MESA 8668 Multivariate Statistics (Dr. Li)

Lab: Structural Equation Modeling

Structural equation modeling is known as latent variable modeling,latent variable path analysis, covariance structure analysis, causal modeling, etc. It is a family of statistical models that seek to explain the relationship between latent(unobserved) variables or constructs that are measured by manifest(observed) variables or indictors. It can be thought of as a combination of regression analysis and factor analysis. SEM estimates a series of separate, but interdependent, multiple regression equations **simultaneously** by specifying the **structural model** used by statistical program.

A full structural equation model contains both measurement models and structural models. Although the measurement model part is covered in the previous lab as confirmatory factor analysis, let's still start with basic components and path analysis in the SEM.

1.Basic Structural Model of observed variables / Path analysis

Path analysis is the oldest member of the SEM family, but it is not obsolete. About 25% of articles reviewed by MacCallum and Austin (2000) concern path models, so path analysis is still widely used. There are also times when there is just a single observed measure of each construct, and path analysis is a single-indicator technique (Kline 2015). A path model contains only observed variable. It has no latent variables except for latent residuals. It is assumed that all observed variables are measured without measurement error.In this section, I will introduce the basic components of path analysis with fomo.csv.

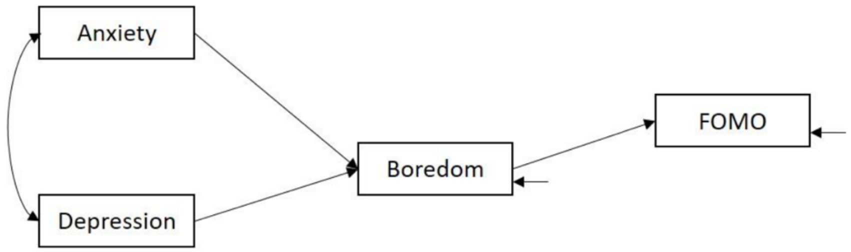
Data introduction and variables

The dataset *fomo.csv* was artificially generated based on the research scenario and descriptive statistics in Holte and Ferraro (2020). It has 250 observations. The variables are shown in the table below:

AnxAtt	Anxiety attachment, sum of 15 items (1-6)
Boredom	Boredom proneness,sum of 8 items (1-7)
Anxiety	Anxiety, sum of 7 items (0-3)
Dep	Depression, sum of 7 items (0-3)
FOMO	Fear of Missing Out, sum of 10 items (1-5)

1.1 Regression equivalency

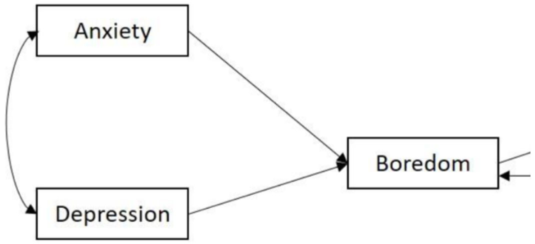
Structural models are basically linear regression models. The purpose of the following example is only to prove that the **regression coefficients estimated in linear regression models and path estimates in structural models are exactly the same.**The following diagram presents a hypothetical model between variables. We will only be looking at the left part to prove our point.



Research question:
How do anxiety and depression predict boredom proneness?

Method 1: linear regression
$Y_{boredom} = B_0 + B_1X_{Anxiety} + B_2X_{Depression} + e$

Method2: path analysis



Let's look at R codes and the output for both methods.

```
#linear regression
lm1<- lm(Boredom~ Anxiety + Dep, data=fomo)
summary(lm1)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.9558     0.9696  18.519  < 2e-16 ***
## Anxiety       0.3412     0.1003   3.403 0.000778 ***
## Dep           0.5582     0.1055   5.293 2.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.311 on 247 degrees of freedom
## Multiple R-squared:  0.4439, Adjusted R-squared:  0.4394
## F-statistic: 98.58 on 2 and 247 DF,  p-value: < 2.2e-16
#path analysis
#model specification
```

```
m1<-'
Boredom~ Anxiety + Dep
Anxiety ~~ Dep
Anxiety ~~ Anxiety
Dep ~~ Dep
'
#model fit
fit1<- sem(m1,data=fomo)
summary(fit1, fit.measures=T, standardized=T)
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Boredom ~
##     Anxiety           0.341    0.100    3.423    0.001    0.341    0.274
##     Dep               0.558    0.105    5.325    0.000    0.558    0.426
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Anxiety ~~
##     Dep              34.908    3.515    9.932    0.000   34.908    0.807
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##     Anxiety          45.474    4.067   11.180    0.000   45.474    1.000
##     Dep              41.122    3.678   11.180    0.000   41.122    1.000
##    .Boredom          39.350    3.520   11.180    0.000   39.350    0.556
##R-Square:
##                   Estimate
##    Boredom           0.444
```

Compare the regression coefficients and path estimate, and you may find that they are the same in the two models, even though in multiple linear models, the coefficients are estimated using OLS method, while in SEM, they are estimated using maximum likelihood method. The path coefficients in SEM are interpreted just as regression coefficients, no matter whether they are standardized. The interpretation of R square is also the same as in linear regression: the ratio of explained variance to the total observed variance in the endogenous variable.

Why choose path analysis over multiple regression?

- Path analysis enables us to analyze models that are more complex (and realistic) than multiple regression. For example, it can examine situations in which there are several dependent variables and those in which there are “chains” of influence, in that variable A influences variable B, which in turn affects variable C.
- Path analysis enables us to compare different models to determine which one best fits the data.

1.2Exogenous and endogenous variables

Exogenous and endogenous variables in SEM are like independent and dependent variables respectively in a regression model, but it is not entirely the same.
In short, if a variable is only a predictor and is not the outcome variable for any other variable, then it is an exogenous variable. In the model diagram, no single-headed arrows point to exogenous variables; endogenous variables can be used to predict other endogenous variables. The detailed explanations are as follows.

- Endogenous variables: **all variables with a single-headed arrow pointing in are endogenous variables.** The word “endogenous” implies “of internal origin, i.e. the presumed causes of endogenous variables are explicitly represented in the model.

  Therefore, **they cannot freely vary or covary**. In other words, in model diagrams, (the symbol that represents covariances or correlations between independent variables)

  does not directly connect two different endogenous variables, and ↺ (the symbol that represents the variance of an exogenous variable) will not originate from and end with any endogenous variables. **Endogenous variables can predict other variables in the model too.** In model diagrams, **they can have arrows pointing in and pointing out.**

  Every endogenous variable has a disturbance/residual/error term that points at them. They are like the error term in regression analysis. The computer can estimate the variance of the error term, thus enabling us to get R square (the proportion of explained variance).

- Exogenous variables: in contrast**, the causes of exogenous variables are not included in the model**. Exogenous variables are like independent variables in regression. “Exogenous” implies "of external origin”. In model diagram, **no single-headed arrows point to exogenous variables**. In RAM symbolism (McArdle–McDonald reticular action model), ↔ (**the symbol that represents covariances or correlations between independent variables) will connect every pair of observed exogenous variables,** and

  (the symbol that represents the variance of an exogenous variable) will connect every observed or latent exogenous variable to itself. Although the correlation between two exogenous variables are estimated by the computer, no prediction is put forward about why these two variables covary. In this sense, **the association is unanalyzed**. In other symbolism, the ↔ symbol may be omitted.

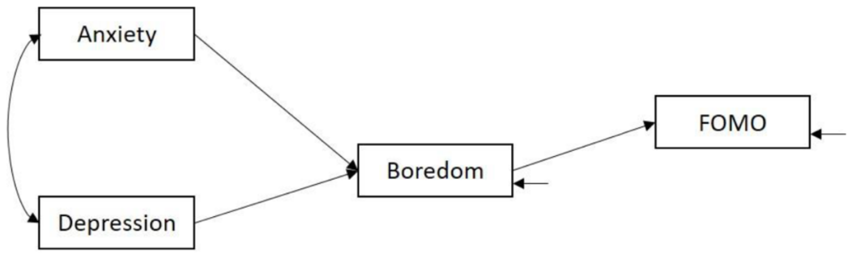
1.3 Model diagram

▭	Observed variable
⬭	Latent variable
→	Structural path or factor loading
↔	Covariance/correlation/variance

How to find all the relationships in this diagram? Just follow the following steps:

- First, find the endogenous variables. Endogenous variables have single-headed arrows pointing in.
- Second, find the corresponding predictors for each endogenous variable. They are the variables from which the arrows pointing at the endogenous variable. In other words, the origins of the directional arrows are predictors.
- Third, find the covariance relationship by looking for the ↔ symbol.
- Fourth, find the error term for each endogenous variable. The ⬭ that represents a latent error variable can be replaced with omitted with only arrows showing ←

Let's implement the method for the following path model.



First, we can see that Boredom and FOMO have arrows pointing in. They are endogenous variables.

Second, Anxiety and Depression have arrows pointing into Boredom. Hence, Anxiety and Depression are two exogenous variables for Boredom. Boredom has an arrow pointing into

FOMO. Hence, Boredom is the predictor for FOMO. However, Boredom is still counted as endogenous variable since it has two predictors.

Third, Anxiety and Depression covary with each other. ↔ for Anxiety and Depression are omitted in the diagram.It is common in publications to omit variance symbol for exogenous variables.

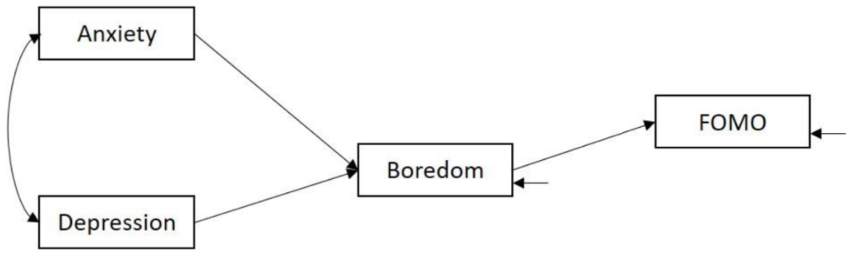
Lastly, there are two latent error variables in this model. The oval symbol is replaced with arrow in the presented diagram which is commonly used.

Note: Do not think that disturbances in path models are equivalent to residuals in multiple regression. To do so would be to confuse a causal model (path analysis) with a statistical model (regression analysis). Regression residuals are artifacts of least squares estimation such that those residuals, by definition, are uncorrelated with the predictors. But disturbances are not analysis artifacts; instead, they are determined by physical reality, including social or genetic factors that affect the corresponding endogenous variable but are unmeasured (Pearl, 2012).

1.4 Model specification

Unlike EFA as an exploratory method, covariance-based SEM is a confirmatory method. Because the literature for newer research areas can be limited, decisions about what to include in the model must sometimes be guided more by the researcher’s expertise than by published reports. Consulting with experts in the field about possible specifications may also help. In this part, we will not discuss variable selection but model equations and model identification based on theoritical research assumptions.

Research question: How do proposed variables affect fear of missing out?



For simplicity, lets set:
X1=Anxiety
X2=Depression
Y1=Boredom
Y2=Fomo

Matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & 0 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix} \Psi = \begin{bmatrix} \psi_{11} & \\ 0 & \psi_{22} \end{bmatrix}$$

The structural equations in matrix form (Lisrel notation) is written with **B**(Beta), **Γ**(Gamma), **Φ**(Phi), and **Ψ**(Psi). **B**(Beta) shows the structural relationship between Ys; **Γ**(Gamma) shows the structural relationship between X and Y; $\zeta(Zeta)$ is the residual for Y; **Φ**(Phi) matrix shows the variance and covariance of X; **Ψ**(Psi) matrix shows the variance and covariance of Y.

The **B**(Beta) matrix is 4 by 4 because we have 2 endogenous variables. The **Γ**(Gamma) matrix is 2 by 2 because we have 2 exogenous variables and 2 endogenous variables. Columns represent X variables and rows represent Y variable. The general rule for the matrix equation is that the subcripts of the parameters go from the column to the row. For example, based on the path model presented, there is a relationship between X2 and Y1. The Gamma parameter is written as γ_{12} meaning the relationship is from X2 to Y1. There are 2 $\zeta s(Zeta)$ due to the existance of 2 Ys. The **Φ**(Phi) has 3 elements with variance parameters for X1 and X2, and covariance between X1 and X2. **Ψ**(Psi) has 2 elements with variance parameters for Y1 and Y2.

1.5 Model identification

A model being “identified” means that it is possible to uniquely solve for the parameters. If we cannot uniquely find a set of parameter values, it then becomes impossible to claim to have found “the values” of the parameters. Identification is a mathematical property of the model depending on the structure of the hypothesized model. It is determined prior to having any actual data. An unknown parameter is identified if that parameter can be written as a function of one or more elements in covariance matrix. If all unknown parameters are identified then the model is identified.

- A **just-identified** structural equation model is identified and has the same number of observations as free parameters ($df_M = 0$).
- An **overidentified** structural equation model is identified and has more observations than free parameters ($df_M > 0$).
- An **underidentified** structural equation model is one for which it is not possible to uniquely estimate all of its free parameters. **(Identification problem)**

<u>Rule of thumbs for model identification</u>

t-rule(necessary condition)

- A necessary but not sufficient condition
- t = number of independent parameters estimated
  p = number of endogenous variables (Y)
  q = number of exogenous variables (X)

- t rule for identification is that the number of non-redundant (or unique) elements in covariance matrix of the observed variables must be greater than or equal to the number of parameters.

$$t \leq \frac{(p+q)(p+q+1)}{2}$$

For our model, p=2,q=2. There should be (2+2)(2+2+1)/2= 10 unique parameters for given variables. By unique parameters, I mean each variable is allowed to covary with other variables. The parameter yielded from the proposed model is 8. Therefore, we get degrees of freedom of 10-8=2. T-rule is satisfied.

Null B rule(sufficient condition)

- A sufficient condition
- Any path model with B = 0 is identified
- When there is no directional path between endogenous variables

** This is a sufficient condition. There are identified models that do not meet null beta rule
For our model, we do have B_{21}, the null B rule is not satisfied. But null B rule is a sufficient condition, we do not need to worry about it here.

Recursive rule(sufficient condition)

- A sufficient condition
- If B matrix can be written as a lower triangular matrix, and Ψ is diagonal, the model is identified

** This is a sufficient condition.There are identified models that do not meet recursive rule
For our model, there is no parameter in the upper trangular matrix for B matrix and Ψ is diagonal. Therefore, recursive rule is satisfied.

Order rule(necessary condition) – apply to nonrecursive models

- A necessary condition
- Identification can be checked for each equation following order rule
- Order rule is that at least (p – 1) variables must be excluded from the equation to make the parameters in that equation identified (where p = number of endogenous variables)
- Order rule assumes that there is no restriction in Ψ matrix (i.e., you can use the order rule only when no restriction is imposed in Ψ matrix)

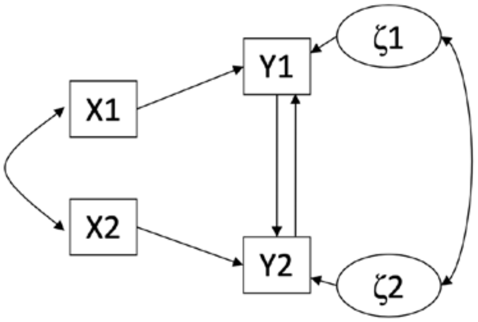
For our model, this rule is not applicable.

Recursive models

- Recursive models are systems of equations that contain no reciprocal causation or feedback loops
- For recursive models,
  - B matrix can always be written as a lower triangular matrix
  - Ψ is a diagonal matrix (i.e. no residual covariance between residual terms of endogenous variables)

Non-recursive models

- Nonrecursive models contain reciprocal causation, feedback loops, (or correlated residuals)
- For nonrecursive models,
  - B matrix canNOT be written as a lower triangular matrix, or
  - Ψ is not a diagonal matrix



1.6 Assumption checking

Multivariate normality

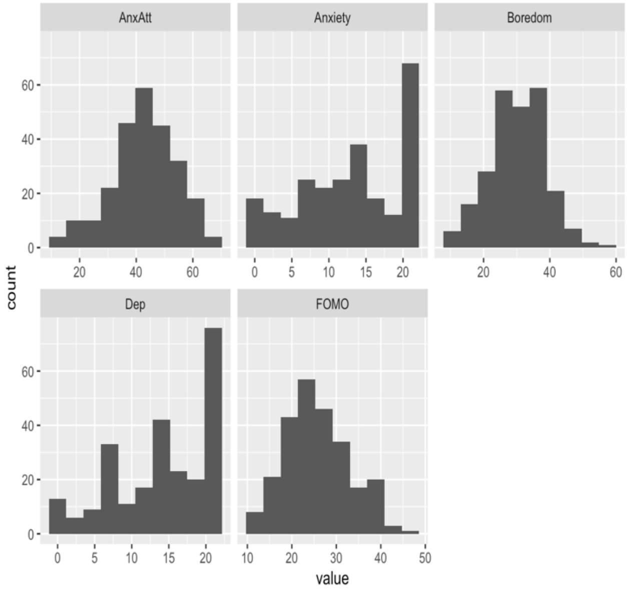
The default estimation method in SEM, maximum likelihood, assumes multivariate normality (multinormality) for continuous outcome variables. This means that

1. all the individual univariate distributions are normal;
2. all joint distributions of any pair of variables is bivariate normal; that is, each variable is normally distributed for each value of every other variable; and
3. all bivariate scatterplots are linear with homoscedastic residuals.

Most of the time, multivariate nonnormality can be detected through inspection of univariate distributions. For variables that voilated normality assumption, we can transform the variables as what we did in GLM class.

In R, we can inspect the univariate distribution through histogram. (option1)

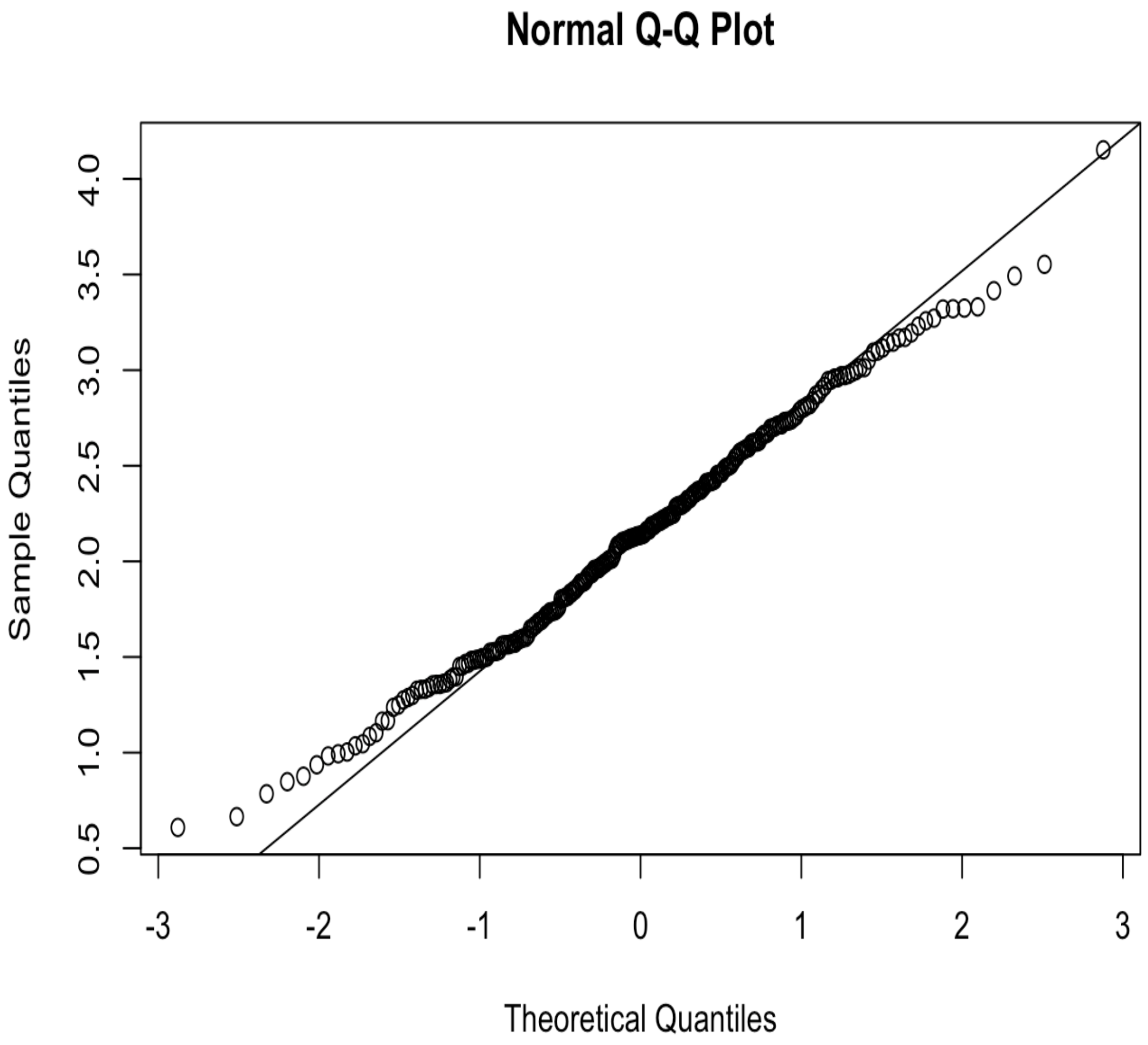
```
# assumption1 multivariate normality with histogram
ggplot(gather(fomo), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')
```



Interpretation: From the output, we can see that variable Anxiety and Depression are not normally distributed. We can consider transformations for Anxiety and Depression refering to previous labs covered in GLM classes.

Or we can conduct Mardia test of multivariate normality. It checks whether the multivariate skewness and kurtosis are consistent with a multivariate normal distribution. The null hypothesis is that the sample comes from a multivariate normal distribution. The alternative is that the sample does not come from a multivariate normal distribution.

```
mardia(fomo)
```



For multivariate normality, both p-values of skewness and kurtosis statistics should be greater than **0.05**.

```
## Call: mardia(x = fomo)
##
## Mardia tests of multivariate skew and kurtosis
## Use describe(x) the to get univariate tests
## n.obs = 250   num.vars =  5
## b1p =  1.38   skew =  57.46  with probability =  0.0097
##  small sample skew =  58.38  with probability =  0.0079
## b2p =  32.5   kurtosis =  -2.36  with probability =  0.018
```

Mardia's multivariate skewness statistic

Chi-square value of skewness statistic

Mardia's multivariate kurtosis statistic

Sample size

When the multivariate normality is satisfied for all endogenous variables, the required sample

size can be smaller. However, this is often not the case.
A rule of thumb when ML (maximum likelihood) method is used for estimation is:
The ratio of sample size (N) to # parameters (q) should be larger than 5:1 or 10:1. The ideal scenario is N:q>20:1, but most analysts now think that is unrealistically high. N:q>5:1 is satisfied in most published studies. For example, if we want to estimate 10 parameters, we need at least 50 samples.
In absolute terms, a "typical" sample size for studies that use SEM is 200. This is approximately the median sample size for published papers. But again, we should also consider the level of complexity of our model.
If the sample size is too small, statistical power will be low. Another way to determine how many samples we need is to conduct a power analysis. For SEM, this website can be used to determine sample size when the desired power is specified:
http://www.quantpsy.org/rmsea/rmsea.htm

```
#assumption2 sample size

str(fomo)
```

```
## 'data.frame':    250 obs. of  5 variables:
##  $ AnxAtt : int  34 60 54 49 36 48 15 19 42 42 ...
##  $ Boredom: int  26 38 31 39 30 32 8 13 24 22 ...
##  $ Anxiety: int  13 18 12 20 7 17 3 4 15 21 ...
##  $ Dep    : int  3 21 14 21 6 21 2 6 21 15 ...
##  $ FOMO   : int  24 33 26 28 23 28 11 16 19 32 ...
```

Interpretation: The data contain 250 observations which is 5 times the variable numbers. Therefore, the assumption of sample size is met.

Outliers

A multivariate outlier has extreme scores on two or more variables, or a pattern of scores that is atypical. For example, a case may have scores between two and three standard deviations above the mean on all variables. Although no individual score might be considered extreme, the case could be a multivariate outlier if this pattern is unusual. The outliers can be examine by calculating for each case its squared Mahalanobis distance(D_M^2), which indicates the distance in variance units between the profile of scores for that case and the vector of sample means, or centroid, correcting for intercorrelations.
In large samples with normal distributions, $D_M^2$is distributed as central chi-square with degrees of freedom equal to the number of variables, or v. A relatively high D_M^2 with a low p value in the corresponding chi-square distribution may lead to the rejection of the null hypothesis that the case comes from the same population as the rest. A conservative level of sta- tistical significance is usually recommended for this test, such as .001.

```
#Typically a p-value that is less than .001 is considered to be an outlier ba
sed on Mahalanobis distance

fomo$mahal<-mahalanobis(fomo, colMeans(fomo), cov(fomo))
```

```
fomo$p <- pchisq(fomo$mahal, df=4, lower.tail=FALSE)
fomo$p[fomo$p<.001]
## numeric(0)
which(fomo$p<.001)
## integer(0)
```

Interpretation: As we can see from the result, we have no outlier.

Missing Data

Variables in study should be complete in data forms. Simply there is no missing data in any variable. In the real world, missing values occur in many, if not most, data sets, despite the best efforts at prevention. Missing data occur for many reasons, including hardware failure, missed appointments, and item non-response. A few missing values, such as < 5% in the total data set, may be of little concern. Recall what we have learned in GLM class, missing data can be **missing completely at random (MCAR), missing at random (MAR)** or **missing not at random (MNAR)**. For MCAR and MAR, available case methods and imputations can be used.

```
#assumption4 missing data
summary(is.na(fomo))
##    AnxAtt         Boredom         Anxiety           Dep
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:250       FALSE:250       FALSE:250       FALSE:250
##     FOMO           mahal             p
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:250       FALSE:250       FALSE:250
```

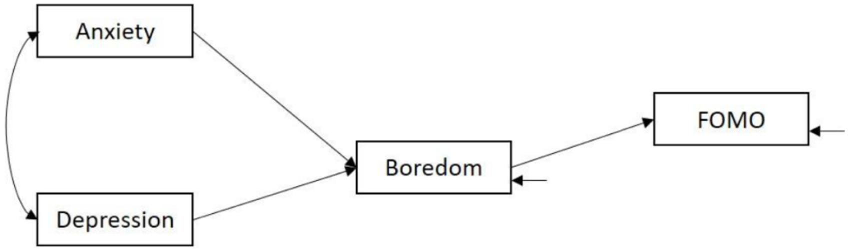
Interpretation: There is no missing data in fomo dataset. We shall proceed with our analysis.

1.7 Model building in R

To conduct path analysis and full SEM, we will use “lavaan” and “semPlot” packages. 2 steps are required in the analysis. We need to specify the model with single quotation marks followed by fitting the model with sem() function. In reality, we can also use cfa() function in R. These two functions will yield identical results. They share the same default model specifications. So it seems that the distinction is currently not all that useful, but that the two commands may eventually possess different functionality. Perhaps the package developer simply wanted to support this future possibility, early on.

```
library(lavaan)
library(semPlot)
```

```
#set work directory
setwd("/your working directory")
# read data
fomo<- read.csv("fomo.csv")
#check structure
str(fomo)
## 'data.frame':    250 obs. of  5 variables:
##  $ AnxAtt : int  34 60 54 49 36 48 15 19 42 42 ...
##  $ Boredom: int  26 38 31 39 30 32 8 13 24 22 ...
##  $ Anxiety: int  13 18 12 20 7 17 3 4 15 21 ...
##  $ Dep    : int  3 21 14 21 6 21 2 6 21 15 ...
##  $ FOMO   : int  24 33 26 28 23 28 11 16 19 32 ...
names(fomo)
## [1] "AnxAtt"  "Boredom" "Anxiety" "Dep"     "FOMO"
```



To define manifest endogenous variable, we use ~ (tilda). For example, anxiety and depression are 2 predictors for boredom. Hence we use ~ to represent the relationship. We use ~~ to represent the variance/covariance relationship.

In the model, we have:

2 structural models:

- Boredom ~ Anxiety + Dep
- FOMO ~ Boredom

1 covariance between exogenous variables

- Anxiety ~~ Dep

2 residuals for endogenous variabls

- Boredom ~~ Boredom
- FOMO ~~ FOMO

2 variance for exogenous variables

    - Anxiety ~~ Anxiety
    - Dep ~~ Dep

As we see in the following codes, the model is specified in the single quotation mark. We did not specify the variances for exogenous variables and endogenous variables. In lavaan, variances for each variable is specified by default. If you are concerned, you can also specify all the relationships as shown.

```
#model specification
#model specification
m2<- '
Boredom ~ Anxiety + Dep
FOMO ~ Boredom
Anxiety ~~ Dep
'
#model fit
fit2<- sem(m2, data=fomo) #fit the model

#Alternatively, you can write the model as:
m2<- '
Boredom ~ Anxiety + Dep
FOMO ~ Boredom
Anxiety ~~ Dep
Anxiety ~~ Anxiety
Dep ~~ Dep
Boredom ~~ Boredom
FOMO ~~ FOMO
'
```

To ensure that we have correctly specify the model, we can use inspect() to check model specification. It allows us to examine the number of estimated parameters in the model.

```
inspect(fit2) #inspect model specification
## $lambda
##          Boredm FOMO Anxity Dep
## Boredom       0    0      0   0
## FOMO          0    0      0   0
## Anxiety       0    0      0   0
## Dep           0    0      0   0
##
## $theta
##          Boredm FOMO Anxity Dep
## Boredom 0
## FOMO    0      0
## Anxiety 0      0    0
## Dep     0      0    0      0
##
## $psi
##          Boredm FOMO Anxity Dep
## Boredom 5
## FOMO    0      6
## Anxiety 0      0    7
## Dep     0      0    4      8
##
```

$psi presents the phi and psi parameters

```
## $beta
##          Boredm FOMO Anxity Dep
## Boredom       0    0      1   2
## FOMO          3    0      0   0
## Anxiety       0    0      0   0
## Dep           0    0      0   0
```

$beta presents the beta and gamma parameters

1.8 Parameter Interpretation

In this part, let's look at how to interpret estimated parameter before model fit evaluation. Use the summary() function, we are able to inspect model parameters. `fit.measures=TRUE` allows us to look at fit indices, `standardized=TRUE` allows us to look at standardized parameters. `rsquare=TRUE` allows us to examine variances accounted in the endogenous variables.

```
summary(fit2, fit.measures=TRUE,standardized=TRUE,rsquare=TRUE) #examine fit
indices
## lavaan 0.6-7 ended normally after 35 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
```

Recall from the model specification part, we have 8 parameters

```
##   Number of free parameters                          8
##
##   Number of observations                           250
##
## Model Test User Model:
##
##   Test statistic                                64.158
##   Degrees of freedom                                 2
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                               591.315
##   Degrees of freedom                                 6
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.894
##   Tucker-Lewis Index (TLI)                       0.681
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -3126.825
##   Loglikelihood unrestricted model (H1)      -3094.746
##
##   Akaike (AIC)                                6269.650
##   Bayesian (BIC)                              6297.821
##   Sample-size adjusted Bayesian (BIC)         6272.461
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.353
##   90 Percent confidence interval - lower         0.281
```

Df=10-8=2

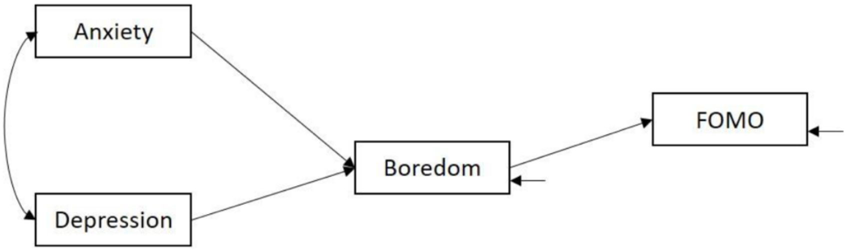
Estimated parameters

```
##   90 Percent confidence interval - upper         0.429
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.114
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Boredom ~
##     Anxiety            0.341    0.100    3.423    0.001    0.341    0.274
##     Dep                0.558    0.105    5.325    0.000    0.558    0.426
##   FOMO ~
##     Boredom            0.531    0.044   12.203    0.000    0.531    0.611
##
## Covariances:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Anxiety ~~
##     Dep               34.908    3.515    9.932    0.000   34.908    0.807
##
## Variances:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##    .Boredom           39.350    3.520   11.180    0.000   39.350    0.556
##    .FOMO              33.498    2.996   11.180    0.000   33.498    0.627
##     Anxiety           45.474    4.067   11.180    0.000   45.474    1.000
##     Dep               41.122    3.678   11.180    0.000   41.122    1.000
##
## R-Square:
```

```
##                     Estimate
##     Boredom            0.444
##     FOMO               0.373
```

By default, the parameters are estimated using the maximum likelihood method.

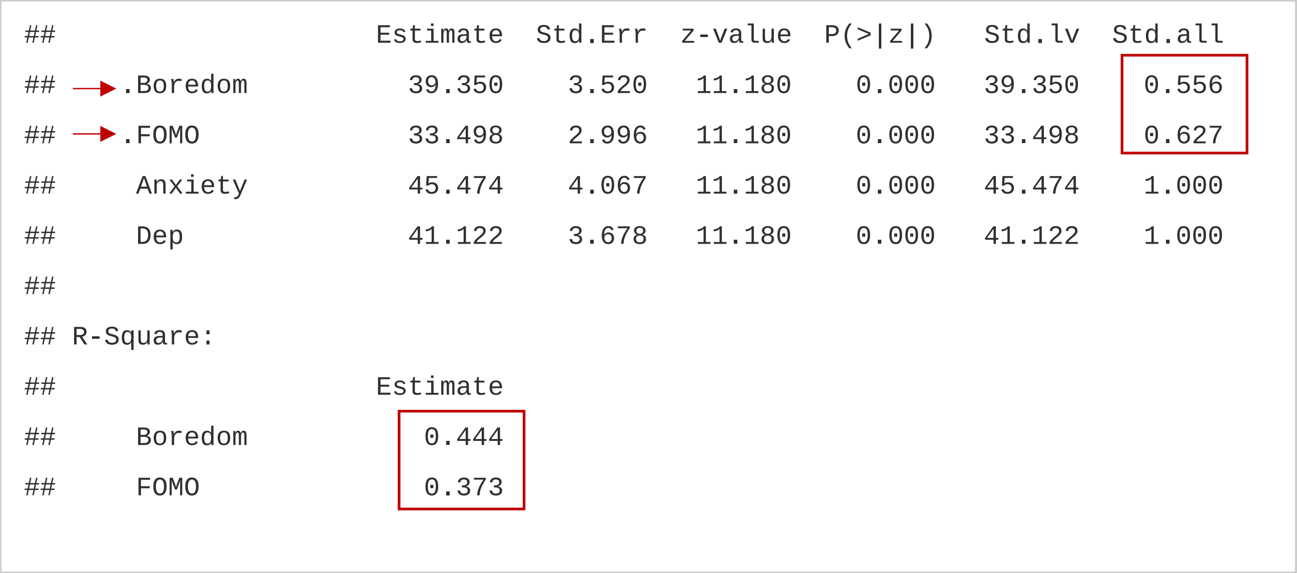
- The “Estimate” column gives the unstandardized regression coefficient.
- The “Std. Err” column gives the standard error of estimate.
- “z-value” is the estimate divided by standard error.
- The “z-value” gives rise to the p-value. Here for α=0.05, the estimates are all statistically significant. In other words, all the coefficients are statistically significant from 0. All the paths are significant.
- “Std.lv” means only the latent variables are standardized. Since we don’t have latent variables here, the numbers in this column are the same as the “Estimate” column.
- “Std.all” means both latent and observed variables are standardized. It is often called the “completely standardized solution”. The results in this column are what we want in most cases.



```
Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   Boredom ~
##     Anxiety            0.341    0.100    3.423    0.001    0.341    0.274
##     Dep                0.558    0.105    5.325    0.000    0.558    0.426
##   FOMO ~
##     Boredom            0.531    0.044   12.203    0.000    0.531    0.611
```

Interpretation:
We can interpret the path coefficient as what we did in multiple regression. For example, in the model, there is a direct effect of Anxiety on Boredom. P-value is .001 less than .05. We reject the null hypothesis and conclude there is a significant relationship between Anxiety and Boredom. One unit increase on the Anxiety scale will yield .341 unit increase on the Boredom scale. We can also use the standardized value. One standard deviation increase in Anxiety will yield .274 standard deviation increase in Boredom.

```
Variances:
```



```
##                        Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##  →.Boredom               39.350    3.520   11.180    0.000   39.350    0.556
##  →.FOMO                  33.498    2.996   11.180    0.000   33.498    0.627
##     Anxiety              45.474    4.067   11.180    0.000   45.474    1.000
##     Dep                  41.122    3.678   11.180    0.000   41.122    1.000
##
## R-Square:
##                        Estimate
##     Boredom               0.444
##     FOMO                  0.373
```

The variance output shows the estimated variances for each exogenous variable and estimated residual variances for each endogenous variable. Lavaan automatically makes the distinction between variances and residual variances. Lavaan indicates the endogenous variables by adding a dot in front of the variable. Same as the regression part, the Estimate column shows the unstandardized values and Std.all column shows the standardized values.

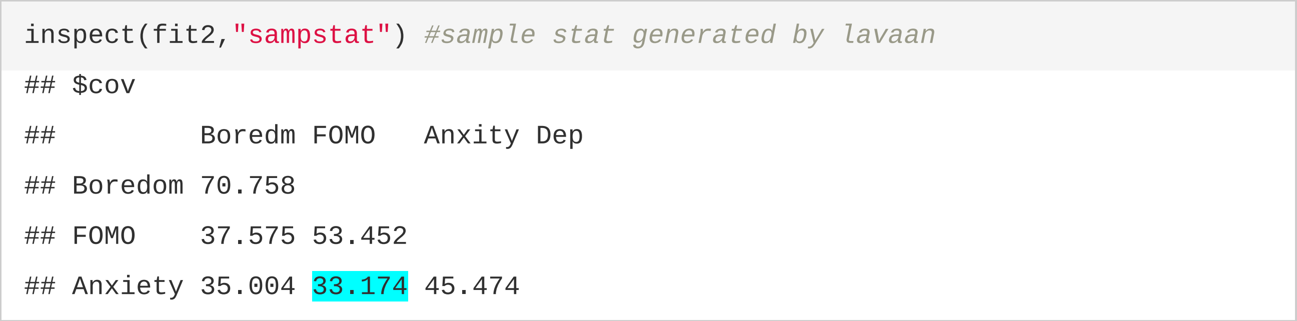
The R-Square output shows the variance in each endogenous variable explained by exogenous variables in the path model. For example, 44.4% variances in Boredom are explained by Anxiety and Depression.

*We may notice that the r-square value and standardized residual variance value add up to 1 for endogenous variables.

1.9 Evaluating Model fit

The idea of covariance-based SEM is to estimate parameters so that the discrepency between the sample covariance matrix and the implied covariance matrix is minimal. Model fit evaluation is to assess how close (or how far) the implied covariance matrix is to (or from) the sample covariance matrix.

We can look at the sample covariance matrix and implied covariance matrix. Two matrices look close to each other with only different two values. Let’s next look at the model fit.



```
inspect(fit2,"sampstat") #sample stat generated by lavaan
## $cov
##         Boredm FOMO   Anxity Dep
## Boredom 70.758
## FOMO    37.575 53.452
## Anxiety 35.004 33.174 45.474
## Dep     34.867 28.250 34.908 41.122
fitted(fit2) #fitted stats from our model
## $cov
##         Boredm FOMO   Anxity Dep
## Boredom 70.758
## FOMO    37.575 53.452
## Anxiety 35.004 18.588 45.474
## Dep     34.867 18.516 34.908 41.122
```

Model fit is evaluated through 2 scopes:

1. Test of overall model fit (likelood ratio test)
2. Fit indices

* Something to remember here:

1. Just identified (or saturated) models always have the perfect fit by its nature
2. Model fit evaluation is meaningful for over-identified models

Test of overall model fit

Chi-square test of exact fit

It is a likelihood ratio test of a hypothesized model (over-identified) to the saturated (perfectly fitting) model. Recall from what we learn from logistic regression, the -2 log likelihood ratio value asymptotically follows a chi-square distribution, here, with (p(p+1)/2 – number of free parameters) degrees of freedom.

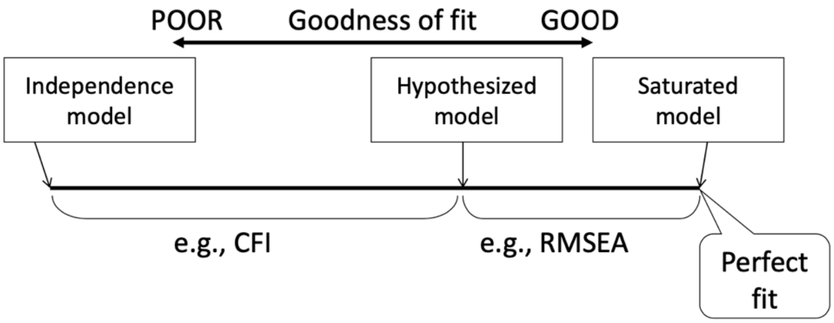
$$H0 :\Sigma=\Sigma(\theta)$$

(population covariance matrix = covariance matrix expressed in terms of model parameters θ)

If H0 is rejected, it means that the lack of fit is significant, i.e., the hypothesized model fits significantly worse than the saturated model. Therefore, we want chi-square to be not significant. But remember that the significance of chi-square statistic is sensitive to sample size. For models with about 75 to 200 cases, the chi square test is generally a reasonable measure of fit. But for models with more cases (400 or more), the chi square is almost always statistically significant. Chi-square is also affected by the size of the correlations in the model: the larger the correlations, the poorer the fit.

Fit indices

Since chi-square is sensitive to sample size and size of correlations, alternative measures of fit have been developed. Fit indices aim to gauge the "closeness" or "discrepancy" of implied covariance matrix to sample covariance matrix. The most commonly used ones are CFI, RMSEA, and SRMR. But we will also cover other frequently reported fit indices.
Here is a graph visualizing the goodness of fit I borrowed from SEM class taught by Dr.Ehri.Ryu.



- **Saturated model**: all measured variables are allowed to covary with one another (i.e., all the elements in the covariance matrix are freely estimated)
- **Independence model** (Null model or Baseline model): all measured variables are mutually uncorrelated (i.e., no covariance between variables)

Goodness of Fit Index(GFI)

GFI is the proportion of variance accounted for by the estimated population covariance. Analogous to R2. A lot of research shows that this index is positively biased (which means a lot of times GFI indicates the model fits the data well, when in fact it doesn't). We want it to be high.

Comparative fit index (CFI)

CFI (Bentler, 1990) provides a measure of goodness of fit of the hypothesized model compared to an independence model. It is defined as:

$$CFI = 1 - \frac{\Delta_1}{\Delta_2} \qquad \Delta_1 = max\left[\left(\chi^2_H - df_H\right), 0\right]$$

$$\Delta_2 = max\left[\left(\chi^2_I - df_I\right), \Delta_1\right]$$

CFI ranges from 0 to 1. Larger value (closer to 1) indicates good fit.

Bentler-Bonett Index or Normed Fit index(NFI)

This is the very first measure of fit proposed in the literature (Bentler & Bonett, 1980) and it is an incremental measure of fit.It is defined as:

$$\frac{x^2(I) - x^2(H)}{x^2(I)}$$

Tucker-Lewis Index(TLI) OR Non-normed fit index(NNFI)

The Tucker-Lewis index (also called the non-normed fit index or NNFI), another incremental fit index, does have such a penalty. Let χ2/df be the ratio of chi square to its degrees of freedom, and the TLI is computed as follows:

$$\frac{\frac{x^2}{df(I)} - \frac{x^2}{df(H)}}{\frac{x^2}{df(I)} - 1}$$

If the index is greater than one, it is set at one. Note that for a given model, a lower chi square to df ratio (as long as it is not less than one) implies a better fitting model. Its penalty for complexity is χ2/df. That is, if the chi square to df ratio does not change, the TLI does not change.

Standardized root mean square residual (SRMR)

SRMR is the standardized version of RMR. RMR is the average of squared value of each element in residual covariance matrix. Since SRMR is standardized, SRMR is the average of squared residual in residual correlation matrix.

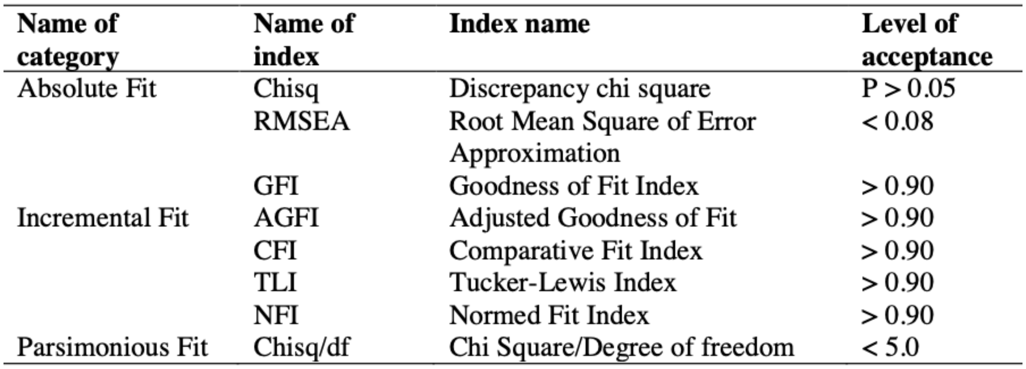
Root mean square error of approximation (RMSEA)

RMSEA (Browne & Cudeck, 1993; Steiger, 1990) provides a measure of discrepancy between the covariance matrix (Σ) and the implied covariance matrix (Σ(θ)) in the population. It is defined as:

$$\text{RMSEA} = \sqrt{\max\left[\left(\frac{\chi^2 - \text{df}}{\text{df}(N-1)}\right), 0\right]}$$

RMSEA attempts to gauge the error of approximation in the population (i.e., lack of model fit) apart from the sampling error.RMSEA includes an adjustment for parsimony of the model by penalizing for number of estimated parameters. In sum, RMSEA is a measure of lack of fit per degree of freedom in the population.

Rule of thumb criterion:



Name of category	Name of index	Index name	Level of acceptance
Absolute Fit	Chisq	Discrepancy chi square	P > 0.05
	RMSEA	Root Mean Square of Error Approximation	< 0.08
	GFI	Goodness of Fit Index	> 0.90
Incremental Fit	AGFI	Adjusted Goodness of Fit	> 0.90
	CFI	Comparative Fit Index	> 0.90
	TLI	Tucker-Lewis Index	> 0.90
	NFI	Normed Fit Index	> 0.90
Parsimonious Fit	Chisq/df	Chi Square/Degree of freedom	< 5.0

Ahmad 2016

Other fit indices: http://davidakenny.net/cm/fit.htm

* Note: Fit indices do not necessarily agree with each other.CFI is influenced by the degree of associations of variables (all things being equal). For example, if correlations are high, that makes the baseline model look worse, which will result in higher CFI value. SRMR is affected by sampling error (to a greater degree than RMSEA).

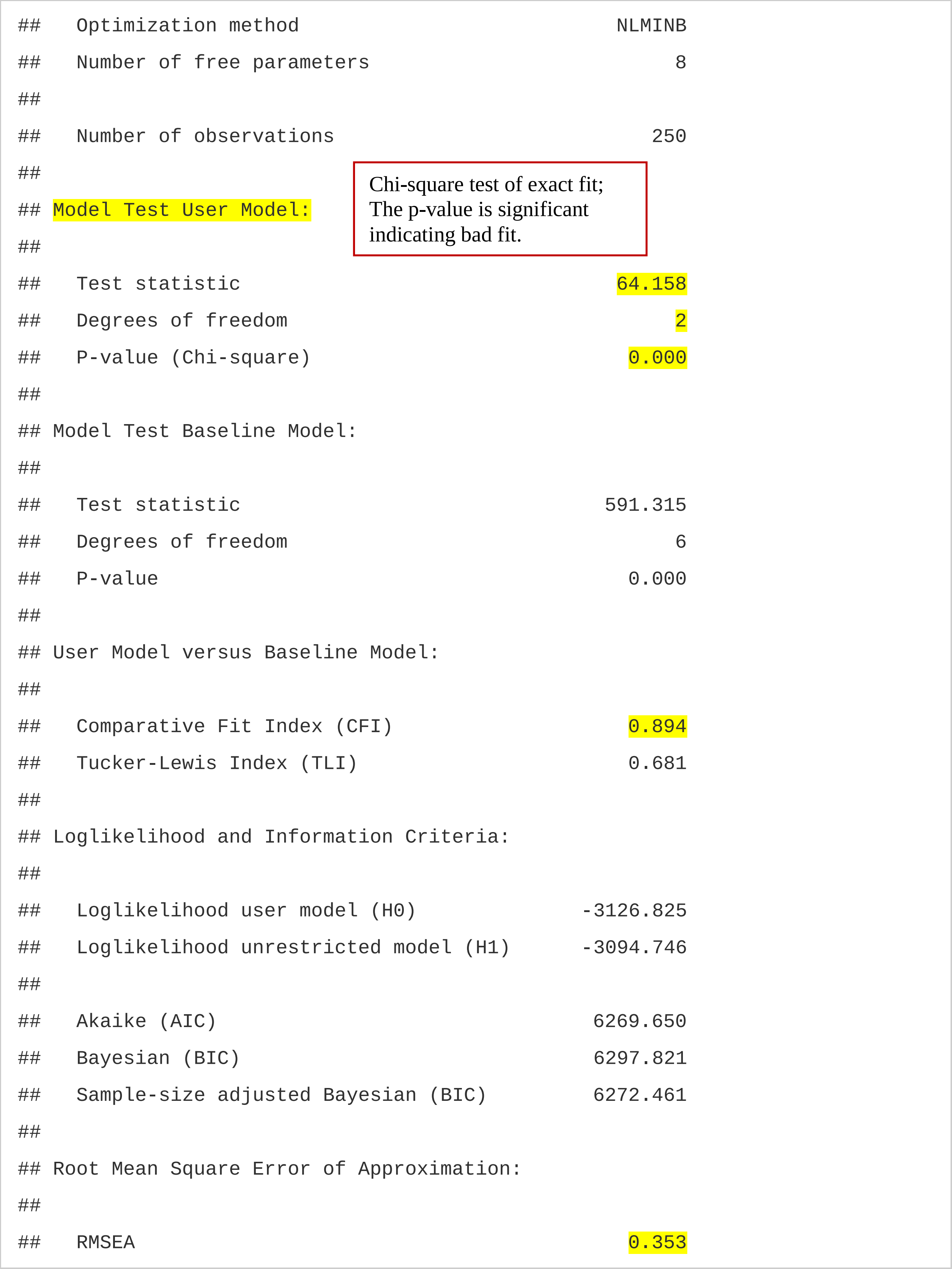
Now, let’s look at our model fit in r output!

```
summary(fit2, fit.measures=TRUE,standardized=TRUE,rsquare=TRUE) #examine fit
indices
```

```
## lavaan 0.6-7 ended normally after 35 iterations

##

##   Estimator                                         ML
```



```
##   Optimization method                           NLMINB
##   Number of free parameters                          8
##
##   Number of observations                           250
##
## Model Test User Model:
##
##   Test statistic                                64.158
##   Degrees of freedom                                 2
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                               591.315
##   Degrees of freedom                                 6
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.894
##   Tucker-Lewis Index (TLI)                       0.681
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -3126.825
##   Loglikelihood unrestricted model (H1)      -3094.746
##
##   Akaike (AIC)                                6269.650
##   Bayesian (BIC)                              6297.821
##   Sample-size adjusted Bayesian (BIC)         6272.461
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.353
```

Chi-square test of exact fit;
The p-value is significant
indicating bad fit.

```
##   90 Percent confidence interval - lower         0.281
##   90 Percent confidence interval - upper         0.429
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.114
##
```

We can also use fitmeasures() for all fit indices.

```
fitmeasures(fit2)
```

```
##                npar                fmin               chisq                  df
##               8.000               0.128              64.158               2.000
##              pvalue      baseline.chisq         baseline.df     baseline.pvalue
##               0.000             591.315               6.000               0.000
##                 cfi                 tli                nnfi                 rfi
##               0.894               0.681               0.681               0.675
##                 nfi                pnfi                 ifi                 rni
##               0.892               0.297               0.895               0.894
##                logl   unrestricted.logl                 aic                 bic
##           -3126.825           -3094.746            6269.650            6297.821
##              ntotal                bic2               rmsea      rmsea.ci.lower
##             250.000            6272.461               0.353               0.281
##      rmsea.ci.upper        rmsea.pvalue                 rmr          rmr_nomean
##               0.429               0.000               5.545               5.545
##                srmr        srmr_bentler srmr_bentler_nomean                crmr
##               0.114               0.114               0.114               0.148
##         crmr_nomean          srmr_mplus   srmr_mplus_nomean               cn_05
##               0.148               0.114               0.114              24.347
##               cn_01                 gfi                agfi                pgfi
##              36.890               0.898               0.492               0.180
##                 mfi                ecvi
##               0.883               0.321
```

Interpretation: The p-value for chi-square is less than .001 which is significant. We reject the null hypothesis and conclude the model does not fit the data well. However, with large sample size, chi-square tends to be significant.

The CFI=.894 which is less than .90 indicating our proposed model does not good a better job than the null model.

The RMSEA=.353 which is greater than .05. The p-value is less than .001. Hence, we reject the null hypothesis and conclude the RMSEA value is greater than .05 at 10% significance level.

SRMR is .114 which is greater than .05. This is potentially caused by the large value in the residual matrix.

In conclusion, the model does not fit the data well.

To understand why our model does not fit the data well, we can examine the residuals. In the standardized and normalized residual, we can see the residual of covariance between fomo and anxiety is the largest, which indicates potential model improvement. In the next section, we will look at model respecification and comparison.

```
resid(fit2)
## $type
## [1] "raw"
##
## $cov
##         Boredm FOMO   Anxity Dep
## Boredom  0.000
## FOMO     0.000  0.000
## Anxiety  0.000 14.586  0.000
## Dep      0.000  9.734  0.000  0.000
resid(fit2,type="standardized")
## $type
## [1] "standardized"
##
## $cov
##         Boredm FOMO  Anxity Dep
## Boredom 0.000
## FOMO    0.000  0.000
## Anxiety 0.000  6.783 0.000
## Dep     0.000  5.140 0.000  0.000
resid(fit2,type="normalized")
## $type
## [1] "normalized"
##
## $cov
##         Boredm FOMO  Anxity Dep
## Boredom 0.000
## FOMO    0.000  0.000
## Anxiety 0.000  3.881 0.000
## Dep     0.000  2.812 0.000  0.000
```

1.10 Model respecification and comparision

Nested Models
As we can see, our model does not fit the data well. We may build a new model if we have an alternative theory about the relationship between the variables. Or we can respecify our model by adding or reducing some paths in the model. These models are called nested models. Model A is nested within Model B if Model A can be specified by adding additional constraints on Model B.

Modification indices(MI)
Modification indices (MI) are suggestions for potential model specifications (through the inclusion of additional parameters) that may result in an increase in model fit.

The MI value indicates the degree of the chi-square goodness of fit value is expected to decrease as a result of adding a suggested parameter (NOTE: smaller chi-square values are associated with increased fit). A value > 3.84 (p=.05) suggests that including a suggested parameter will result in a significant improvement in model fit. The P>MI column is the projected p-value for the increment in fit as a result of adding in the suggested parameter.

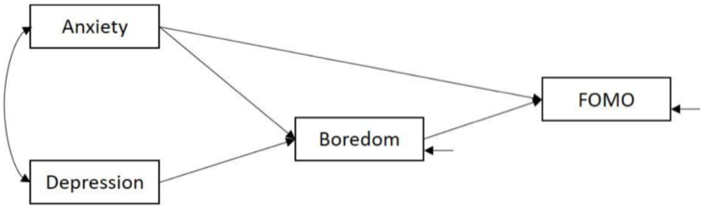
Modification indices are also sensitive to sample size (just like the chi-square statistic for test of overall model fit).Also, MI does not reflect what will happen if more than one parameter constraints are relaxed.

Obviously, we do not want to add all suggested parameters. However, we might consider identifying those MI values that are the largest (indicating the additions that are most likely to yield the greatest increments in fit).However, we should not blindly follow MI statistics to modify your model.

For our model, the MI for FOMO ~ Anxiety is the largest suggesting by adding the path relationship from Anxiety to FOMO will improve our model fit.

```
modindices(fit2)
##        lhs op      rhs     mi     epc sepc.lv sepc.all sepc.nox
## 9  Boredom ~~    FOMO 46.311 -23.454 -23.454    -0.646    -0.646
## 12    FOMO ~~ Anxiety 19.338   6.492   6.492     0.166     0.166
## 13    FOMO ~~     Dep  1.190  -1.561  -1.561    -0.042    -0.042
## 14 Boredom  ~    FOMO 46.311  -0.700  -0.700    -0.609    -0.609
## 15    FOMO  ~ Anxiety 56.387   0.518   0.518     0.478     0.478
## 16    FOMO  ~     Dep 29.537   0.407   0.407     0.357     0.357
## 18 Anxiety  ~    FOMO 19.338   0.194   0.194     0.210     0.210
## 21     Dep  ~    FOMO  1.190  -0.047  -0.047    -0.053    -0.053
```

Model respecification
Given the MI results, we can add the suggested path and evaluate the model fit. The new model is proposed by freeing the parameter between Anxiety and FOMO. By definition, the new model and our old model are nested models.



```
#model comparison
m3<- '
Boredom ~ Anxiety + Dep
FOMO ~ Boredom + Anxiety  #adding path between FOMO and Anxiety
Anxiety ~~ Dep
'
fit3<- sem(m3, data=fomo)
```

Likelihood ratio test aka chi-square difference test
Likelihood ratio test compares M1 and M2(M2 is nested with M1). If the chi-square values of the two models are significantly different, then we choose the model with lower chi-square; if they are the same, we usually choose the simpler model.

Let’s look at r codes and outputs. We can use anova() or lavTestLRT(). They will give us the same results. The p-value is less than .001.Hence, we reject the null hypothesis and conclude that there is significant differences between our two models by giving up 1 degree of freedom.

```
anova(fit2, fit3)
## Chi-Squared Difference Test
```

```
##
##      Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
## fit3  1 6207.7 6239.4  0.2577
## fit2  2 6269.6 6297.8 64.1575       63.9       1  1.309e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
lavTestLRT(fit2, fit3)
## Chi-Squared Difference Test
##
##      Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
## fit3  1 6207.7 6239.4  0.2577
## fit2  2 6269.6 6297.8 64.1575       63.9       1  1.309e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit indices

We can also look at fit indice to judge whether our model is improved.

	Chi-square	Df(p)	CFI	RMSEA(p)	LOWER	UPPER	SRMR
Model1	64.158	2(.000)	.894	.353	.281	.429	.114
Model2	.258	1(.612)	1.000	.000	.000	.134	.004

Chi-square descreased

P-value is no more significant which is what we want

CFI increased
RMSEA decreased

SRMR decreased

Nested and Non-nested models

Information criteria

The AIC and BIC are information indices.They penalize for model complexity. These are not useful for evaluating the fit of one specific model (as we were doing with the other indices in this table). However, they are useful for identifying the best fitting model out of a candidate set. The better fitting model(s) out of a candidate set of models (2 or more models) are those with **AIC or BIC values that are lowest (i.e., lower is better).**

These indices are generally most useful for identifying the best fitting model out of a candidate set of non-nested models. However, they can also be useful when identifying fit even with nested models (although candidate nested models are most often evaluated using likelihood ratio chi-square tests).

AIC (Akaike’s Information Criterion) AIC = -2 log L + 2k

BIC (Bayesian Information Criterion) BIC =-2logL+k*logN
k = number of free parameters N = sample size

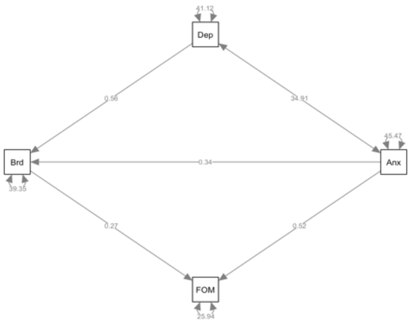
```
anova(fit2, fit3)
## Chi-Squared Difference Test
##
##      Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
## fit3  1 6207.7 6239.4  0.2577
## fit2  2 6269.6 6297.8 64.1575       63.9       1  1.309e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can find that model 2(fit3) has lower AIC and BIC values. Along with likelihood ratio test, modification idex, fit indices, model 2 shows better fit than model 1 meaning it is worthy to give up 1 dgree of freedom to add path relationship between Anxiety and FOMO.

But we need to always keep in mind that the model respecification should not purely rely on statistical outputs but theoretical support.

1.11 Model diagram
We could utilize the semPlot package to build the plot of the estimated model with estimated path coefficients shown. We will generate our path diagram based on model 2. The diagram could look slightly different from the original one. But we can always use different layout or style in the r code to make modification. (Type ?semPath in r for more information)

```
#model diagram with parameters
semPaths(fit3,what = "paths",whatLabels = "par",layout="tree3",residuals=T)
```



Since we are familiar with path analysis now, we will demonstrate full structural equation modeling following steps proposed by Hair et al. As Dr.Li introduced in class, the full structural equation model has both measurement model and structure model.

2.Full SEM

Introduction to dataset

Data Set: HBAT_SEM

HBAT is a common dataset developed and included in the textbook with many of the techniques to allow students to see the interrelationships among techniques as well as techniques themselves. HBAT_SEM with data responses from 400 individuals is specifically developed to demonstrate structural equation modeling.

Data Set Variables:

Latent Constructs:

Job satisfaction (JS). Reactions resulting from an appraisal of one’s job situation.
Organizational commitment (OC). The extent to which an employee identifies and feels part of HBAT.
Staying intentions (SI). The extent to which an employee intends to continue working for HBAT and is not participating in activities that make quitting more likely.
Environmental perceptions (EP). Beliefs an employee has about day-to-day, physical working conditions.
Attitudes toward coworkers (AC). Attitudes an employee has toward the coworkers he/she interacts with on a regular basis.

Observed Indicators:

item	**Scale type**	**Description**	**construct**
JS_1	0–10 Likert Disagree–Agree	All things considered, I feel very satisfied when I think about my job.	JS
OC_1	0–10 Likert Disagree–Agree	My work at HBAT gives me a sense of accomplishment.	OC
OC_2	0–10 Likert Disagree–Agree	I am willing to put in a great deal of effort beyond that normally expected to help HBAT be successful.	OC
EP_1	0–10 Likert Disagree–Agree	I am comfortable with my physical work environment at HBAT.	EP
OC_3	0–10 Likert Disagree–Agree	I have a sense of loyalty to HBAT.	OC
OC_4	0–10 Likert Disagree–Agree	I am proud to tell others that I work for HBAT.	OC
EP_2	0–10 Likert Disagree–Agree	The place I work in is designed to help me do my job better.	EP
EP_3	0–10 Likert Disagree–Agree	There are few obstacles to make me less productive in my workplace.	EP
AC_1	5-point Likert	How happy are you with the work of your coworkers? ___ Not happy ___Somewhat happy ___ Happy ___ Very happy ___ Extremely happy	AC
EP_4	7-point Semantic Differential	What term best describes your work environment at HBAT? Too hectic ________ Very soothing	EP
JS_2	7-point Semantic Differential	When you think of your job, how satisfied do you feel? Not at all satisfied ___________________ Very much satisfied	JS
JS_3	7-point Semantic Differential	How satisfied are you with your current job at HBAT? Very unsatisfied ___________________ Very satisfied	JS
AC_2	7-point Semantic Differential	How do you feel about your coworkers? Very unfavorable ___________ Very favorable	AC
SI_1	5-point Likert Disagree–Agree	I am not actively searching for another job. Strongly disagree ___________ Strongly agree	SI

JS_4	5-point Likert	How satisfied are you with HBAT as an employer? ___ Not at all ___ Little ___ Average ___ A lot ___ Very much	JS
SI_2	5-point Likert Disagree–Agree	I seldom look at the job listings on monster.com. Strongly disagree ___________ Strongly agree	SI
JS_5	Percent Satisfaction	Indicate your satisfaction with your current job at HBAT by placing a percentage in the blank, with 0% 5 Not satisfied at all, and 100% 5 Highly satisfied. ________	JS
AC_3	5-point Likert	How often do you do things with your coworkers on your days off?___ Never ___ Rarely ___ Occasionally ___ Often ___ Very often	AC
SI_3	5-point Likert Disagree–Agree	I have no interest in searching for a job in the next year. Strongly disagree ____________ Strongly agree	SI
AC_4	6-point Semantic Differential	Generally, how similar are your coworkers to you? Very different ___________ Very similar	AC
SI_4	5-point Likert	How likely is it that you will be working at HBAT one year from today? ___Very unlikely ___ Unlikely ___ Somewhat likely ___ Likely ___ Very likely	SI

Research Context:

HBAT employs thousands of workers in different operations around the world. Like many firms, one of its biggest management problems is attracting and keeping productive employees. The cost to replace and retrain employees is high. Yet the average new person hired works for HBAT less than three years. In most jobs, the first year is not productive, meaning the employee is not contributing as much as the costs associated with employing him/her. After the first year, most employees become productive. HBAT management would like to understand the factors that contribute to employee retention. A better understanding can be gained by learning how to measure the key constructs. Thus, HBAT is interested in developing and testing a measurement model made up of constructs that affect employees' attitudes and behaviors about remaining with HBAT.

Research questions

How do the proposed constructs affect employees' attitudes and behaviors about remaining with HBAT?

To answer our research question, we will follow the six-stage SEM process proposed by Hair et al.(2018) to demonstrate a full structural equation model in R using lavaan package.

Six-stage SEM process defined in Hair et al. (2018)

Stage 1: Defining individual constructs

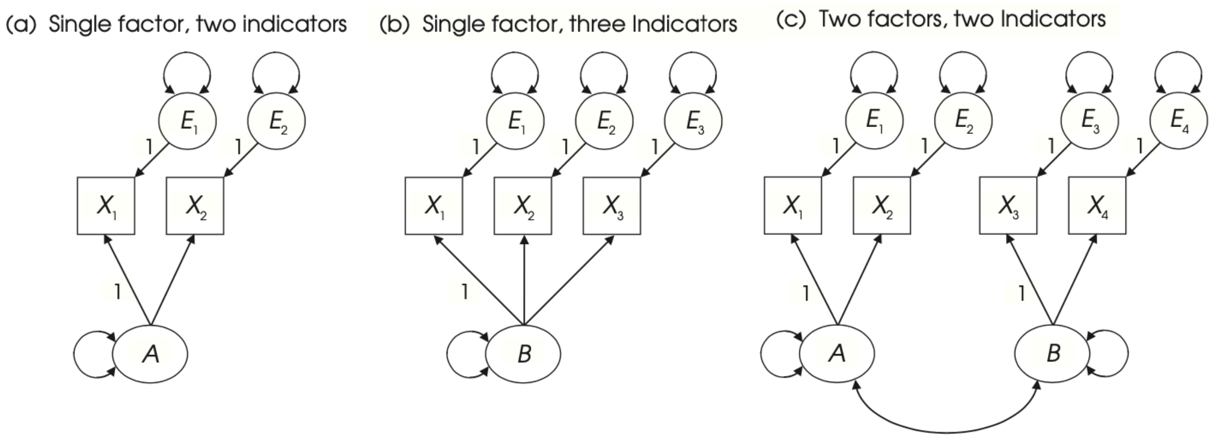
Based on published literature and some preliminary interviews with employees, HBAT initiated the research project focusing on five key constructs to study employee turnover problem. The five constructs of interest are *Job satisfaction (JS)*, *Organizational commitment (OC), Staying intentions (SI), Environmental perceptions (EP),* and *Attitudes toward coworkers (AC).*

Stage 2: Developing the overall measurement model

In this stage, researchers must specify the measurement model to be tested including relationships among constructs defined and the nature of each construct (reflective versus formative). Five constructs and 21 corresponding observed indictors are illustrated in the previous introduction section. Each construct is allowed to covary with another.

The following diagram shows the proposed measurement model from Hair et al(2018). Job Satisfaction (JS) contains 5 observed indicators; Organizational Commitment (OC) contains 4 observed indicators; Staying intentions (SI) contains 4 observed indicators; Environmental perceptions (EP) contains 4 observed indicators; and Attitudes toward coworkers (AC) contains 4 observed indicators. Since the focus of this lab is full SEM model, the matrix equation of the model will be shown in the full SEM part.

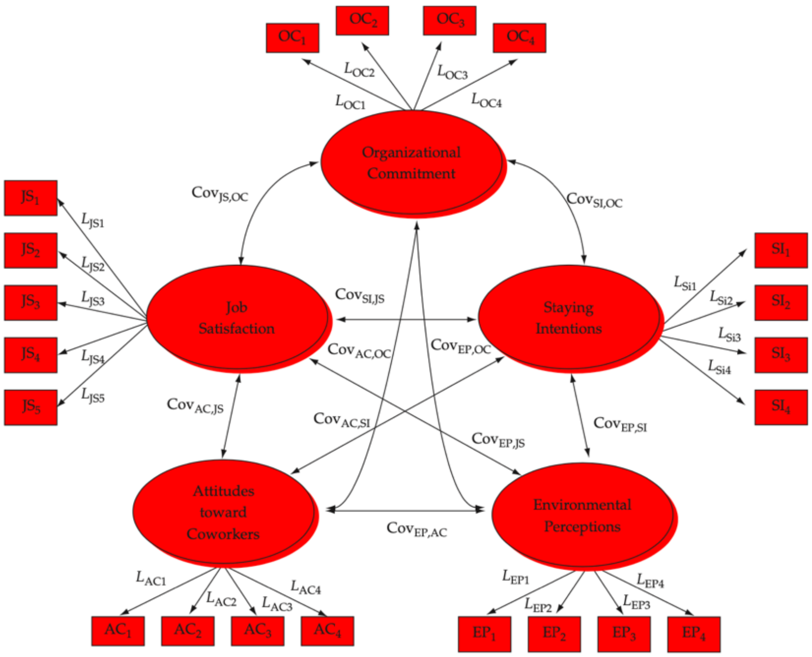
CFA model identification rule of thumb:
If a CFA model
1. with a single factor has at least three indicators, or
2. has two or more factors where each factor has two or more indicators, then the model is identified.


(a) Single factor, two indicators
(b) Single factor, three Indicators
(c) Two factors, two Indicators
E_1
E_2
E_3
E_4
X_1
X_2
X_3
X_4
1
A
B

Kline 2015,p202

As we can see, the manifest variables are presented in rectangles with error term pointing in. All the manifest variables in this model are endogenous variables that are predicted by their corresponding constructs. We call this kind of model **reflective model.**

Figure 10.5
Measurement Theory Model (CFA) for HBAT Employees



Stage 3: Designing a study to produce empirical results

HBAT conducted a study and four hundred completed responses were obtained. In this stage, models are specified in R followed by **model identification**. In the model, there are 52 parameters to be estimated including 16 factor loadings (a single loading estimate per construct is fixed to 1 to set the scale), 15 represent factor variance and covariance term, and 21 represent error variance term. The total number of unique variance and covariance terms equal the initial degrees of freedom: $df_{initial}$=p(p+1)/2 =21(21+1)/2=231. The number of unique variance and covariance term is greater the free parameter. This model has degrees of freedom= 231-52=179. Therefore, we conclude this model is overidentified.

In R, we specify our model with the following codes.

```
mcfa<- '
JS =~ JS1+JS2+JS3+JS4+JS5
AC =~ AC1+AC2+AC3+AC4
OC =~ OC1+OC2+OC3+OC4
EP =~ EP1+EP2+EP3+EP4
SI =~ SI1+SI2+SI3+SI4
'
fcfa<- cfa(mcfa,data=hbat.sem)
```

Matrix equation

To write our measure model in matrix form:

- We will first set EP =$\xi 1$(Ksi/Xi), AC= $\xi 2$, OC=$\xi 3$, JS= $\xi 4$, and SI= $\xi 5$ since they are all exogenous constructs in CFA model
- For $\xi 1$, let's set EP1- EP4 to X1 – X4
- For $\xi 2$, let's set AC1- AC4 to X5 – X8
- For $\xi 3$, let's set OC1-OC4 to X9-Y12
- For $\xi 4$, let's set JS1-JS5 to X13-Y17
- For $\xi 5$, let's set SI1-SI4 to X18-X21
- For X1-X21, their corresponding variance will be $\delta 1$- $\delta 21$ (delta)
- For X1-X21, their corresponding path parameters will be λ_x

Now, we are ready to write our matrix equation!

$$\begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \\ X5 \\ X6 \\ X7 \\ X8 \\ X9 \\ X10 \\ X11 \\ X12 \\ X13 \\ X14 \\ X15 \\ X16 \\ X17 \\ X18 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \lambda_{x21} & 0 & 0 & 0 & 0 \\ \lambda_{x31} & 0 & 0 & 0 & 0 \\ \lambda_{x41} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \lambda_{x62} & 0 & 0 & 0 \\ 0 & \lambda_{x72} & 0 & 0 & 0 \\ 0 & \lambda_{x82} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \lambda_{x103} & 0 & 0 \\ 0 & 0 & \lambda_{x113} & 0 & 0 \\ 0 & 0 & \lambda_{x123} & 0 & 0 \\ 0 & 00 & 0 & 1 & 0 \\ 0 & 0 & 0 & \lambda_{x144} & 0 \\ 0 & 0 & 0 & \lambda_{x154} & 0 \\ 0 & 0 & 0 & \lambda_{x164} & 0 \\ 0 & 0 & 0 & \lambda_{x174} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \lambda_{x195} \\ 0 & 0 & 0 & 0 & \lambda_{x205} \\ 0 & 0 & 0 & 0 & \lambda_{x215} \end{bmatrix} \begin{bmatrix} \xi 1 \\ \xi 2 \\ \xi 3 \\ \xi 4 \\ \xi 5 \end{bmatrix} + \begin{bmatrix} \delta 1 \\ \delta 2 \\ \delta 3 \\ \delta 4 \\ \delta 5 \\ \delta 6 \\ \delta 7 \\ \delta 8 \\ \delta 9 \\ \delta 10 \\ \delta 11 \\ \delta 12 \\ \delta 13 \\ \delta 14 \\ \delta 15 \\ \delta 16 \\ \delta 17 \\ \delta 18 \end{bmatrix}$$

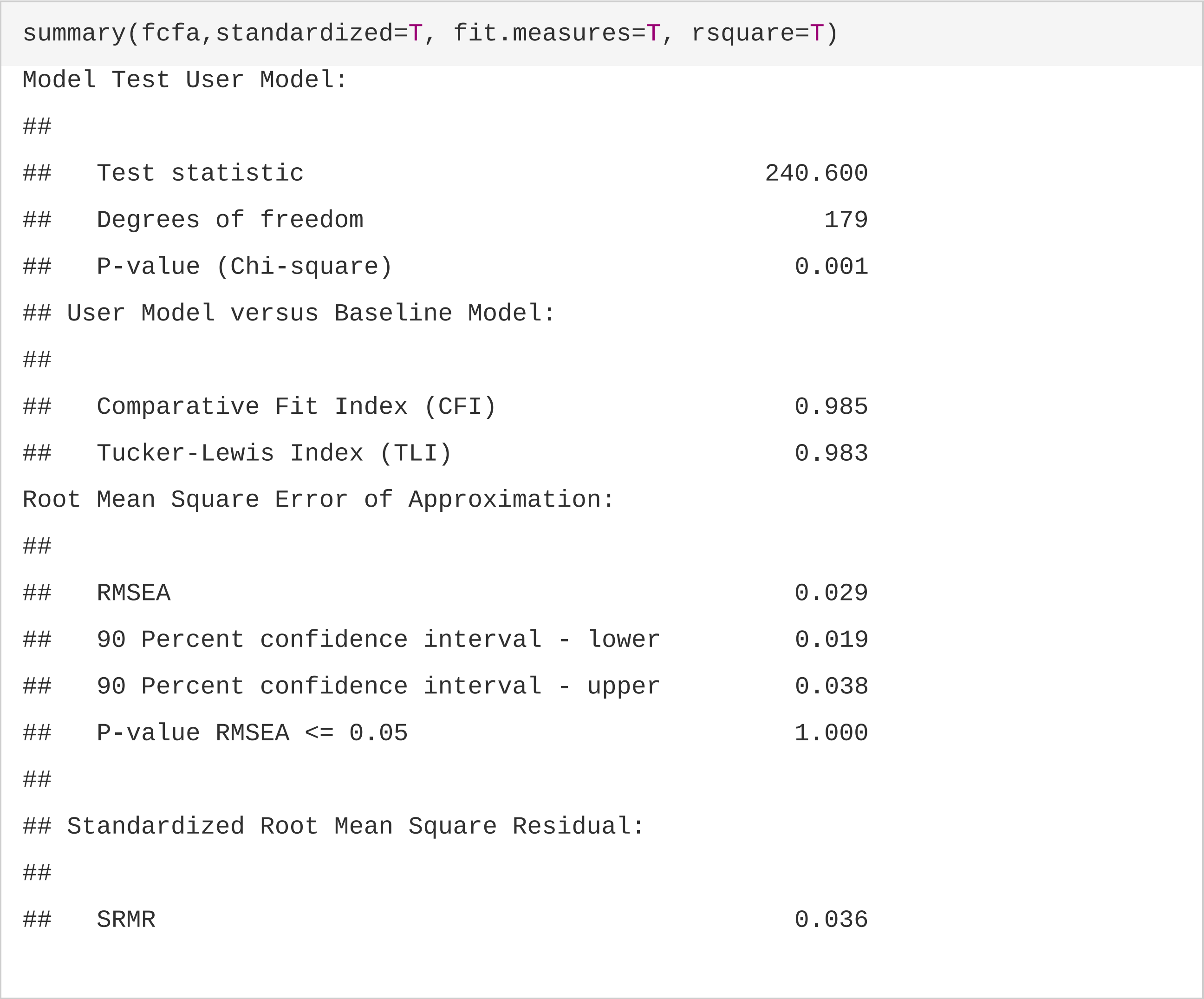
$$\Phi = \begin{bmatrix} \Phi_{11} & & & & \\ \Phi_{21} & \Phi_{22} & & & \\ \Phi_{31} & \Phi_{32} & \Phi_{33} & & \\ \Phi_{41} & \Phi_{42} & \Phi_{43} & \Phi_{44} & \\ \Phi_{51} & \Phi_{52} & \Phi_{53} & \Phi_{54} & \Phi_{55} \end{bmatrix}$$

$$\Theta_\delta = \begin{bmatrix} \theta_{\delta 11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{\delta 2121} \end{bmatrix}$$

Stage 4: Assessing measurement model validity

In this stage, we examined the results of testing measurement theory by comparing the theoretical measurement model against reality, as represented by the sample covariance matrix. We first examined the overall fit through key GOF values, construct validity consisting of convergent validity, discriminant validity, and nomological validity. Then, path estimates, standardized residuals, and modification indices were further examined for model improvement.

Overall fit: CFA output includes many fit indices. We did not present all possible fit indices. Rather, we will focus on the key GOF values using our rules of thumb to provide some assessment of fit. We will discuss the key values such as x^2 statistic, as well as the CFI and RMSEA.



```
summary(fcfa,standardized=T, fit.measures=T, rsquare=T)
Model Test User Model:
##
##   Test statistic                              240.600
##   Degrees of freedom                              179
##   P-value (Chi-square)                          0.001
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                   0.985
##   Tucker-Lewis Index (TLI)                      0.983
Root Mean Square Error of Approximation:
##
##   RMSEA                                         0.029
##   90 Percent confidence interval - lower        0.019
##   90 Percent confidence interval - upper        0.038
##   P-value RMSEA <= 0.05                         1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                          0.036
```

The overall model x^2 is 240.600 with 179 degrees of freedom. The p-value associated with this result is .001. This p-value is significant using a type I error rate of .05. Thus, the x^2 goodness of fit statistic does not indicate that the observed covariance matrix matches the estimated covariance matrix within sampling variance. However, given the problems associated with statistical power, and the effective sample size of 399, we examine other fit statistics closely as well.

Next we look at several other fit indices. Our rule of thumb suggests that we rely on at least one absolute fit index and one incremental fit index, in addition to the x^2 results. The value for RMSEA, an absolute fit index, is 0.029. This value appears quite low and is below the .08 guideline for a model with 21 measured variables and a sample size of 400. Using the 90 percent confidence interval for this RMSEA, we conclude the true value of RMSEA is between 0.019 and 0.038. Thus, even the upper bound of RMSEA is low in this case. The RMSEA therefore provides additional support for model fit. Next we see the standardized root mean square residual(SRMR) with a value of .036, below the conservative cut-off of .05.

Moving to the incremental fit indices, the CFI is the most widely used index. In our HBAT CFA model the CFI has a value of 0.985, which, like the RMSEA, exceeds the CFI guidelines of greater than .94 for a model of this complexity and sample size.

The CFA results suggest the HBAT measurement model provides a reasonably good fit, and thus it is suitable to proceed to further examination of the model results. Issues related to construct validity will be examined next and then the focus shifts to model diagnostics aimed at improving the specified model.

Construct validity

To assess construct validity, we examine convergent, discriminant, and nomological validity.

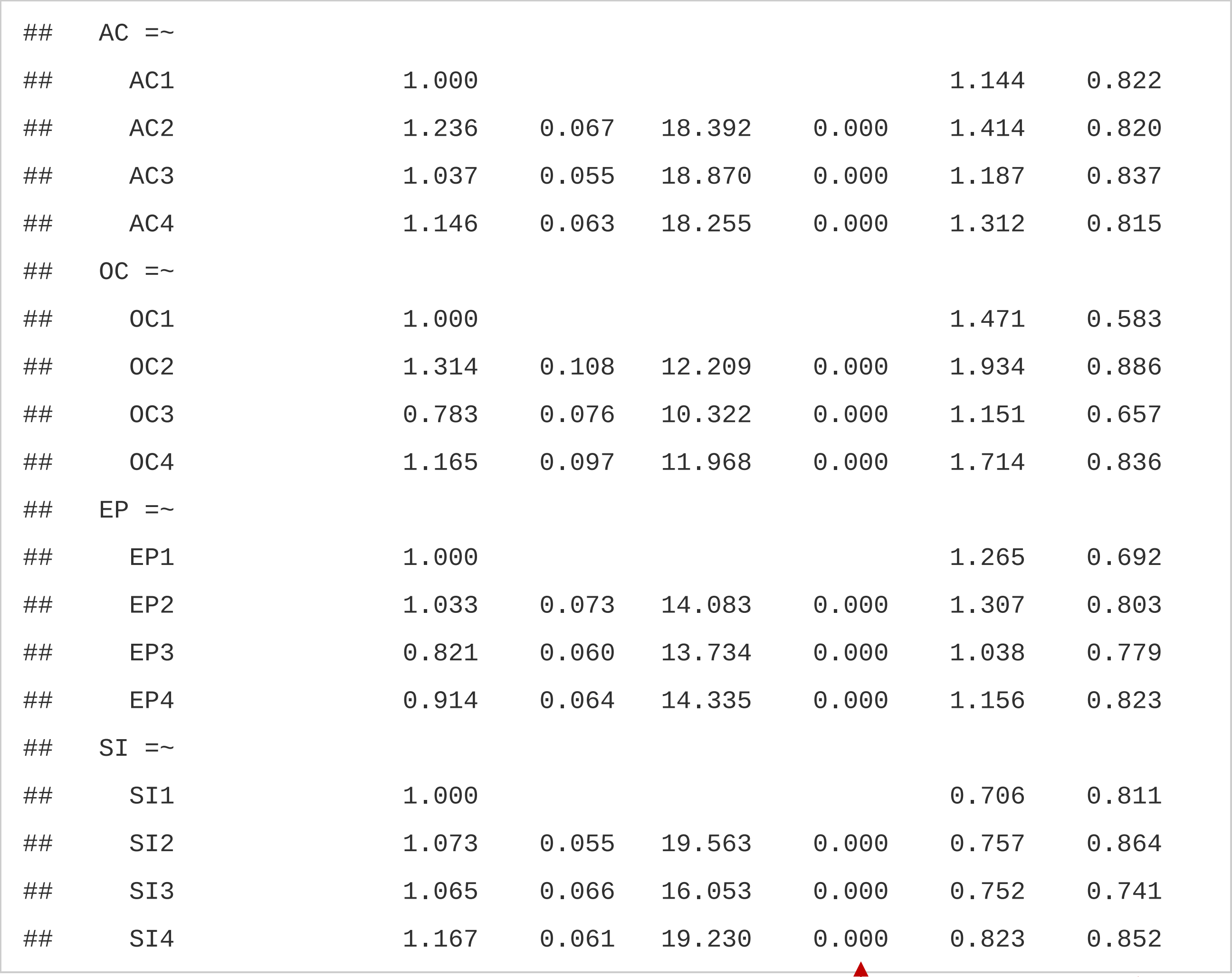
- Convergent validity

  The items that are indicators of a specific construct should converge or share a high proportion of variance in common, known as convergent validity. Several ways are available to estimate the relative amount of convergent validity among item measures.

  Factor loading

  At a minimum, all factor loadings should be statistically significant. Because a statistically significant loading could still be very weak in strength, particularly with large samples, a good rule of thumb is that standardized loading estimates should be .5 or higher, and ideally .7 or higher. In most cases, researchers should interpret standardized parameter estimates, because they are constrained to range between 1.0 and -1.0.

```
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   JS =~
##     JS1               1.000                               0.990    0.740
##     JS2               1.033    0.076   13.682    0.000    1.023    0.748
##     JS3               0.903    0.072   12.515    0.000    0.894    0.680
##     JS4               0.910    0.070   12.953    0.000    0.901    0.705
##     JS5              15.190    1.133   13.410    0.000   15.042    0.731
```



```
##   AC =~
##     AC1               1.000                               1.144    0.822
##     AC2               1.236    0.067   18.392    0.000    1.414    0.820
##     AC3               1.037    0.055   18.870    0.000    1.187    0.837
##     AC4               1.146    0.063   18.255    0.000    1.312    0.815
##   OC =~
##     OC1               1.000                               1.471    0.583
##     OC2               1.314    0.108   12.209    0.000    1.934    0.886
##     OC3               0.783    0.076   10.322    0.000    1.151    0.657
##     OC4               1.165    0.097   11.968    0.000    1.714    0.836
##   EP =~
##     EP1               1.000                               1.265    0.692
##     EP2               1.033    0.073   14.083    0.000    1.307    0.803
##     EP3               0.821    0.060   13.734    0.000    1.038    0.779
##     EP4               0.914    0.064   14.335    0.000    1.156    0.823
##   SI =~
##     SI1               1.000                               0.706    0.811
##     SI2               1.073    0.055   19.563    0.000    0.757    0.864
##     SI3               1.065    0.066   16.053    0.000    0.752    0.741
##     SI4               1.167    0.061   19.230    0.000    0.823    0.852
```

The p-values for all loadings are less than .05 meaning they are significant

All standardized loadings are above .5

Average variance extracted

The square of a standardized factor loading represents how much variation in an item is explained by the latent factor and is termed the variance extracted of the item. Thus, a loading of .7 squared yields a communality of .5. In short, the factor is explaining half the variation in the item with the other half being error variance.

With CFA, the average variance extracted (AVE) is calculated as the mean variance extracted for the items loading on a construct and is a summary indicator of convergence. This value can be calculated using standardized loadings:

$$AVE = \frac{\sum_{i=1}^{n} L_i^2}{n}$$

L_i represents the completely standardized factor loading for the ith measured variable and n is the number of item indicators for a construct. In words, AVE is computed as the total of all squared standardized factor loadings (squared multiple correlations) divided by the number of items. An AVE of less than .5 indicates that, on average, more error remains in the items than variance held in common with the latent factor upon which they load.

```
#AVE
(0.740^2+0.748^2+0.680^2+0.705^2+0.731^2)/5 #JS  52.02%
(0.822^2+0.820^2+0.837^2+0.815^2)/4 #AC  67.82%
(0.583^2+0.886^2+0.657^2+0.836)/4 #OC  59.81%
(0.692^2+0.803^2+0.779^2+0.823)/4 #EP  63.84%
(0.811^2+0.864^2+0.741^2+0.852)/4 #SI  70.13%
All exceed the 50% rule of thumb
```

- Discriminant validity
  Discriminant validity is the extent to which a construct or variable is truly distinct from other constructs or variables. Thus, high discriminant validity provides evidence that a construct is unique and captures some phenomena other measures do not.

  A rigorous test is to compare the average variance-extracted values for any two constructs with the square of the correlation estimate between these two constructs. The variance-extracted estimates should be greater than the squared correlation estimate. The logic here is based on the idea that a latent construct should explain more of the variance in its item measures that it shares with another construct. Passing this test provides good evidence of discriminant validity.(Hair 2018)

```
lvcorr<- inspect(fcfa, "cor.lv")
lvcorr^2
##    JS    AC    OC    EP    SI
## JS 1.000
## AC 0.002 1.000
## OC 0.044 0.094 1.000
## EP 0.058 0.066 0.247 1.000
## SI 0.053 0.095 0.306 0.316 1.000
```

All AVE estimates are greater than the corresponding interconstruct squared correlation estimates. Therefore, this test indicates there are no problems with discriminant validity for the HBAT CFA model.

```
inspect(fcfa, "cor.lv")
```

```
##    JS    AC    OC    EP    SI
## JS 1.000
## AC 0.050 1.000
## OC 0.209 0.307 1.000
## EP 0.242 0.257 0.497 1.000
## SI 0.230 0.309 0.553 0.562 1.000
```

Ahmad et al.(2016) suggest that the discriminant validity is achieved when the measurement model is free from redundant items. Another requirement for discriminant validity is the correlation between each pair of latent exogenous construct should be less than 0.85. Other than that, the square root of AVE for the construct should be higher than the correlation between the respective constructs(Almad et al.2016)

- Nomological validity
  Nomological validity refers to the degree that the summated scale makes accurate predictions of other concepts in a theoretically based model. The correlation matrix provides a useful start in this effort to the extent that the constructs are expected to relate to one another. Previous organizational behavior research suggests that more favorable evaluations of all constructs are generally expected to produce positive employee outcomes. Correlations between the factor scores for each construct support the prediction that these constructs are positively related to one another.

```
#Nomological validity
inspect(fcfa, "cor.lv")
##    JS    AC    OC    EP    SI
## JS 1.000
## AC 0.050 1.000
## OC 0.209 0.307 1.000
## EP 0.242 0.257 0.497 1.000
## SI 0.230 0.309 0.553 0.562 1.000
```

Reliability

- Construct reliability
  Coefficient alpha remains a commonly applied estimate, although it may understate reliability. Different reliability coefficients do not produce dramatically different reliability estimates, but a slightly different **construct reliability (CR).** High construct reliability indicates that internal consistency exists (≥|.7|), meaning that the measures all consistently represent the same latent construct.
  It is computed from the squared sum of factor loadings (L_i) for each construct and the sum of error variance term for a construct(e_i) as:

$$CR = \frac{(\sum_{i=1}^{n} L_i)^2}{(\sum_{i=1}^{n} L_i)^2 + (\sum_{i=1}^{n} e_i)}$$

It is a little more complicated to calculate manually. Let use R commands to do this. As we can see here, the reliability alpha is low for JS construct.

```
#construct reliability
semTools::reliability(fcfa)
##                JS        AC        OC        EP        SI
## alpha  0.2809951 0.8907652 0.8227408 0.8474340 0.8863459(Cronbach alpha, 1951)
## omega  0.6396532 0.8928576 0.8267927 0.8496932 0.8871747(CR)
## omega2 0.6396532 0.8928576 0.8267927 0.8496932 0.8871747
## omega3 0.6404932 0.8928422 0.8180947 0.8498606 0.8870258
## avevar 0.5345483 0.6772156 0.5524543 0.5874690 0.6635062 (AVE)
```

- Internal reliability
  Internal reliability is achieved when the Cronbach's Alpha value is 0.6 or higher. Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability.(Ahmad et al.,2016)
  As we can see, the alpha for JS construct did not reach expected level. The internal reliability is not hold. (This result is not aligned with textbook results but consistent across R and Stata. I do not know the reason.)

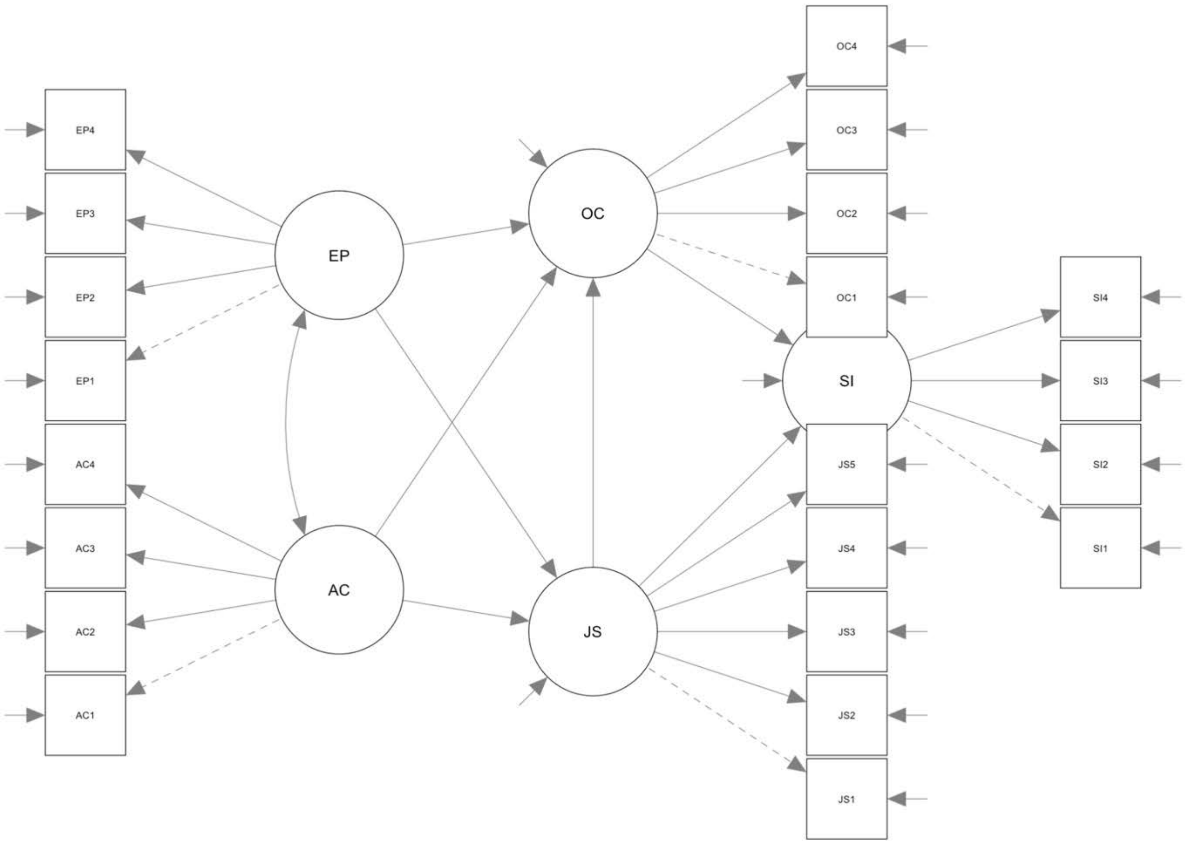
Stage 5: Specifying the structural model

In this stage, HBAT research team proposed a theory based on literature and collective experience. The theory implies the following piecemeal hypotheses:

H1: Environmental perceptions(EP) are positively related to job satisfaction(JS).
H2: Environmental perceptions(EP) are positively related to organizational commitment(OC).
H3: Attitudes toward coworkers(AC) are positively related to job satisfaction(JS).
H4: Attitudes toward coworkers(AC) are positively related to organizational commitment(OC).
H5: Job satisfaction(JS) is related positively to organizational commitment(OC).
H6: Job satisfaction(JS) is related positively to staying intentions(SI).
H7: Organizational commitment(OC) is related positively to staying intentions(SI).

Hypothesis
H1: EP +→JS
H2: EP +→ OC
H3: AC +→ JS
H4: AC +→ OC

H5: JS + →OC
H6: JS +→ SI
H7: OC +→ SI



Matrix equation
To write our model in matrix form:

- We will first set EP =$\xi 1$(Ksi/Xi), AC= $\xi 2$ since they are exogenous constructs.
- Then we will set OC=$\eta 1(Eta)$, JS= $\eta 2$, and SI= $\eta 3$ since they are endogenous constructs.

- For $\xi 1$, let's set EP1- EP4 to X1 – X4
- For $\xi 2$, let's set AC1- AC4 to X5 – X8
- For X1-X8, their corresponding variance will be $\delta 1$- $\delta 8$ (delta)
- For X1-X8, their corresponding path parameters will be λ_x

- For $\eta 1$, let's set OC1-OC4 to Y1-Y4
- For $\eta 2$, let's set JS1-JS5 to Y5-Y9
- For $\eta 3$, let's set SI1-SI4 to Y10-Y13
- For Y1-Y13, their corresponding variance will be $\epsilon 1$- $\epsilon 13$ (epsilon)
- For Y1-Y13, their corresponding path parameters will be λ_y

- The parameter between ξ and η will be written as γ (gamma); The parameter between η and η will be written as β (beta);

Now, we are ready to write our matrix equation!

Measurement model 1

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{x21} & 0 \\ \lambda_{x31} & 0 \\ \lambda_{x41} & 0 \\ 0 & 1 \\ 0 & \lambda_{x62} \\ 0 & \lambda_{x72} \\ 0 & \lambda_{x82} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \end{bmatrix}$$

Measurement model 2

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{Y21} & 0 & 0 \\ \lambda_{Y31} & 0 & 0 \\ \lambda_{Y41} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{Y52} & 0 \\ 0 & \lambda_{Y62} & 0 \\ 0 & \lambda_{Y72} & 0 \\ 0 & \lambda_{Y82} & 0 \\ 0 & \lambda_{Y92} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{Y113} \\ 0 & 0 & \lambda_{Y123} \\ 0 & 0 & \lambda_{Y133} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \end{bmatrix}$$

Structural Model

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} 0 & \beta_{12} \\ 0 & 0 \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix} \quad \Psi = \begin{bmatrix} \psi_{11} & & \\ 0 & \psi_{22} & \\ 0 & 0 & \psi_{33} \end{bmatrix}$$

$$\Theta_\delta = \begin{bmatrix} \theta_{\delta 11} & & & & & & & \\ 0 & \theta_{\delta 22} & & & & & & \\ 0 & 0 & \theta_{\delta 33} & & & & & \\ 0 & 0 & 0 & \theta_{\delta 44} & & & & \\ 0 & 0 & 0 & 0 & \theta_{\delta 55} & & & \\ 0 & 0 & 0 & 0 & 0 & \theta_{\delta 66} & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{\delta 77} & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{\delta 88} \end{bmatrix}$$

$$\Theta_\epsilon = \begin{bmatrix} \theta_{\epsilon 11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_{\epsilon 1313} \end{bmatrix}$$

$$\Lambda_x = \begin{bmatrix} 1 & 0 \\ \lambda_{x21} & 0 \\ \lambda_{x31} & 0 \\ \lambda_{x41} & 0 \\ 0 & 1 \\ 0 & \lambda_{x62} \\ 0 & \lambda_{x72} \\ 0 & \lambda_{x82} \end{bmatrix} \Lambda_y = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{Y21} & 0 & 0 \\ \lambda_{Y31} & 0 & 0 \\ \lambda_{Y41} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{Y52} & 0 \\ 0 & \lambda_{Y62} & 0 \\ 0 & \lambda_{Y72} & 0 \\ 0 & \lambda_{Y82} & 0 \\ 0 & \lambda_{Y92} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{Y113} \\ 0 & 0 & \lambda_{Y123} \\ 0 & 0 & \lambda_{Y133} \end{bmatrix}$$

Stage6: Assessing structural model validity

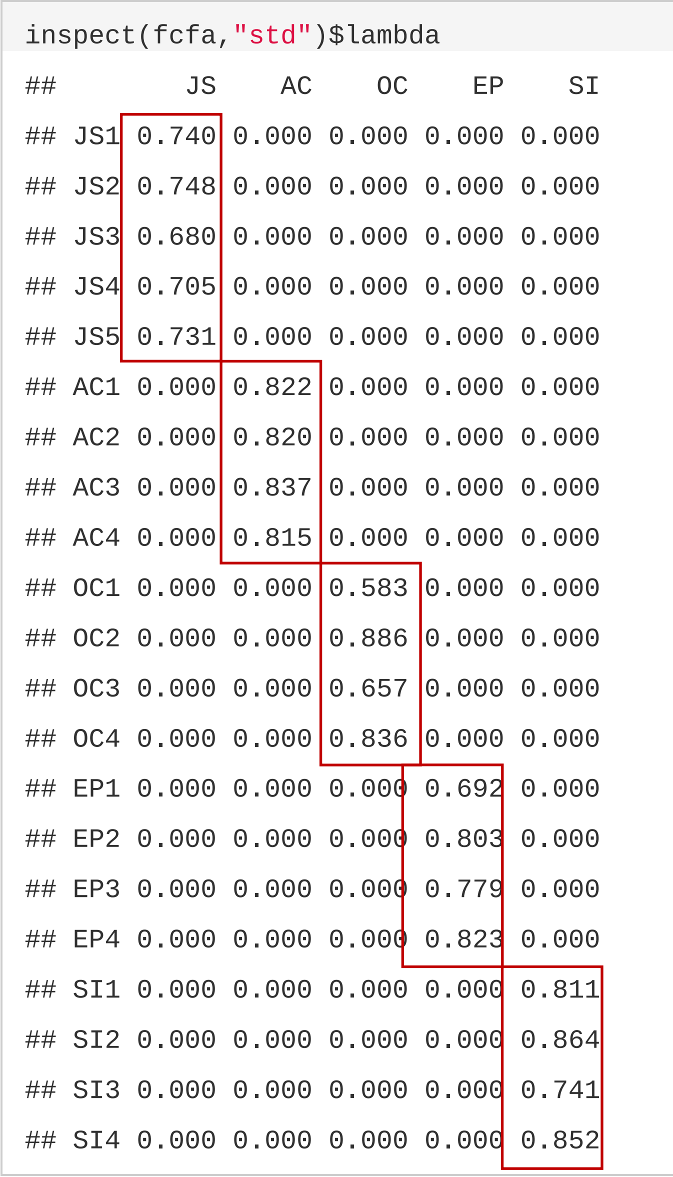
In this stage, our emphasis first will be on SEM model fit and then whether the structural relationships are consistent with theoretical expectations. We will examine the key GOF (goodness of fit) values as in stage 4. Besides, we examined the loading estimates to make sure they are virtually unchanged from the CFA results. Next, we examined the model diagnostics.

GOF index

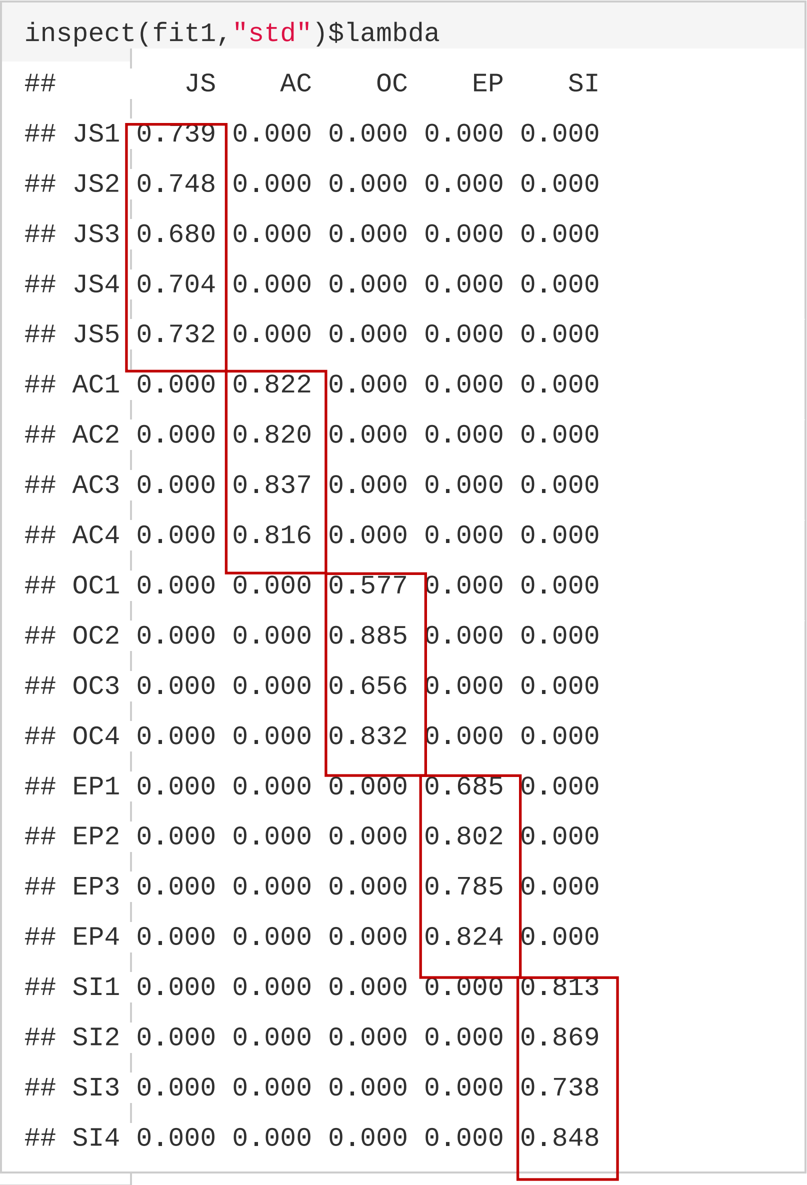
The table below shows the overall fit statistics from testing the Employee Retention Model. The chi-square is 287.04 with 181 degrees of freedom(p<.05), meaning the normed chi-square is 287.04/181=1.59.The CFI is .975 with a RMSEA of .038, which corresponds to a 90 percent confidence interval of .030 to .046. Although 0 is not in the confidence interval, the RMSEA value is relatively low. All measures are within a range that would be associated with good fit.

```
##                 cfa model employee retention model
## chisq             240.600                  287.040
## df                179.000                  181.000
## pvalue              0.001                    0.000
## gfi                 0.947                    0.938
## rmsea               0.029                    0.038
## rmsea.ci.lower      0.019                    0.030
## rmsea.ci.upper      0.038                    0.046
## rmr                 0.414                    0.410
## srmr                0.036                    0.060
## nfi                 0.946                    0.936
## nnfi                0.983                    0.971
## cfi                 0.985                    0.975
## rfi                 0.937                    0.925
## agfi                0.932                    0.921
## pnfi                0.806                    0.806
```

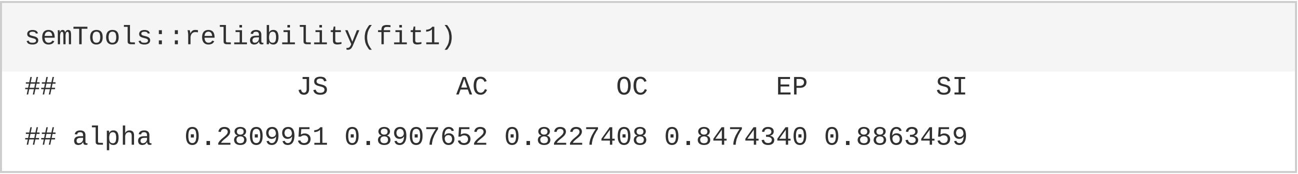
*Loading: Individual standardized factor loadings (regression weights) are above .5.*The loadings estimates are virtually unchanged from the CFA results. The estimated standardized loadings changed slightly.



```
inspect(fcfa,"std")$lambda
##        JS    AC    OC    EP    SI
## JS1 0.740 0.000 0.000 0.000 0.000
## JS2 0.748 0.000 0.000 0.000 0.000
## JS3 0.680 0.000 0.000 0.000 0.000
## JS4 0.705 0.000 0.000 0.000 0.000
## JS5 0.731 0.000 0.000 0.000 0.000
## AC1 0.000 0.822 0.000 0.000 0.000
## AC2 0.000 0.820 0.000 0.000 0.000
## AC3 0.000 0.837 0.000 0.000 0.000
## AC4 0.000 0.815 0.000 0.000 0.000
## OC1 0.000 0.000 0.583 0.000 0.000
## OC2 0.000 0.000 0.886 0.000 0.000
## OC3 0.000 0.000 0.657 0.000 0.000
## OC4 0.000 0.000 0.836 0.000 0.000
## EP1 0.000 0.000 0.000 0.692 0.000
## EP2 0.000 0.000 0.000 0.803 0.000
## EP3 0.000 0.000 0.000 0.779 0.000
## EP4 0.000 0.000 0.000 0.823 0.000
## SI1 0.000 0.000 0.000 0.000 0.811
## SI2 0.000 0.000 0.000 0.000 0.864
## SI3 0.000 0.000 0.000 0.000 0.741
## SI4 0.000 0.000 0.000 0.000 0.852
```



```
inspect(fit1,"std")$lambda
##        JS    AC    OC    EP    SI
## JS1 0.739 0.000 0.000 0.000 0.000
## JS2 0.748 0.000 0.000 0.000 0.000
## JS3 0.680 0.000 0.000 0.000 0.000
## JS4 0.704 0.000 0.000 0.000 0.000
## JS5 0.732 0.000 0.000 0.000 0.000
## AC1 0.000 0.822 0.000 0.000 0.000
## AC2 0.000 0.820 0.000 0.000 0.000
## AC3 0.000 0.837 0.000 0.000 0.000
## AC4 0.000 0.816 0.000 0.000 0.000
## OC1 0.000 0.000 0.577 0.000 0.000
## OC2 0.000 0.000 0.885 0.000 0.000
## OC3 0.000 0.000 0.656 0.000 0.000
## OC4 0.000 0.000 0.832 0.000 0.000
## EP1 0.000 0.000 0.000 0.685 0.000
## EP2 0.000 0.000 0.000 0.802 0.000
## EP3 0.000 0.000 0.000 0.785 0.000
## EP4 0.000 0.000 0.000 0.824 0.000
## SI1 0.000 0.000 0.000 0.000 0.813
## SI2 0.000 0.000 0.000 0.000 0.869
## SI3 0.000 0.000 0.000 0.000 0.738
## SI4 0.000 0.000 0.000 0.000 0.848
```



```
semTools::reliability(fit1)
##                JS        AC        OC        EP        SI
## alpha  0.2809951 0.8907652 0.8227408 0.8474340 0.8863459
## omega  0.6401053 0.8928834 0.8237614 0.8487858 0.8868089
## omega2 0.6401053 0.8928834 0.8237614 0.8487858 0.8868089
## omega3 0.6409479 0.8928948 0.8101963 0.8477886 0.8860906
## avevar 0.5351300 0.6772825 0.5473606 0.5855465 0.6626318
```

Regression coefficients
Let's recall that our model hypotheses:
H1: EP +→JS
H2: EP +→ OC
H3: AC +→ JS
H4: AC +→ OC
H5: JS + →OC
H6: JS +→ SI
H7: OC +→ SI
All but two structural path estimates are significant and in the expected direction. The exceptions are the estimates between AC and JS and between JS and OC. Both estimates have significance below the critical t-value for a Type I error of .05. Therefore, although the esti- mate is in the hypothesized direction, it is not supported. Overall, however, given that five of seven estimates are consistent with individual hypotheses, these results support the theoretical model, with a caveat for the two insignificant paths.

```
Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   JS ~
##     EP                 0.199    0.049    4.040    0.000    0.252    0.252
##     AC                -0.010    0.051   -0.188    0.851   -0.011   -0.011
##   OC ~
##     EP                 0.523    0.079    6.627    0.000    0.450    0.450
##     AC                 0.255    0.068    3.745    0.000    0.200    0.200
##     JS                 0.126    0.078    1.608    0.108    0.085    0.085
##   SI ~
##     JS                 0.087    0.036    2.383    0.017    0.121    0.121
##     OC                 0.269    0.032    8.284    0.000    0.553    0.553
##
## Covariances:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   AC ~~
##     EP                 0.368    0.087    4.244    0.000    0.257    0.257
```

Model diagnostics

Several diagnostic measures are available for researchers to evaluate SEM models. They range from fit indices to standardized residuals and modification indices. Each of these will be examined in the following discussion to determine if model respecification should be considered.

Chi-square

The chi-square difference is 287.04-240.60=46.44 with two degrees of freedom. The difference in degrees of freedom is two, which is due to the fact that the model includes all but two of the possible structural relationships. The possibility of another meaningful structural path should be explored, particularly if other diagnostic information points in that direction.

Modification indices

Examination of the two modification indices for the direct paths of EP→SI and AC →SI shows that both have values over 4.0, although the EP →SI value is much higher (39.855 versus 9.011). This strongly supports the addition of the EP S SI relationship if it can be supported theoretically.

Note that modification indices can be estimated for the “second half” of the recursive relationships between SI -> JS and SI -> OC. As we can see, both indicate substantial improvement in model fit. But more important, because they have no theoretical basis for inclusion in the model, they highlight the potential dangers. (Hair2018)

```
#modification indices
modindices(fit1)
```

```
## 353   JS ~~   SI 38.227 -0.982  -1.797   -1.797   -1.797
## 355   AC ~~   SI  2.688  0.063   0.097    0.097    0.097
## 357   OC ~~   SI 44.753 -0.536  -0.781   -0.781   -0.781
## 358   EP ~~   SI 28.333  0.264   0.369    0.369    0.369
## 360   JS  ~   SI 38.226 -3.010  -2.155   -2.155   -2.155
## 361   OC  ~   SI 44.755 -1.643  -0.799   -0.799   -0.799
## 362   SI  ~   EP 39.855  0.217   0.383    0.383    0.383
## 363   SI  ~   AC  9.011  0.095   0.153    0.153    0.153
## 366   EP  ~   SI 28.333  0.809   0.457    0.457    0.457
## 370   AC  ~   SI  2.688  0.194   0.120    0.120    0.120
```

Model respecification

Any respecification must have strong theoretical as well as empirical support of the nature that would allow an a priori alternative model. Model respecification should not be the result of searching for relationships, but rather for improving model fit that is theoretically justified. To further assess the SEM model, the HBAT research team conducts a post hoc analysis adding a direct relationship between EP and SI.

```
m2<- '
JS =~ JS1+JS2+JS3+JS4+JS5
AC =~ AC1+AC2+AC3+AC4
OC =~ OC1+OC2+OC3+OC4
EP =~ EP1+EP2+EP3+EP4
SI =~ SI1+SI2+SI3+SI4
JS ~ EP +AC
OC ~ EP + AC + JS
SI ~ JS + OC +EP
EP ~~ AC

'
fit2<- sem(model=m2, data=hbat.sem)
summary(fit2, standardized=T, rsquare=T, fit.measures=T)
```

```
##                    cfa retention revised
## chisq          240.600   287.040 246.102
## df             179.000   181.000 180.000
## pvalue           0.001     0.000   0.001
## gfi              0.947     0.938   0.945
## rmsea            0.029     0.038   0.030
## rmsea.ci.lower   0.019     0.030   0.020
## rmsea.ci.upper   0.038     0.046   0.039
## rmr              0.414     0.410   0.412
## srmr             0.036     0.060   0.040
## nfi              0.946     0.936   0.945
## nnfi             0.983     0.971   0.982
## cfi              0.985     0.975   0.984
## rfi              0.937     0.925   0.936
## agfi             0.932     0.921   0.930
## pnfi             0.806     0.806   0.810
```

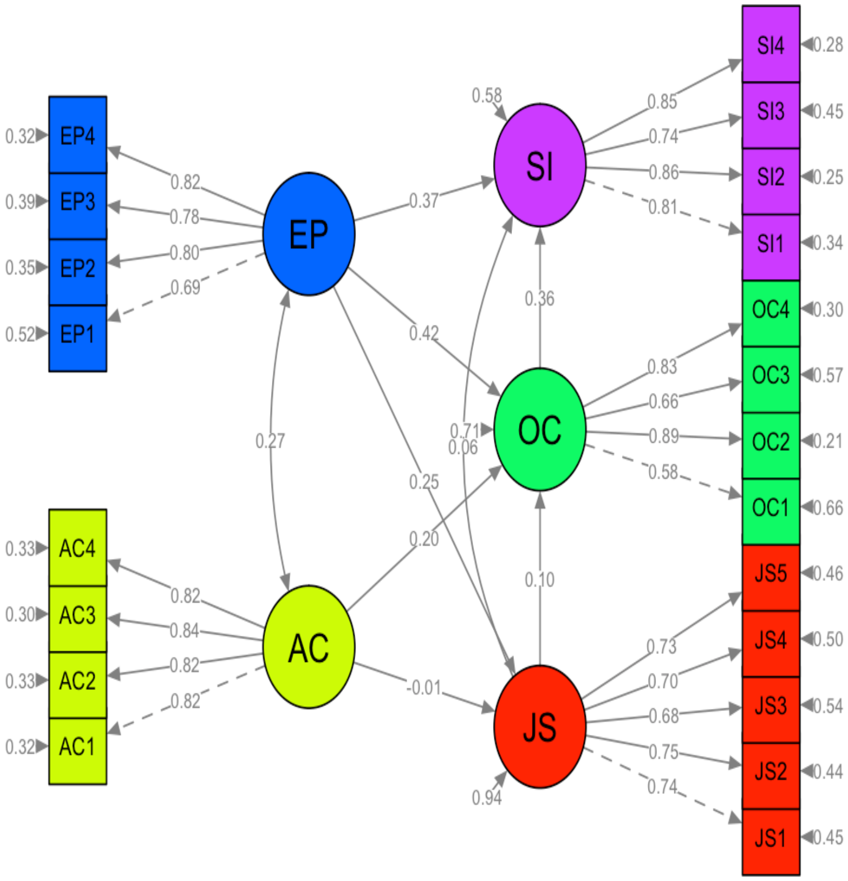
```
Regressions:
```

```
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   JS ~
##     EP                 0.192    0.049    3.934    0.000    0.245    0.245
##     AC                -0.012    0.051   -0.229    0.819   -0.013   -0.013
##   OC ~
##     EP                 0.488    0.077    6.309    0.000    0.421    0.421
##     AC                 0.253    0.070    3.640    0.000    0.198    0.198
##     JS                 0.144    0.080    1.803    0.071    0.097    0.097
##   SI ~
##     JS                 0.046    0.035    1.330    0.183    0.065    0.065
##     OC                 0.172    0.030    5.777    0.000    0.359    0.359
##     EP                 0.207    0.034    6.094    0.000    0.371    0.371
##
## Covariances:
##                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   AC ~~
##     EP                 0.384    0.088    4.376    0.000    0.265    0.265
##
```

The resulting standardized parameter estimate for $P_{SI,EP}$ is.37(p<.001) which is significant. The overall fit reveals a x^2 value of 246.102 with 180 degrees of freedom and a normed x^2 value of 1.37. The CFI goes up to .984 which is practically the same as the CFA model. This is a better fit than the original structrural model. We can also look at the likelihood ratio test. The p value is less than .05 meaning our revised model is significantly different from the original one.

```
anova(fit1,fit2)
## Chi-Squared Difference Test
##
##       Df   AIC   BIC  Chisq Chisq diff Df diff      Pr(>Chisq)
## fit2 180 27941 28144 246.10
## fit1 181 27980 28179 287.04     40.938       1 0.0000000001572 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
semPaths(fit2,what = "path", whatLabels = "std", residuals = T,rotation =2,la
yout = "tree2",style = "lisrel",groups = "latents")
```



Reference

Holte, A. J. & Ferraro, F. R.(2020).Anxious, bored, and (maybe) missing out: Evaluation of anxiety attachment, boredom proneness, and fear of missing out (FoMO).Computers in Human Behavior, 112,106465.

Hair, Joseph F., Jr; Black, William C.; Babin, Barry J.; and Anderson, Rolph E. (2018) Multivariate Data Analysis, Eighth Edition. Hamphire, United Kingdom : CENGAGE INDIA. ISBN: 978-1-4737-5654-0.

Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). New York: Guilford Press.

Ahmad, S., Zulkurnain, N., & Khairushalimi, F. (2016). Assessing the Validity and Reliability of a Measurement Model in Structural Equation Modeling (SEM). Journal of Advances in Mathematics and Computer Science, 15(3), 1-8. https://doi.org/10.9734/BJMCS/2016/25183

Trujillo-Ortiz, A. and R. Hernandez-Walls. (2003). Mskekur: Mardia's multivariate skewness and kurtosis coefficients and its hypotheses testing. A MATLAB file. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3519

Mardia, K. V. (1970), Measures of multivariate skewnees and kurtosis with applications. Biometrika, 57(3):519-530. Mardia, K. V. (1974), Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. Sankhy A, 36:115-128.

Stevens, J. (1992), Applied Multivariate Statistics for Social Sciences. 2nd. ed. New-Jersey:Lawrance Erlbaum Associates Publishers. pp. 247-248.