

SEM, Revealed.

Modeling connections with latent variables and regression pathways.



Slides available at <https://tinyurl.com/CHBD-slides24>

PDF slides at <https://tinyurl.com/CHBD-pdf24>

Waylon Howard | Biostatistician @
BEAR/CHBD



was founded with a collaborative spirit and lofty objective: to deliver comprehensive data management, advanced analytics, and expert statistical, epidemiologic, and qualitative support...

By alleviating external consultants, designing studies in-house, and engaging investigators directly, we're able to provide responsive, efficient, and high-quality data support at a fraction of the going cost.

Biostatistics @ BEAR

Supports stakeholders throughout SCRI by helping them make better decisions using data

Topics

- How to measure it?
- Fitting a CFA model.
- Estimators.
- Model Fit.
- Statistical Code.
- Annotated Results.

Perceived Social Support

How to measure it?

Self-report questions

1. My friends really try to help me.
 2. I can count on my friends when things go wrong.
 3. I can talk about my problems with my friends.
-

Very Strongly Disagree	Strongly Disagree	Mildly Disagree	Neutral	Mildly Agree	Strongly Agree	Very Strongly Agree
1	2	3	4	5	6	7

Higher scores = More Perceived Social Support

1. My friends really try to help me.

Very Strongly Disagree	Strongly Disagree	Mildly Disagree	Neutral	Mildly Agree	Strongly Agree	Very Strongly Agree
1	2	3	4	5	6	7

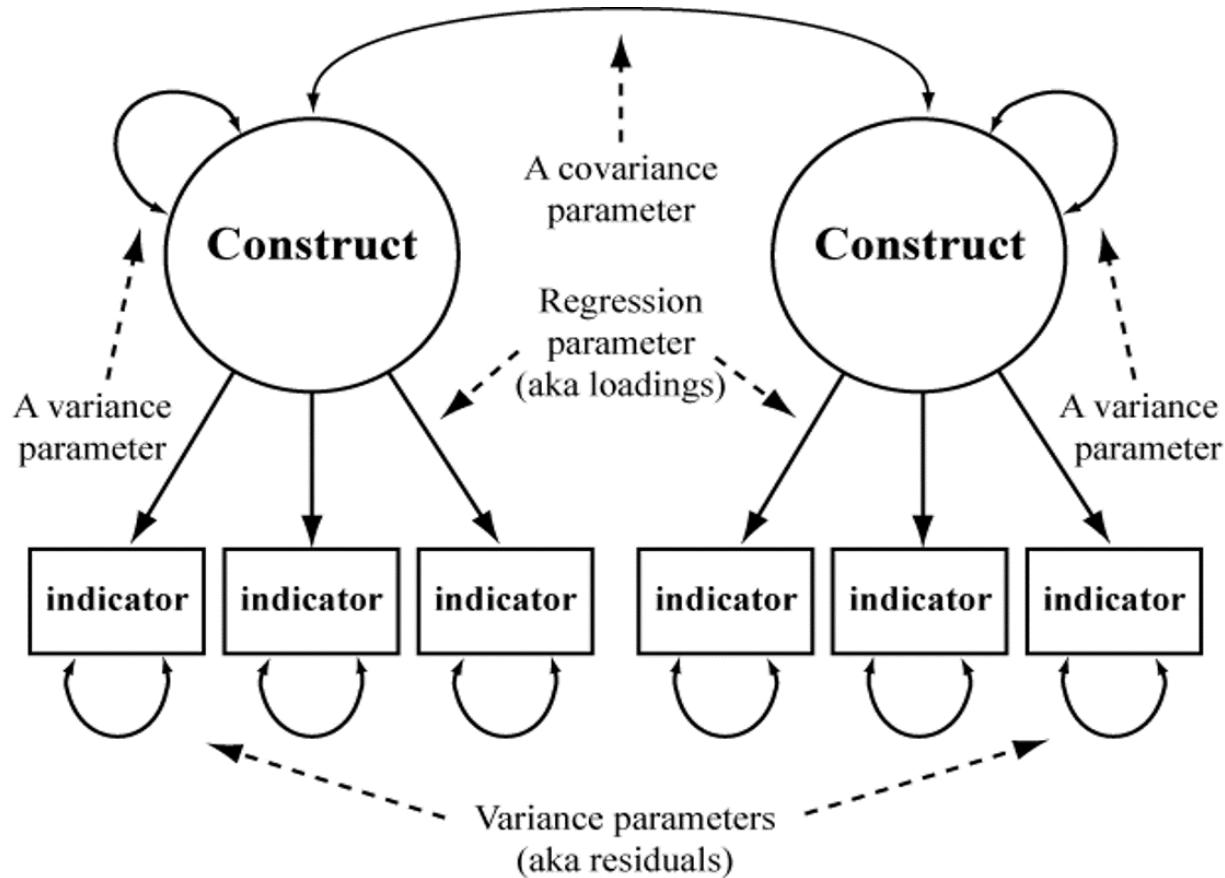
$$X_i = T_i + (S_i + e_i)$$

T_i is the 'true' score

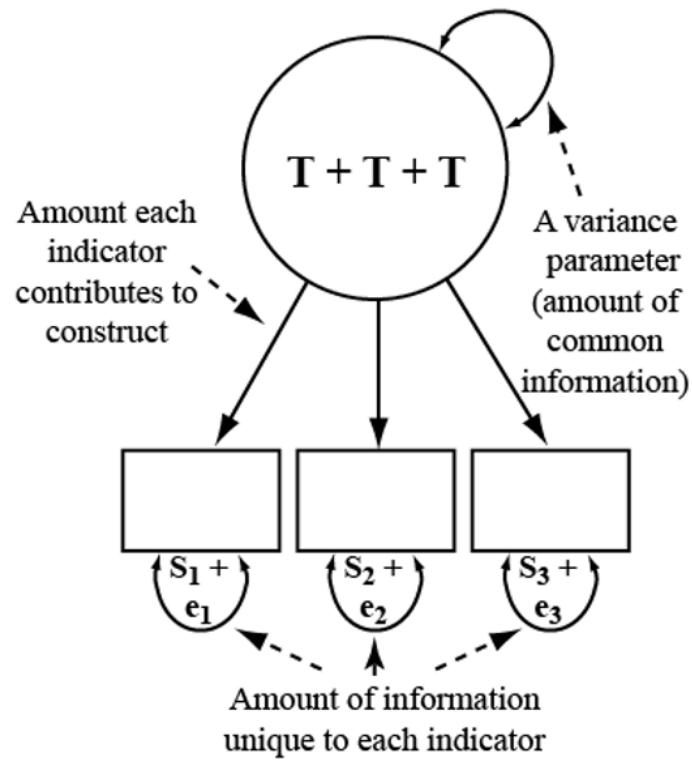
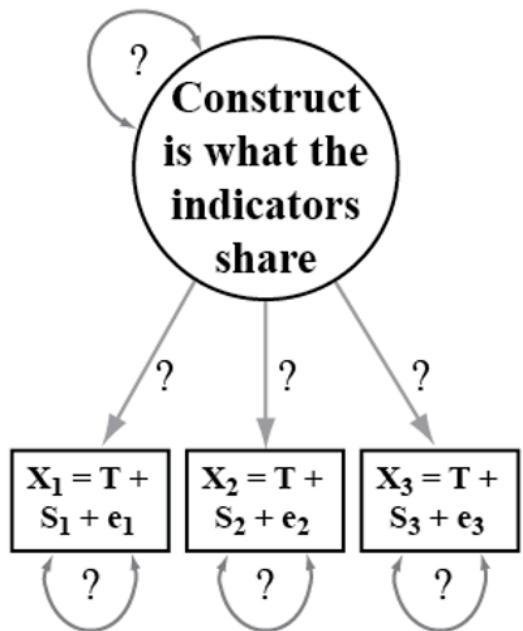
S_i is item-specific, yet reliable

e_i is random error, or noise

Use the scoring procedure: No measurement error (*always perfectly measured*), uniform effectiveness of items (*equal items*), invariance across groups and time.

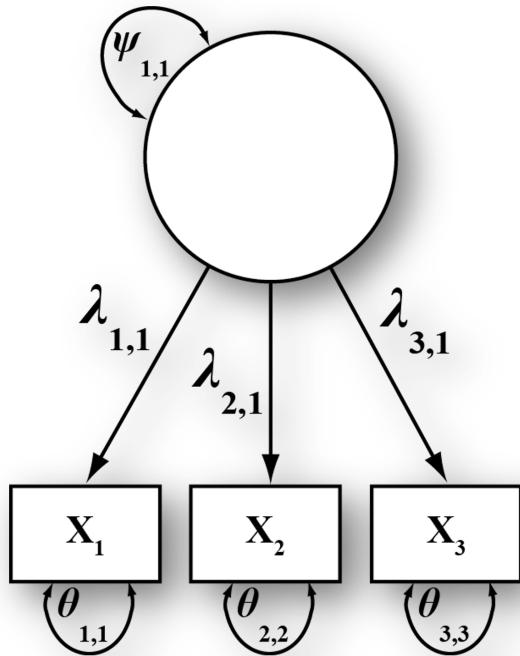


$$X_i = T_i + (S_i + e_i)$$



How do we get the true score?

Fitting a CFA model.



Estimated Parameters: 7

Observed Information: 6

Matrix Formula:

$$\Sigma = \Lambda \Psi \Lambda' + \Theta$$

Σ = Variance/Covariance Matrix

	X1	X2	X3
X1	5.66		
X2	4.90	5.50	
X3	4.33	4.38	5.63

Implied Variance/Covariance Matrix

	X1	X2	X3
X1	$\lambda_{11} \psi_{11} \lambda_{11} + \theta_{11}$		
X2	$\lambda_{11} \psi_{11} \lambda_{21}$	$\lambda_{21} \psi_{11} \lambda_{21} + \theta_{22}$	
X3	$\lambda_{11} \psi_{11} \lambda_{31}$	$\lambda_{21} \psi_{11} \lambda_{31}$	$\lambda_{31} \psi_{11} \lambda_{31} + \theta_{33}$

Underidentified: $X + Y = 20$

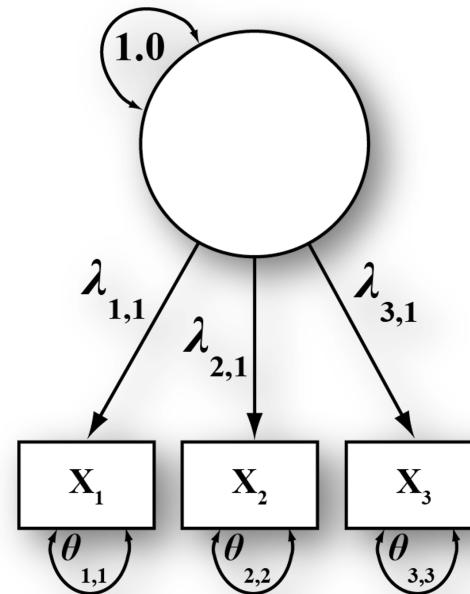
Variance /Covariance Matrix

	X1	X2	X3
X1	5.66		
X2	4.90	5.50	
X3	4.33	4.38	5.63

Just Identified.

	X1	X2	X3
X1	$\lambda_{11} \lambda_{11} + \theta_{11}$		
X2	$\lambda_{11} \lambda_{21}$	$\lambda_{21} \lambda_{21} + \theta_{22}$	
X3	$\lambda_{11} \lambda_{31}$	$\lambda_{21} \lambda_{31}$	$\lambda_{31} \lambda_{31} + \theta_{33}$

Fix the latent variance to 1.0



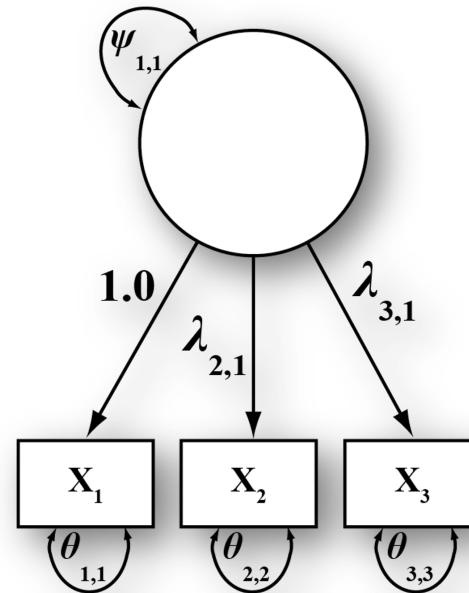
Variance /Covariance Matrix

	X1	X2	X3
X1	5.66		
X2	4.90	5.50	
X3	4.33	4.38	5.63

Just Identified.

	X1	X2	X3
X1	$\psi_{11} + \theta_{11}$		
X2	$\psi_{11} \lambda_{21}$	$\lambda_{21} \psi_{11} \lambda_{21} + \theta_{22}$	
X3	$\psi_{11} \lambda_{31}$	$\lambda_{21} \psi_{11} \lambda_{31}$	$\lambda_{31} \psi_{11} \lambda_{31} + \theta_{33}$

Fix the loading to 1.0



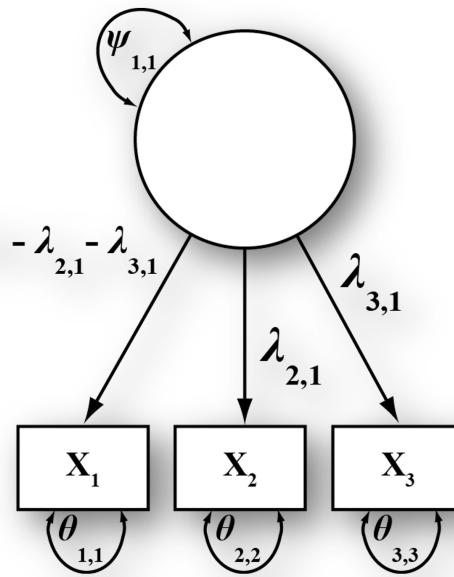
Variance /Covariance Matrix

	X1	X2	X3
X1	5.66		
X2	4.90	5.50	
X3	4.33	4.38	5.63

Just Identified.

	X1	X2	X3
X1	$(3 - \lambda_{21} - \lambda_{31})$ $\psi_{11} (3 - \lambda_{21} - \lambda_{31}) + \theta_{11}$		
X2	$(3 - \lambda_{21} - \lambda_{31})$ $\psi_{11} \lambda_{21}$	$\lambda_{21} \psi_{11} \lambda_{21} + \theta_{22}$	
X3	$(3 - \lambda_{21} - \lambda_{31})\psi_{11}$ λ_{31}	$\lambda_{21} \psi_{11} \lambda_{31}$	$\lambda_{31} \psi_{11} \lambda_{31} + \theta_{33}$

Constrain loading to average 1.0

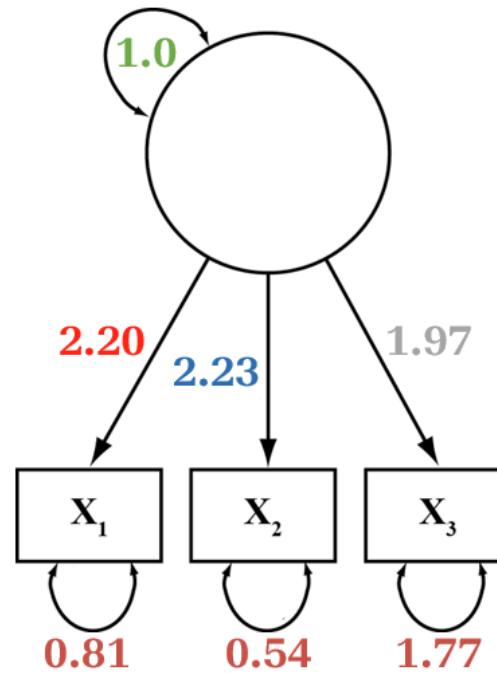


$$\lambda_{1,1} = 3 - \lambda_{2,1} - \lambda_{3,1}$$

Variance /Covariance Matrix

	X1	X2	X3
X1	5.66		
X2	4.90	5.50	
X3	4.33	4.38	5.63

	X1	X2	X3
X1	$2.20 * 1.0 *$ $2.20 + 0.81$ $= 5.66$		
X2	$2.20 * 1.0 *$ $2.23 = 4.90$	$2.23 * 1.0 *$ $2.23 + 0.54$ $= 5.50$	
X3	$2.20 * 1.0 *$ $1.97 = 4.33$	$2.23 * 1.0 *$ $1.97 = 4.38$	$1.97 * 1.0 *$ $1.97 + 1.77$ $= 5.63$



How do we get the numbers?
Estimators.

Maximum Likelihood

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \mu)^2}{\sigma^2}}$$

ML identifies the population parameters that are most likely given the observed data

A likelihood (or log likelihood) function quantifies the fit of the data to the parameters

ML requires a population distribution (normal)

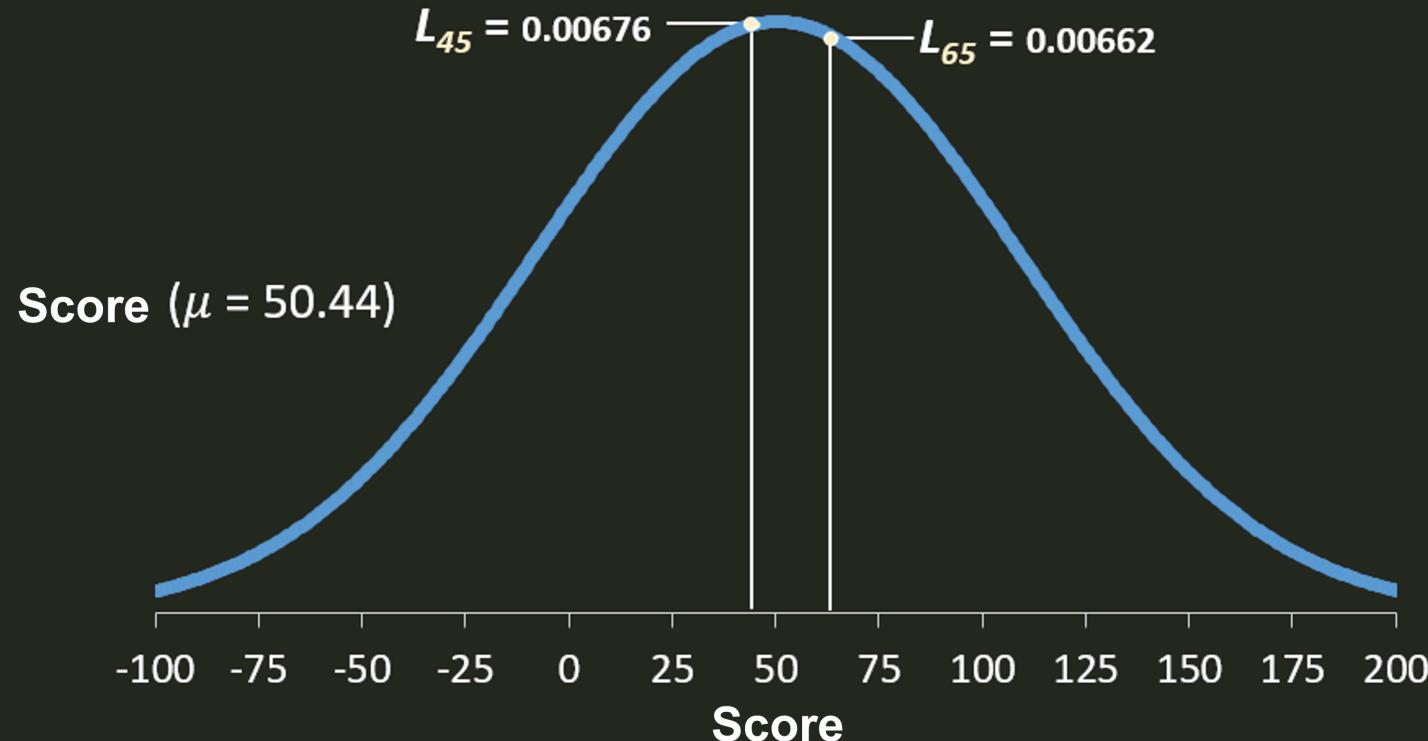
$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-.5 \frac{(Y_i - \mu)^2}{\sigma^2}]}$$

Applying the density function gives the relative probability (L_i) of each score from this normal distribution.

Score ($\mu = 50.44, \sigma = 58.68$)

Person ID	Score	Likelihood
1	36.6	0.00661212
2	41.8	0.006725313
3	42.6	0.006738201
4	43.1	0.006745631
5	43.4	0.006749858
6	44.2	0.006760279
7	44.9	0.006768379
8	46.3	0.006781711
9	48.6	0.006795269
10	49.0	0.006796563
11	50.0	0.006798419
12	51.6	0.006797282
13	54.6	0.006781547
14	54.8	0.00677987
15	55.7	0.006771351
16	57.2	0.006753646
17	57.6	0.006748188
18	60.3	0.006703308
19	60.9	0.006691451
20	65.3	0.006584073

Largest relative probability is for ID 11



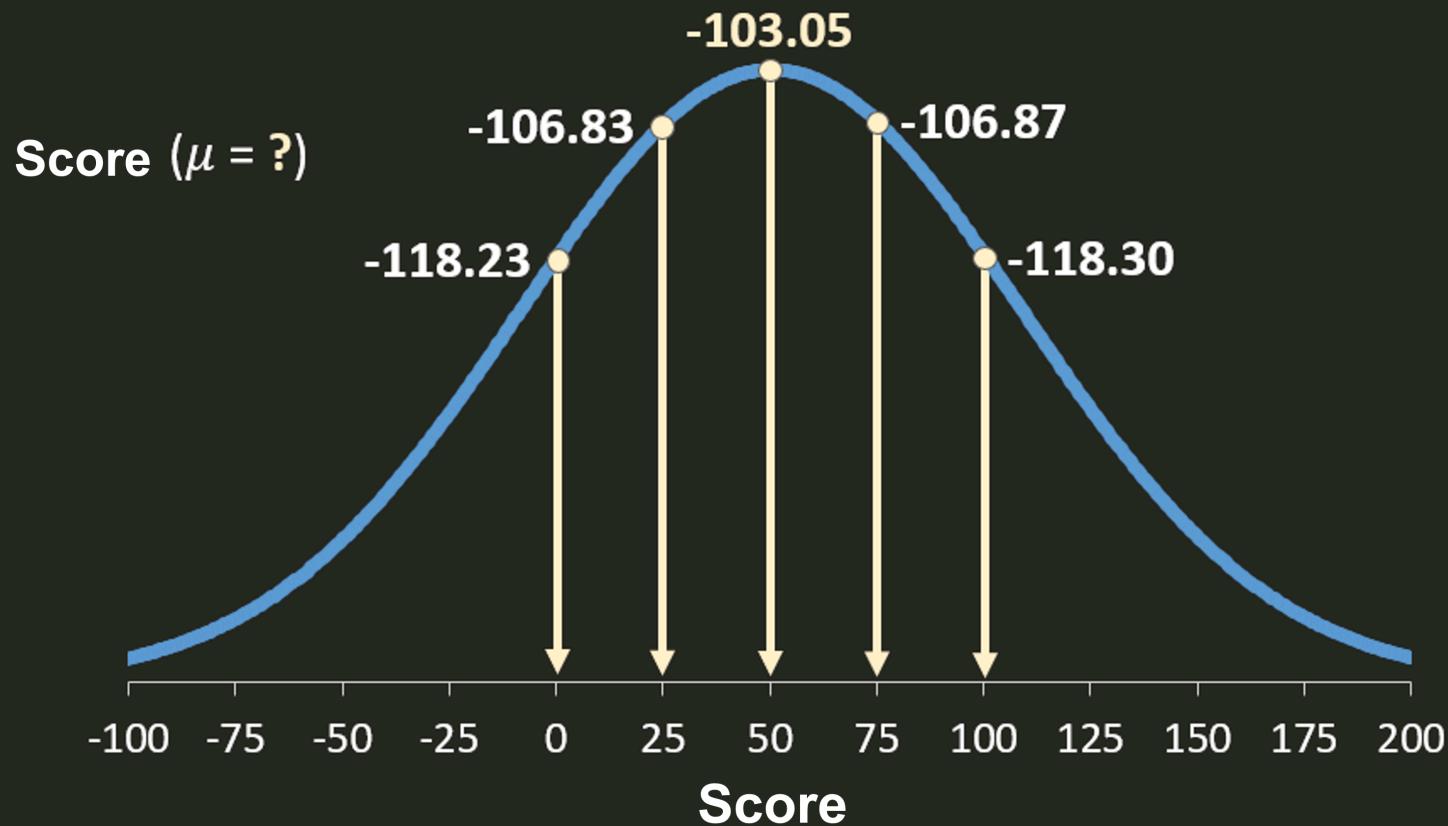
Multiply each L_i to get
sample likelihood.

To avoid small numbers, we add the log of the likelihood.

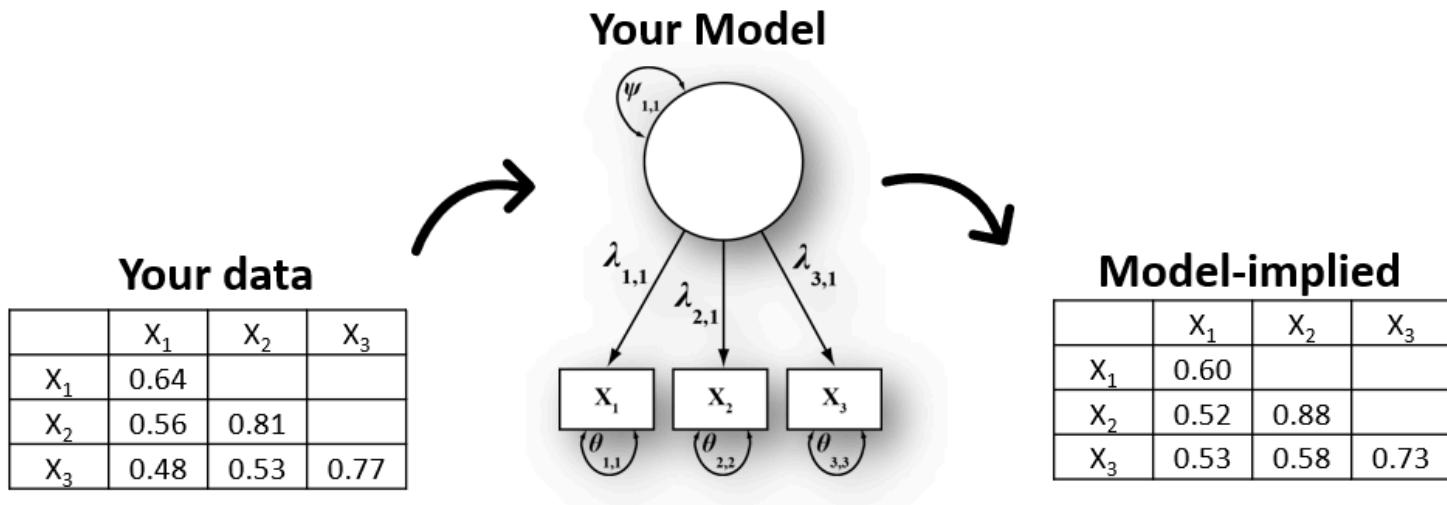
-99.9824

Person ID	HRQoL	Likelihood	LogLikelihood
1	36.6	0.00661212	-5.0189
2	41.8	0.006725313	-5.0019
3	42.6	0.006738201	-5.0000
4	43.1	0.006745631	-4.9989
5	43.4	0.006749858	-4.9982
6	44.2	0.006760279	-4.9967
7	44.9	0.006768379	-4.9955
8	46.3	0.006781711	-4.9935
9	48.6	0.006795269	-4.9915
10	49.0	0.006796563	-4.9913
11	50.0	0.006798419	-4.9911
12	51.6	0.006797282	-4.9912
13	54.6	0.006781547	-4.9935
14	54.8	0.00677987	-4.9938
15	55.7	0.006771351	-4.9951
16	57.2	0.006753646	-4.9977
17	57.6	0.006748188	-4.9985
18	60.3	0.006703308	-5.0052
19	60.9	0.006691451	-5.0069
20	65.3	0.006584073	-5.0231

Largest relative probability is for ID 11

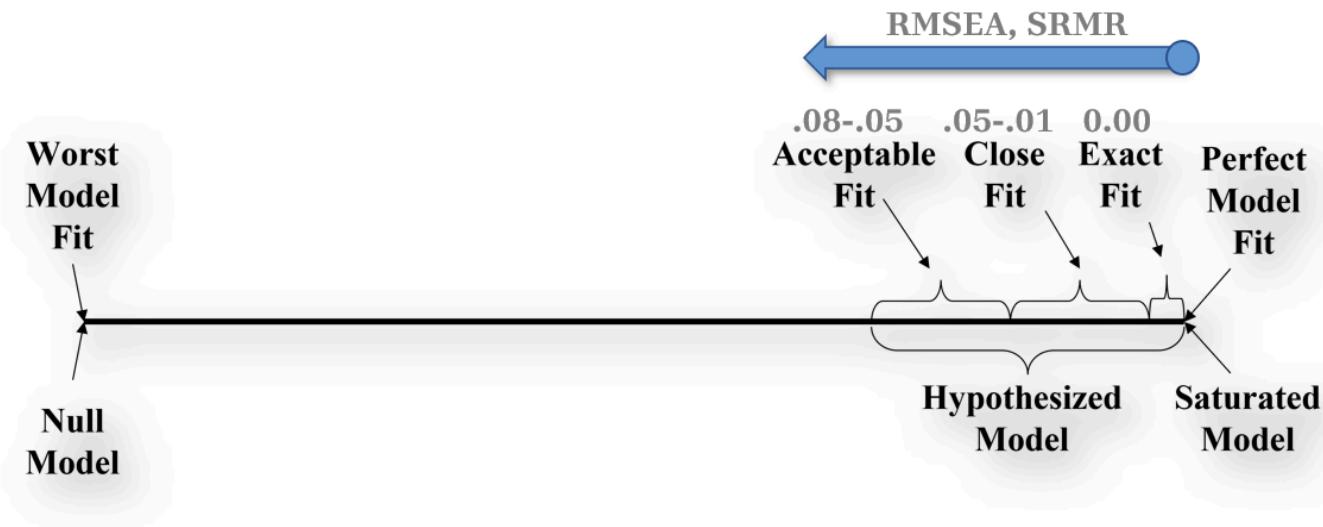


Model Fit



Your data = Model-implied?

$$\text{Chi-square } (\chi^2) = -2 * (\text{Null Loglikelihood} - \text{Alternative Loglikelihood})$$



Absolute Model Fit:

$> .10$ poor fit

.08 - .10 mediocre fit

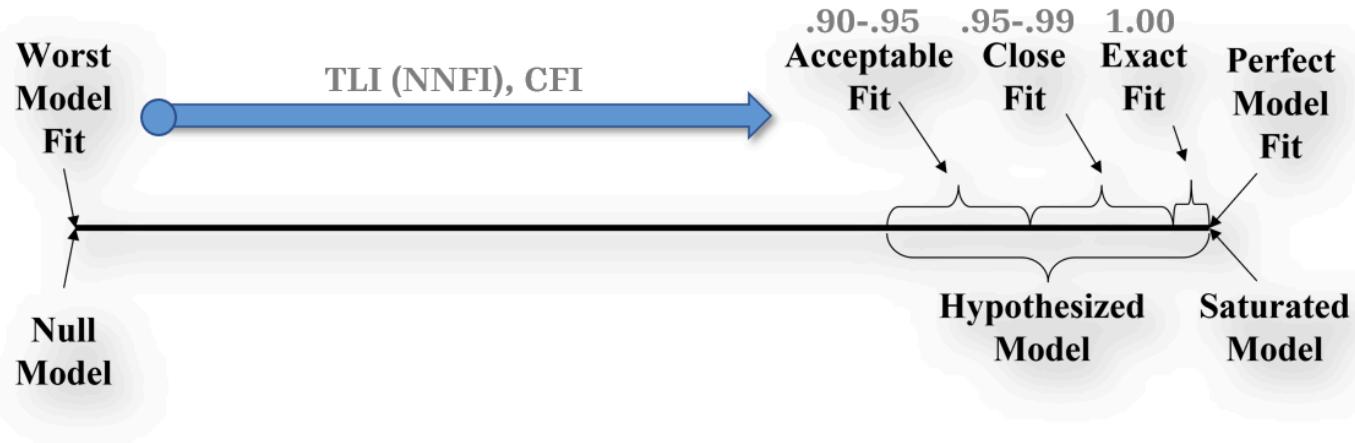
.05 - .08 acceptable fit

.01 - .05 close fit

.00 exact fit

How far from perfect:

RMSEA, SRMR



Relative Model Fit:

< .85 poor fit

.85-.90 mediocre fit

.90-.95 acceptable fit

.95-.99 close fit

1.00 exact fit

How far from worst:

TLI (NNFI), CFI...

Also: Modification indices, Fitted residual matrix, Parameter estimates...

MODEL FIT INFORMATION

Number of Free Parameters

19

$$\text{Chi-Square} = -2[(-1365.848) - (-1351.359)] = \mathbf{28.978}$$

Loglikelihood

H0 Value	-1365.848
H1 Value	-1351.359

$$DF = \frac{v(v+1)}{2} - p = \frac{6(6+1)}{2} - 13 = \mathbf{8}$$

Information Criteria

Akaike (AIC)	2769.696
Bayesian (BIC)	2844.509
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	2784.226

$$\text{RMSEA} = \sqrt{\frac{\frac{\chi_T^2 - df_T}{N}}{df_T}} = \sqrt{\frac{\frac{28.978 - 8}{379}}{8}} = \mathbf{0.083}$$

Chi-Square Test of Model Fit

Value	28.978
Degrees of Freedom	8
P-Value	0.0003

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.083
90 Percent C.I.	0.052 0.117
Probability RMSEA <= .05	0.041

$$\begin{aligned} \text{CFI} &= \frac{(\chi_0^2 - df_0) - (\chi_T^2 - df_T)}{(\chi_0^2 - df_0)} \\ &= \frac{(1939.234 - 15) - (28.978 - 8)}{(1939.234 - 15)} = \mathbf{0.989} \end{aligned}$$

CFI/TLI

CFI	0.989
TLI	0.980

$$\text{TLI} = \frac{\left(\frac{\chi_0^2}{df_0}\right) - \left(\frac{\chi_T^2}{df_T}\right)}{\left(\frac{\chi_0^2}{df_0}\right) - 1} = \frac{\left(\frac{1939.234}{15}\right) - \left(\frac{28.978}{8}\right)}{\left(\frac{1939.234}{15}\right) - 1} = \mathbf{0.980}$$

Chi-Square Test of Model Fit for the Baseline Model

Value	1939.234
Degrees of Freedom	15
P-Value	0.0000

SRMR (Standardized Root Mean Square Residual)

Value	0.030
-------	-------

How do we estimate a model?

Statistical software (Mplus, R, SAS).

Sample CFA Mplus Code

DATA: FILE = mydata.dat;

VARIABLE:

NAMES = SUP1 SUP2 SUP3;

MODEL:

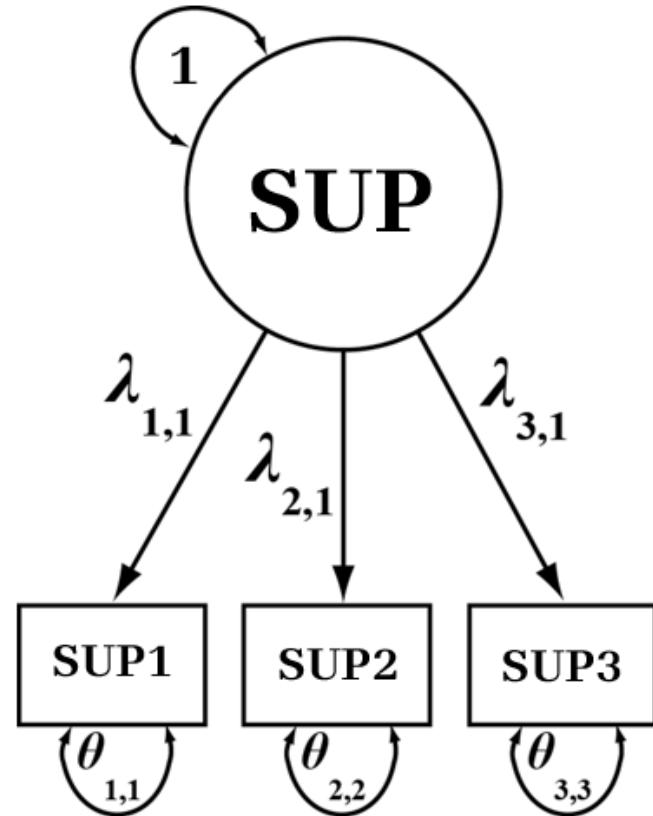
SUP by SUP1*

SUP2

SUP3;

SUP@1;

OUTPUT: TECH1;



Sample CFA Mplus Estimates

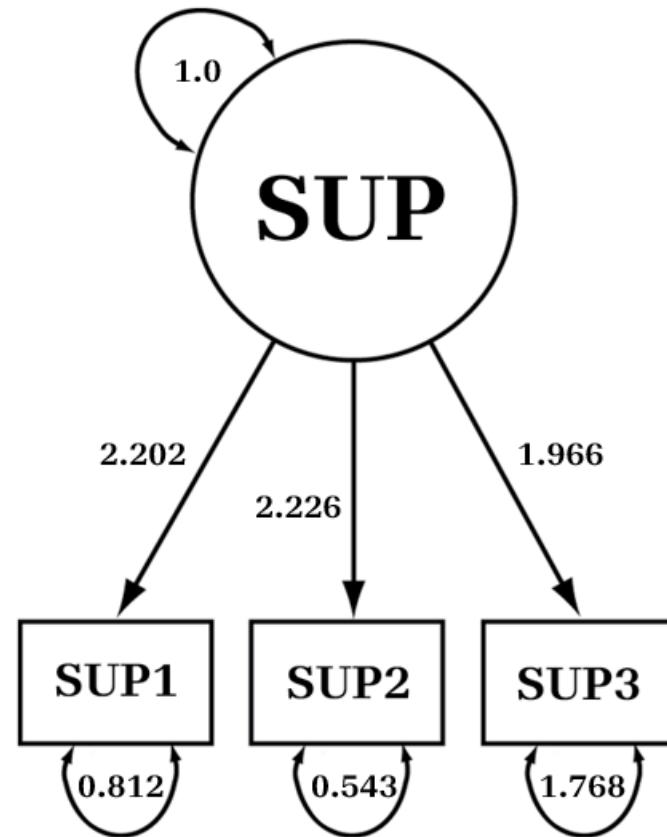
MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SUP	BY				
	SUP1	2.202	0.155	14.246	0.000
	SUP2	2.226	0.150	14.879	0.000
	SUP3	1.966	0.164	11.990	0.000
Intercepts					
	SUP1	3.287	0.199	16.522	0.000
	SUP2	2.990	0.196	15.239	0.000
	SUP3	3.322	0.198	16.739	0.000
Variances					
	SUP	1.000	0.000	999.000	999.000
Residual Variances					
	SUP1	0.812	0.183	4.429	0.000
	SUP2	0.543	0.171	3.176	0.001
	SUP3	1.768	0.243	7.265	0.000

Sample CFA Mplus Estimates

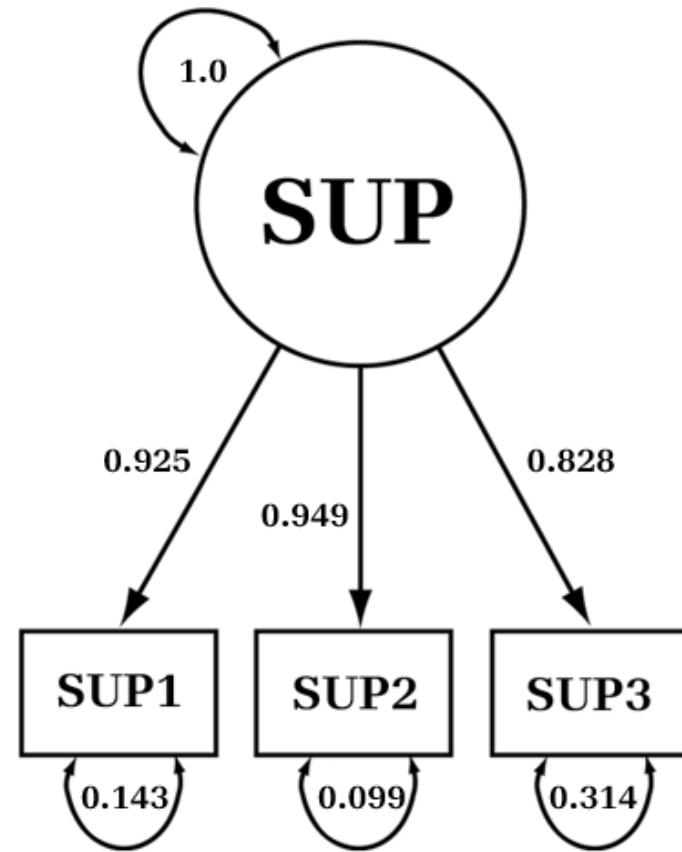
MODEL RESULTS

	Estimate	S.E.	Est./S.E.	P-Value
SUP BY				
SUP1	2.202	0.155	14.246	0.000
SUP2	2.226	0.150	14.879	0.000
SUP3	1.966	0.164	11.990	0.000
Intercepts				
SUP1	3.287	0.199	16.522	0.000
SUP2	2.990	0.196	15.239	0.000
SUP3	3.322	0.198	16.739	0.000
Variances				
SUP	1.000	0.000	999.000	999.000
Resid Var				
SUP1	0.812	0.183	4.429	0.000
SUP2	0.543	0.171	3.176	0.001
SUP3	1.768	0.243	7.265	0.000



STDYX Standardization

	Estimate	S.E.	Est./S.E.	P-Value
SUP BY				
SUP1	0.925	0.019	48.366	0.000
SUP2	0.949	0.017	54.928	0.000
SUP3	0.828	0.029	28.129	0.000
Intercepts				
SUP1	1.382	0.117	11.818	0.000
SUP2	1.275	0.113	11.318	0.000
SUP3	1.400	0.118	11.897	0.000
Variances				
SUP	1.000	0.000	999.000	999.000
Resid Var				
SUP1	0.143	0.035	4.050	0.000
SUP2	0.099	0.033	3.011	0.003
SUP3	0.314	0.049	6.435	0.000

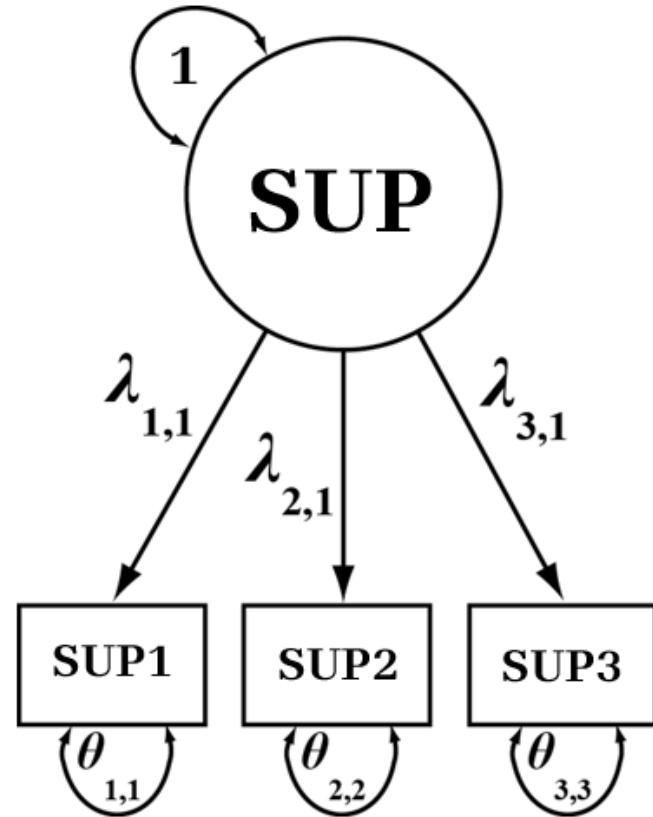


Perceived Social Support (latent SUP) accounts for **85.6%** ($0.925^2 = 0.856$)

of the variance in the indicator SUP1. Also, $0.856 + 0.143 = 1.0$

Sample CFA R (Lavaan) Code

```
library(lavaan)  
  
m1 ← '  
  
SUP =~ NA*SUP1 + SUP2 + SUP3  
  
SUP ~~ 1*SUP  
  
'  
  
fit1 ← cfa(m1, data=mydata, std.lv=T)  
  
summary(fit1, standardized=T,  
  
fit.measures=T, rsquare=T)
```



Sample CFA R (Lavaan) Estimates

Latent variables:

	Estimate	Std. Err	z-value	P(> z)	std.lv	std.all
SUP =~						
SUP1	2.211	0.155	14.262	0.000	2.211	0.927
SUP2	2.230	0.150	14.821	0.000	2.230	0.949
SUP3	1.978	0.164	12.039	0.000	1.978	0.832

Variances:

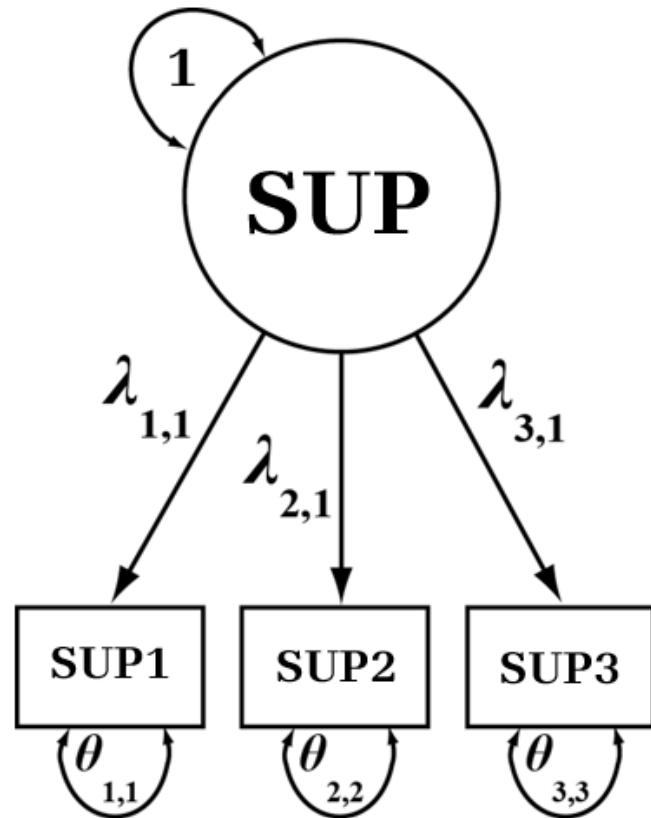
	Estimate	Std. Err	z-value	P(> z)	std.lv	std.all
SUP	1.000				1.000	1.000
.SUP1	0.797	0.180	4.419	0.000	0.797	0.140
.SUP2	0.554	0.170	3.268	0.001	0.554	0.100
.SUP3	1.739	0.240	7.240	0.000	1.739	0.308

R-Square:

	Estimate
SUP1	0.860
SUP2	0.900
SUP3	0.692

Sample CFA SAS (Proc Calis) Code

```
proc calis data=mydata method=ml;  
path SUP → SUP1 SUP2 SUP3 =  
ly1 - ly3;  
pvar SUP = 1,  
SUP1 SUP2 SUP3 = te1 - te3;  
run;
```



Sample CFA SAS (Proc Calis) Estimates

The SAS System						
The CALIS Procedure						
Covariance Structure Analysis: Maximum Likelihood Estimation						
PATH List						
Path		Parameter	Estimate	Standard Error	t Value	Pr > t
SUP	==>	SUP1	ly1	2.21910	0.15515	14.2114 <.0001
SUP	==>	SUP2	ly2	2.23813	0.15155	14.7684 <.0001
SUP	==>	SUP3	ly3	1.98466	0.16544	11.9962 <.0001

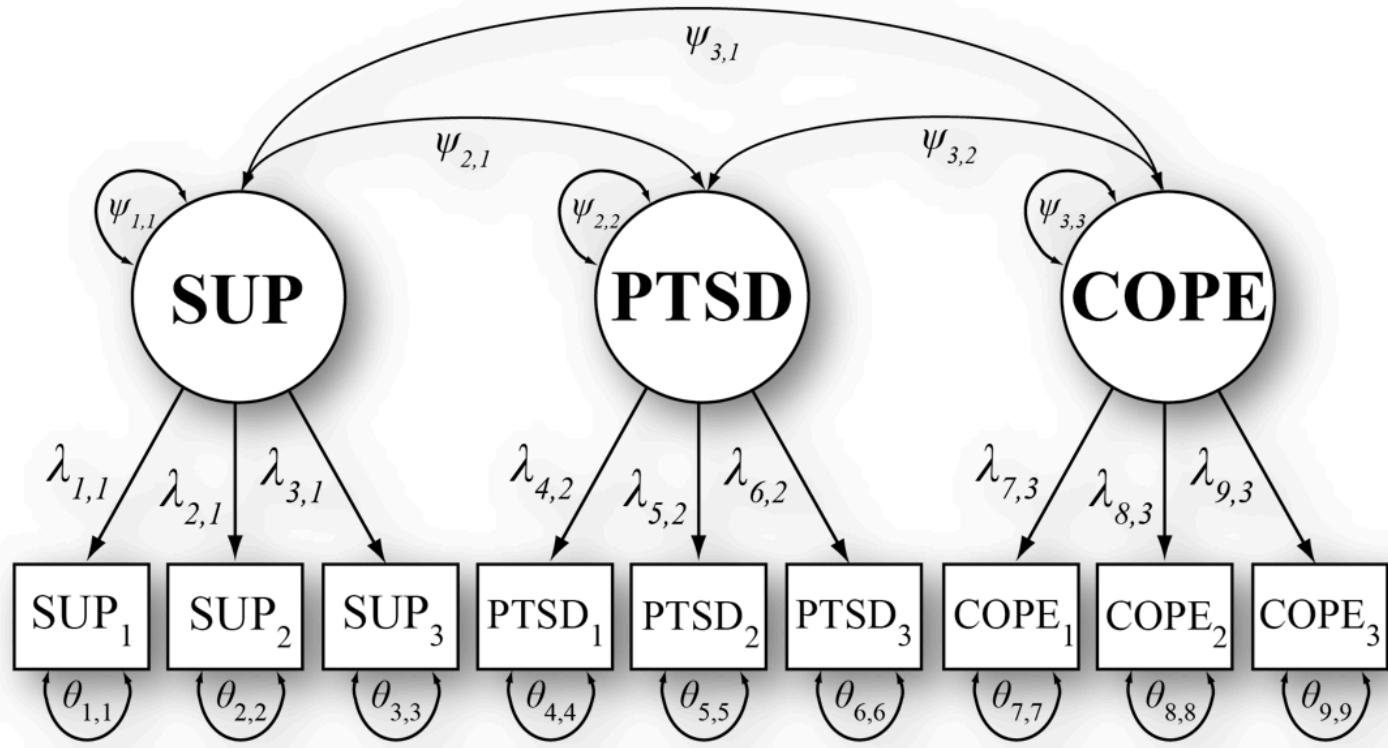
Variance Parameters						
Variance Type	Variable	Parameter	Estimate	Standard Error	t Value	Pr > t
Exogenous	SUP		1.00000			
Error	SUP1	theta1	0.80308	0.18240	4.4030 <.0001	
	SUP2	theta2	0.55811	0.17139	3.2563 0.0011	
	SUP3	theta3	1.75168	0.24281	7.2143 <.0001	

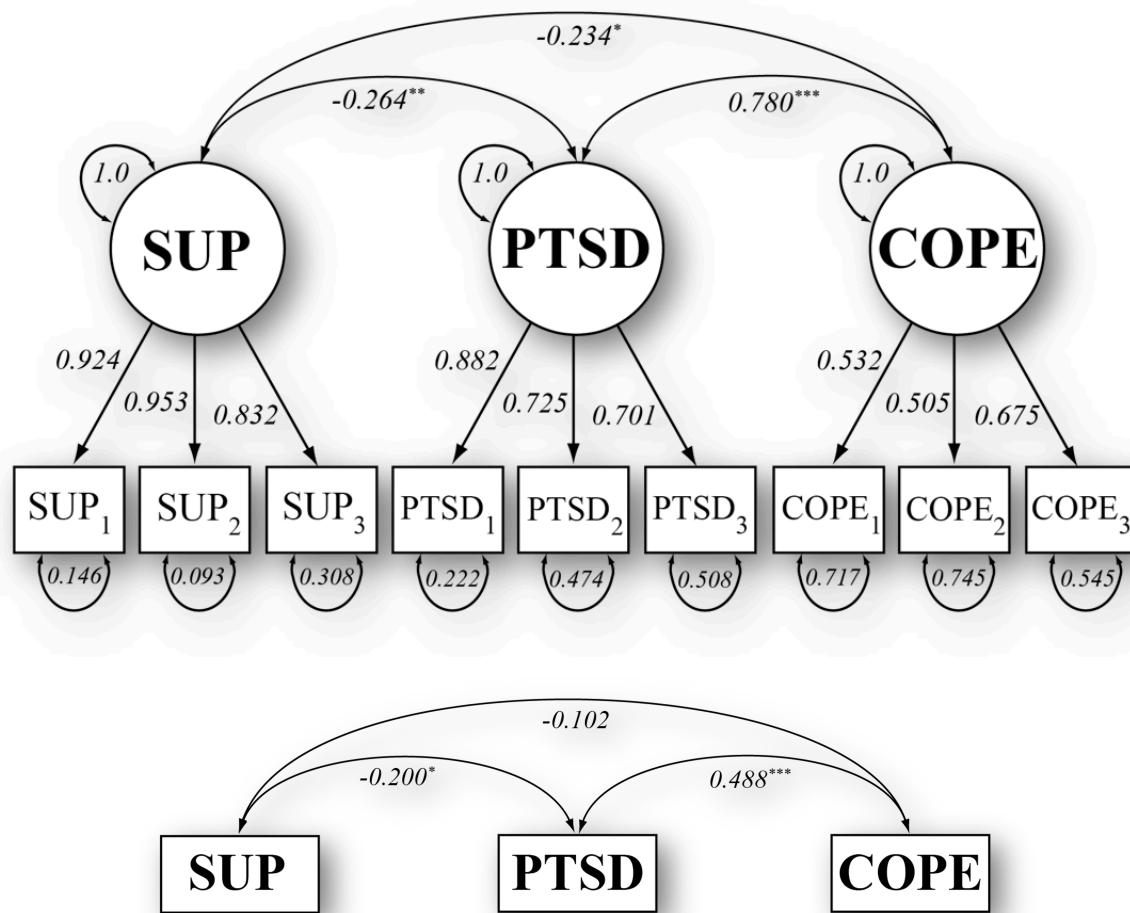
Squared Multiple Correlations				
Variable	Error Variance	Total Variance	R-Square	
SUP1	0.80308	5.72750	0.8598	
SUP2	0.55811	5.56733	0.8998	
SUP3	1.75168	5.69054	0.6922	

The SAS System						
The CALIS Procedure						
Covariance Structure Analysis: Maximum Likelihood Estimation						
Standardized Results for PATH List						
Path		Parameter	Estimate	Standard Error	t Value	Pr > t
SUP	==>	SUP1	ly1	0.92725	0.01879	49.3364 <.0001
SUP	==>	SUP2	ly2	0.94855	0.01718	55.2264 <.0001
SUP	==>	SUP3	ly3	0.83197	0.02905	28.6353 <.0001

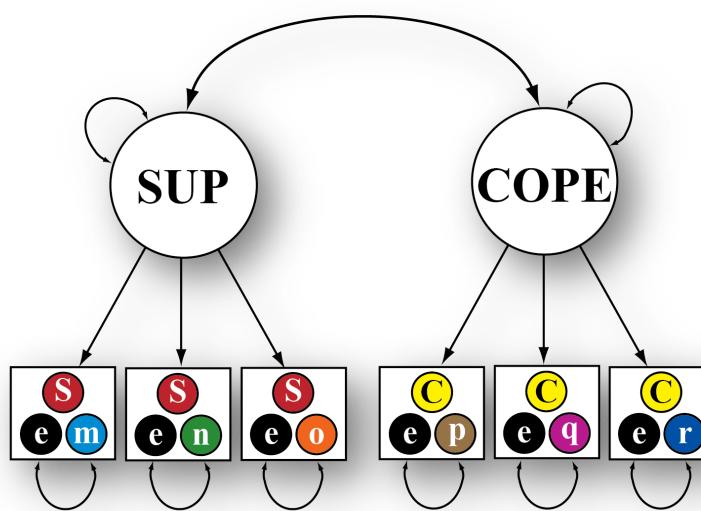
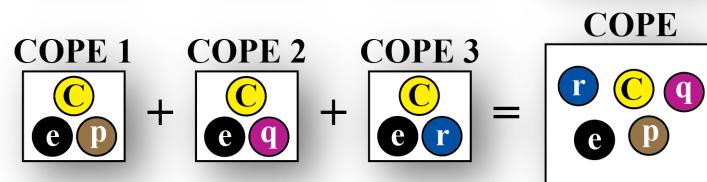
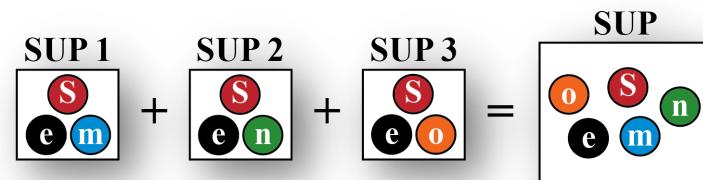
Standardized Results for Variance Parameters						
Variance Type	Variable	Parameter	Estimate	Standard Error	t Value	Pr > t
Exogenous	SUP		1.00000			
Error	SUP1	theta1	0.14022	0.03485	4.0229 <.0001	
	SUP2	theta2	0.10025	0.03258	3.0766 0.0021	
	SUP3	theta3	0.30782	0.04834	6.3673 <.0001	

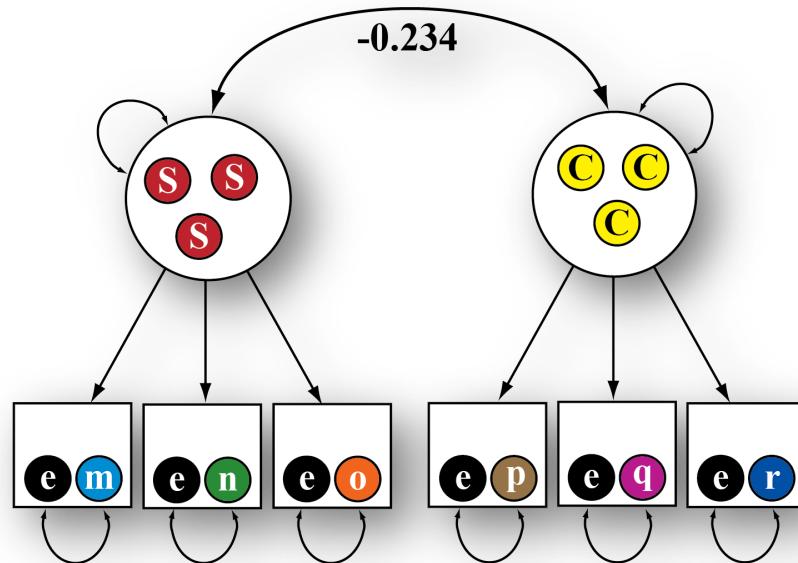
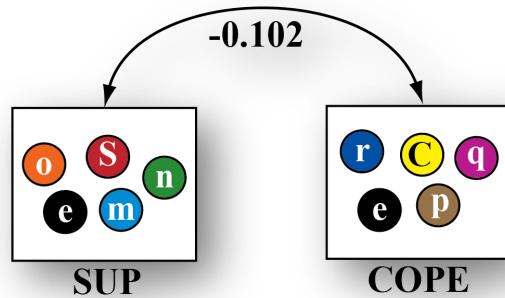
Advantages of CFA? observed vs. latent correlations.





CFA reveals subtle relationships between constructs that may be obscured by scale scores.



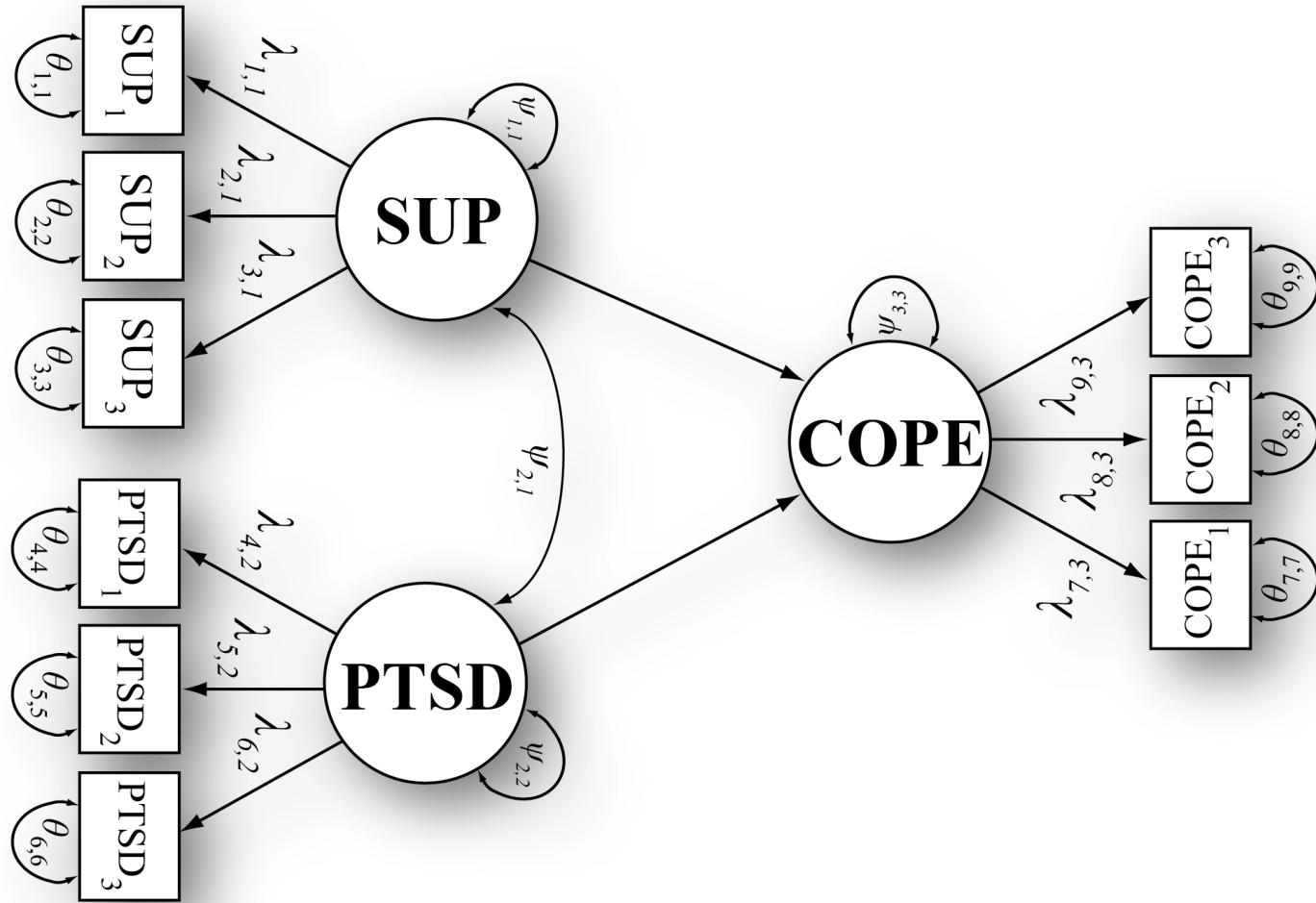


Disattenuated correlation (controlling for measurement error and specific variance)

40 / 53

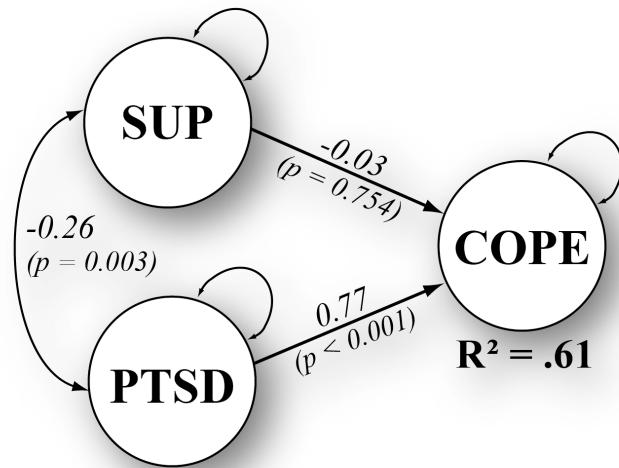
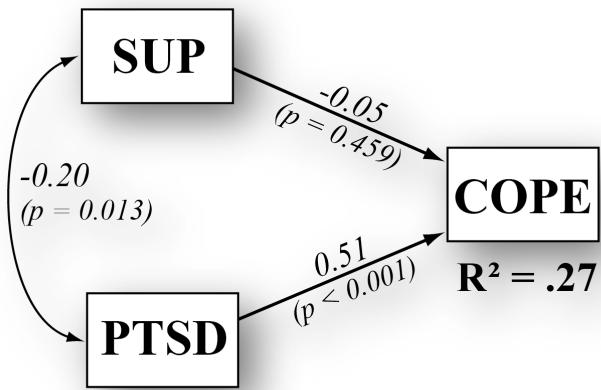
How about latent regression?

Structural Equation Modeling.



CFA reveals subtle relationships between constructs that may be obscured by scale scores.

Latent vs. Manifest/Observed



Measurement error in the observed variables can reduce the accuracy of the regression model.

Model Fit: $\chi^2(24, n=144) = 36.14$; RMSEA = .059(.000;.097) ; CFI = .980; TLI/NNFI = .970

Does the relationship between X
and Y depend on Z?

Moderation.

ANALYSIS:

```
TYPE = GENERAL RANDOM;
ESTIMATOR = ML;
ALGORITHM = INTEGRATION;
```

MODEL:

```
Z by PTSD_1* PTSD_2 PTSD_3 ;
Z@1;
X by BCOPE_1* BCOPE_6 BCOPE_13 ;
X@1;
Y by SUP_1* SUP_2 SUP_3 ;
Y@1;

XZ | X XWITH Z;

Y ON X (b1);
Y ON Z (b2);
Y ON XZ (b3);
```

MODEL CONSTRAINT:

```
NEW(LOW_Z MED_Z HIGH_Z SIMP_LO SIMP_MED SIMP_HI);

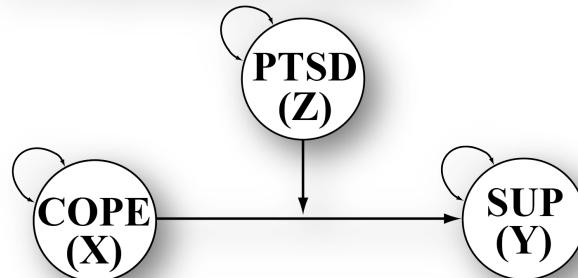
LOW_Z = 0;
MED_Z = 1.5;
HIGH_Z = 3;

SIMP_LO = b1 + b3*LOW_Z;
SIMP_MED = b1 + b3*MED_Z;
SIMP_HI = b1 + b3*HIGH_Z;

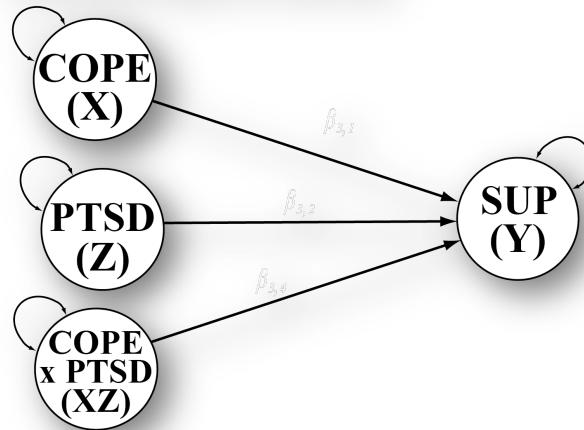
PLOT (LOMOD MEDMOD HIMOD);
LOOP(XVAL,1,4,0.5);
LOMOD = (b1 + b3*LOW_Z)*XVAL;
MEDMOD = (b1 + b3*MED_Z)*XVAL;
HIMOD = (b1 + b3*HIGH_Z)*XVAL;

PLOT: TYPE = plot2;
OUTPUT: CINT STAND;
```

(A) Conceptual diagram for moderation

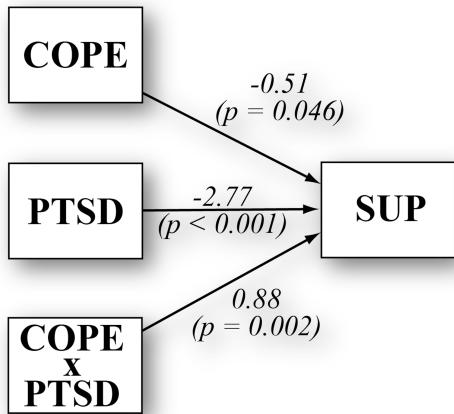


(B) Statistical diagram of moderation



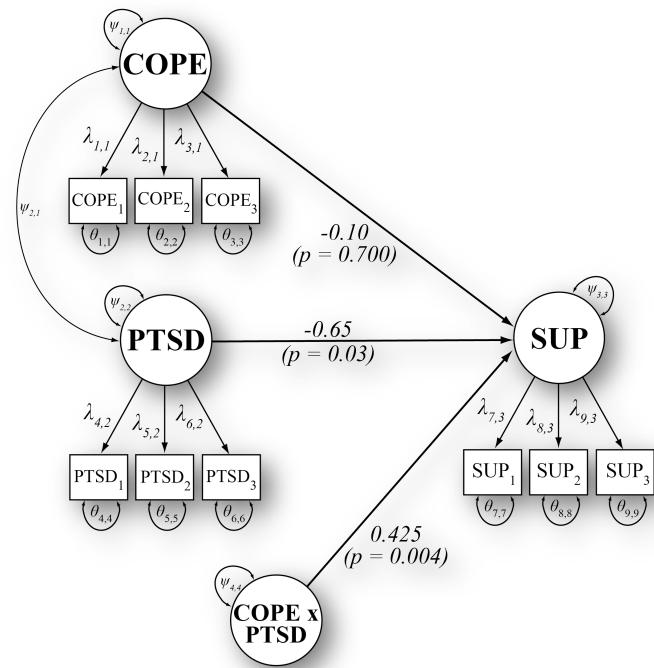
Mplus code using the Latent Moderated Structural Equations (LMS) Method.

Regular Moderation



$$R^2 = 0.11$$

Latent Moderation



$$R^2 = 0.46$$

Model differences: In *regular moderation*, measurement error can inflate predictive effects

(i.e., making them seem larger). *Latent moderation* provides more accurate estimates.

Does the effect of X on Y operate through M?

Mediation.

ANALYSIS:

```
TYPE = GENERAL;  
ESTIMATOR = ML;  
BOOTSTRAP = 10000;
```

MODEL:

```
X by PTSD_1* PTSD_2 PTSD_3 ;  
X@1;  
M by BCOPE_1* BCOPE_6 BCOPE_13 ;  
M@1;  
Y by SUP_12* SUP_6 SUP_7 ;  
Y@1;
```

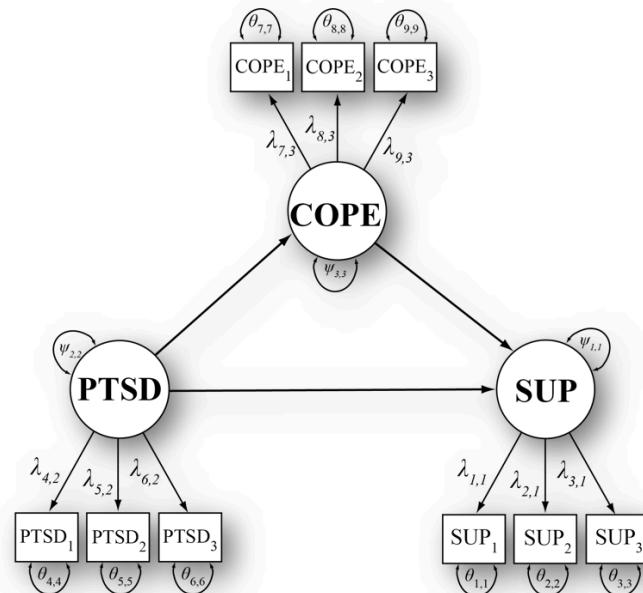
```
Y ON X M;  
M ON X;
```

```
MODEL INDIRECT:  
Y IND X;
```

OUTPUT:

```
STAND CINT (bcbootstrap);
```

Latent Mediation



Do I have enough power for this?

SEM Power Analysis.

Parameter Power

```
TITLE: Monte Carlo Simulation for Power;
MONTECARLO:
  NAMES ARE sup1 sup2 sup3 copel cope2 cope3;
  NOBSERVATIONS = 150;
  NREPS = 2000;
  SEED = 461981;
  !save = mypower.dat;

MODEL POPULATION:
  SUP BY sup1*0.924 sup2*0.953 sup3*0.832;
  COPE BY copel*0.532 cope2*0.505 cope3*0.675;
  !SUP WITH COPE*-0.234;
  SUP WITH COPE*-0.300;
  sup1*0.146 sup2*0.093 sup3*0.308;
  copel*0.717 cope2*0.745 cope3*0.545;
  SUP@1; COPE@1;

MODEL:
  SUP BY sup1*0.924 sup2*0.953 sup3*0.832;
  COPE BY copel*0.532 cope2*0.505 cope3*0.675;
  !SUP WITH COPE*-0.234;
  SUP WITH COPE*-0.300;
  sup1*0.146 sup2*0.093 sup3*0.308;
  copel*0.717 cope2*0.745 cope3*0.545;
  SUP@1; COPE@1;

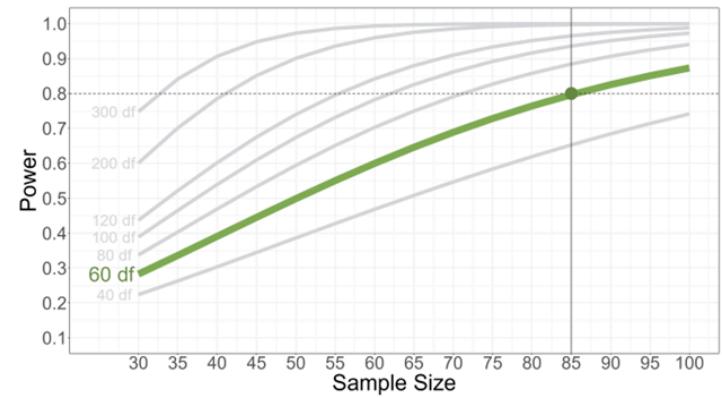
ANALYSIS: ESTIMATOR = ML;
OUTPUT: TECH9;
```

		MODEL RESULTS			
		Population	ESTIMATES	95%	% Sig
		Average	Cover	Coeff	
SUP	BY				
SUP1		0.924	0.9203	0.955	1.000
SUP2		0.953	0.9491	0.945	1.000
SUP3		0.832	0.8293	0.957	1.000
COPE	BY				
COPE1		0.532	0.5300	0.944	0.995
COPE2		0.505	0.4991	0.949	0.993
COPE3		0.675	0.6837	0.952	0.997
SUP	WITH				
COPE		-0.300	-0.2928	0.940	0.797

80% power for a correlation of at least -0.30 given $N = 150$.

Model Power

```
w = NULL # We want to store results in here  
  
alpha = .05  
d_list = c(40, 60, 80, 100, 120, 200, 300) #degrees of freedom  
n_list = c(seq(30, 100, by = 5)) #sample size  
rmsea0 <- 0.05 #null hypothesized RMSEA (.05 close fit, .00 exact fit)  
rmseaa_list = c(.10)  
  
for (d in d_list){  
  for (n in n_list){  
    for (rmseaa in rmseaa_list){  
  
      ncp0 <- (n-1)*d*rmsea0^2  
      ncpa <- (n-1)*d*rmseaa^2  
  
      #compute power  
      if(rmseaa<rmsea0) {  
        cval <- qchisq(alpha,d,ncp=ncp0,lower.tail=F)  
        pow <- pchisq(cval,d,ncp=ncpa,lower.tail=F)  
      }  
      if(rmseaa>rmsea0) {  
        cval <- qchisq(1-alpha,d,ncp=ncp0,lower.tail=F)  
        pow <- 1-pchisq(cval,d,ncp=ncpa,lower.tail=F)  
      }  
  
      w = rbind(w, c(d, n, rmseaa, pow))  
    }  
  }  
}  
  
colnames(w) = c('df', 'n', 'rmseaa', 'Power')  
simdf <- as.data.frame(w)  
simdf$grp <- as.factor(simdf$df)
```



80% power for a CFA with at least 60 df given $N = 85$.

Thanks!

Any further questions?

Feel free to join the coding session.

