Path Analysis

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #14

In This Lecture...

- Path analysis: Multivariate Linear Models Where Outcomes Can Be Also Predictors

- Path analysis details:
  - Model identification
  - Modeling workflow

- Example Analyses

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.

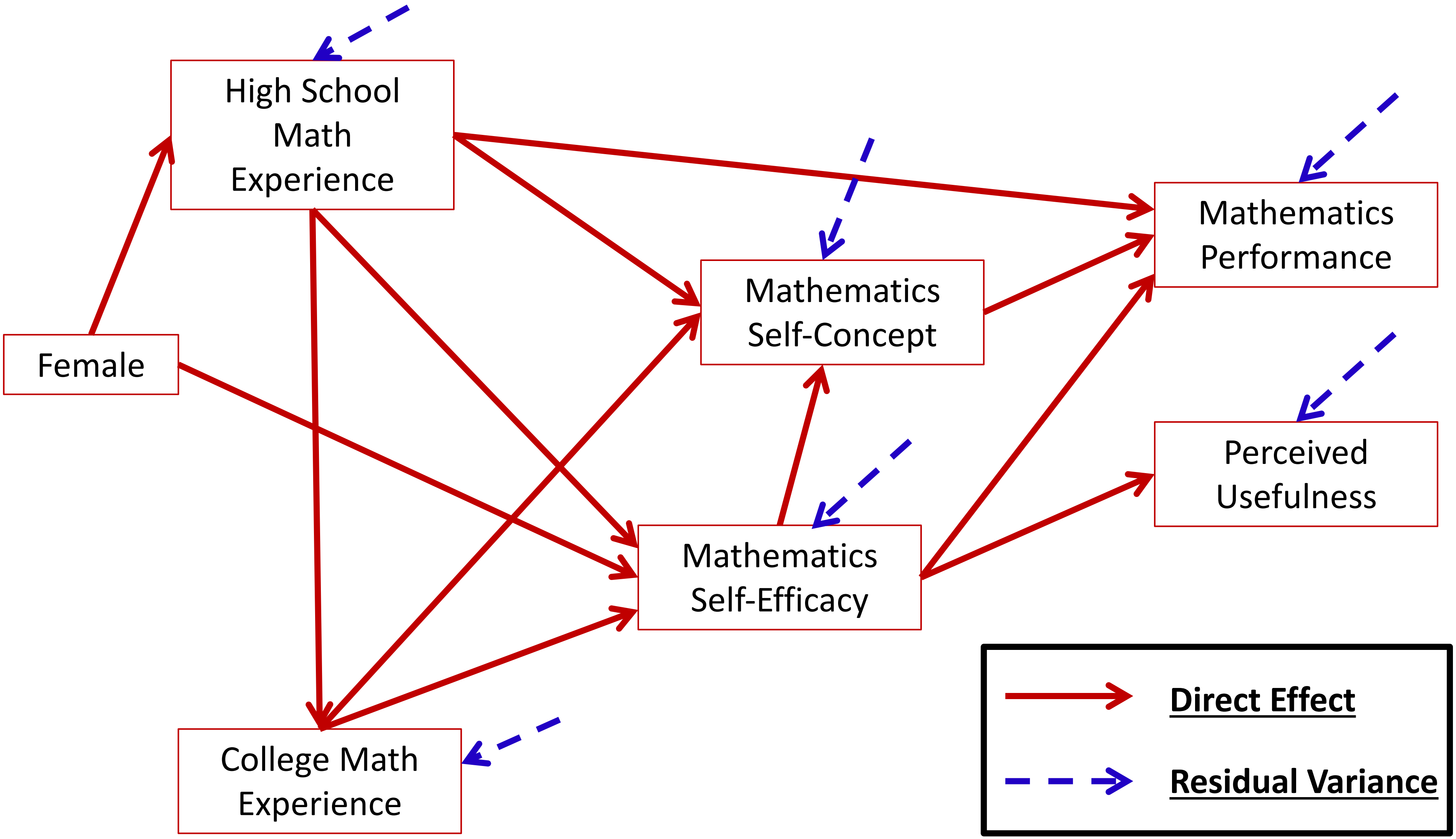
- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- 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

- Female (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

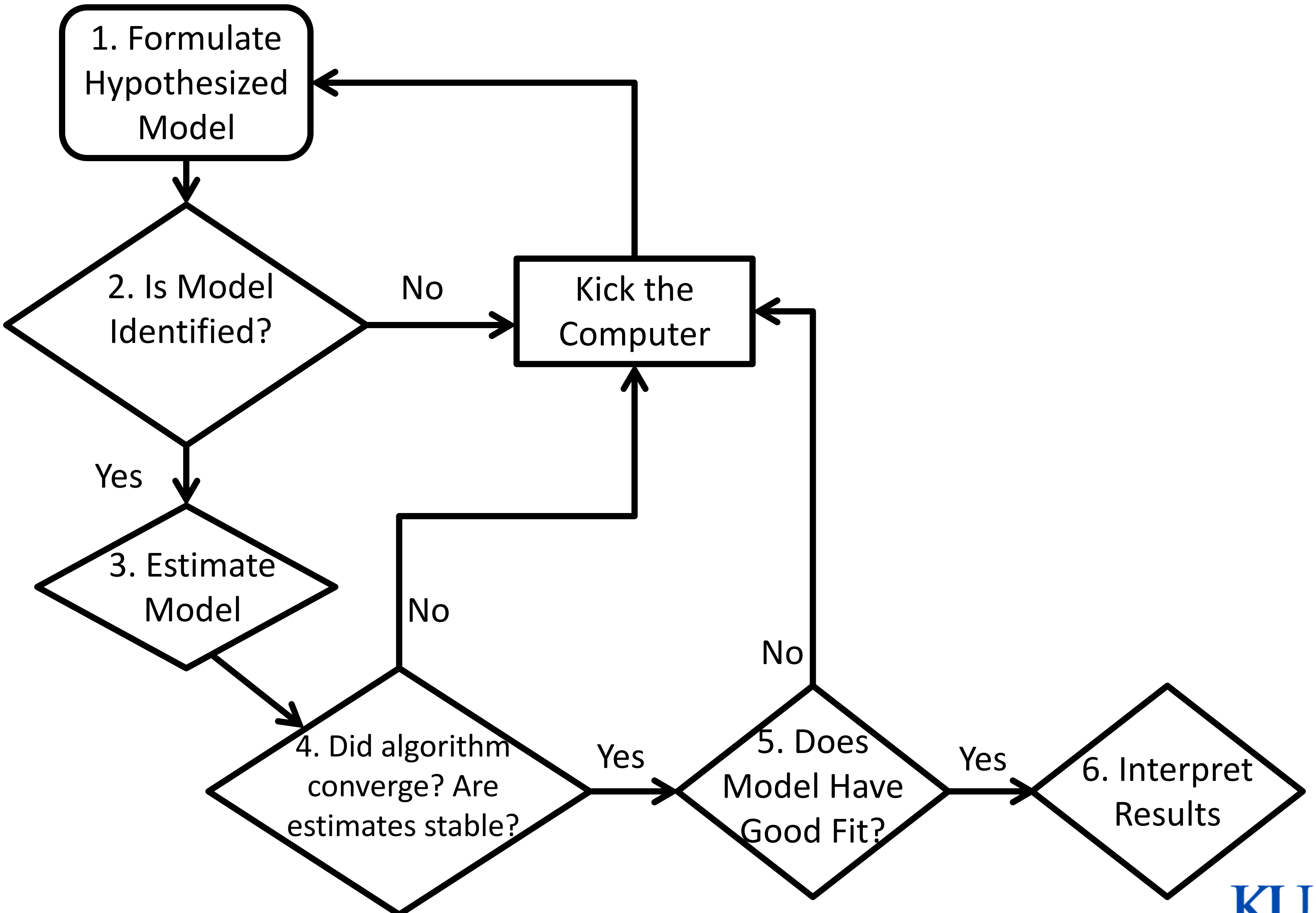
- By specifying simultaneous regression equations (the core of path models), a very specific covariance matrix is implied
  - This is where things deviate from our familiar R matrix

- 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

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

- For path models, identification can be very difficult

- 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

- Model 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
  - 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 likelihood function, but are considered "saturated" so no additional parameters can be added (all parameters are estimated)
  - These do not enter into the equation for the covariance matrix of the endogenous and exogenous variables

T-rule Identification Status

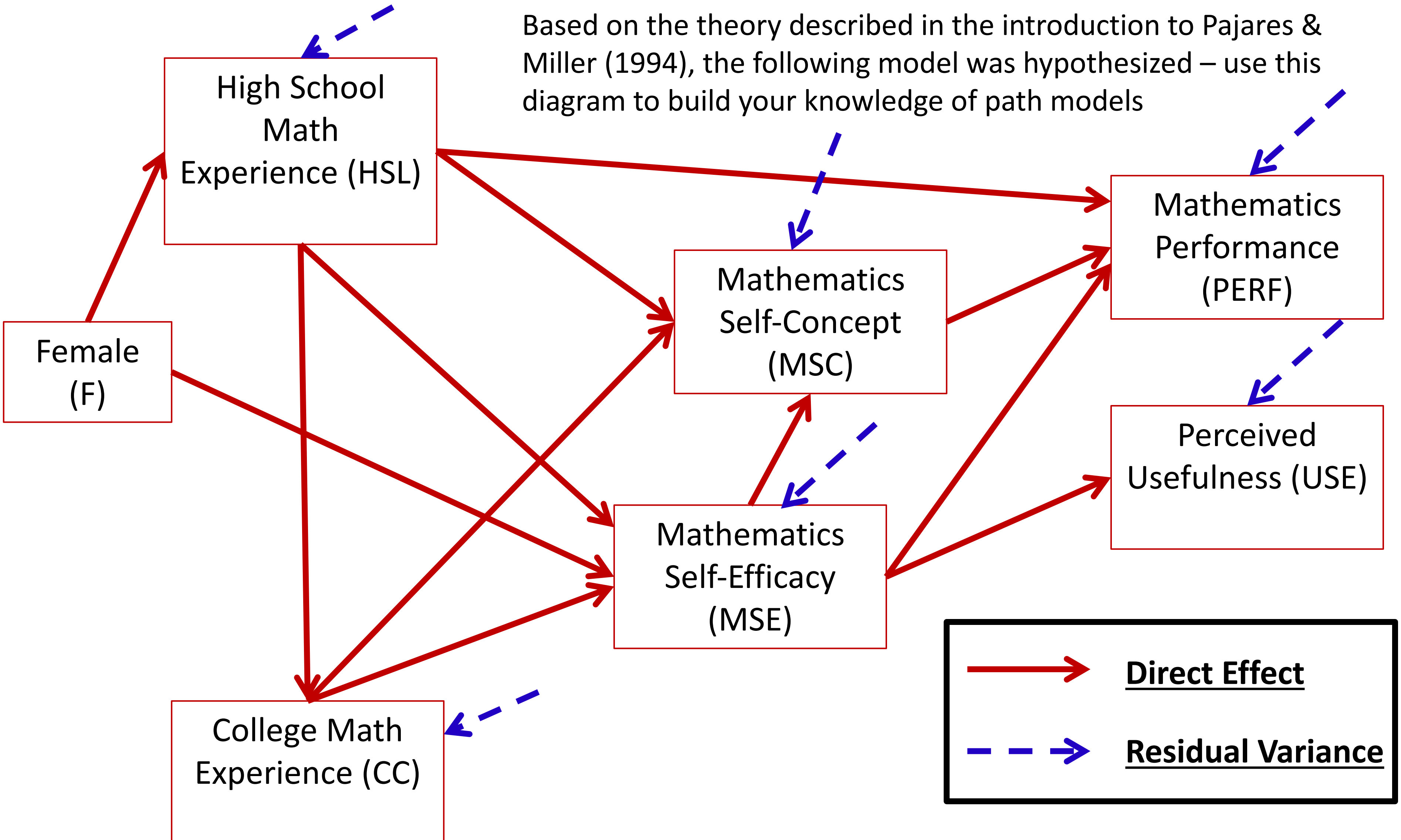
- **Just-Identified:** number of observed covariances = number of model parameters
  - Necessary for identification, but no model fit indices available

- **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

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 R

- **The model is over-identified**
  - 28 total variance/covariances but 19 model parameters
  - We can use R 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:

$$HSL_i = \beta_{0,HSL} + \beta_{F,HSL}F_i + e_{i,HSL}$$

$$CC_i \quad = \beta_{0,CC} + \beta_{HSL,CC}HSL_i + e_{i,CC}$$

$$MSE_i = \beta_{0,MSE} + \beta_{F,MSE}F_{\mathrm{i}} + \beta_{HSL,MSE}HSL_i + \beta_{CC,MSE}CC_i + e_{i,MSE}$$

$$MSC_i = \beta_{0,MSC} + \beta_{HSL,MSC}HSL_i + \beta_{CC,MSC}CC_i + \beta_{MSE,MSC}MSE_i + e_{i,MSC}$$

$$USE_i = \beta_{0,USE} + \beta_{MSE,USE}MSE_i + e_{i,USE}$$

$$PERF_i = \beta_{0,PERF} + \beta_{HSL,PERF}HSL_i + \beta_{MSE,PERF}MSE_i + \beta_{MSC,PERF}MSC_i + e_{i,PERF}$$

Path Model Estimation

- 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 R

```
#model 01---------------------------------------------------------------------
model01.syntax =
"
#endogenous variable equations
perf ~ hsl + msc + mse
use  ~ mse
mse  ~ hsl + cc + female
msc  ~ mse + cc + hsl
cc   ~ hsl
hsl  ~ female

#endogenous variable intercepts
perf ~ 1
use  ~ 1
mse  ~ 1
msc  ~ 1
cc   ~ 1
hsl  ~ 1

#endogenous variable residual variances
perf ~~ perf
use  ~~ use
mse  ~~ mse
msc  ~~ msc
cc   ~~ cc
hsl  ~~ hsl

#endogenous variable residual covariances
#none specfied in the original model so these have zeros:
perf ~~ 0*use + 0*mse + 0*msc + 0*cc + 0*hsl
use  ~~ 0*mse + 0*msc + 0*cc + 0*hsl
mse  ~~ 0*msc + 0*cc + 0*hsl
msc  ~~ 0*cc + 0*hsl
cc   ~~ 0*hsl
"
```

Model Fit Evaluation

- First, we check convergence:

```
> inspect(model01.fit, what="converged")
[1] TRUE
>
```

  - lavaan's 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

```
Estimator                                          ML       Robust
Minimum Function Test Statistic                58.896       58.913
Degrees of freedom                                  9            9
P-value (Chi-square)                            0.000        0.000
Scaling correction factor                                    1.000
  for the Yuan-Bentler correction (Mplus variant)
```

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).

```
Root Mean Square Error of Approximation:

  RMSEA                                         0.126        0.126
  90 Percent Confidence Interval         0.096  0.157        0.096  0.157
  P-value RMSEA <= 0.05                         0.000        0.000
```

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

```
User model versus baseline model:

  Comparative Fit Index (CFI)                   0.917        0.918
  Tucker-Lewis Index (TLI)                      0.806        0.809
```

The CFI estimate is .917 and the TLI is .806. Good fit is considered 0.95 or higher.

```
Model test baseline model:

  Minimum Function Test Statistic             619.926      629.882
  Degrees of freedom                               21           21
  P-value                                       0.000        0.000
```

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

```
Standardized Root Mean Square Residual:

  SRMR                                          0.056        0.056
```

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
- ***<u>Generally, model modification should be guided by theory</u>***
  - However, we can inspect the normalized residual covariance matrix (like z-scores) to see where our biggest misfit occurs

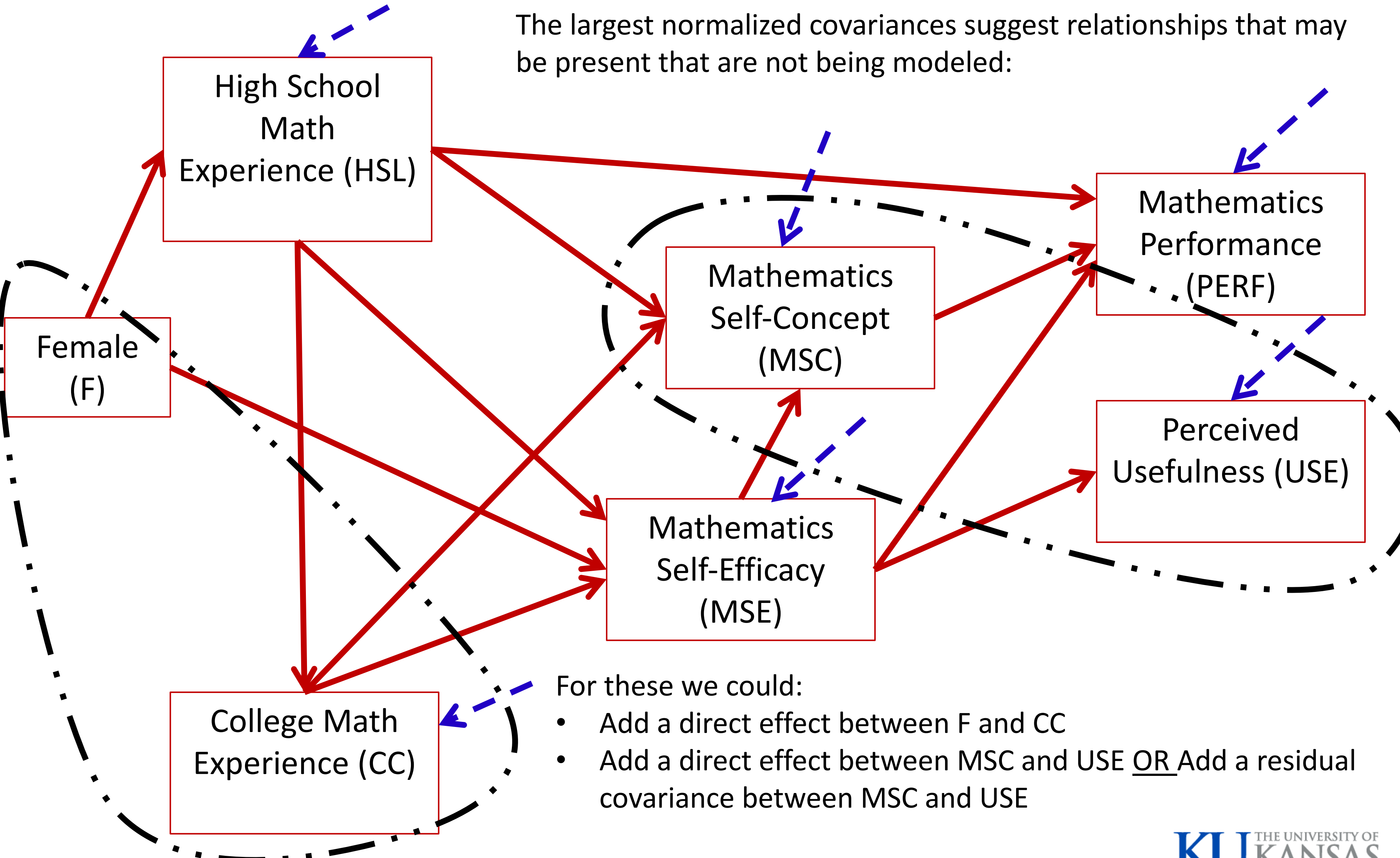
```
> residuals(model01.fit ,type="normalized")
$type
[1] "normalized"

$cov
       perf   use    mse    msc    cc     hsl    female
perf   -0.076
use    -0.159  0.041
mse    -0.071  0.110  0.086
msc     0.059  5.051 -0.039  0.043
cc     -0.028  0.720 -0.377 -0.161  0.046
hsl     0.006  0.559  0.085  0.105  0.034  0.039
female -1.522 -0.027 -0.422 -1.452 -2.567  0.091  0.000

$mean
  perf    use    mse    msc     cc    hsl female
-0.014  0.126  0.012  0.211  0.009  0.004  0.000
```

One normalized residual covariance is bigger than +/-1.96:
MSC with USE and
CC with Female

THE UNIVERSITY OF KANSAS

Our Destination: Overall Path Model

The largest normalized covariances suggest relationships that may be present that are not being modeled:



For these we could:

- Add a direct effect between F and CC
- Add a direct effect between MSC and USE <u>OR</u> Add a residual covariance between MSC and USE

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)

```
> model01.mi
     lhs op    rhs     mi mi.scaled    epc sepc.lv sepc.all sepc.nox
54   msc  ~    use 41.517    41.529  0.299   0.299    0.275    0.275
31   use ~~    msc 41.517    41.529 70.912  70.912    0.262    0.262
46   use  ~    msc 40.032    40.044  0.451   0.451    0.490    0.490
63   hsl  ~    mse  6.477     6.479  1.138   1.138   10.258   10.258
60    cc  ~ female  6.477     6.478 -1.756  -1.756   -0.142   -0.298
65   hsl  ~     cc  6.477     6.478  0.447   0.447    1.992    1.992
39    cc ~~    hsl  6.477     6.478 15.131  15.131    1.945    1.945
58    cc  ~    mse  6.476     6.478 -0.410  -0.410   -0.829   -0.829
59    cc  ~    msc  6.476     6.478 -0.568  -0.568   -1.654   -1.654
```


THE UNIVERSITY OF KANSAS

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

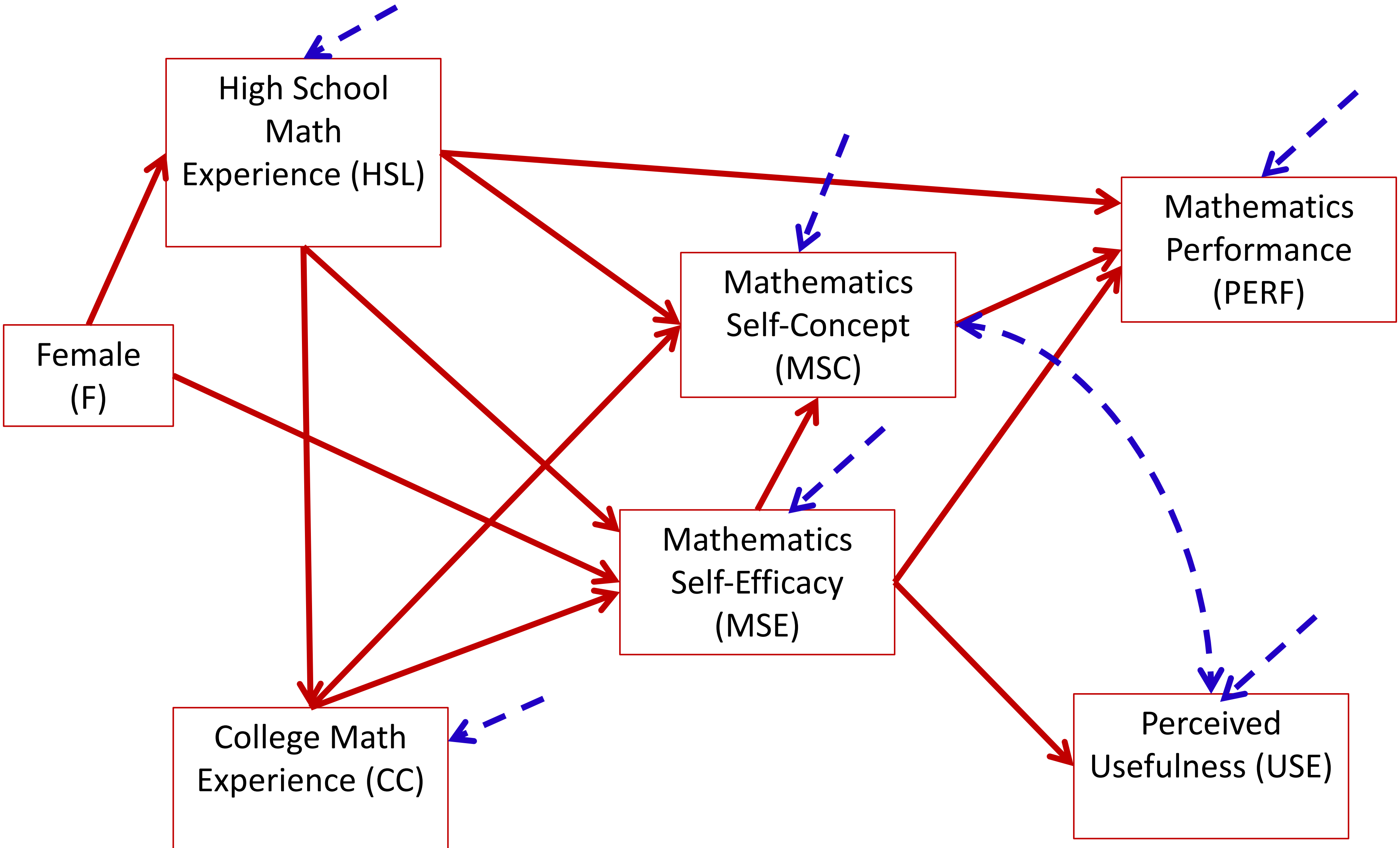
- All three are for the same variable: so we can only choose one
  - This is where theory would help us decide

- As we do not know theory, we will choose to add a residual covariance between USE and MSC (the "~~" symbol)
  - 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

New Model Syntax

```
model02.syntax =
"
#endogenous variable equations
perf ~ hsl + msc + mse
use  ~ mse
mse  ~ hsl + cc + female
msc  ~ mse + cc + hsl
cc   ~ hsl
hsl  ~ female

#endogenous variable intercepts
perf ~ 1
use  ~ 1
mse  ~ 1
msc  ~ 1
cc   ~ 1
hsl  ~ 1

#endogenous variable residual variances
perf ~~ perf
use  ~~ use
mse  ~~ mse
msc  ~~ msc
cc   ~~ cc
hsl  ~~ hsl

#endogenous variable residual covariances
#none specfied in the original model so these have zeros:
perf ~~ 0*use + 0*mse + 0*msc + 0*cc + 0*hsl
use  ~~ 0*mse + msc + 0*cc + 0*hsl      #<- the changed part of syntax here (no 0* in front of msc)
mse  ~~ 0*msc + 0*cc + 0*hsl
msc  ~~ 0*cc + 0*hsl
cc   ~~ 0*hsl
"
```

Modified Model #02



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)
  - Estimation converged; Standard errors look acceptable

```
Estimator                                         ML      Robust
Minimum Function Test Statistic               14.827      14.393
Degrees of freedom                                 8           8
P-value (Chi-square)                           0.063       0.072
Scaling correction factor                                  1.030
  for the Yuan-Bentler correction (Mplus variant)
```

The comparison with the saturated model suggests our model fits statistically

```
Root Mean Square Error of Approximation:

  RMSEA                                         0.049       0.048
  90 Percent Confidence Interval          0.000  0.088       0.000  0.086
  P-value RMSEA <= 0.05                         0.457       0.484
```

The RMSEA is 0.048, which indicates good fit

```
User model versus baseline model:

  Comparative Fit Index (CFI)                    0.989       0.990
  Tucker-Lewis Index (TLI)                       0.970       0.972
```

The CFI and TLI both indicate good fit

```
Standardized Root Mean Square Residual:

  SRMR                                           0.035       0.035
```

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 Female
  - Given the number of covariances we have, this is likely okay

```
> residuals(model02.fit ,type="normalized")
$type
[1] "normalized"

$cov
       perf   use    mse    msc    cc     hsl    female
perf   -0.062
use    -0.990  0.020
mse    -0.113  0.064 -0.103
msc    -0.003  0.337 -0.104  0.054
cc      0.018  0.771 -0.356  0.050  0.034
hsl     0.062  0.638  0.020  0.154  0.037  0.017
female -1.499  0.026 -0.359 -1.456 -2.568  0.051  0.000

$mean
  perf    use    mse    msc     cc    hsl female
-0.013  0.065  0.027  0.044  0.003 -0.015  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)

```
> model02.mi
      lhs op    rhs    mi mi.scaled    epc sepc.lv sepc.all sepc.nox
39     cc ~~    hsl 6.697     6.501 14.964  14.964    1.922    1.922
60     cc  ~ female 6.697     6.501 -1.788  -1.788   -0.144   -0.304
65    hsl  ~     cc 6.697     6.501  0.441   0.441    1.965    1.965
58     cc  ~    mse 6.697     6.501 -0.429  -0.429   -0.868   -0.868
63    hsl  ~    mse 6.697     6.501  1.124   1.124   10.128   10.128
61    hsl  ~   perf 4.410     4.281  0.774   0.774    1.739    1.739
42   perf  ~    use 3.208     3.114 -0.015  -0.015   -0.081   -0.081
25   perf ~~    use 3.148     3.056 -3.087  -3.087   -0.066   -0.066
45    use  ~   perf 2.565     2.490 -0.732  -0.732   -0.138   -0.138
44   perf  ~ female 1.981     1.923 -0.350  -0.350   -0.056   -0.118
29   perf ~~    hsl 1.981     1.923  2.933   2.933    0.747    0.747
```

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: 70.249

- The difference in log-likelihoods is:
  - Original Model: -6,126.013
  - Modified Model: -6,103.978
  - -2*(change) = 58.279

- The values given by the MI and EPC are approximations

Model Parameter Investigation

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  perf ~
    hsl               0.153    0.107    1.432    0.152    0.153    0.068
    msc               0.037    0.009    4.147    0.000    0.037    0.215
    mse               0.139    0.013   10.700    0.000    0.139    0.557
  use ~
    mse               0.277    0.073    3.803    0.000    0.277    0.209
  mse ~
    hsl               4.138    0.406   10.203    0.000    4.138    0.459
    cc                0.393    0.105    3.723    0.000    0.393    0.194
    female            4.168    1.160    3.593    0.000    4.168    0.166
  msc ~
    mse               0.736    0.066   11.119    0.000    0.736    0.512
    cc                0.519    0.117    4.434    0.000    0.519    0.179
    hsl               2.824    0.593    4.764    0.000    2.824    0.218
  cc ~
    hsl               0.662    0.247    2.686    0.007    0.662    0.149
  hsl ~
    female            0.208    0.154    1.348    0.178    0.208    0.075
```

There are two direct effects that are non-significant:

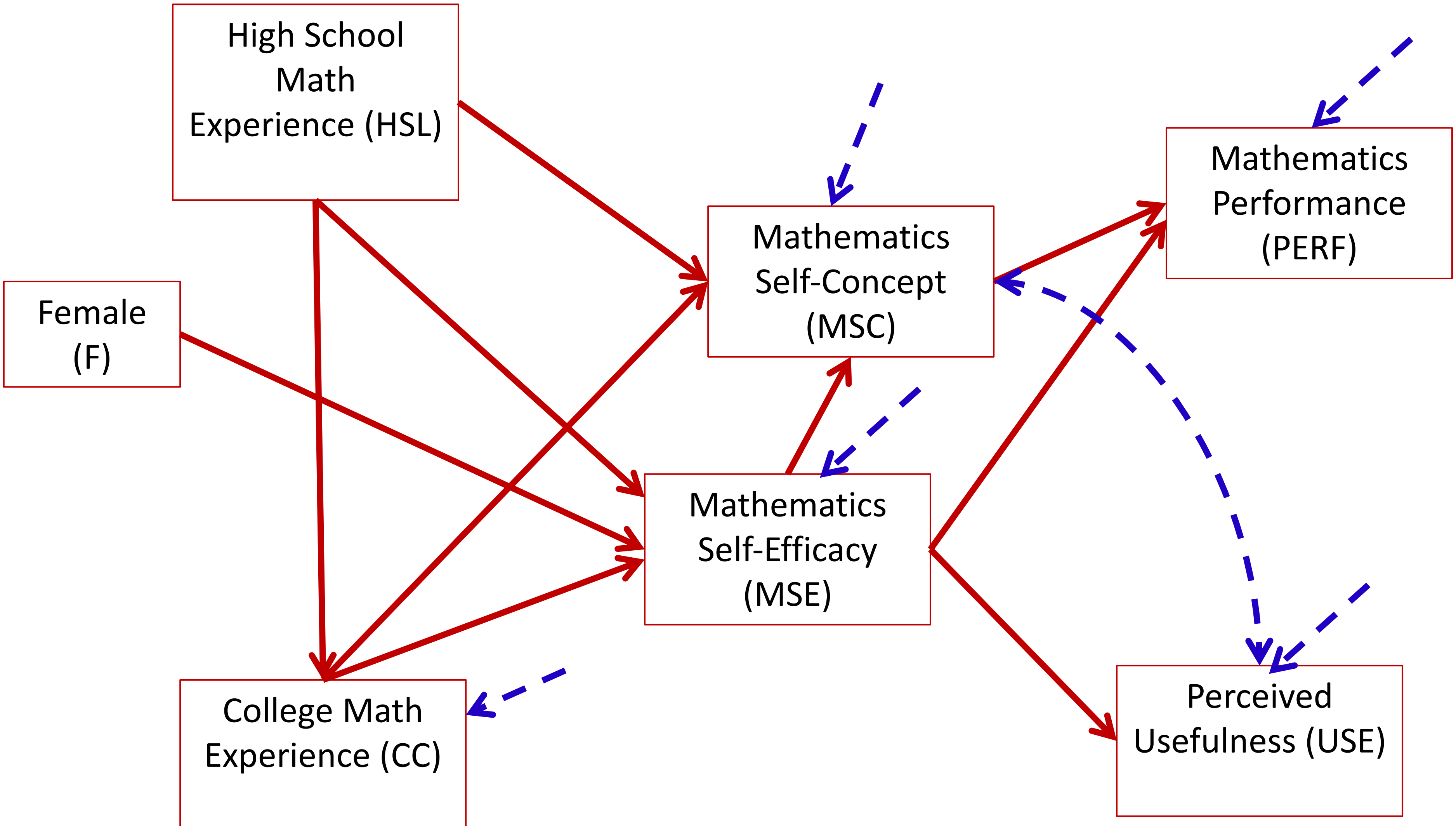
$$\beta_{F,HSL} = 0.208$$
$$\beta_{HSL,PERF} = 0.153$$

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

THE UNIVERSITY OF KANSAS

Modified Model #03



Model #03 Syntax

```
model03.syntax =
  "
#endogenous variable equations
perf ~ msc + mse
use  ~ mse
mse  ~ hsl + cc + female
msc  ~ mse + cc + hsl
cc   ~ hsl

#endogenous variable intercepts
perf ~ 1
use  ~ 1
mse  ~ 1
msc  ~ 1
cc   ~ 1

#endogenous variable residual variances
perf ~~ perf
use  ~~ use
mse  ~~ mse
msc  ~~ msc
cc   ~~ cc

#endogenous variable residual covariances
#none specfied in the original model so these have zeros:
perf ~~ 0*use + 0*mse + 0*msc + 0*cc
use  ~~ 0*mse + msc + 0*cc
mse  ~~ 0*msc + 0*cc
msc  ~~ 0*cc
"
```

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 not be as good as we would think

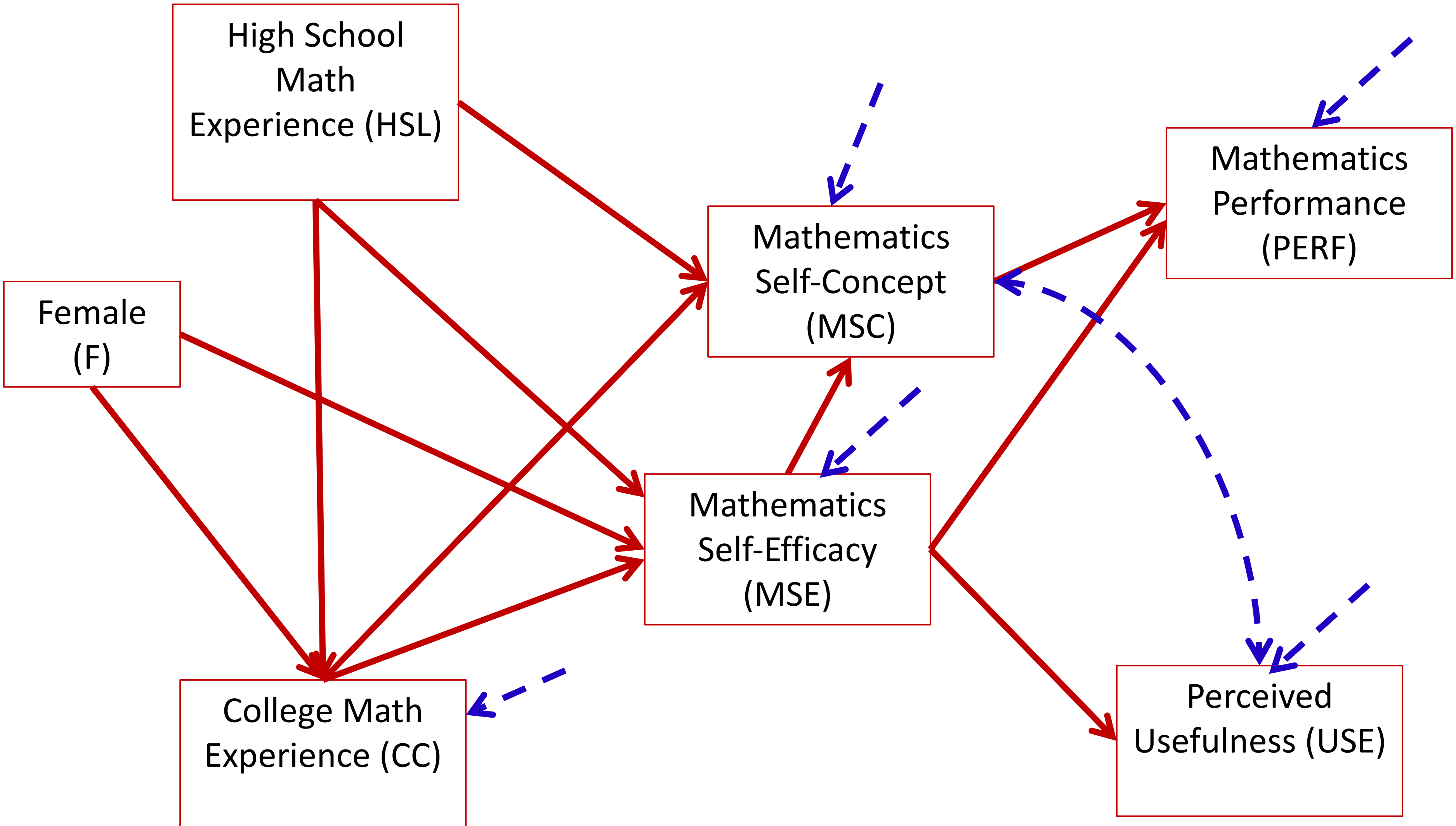
```
Estimator                                         ML       Robust
Minimum Function Test Statistic               16.687       16.443
Degrees of freedom                                 9            9
P-value (Chi-square)                           0.054        0.058
Scaling correction factor                                   1.015
  for the Yuan-Bentler correction (Mplus variant)

 Root Mean Square Error of Approximation:

  RMSEA                                        0.049        0.049
  90 Percent Confidence Interval        0.000  0.086        0.000  0.085
  P-value RMSEA <= 0.05                        0.460        0.475
```

- Again, the largest normalized residual covariance is that of Female and CC
  - MI for direct effect of Female on CC is 6.706, indicating that adding this parameter may improve model fit

- So, we will now add a direct effect of Female on CC

THE UNIVERSITY OF KANSAS

Modified Model #04



Model 04 Syntax

```
model04.syntax = "
#endogenous variable equations
perf ~ msc + mse
use  ~ mse
mse  ~ (b_hsl_mse)*hsl + (b_cc_mse)*cc + female
msc  ~ mse + cc + hsl
cc   ~ (b_hsl_cc)*hsl + female

#endogenous variable intercepts
perf ~ 1
use  ~ 1
mse  ~ 1
msc  ~ 1
cc   ~ 1

#endogenous variable residual variances
perf ~~ perf
use  ~~ use
mse  ~~ mse
msc  ~~ msc
cc   ~~ cc

#endogenous variable residual covariances
#none specfied in the original model so these have zeros:
perf ~~ 0*use + 0*mse + 0*msc + 0*cc
use  ~~ 0*mse + msc + 0*cc
mse  ~~ 0*msc + 0*cc
msc  ~~ 0*cc

#indirect effect of interest:
ind_hsl_mse := b_hsl_cc*b_cc_mse

#total effect of interest:
tot_hsl_mse := b_hsl_mse + (b_hsl_cc*b_cc_mse)
"
```

Model #04: 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

```
Estimator                                         ML       Robust
Minimum Function Test Statistic                9.923        9.694
Degrees of freedom                                 8            8
P-value (Chi-square)                           0.270        0.287
Scaling correction factor                                   1.024
  for the Yuan-Bentler correction (Mplus variant)


 Root Mean Square Error of Approximation:

  RMSEA                                         0.026       0.025
  90 Percent Confidence Interval       0.000    0.071       0.000   0.070
  P-value RMSEA <= 0.05                         0.764       0.781
```

- 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 #6 Parameter Interpretation

Interpret each of these parameters as you would in regression:

A one-unit increase in HSL brings about a .704 unit increase in CC, holding Female constant

We can interpret the standardized parameter estimates for all variables except gender

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

```
Regressions:
                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  perf ~
    msc                 0.041    0.009    4.670    0.000    0.041    0.234
    mse                 0.145    0.013   11.425    0.000    0.145    0.579
  use ~
    mse                 0.276    0.073    3.785    0.000    0.276    0.207
  mse ~
    hsl   (b_hsl_m)     4.162    0.403   10.329    0.000    4.162    0.464
    cc       (b_c_)     0.398    0.105    3.802    0.000    0.398    0.198
    feml                4.222    1.154    3.657    0.000    4.222    0.169
  msc ~
    mse                 0.731    0.066   11.036    0.000    0.731    0.508
    cc                  0.529    0.116    4.556    0.000    0.529    0.183
    hsl                 2.851    0.591    4.821    0.000    2.851    0.221
  cc ~
    hsl   (b_hsl_c)     0.704    0.245    2.869    0.004    0.704    0.158
    feml               -1.790    0.671   -2.667    0.008   -1.790   -0.144
```


THE UNIVERSITY OF KANSAS

Overall Model Interpretation

- High School Experience and Female are significant predictors of College Experience
  - Females lower than males 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
  - Females have higher Math Self-Efficacy than Men

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

THE UNIVERSITY OF KANSAS

Model Interpretation: Explained Variability

- The R^2 for each endogenous variable:
  - CC – 0.042
  - MSE – 0.313
  - MSC – 0.511
  - USE – 0.043
  - PERF – 0.570

```
> inspect(model04.fit, what="r2") #r-squared values for DVs
 perf   use   mse   msc    cc
0.570 0.043 0.313 0.511 0.042
```

- 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

THE UNIVERSITY OF KANSAS

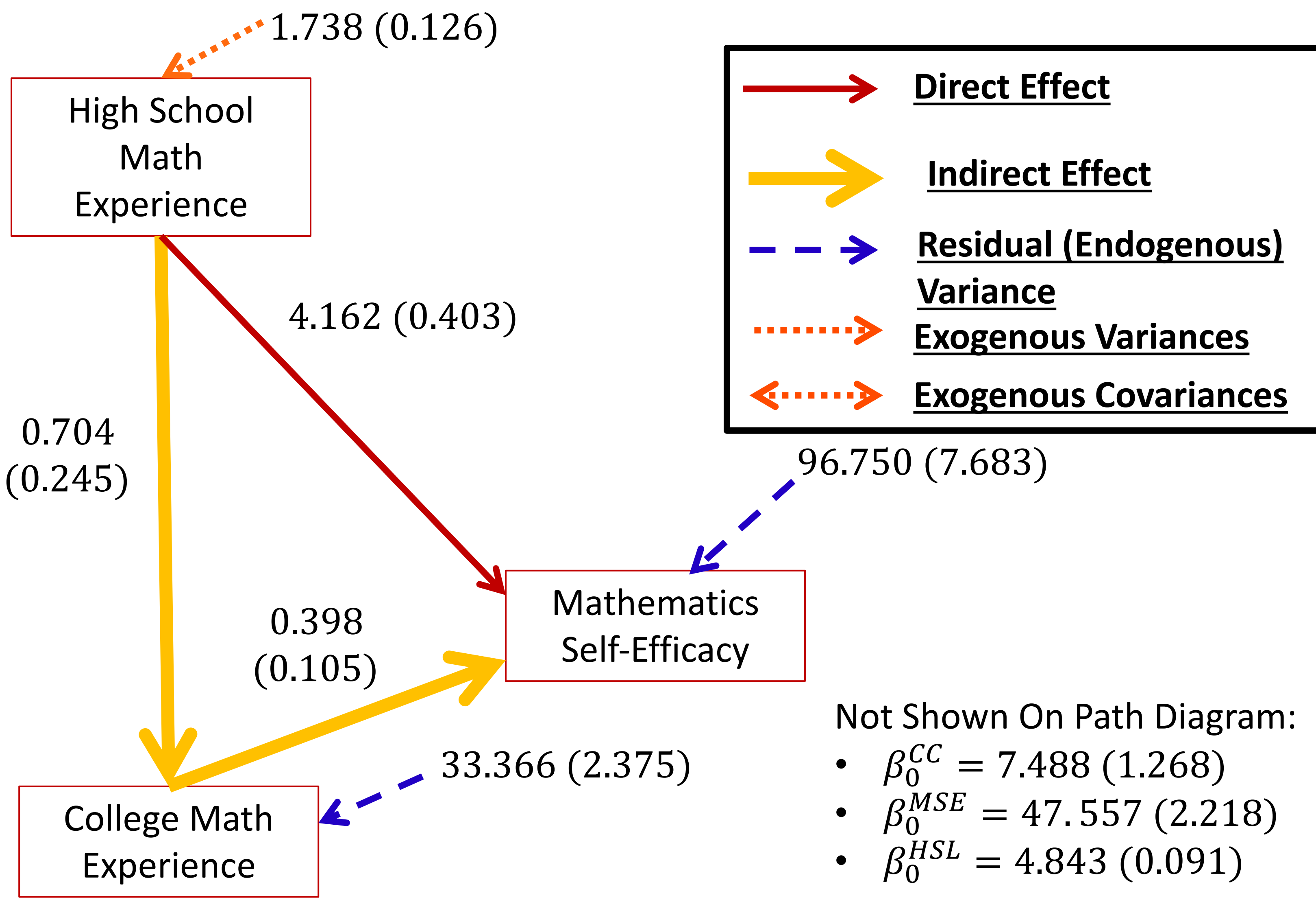
Indirect Paths

- 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

- A number of other indirect paths exist in the model

Direct and Indirect Effects of HSL on MSE (Part of Model 3)



Not Shown On Path Diagram:

- $\beta_0^{CC} = 7.488\ (1.268)$
- $\beta_0^{MSE} = 47.557\ (2.218)$
- $\beta_0^{HSL} = 4.843\ (0.091)$

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: $\beta_{HSL}^{CC} = 0.704$
  - The path between College Experience (CC) and Math Self-Efficacy (MSE): $\beta_{CC}^{MSE} = 0.398$

- The **indirect effect** of HSL on MSE is the product of these two terms: $\beta_{HSL}^{CC}\beta_{CC}^{MSE} = 0.704 * 0.398 = 0.280$

- 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 0.704 (the direct effect of HSL on CC)
  - Then, as CC increases by 0.704, HSL increases by 0.398 (the direct effect of CC on MSE)

- Indirectly, MSE increases by 0.280 (the multiplication of the two direct effects) for every one unit increase of HSL

THE UNIVERSITY OF KANSAS

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:

$$\beta_{HSL}^{MSE} + \beta_{HSL}^{CC}\beta_{CC}^{MSE} = 4.162 + 0.704^{*}0.398 = 4.443$$

- This means that for every one-unit increase in HSL, the total increase in MSE is 4.443
  - The direct effect represents the increase holding CC constant, which is implausible in this model

Hypothesis Tests for Indirect and Total Effects in lavaan

- 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
  - This test uses a bootstrap (resampling) technique to get the p-value

- In lavaan, first label parameters:

- Then add effects:

```
#indirect effect of interest:
  ind_hsl_mse := b_hsl_cc*b_cc_mse

#total effect of interest:
  tot_hsl_mse := b_hsl_mse + (b_hsl_cc*b_cc_mse)
```

lavaan Output

- Lavaan provides the total and indirect effects between terminating and originating variables
  - If the standardized=TRUE command is included in the summary() function call, the standardized versions of these effects are also given (the increase in standard deviations)

```
Defined Parameters:
                   Estimate  Std.Err  Z-value  P(>|z|)   Std.lv  Std.all
    ind_hsl_mse       0.280    0.114    2.452    0.014    0.280    0.031
    tot_hsl_mse       4.443    0.414   10.741    0.000    4.443    0.495
```

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

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

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

KU THE UNIVERSITY OF KANSAS

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