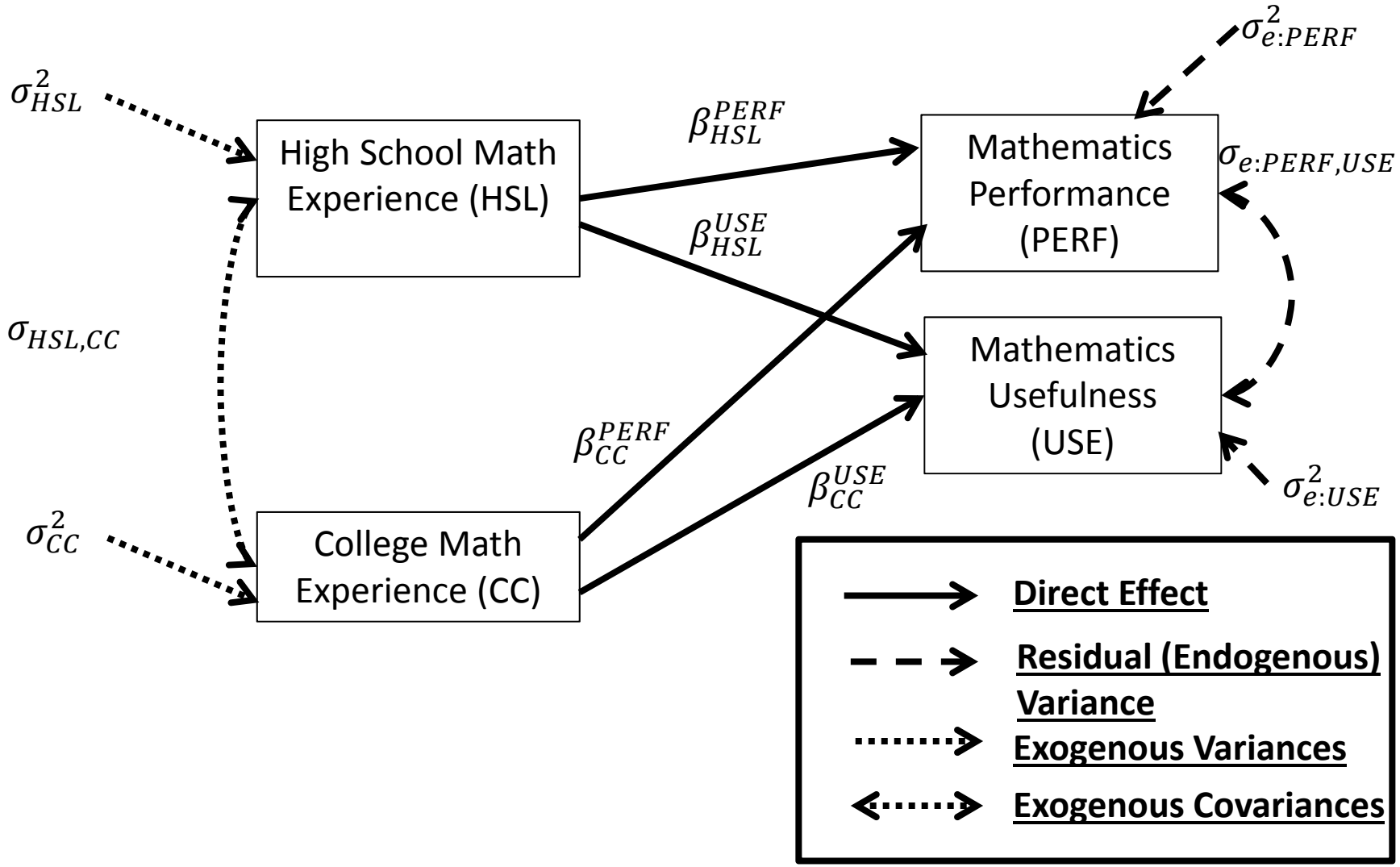
- To transition from regression to path analysis, we will now try to predict two variables simultaneously:
  - Predicting mathematics performance (PERF) with high school (HSL) and college (CC) experience
  - Predicting perceived usefulness (USE) with high school (HSL) and College (CC) experience

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$
$$USE_i = \beta_0^{USE} + \beta_{HSL}^{USE} HSL_i + \beta_{CC}^{USE} CC_i + e_i^{USE}$$

- We denote the residual for PERF as e_i^{PERF} and the residual for USE as e_i^{PERF}
  - Here, we assume the residuals are Multivariate Normal:

$$\begin{bmatrix} e_i^{PERF} \\ e_i^{USE} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e:PERF}^2 & \sigma_{e:PERF,USE} \\ \sigma_{e:PERF,USE} & \sigma_{e:USE}^2 \end{bmatrix} \right)$$

Multivariate Linear Regression Path Diagram



Labeling Variables

- The endogenous (dependent) variables are:
  - Performance (PERF) and Usefulness (USE)

- The exogenous (independent) variables are:
  - High school (HSL) and college (CC) experience

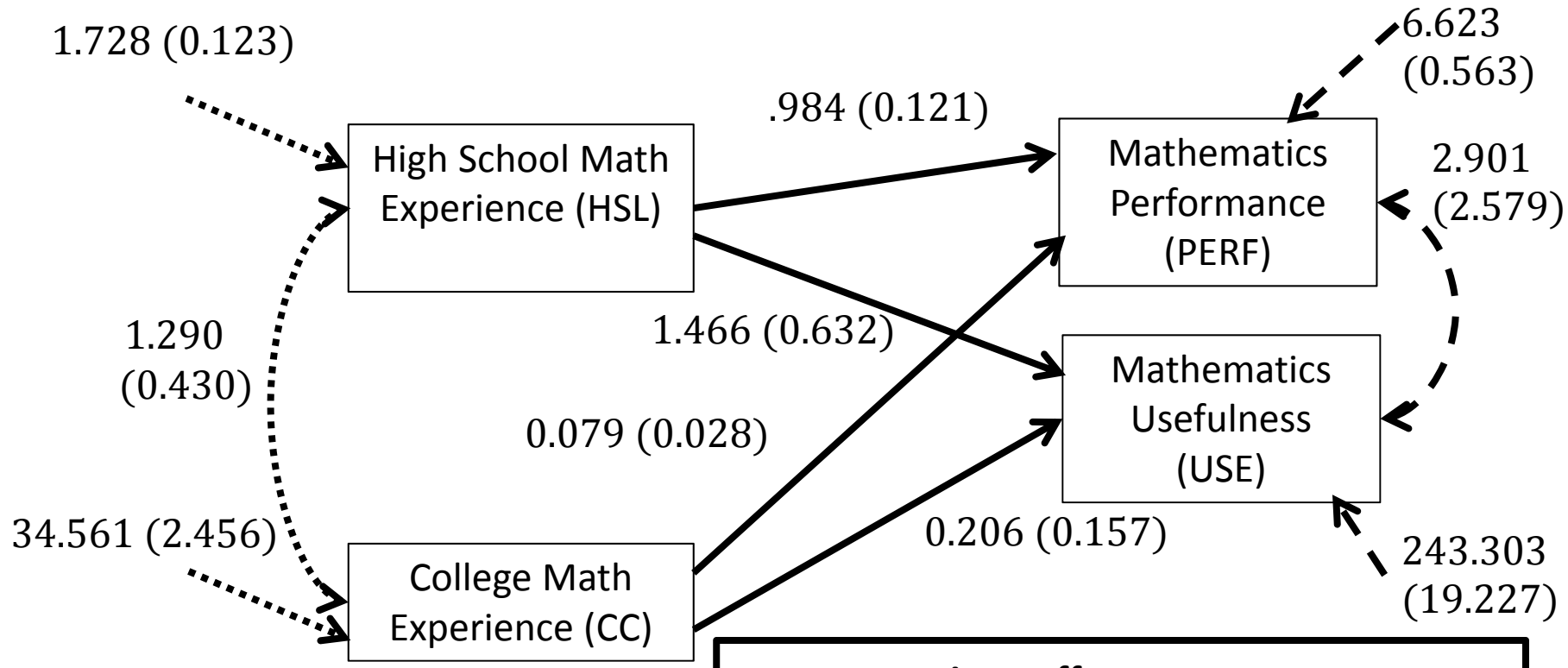
Multivariate Regression Model Parameters

- If we considered all four variables to be part of a multivariate normal distribution, our unstructured (saturated) model would have 14 parameters:
  - 4 means
  - 4 variances
  - 6 covariances (4-choose-2 or 4*(4-1)/2))

- The model itself has 14 parameters:
  - 4 intercepts
  - 4 slopes
  - 2 residual variances
  - 1 residual covariance
  - 2 exogenous variances
  - 1 exogenous covariance

- Therefore, this model will fit perfectly – no model fit statistics will be available
  - Even without model fit, interpretation of parameters can proceed

Multivariate Linear Regression Path Diagram



Not Shown On Path Diagram:

- $\beta_0^{PERF} = 8.264\ (0.594)$
- $\beta_0^{USE} = 43.129\ (3.338)$
- $\mu_{HSL} = 4.922\ (0.073)$
- $\mu_{CC} = 10.330\ (0.331)$

Interpreting Multivariate Regression Results for PERF (nearly identical results)

- $\beta_0^{PERF} = 8.264$: the intercept for PERF – the value of PERF when all predictors are zero (HSL = 0 and CC = 0)

- $\beta_{HSL}^{PERF} = 0.986$: the slope for HSL predicting PERF. Indicates that for every one-unit increase in HSL (holding CC constant), PERF increases by .986
  - The standardized coefficient was .438

- $\beta_{CC}^{PERF} = 0.079$: the slope for CC predicting PERF. Indicates that for every one-unit increase in CC (holding HSL constant), PERF increases by .079
  - The standardized coefficient was .157

Interpreting Multivariate Regression Results for USE

- $\beta_0^{USE} = 43.129$: the intercept for USE – the value of USE when all predictors are zero (HSL = 0 and CC = 0)

- $\beta_{HSL}^{USE} = 1.466$: the slope for HSL predicting USE. Indicates that for every one-unit increase in HSL (holding CC constant), USE increases by 1.466
  - The standardized coefficient was .122

- $\beta_{CC}^{USE} = 0.206$: the slope for CC predicting USE. Indicates that for every one-unit increase in CC (holding HSL constant), USE increases by .206. This was found to be not significant, meaning college experience did not predict perceived usefulness
  - The standardized coefficient was .077

Interpretation of Residual Variances and Covariances

- $\sigma^2_{e:PERF} = 6.623$: the residual variance for PERF
  - The R^2 for PERF was .240 (the same as before)

- $\sigma^2_{e:USE} = 243.303$: the residual variance for USE
  - The R^2 for USE was .024 (a very small effect)

- $\sigma_{e:PERF,USE} = 2.901$: the residual covariance between USE and PERF
  - This value was not significant, meaning we can potentially set its value to zero and re-estimate the model

- Each of these variance describes the amount of variance not accounted for in each dependent (endogenous) variable

Overall Model R^2 for All Endogenous Variables

- Although the residual variance and R^2 values for PERF and USE describe how each variable is explained individually, we can use multivariate statistics to describe the joint explanation of both
  - R^2 comparing the generalized variances (determinant of covariance matrix)

- The overall generalized variance of the endogenous variables without the model was $|\boldsymbol{\Sigma}| = \left|\begin{bmatrix} 8.709 & 6.362 \\ 6.362 & 249.254 \end{bmatrix}\right| = 2{,}130.28$

- The generalized **residual** variance of the endogenous variables was $|\widehat{\boldsymbol{\Sigma}}| = \left|\begin{bmatrix} 6.623 & 2.901 \\ 2.901 & 243.303 \end{bmatrix}\right| = 1{,}602.98$

- Therefore, the generalized R^2 was $\frac{2{,}130.28 - 1{,}602.98}{2{,}130.28} = .248$
  - Most of that came from the PERF variable

Comparison of Model Output from Linear and Multivariate Regression Models

Linear Regression

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.986	0.122	8.103	0.000
CC	0.079	0.028	2.861	0.004
HSL WITH				
CC	1.275	0.430	2.967	0.003
Means				
HSL	4.925	0.073	67.051	0.000
CC	10.331	0.331	31.189	0.000
Intercepts				
PERF	8.253	0.598	13.807	0.000
Variances				
HSL	1.726	0.123	14.039	0.000
CC	34.556	2.456	14.069	0.000
Residual Variances				
PERF	6.631	0.564	11.759	0.000

- Results for linear regression parameters will be virtually unchanged
- Here, they differ due to one extra observation included in model

Multivariate Regression

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.984	0.121	8.115	0.000
CC	0.079	0.028	2.865	0.004
USE ON				
HSL	1.466	0.632	2.320	0.020
CC	0.206	0.157	1.305	0.192
HSL WITH				
CC	1.290	0.430	3.001	0.003
USE WITH				
PERF	2.901	2.579	1.125	0.261
Means				
HSL	4.922	0.073	67.094	0.000
CC	10.330	0.331	31.194	0.000
Intercepts				
PERF	8.264	0.594	13.911	0.000
USE	43.129	3.338	12.920	0.000
Variances				
HSL	1.728	0.123	14.027	0.000
CC	34.561	2.456	14.069	0.000
Residual Variances				
PERF	6.623	0.563	11.755	0.000
USE	243.303	19.227	12.654	0.000

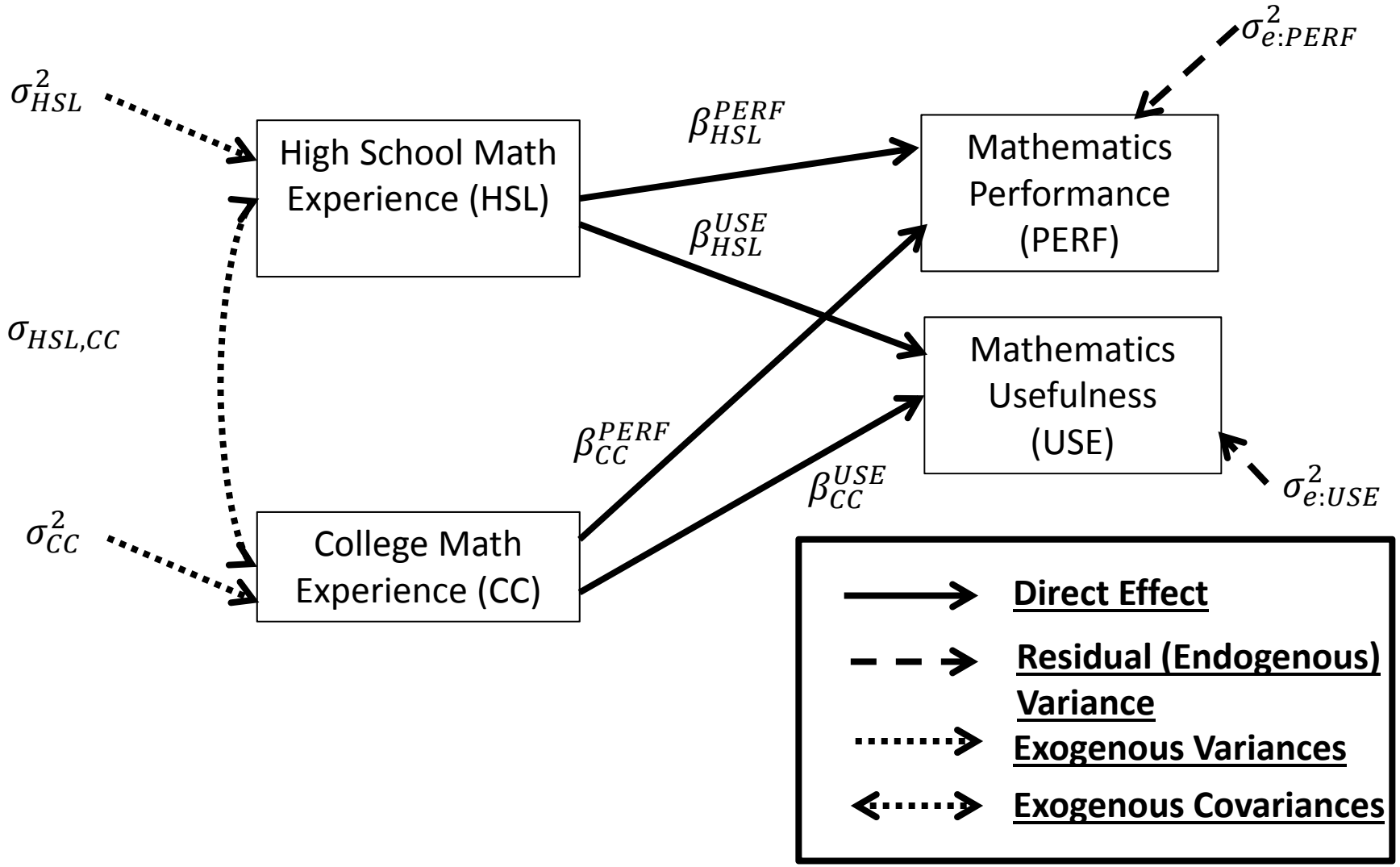
Model Modification

- The residual covariance parameter (between PERF and USE) was not significant
- This means that after accounting for the relationship between HSL and CC with PERF along with HSL and CC with USE, the correlation between these two is zero
  - Meaning we can likely remove the parameter from the model

```
MODEL:
    perf ON hsl cc;
    use ON hsl cc;
    hsl WITH cc;

    perf WITH use @0;
```

- Removal of the parameter from the model would reduce the number of estimated parameters from 14 to 13
  - And would provide a mechanism to inspect goodness of fit of the reduced model

Reduced Model Path Diagram



Model Fit Information

- The Mplus Model Fit Information section provides model fit statistics that can help judge the fit of a model
  - More frequently used in models with latent variables, but sometimes used in path analysis

- The important thing to note is that not all "good-fitting" models are useful…
  - More on this topic in two weeks

- The next few slides describe the statistics reported in this section of Mplus output

Log-likelihood Output

- The log-likelihood output section provides two log-likelihood values:
  - H0: the log-likelihood from the model run in the analysis
  - H1: the log-likelihood from the saturated (unstructured) model

```
Loglikelihood

          H0 Value                          -3573.439
          H0 Scaling Correction Factor          0.9584
            for MLR
          H1 Value                          -3572.730
          H1 Scaling Correction Factor          0.9714
            for MLR
```

- The log-likelihood is the log of the value from the likelihood function (the function that finds the parameters), evaluated at the peak
- If these statistics are identical, then you are running a model equivalent to the saturated model
  - No other model fit will be available or useful

Information Criteria Output

- The information criteria output provides relative fit statistics:

```
Information Criteria

          Akaike (AIC)                    7172.878
          Bayesian (BIC)                  7223.031
          Sample-Size Adjusted BIC        7181.790
            (n* = (n + 2) / 24)
```

  - AIC: Akaike Information Criterion
  - BIC: Bayesian Information Criterion (also called Schwarz's criterion)
  - Sample-size Adjusted BIC

- These statistics weight the information given by the parameter values by the parsimony of the model (the number of model parameters)
  - For all statistics, the smaller number is better
- The core of these statistics is -2*log-likelihood

Comparing Information Criteria

- The information from our reduced model (without the residual covariance):

```
Information Criteria

          Akaike (AIC)                       7172.878
          Bayesian (BIC)                     7223.031
          Sample-Size Adjusted BIC           7181.790
            (n* = (n + 2) / 24)
```

- The information criteria from our full model (with the residual covariance estimated):

```
Information Criteria

          Akaike (AIC)                       7173.459
          Bayesian (BIC)                     7227.470
          Sample-Size Adjusted BIC           7183.057
            (n* = (n + 2) / 24)
```

- For each statistic, the reduced model is preferred because values are smaller (so it fits better, relative to the full model)

Chi-Square Test of Model Fit

- The Chi-Square Test of Model Fit provides a likelihood ratio test comparing the current model to the **saturated (unstructured) model**:

Chi-Square Test of Model Fit

Value	1.245*
Degrees of Freedom	1
P-Value	0.2645
Scaling Correction Factor for MLR	1.1393

  - The value is -2 times the difference in log-likelihoods
  - The degrees of freedom is the difference in the number of estimated model parameters
  - The p-value is from the Chi-square distribution
- If this test has a significant p-value:
  - The current model (H0) is rejected – the model fit is significantly worse than the full model
  - However, in latent variable models, this test is usually ignored
    - Said to be overly sensitive
- If this test does not have a significant p-value:
  - The current model (H0) is not rejected – fits equivalently to full model

RMSEA
(Root Mean Square Error of Approximation)

- The RMSEA is an index of model fit where 0 indicates perfect fit (smaller is better):

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

          Estimate                           0.026
          90 Percent C.I.                    0.000  0.148
          Probability RMSEA <= .05           0.449
```

- RMSEA is based on the approximated covariance matrix
  - More on this in two weeks

- The goal is a model with an RMSEA less than .05
  - Although there is some flexibility

- The result above indicates our model fits well (RMSEA of .026)
  - Expected for 13 parameters (out of 14 possible)

CFI/TLI

- The CFI/TLI section provides two additional measures of model fit:

CFI/TLI

CFI	0.997
TLI	0.984

- CFI stands for Comparative Fit Index
  - Higher is better (above .95 indicates good fit)
  - Compares fit to independence model (uncorrelated variables)

- TLI stands for Tucker Lewis Index
  - Higher is better (above .95 indicates good fit)

- Both measures indicate good model fit (as they should for 13 parameters out of 14 possible)

Chi-Square Test of Model Fit for the Baseline Model

- The Chi-Square test of model fit for the baseline model provides a likelihood ratio test comparing **the saturated (unstructured) model** with an **independent variables model** (called the baseline model)

Chi-Square Test of Model Fit for the Baseline Model

Value	82.959
Degrees of Freedom	5
P-Value	0.0000

- Here, the "null" model is the baseline (the independent endogenous variables model)
  - If the test is significant, this means that at least one (and likely more than one) variable has a significant covariance
  - If the test is not significant, this means that the independence model is appropriate
    - This is not likely to happen
    - But if it does, there are virtually no other models that will be significant

Standardized Root Mean Squared Residual

- The SRMR (standardized root mean square residual) provides the average standardized difference between the observed correlation and the model-predicted correlation

```
SRMR (Standardized Root Mean Square Residual)

          Value                                   0.016
```

- Lower is better (some suggest less than 0.08)
- This indicates our model fits the data well (as it should for 13 out of 14 possible parameters in use)

Comparing Our Full and Reduced Multivariate Regression Models

Full Model

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two P
PERF ON				
HSL	0.984	0.121	8.115	
CC	0.079	0.028	2.865	
USE ON				
HSL	1.466	0.632	2.320	
CC	0.206	0.157	1.305	
HSL WITH				
CC	1.290	0.430	3.001	
USE WITH				
PERF	2.901	2.579	1.125	
Means				
HSL	4.922	0.073	67.094	
CC	10.330	0.331	31.194	
Intercepts				
PERF	8.264	0.594	13.911	
USE	43.129	3.338	12.920	
Variances				
HSL	1.728	0.123	14.027	
CC	34.561	2.456	14.069	
Residual Variances				
PERF	6.623	0.563	11.755	
USE	243.303	19.227	12.654	

Reduced Model

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.988	0.121	8.149	0.000
CC	0.079	0.028	2.861	0.004
USE ON				
HSL	1.473	0.632	2.331	0.020
CC	0.210	0.157	1.341	0.180
HSL WITH				
CC	1.293	0.430	3.008	0.003
PERF WITH				
USE	0.000	0.000	999.000	999.000
Means				
HSL	4.921	0.073	67.085	0.000
CC	10.331	0.331	31.196	0.000
Intercepts				
PERF	8.242	0.596	13.839	0.000
USE	43.074	3.334	12.918	0.000
Variances				
HSL	1.729	0.123	14.027	0.000
CC	34.563	2.457	14.068	0.000
Residual Variances				
PERF	6.617	0.563	11.747	0.000
USE	243.191	19.243	12.638	0.000

Multivariate Regression in Matrices

- Our linear regression equation was given by:

$$PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{CC}^{PERF} CC_i + e_i^{PERF}$$
$$USE_i = \beta_0^{USE} + \beta_{HSL}^{USE} HSL_i + \beta_{CC}^{USE} CC_i + e_i^{USE}$$

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

- Alternatively, we could rephrase this:

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

Where:

- $\boldsymbol{\zeta}_i = \begin{bmatrix} e_i^{PERF} \\ e_i^{USE} \end{bmatrix} \sim N_2(\mathbf{0}, \boldsymbol{\Psi})$ ($\boldsymbol{\Psi}$ is the $p\ x\ p$ residual covariance matrix)
- $\boldsymbol{\Gamma} = \begin{bmatrix} \beta_{HSL}^{PERF} & \beta_{CC}^{PERF} \\ \beta_{HSL}^{USE} & \beta_{CC}^{USE} \end{bmatrix}$ (matrix of size $p\ x\ q$ relating exogenous variables to endogenous variable(s))
- $\boldsymbol{x}_i = \begin{bmatrix} HSL_i \\ CC_i \end{bmatrix}$ (matrix of size $q\ x\ 1$ containing observed exogenous variables)
- $\boldsymbol{y}_i = \begin{bmatrix} PERF_i \\ USE_i \end{bmatrix}$ (matrix of size $p\ x\ 1$ containing observed endogenous variables)
- $\boldsymbol{\alpha} = \begin{bmatrix} \beta_0^{PERF} \\ \beta_0^{USE} \end{bmatrix}$ (matrix of size $p\ x\ 1$ containing intercepts for endogenous variables)

More Regression with Matrices

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

$$\boldsymbol{\mu}_x = \begin{bmatrix} \mu_{HSL} \\ \mu_{CC} \end{bmatrix}$$

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

$$\boldsymbol{\Phi} = \begin{bmatrix} \sigma^2_{HSL} & \sigma_{HSL,CC} \\ \sigma_{HSL,CC} & \sigma^2_{CC} \end{bmatrix}$$

- 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 conditional mean of the endogenous variables is:

$$\hat{\boldsymbol{\mu}}_y = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_x$$

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

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

- This is given in Mplus output:

$\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi}$

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	USE	HSL	CC
PERF	8.722			
USE	3.509	249.274		
HSL	1.611	2.818	1.729	
CC	4.001	9.177	1.293	34.563

$\boldsymbol{\Phi}$

$\boldsymbol{\Phi\Gamma}^T$

Matching Matrices with Results

- To more specifically link our results (for both the full and reduced model) to the matrices from the previous page:
  - Note: difference between models is one parameter – the covariance between residuals is 0 in the reduced model

Name	Matrix	Full Model Estimates	Reduced Model Estimates
Residual Covariance Matrix	$\boldsymbol{\Psi}$	$\begin{bmatrix} 6.623 & 2.901 \\ 2.901 & 243.303 \end{bmatrix}$	$\begin{bmatrix} 6.617 & 0 \\ 0 & 243.191 \end{bmatrix}$
Regression Weights of Exogenous onto Endogenous	$\boldsymbol{\Gamma}$	$\begin{bmatrix} 0.984 & 0.079 \\ 1.466 & 0.206 \end{bmatrix}$	$\begin{bmatrix} 0.988 & 0.079 \\ 1.473 & 0.210 \end{bmatrix}$
Covariance Matrix of Exogenous Variables	$\boldsymbol{\Phi}$	$\begin{bmatrix} 1.728 & 1.290 \\ 1.290 & 34.561 \end{bmatrix}$	$\begin{bmatrix} 1.729 & 1.293 \\ 1.293 & 34.563 \end{bmatrix}$
Mean Vector of Exogenous Variables	$\boldsymbol{\mu}_x$	$\begin{bmatrix} 4.922 \\ 10.330 \end{bmatrix}$	$\begin{bmatrix} 4.921 \\ 10.331 \end{bmatrix}$
Vector of Endogenous Variable Intercepts	$\boldsymbol{\alpha}$	$\begin{bmatrix} 8.264 \\ 43.129 \end{bmatrix}$	$\begin{bmatrix} 8.242 \\ 43.074 \end{bmatrix}$

Predicted Model Mean Vectors

- **FULL MODEL:** The unconditional mean vector of the endogenous variables is:

$$\boldsymbol{\mu}_y = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_x = \begin{bmatrix} 8.264 \\ 43.129 \end{bmatrix} + \begin{bmatrix} 0.984 & 0.079 \\ 1.466 & 0.206 \end{bmatrix} \begin{bmatrix} 4.922 \\ 10.330 \end{bmatrix} = \begin{bmatrix} 13.919 \\ 52.466 \end{bmatrix}$$

The **FULL MODEL *Perfectly*** Reproduces "Saturated" Mean Vector

Model Estimated Means/Intercepts/Thresholds

	PERF	USE	HSL	CC
1	13.919	52.466	4.922	10.330

Residuals for Means/Intercepts/Thresholds

	PERF	USE	HSL	CC
1	0.000	0.000	0.000	0.000

- **REDUCED MODEL:** The unconditional mean vector of the endogenous variables is:

$$\boldsymbol{\mu}_y = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\mu}_x = \begin{bmatrix} 8.242 \\ 43.074 \end{bmatrix} + \begin{bmatrix} 0.988 & 0.079 \\ 1.473 & 0.210 \end{bmatrix} \begin{bmatrix} 4.921 \\ 10.331 \end{bmatrix} = \begin{bmatrix} 13.920 \\ 52.497 \end{bmatrix}$$

The **REDUCED MODEL *NEARLY*** Reproduces "Saturated" Mean Vector

Model Estimated Means/Intercepts/Thresholds

	PERF	USE	HSL	CC
1	13.920	52.497	4.921	10.331

Residuals for Means/Intercepts/Thresholds

	PERF	USE	HSL	CC
1	-0.002	-0.030	0.000	0.000

Full Model Predicted Covariance Matrices

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

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

$$\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi} = \begin{bmatrix} 0.984 & 0.079 \\ 1.466 & 0.206 \end{bmatrix} \begin{bmatrix} 1.728 & 1.290 \\ 1.290 & 34.561 \end{bmatrix} \begin{bmatrix} 0.984 & 1.466 \\ 0.079 & 0.206 \end{bmatrix} + \begin{bmatrix} 6.623 & 2.901 \\ 2.901 & 243.303 \end{bmatrix} = \begin{bmatrix} 8.709 & 6.362 \\ 6.362 & 249.252 \end{bmatrix}$$

$$\boldsymbol{\Gamma\Phi} = \begin{bmatrix} 0.984 & 0.079 \\ 1.466 & 0.206 \end{bmatrix} \begin{bmatrix} 1.728 & 1.290 \\ 1.290 & 34.561 \end{bmatrix} = \begin{bmatrix} 1.801 & 3.992 \\ 2.798 & 8.994 \end{bmatrix}$$

$$\boldsymbol{\Phi\Gamma}^T = \begin{bmatrix} 1.728 & 1.290 \\ 1.290 & 34.561 \end{bmatrix} \begin{bmatrix} 0.984 & 1.466 \\ 0.079 & 0.206 \end{bmatrix} = \begin{bmatrix} 1.801 & 2.798 \\ 3.992 & 8.994 \end{bmatrix}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} 1.728 & 1.290 \\ 1.290 & 34.561 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{y,x} = \begin{bmatrix} 8.709 & 6.362 & 1.801 & 3.992 \\ 6.362 & 249.252 & 2.798 & 1.728 \\ 1.801 & 2.798 & 1.728 & 1.290 \\ 3.992 & 8.994 & 1.290 & 34.561 \end{bmatrix}$$

Full Model Predicted and Residual Covariance Matrices – in Mplus

- The FULL MODEL exactly reproduces the covariance matrix of endogenous and exogenous variables:

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	USE	HSL	CC
PERF	8.709			
USE	6.362	249.252		
HSL	1.801	2.798	1.728	
CC	3.992	8.994	1.290	34.561

Residuals for Covariances/Correlations/Residual Correlations

	PERF	USE	HSL	CC
PERF	0.000			
USE	0.000	0.002		
HSL	0.000	0.000	0.000	
CC	0.000	0.000	0.000	0.000

- Sense a trend? This will always be the case – a model with 0 degrees of freedom will always reproduce the covariance matrix of the saturated (unstructured) model
  - The parameters explain the covariances in more meaningful ways

Reduced Model Predicted Covariance Matrices

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

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

$$\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi} = \begin{bmatrix} 0.988 & 0.079 \\ 1.473 & 0.210 \end{bmatrix} \begin{bmatrix} 1.729 & 1.293 \\ 1.293 & 34.563 \end{bmatrix} \begin{bmatrix} 0.988 & 1.473 \\ 0.079 & 0.210 \end{bmatrix} + \begin{bmatrix} 6.617 & 0 \\ 0 & 243.191 \end{bmatrix} = \begin{bmatrix} 8.722 & 3.509 \\ 3.509 & 249.274 \end{bmatrix}$$

$$\boldsymbol{\Gamma\Phi} = \begin{bmatrix} 0.988 & 0.079 \\ 1.473 & 0.210 \end{bmatrix} \begin{bmatrix} 1.729 & 1.293 \\ 1.293 & 34.563 \end{bmatrix} = \begin{bmatrix} 1.811 & 4.001 \\ 2.818 & 9.177 \end{bmatrix}$$

$$\boldsymbol{\Phi\Gamma}^T = \begin{bmatrix} 1.729 & 1.293 \\ 1.293 & 34.563 \end{bmatrix} \begin{bmatrix} 0.988 & 1.473 \\ 0.079 & 0.210 \end{bmatrix} = \begin{bmatrix} 1.811 & 2.818 \\ 4.001 & 9.177 \end{bmatrix}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} 1.729 & 1.293 \\ 1.293 & 34.563 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{y,x} = \begin{bmatrix} 8.722 & 3.509 & 1.811 & 4.001 \\ 3.509 & 249.274 & 2.818 & 9.177 \\ 1.811 & 2.818 & 1.729 & 1.293 \\ 4.001 & 9.177 & 1.293 & 34.563 \end{bmatrix}$$

Reduced Model Predicted and Residual Covariance Matrices – in Mplus

- The REDUCED MODEL does not exactly reproduce the covariance matrix of endogenous and exogenous variables:

Model Estimated Covariances/Correlations/Residual Correlations

	PERF	USE	HSL	CC
PERF	8.722			
USE	3.509	249.274		
HSL	1.811	2.818	1.729	
CC	4.001	9.177	1.293	34.563

Residuals for Covariances/Correlations/Residual Correlations

	PERF	USE	HSL	CC
PERF	-0.013			
USE	2.853	-0.020		
HSL	-0.010	-0.021	-0.001	
CC	-0.009	-0.183	-0.003	-0.002

- Note: the position of greatest discrepancy is for the covariance of PERF and USE
  - The location where the residual covariance would matter
- The question of model fit statistics is whether "close fit" is close enough – does the model fit well enough

“BASIC” PATH ANALYSIS

Basic Path Model

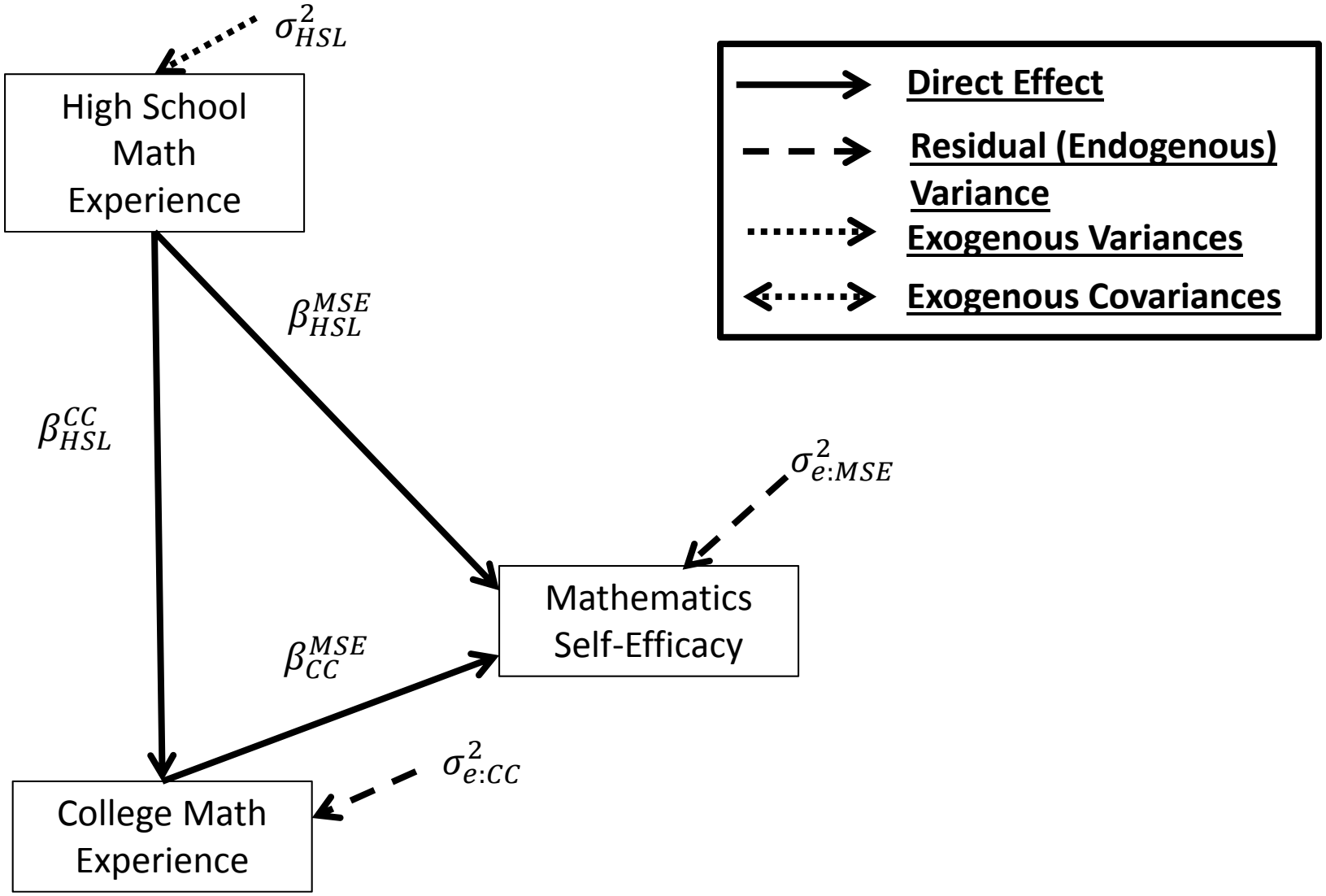
- To demonstrate path analysis using a limited set of variables, we will examine the relationship between three of our observed variables: HSL (high school experience), CC (college experience), and MSE (math self-efficacy)

- Specifically, we seek to investigate the following regression equations, simultaneously:

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

- We are hypothesizing that:
  - College experience is predicted by high school experience
  - Math self efficacy is predicted by high school experience (directly and indirectly) and college experience directly

Basic Path Model



Labeling Variables

- The independent (exogenous) variable in our analysis is:
  - High School Experience (HSL) – nothing explains why it varies

- The dependent (endogenous) variables in our analysis are:
  - College Experience (CC) – is explained (predicted) by HSL
  - Math Self-Efficacy (MSE) – is explained (predicted) by HSL and CC

- College Experience (CC) is both predicted and a predictor variable
  - If any variable is predicted at all, it is endogenous
  - Path models allow endogenous variables to predict other variables

- Also note that High School Experience (HSL) predicts Math Self Efficacy (MSE) directly and indirectly
  - Direct examines how HSL predicts MSE by itself
  - Indirect examines how relationship between HSL and MSE **is mediated** by CC (more on mediation later in course)
  - Combination of effects speaks of how variables interact with other variables in a path model

Path Model Parameters

- If we considered all three variables to be part of a multivariate normal distribution, our unstructured (saturated) model would have 9 parameters:
  - 3 means
  - 3 variances
  - 3 covariances (3-choose-2 or 3*(3-1)/2))

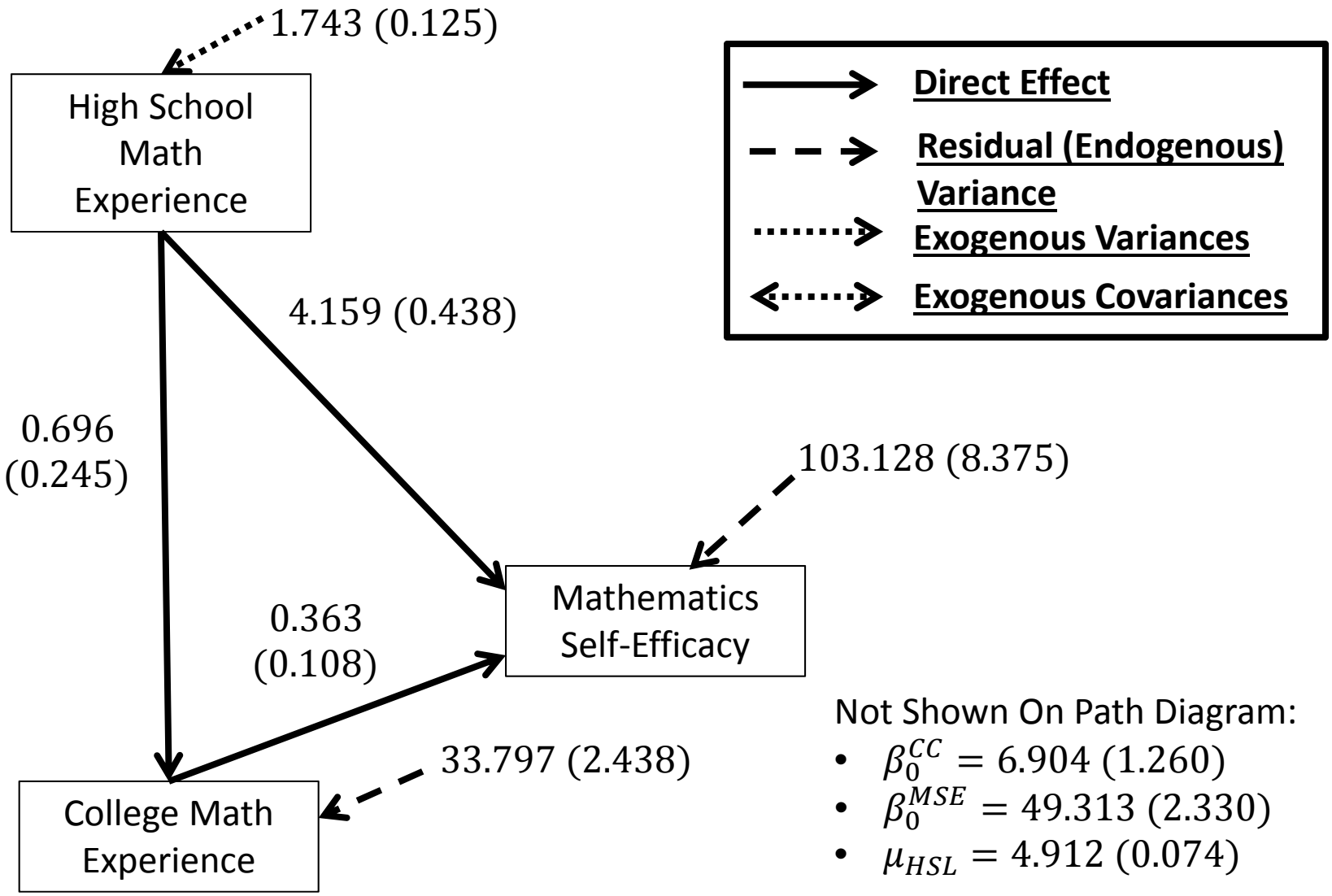
- The model itself has 9 parameters:
  - 2 intercepts
  - 3 slopes
  - 2 residual variances
  - 1 exogenous variance
  - 1 exogenous mean

- Therefore, this model will fit perfectly – no model fit statistics will be available
  - Even without model fit, interpretation of parameters can proceed

Mplus Syntax for Path Model

```
TITLE:
    Basic Path Model
    Predicting MSE and CC

DATA:
    FILE = mathdata.csv;

VARIABLE:
    NAMES = id gender hsl cc use msc mas mse perf;
    USEVARIABLE = hsl cc mse;
    IDVARIABLE = id;
    MISSING = .;

ANALYSIS:
    ESTIMATOR = MLR;

MODEL:
    cc ON hsl;
    mse ON hsl cc;
    hsl; !provides variance estimate for HSL (puts values into likelihood)

MODEL INDIRECT:
    mse IND hsl; !requests calculation of direct and indirect effects

OUTPUT:
    STANDARDIZED RESIDUAL SAMPSTAT;
```

Basic Path Model Results



Not Shown On Path Diagram:

- $\beta_0^{CC} = 6.904\ (1.260)$
- $\beta_0^{MSE} = 49.313\ (2.330)$
- $\mu_{HSL} = 4.912\ (0.074)$

Interpretation of Path Model Parameters: CC

- Because no test of model fit is possible, we can immediately go onto examination of the parameters:
  - $\beta_0^{CC} = 6.904$: the intercept for CC; the predicted value of CC when all predictors of CC are zero (HSL = 0)
  - $\beta_{HSL}^{CC} = 0.696$: the direct effect of HSL on CC (analogous to a regression slope). For every one-unit increase in HSL, CC increases by 0.696
    - The standardized coefficient was 0.156

Interpretation of Path Model Parameters: MSE

- For the Math Self-Efficacy variable, the path model parameters were:
  - $\beta_0^{MSE} = 49.313$: the intercept for MSE; the predicted value of MSE when all predictors of MSE are zero (HSL = 0 and CC = 0)
  - $\beta_{HSL}^{MSE} = 4.159$: the direct effect of HSL on MSE. For every one-unit increase in HSL, MSE increases by 4.159 (holding CC constant – an important distinction)
    - The standardized coefficient was .462
  - $\beta_{CC}^{MSE} = 0.363$: the direct effect of CC on MSE. For every one-unit increase in CC, MSE increases by 0.363 (holding HSL constant)
    - The standardized coefficient was .180

Interpretation of Residual Variances

- $\sigma^2_{e:CC} = 33.797$: the residual variance for CC
  - The R^2 for CC was .024 (a very small effect)

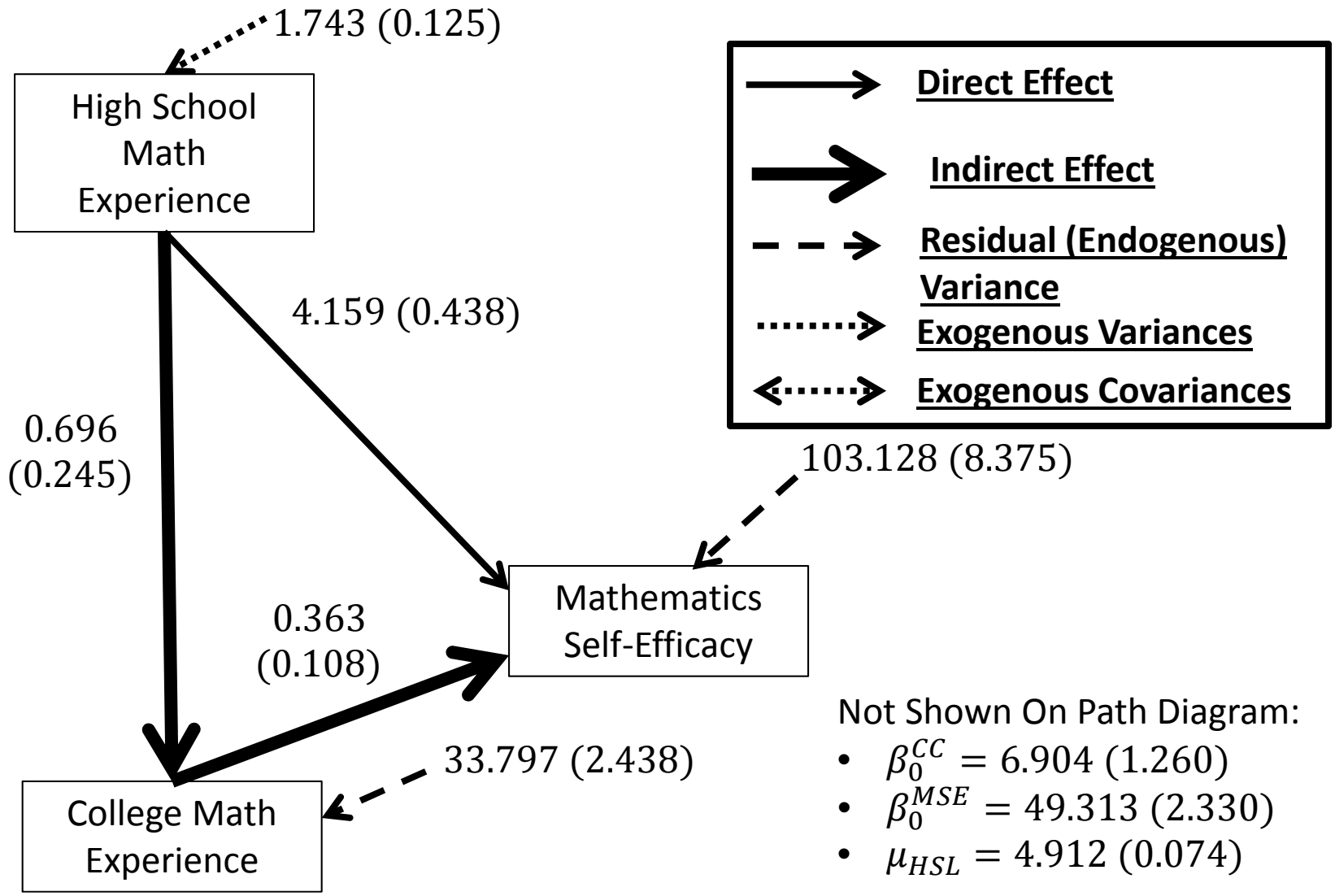
- $\sigma^2_{e:MSE} = 103.128$: the residual variance for MSE
  - The R^2 for MSE was .271

Indirect Paths

- Because High School Experience (HSL) predicted College Experience (CC) and College Experience (CC) predicted Math Self-Efficacy (MSE), an indirect path between HSL and MSE exists
  - An indirect path represents the effect of one variable on another, as mediated by one or more variables

- The indirect path suggests that the relationship between High School Experience (HSL) and Math Self-Efficacy is mediated by College Experience (CC)
  - More formally, the mediational relationship is hypothesized by the path model, a formal test of hypothesis is needed to establish College Experience as a mediator of High School Experience and Math Self-Efficacy

Direct and Indirect Effects of HSL on MSE



Not Shown On Path Diagram:

- $\beta_0^{CC} = 6.904\ (1.260)$
- $\beta_0^{MSE} = 49.313\ (2.330)$
- $\mu_{HSL} = 4.912\ (0.074)$

Calculation of Indirect Effects

- The indirect effect of High School Experience on Math Self-Efficacy is the combination of two path coefficients:
  - The path between High School (HSL) and College (CC) Experience: $\beta_{HSL}^{CC} = 0.696$
  - The path between College Experience (CC) and Math Self-Efficacy (MSE): $\beta_{CC}^{MSE} = 0.363$

- The **indirect effect** of HSL on MSE is the product of these two terms: $\beta_{HSL}^{CC}\beta_{CC}^{MSE} = 0.696^{*}0.363 = 0.253$

- The indirect effect is the amount of increase in the outcome variable (MSE in this case) that comes indirectly by a one-unit increase in the initiating variable (HSL in this case)
  - As HSL increases by one unit, CC increases by 0.696 (the direct effect of HSL on CC)
  - Then, as CC increases by 0.696, HSL increases by 0.393 (the direct effect of CC on MSE)

- Indirectly, MSE increases by 0.253 (the multiplication of the two direct effects) for every one unit increase of HSL

Hypothesis Tests for Indirect Effects

- Of importance in the understanding of mediating variables is the test of hypothesis for the indirect effect
  - If the indirect effect (the product of the two direct effects) is significant, then the third variable is said to be a mediator

- Hypothesis tests for the indirect effect have become a hot topic in recent years
  - We will discuss this more in the mediation lecture later in class
  - For now, we will stick with the test of the indirect effect given to use by Mplus using the "MODEL INDIRECT" command
    - This test uses a delta-method transformation (relevant if you are publishing in this area)

```
MODEL INDIRECT:
    mse IND hsl;
```

Total Effects

- Finally, of concern in mediational models and general path models is the total effect of one variable on another
- The **total effect** is the sum of all direct and indirect effects
  - It represents the **total** increase in the outcome variable for a one-unit increase in the initiating variable

- In our example, the total effect of High School Experience (HSL) on Math Self-Efficacy (MSE) is the sum of the direct and indirect effects:

$$\beta_{HSL}^{MSE} + \beta_{HSL}^{CC}\beta_{CC}^{MSE} = 4.159 + 0.696^{*}0.363 = 4.412$$

- This means that for every one-unit increase in HSL, the total increase in MSE is 4.412
  - The direct effect represents the increase holding CC constant, which is implausible in this model

Mplus Output

- The MODEL INDIRECT command provides the total and indirect effects between terminating and originating variables
  - If the STANDARDIZE command is included in the OUTPUT section, the standardized versions of these effects are also given (the increase in standard deviations)

TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from HSL to MSE				
Total	4.412	0.443	9.969	0.000
Total indirect	0.253	0.109	2.326	0.020
Specific indirect				
MSE				
CC				
HSL	0.253	0.109	2.326	0.020
Direct				
MSE				
HSL	4.159	0.438	9.492	0.000

STANDARDIZED TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from HSL to MSE				
Total	0.490	0.044	11.097	0.000
Total indirect	0.028	0.012	2.360	0.018
Specific indirect				
MSE				
CC				
HSL	0.028	0.012	2.360	0.018
Direct				
MSE				
HSL	0.462	0.045	10.336	0.000

- Here, our output suggest the indirect effect is significant, so we say that CC mediates the relationship between HSL and MSE

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

Rebuilding our Equations From Matrices

- To elaborate on how we get back to our path model equations from matrix form:

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

$$\begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} = \begin{bmatrix} \beta_0^{CC} \\ \beta_0^{MSE} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \beta_{CC}^{MSE} & 0 \end{bmatrix} \begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} + \begin{bmatrix} \beta_{HSL}^{CC} \\ \beta_{HSL}^{MSE} \end{bmatrix} [HSL_i] + \begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix}$$

$$\begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} = \begin{bmatrix} \beta_0^{CC} \\ \beta_0^{MSE} \end{bmatrix} + \begin{bmatrix} 0 \\ \beta_{cc}^{MSE} CC_i \end{bmatrix} + \begin{bmatrix} \beta_{HSL}^{CC} HSL_i \\ \beta_{HSL}^{MSE} HSL_i \end{bmatrix} + \begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix}$$

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

$$\ldots$$

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

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

Rebuilding our Equations From Matrices ...and Estimates

- To elaborate on how we get back to our path model equations from matrix form:

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

$$\begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} = \begin{bmatrix} \beta_0^{CC} \\ \beta_0^{MSE} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \beta_{CC}^{MSE} & 0 \end{bmatrix} \begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} + \begin{bmatrix} \beta_{HSL}^{CC} \\ \beta_{HSL}^{MSE} \end{bmatrix} [HSL_i] + \begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix}$$

$$\begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} = \begin{bmatrix} 6.904 \\ 49.313 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0.363 & 0 \end{bmatrix} \begin{bmatrix} CC_i \\ MSE_i \end{bmatrix} + \begin{bmatrix} 0.696 \\ 4.159 \end{bmatrix} [HSL_i] + \begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix}$$

...

$$CC_i = 6.904 + 0.696HSL_i + e_i^{CC}$$

$$MSE_i = 49.313 + 0.636CC_i + 4.159HSL_i + e_i^{MSE}$$

Where $\begin{bmatrix} e_i^{CC} \\ e_i^{MSE} \end{bmatrix} \sim N_2\left(\mathbf{0}, \boldsymbol{\Psi} = \begin{bmatrix} 33.797 & 0 \\ 0 & 103.128 \end{bmatrix}\right)$

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 (discussed shortly)

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

IDENTIFICATION OF PATH MODELS

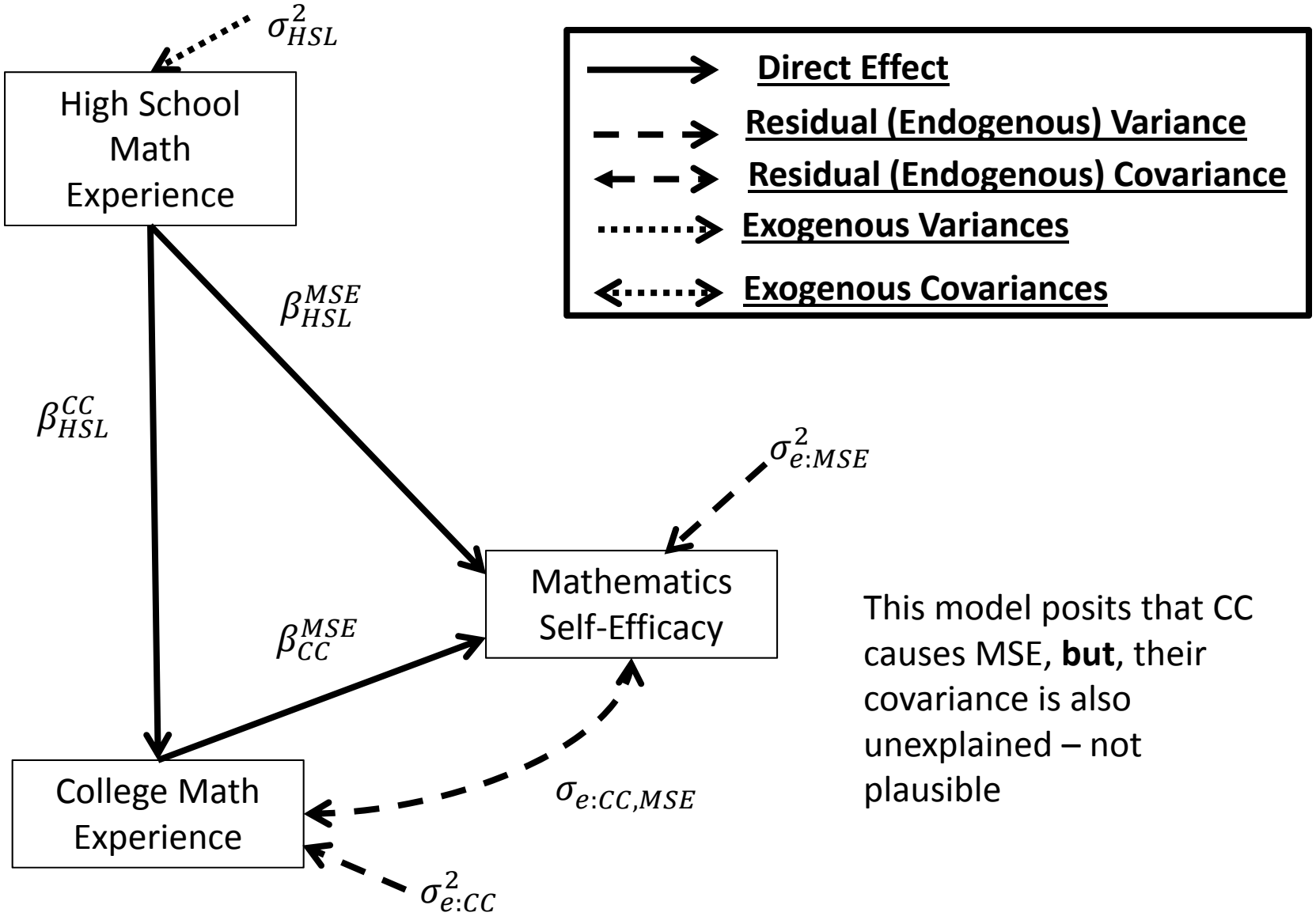
Path Model Identification

- You may have noticed that unlike the multivariate regression analysis, our path model did not have a covariance between the residuals of the endogenous variables
  - The reason for this is that we would have one more parameter than we have degrees of freedom - an unidentified model

- See the error from Mplus when trying to add this parameter:

```
     MAXIMUM LOG-LIKELIHOOD VALUE FOR THE UNRESTRICTED (H1) MODEL IS   -2713.834


     THE DEGREES OF FREEDOM FOR THIS MODEL ARE NEGATIVE.  THE MODEL IS NOT
     IDENTIFIED.  NO CHI-SQUARE TEST IS AVAILABLE.  CHECK YOUR MODEL.

THE MODEL ESTIMATION TERMINATED NORMALLY

     THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES COULD NOT BE
     COMPUTED.  THE MODEL MAY NOT BE IDENTIFIED.  CHECK YOUR MODEL.
     PROBLEM INVOLVING PARAMETER 9.

     THE CONDITION NUMBER IS      -0.335D-10.
```

Basic Path Model + Residual Covariance



This model posits that CC causes MSE, **but**, their covariance is also unexplained – not plausible

Identification of Path Models

- Model identification is necessary for statistical models to have meaningful results
  - From the error on the previous slide, we essentially had too many unknown values (parameters) and not enough places to put the parameters in the model
- For path models, identification can be a very difficult thing to understand
  - We will stick to the basics here
- Because of their unique structure, path models must have identification in two ways:
  - "Globally" – so that the total number of parameters does not exceed the total number of means, variances, and covariances of the endogenous and exogenous variables
  - "Locally" – so that each individual equation is identified
- Identification is guaranteed if a model is both "globally" and "locally" identified

Global Identification: "T-rule"

- A necessary but not sufficient condition for a path models is that of having equal to or fewer model parameters than there are distributional parameters

- As the path models we discuss assume the multivariate normal distribution, we have two matrices of parameters with which to work
  - Distributional parameters: the elements of the mean vector and (or more precisely) the covariance matrix

- For the MVN, the so-called T-rule states that a model must have equal to or fewer parameters than the unique elements of the covariance matrix of all endogenous and exogenous variables (the sum of all variables in the analysis)
  - Let $s = p + q$, the total of all endogenous (p) and exogenous (q) variables
  - Then the total unique elements are $\frac{s(s+1)}{2}$

More on the "T-rule"

- The classical definition of the "T-rule" counts the following entities as model parameters:
  - Direct effects (regression slopes)
  - Residual variances
  - Residual covariances
  - Exogenous variances
  - Exogenous covariances

- Missing from this list are:
  - The set of exogenous variable means
  - The set of intercepts for endogenous variables

- Each of the missing entities are part of the Mplus likelihood function, but are considered "saturated" so no additional parameters can be added
  - These do not enter into the equation for the covariance matrix of the endogenous and exogenous variables

Global Identification of Our Examples

Model	Endogenous Variables (p)	Exogenous Variables (q)	Unique Covariance Matrix Elements	Model Parameters (excluding Exo. Var. means)	Identification Status
Linear Regression (slide 14)	1	2	$\frac{3^*(3+1)}{2} = 6$	6: $\sigma^2_{HSL}, \sigma_{HSL,CC}, \sigma^2_{CC}, \beta^{PERF}_{HSL}, \beta^{PERF}_{CC}, \sigma^2_{e:PERF}$	Just Identified
Multivariate Regression Full Model (slide 39)	2	2	$\frac{4^*(4+1)}{2} = 10$	10: $\sigma^2_{HSL}, \sigma_{HSL,CC}, \sigma^2_{CC}, \beta^{PERF}_{HSL}, \beta^{PERF}_{CC}, \beta^{USE}_{HSL} \beta^{USE}_{CC}, \sigma^2_{e:PERF}, \sigma^2_{e:USE}, \sigma_{e:PERF,USE}$	Just Identified
Multivariate Regression Reduced Model (slide 49)	2	2	$\frac{4^*(4+1)}{2} = 10$	9: $\sigma^2_{HSL}, \sigma_{HSL,CC}, \sigma^2_{CC}, \beta^{PERF}_{HSL}, \beta^{PERF}_{CC}, \beta^{USE}_{HSL} \beta^{USE}_{CC}, \sigma^2_{e:PERF}, \sigma^2_{e:USE}$	Over-Identified
Path Model (slide 71)	2	1	$\frac{3^*(3+1)}{2} = 6$	6:$\sigma^2_{HSL}, \sigma^2_{e:CC}, \sigma^2_{e:MSE} \beta^{CC}_{HSL}, \beta^{MSE}_{CC}, \beta^{MSE}_{HSL}$	Just Identified

T-rule Identification Status

- **Just-Identified:** number of covariances = number of model parameters
  - Necessary for identification, but no model fit indices available

- **Over-Identified:** number of covariances > number of model parameters
  - Necessary for identification; model fit indices available

- **Under-Identified:** number of covariances < number of model parameters
  - **Model is NOT IDENTIFIED**: No results available
  - Do not pass go…do not collect $200

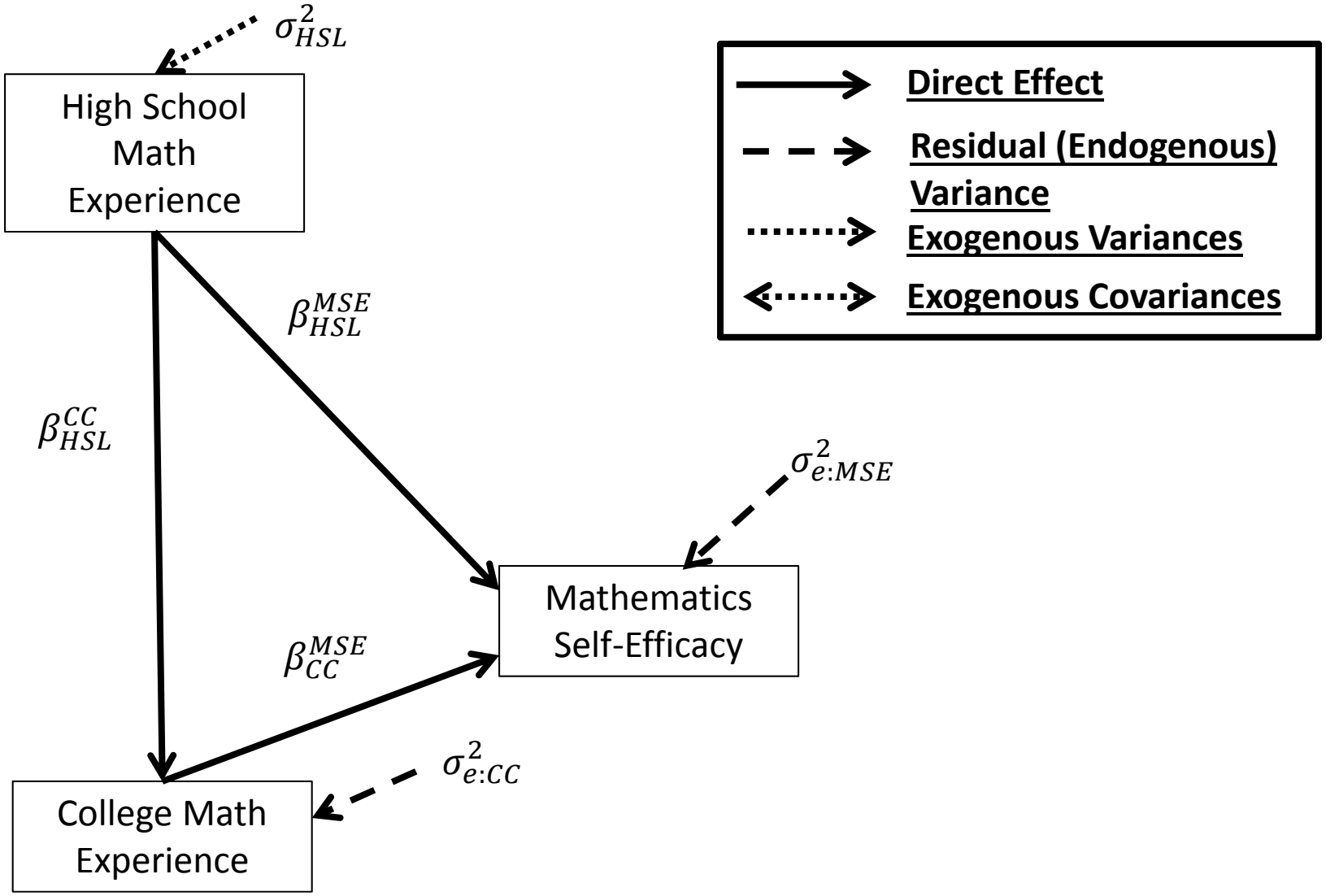
Moving from Global to Local Identification: Types of Path Models

- For most research designs, global identification will suffice
  - For the most part, **recursive path models** will be identified if the "t-rule" is met

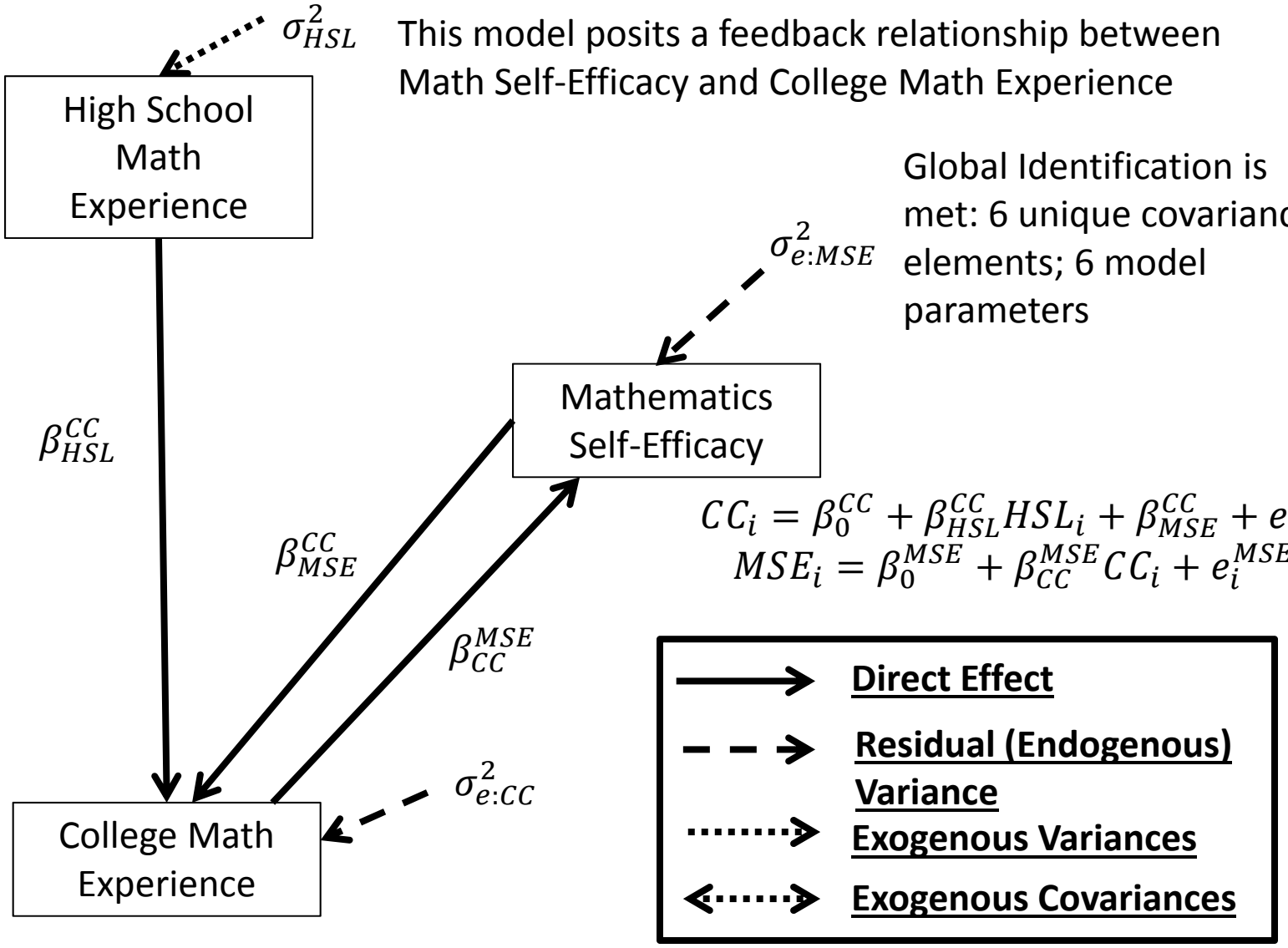
- A **recursive path model** is one where the direct effects are unidirectional – no feedback loops
  - Our path model is an example of a recursive path model

- A **non-recursive path model** is one where the direct effects are bidirectional for some variables – feedback loops are present
  - Difficult to envision using cross-sectional data
  - More frequent in econometrics
  - Different estimation algorithms used (see the next few slides)

Basic Path Model: Recursive



A Non-Recursive Path Model



This model posits a feedback relationship between Math Self-Efficacy and College Math Experience

Global Identification is met: 6 unique covariance elements; 6 model parameters

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

Mplus Estimates

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
CC ON				
HSL	30.536	48.227	0.633	0.527
MSE	-6.764	10.890	-0.621	0.535
MSE ON				
CC	6.335	2.340	2.707	0.007
Means				
HSL	4.912	0.074	66.671	0.000
Intercepts				
CC	357.406	564.278	0.633	0.526
MSE	8.103	24.205	0.335	0.738
Variances				
HSL	1.743	0.140	12.485	0.000
Residual Variances				
CC	5121.175	16119.682	0.318	0.751
MSE	1308.504	949.958	1.377	0.168

Mplus will estimate this model; however, estimates indicate some type of problem in the analysis

Step #2: Local Identification

- If a model is globally identified (just- or over-identified), then the next step is to determine if it is locally identified
  - Mostly an issue for non-recursive models

- Local identification is verified by satisfying the rank condition

- The rank condition starts with the augmented matrix:

$$\mathbf{A} = [(\mathbf{I} - \mathbf{B}) \quad | \quad \mathbf{\Gamma}]$$

- For the previous model, this would be:

$$\mathbf{A} = \begin{bmatrix} 1 & -\beta_{MSE}^{CC} & \beta_{HSL}^{CC} \\ -\beta_{CC}^{MSE} & 1 & 0 \end{bmatrix}$$

Rank Condition, Continued

- Then, remove the non-zero elements of matrix $\mathbf{A}$

$$\mathbf{A}_1^* = \begin{bmatrix} \beta_{HSL}^{CC} \\ 0 \end{bmatrix}$$

- The places where non-zeros occur correspond to the specific equations in the model
  - There is only one here – in the first row of $\mathbf{A}_1^*$
  - Our first hint that local identification is not met

- The row rank (roughly- the number of rows that have some number other than zero in them) of the sub-matrices must be equal to or greater than the number of exogenous variables minus 1 ($p - 1$)
  - For equation 1, the rank is 1 (and $p - 1 = 1$), so it is identified
  - For equation 2, the rank is 0 (no matrix left), so it is **not identified**

- **Local non-identification was the cause of Mplus issues**
  - Mplus still attempted to estimate the model and did not return an error

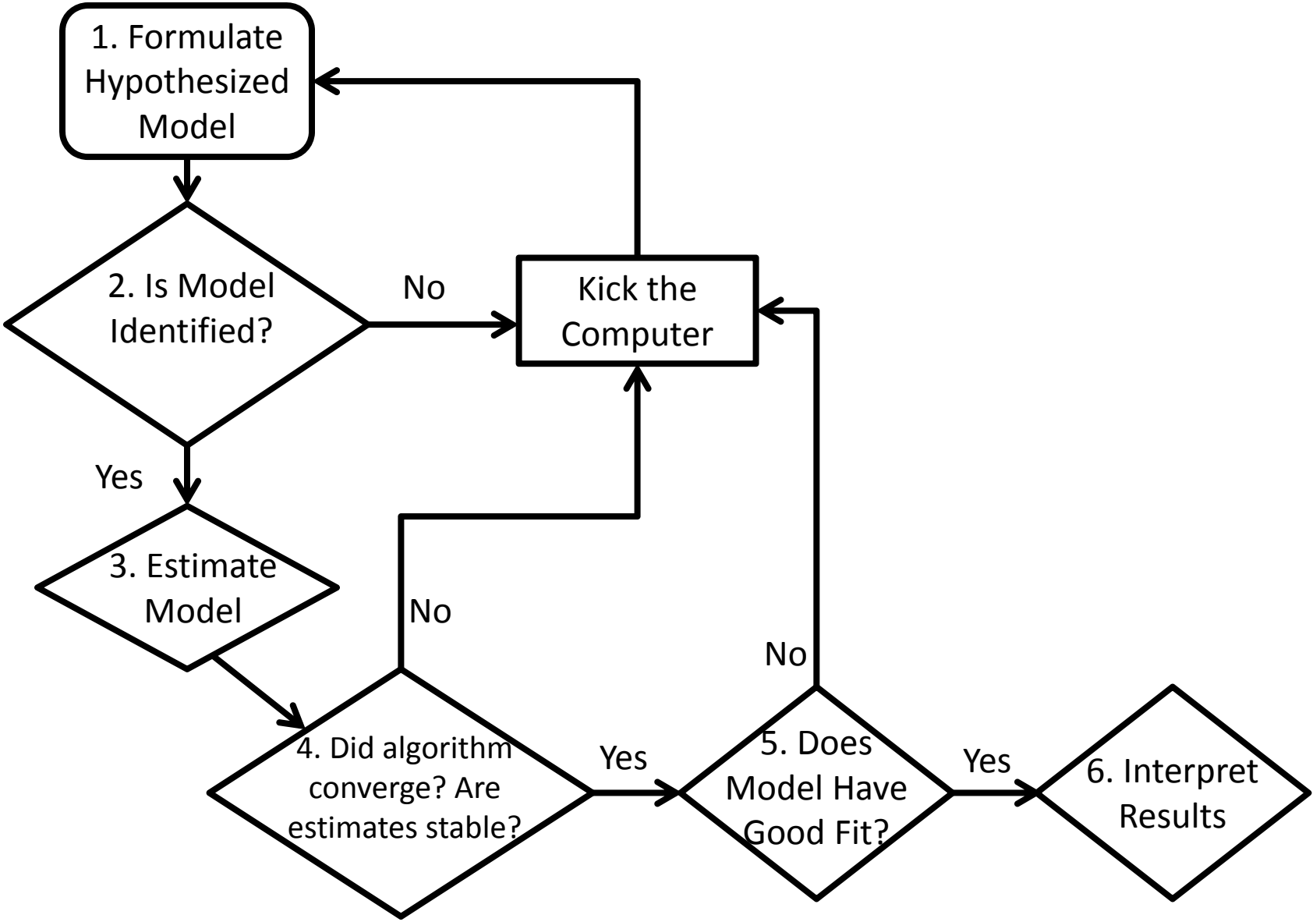
Guiding Identification Principals

- If you have a recursive model (no feedback loops) make sure:
  - # of model parameters $\leq$ # of unique covariance elements
  - No undirected paths (residual covariance) connecting variables with direct effects
    - Does not make sense to say one variable causes another yet their correlation is unexplained

- If you have a non-recursive model (feedback loops):
  - Think critically about whether such a model can be investigated by your data (cross-sectional versus longitudinal)
  - Attempt to determine if the model meets the rank condition
  - Investigate model output for irregularities (very large effects relative to the scale of the variables

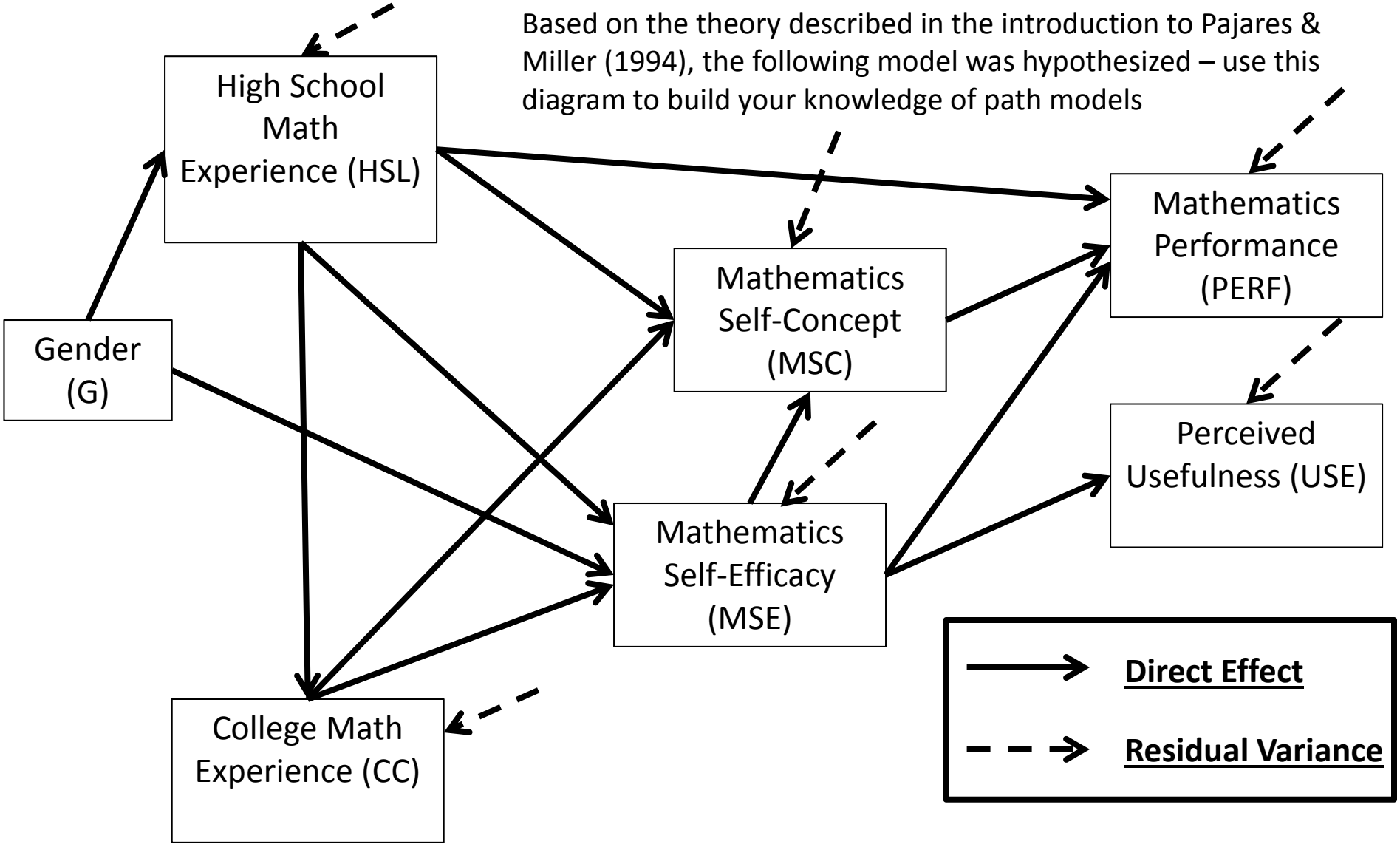
THE FINAL PATH MODEL: PUTTING IT ALL TOGETHER

A Path Model of Path Analysis Steps



Our Destination: Overall Path Model

Based on the theory described in the introduction to Pajares & Miller (1994), the following model was hypothesized – use this diagram to build your knowledge of path models



Path Model Setup – Questions for the Analysis

- How many variables are in our model? 7
  - Gender, HSL, CC, MSC, MSE, PERF, and USE
- How many variables are endogenous? 6
  - HSL, CC, MSC, MSE, PERF, and USE
- How many variables are exogenous? 1
  - Gender
- Is the model recursive or non-recursive?
  - Recursive – no feedback loops present

Path Model Setup – Questions for the Analysis

- Is the model identified?
  - Check the t-rule first (and only as it is recursive)
  - How many covariance terms are there in the all-variable matrix?

$$\frac{7^{*}(7+1)}{2} = 28$$

  - How many model parameters are to be estimated?
    - 12 direct paths
    - 6 residual variances
    - 1 variance of the exogenous variable
    - **(19 model parameters for the covariance matrix)**
    - 6 endogenous variable intercepts
      - Not relevant for t-rule identification, but counted in Mplus
- **The model is over-identified**
  - 28 total variance/covariances but 19 model parameters
  - We can use Mplus to run our analysis

Overall Hypothesized Path Model: Equation Form

- The path model from can be re-expressed in the following 6 endogenous variable regression equations:

1. $HSL_i = \beta_0^{HSL} + \beta_G^{HSL} G_i + e_i^{HSL}$
2. $CC_i \quad = \beta_0^{CC} + \beta_{HSL}^{CC} HSL_i + e_i^{CC}$
3. $MSE_i = \beta_0^{MSE} + \beta_G^{MSE} G_{\mathrm{i}} + \beta_{HSL}^{MSE} HSL_i + \beta_{CC}^{MSE} CC_i + e_i^{MSE}$
4. $MSC_i = \beta_0^{MSC} + \beta_{HSL}^{MSC} HSL_i + \beta_{CC}^{MSC} CC_i + \beta_{MSE}^{MSC} MSE_i + e_i^{MSC}$
5. $USE_i = \beta_0^{USE} + \beta_{MSE}^{UES} MSE_i + e_i^{USE}$
6. $PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{MSE}^{PERF} MSE_i + \beta_{MSC}^{PERF} MSC_i + e_i^{PERF}$

Overall Path Model: Matrix Form

- The general path model equation is:

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

- The vectors related to endogenous variables for an observation i:

$$\boldsymbol{y}_i = \underbrace{\begin{bmatrix} HSL_i \\ CC_i \\ MSE_i \\ MSC_i \\ USE_i \\ PERF_i \end{bmatrix}}_{\textbf{Data}}; \boldsymbol{\alpha} = \underbrace{\begin{bmatrix} \beta_0^{HSL} \\ \beta_0^{CC} \\ \beta_0^{MSE} \\ \beta_0^{MSC} \\ \beta_0^{USE} \\ \beta_0^{PERF} \end{bmatrix}}_{\textbf{Intercepts}}; \boldsymbol{\zeta}_i = \underbrace{\begin{bmatrix} e_i^{HSL} \\ e_i^{CC} \\ e_i^{MSE} \\ e_i^{MSC} \\ e_i^{USE} \\ e_i^{PERF} \end{bmatrix}}_{\textbf{Residuals}}$$

- The vector of exogenous variables for an observation i:

$$\boldsymbol{x}_i = \underbrace{[G_i]}_{\textbf{Data}}$$

Overall Path Model: Matrix Form

- The general path model equation is:

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

- The matrix relating endogenous variables to endogenous variables (with labels and $\boldsymbol{y}_i$ for attempted clarity) is:

$$\mathbf{B} = \begin{array}{c} \begin{matrix} \underline{\text{HSL}} & \underline{\text{CC}} & \underline{\text{MSE}} & \underline{\text{MSC}} & \underline{\text{USE}} & \underline{\text{PERF}} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{HSL}^{CC} & 0 & 0 & 0 & 0 & 0 \\ \beta_{HSL}^{MSE} & \beta_{CC}^{MSE} & 0 & 0 & 0 & 0 \\ \beta_{HSL}^{MSC} & \beta_{CC}^{MSC} & \beta_{MSE}^{MSC} & 0 & 0 & 0 \\ 0 & 0 & \beta_{MSE}^{USE} & 0 & 0 & 0 \\ \beta_{HSL}^{PERF} & 0 & \beta_{MSE}^{PERF} & \beta_{MSC}^{PERF} & 0 & 0 \end{bmatrix} \end{array}; \boldsymbol{y}_i = \begin{bmatrix} HSL_i \\ CC_i \\ MSE_i \\ MSC_i \\ USE_i \\ PERF_i \end{bmatrix}$$

Overall Path Model: Matrix Form

- The general path model equation is:

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

- The matrix relating exogenous variables to endogenous variables (with labels and $\boldsymbol{x}_i$ for attempted clarity) is:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \beta_G^{HSL} \\ 0 \\ \beta_G^{MSE} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \underline{\textbf{HSL}} \\ \underline{\textbf{CC}} \\ \underline{\textbf{MSE}} \\ \underline{\textbf{MSC}} \\ \underline{\textbf{USE}} \\ \underline{\textbf{PERF}} \end{matrix}; \quad \boldsymbol{x}_i = [G_i]$$

Overall Path Model: Matrix Form

- The general path model equation is:

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

- The assumption about the residuals is that:

$$\boldsymbol{\zeta}_i = \begin{bmatrix} e_i^{HSL} \\ e_i^{CC} \\ e_i^{MSE} \\ e_i^{MSC} \\ e_i^{USE} \\ e_i^{PERF} \end{bmatrix} \sim N_6(\mathbf{0}, \boldsymbol{\Psi}) \text{ where } \boldsymbol{\Psi} = \begin{bmatrix} \sigma^2_{e:HSL} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{e:CC} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{e:MSE} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{e:MSC} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{e:USE} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{e:PERF} \end{bmatrix}$$

- Finally, the covariance matrix of the exogenous variables is:

$$\boldsymbol{\Phi} = [\sigma_G^2]$$

- **From these matrices, you can construct the model-implied covariance matrix for all variables ($\boldsymbol{\Sigma}_{\mathbf{y,x}}$)**

Path Model Estimation in Mplus

- Having (1) constructed our model and (2) verified it was identified using the t-rule and that it is a recursive model, the next step is to (3) estimate the model with Mplus

```
MODEL:
    hsl ON gender;
    cc ON hsl;
    mse ON hsl gender cc;
    msc ON hsl cc mse;
    use ON mse;
    perf ON mse msc hsl;

!added because Mplus will default to adding covariances of non-related endogenous variables
    perf WITH use@0;

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

- NOTE: Gender is not listed under the model statement
  - It is a categorical variable (dummy coded 0/1)
- If added, Mplus treats it as continuous and plugs it into the MVN log-likelihood
  - This is a big no-no as it cannot be MVN

Model Fit Evaluation

- First, we check convergence:

```
THE MODEL ESTIMATION TERMINATED NORMALLY
```

  - Mplus' algorithm converged
- Second, we check for abnormally large standard errors
  - None too big, relative to the size of the parameter
  - Indicates identified model
- Third, we look at the model fit statistics:

Model Fit Statistics

```
Chi-Square Test of Model Fit

          Value                             58.913*
          Degrees of Freedom                    9
          P-Value                           0.0000
          Scaling Correction Factor         0.9997
            for MLR

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.126
          90 Percent C.I.                    0.096  0.157
          Probability RMSEA <= .05           0.000

CFI/TLI

          CFI                                0.918
          TLI                                0.809

Chi-Square Test of Model Fit for the Baseline Model

          Value                            629.882
          Degrees of Freedom                    21
          P-Value                           0.0000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.056
```

This is a likelihood ratio (deviance) test comparing our model (H_0) with the saturated model – The saturated model fits much better (but that is typical).

The RMSEA estimate is 0.126. Good fit is considered 0.05 or less.

The CFI estimate is .918 and the TLI is .809. Good fit is considered 0.95 or higher.

This compares the independence model (H_0) to the saturated model (H_1) – it indicates that there is significant covariance between variables

The average standardized residual covariance is 0.056. Good fit is less than 0.05.

Based on the model fit statistics, we can conclude that our model does not do a good job of approximating the covariance matrix – so we cannot make inferences with these results (biased standard errors and effects may occur)

Model Modification

- Now that we have concluded that our model fit is poor we must modify the model to make the fit better
  - Our modifications are purely statistical – which draws into question their generalizability beyond this sample
- ***Generally, model modification should be guided by theory***
  - However, we can inspect the normalized residual covariance matrix (like z-scores) to see where our biggest misfit occurs

Normalized Residuals for Covariances/Correlations/Residual Correlations

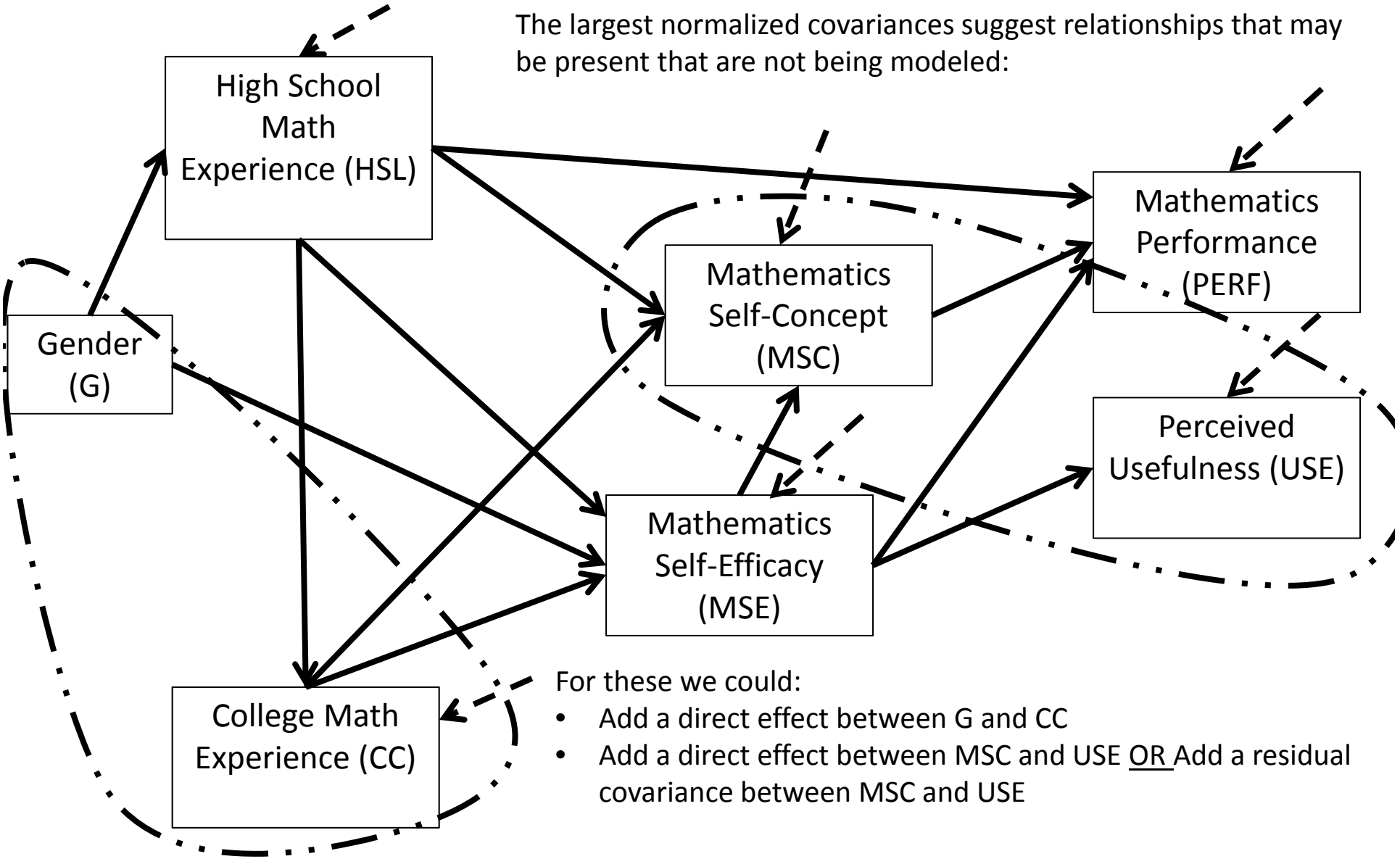
	HSL	CC	MSE	MSC	USE
HSL	0.039				
CC	-0.034	0.046			
MSE	0.085	-0.377	-0.086		
MSC	0.105	-0.161	-0.039	0.043	
USE	0.559	0.720	-0.110	5.051	0.041
PERF	0.006	-0.028	-0.071	0.059	-0.159
GENDER	0.091	-2.567	-0.422	-1.452	-0.027

Normalized Residuals for Covariances/Correlations/Residual Correlations

	PERF	GENDER
PERF	-0.076	
GENDER	-1.522	0.000

Two normalized residual covariances are bigger than +/-1.96:
MSC with USE and
CC with Gender

Our Destination: Overall Path Model



Modification Indices: More Help for Fit

- As we used Maximum Likelihood to estimate our model, another useful feature is that of the modification indices
  - Modification indices are actually Score (LaGrangian Multiplier) tests that attempt to suggest the change in the log-likelihood for adding a given model parameter (larger values indicate a better fit for adding the parameter)

MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index 0.000

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
ON Statements					
HSL	ON CC	6.480	0.447	0.447	1.992
HSL	ON MSE	6.495	1.139	1.139	10.270
HSL	ON MSC	4.093	0.165	0.165	2.146
HSL	ON USE	0.374	0.003	0.003	0.042
HSL	ON PERF	5.293	1.228	1.228	2.763
CC	ON MSE	6.477	-0.410	-0.410	-0.829
CC	ON MSC	6.479	-0.568	-0.568	-1.654
CC	ON USE	0.481	0.016	0.016	0.042
CC	ON PERF	0.059	-0.047	-0.047	-0.024
CC	ON GENDER	6.478	-1.756	-1.756	-0.142
MSE	ON MSC	1.268	0.266	0.266	0.383
MSE	ON USE	0.808	-0.060	-0.060	-0.080
MSE	ON PERF	1.075	1.096	1.096	0.274
MSC	ON USE	41.528	0.299	0.299	0.275
MSC	ON PERF	0.075	-0.414	-0.414	-0.072
MSC	ON GENDER	1.269	-1.669	-1.669	-0.046
USE	ON HSL	0.374	0.482	0.482	0.040
USE	ON CC	0.785	0.141	0.141	0.052
USE	ON MSC	40.043	0.451	0.451	0.490
USE	ON PERF	0.002	0.019	0.019	0.004
PERF	ON CC	0.075	0.006	0.006	0.012
PERF	ON USE	2.573	-0.013	-0.013	-0.067
PERF	ON GENDER	2.219	-0.373	-0.373	-0.060

WITH Statements					
CC	WITH HSL	6.480	15.132	15.132	1.974
MSC	WITH HSL	1.268	14.378	14.378	0.914
USE	WITH HSL	0.362	0.817	0.817	0.040
USE	WITH CC	0.635	4.253	4.253	0.047
USE	WITH MSE	0.808	14.378	14.378	0.094
USE	WITH MSC	41.529	70.912	70.912	0.386
PERF	WITH HSL	2.219	3.213	3.213	1.251
PERF	WITH CC	0.075	0.207	0.207	0.018
PERF	WITH MSE	0.751	3.528	3.528	0.183
PERF	WITH MSC	0.075	-1.567	-1.567	-0.067
PERF	WITH USE	2.573	-2.994	-2.994	-0.100
GENDER	WITH CC	6.479	-0.397	-0.397	-0.144
GENDER	WITH MSC	1.269	-0.378	-0.378	-0.067
GENDER	WITH USE	0.004	0.025	0.025	0.003
GENDER	WITH PERF	2.219	-0.084	-0.084	-0.091

Modification Indices Results

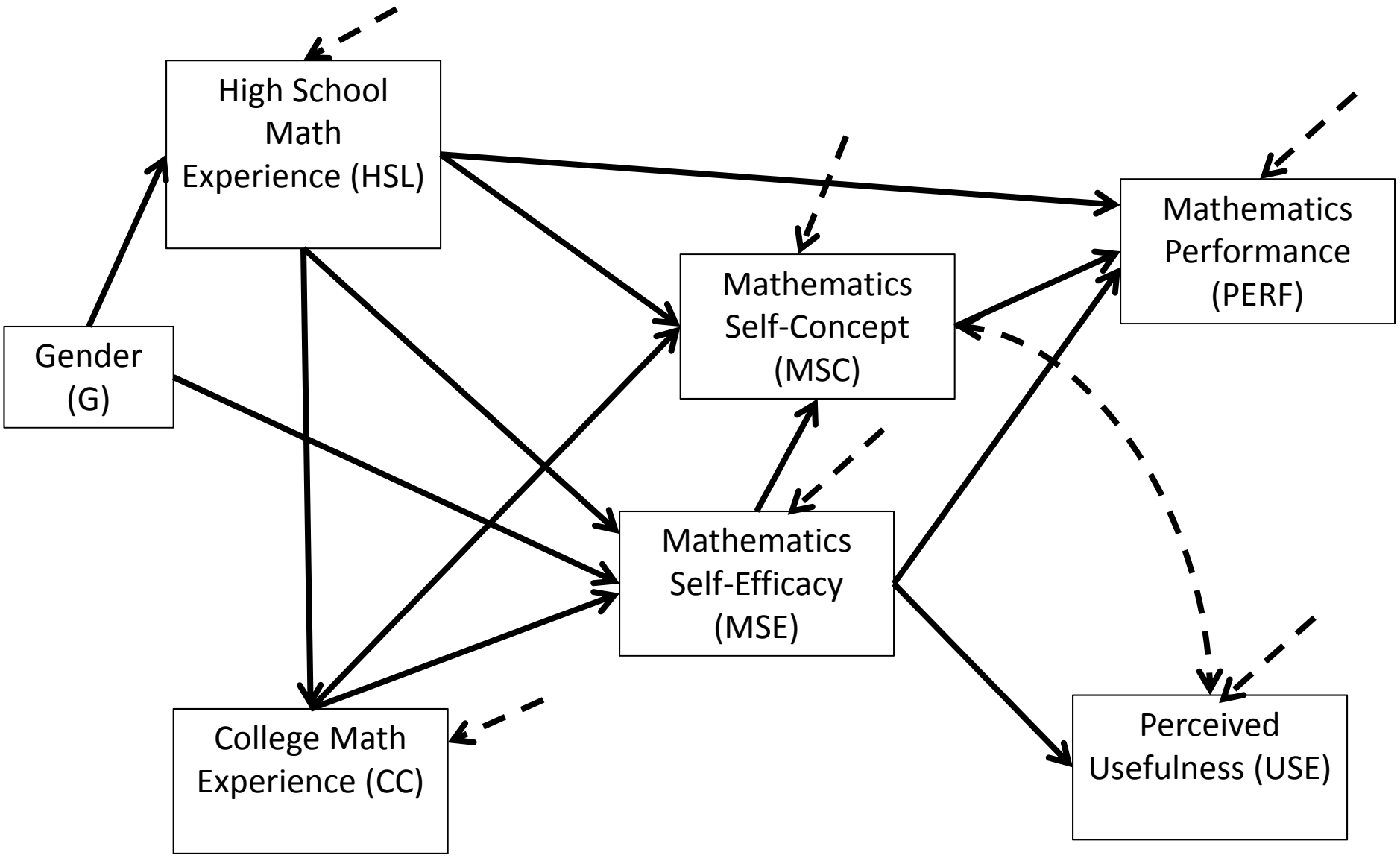
- The modification indices have three large values:
  - A direct effect predicting MSC from USE
  - A direct effect predicting USE from MSC
  - A residual covariance between USE and MSC

- Note: the MI value is -2 times the change in the log-likelihood and the EPC is the expected parameter value
  - The MI is like a 1 DF Chi-Square Deviance test
    - Values greater than 3.84 are likely to be significant changes in the log-likelihood

- Because all three happen for the same variable, we can only choose one
  - This is where theory would help us decide

- As we do not know theory, we will choose to add a residual covariance between USE and MSC
  - Their covariance is **unexplained** by the model – not a great theoretical statement (but will allow us to make inferences if the model fits)
  - MI = 41.529
  - EPC = 70.912

Modified Model



Assessing Model fit of the Modified Model

- Now we must start over with our path model decision tree
  - The model is identified (now 20 parameters < 28 covariances)
  - Mplus estimation converged; Standard errors look acceptable
- Model fit statistics:

```
Chi-Square Test of Model Fit

          Value                             14.393*
          Degrees of Freedom                    8
          P-Value                           0.0721
          Scaling Correction Factor         1.0302
            for MLR

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.048
          90 Percent C.I.                    0.000  0.087
          Probability RMSEA <= .05           0.483

CFI/TLI

          CFI                                0.990
          TLI                                0.972

SRMR (Standardized Root Mean Square Residual)

          Value                              0.035
```

The comparison with the saturated model suggests our model fits statistically

The RMSEA is 0.048, which indicates good fit

The CFI and TLI both indicate good fit

The SRMR also indicates good fit

Therefore, we can conclude the model adequately approximates the covariance matrix – meaning we can now inspect our model parameters…but first, let's check our residual covariances and modification indices

Normalized Residual Covariances

- Only one normalized residual covariance is bigger than +/- 1.96: CC with Gender
  - Given the number of covariances we have, this is likely okay

Normalized Residuals for Covariances/Correlations/Residual Correlations

	HSL	CC	MSE	MSC	USE
HSL	0.017				
CC	0.037	0.034			
MSE	0.020	-0.356	-0.103		
MSC	0.154	0.050	-0.104	0.054	
USE	0.638	0.771	0.064	0.337	0.020
PERF	0.062	0.018	-0.113	-0.003	-0.990
GENDER	0.051	-2.568	-0.359	-1.456	0.026

Normalized Residuals for Covariances/Correlations/Residual Correlations

	PERF	GENDER
PERF	-0.062	
GENDER	-1.499	0.000

Modification Indices

- Now, no modification indices are glaringly large, although some are bigger than 3.84
  - We discard these as our model now fits (and adding them may not be meaningful)

ON Statements

HSL	ON CC	6.503	0.441	0.441	1.965
HSL	ON MSE	6.427	1.117	1.117	10.068
HSL	ON MSC	1.355	0.022	0.022	0.289
HSL	ON USE	0.479	0.004	0.004	0.048
HSL	ON PERF	4.274	0.773	0.773	1.737
CC	ON MSE	6.518	-0.429	-0.429	-0.869
CC	ON MSC	0.012	-0.008	-0.008	-0.023
CC	ON USE	0.423	0.015	0.015	0.040
CC	ON PERF	0.022	-0.029	-0.029	-0.015
CC	ON GENDER	6.501	-1.788	-1.788	-0.144
MSE	ON MSC	0.023	0.026	0.026	0.038
MSE	ON USE	0.904	-0.064	-0.064	-0.085
MSE	ON PERF	0.601	0.831	0.831	0.207
MSC	ON PERF	1.907	1.150	1.150	0.199
MSC	ON GENDER	1.817	-1.887	-1.887	-0.052
USE	ON HSL	0.480	0.554	0.554	0.046
USE	ON CC	0.710	0.135	0.135	0.050
USE	ON MSC	1.148	0.222	0.222	0.241
USE	ON PERF	2.491	-0.732	-0.732	-0.138
USE	ON GENDER	0.295	0.947	0.947	0.028
PERF	ON CC	0.083	0.006	0.006	0.013
PERF	ON USE	3.114	-0.015	-0.015	-0.081
PERF	ON GENDER	1.923	-0.350	-0.350	-0.056

WITH Statements

CC	WITH HSL	6.505	14.968	14.968	1.949
MSC	WITH HSL	1.824	15.821	15.821	1.004
MSC	WITH MSE	1.815	44.019	44.019	0.373
USE	WITH HSL	0.404	0.877	0.877	0.043
USE	WITH CC	0.556	4.039	4.039	0.045
USE	WITH MSE	1.288	-18.046	-18.046	-0.118
PERF	WITH HSL	1.924	2.933	2.933	1.147
PERF	WITH CC	0.083	0.219	0.219	0.019
PERF	WITH MSE	0.580	3.159	3.159	0.165
PERF	WITH MSC	1.903	4.324	4.324	0.186
PERF	WITH USE	3.056	-3.087	-3.087	-0.103
GENDER	WITH CC	6.501	-0.404	-0.404	-0.146
GENDER	WITH MSC	1.817	-0.427	-0.427	-0.075
GENDER	WITH USE	0.295	0.214	0.214	0.029
GENDER	WITH PERF	1.923	-0.079	-0.079	-0.086

More on Modification Indices

- Recall from our original model that we received the following modification index values for the residual covariance between MSC and USE
  - MI = 41.529
  - EPC = 70.912

- The estimated residual covariance between MSC and USE in the modified model is: 70.247

Model Parameter Investigation

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HSL ON				
GENDER	0.208	0.154	1.348	0.178
CC ON				
HSL	0.662	0.247	2.686	0.007
MSE ON				
HSL	4.138	0.406	10.203	0.000
GENDER	4.168	1.160	3.593	0.000
CC	0.393	0.105	3.723	0.000
MSC ON				
HSL	2.823	0.593	4.764	0.000
CC	0.519	0.117	4.433	0.000
MSE	0.736	0.066	11.120	0.000
USE ON				
MSE	0.277	0.073	3.803	0.000
PERF ON				
MSE	0.139	0.013	10.700	0.000
MSC	0.037	0.009	4.147	0.000
HSL	0.153	0.107	1.432	0.152
PERF WITH				
USE	0.000	0.000	999.000	999.000
MSC WITH				
USE	70.247	10.358	6.782	0.000

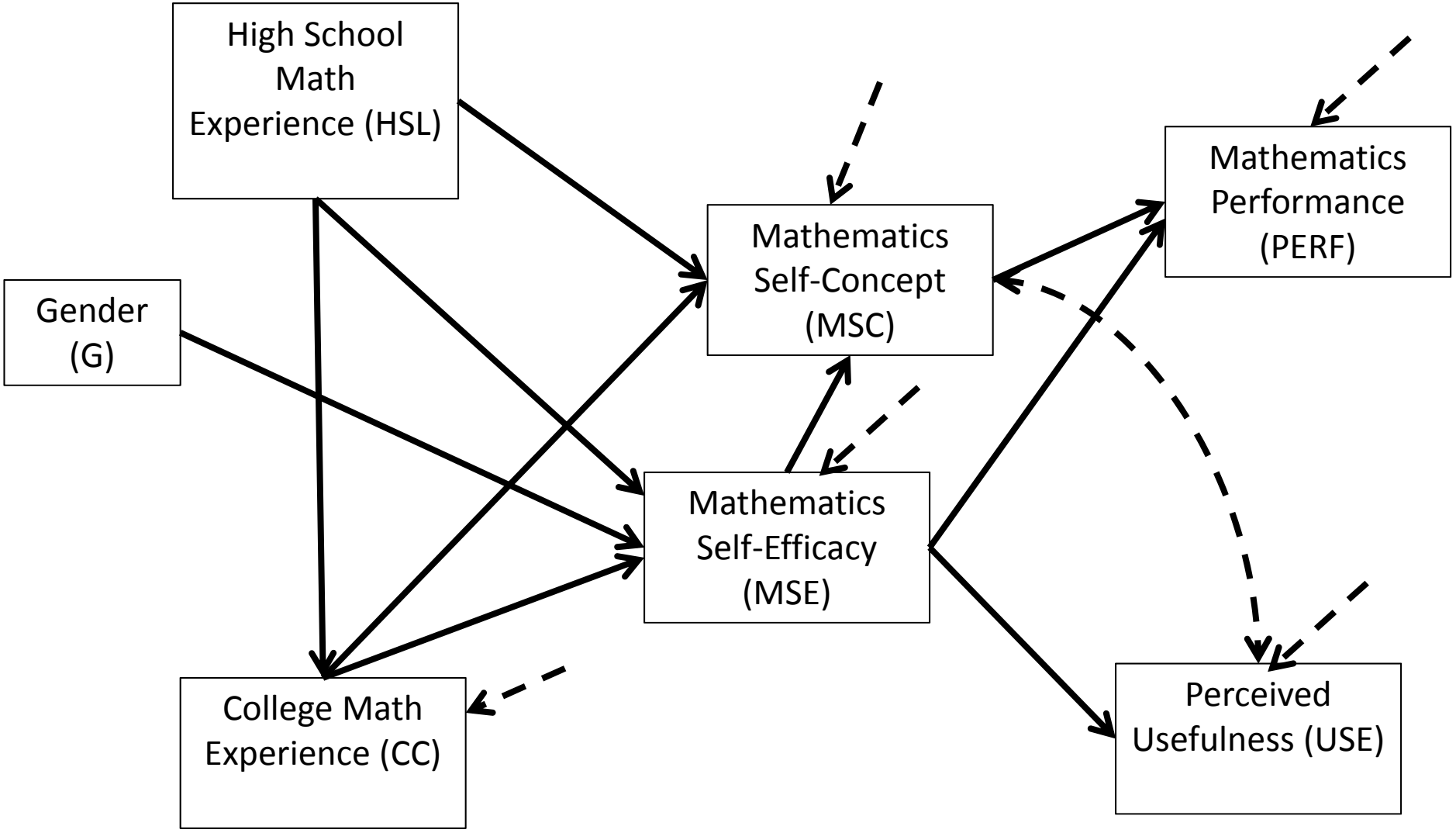
There are two direct effects that are non-significant:

$$\beta_G^{HSL} = 0.208$$
$$\beta_{HSL}^{PERF} = 0.153$$

We can leave these in the model, but the overall path model seems to suggest they are not needed

So, I will remove them and re-estimate the model

Modified Model #2



Model #2: Model Fit Results

- We have: an identified model, a converged algorithm, and stable standard errors, so model fit should be inspected
  - Next – inspect model fit
  - Model fit seems to not be as good as we would think

Chi-Square Test of Model Fit

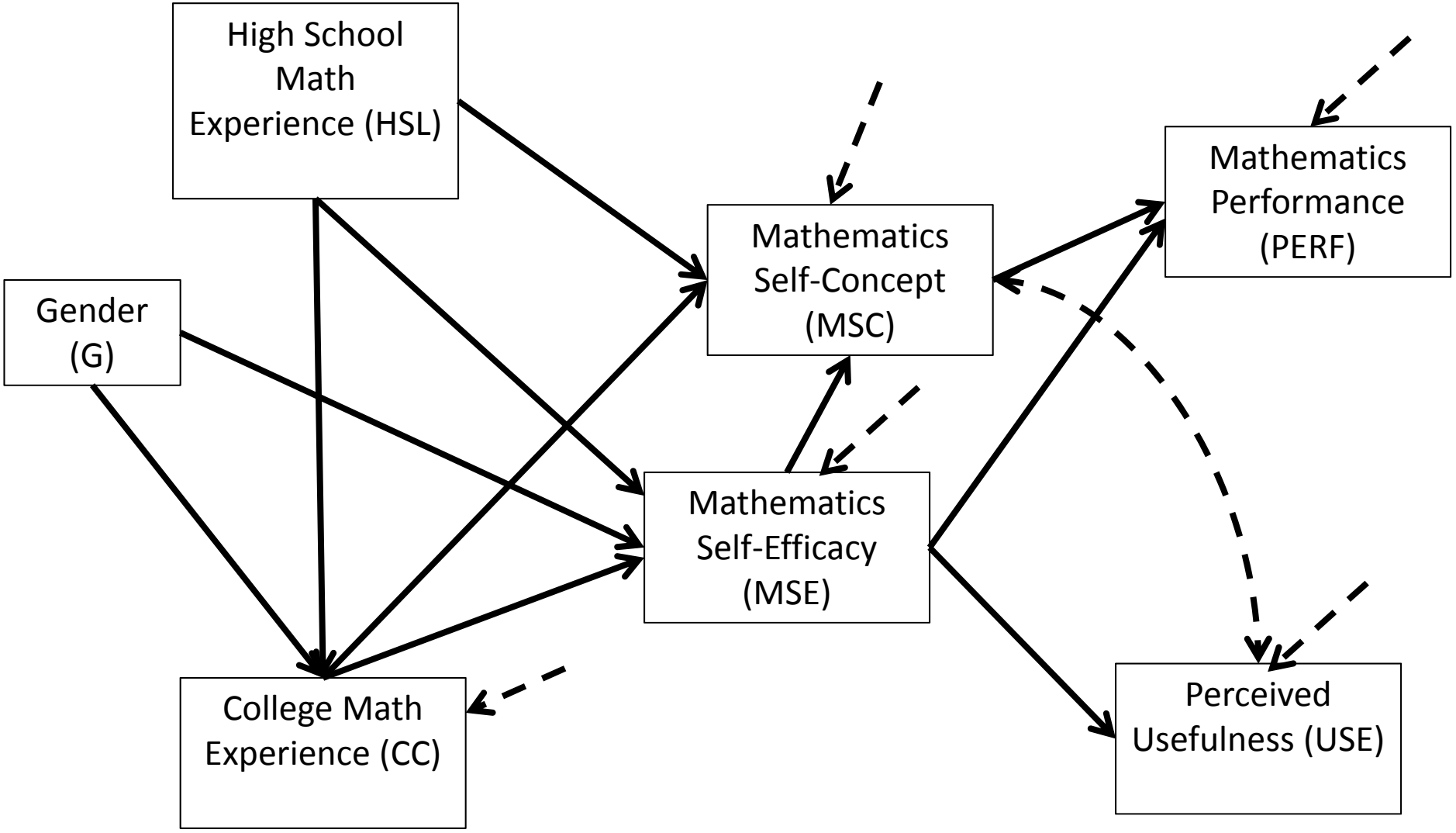
Value	18.293*
Degrees of Freedom	10
P-Value	0.0502
Scaling Correction Factor for MLR	1.0156

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.049	
90 Percent C.I.	0.000	0.083
Probability RMSEA <= .05	0.477	

- Again, the largest normalized residual covariance is that of GENDER and CC
  - MI for direct effect of GENDER on CC is 6.494, indicating that adding this parameter may improve model fit
- So, we will now add a direct effect of Gender on CC

Modified Model #3



Model #3: Model Fit Results

- We have: an identified model, a converged algorithm, and stable standard errors, so model fit should be inspected
  - Next – inspect model fit
  - Model fit seems to be very good

Chi-Square Test of Model Fit

Value	11.616*
Degrees of Freedom	9
P-Value	0.2358
Scaling Correction Factor for MLR	1.0235

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.029	
90 Percent C.I.	0.000	0.070
Probability RMSEA <= .05	0.757	

- No normalized residual covariances are larger than +/- 1.96 – so we appear to have good fit

- No Modification Indices are larger than 3.84
  - We will leave this model as-is and interpret the results

Model #3 Parameter Interpretation

Interpret each of these parameters as you would in regression:

A one-unit increase in HSL brings about a .707 unit increase in CC, holding gender constant

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
CC	ON				
	HSL	0.707	0.246	2.879	0.004
	GENDER	-1.779	0.671	-2.653	0.008
MSE	ON				
	HSL	4.158	0.403	10.305	0.000
	GENDER	4.283	1.154	3.711	0.000
	CC	0.398	0.105	3.796	0.000
MSC	ON				
	HSL	2.831	0.593	4.775	0.000
	CC	0.528	0.116	4.545	0.000
	MSE	0.733	0.066	11.076	0.000
USE	ON				
	MSE	0.276	0.073	3.785	0.000
PERF	ON				
	MSE	0.145	0.013	11.425	0.000
	MSC	0.041	0.009	4.671	0.000
PERF	WITH				
	USE	0.000	0.000	999.000	999.000
MSC	WITH				
	USE	70.596	10.376	6.804	0.000

Model #3 Standardized Parameter Estimates

- We can interpret the STDYX standardized parameter estimates for all variables except gender
  - It is not continuous so SD of gender does not make sense

- A 1-SD increase in HSL means CC increases by 0.158 SD

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
CC ON				
HSL	0.158	0.054	2.911	0.004
GENDER	-0.143	0.053	-2.717	0.007
MSE ON				
HSL	0.466	0.042	11.032	0.000
GENDER	0.172	0.045	3.792	0.000
CC	0.199	0.052	3.795	0.000
MSC ON				
HSL	0.220	0.045	4.851	0.000
CC	0.183	0.040	4.587	0.000
MSE	0.508	0.042	12.136	0.000
USE ON				
MSE	0.206	0.054	3.815	0.000
PERF ON				
MSE	0.578	0.046	12.680	0.000
MSC	0.234	0.049	4.758	0.000

Model #3 STDY Interpretation

- The STDY standardization does not standardize by the SD of the X variable
  - So it's interpretation makes sense for Gender (1 = male)

Here, males have an average CC (intercept) that is -.301 SD lower than females

STDY Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
CC ON				
HSL	0.158	0.054	2.911	0.004
GENDER	-0.301	0.111	-2.719	0.007
MSE ON				
HSL	0.466	0.042	11.032	0.000
GENDER	0.363	0.095	3.800	0.000
CC	0.199	0.052	3.795	0.000

Overall Model Interpretation

- High School Experience and Gender are significant predictors of College Experience
  - Men lower than women in College Experience
  - More High School Experience means more College Experience

- High School Experience, College Experience, and Gender are significant predictors of Math Self-Efficacy
  - More High School and College Experience means higher Math Self-Efficacy
  - Men have higher Math Self-Efficacy than Women

Overall Model Interpretation, Continued

- High School Experience, College Experience, and Math Self-Efficacy are significant predictors of Math Self-Concept
  - More High School and College Experience and higher Math Self-Efficacy mean higher Math Self-Concept

- Higher Math Self-Efficacy means significantly higher Perceived Usefulness

- Higher Math Self-Efficacy and Math Self-Concept result in higher Math Performance scores

- Math Self-Concept and Perceived Usefulness have a significant residual covariance

Model Interpretation: Explained Variability

- The R^2 for each endogenous variable:
  - CC – 0.046
  - MSE – 0.306
  - MSC – 0.509
  - USE – 0.042
  - PERF – 0.568

- Note how college experience and perceived usefulness both have low percentages of variance accounted for by the model
  - We could have increased the R^2 for USE by adding the direct path between MSC and USE instead of the residual covariance

Looking at the Indirect Effect of Gender on Performance

- Gender had a significant indirect effect on performance:

TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from GENDER to PERF				
Total	0.586	0.212	2.771	0.006
Total indirect	0.586	0.212	2.771	0.006

- This means that overall, men have a PERF score that is .586 higher than women (or .198 SD of PERF from below)

STDY Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Effects from GENDER to PERF				
Total	0.198	0.071	2.793	0.005
Total indirect	0.198	0.071	2.793	0.005

ADDITIONAL MODELING CONSIDERATIONS IN PATH ANALYSIS

Additional Modeling Considerations

- The path analysis we just ran was meant to be an introduction to the topic and the field
  - It is much more complex than what was described

- In particular, our path analysis assumed all variables to be
  - Continuous and Multivariate Normal
  - Measured with perfect reliability

- In reality, neither of these are true

- Structural equation models (path models with latent variables) will help with variables with measurement error
  - We begin next week
- Modifications to model likelihoods or different distributional assumptions will help with the normality assumption
  - Last week of class

About Causality

- You will read a lot of talk about path models indicating causality, or how path models are causal models

- It is important to note that causality can rarely, if ever, be inferred on the basis of observational data
  - Experimental designs with random assignment and manipulations of factors will help detect causality

- With observational data, about the best you can say is that IF your model fits, then causality is ONE reason
  - But realistically, you are simply describing covariances of variables in more fancy ways/parameters

- If your model does not fit, the causality is LIKELY not occurring
  - But still could be possible if important variables are omitted

CONCLUDING REMARKS

Path Analysis: An Introduction

- In this lecture (spanning multiple weeks), we discussed the basics of path analysis
  - Model specification/identification
  - Model estimation
  - Model fit (necessary, but not sufficient)
  - Model modification and re-estimation
  - Final model parameter interpretation

- There is a lot to the analysis – but what is important to remember is the over-arching principal of multivariate analyses: covariance between variables is important
  - Path models imply very specific covariance structures
  - The validity of the results hinge upon accurately finding an approximation to the covariance matrix

Where We Are Heading…

- Over the next few weeks, we will be doing path models, but with unobserved latent variables
  - These are more commonly called factor models or structural equation models

- As with path models, structural equation models are multivariate analysis techniques
  - Models make specific implications for the covariance matrix

- Factor models shift the focus from prediction of observed variables to measurement of unobserved variables

- In the end, we will combine both – factor models for measuring unobserved variables and path models for predicting observed and unobserved variables
  - But all will fall under a common multivariate framework