An Introduction to Mplus and Path Analysis

PSYC 943 (930): Fundamentals of Multivariate Modeling

Lecture 19: November 7, 2012

Today's Lecture

- A brief intro to Mplus

- Path analysis
  - …starting with multivariate regression…
  - …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)

- 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.

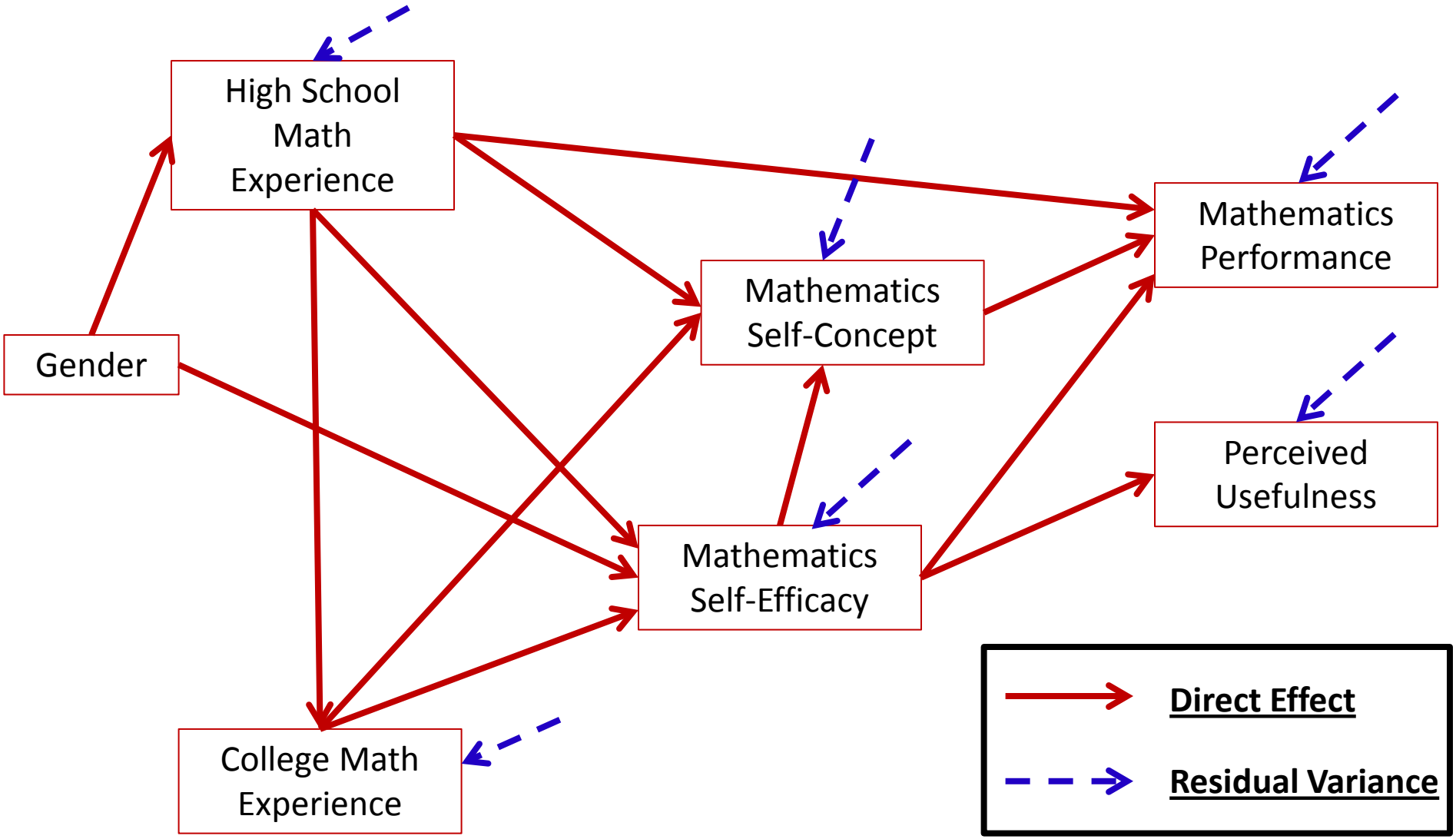
- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- 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, when using an identity link, 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
  - This is where things deviate from our familiar R matrix
- Like multivariate models, the key to path analysis is finding an 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

INTRODUCTION TO MPLUS

The Mplus Statistical Package

- Mplus provides a general latent variable modeling framework that allows for combinations of:
  - Continuous or categorical latent variables
  - Continuous, categorical, count, nominal or censored data

- Mplus is commercial software that is available on Windows machines in the Burnett computer labs

- Mplus is also available for purchase:
  - Available at http://www.statmodel.com
  - From $195 to $350 (student)

Mplus Data File Input Format

- Mplus input files must be ASCII text based (so not binary)
  - Text-based file formats: *.txt, *.dat, *.csv
  - Not-text-based file formats: *.xlsx, *.sas7bdat, *.sav

- The easiest way to get data files into Mplus is to use "free-formatting" (some type of delimiter between columns)
  - I prefer comma-delimited files and will only use those in this course
  - **Cannot start with variable names in first row of data**

- Typically, I store data in Excel and save as a comma-delimited file
  - Save As…*.csv…
    - (then click OK to the first question)…(then click YES to the second)
    - Ignore the warning (click NO) to re-save when closing the Excel Workbook

Mplus Syntax Conventions

- Most syntax must have a semi-colon end each line (;)
  - Exceptions: TITLE section, comments, and continuing lines
- Comments are denoted with an exclamation point (!)
- Syntax is organized by sections; headings of sections end with colons (:)
  - TITLE:, DATA:, VARIABLE:, DEFINE:, and MODEL: are what we use this week
- Syntax cannot exceed 90 characters per row (‽)
- Mplus input files are typically saved with the extension *.inp
- Mplus output files are typically saved with the extension *.out
  - Both are ASCII text (i.e., you can open with text editors)
- The default location for the data file and output file are the folder containing the input file

Mplus TITLE Section

- The TITLE section contains the label of the analysis
- You can type whatever you want here…it will appear verbatim at the top of the output file
- You do not have to terminate this section with a semi-colon
- This section is optional

Mplus DATA Section

- The DATA section is where data files are defined

- Define the name (and path if different from input file folder) by using the command:

FILE = mydata.csv

- POTENTIAL MISTAKES:
  - Data must be numeric – if not errors happen
  - First row of data should not contain variable names

- This section is NOT OPTIONAL

Mplus VARIABLE Section

- The Mplus VARIABLE section defines the names of the variables in the data file, variable types, and variables in your analysis

  - NAMES = provides variable names
    - Names cannot be more than 8 characters
    - Lists of variables can be created (i.e., X1-X10 makes 10 variables)
    - By default all variables listed in the NAMES section are assumed to be part of the analysis
  - USEVARIABLE = provides names of variables used in the analysis (optional)
  - IDVARIABLE = provides the ID variable name (optional)

- This section is NOT OPTIONAL

Mplus DEFINE Section

- The DEFINE section is where new variables are created

- To test our equal slopes hypothesis we created interaction variables by the following syntax:

```
TITLE: !title section puts text below at the top of the output file
    ANCOVA MODEL WITH GENDER, MOTIVATION, AND ACHIEVEMENT
    TESTING INTERACTIONS - DIFFERENT SLOPES WITHIN GENDER GROUP
DATA: !data section defines data file
    FILE = exampledata.csv;

VARIABLE: !variable section defines varibles in data file
    NAMES = ID achieve selfest motivate gender se1-se5
            motiv1-motiv5 ach1-ach20;
    IDVARIABLE = ID;
    USEVARIABLE = achieve selfest gender motivate femaleSE femaleM;

DEFINE:
    femaleSE = gender*selfest;
    femaleM = gender*motivate;

MODEL:
    achieve ON selfest gender motivate femaleSE femaleM;
```

Mplus MODEL Section

- The MODEL section is where you define the model

  These models use the ON statement (ON = REGRESSION)

  achieve ON selfest motivate

- And an empty model to figure out the covariance matrix of the MOTIVATION items
  - This used the WITH statement (WITH = COVARIANCE)

  motiv1-motive5 WITH motive1-motive5

MULTIVARIATE REGRESSION

Multivariate Regression

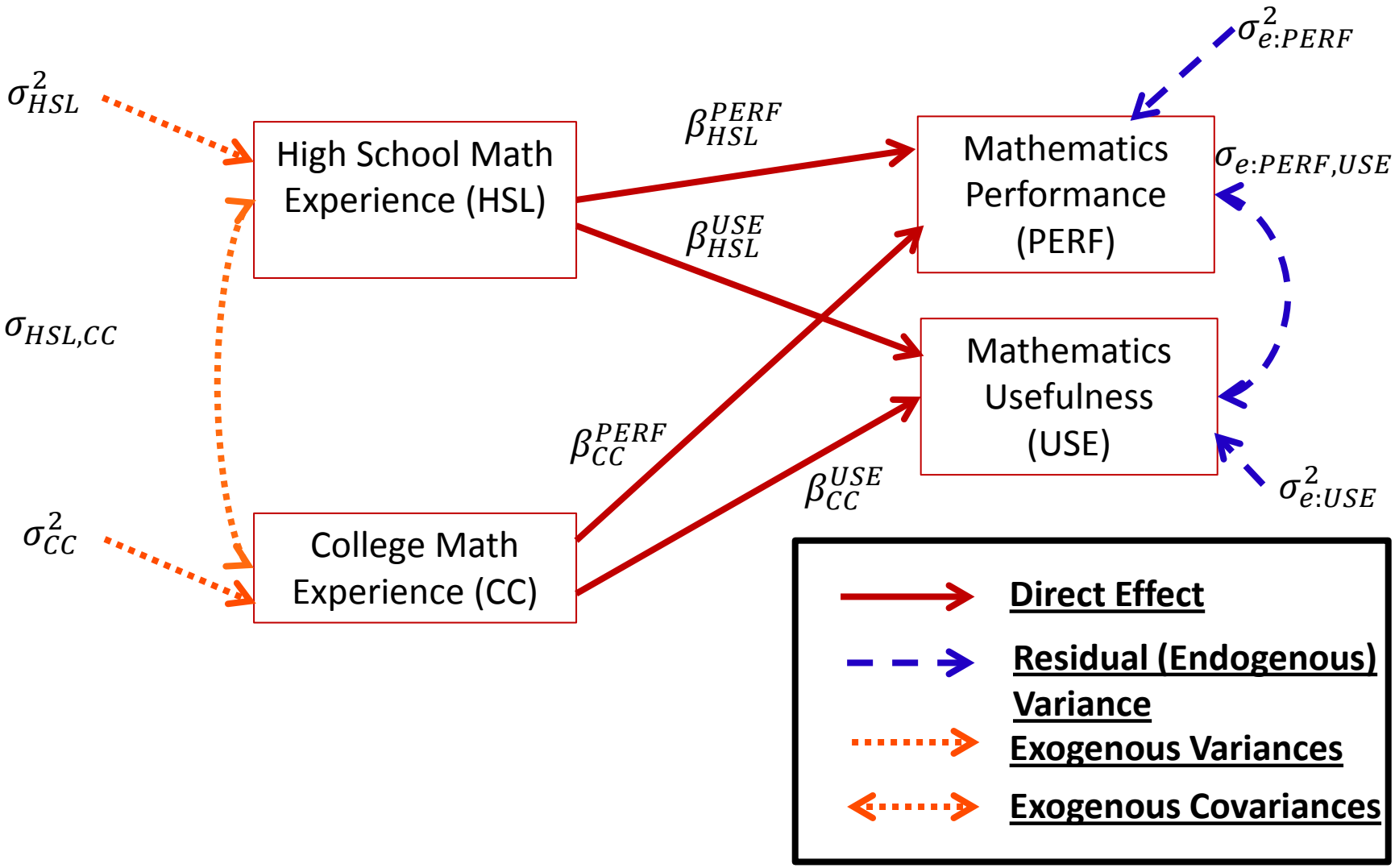
- Before we dive into path analysis, we will begin with a multivariate regression model:
  - Predicting mathematics performance (PERF) with high school (HSL) and college (CC) experience
  - Predicting perceived usefulness (USE) with high school (HSL) and College (CC) experience

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

- We denote the residual for PERF as e_i^{PERF} and the residual for USE as e_i^{USE}
  - We also 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



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) and Mathematics Usefulness (USE)

- 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

Multivariate Regression in Mplus

- The basic code for linear regression in Mplus uses the ON statement
- The WITH statement estimates a covariance between the two variables

```
TITLE:
    MULTIVARIATE Multiple Regression Analysis
    Predicting Performance and Perceived Usefulness
    NOTE: NO LISTWISE DELETION OF INCOMPLETE CASES
    THIS IS DUE TO ADDING THE WITH STATEMENT (INSERTS INTO LIKELIHOOD FUNCTION)

DATA:
    FILE = mathdata.csv;

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

MODEL:
    perf ON hsl cc;
    use ON hsl cc;
    hsl; cc; hsl WITH cc; !adds exogenous variables to likelihood

OUTPUT:
    STANDARDIZED RESIDUAL SAMPSTAT;
```

Labeling Variables

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

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

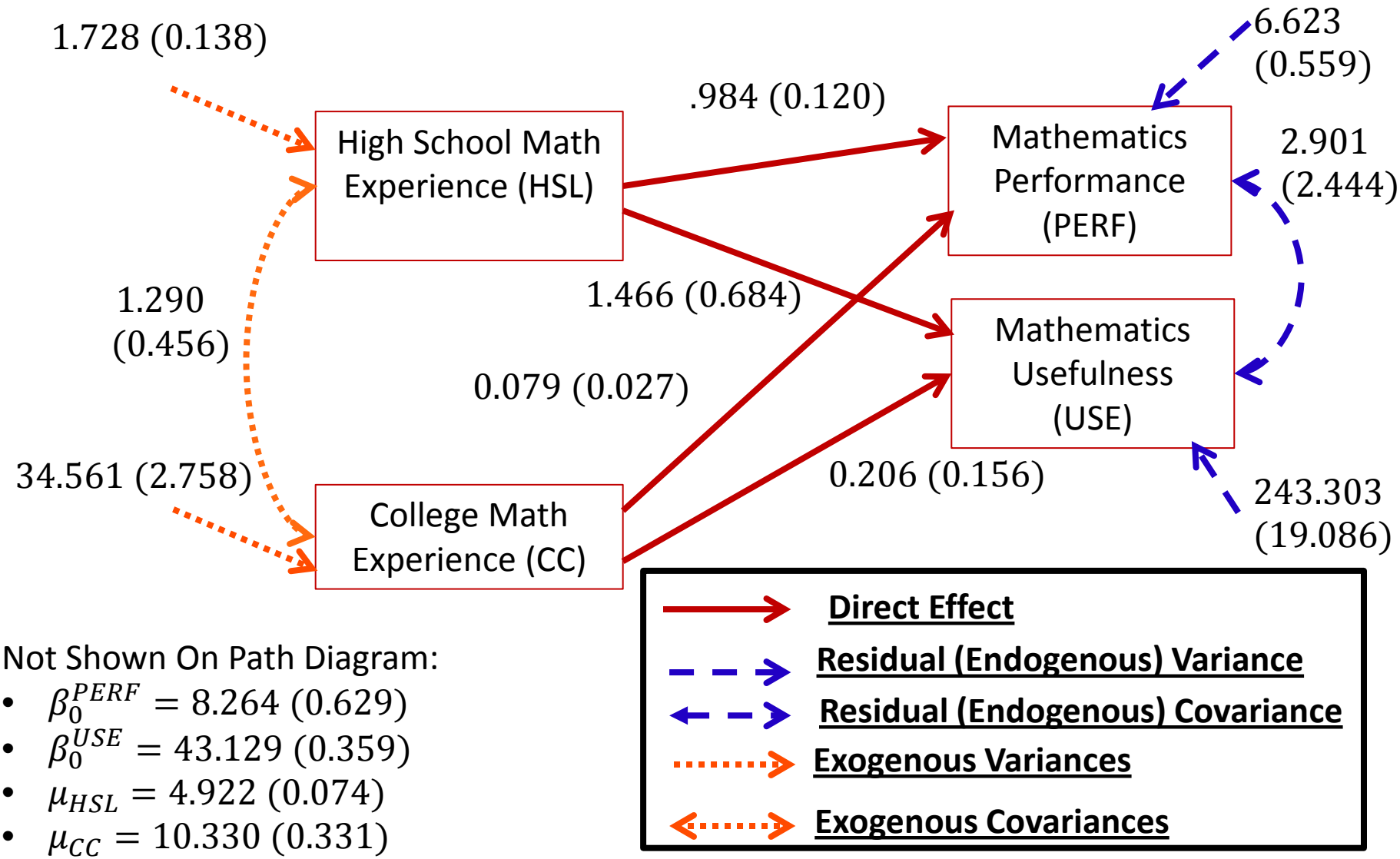
Multivariate Regression Model Parameters

- If we considered all four variables to be part of a multivariate normal distribution, our unstructured (saturated) model would have a total of 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 (Unstandardized Coefficients)



Not Shown On Path Diagram:

- $\beta_0^{PERF} = 8.264\ (0.629)$
- $\beta_0^{USE} = 43.129\ (0.359)$
- $\mu_{HSL} = 4.922\ (0.074)$
- $\mu_{CC} = 10.330\ (0.331)$

Interpreting Multivariate Regression Results for PERF

- $\beta_0^{PERF} = 8.264$: 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	0.120	8.191	0.000
CC	0.079	0.027	2.930	0.003
HSL WITH				
CC	1.275	0.456	2.796	0.005
Means				
HSL	4.925	0.073	67.022	0.000
CC	10.331	0.331	31.170	0.000
Intercepts				
PERF	8.253	0.631	13.084	0.000
Variances				
HSL	1.726	0.137	12.573	0.000
CC	34.556	2.757	12.534	0.000
Residual Variances				
PERF	6.631	0.560	11.841	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	0.120	8.183	0.000
CC	0.079	0.027	2.934	0.003
USE ON				
HSL	1.466	0.684	2.143	0.032
CC	0.206	0.156	1.317	0.188
HSL WITH				
CC	1.290	0.456	2.827	0.005
USE WITH				
PERF	2.901	2.444	1.187	0.235
Means				
HSL	4.922	0.074	66.952	0.000
CC	10.330	0.331	31.174	0.000
Intercepts				
PERF	8.264	0.629	13.129	0.000
USE	43.129	3.590	12.014	0.000
Variances				
HSL	1.728	0.138	12.562	0.000
CC	34.561	2.758	12.533	0.000
Residual Variances				
PERF	6.623	0.559	11.846	0.000
USE	243.303	19.086	12.748	0.000

Model Modification

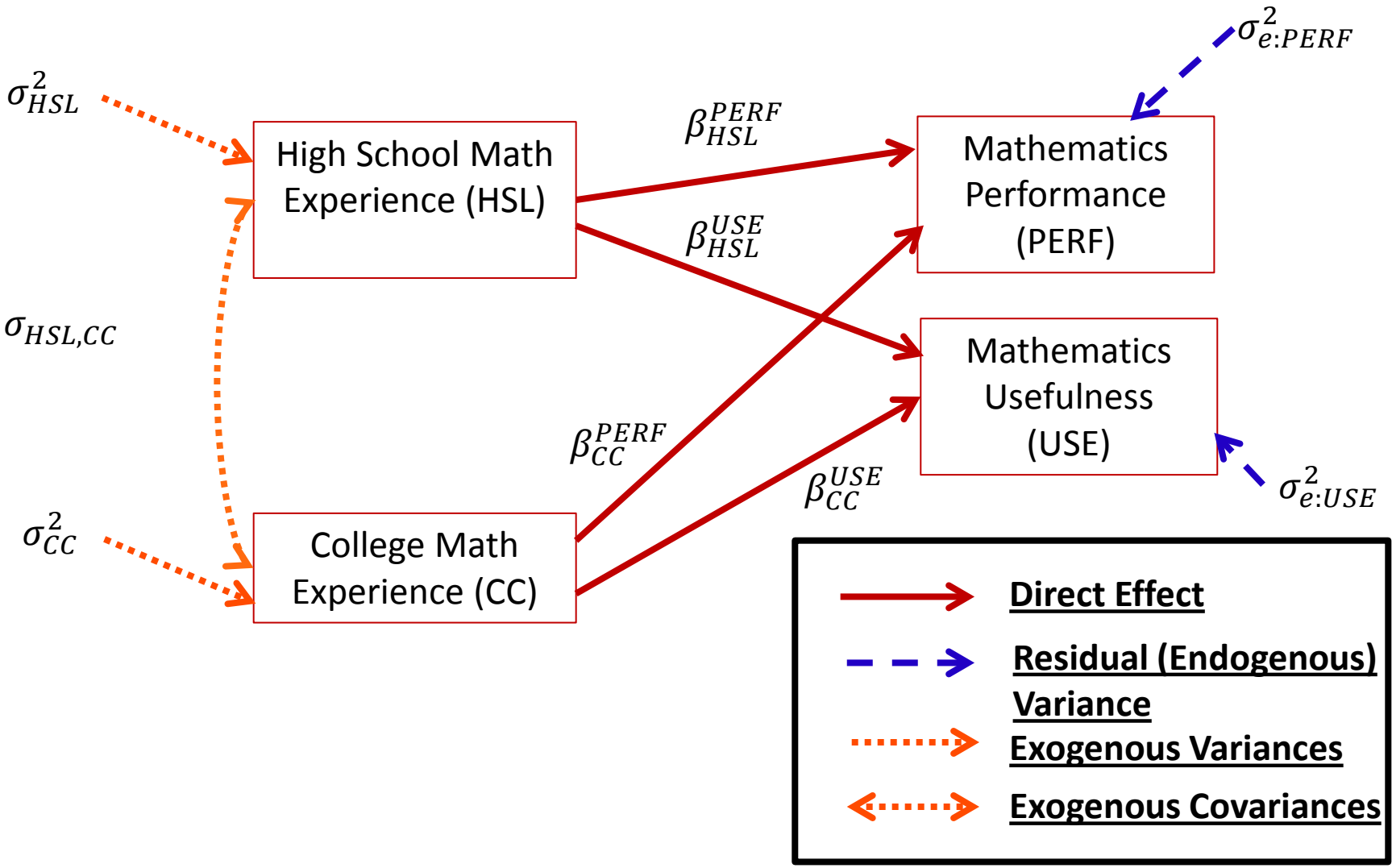
- The residual covariance parameter (between PERF and USE) was not significant
- This means that after accounting for the relationship between HSL and CC with PERF along with HSL and CC with USE, the correlation between these two is zero
  - Meaning we can likely remove the parameter from the model

```
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…

- 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
          H1 Value                       -3572.730
```

- The saturated model (H1):
  - All variances, covariances, and means estimated from Multivariate Normal
  - Cannot do any better than the saturated model when using MVN
  - If not all variables are normal, saturated model is harder to understand

- 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

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	1.419
Degrees of Freedom	1
P-Value	0.2336

  - 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

- 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                           0.035
          90 Percent C.I.                    0.000  0.152
          Probability RMSEA <= .05           0.416
```

- RMSEA is based on the approximated covariance matrix

- 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 .035)
  - Expected for 13 parameters (out of 14 possible)

CFI/TLI

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

CFI/TLI

CFI	0.995
TLI	0.973

- 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	83.781
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-Tailed P-Value
PERF ON				
HSL	0.984	0.120	8.183	0.000
CC	0.079	0.027	2.934	0.003
USE ON				
HSL	1.466	0.684	2.143	0.032
CC	0.206	0.156	1.317	0.188
HSL WITH				
CC	1.290	0.456	2.827	0.005
USE WITH				
PERF	2.901	2.444	1.187	0.235
Means				
HSL	4.922	0.074	66.952	0.000
CC	10.330	0.331	31.174	0.000
Intercepts				
PERF	8.264	0.629	13.129	0.000
USE	43.129	3.590	12.014	0.000
Variances				
HSL	1.728	0.138	12.562	0.000
CC	34.561	2.758	12.533	0.000
Residual Variances				
PERF	6.623	0.559	11.846	0.000
USE	243.303	19.086	12.748	0.000

Reduced Model

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERF ON				
HSL	0.988	0.120	8.236	0.000
CC	0.079	0.027	2.934	0.003
USE ON				
HSL	1.473	0.684	2.155	0.031
CC	0.210	0.156	1.349	0.177
HSL WITH				
CC	1.293	0.456	2.833	0.005
PERF WITH				
USE	0.000	0.000	999.000	999.000
Means				
HSL	4.921	0.074	66.946	0.000
CC	10.331	0.331	31.177	0.000
Intercepts				
PERF	8.242	0.629	13.111	0.000
USE	43.074	3.587	12.008	0.000
Variances				
HSL	1.729	0.138	12.558	0.000
CC	34.563	2.758	12.532	0.000
Residual Variances				
PERF	6.617	0.559	11.847	0.000
USE	243.191	19.076	12.749	0.000

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

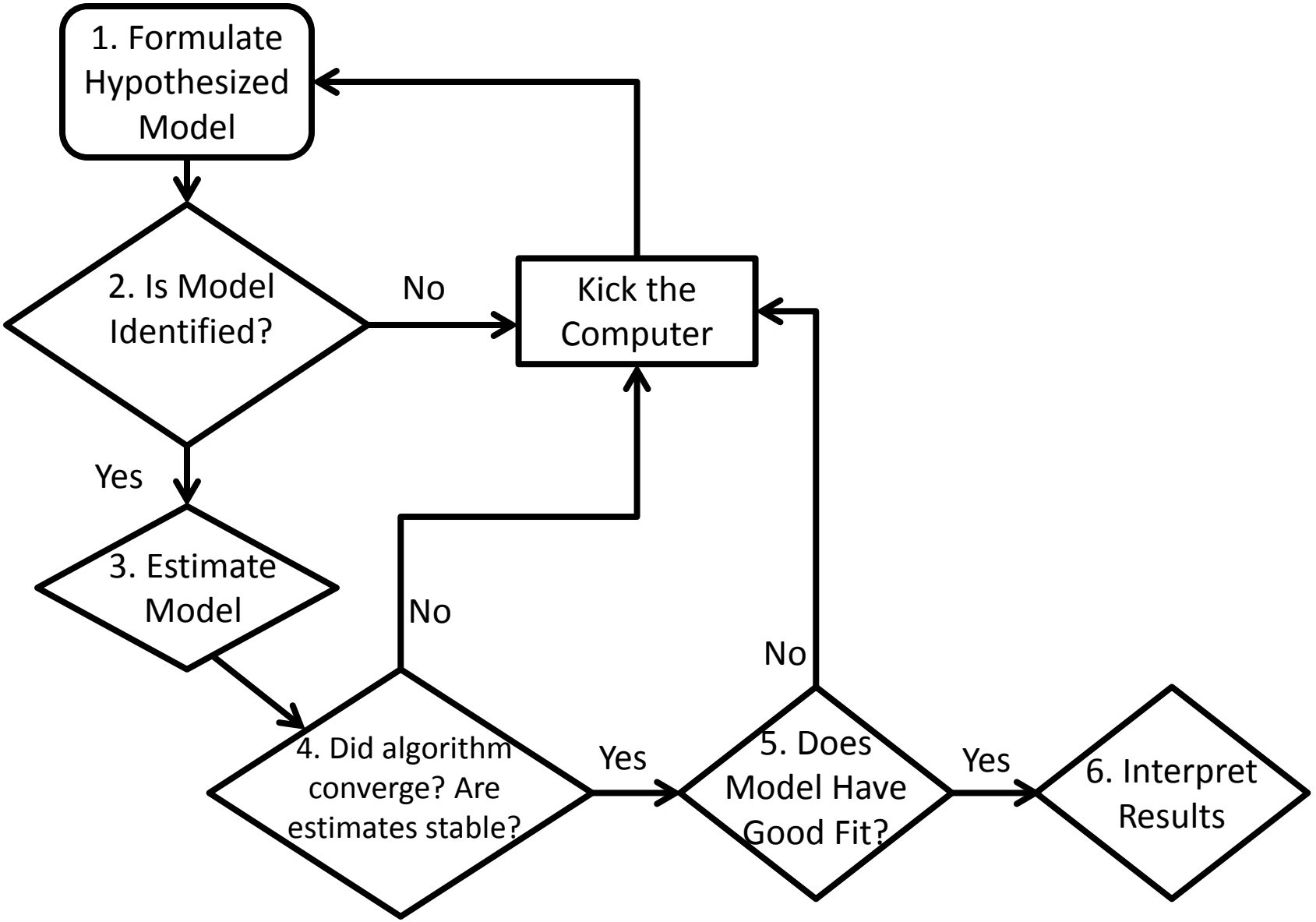
	PERF	USE	HSL	CC
PERF	-0.013			
USE	2.853	-0.020		
HSL	-0.010	-0.021	-0.001	
CC	-0.009	-0.183	-0.003	-0.002

- 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

THE FINAL PATH MODEL: PUTTING IT ALL TOGETHER

A Path Model of Path Analysis Steps



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

T-rule Identification Status

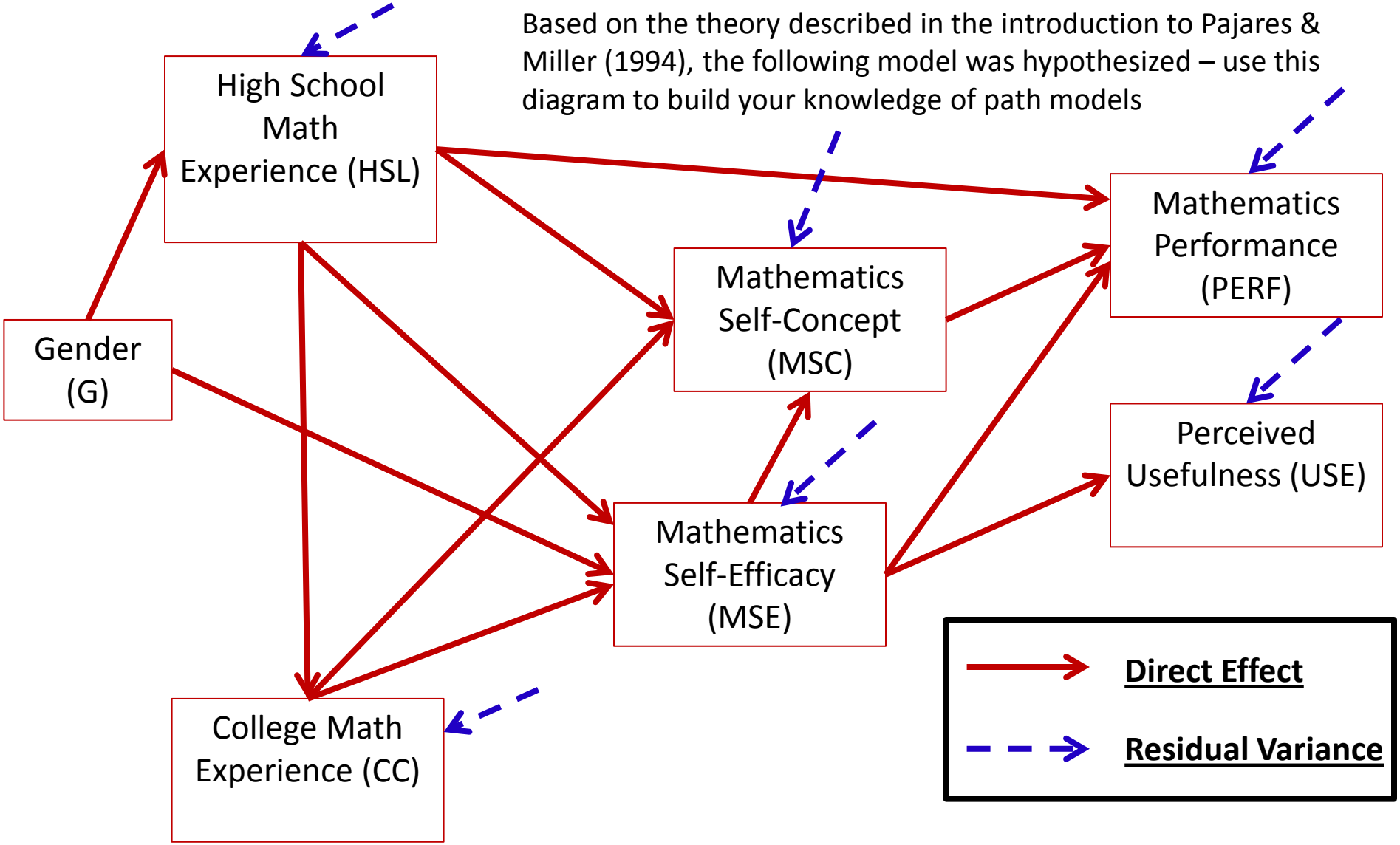
- **Just-Identified:** number of observed covariances = number of model parameters
  - Necessary for identification, but no model fit indices available

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

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

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

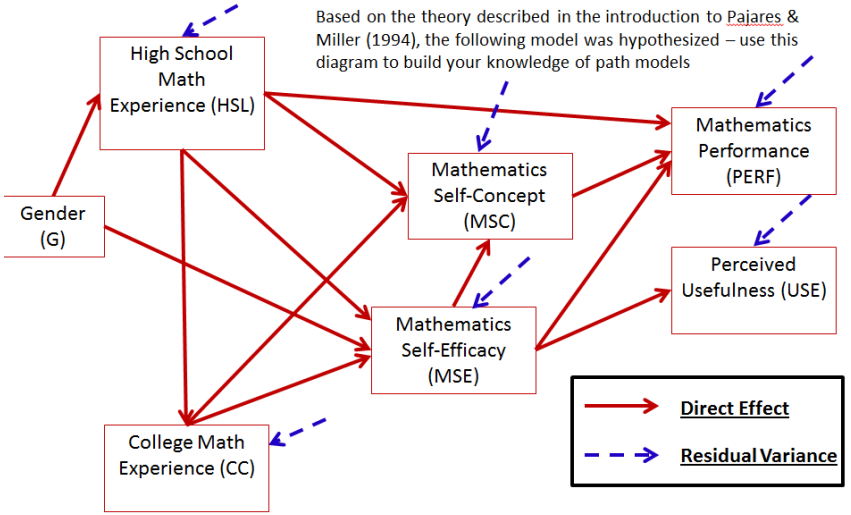
- Is the model identified?
  - Check the t-rule first (and only as it is recursive)
  - How many covariance terms are there in the all-variable matrix?
    - $\frac{7^{*}(7+1)}{2} = 28$
  - How many model parameters are to be estimated?
    - 12 direct paths
    - 6 residual variances
    - 1 variance 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. $CC_i = \beta_0^{CC} + \beta_{HSL}^{CC} HSL_i + e_i^{CC}$
3. $MSE_i = \beta_0^{MSE} + \beta_G^{MSE} G_i + \beta_{HSL}^{MSE} HSL_i + \beta_{CC}^{MSE} CC_i + e_i^{MSE}$
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. $PERF_i = \beta_0^{PERF} + \beta_{HSL}^{PERF} HSL_i + \beta_{MSE}^{PERF} MSE_i + \beta_{MSC}^{PERF} MSC_i + e_i^{PERF}$



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                             58.896
          Degrees of Freedom                    9
          P-Value                           0.0000

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                                0.917
          TLI                                0.806

Chi-Square Test of Model Fit for the Baseline Model

          Value                            619.926
          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 .917 and the TLI is .806. 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

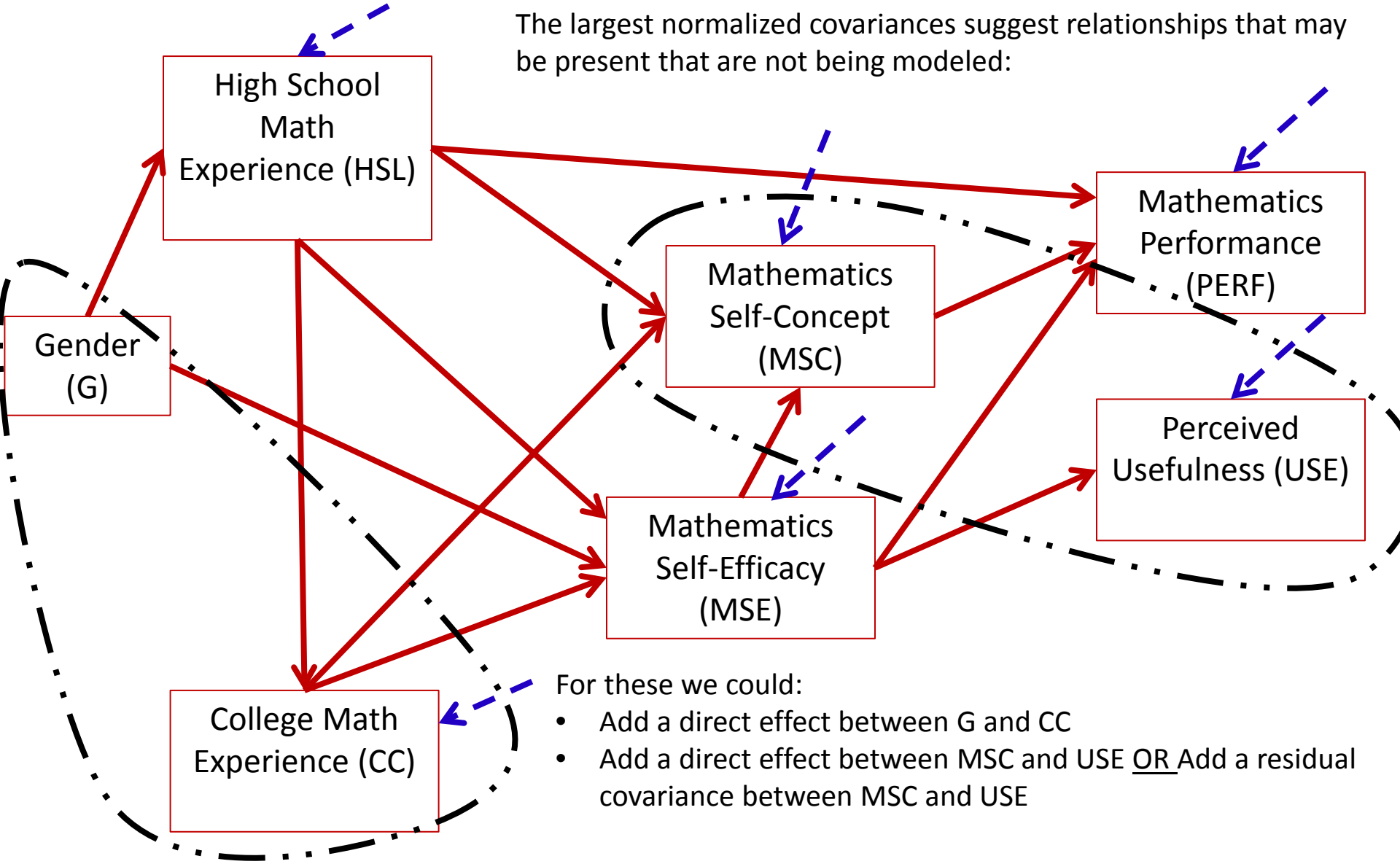
	HSL	CC	MSE	MSC	USE
HSL	0.035				
CC	-0.032	0.041			
MSE	0.080	-0.374	-0.083		
MSC	0.109	-0.161	-0.038	0.042	
USE	0.523	0.723	-0.108	4.442	0.040
PERF	0.005	-0.028	-0.068	0.060	-0.162
GENDER	0.091	-2.475	-0.408	-1.491	-0.026

Normalized Residuals for Covariances/Correlations/Residual Correlations

	PERF	GENDER
PERF	-0.079	
GENDER	-1.516	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 (also called Score or 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)

		M.I.	E.P.C.	Std E.P.C.	StdYX E.P.C.
ON Statements					
HSL	ON CC	6.478	0.447	0.447	1.992
HSL	ON MSE	6.493	1.139	1.139	10.270
HSL	ON MSC	4.091	0.165	0.165	2.146
HSL	ON USE	0.374	0.003	0.003	0.042
HSL	ON PERF	5.292	1.228	1.228	2.763
CC	ON MSE	6.475	-0.410	-0.410	-0.829
CC	ON MSC	6.477	-0.568	-0.568	-1.654
CC	ON USE	0.480	0.016	0.016	0.042
CC	ON PERF	0.058	-0.047	-0.047	-0.024
CC	ON GENDER	6.477	-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	1.074	1.096	1.096	0.274
MSC	ON USE	41.516	0.299	0.299	0.275
MSC	ON PERF	0.075	-0.414	-0.414	-0.072
MSC	ON GENDER	1.268	-1.669	-1.669	-0.046
USE	ON HSL	0.374	0.482	0.482	0.040
USE	ON CC	0.784	0.141	0.141	0.052
USE	ON MSC	40.032	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	2.572	-0.013	-0.013	-0.067
PERF	ON GENDER	2.219	-0.373	-0.373	-0.060

WITH Statements					
CC	WITH HSL	6.478	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	-14.272	-14.272	-0.094
USE	WITH MSC	41.517	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	0.750	3.528	3.528	0.183
PERF	WITH MSC	0.075	-1.567	-1.567	-0.067
PERF	WITH USE	2.572	-2.994	-2.994	-0.100
GENDER	WITH CC	6.477	-0.397	-0.397	-0.144
GENDER	WITH MSC	1.268	-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

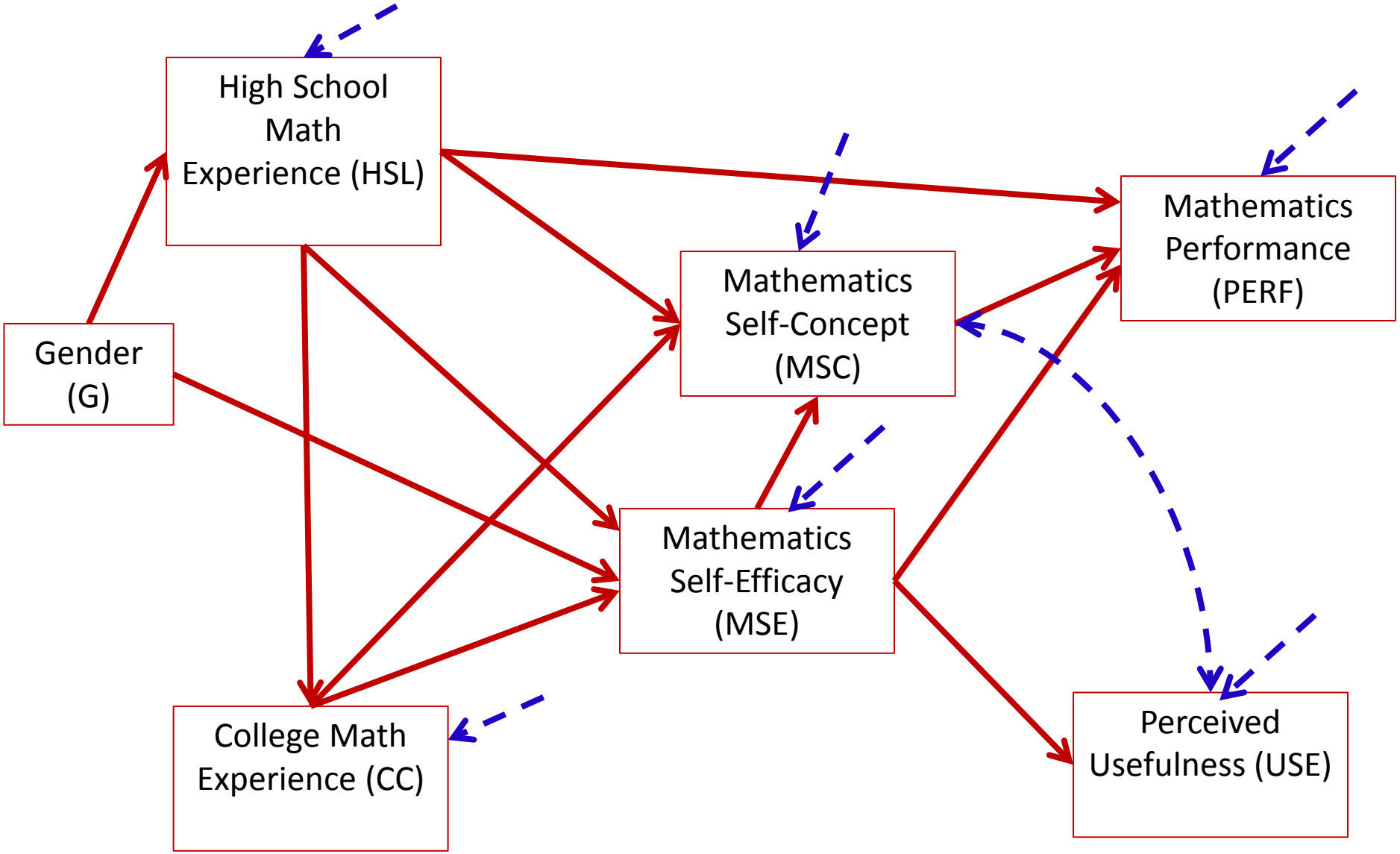
- The modification indices have three large values:
  - A direct effect predicting MSC from USE
  - A direct effect predicting USE from MSC
  - A residual covariance between USE and MSC

- 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 = 41.517
  - 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                             14.827
          Degrees of Freedom                    8
          P-Value                           0.0626

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.049
          90 Percent C.I.                    0.000  0.088
          Probability RMSEA <= .05           0.457

CFI/TLI

          CFI                                0.989
          TLI                                0.970

SRMR (Standardized Root Mean Square Residual)

          Value                              0.035
```

The comparison with the saturated model suggests our model fits statistically

The RMSEA is 0.049, 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	0.015				
CC	0.035	0.030			
MSE	0.019	-0.353	-0.099		
MSC	0.160	0.050	-0.104	0.053	
USE	0.597	0.774	0.063	0.296	0.020
PERF	0.060	0.017	-0.109	-0.004	-1.012
GENDER	0.051	-2.476	-0.347	-1.495	0.025

Normalized Residuals for Covariances/Correlations/Residual Correlations

	PERF	GENDER
PERF	-0.064	
GENDER	-1.492	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)

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	6.699	0.441	0.441	1.965
HSL	ON MSE	6.621	1.117	1.117	10.068
HSL	ON MSC	1.395	0.022	0.022	0.289
HSL	ON USE	0.494	0.004	0.004	0.048
HSL	ON PERF	4.403	0.773	0.773	1.737
CC	ON MSE	6.714	-0.429	-0.429	-0.869
CC	ON MSC	0.012	-0.008	-0.008	-0.023
CC	ON USE	0.435	0.015	0.015	0.040
CC	ON PERF	0.022	-0.029	-0.029	-0.015
CC	ON GENDER	6.697	-1.788	-1.788	-0.144
MSE	ON MSC	0.024	0.026	0.026	0.038
MSE	ON USE	0.932	-0.064	-0.064	-0.085
MSE	ON PERF	0.619	0.831	0.831	0.207
MSC	ON PERF	1.965	1.150	1.150	0.199
MSC	ON GENDER	1.872	-1.887	-1.887	-0.052
USE	ON HSL	0.494	0.554	0.554	0.046
USE	ON CC	0.732	0.135	0.135	0.050
USE	ON MSC	1.183	0.222	0.222	0.241
USE	ON PERF	2.566	-0.732	-0.732	-0.138
USE	ON GENDER	0.304	0.947	0.947	0.028
PERF	ON CC	0.085	0.006	0.006	0.013
PERF	ON USE	3.208	-0.015	-0.015	-0.081
PERF	ON GENDER	1.981	-0.350	-0.350	-0.056

WITH Statements					
CC	WITH HSL	6.701	14.968	14.968	1.949
MSC	WITH HSL	1.878	15.821	15.821	1.004
MSC	WITH MSE	1.870	44.019	44.019	0.373
USE	WITH HSL	0.417	0.877	0.877	0.043
USE	WITH CC	0.573	4.039	4.039	0.045
USE	WITH MSE	1.327	-18.046	-18.046	-0.118
PERF	WITH HSL	1.982	2.933	2.933	1.147
PERF	WITH CC	0.085	0.219	0.219	0.019
PERF	WITH MSE	0.597	3.159	3.159	0.165
PERF	WITH MSC	1.960	4.324	4.324	0.186
PERF	WITH USE	3.148	-3.087	-3.087	-0.103
GENDER	WITH CC	6.697	-0.404	-0.404	-0.146
GENDER	WITH MSC	1.872	-0.427	-0.427	-0.075
GENDER	WITH USE	0.304	0.214	0.214	0.029
GENDER	WITH PERF	1.981	-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 = 41.517
  - EPC = 70.912

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

- The difference in log-likelihoods is:
  - Original Model: -5,889.496
  - Modified Model: -5,867.461
  - -2*(change) = 44.07

- These are approximately the values given by the MI and EPC

Model Parameter Investigation

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HSL	ON				
GENDER		0.208	0.154	1.350	0.177
CC	ON				
HSL		0.662	0.257	2.578	0.010
MSE	ON				
HSL		4.138	0.436	9.489	0.000
GENDER		4.168	1.181	3.529	0.000
CC		0.393	0.102	3.867	0.000
MSC	ON				
HSL		2.823	0.582	4.849	0.000
CC		0.519	0.119	4.356	0.000
MSE		0.736	0.068	10.794	0.000
USE	ON				
MSE		0.277	0.074	3.732	0.000
PERF	ON				
MSE		0.139	0.015	9.564	0.000
MSC		0.037	0.010	3.815	0.000
HSL		0.153	0.112	1.366	0.172
PERF	WITH				
USE		0.000	0.000	999.000	999.000
MSC	WITH				
USE		70.247	11.526	6.095	0.000

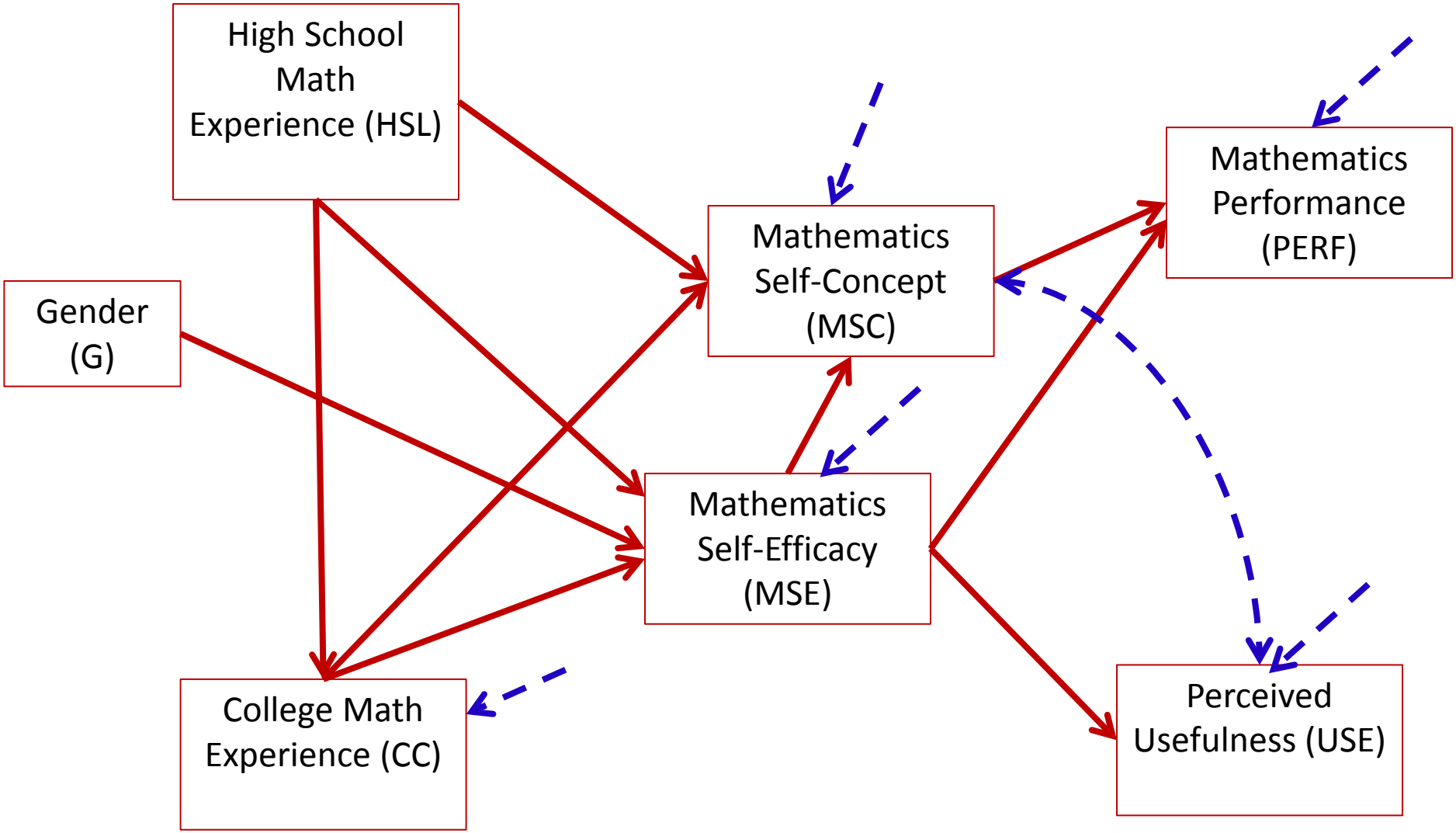
There are two direct effects that are non-significant:

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

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	18.579
Degrees of Freedom	10
P-Value	0.0460

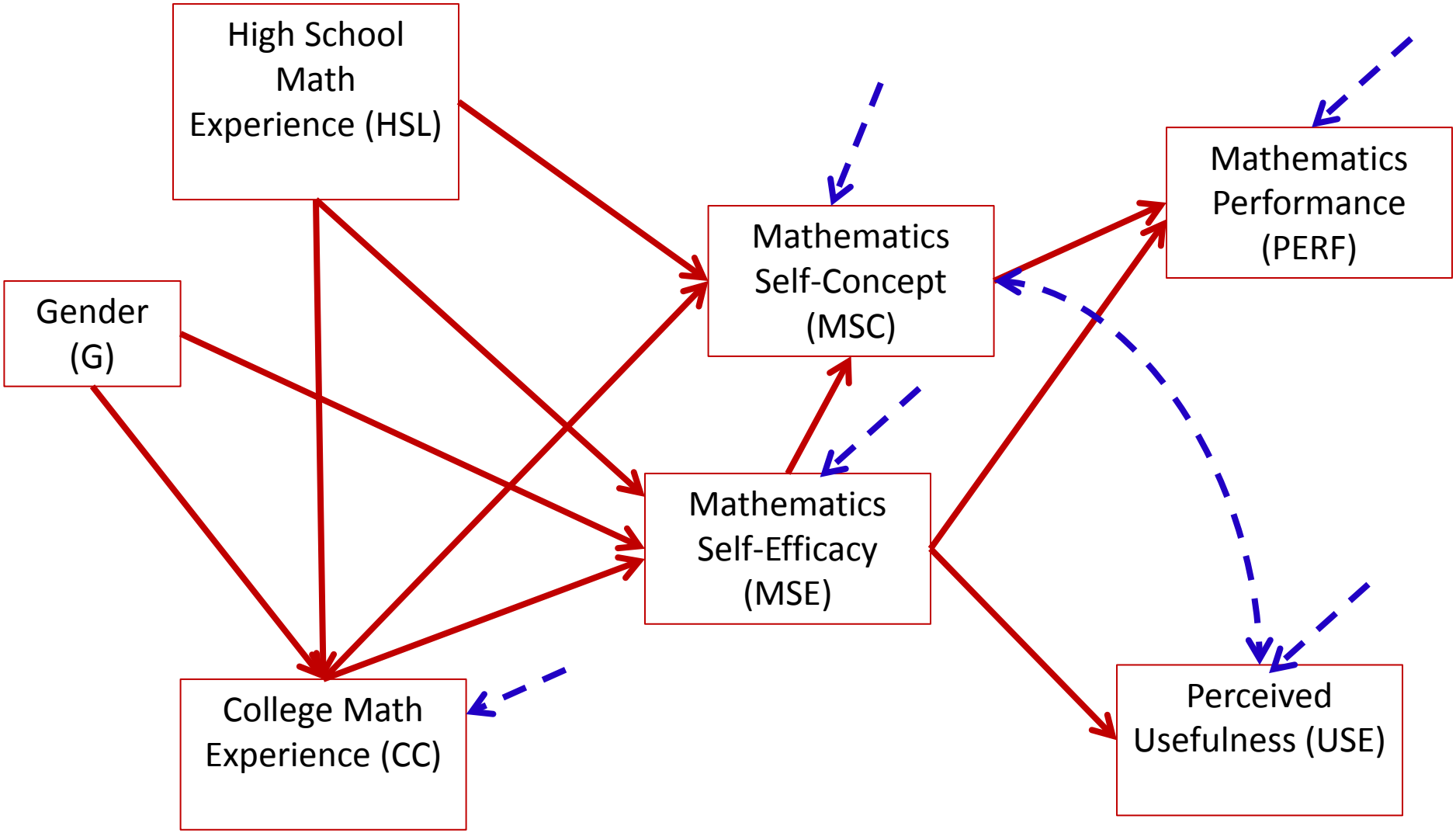
RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.050	
90 Percent C.I.	0.007	0.084
Probability RMSEA <= .05	0.462	

- Again, the largest normalized residual covariance is that of GENDER and CC
  - MI for direct effect of GENDER on CC is 6.595, 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	11.889
Degrees of Freedom	9
P-Value	0.2196

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.030	
90 Percent C.I.	0.000	0.071
Probability RMSEA <= .05	0.742	

- 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	0.255	2.775	0.006
	GENDER	-1.779	0.686	-2.595	0.009
MSE	ON				
	HSL	4.158	0.434	9.589	0.000
	GENDER	4.283	1.180	3.631	0.000
	CC	0.398	0.101	3.937	0.000
MSC	ON				
	HSL	2.831	0.581	4.874	0.000
	CC	0.528	0.118	4.466	0.000
	MSE	0.733	0.068	10.759	0.000
USE	ON				
	MSE	0.276	0.074	3.720	0.000
PERF	ON				
	MSE	0.145	0.014	10.441	0.000
	MSC	0.041	0.009	4.304	0.000
PERF	WITH				
	USE	0.000	0.000	999.000	999.000
MSC	WITH				
	USE	70.596	11.518	6.129	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	0.056	2.808	0.005
	GENDER	-0.143	0.054	-2.632	0.008
MSE	ON				
	HSL	0.466	0.044	10.541	0.000
	GENDER	0.172	0.047	3.646	0.000
	CC	0.199	0.050	3.958	0.000
MSC	ON				
	HSL	0.220	0.045	4.873	0.000
	CC	0.183	0.041	4.461	0.000
	MSE	0.508	0.042	11.953	0.000
USE	ON				
	MSE	0.206	0.054	3.811	0.000
PERF	ON				
	MSE	0.578	0.050	11.548	0.000
	MSC	0.234	0.054	4.312	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	0.056	2.808	0.005
GENDER	-0.301	0.114	-2.645	0.008
MSE ON				
HSL	0.466	0.044	10.541	0.000
GENDER	0.363	0.099	3.679	0.000
CC	0.199	0.050	3.958	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

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
  - See PSYC 948 in the Spring

- Modifications to model likelihoods or different distributional assumptions will help with the normality assumption
  - See next lecture

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 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