Structural Equation Modeling in Criminal Justice Research

Travis M. Carter

Michigan State University

11/29/2021

What is Structural Equation Modeling (SEM)?

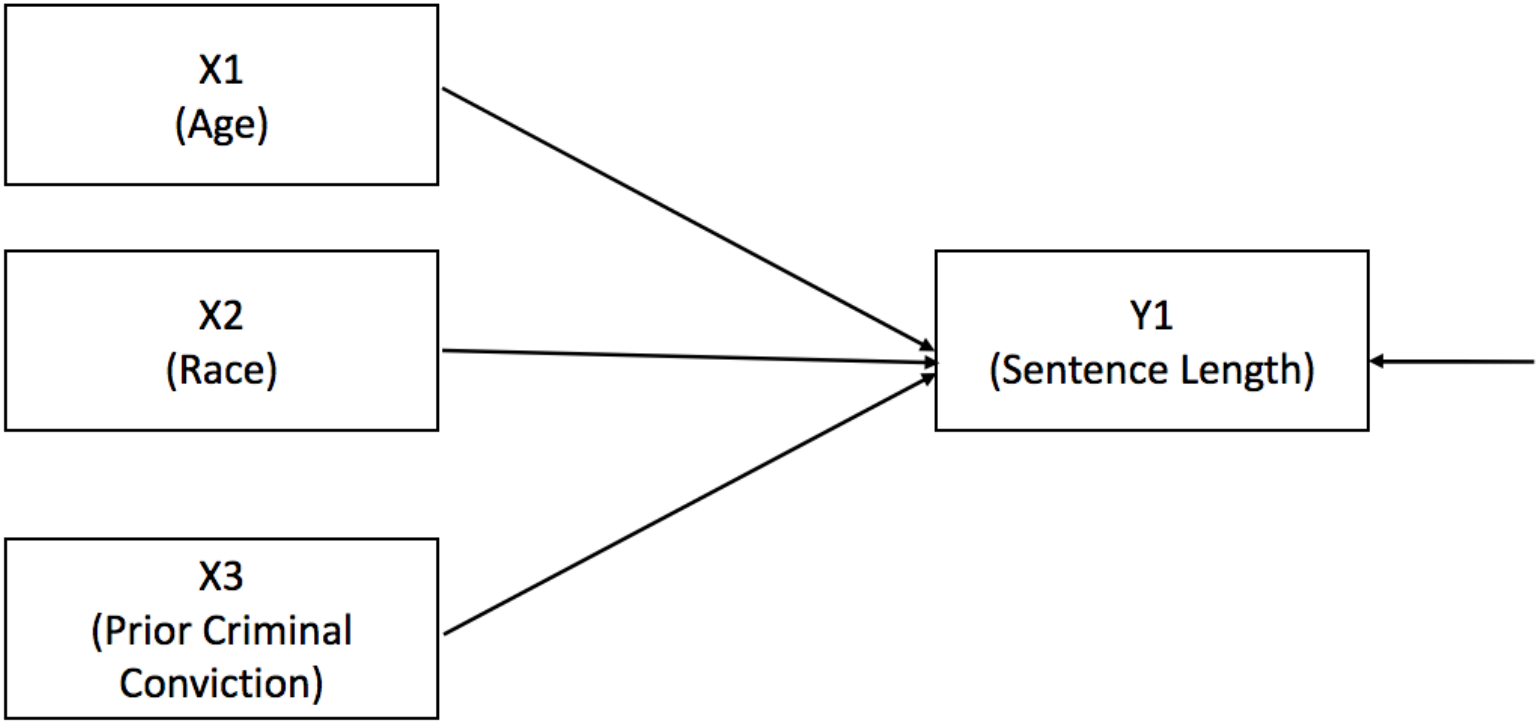
The basics

- SEM is an advanced statistical method used for testing relationships between *observed* and/or *latent* variables.
  - *Observed variables*: Directly measurable things (e.g., age, height, years in school)
  - *Latent variables*: Theoretical constructs measured indirectly via observed variables (e.g., low self-control, anxiety, social capital)

Types of SEM techniques

Path Analysis

- Path analysis (PA) is essentially univariate and multivariate regression analysis wiht only observed variables.



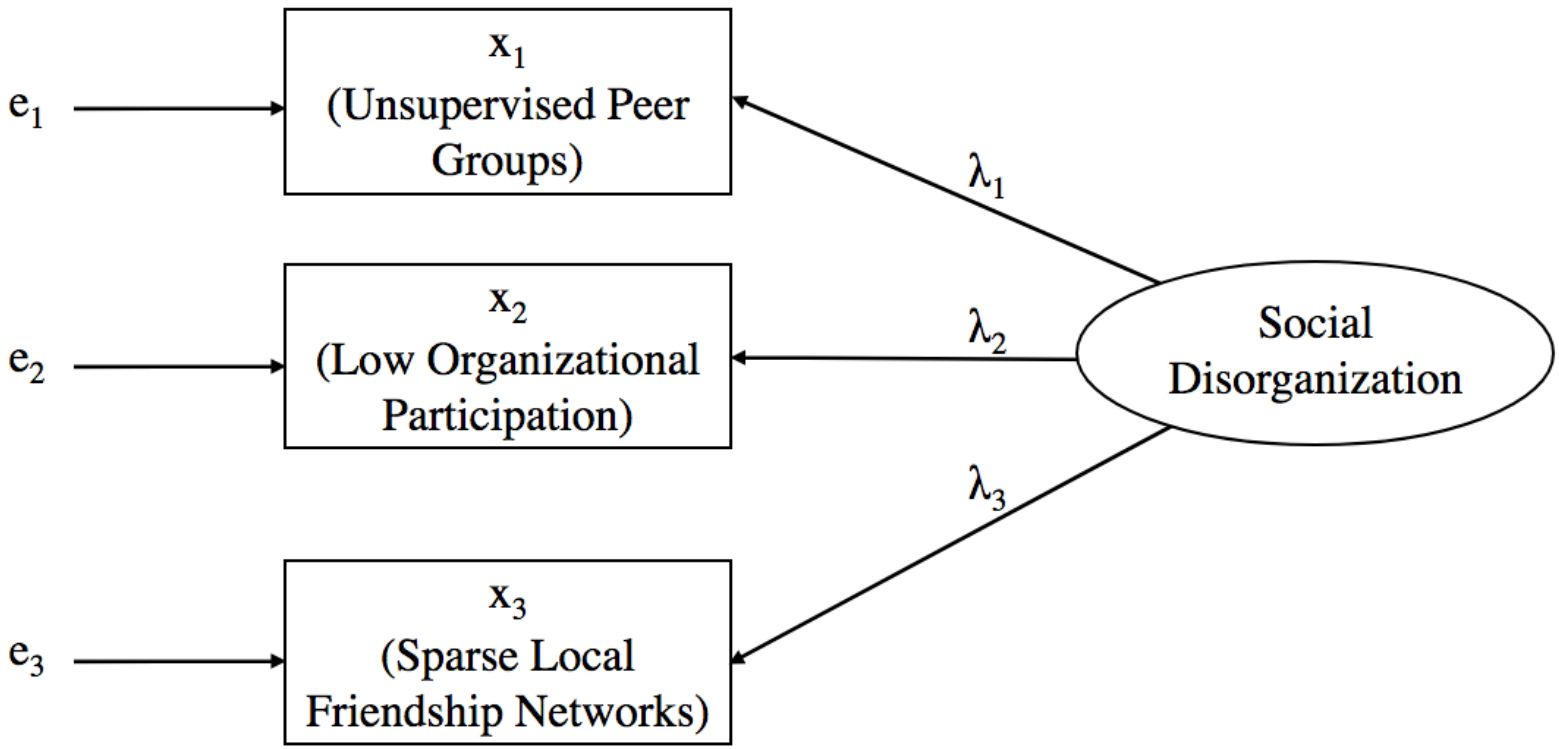
Exploratory and Confirmatory Factor Analysis

- Exploratory Factor Analysis (EFA) is a way to identify the number of potential *latent variables* (constructs) that are manifested through a set of *observed variables*.
- Check out this link to learn more about the nuts and bolts of EFA.[1]
- Check this link out to learn about what are the best practices for conducting EFA in your research.[2] in your research.

[1] https://scholar.google.com/scholar?hl=en&as_sdt=0%2C14&q=A+Beginner%E2%80%99s+Guide+to+Factor+Analysis%3A++Focusing+on+Exploratory+Factor+Analysis+&btnG=

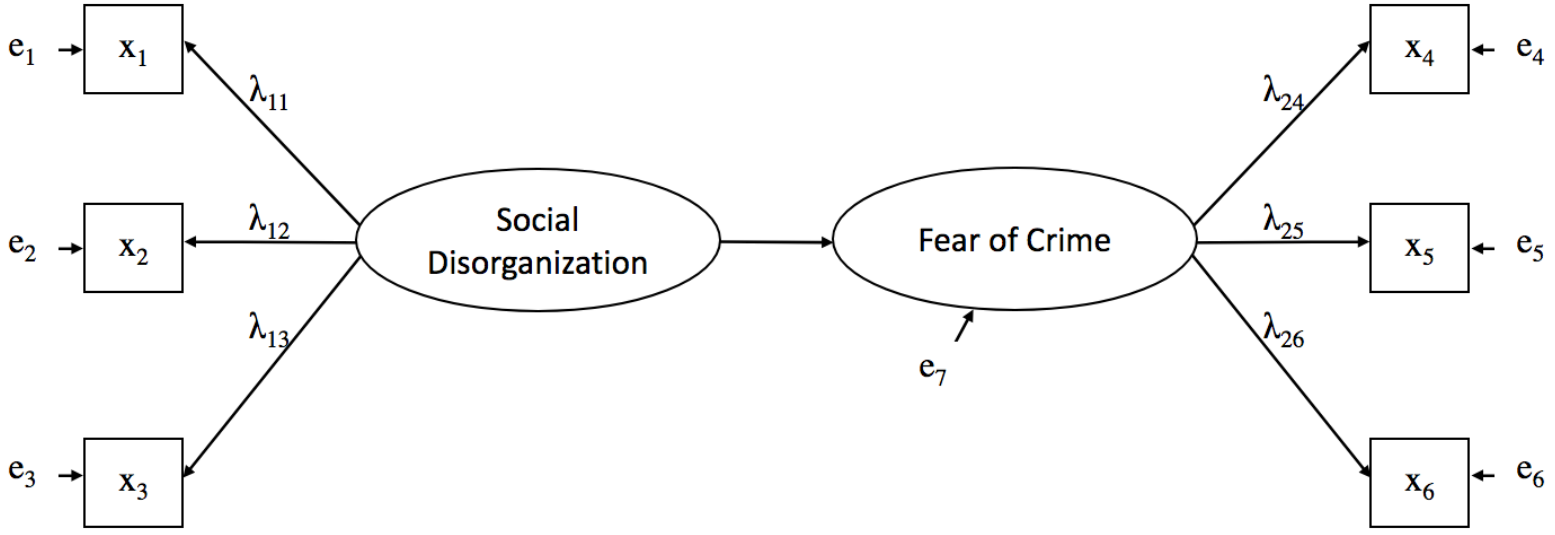
[2] https://scholar.google.com/scholar?hl=en&as_sdt=0%2C14&q=Best+practices+in+exploratory+factor+analysis%3A+four+recommendations+for+getting+the+most+from+your+analysis+recommendations+for+getting+the+most+from+your+analysis&btnG=

- Confirmatory Factor Analysis (CFA) takes what we know from EFA and **tests** whether these theoretical constructs are indeed manifestations (in the aggregate) of these observed variables.
- We do this by estimating relationships between observed measures that we think might collectively reflect some theoretical construct.



Structural Regression Models

- Structural Regression Models (SRM) take what we know about CFA and extends it by allowing us to model the magnitude and direction of relationships between latent variables.
- Anything you would do in a multivariate regression environment, you can do in SRM as well–including mediation and moderation!



Latent Change Analysis

- Latent Change Analysis (LCA) is a longitudinal extentions of CFA.
- We can use it to determine if latent constructs change over time within and between respondents (think growth modeling with IQ).
  - Do folks change in IQ over time within themselves?
  - Do folks differ in growth rates of IQ between each other?

Why is SEM important?

Why is SEM important?

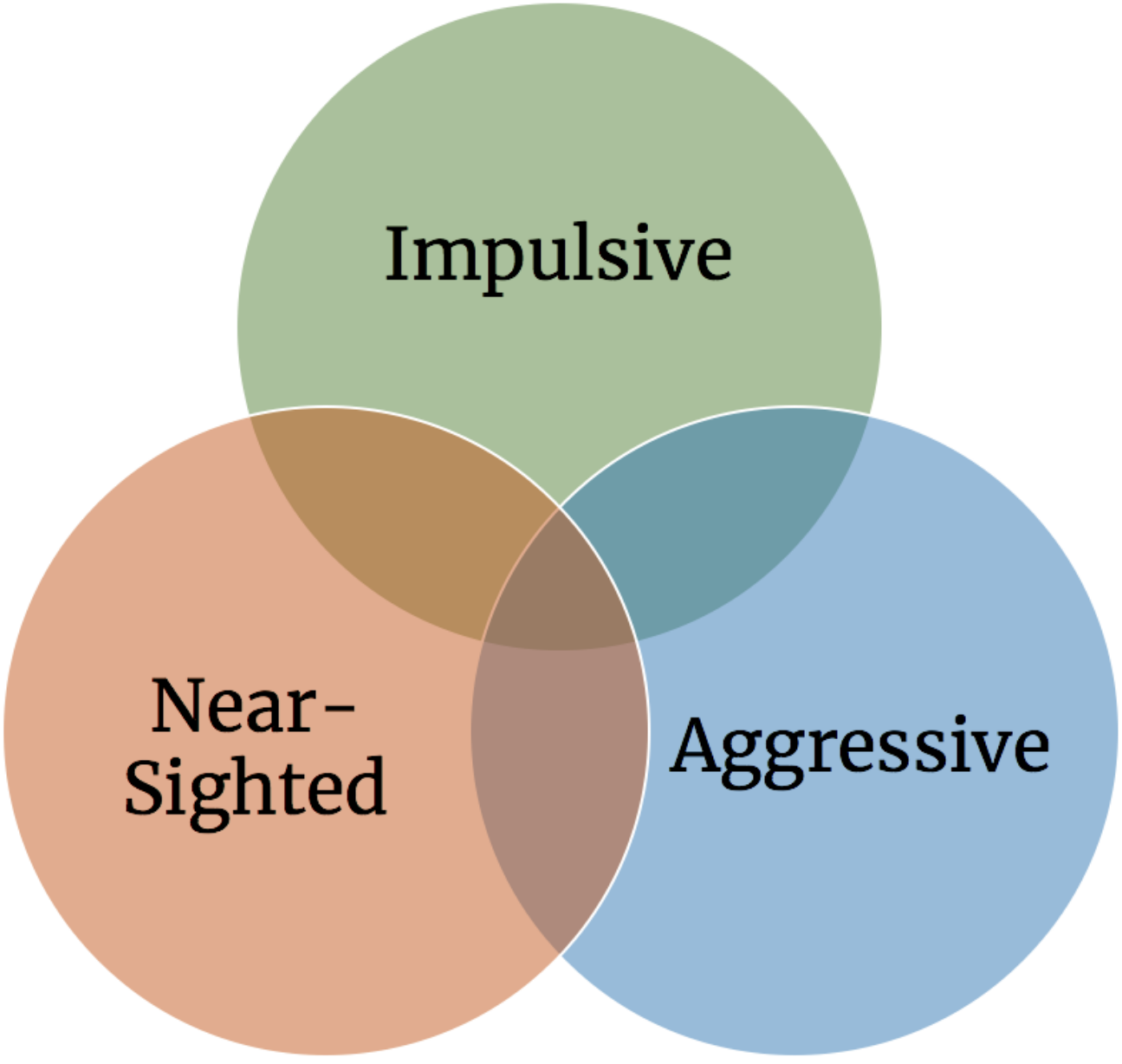
- Recall the general OLS regression equation we all know and love:

$$Y = \mu_1 + X_1\beta_1 + ...X_k\beta_k + \epsilon$$

- Here, μ is a constant, X is an observed (independent) variable, Y is our outcome variable, and ϵ is our error term.
- According to OLS assumptions, X is to be measured perfectly. If not, β will be inconsistent to the extent that X is unreliable.[3]

[3] Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling.* Routledge.

- Let's say X is an index scale based on 3 survey items, all of which tap into different elements low self-control.
- Each item, however, has some variance that is unshared by all other items.
- Multiply this by 3 and you have three times the amount of unshared variance that is captured in X but is entirely irrelevant for understanding the true form and function of low self-control.


Impulsive
Near-
Sighted
Aggressive

- SEM resolves this issue by allowing us to take into account **measurement error** in our theoretical variables of interest.[4]

[4] Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling.* Routledge.

How does SEM work?

How does SEM work?

- SEM is based on a system of equations, not a single equation like we see in regular OLS regression models.
- We determine whether a latent variable exists by regressing our observed indicators on that latent construct.
- We then evaluate fit indices to confirm it exists based on the data at our disposal.

- Remember that CFA from earlier? Here is what a model definition of equations might look like for a latent variable with 3 observed variables.

$$Item_1 = \mu_1 + \lambda_1 LSC + \epsilon_1$$

$$Item_2 = \mu_2 + \lambda_2 LSC + \epsilon_2$$

$$Item_3 = \mu_3 + \lambda_3 LSC + \epsilon_3$$

- Where μ is our constant, λ is our factor loading, LSC_1 is our latent variable, and our outcome is each observed variable tapping into low self-control.
- We have rules for determining which parameters are estimated in these equations for those interested in learning the nuts and bolts, **which I highly recommend** (See Supplementary materials at the back of my presentation).

How do we use SEM in CJ research: An example

How do we use SEM in CJ research: An example

- Say we have a sample of survey responses from a large victimization survey.
- We are interested in understanding whether people's perceived risk of being victimized drives their concern for personal safety.
- Here's a snapshot of our survey items.

Please indicate your level of agreement or disagreement with each of the following statements about the police in your neighborhood.

11. The police in my neighborhood treat people with dignity and respect.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

12. The police in my neighborhood take time to listen to people.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

13. The police in my neighborhood explain their decisions to people they deal with.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

For the next set of questions, please indicate how afraid you are of each happening to you in your neighborhood.

14. How afraid are you of someone breaking into your house while you are home?
- ☐ Not at all Afraid
- ☐ Somewhat Afraid
- ☐ Afraid
- ☐ Very Afraid

15. How afraid are you of someone robbing you with a gun or knife in your neighborhood?
- ☐ Not at all Afraid
- ☐ Somewhat Afraid
- ☐ Afraid
- ☐ Very Afraid

16. How afraid are you of someone assaulting you in your neighborhood?
- ☐ Not at all Afraid
- ☐ Somewhat Afraid
- ☐ Afraid
- ☐ Very Afraid

Please indicate your level of agreement or disagreement with each of the following statements about the law enforcement in your neighborhood.

17. I have a great deal of respect for the police.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

18. I feel proud of the police.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

19. Overall, the police are honest.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

20. The police enforce laws consistently when dealing with all people in my neighborhood.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

21. People in my neighborhood are likely to call the police to report an accident.
- ☐ Strongly Agree
- ☐ Somewhat Agree
- ☐ Neither Agree nor Disagree
- ☐ Somewhat Disagree
- ☐ Strongly Disagree

Start Here

- Please use a black or blue pen to complete this form.
- Mark ☒ to indicate your answer. If you want to change your answer, darken the box ■ and mark the correct answer.

Your Community

1. In general, how do you rate your neighborhood as a place to live? Would you say:
   - ☐ Excellent
   - ☐ Good
   - ☐ Fair
   - ☐ Poor

Please indicate your level of agreement or disagreement with each of the following statements about your neighborhood.

2. People that live in my neighborhood are generally friendly.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

3. I am happy I live in this neighborhood.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

4. People around here take care of each other.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

5. People in this neighborhood can be trusted.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

6. People around here are willing to help their neighbors.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

7. This is a close-knit neighborhood.
   - ☐ Strongly Agree
   - ☐ Somewhat Agree
   - ☐ Neither Agree nor Disagree
   - ☐ Somewhat Disagree
   - ☐ Strongly Disagree

For the next set of questions, please indicate how likely you think each might happened to you in your neighborhood.

8. How likely do you think it would be for someone to break into your house while you are home?
   - ☐ Not at all Likely
   - ☐ Somewhat Likely
   - ☐ Likely
   - ☐ Very Likely

9. How likely do you think it is that someone who has a gun or knife would try to rob you in your neighborhood?
   - ☐ Not at all Likely
   - ☐ Somewhat Likely
   - ☐ Likely
   - ☐ Very Likely

10. How likely do you think it is that someone will assault you in your neighborhood?
    - ☐ Not at all Likely
    - ☐ Somewhat Likely
    - ☐ Likely
    - ☐ Very Likely

EFA

- First, let's say we were interested in determining whether these items hang together, or if they tap into different constructs.
  - Let's run an EFA on all six of our survey items.

EFA R output: Factor Loadings

```
TABH <- psych::fa(FOC_AND_ROV,
                  nfactors = 2,
                  rotate = "oblimin")
```

Table 1: EFA Results: Oblique Rotated Factor Loadings

	1-Factor Model	2-Factor Model
DTA.FV1	0.7805331	0.0308348
DTA.FV2	0.9653352	-0.0346979
DTA.FV3	0.9007149	0.0288003
DTA.PR1	0.1150691	0.5565472
DTA.PR2	-0.0307854	0.9267931
DTA.PR3	0.0154627	0.8644993

EFA R output: Eigenvalues

```
TABH <- psych::fa(FOC_AND_ROV, nfactors = 2,
                  rotate = "oblimin")
```

Table 2: EFA Results: Eigenvalues and Variances

	1-Factor Model	2-Factor Model
SS loadings	2.41	1.97
Proportion Var	0.40	0.33
Cumulative Var	0.40	0.73
Proportion Explained	0.55	0.45
Cumulative Proportion	0.55	1.00

CFA

- OK, looks like they fall into diff constructs based on rotated factor loadings... what if we want to confirm whether the data fits this well? Enter in CFA.

CFA R output: Create the model

- The first step in developing our CFA is to write out our model definition of equations.

```
attach(DTA)
MEASUREMENT_MODEL <- '
# LATENT VARIABLE DEFINITION OF EQS
  FEAR       =~ FV1 + FV2 + FV3
  RISK       =~ PR1 + PR2 + PR3
'
```

CFA R output: Estimate the model

```
fit_MM <- cfa(MEASUREMENT_MODEL,
              data = DTA,
              ordered = T)
```

CFA R output: How to evaluate fit of model

- According to Raykov and Marcoulides (2006) A good fitting model has the following.[5]
  - An RMSEA value <= .05
  - Alternative for RMSEA: The lower bound of the 90% confidence interval is <= .05 and its upper bound is <= .08.
  - CFI and TLI are >= .95 (>= .90 for acceptable fit)
  - A non-significant χ^2 value (this is biased towards the null in large samples).
  - No negative residual variances in our model (more on this later).

[5] Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling.* Routledge.

CFA R output: Now let's evaluate the fit of our model

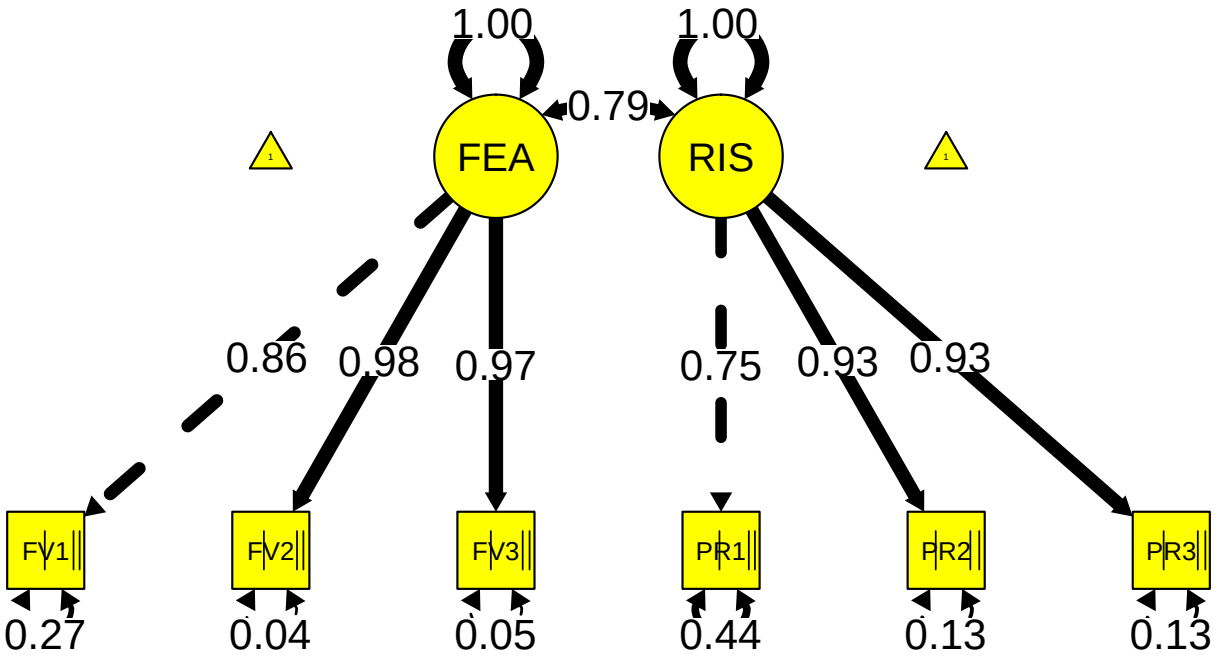
```
FIT_INDICES_MM <- fitMeasures(fit_MM,
                              c("chi" , "rmsea",
                                "rmsea.ci.upper",
                                "rmsea.ci.lower",
                                "tli","cfi"))
```

Table 3: Fit Indices for Measurement Model

	Fit Stats
rmsea	0.093
rmsea.ci.upper	0.104
rmsea.ci.lower	0.082
tli	0.999
cfi	0.999

CFA R output: Make a plot of our results

```
semPaths(fit_MM, residuals = T,
         "std", fade = FALSE, posCol = c("black"),
         color = c("yellow"),
         edge.label.cex = 1.25, edge.label.font = 1.25)
```



SRM

- Ok, looks like we are dealing with 2 different constructs... potentially. Without getting into the weeds, let's see if risk has any predictive influence on fear.

- Here is the input for our model:

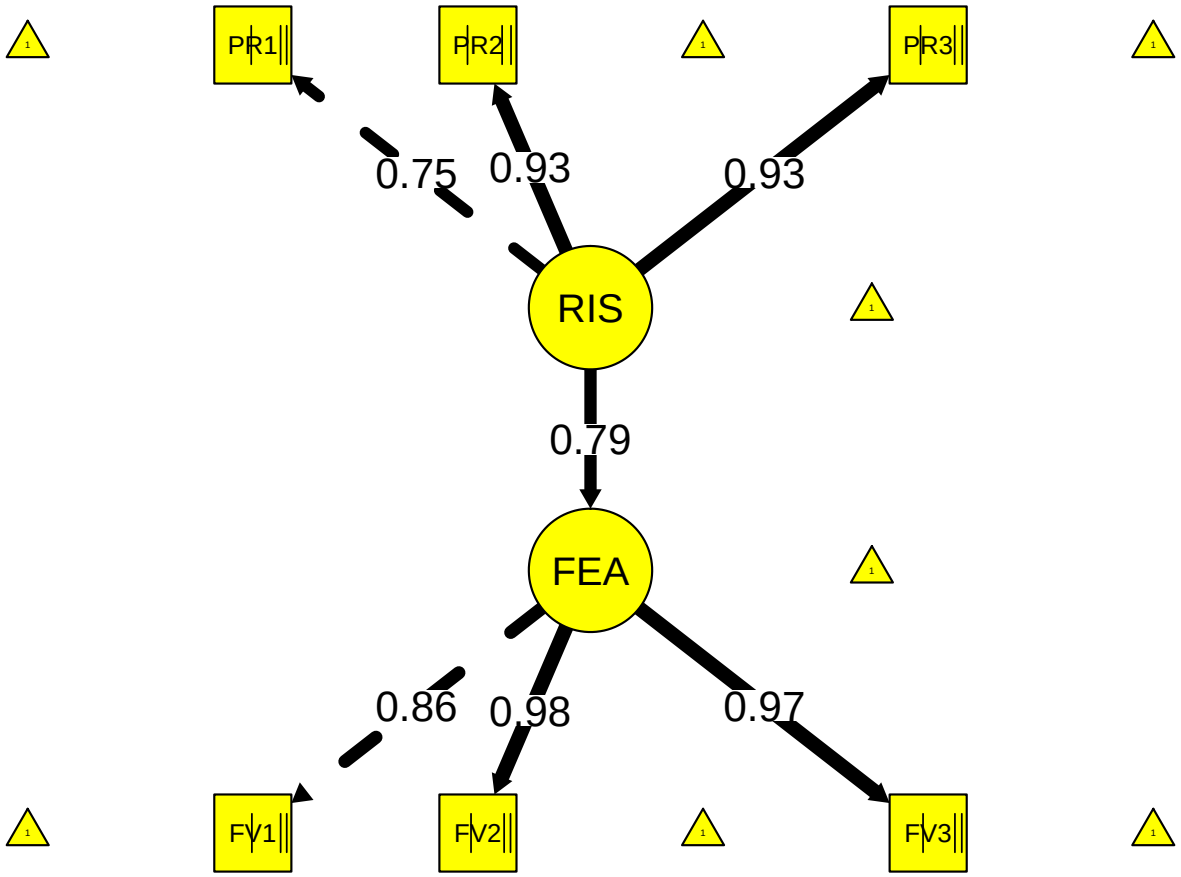
```
STRUCTURAL_MODEL <- '
# LATENT VARIABLE DEFINITION OF EQS
  FEAR       =~ FV1 + FV2 + FV3
  RISK       =~ PR1 + PR2 + PR3

# STRUCTURAL REGRESSION MODEL
  FEAR        ~ RISK
'
fit_SM <- sem(STRUCTURAL_MODEL,
              data = DTA,
              ordered = T)
```

Here is the fit of our model. Do we like what we see?

Table 4: Fit Indices for Structural Model

	Fit Stats
rmsea	0.093
rmsea.ci.upper	0.104
rmsea.ci.lower	0.082
tli	0.999
cfi	0.999

Here is a figure for our model.



Here is the output for our SRM.

Table 5: Structural Model Results

LHS	OPERATOR	RHS	EXO	Est.	S.E.	Z Stat.	p
FEAR	=~	FV1	0	1.00	0.00	NA	NA
FEAR	=~	FV2	0	1.14	0.01	130.32	0
FEAR	=~	FV3	0	1.14	0.01	137.86	0
RISK	=~	PR1	0	1.00	0.00	NA	NA
RISK	=~	PR2	0	1.24	0.02	67.03	0
RISK	=~	PR3	0	1.24	0.02	68.16	0
FEAR	~	RISK	0	0.90	0.02	57.07	0

SEM in CJ wrap-up

SEM in CJ wrap-up

- What is SEM?
- Why do we use SEM?
- What applications does SEM have in our research?
- What if you want to learn more?
  - Books:
    - Raykov, T., & Marcoulides, G. A. (2012). *A first course in structural equation modeling*. Routledge.
    - Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
  - Websites:
    - Lavaan resource page **https://lavaan.ugent.be/**
    - UCLA IDRE resource page **https://stats.idre.ucla.edu/r/seminars/rsem/**

- Articles:
  - Jacinta Gau has a great best practices article
    - Gau, J. M. (2010). Basic principles and practices of structural equation modeling in criminal justice and criminology research. *Journal of Criminal Justice Education*, 21(2), 136-151.
  - *Lavaan* Package article
    - Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *Journal of Statistical Software*, 48(2), 1-36.
- Researchers:
  - Dr. Joe Hamm
  - Dr. George Burruss
  - Dr. Johnathan Jackson
- Courses @ MSU:
  - **CEP 938** *Latent Variable Modeling* by Dr. Tenko Raykov
  - **SOC 883** *Multi-Equation Quantitative Models* by Dr. Sandra Marquart-Pyatt
  - **HDFS 961** *Applied Structural Equation Modeling* by Dr. Amy Nuttall
  - **HDFS 962** *Longitudinal Structural Equation Modeling* by Dr. Amy Nuttall

ICPSR

ICPSR

- One alternative to taking a course here at MSU is to take one through the ICPSR summer program.
  - https://www.icpsr.umich.edu/web/pages/sumprog/
- They host tons of courses ranging in methodological and quantitative topics
  - Regression Analysis 1-3
  - Generalized Linear Models 1-2
  - Measurement, scaling, etc.
  - Time Series Analysis 1-2
  - Multilevel Modeling 1-2
  - Bayesian Stats
  - Network Analysis
- Great use of your summer fellowship!

Thanks for listening!

Thanks for listening!

- You can find me on these platforms
  - **Twitter**: https://twitter.com/tcarter_MSU
  - **GitHub**: https://github.com/carte475
  - **SCJ Directory**: https://cj.msu.edu/directory/carter-travis.html
- If you have any further questions or want to meet up and nerd-out about stats, here is my email: carte475@msu.edu

Supplemental materials

How do we identify a model in SEM?

- In OLS regression, we generally need not worry about identifying a model (specifying which parameters to be estimated) apart from our exogenous (independent) and endogenous (dependent) variables.
- In SEM, we must carefully consider what is and is not to be estimated.
- Remember our CFA model with 3 observed variables and 1 latent variable? Here it is again as written out in the form of equations.

$$Item_1 = \mu_1 + \lambda_1 LSC + \epsilon_1$$

$$Item_2 = \mu_2 + \lambda_2 LSC + \epsilon_2$$

$$Item_3 = \mu_3 + \lambda_3 LSC + \epsilon_3$$

- Remember, μ is our constant, λ is our factor loading, LSC_1 is our latent variable, and our outcome is each observed variable tapping into low self-control.

- In this structural equation model, we have 3 equations as opposed to one equation to worry about.
  - How? We have three outcome variables based on the three survey items tapping into low self-control.
  - **Note**: In SEM, our observed variables tapping into the latent variable are considered endogenous.
- Based on this information, we need to determine which relationships between observed and unobserved variables will be estimated via the parameters in our model and which and which will not be estimated.
  - Why not estimate them all? We have only so many observed variables to then determine potential relationships with latent variables! They are a finite resource!
- **QUESTION**: How many exogenous and endogenous variables are in the model above?

- **ANSWER**:
  - We have 6 exogenous variables (3 error terms + 3 latent variable factor loadings = 6).
  - We have 3 endogenous variables (3 observed items) and 3 constants.
- Fun fact: In SEM, our constants (μ) are the observed item means across our entire sample.
  - If $Item_1$ measured respondent aggressiveness and the mean score across our sample for that variable was (out of 10) $\bar{x} = 3.5$, then $\mu_1 = 3.5$.
- Fun fact 2: In SEM, our residuals are independent variables.
  - A residual captures the amount of variation in an observed variable that is due to measurement error.
  - In other words, a residual is the amount of item variance *unshared* with all other measures of a common factor (i.e., LSC)

- Now let's figure out how to identify a model based on Raykov and Marcoulides (2006) 6 rules for determining model parameters.

1. All variances of our exog. variables are model parameters.
2. All covariances between exog. variables are model parameters.
3. All factor loadings (λ) are model parameters.
4. All regression coeffs are model parameters.
5. The variances and covariances between and of our endog. variables (and between our exog. and endog. variables) are not model parameters.
   5.1 Why? B/c these variances and covariances are explained by model parameters already.
6. For each latent variable, we must set it's scale to identify a model.
   6.1 We can do this by "fixing" one of it's factor loadings to a constant such as 1, or we can set it's variance to 1.

Based on what we now know, let's go about determining the number of estimable parameters in our model. Given our rules... how many do we need to estimate?

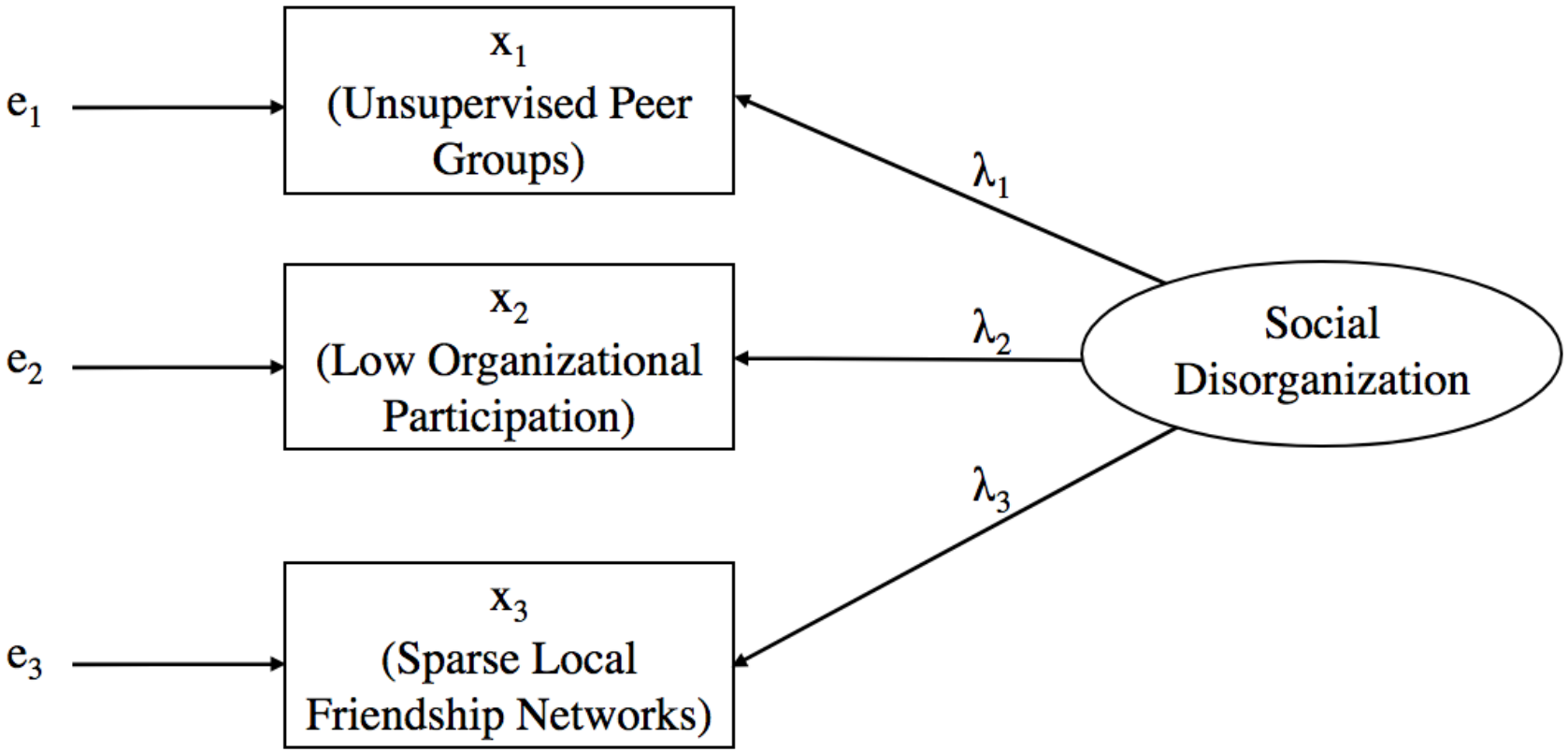


Figure 4: CFA Example.

Let's go through the rules.

- Rule 1 says we should estimate 4 parameters.
- Rule 2 says we should estimate 0 parameters.
  - Why? It is assumed our residuals are uncorrelated! Also, we have only one latent variable, if there were 2 or more, then we would estimate covariance parameters.
- Rule 3 says we should estimate 3 parameters.
- Rule 4 says we should estimate 0 parameters.
  - Why? No regression coeffs. This is a CFA, not a PA or SRM.
- Rule 5 is acknowledged.
- Rule 6 says we should subtract 1 parameter from the total number of estimable parameters.
  - We can do this by fixing the variance of Social Disorganization or by fixing one of it's three factor loadings to $\lambda = 1$.

This gives us a total of 6 "free" parameters to be freely estimated.

How do we estimate model parameters in SEM?

- Without getting into the nitty-gritty details, parameter estimation in SEM can be **roughly** defined through the following steps.

1. Identify your model by determining which parameters are free (estimable) and which are not. We just did this!
2. Create a matrix of the covariance relationships between your model parameters to be estimated.
3. Rinse and repeat a process of defining the values for those relationships based on how well they fit the observed data at your disposal.
   3.1 The process of estimation is usually done through Maximum Likelihood expectation, which explains the iterative estimation process described above.

4. Upon finding a model that minimizes the "distance" between our observed matrix of data and the specified matrix of relationships between parameters, we then use goodness-of-fit (GOF) indices to determine which values and models fit best to our data.

A note on estimating SEM parameters.

We mentioned one cannot estimate all parameters in a model because we have so many degrees of freedom to give.

Here is a further note on what that really means.

- In our CFA model example, we have 3-observed variables and 1-latent variable.
- Based on this information, we have approximately 3 non-redundant elements (observed variables) $p = 3$.
- The reason we care about non-redundant elements is that we cannot estimate more relationships than it is humanly possible given the number of observed variables in our proposed model.
- To determine the max number number of relationships possibly estimable, we use the following formula $p(p+1)/2$.

- This tells us that, if $p = 3$, then we can estimate at most $3(3 + 1)/2 = 6$ six parameters. To get our degrees of freedom we simply subtract this from our model parameters ($q = 6$), which means we have 0 degrees of freedom.
- When $df >= 0$ we have an identified model!
- If $df < 0$ we have too many parameters and therefore cannot identify a unique solution for our estimated parameters. That is like having more questions than you have answers!
- Solution? You probably goofed up determining the model parameters (p) and should carefully go through the steps outlined by Raykov and Marcoulides (2006).