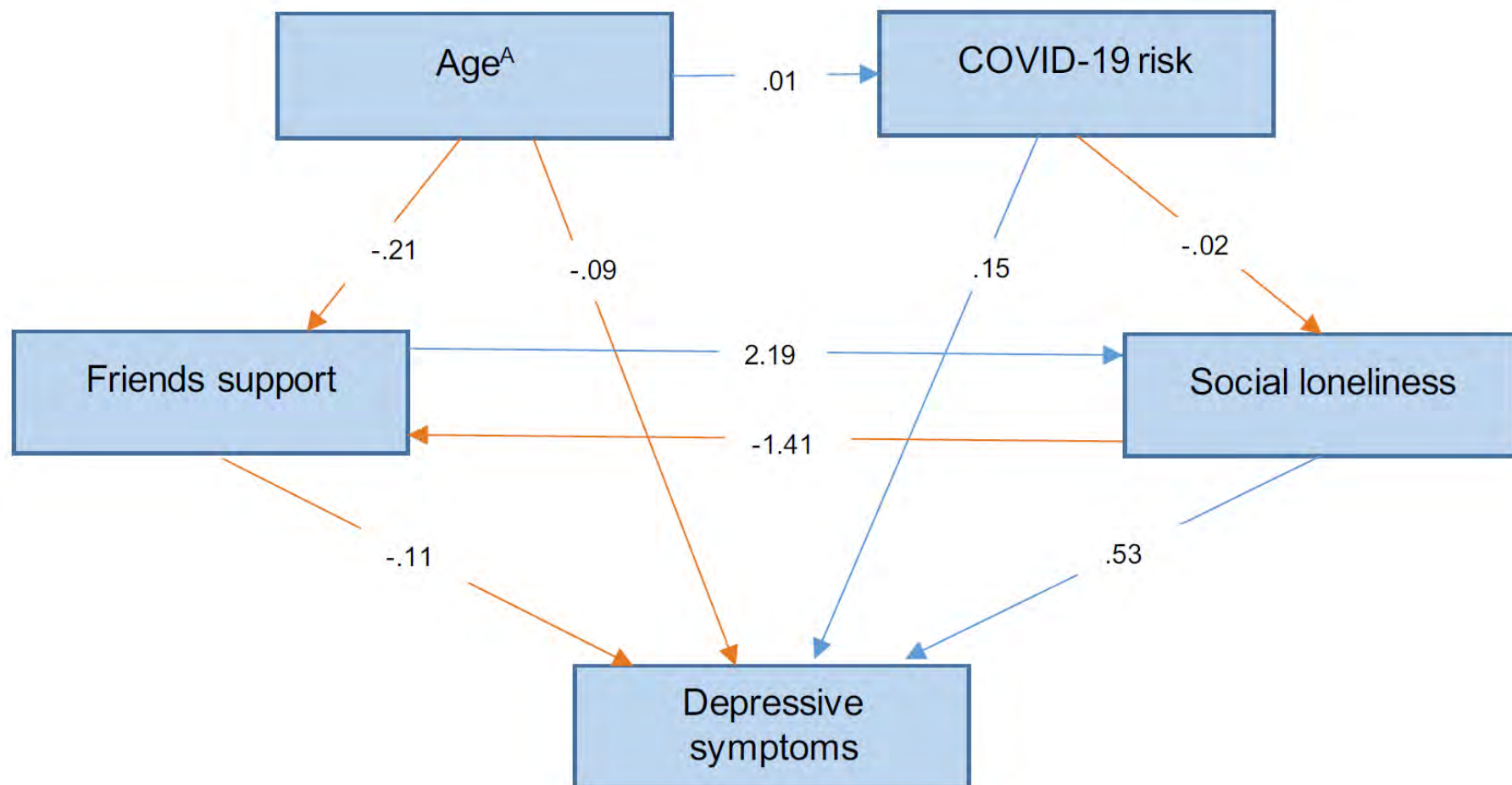


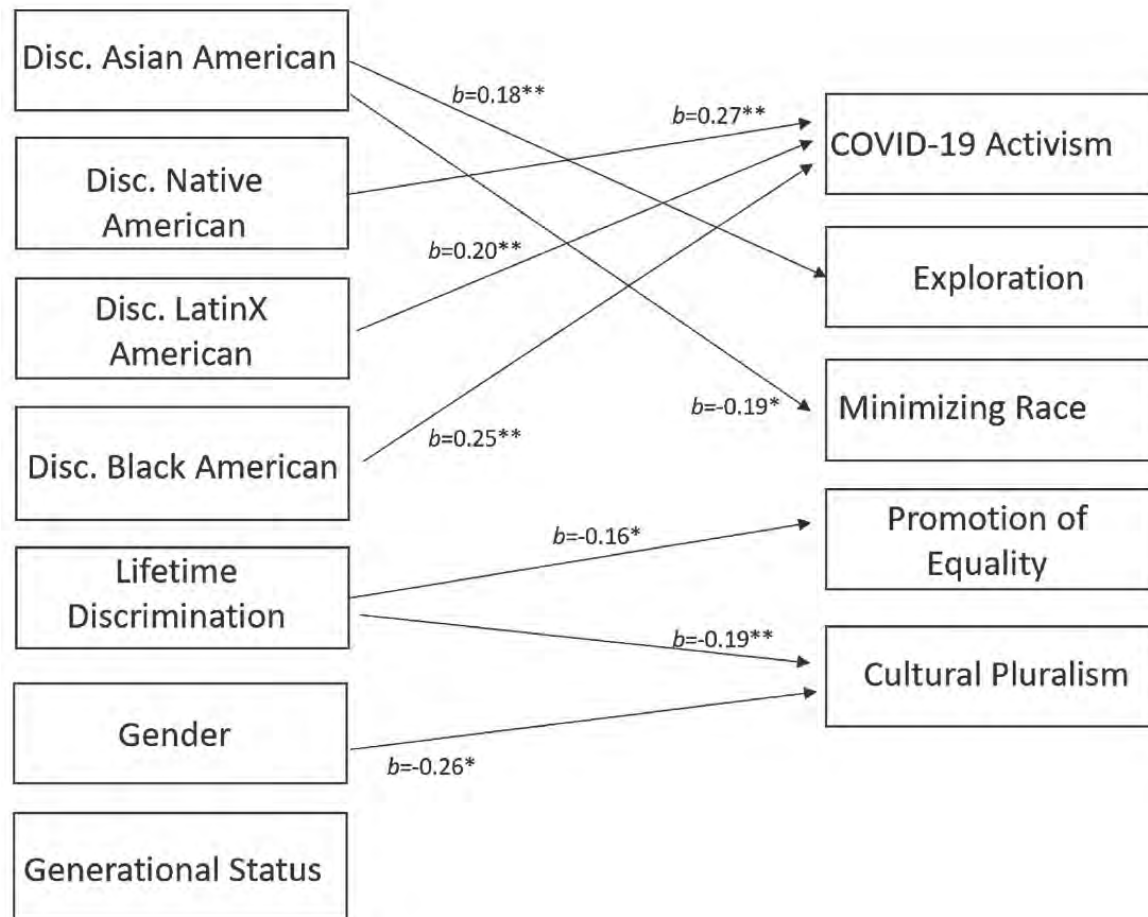
Evaluating Measurement Invariance in Structural Equation Models

Meghan K. Cain, Ph.D., StataCorp LLC

March 16th, 2023

American Psychological Association





Binary Group Variables

$EVENT_{mass}$

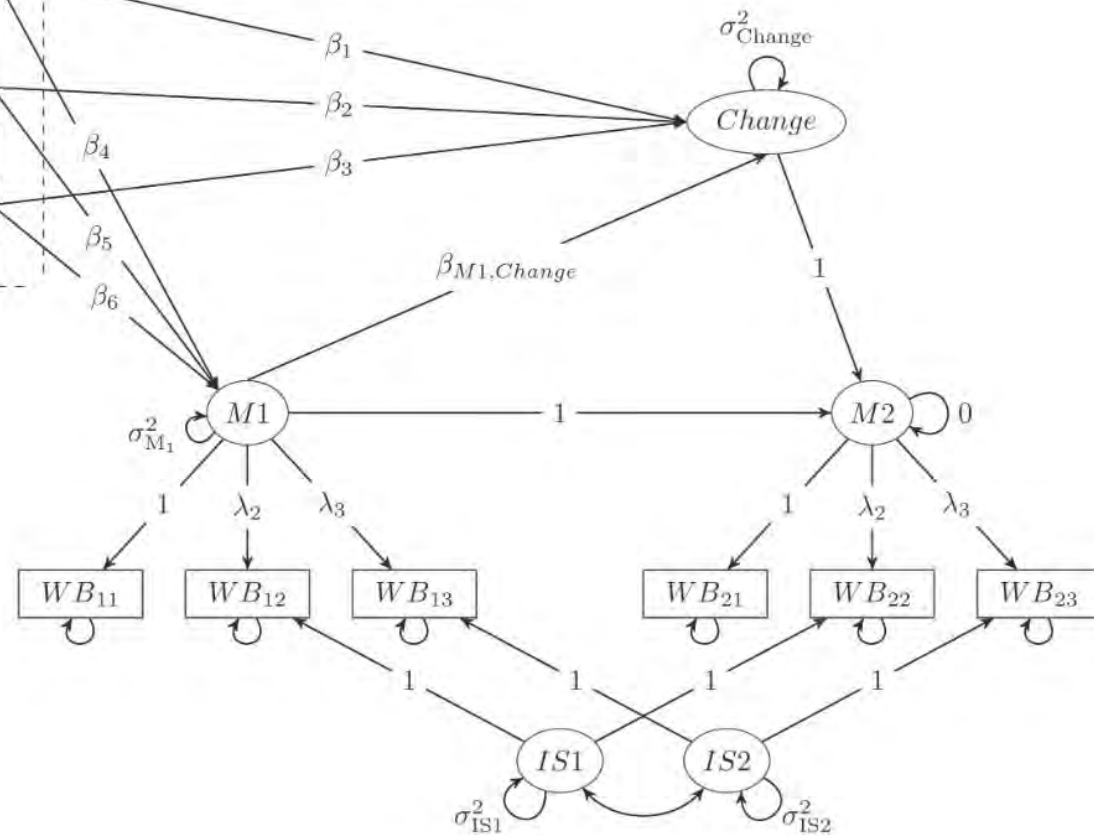
$EVENT_{other}$

$CONTROL_{other}$

Definition of Immediate Effects:

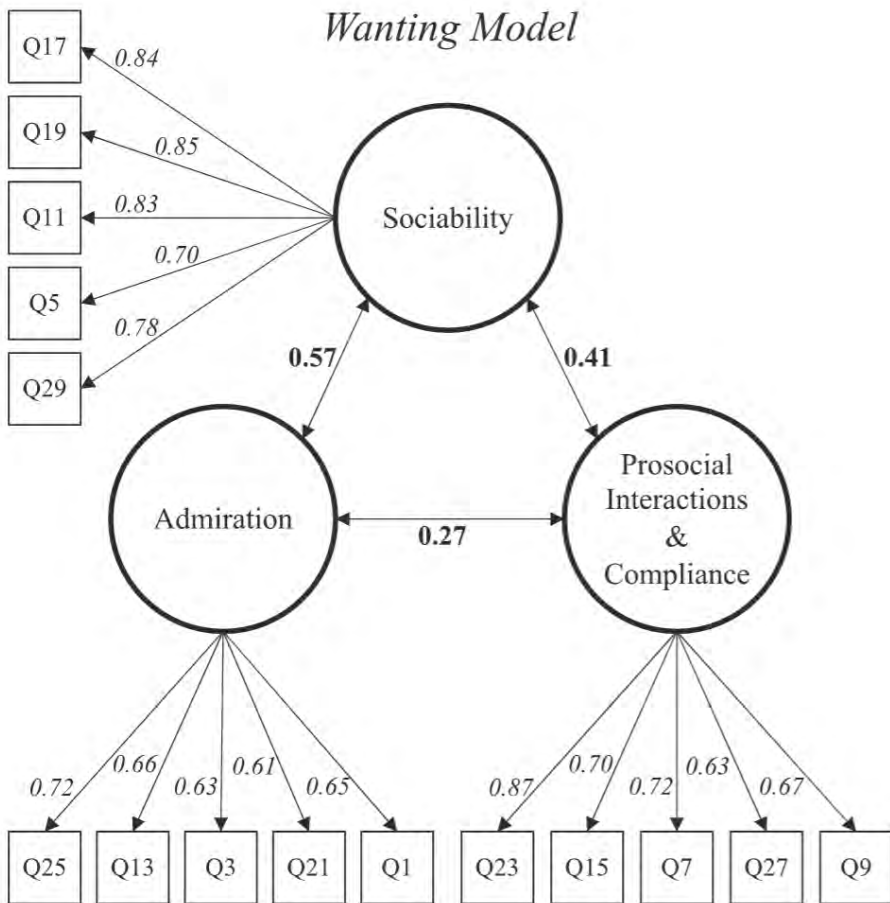
$$DiD_{mass} = \beta_1$$

$$DiD_{other} = \beta_2 - \beta_3$$

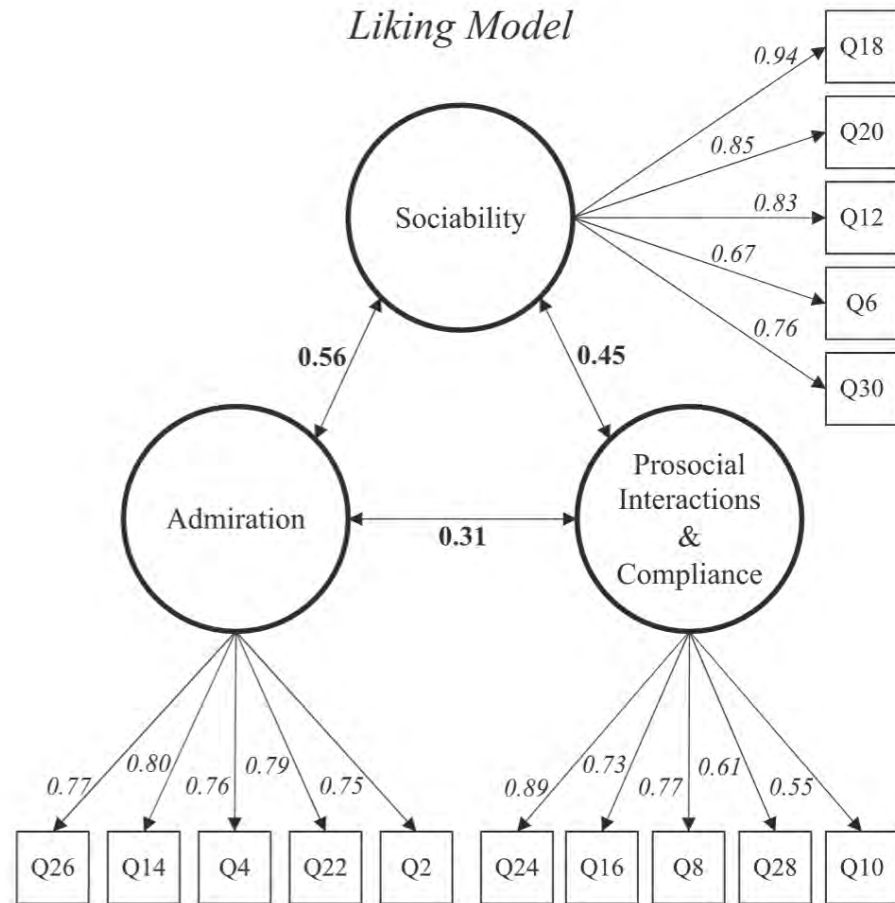


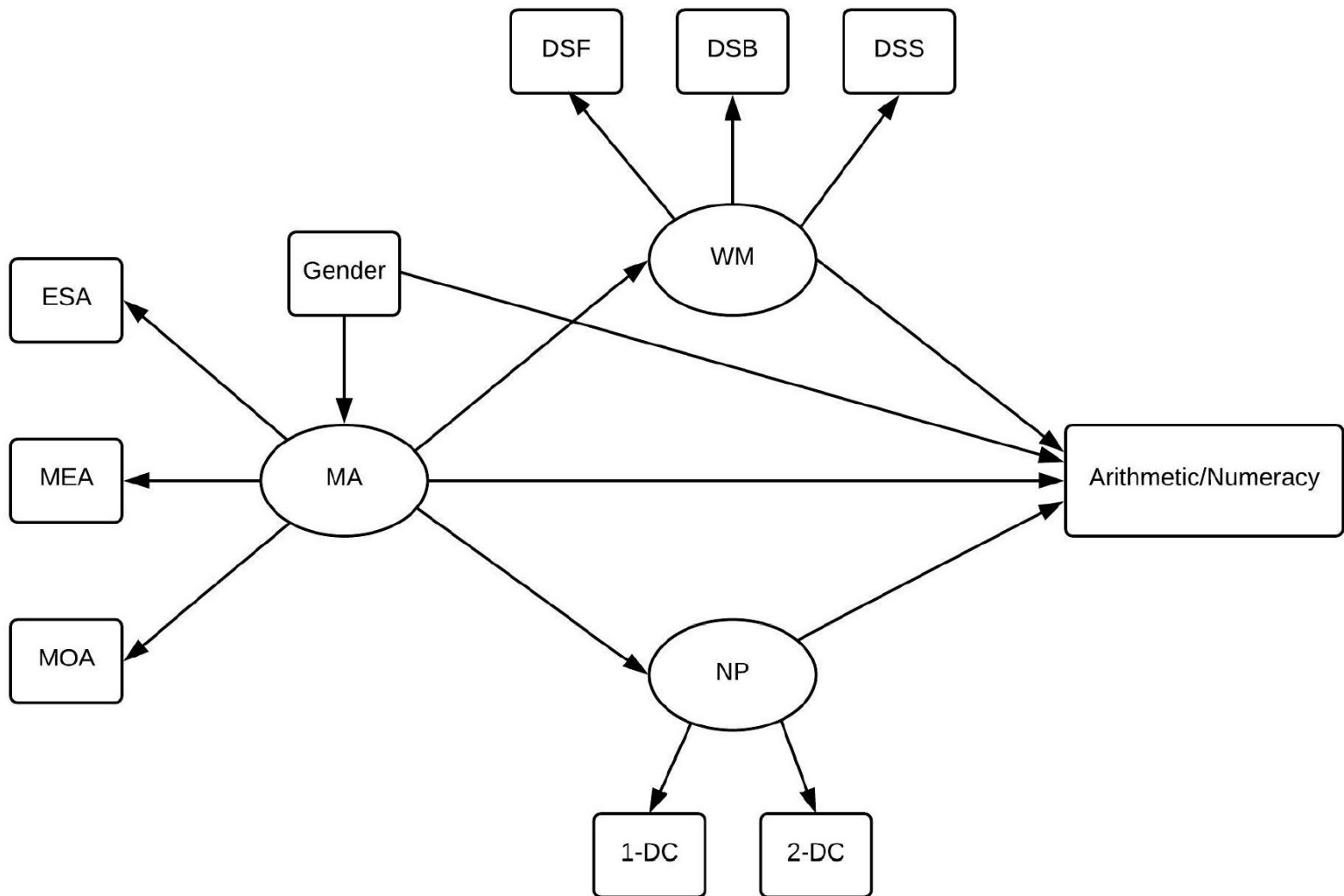


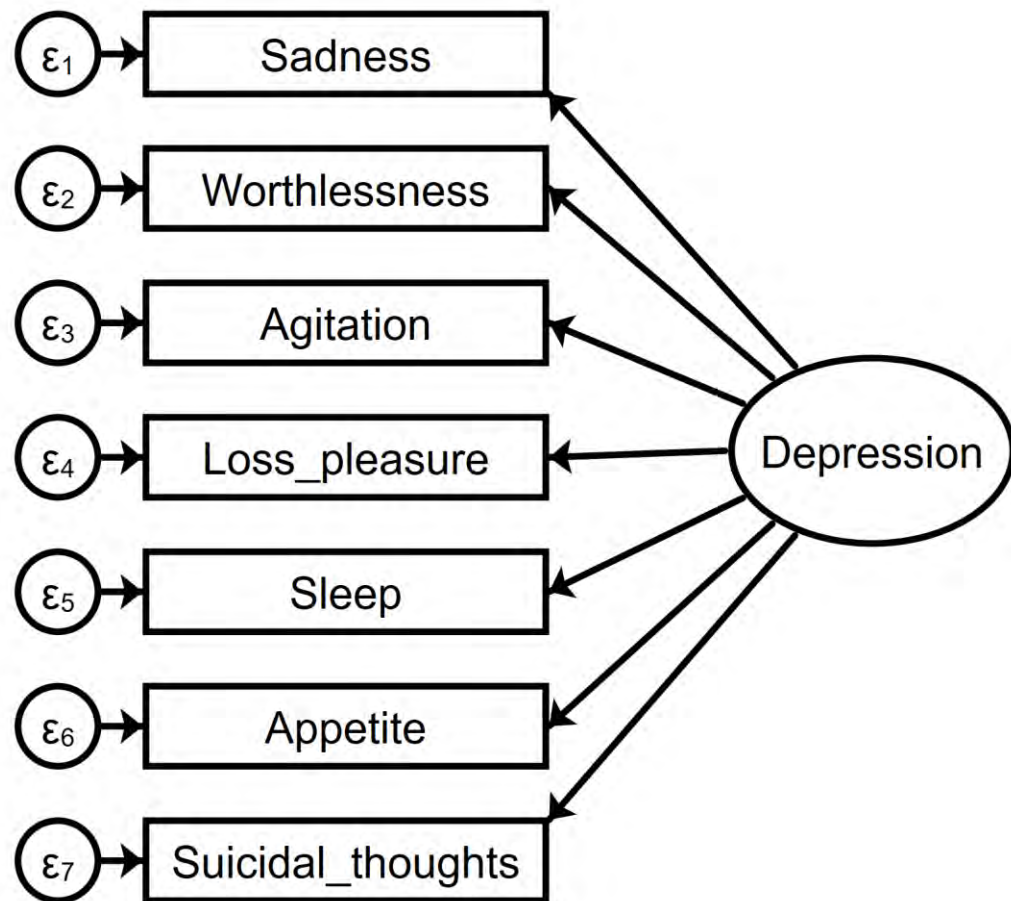
Wanting Model

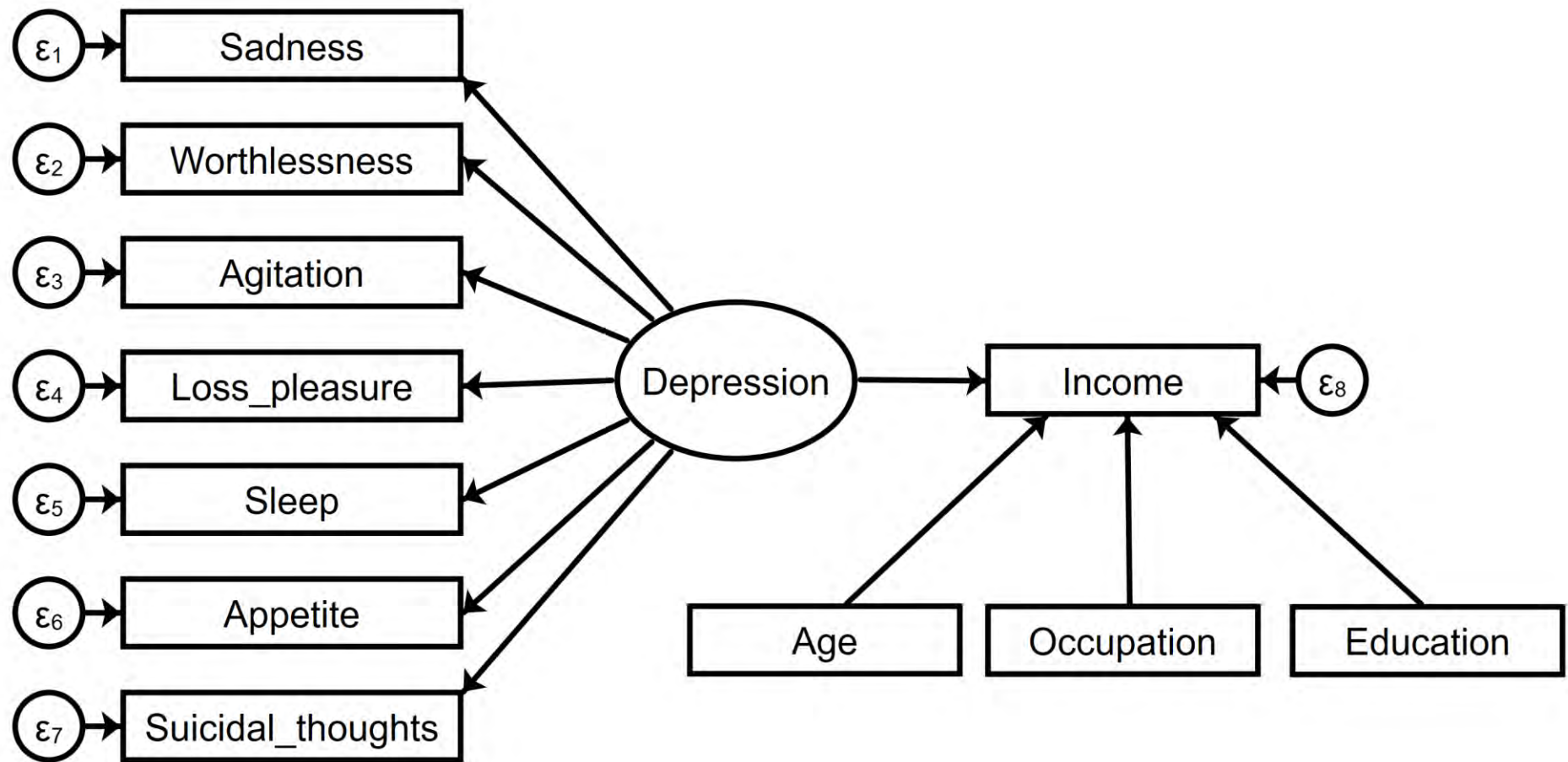


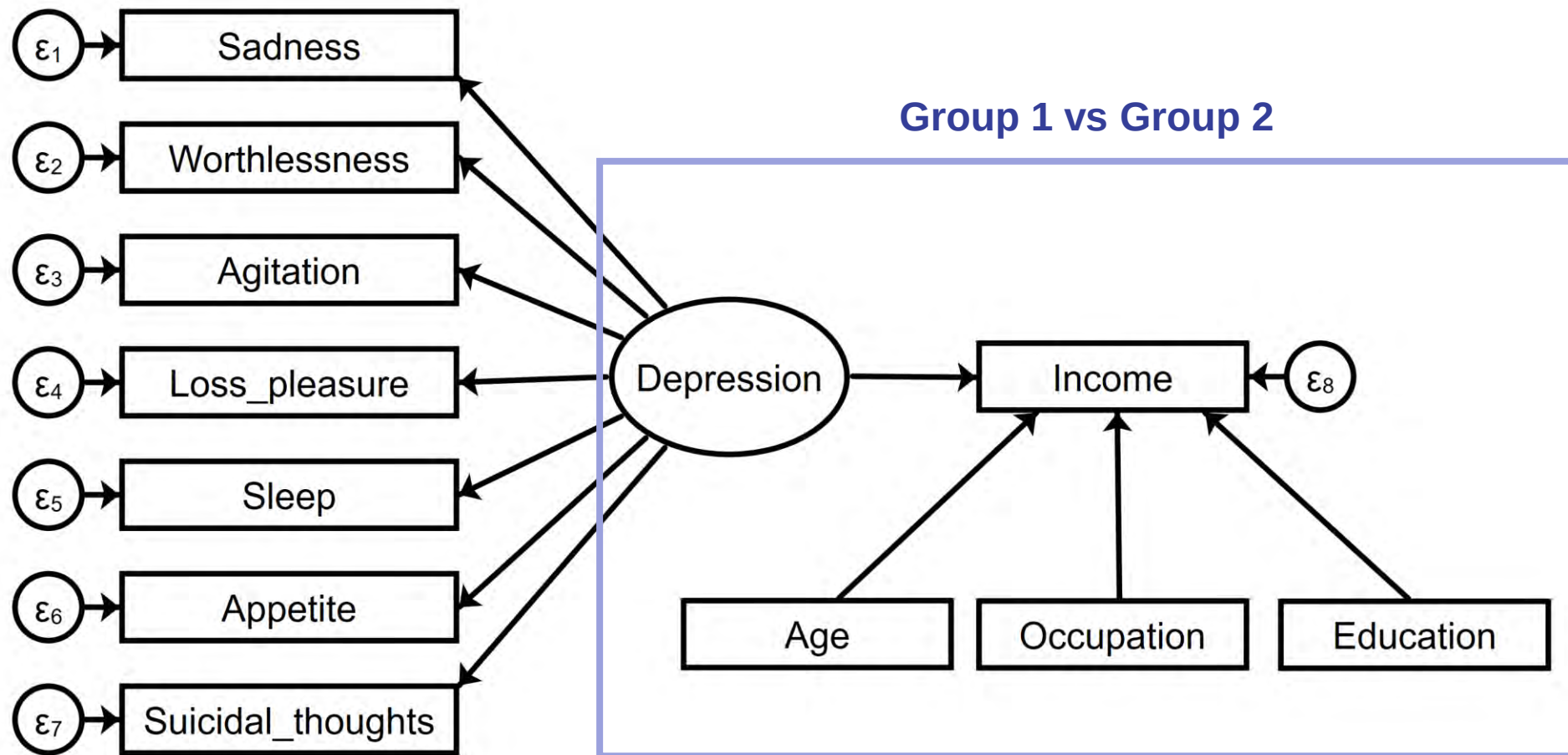
Liking Model



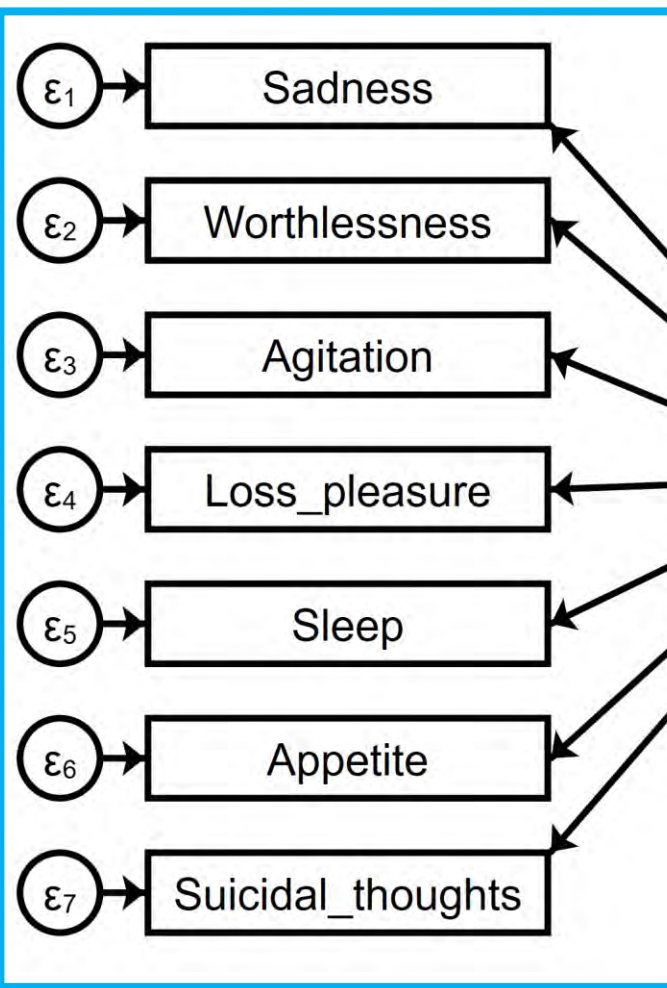




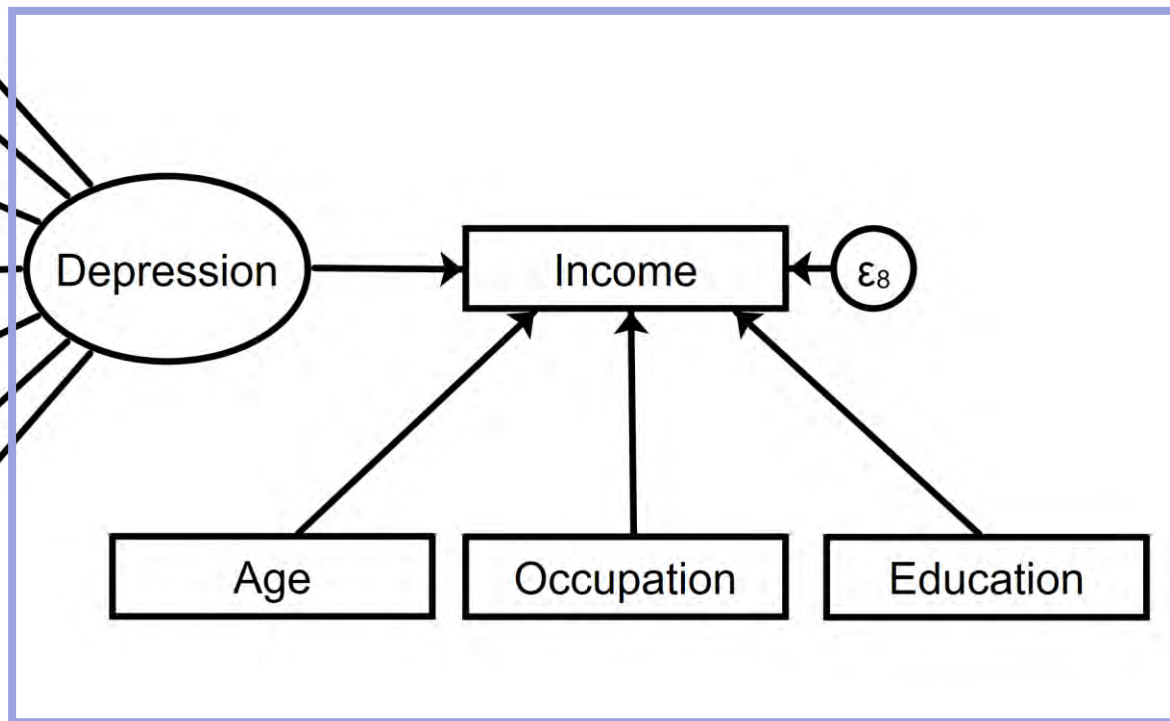




Group 1 = Group 2



Group 1 vs Group 2





Measurement invariance/ equivalence (MI/E)

“The general question of invariance of measurement is one of whether or not, under different conditions ... measurement operations yield measures of the same attribute.” – Horn and McArdle (1992)

“Are there gender, ethnic, or other individual differences that preclude responding to instruments in similar ways?”
– Vandenberg and Lance (2000)

“Measurement noninvariance suggests that ... the construct cannot be meaningfully tested or construed across groups or across time.”
– Putnick and Bornstein (2016)

SAT-Verbal

Although cacti are not _____ Hawaii, several species have been introduced there and are flourishing.

- A. adaptable to
- B. native to*
- C. excluded from
- D. compatible with
- E. limited to

COCKPIT:PILOT::

- A. booth:toll collector*
- B. somersault:acrobat
- C. baton:drum major
- D. sidewalk:pedestrian

Noting the potential danger involved in producing nuclear energy as a source of power, many people argue that we should be more systematic in _____ the Sun's energy.

- A. heating
- B. magnifying
- C. dispelling
- D. tapping*
- E. discovering

RIVET:METAL::

- A. needle:thimble
- B. cork:bottle
- C. nail:hammer
- D. staple:paper*
- E. rope:swing

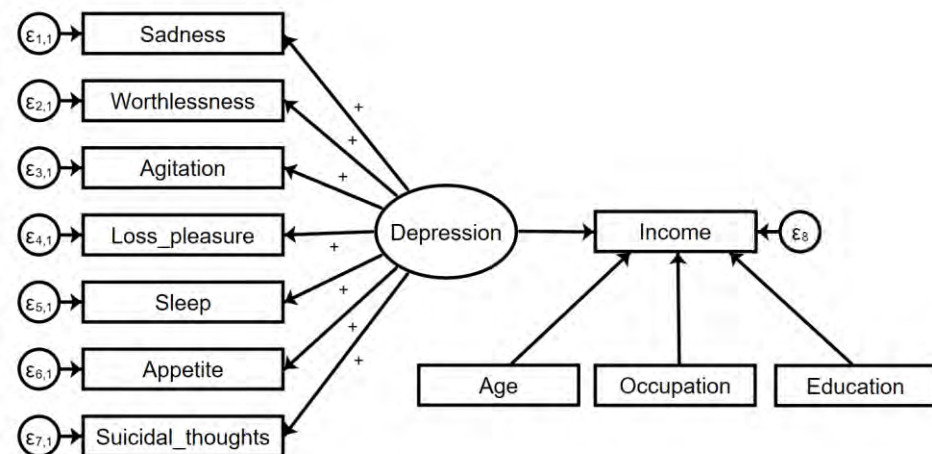


How to assess MI/E?

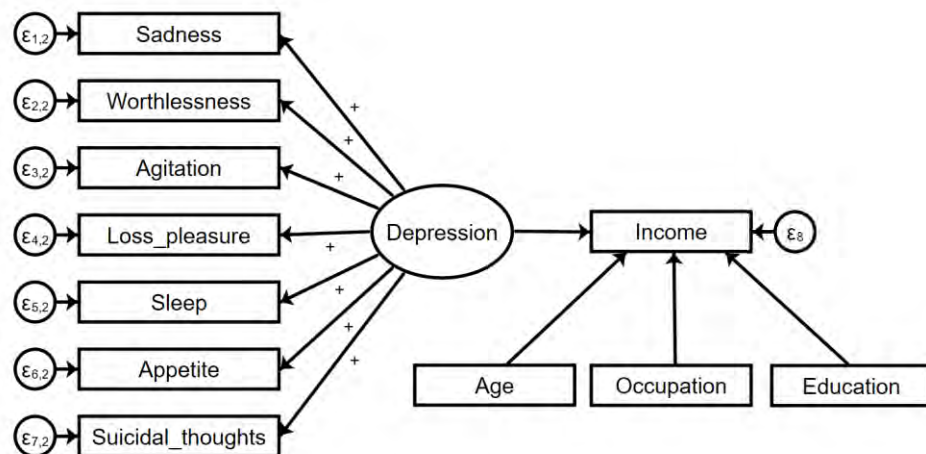
- Item response theory:
 - Differential item functioning (DIF)
 - Differential test functioning (DTF)
- Structural equation modeling:
 - Factor invariance (Widaman and Reiss, 1997)
 - Configural invariance
 - Weak factor invariance (metric invariance)
 - Strong factor invariance (scale invariance)
 - Strict factor invariance
 - For reviews, see Putnick and Bornstein (2016) or Vandenberg and Lance (2000)

What MI/E tells you

- Configural invariance:
same factor loading
configuration



Group 1



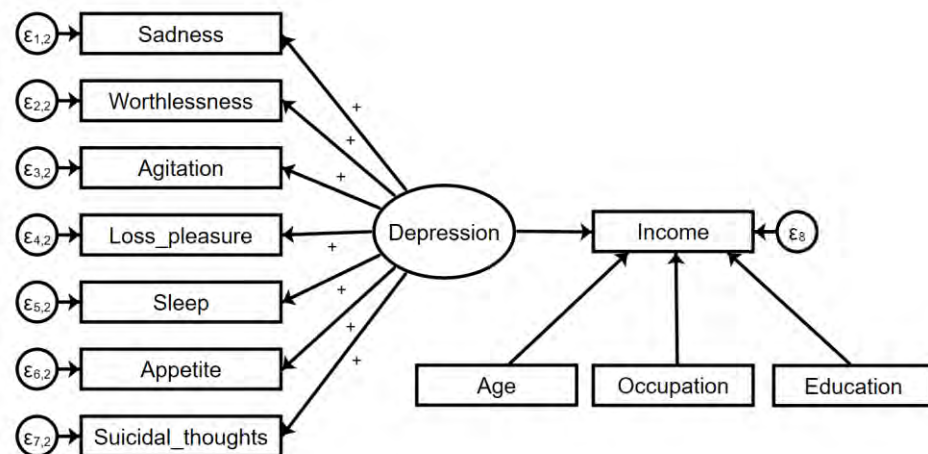
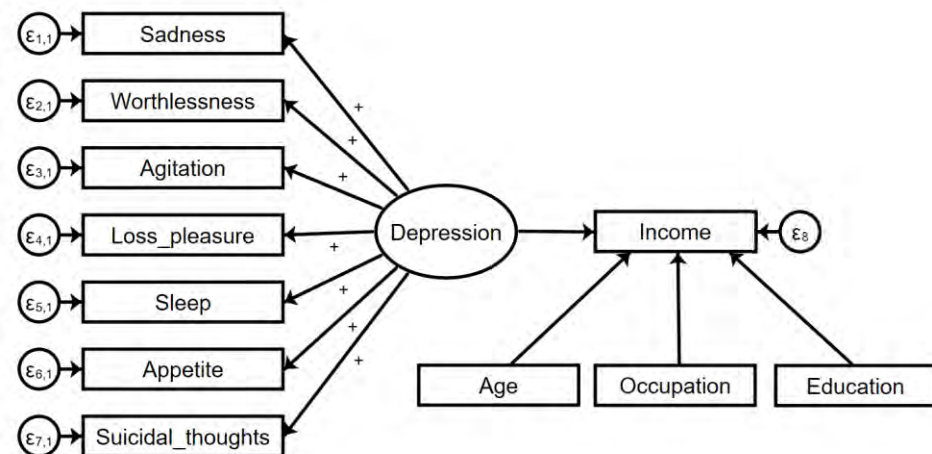
Group 2

What MI/E tells you

- Configural invariance: same factor loading configuration

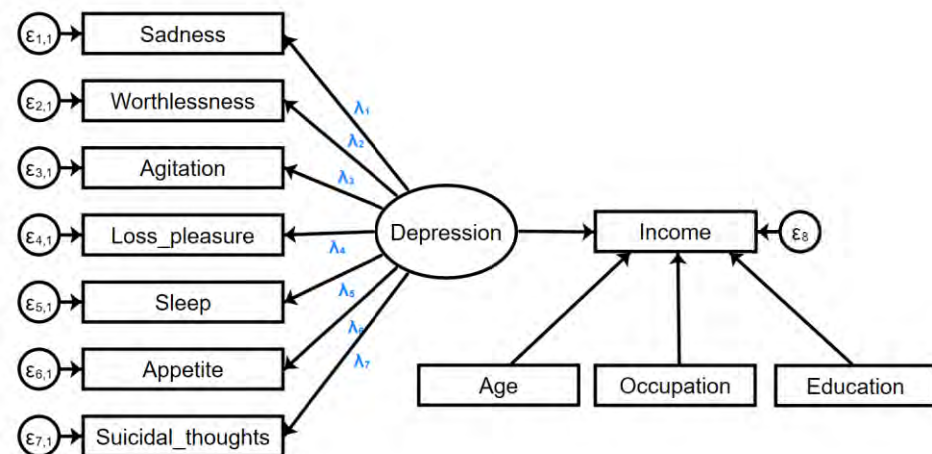


Latent factors have the same conceptual meaning in each group

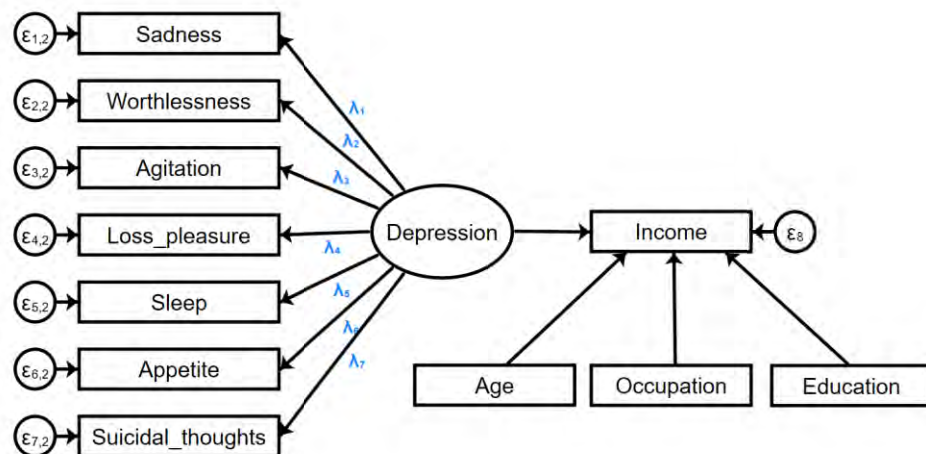


What MI/E tells you

- Weak factor invariance:
same factor loadings



Group 1



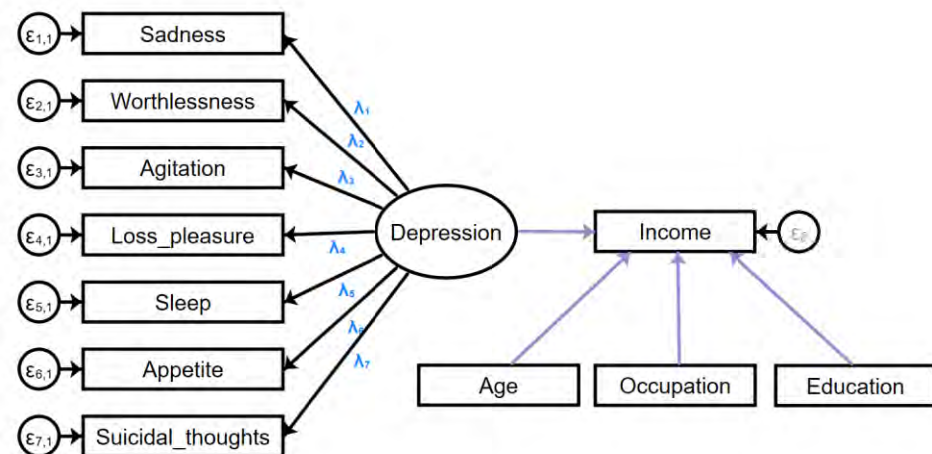
Group 2

What MI/E tells you

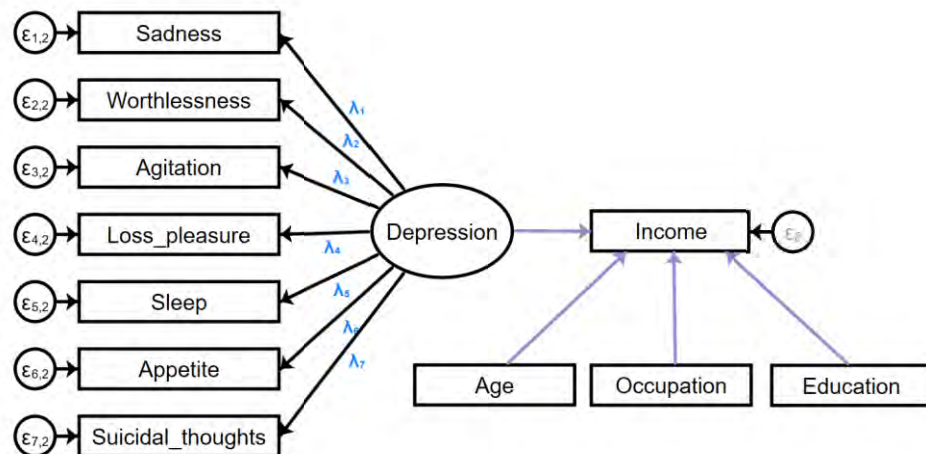
- Weak factor invariance:
same factor loadings



Covariance parameter
comparisons are
meaningful



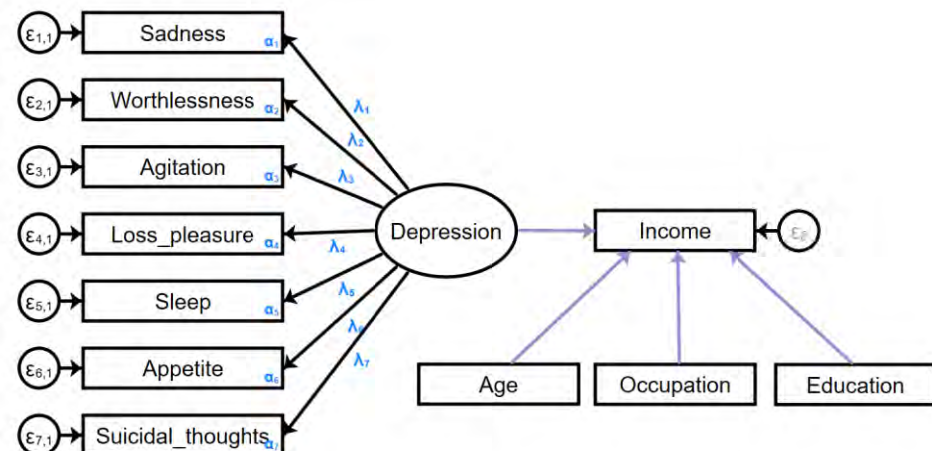
Group 1



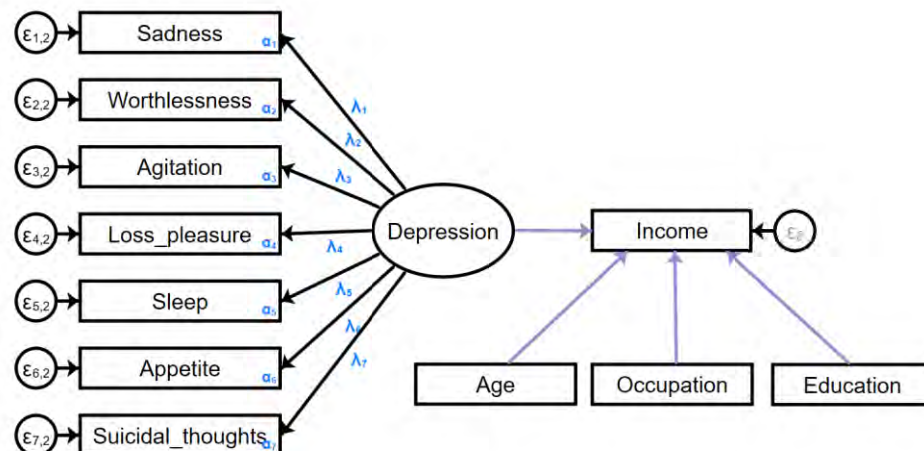
Group 2

What MI/E tells you

- Strong factor invariance:
same factor loadings and
measurement intercepts



Group 1



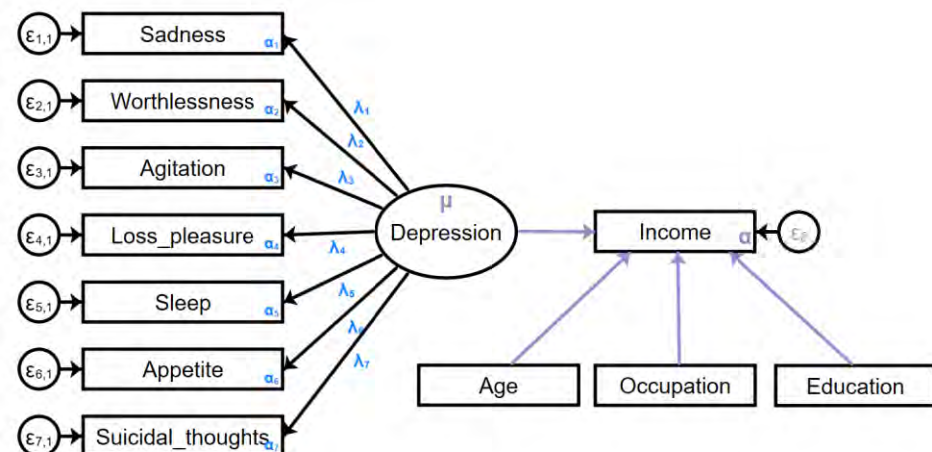
Group 2

What MI/E tells you

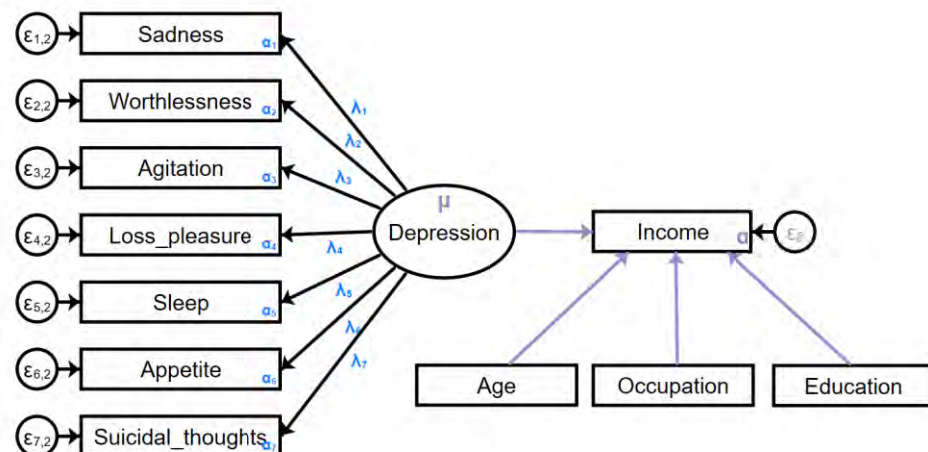
- Strong factor invariance:
same factor loadings and
measurement intercepts



Mean parameter
comparisons are
meaningful



Group 1



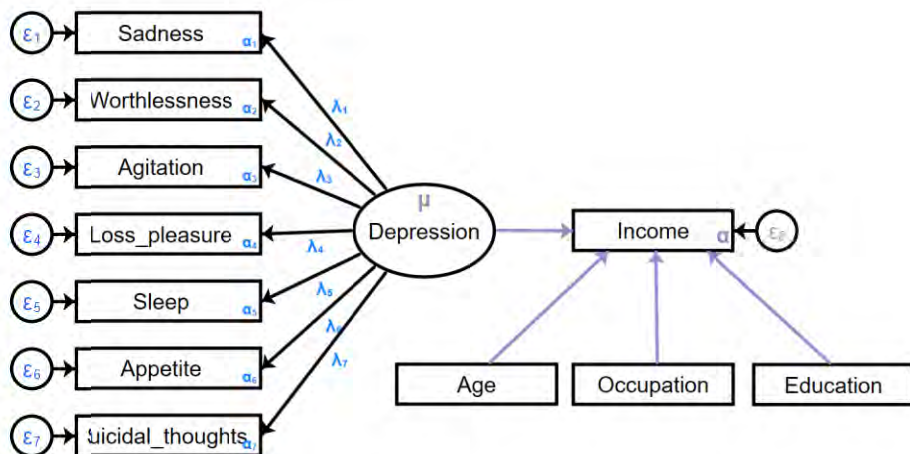
Group 2

What MI/E tells you

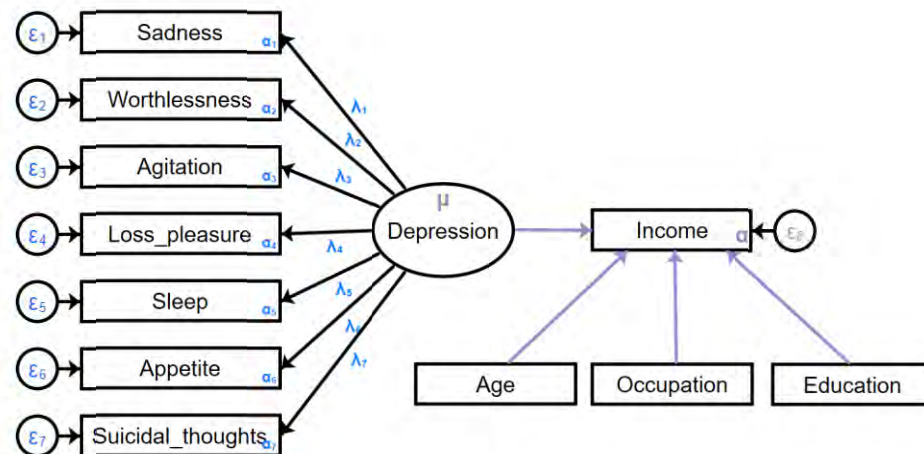
- **Strict factor invariance:**
same factor loadings,
measurement intercepts, and
measurement error variances



Homogeneity of
variance



Group 1



Group 2

Testing MI/E

- Likelihood ratio test
- AIC and BIC
- $\Delta CFI < 0.01$ (Cheung & Rensvold, 2002)
- $\Delta RMSEA < 0.015$, $\Delta SRMR < 0.30/0.15$ (Chen, 2007)
- Equivalence testing (Yuan & Chan, 2016)
- Approximate measurement invariance (i.e., de Schoot et al., 2013)

Application of Confirmatory Factor Analysis to the Study of Self-Concept: First- and Higher Order Factor Models and Their Invariance Across Groups

Herbert W. Marsh
University of Sydney
Sydney, New South Wales, Australia

Dennis Hocevar
University of Southern California

The purpose of this article is to demonstrate the application of confirmatory factor analysis (CFA) in testing first and higher order factor models and their invariance across independent groups. These analyses were performed within the LISREL framework developed by Jöreskog and Sörbom (Jöreskog, 1969, 1970, 1971, 1981; Jöreskog & Sörbom, 1983). LISREL (linear structural relations) is a statistical tool for analyzing covariance matrices according to systems of structural equations. Data from a study designed to test Shavelson's (Shavelson & Bolus, 1982; Shavelson, Hubner, & Stanton, 1976) theoretical predictions about the structure of self-concept were used in the demonstration. The advantages of confirmatory factor analytic procedures in the analyses of higher order factor models and factorial invariance are illustrated and discussed. Issues and problems encountered in the analyses are also discussed, and new solutions to some of these are proposed.

Self-concept

Self-description questionnaire (SDQ; Marsh et al., 1983)

First-order Factors

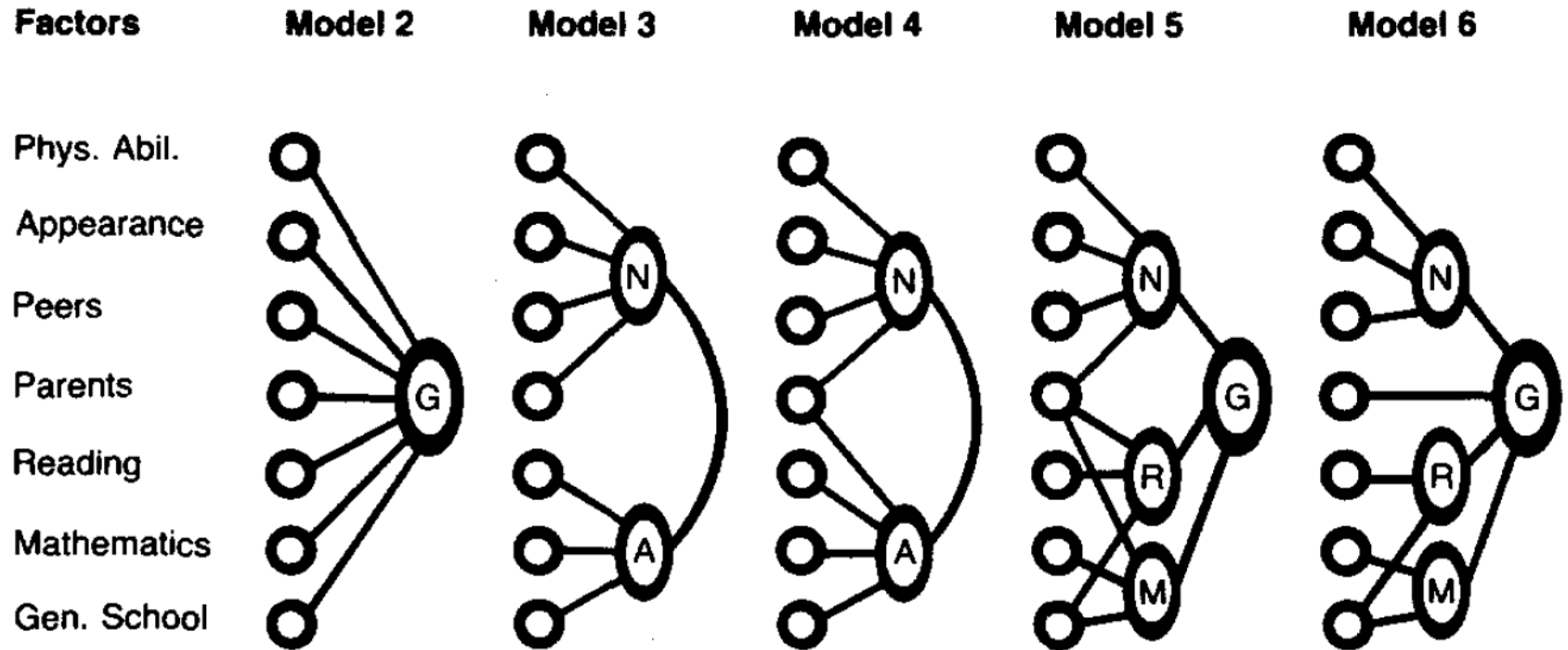


Figure 1. Higher order factor structures for Models 2 to 6.

The dataset

```
. webuse sem_2fmmby  
(Two-factor CFA)
```

```
. notes
```

```
_dta:
```

1. Source: Summary statistics data from Marsh, H. W., and D. Hocevar. 1985. Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. Psychological Bulletin 97: 562-582. <http://doi.org/10.1037/0033-2909.97.3.562>.
2. Summary statistics based on 134 students in grade 4 and 251 students in grade 5 from Sydney, Australia.
3. Group 1 is grade 4, group 2 is grade 5.
4. Data collected using the Self-Description Questionnaire and includes sixteen subscales designed to measure nonacademic status: four intended to measure physical ability, four intended to measure physical appearance, four intended to measure relations with peers, and four intended to measure relations with parents.

The dataset

```
. ssd describe
```

Summary statistics data from https://www.stata-press.com/data/r17/sem_2fmmby.dta

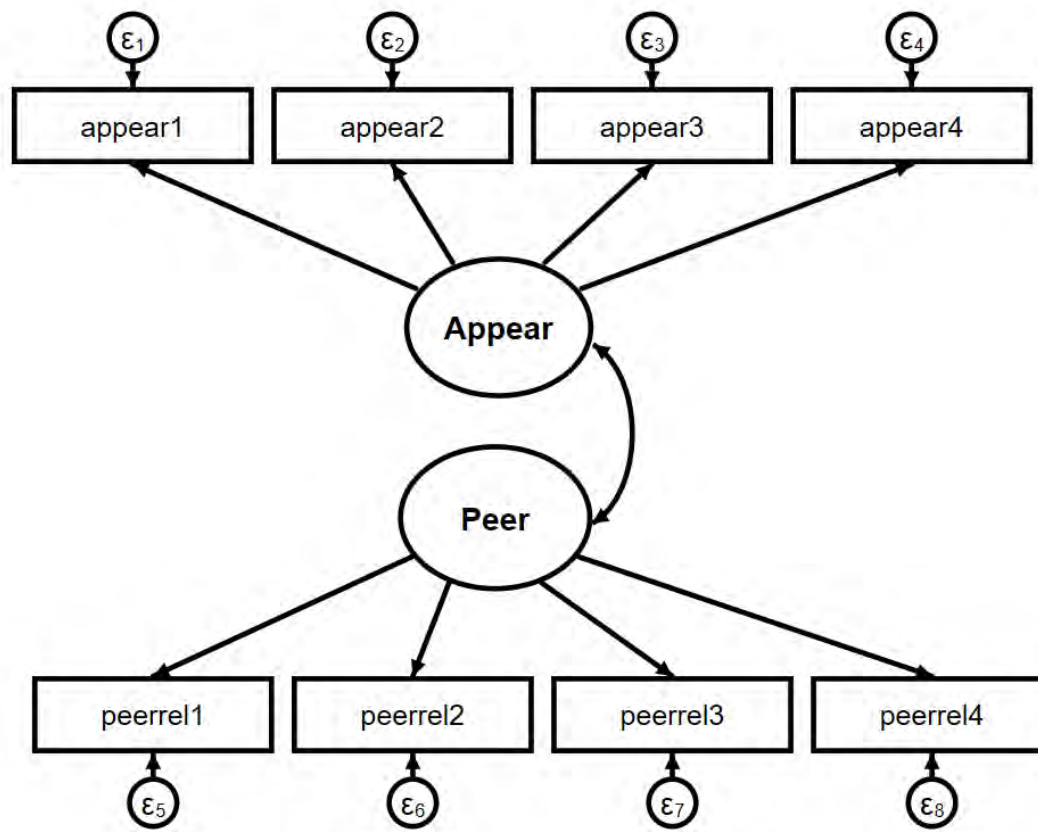
```
Observations:      385      Two-factor CFA
Variables:         16      25 May 2020 11:11
                        (_dta has notes)
```

Variable name	Variable label
phyab1	Physical ability 1
phyab2	Physical ability 2
phyab3	Physical ability 3
phyab4	Physical ability 4
appear1	Appearance 1
appear2	Appearance 2
appear3	Appearance 3
appear4	Appearance 4
peerrel1	Relationship w/ peers 1
peerrel2	Relationship w/ peers 2
peerrel3	Relationship w/ peers 3
peerrel4	Relationship w/ peers 4
parrel1	Relationship w/ parent 1
parrel2	Relationship w/ parent 2
parrel3	Relationship w/ parent 3
parrel4	Relationship w/ parent 4

Group variable: grade (2 groups)

Obs by group: 134, 251

Model





SDQ items

FALSE

MOSTLY FALSE

SOMETIMES FALSE
SOMETIMES TRUE

MOSTLY TRUE

TRUE

Appearance

I am good looking

I'm an attractive person

I'm better looking than most of my friends

I like the way I look

Peer relationships

I make friends easily

I have lots of friends

I am popular with kids my own age

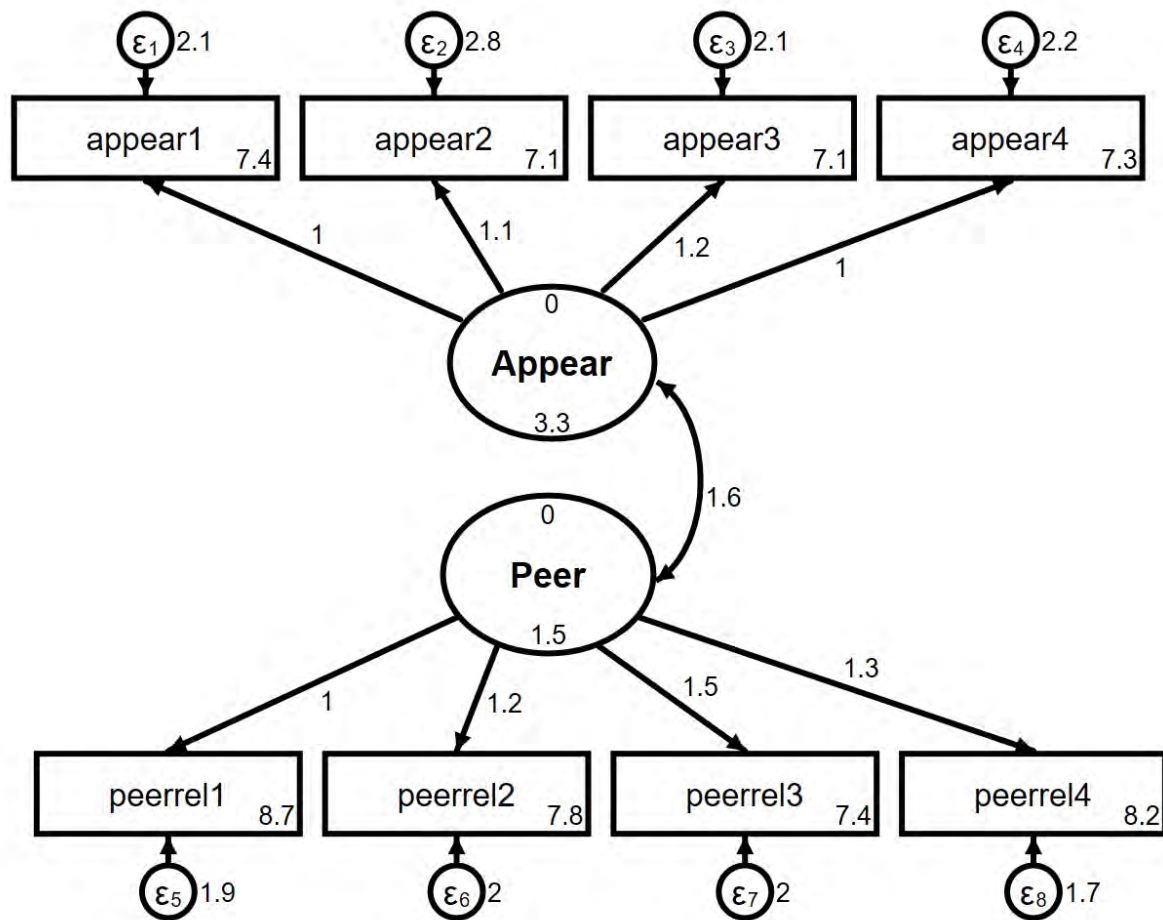
I am easy to like



Research questions

1. Are relationships with peers stronger in 5th grade than in 4th grade?
2. Are physical appearance ratings higher in 5th grade than in 4th grade?
3. Is physical appearance more strongly associated with peer relationships in 5th grade than in 4th grade?

Overall model

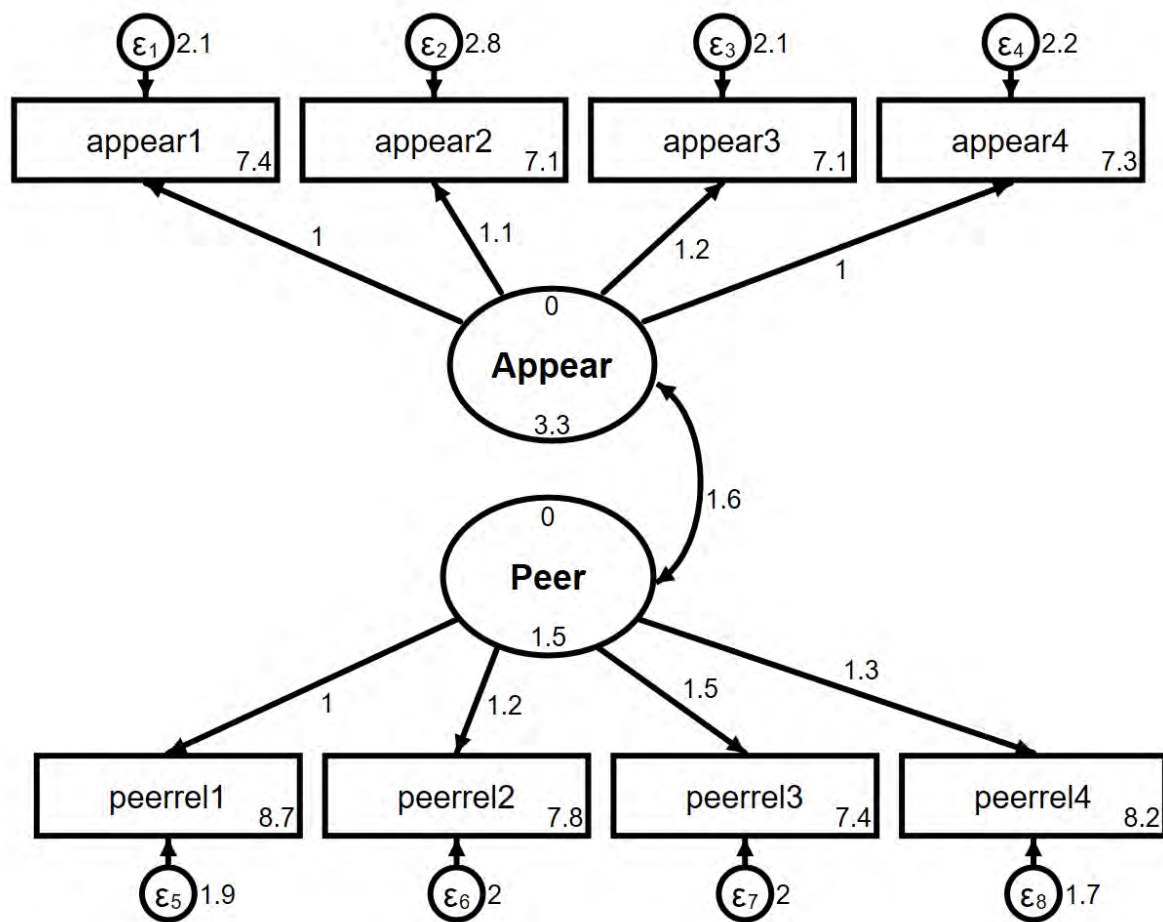


Overall model

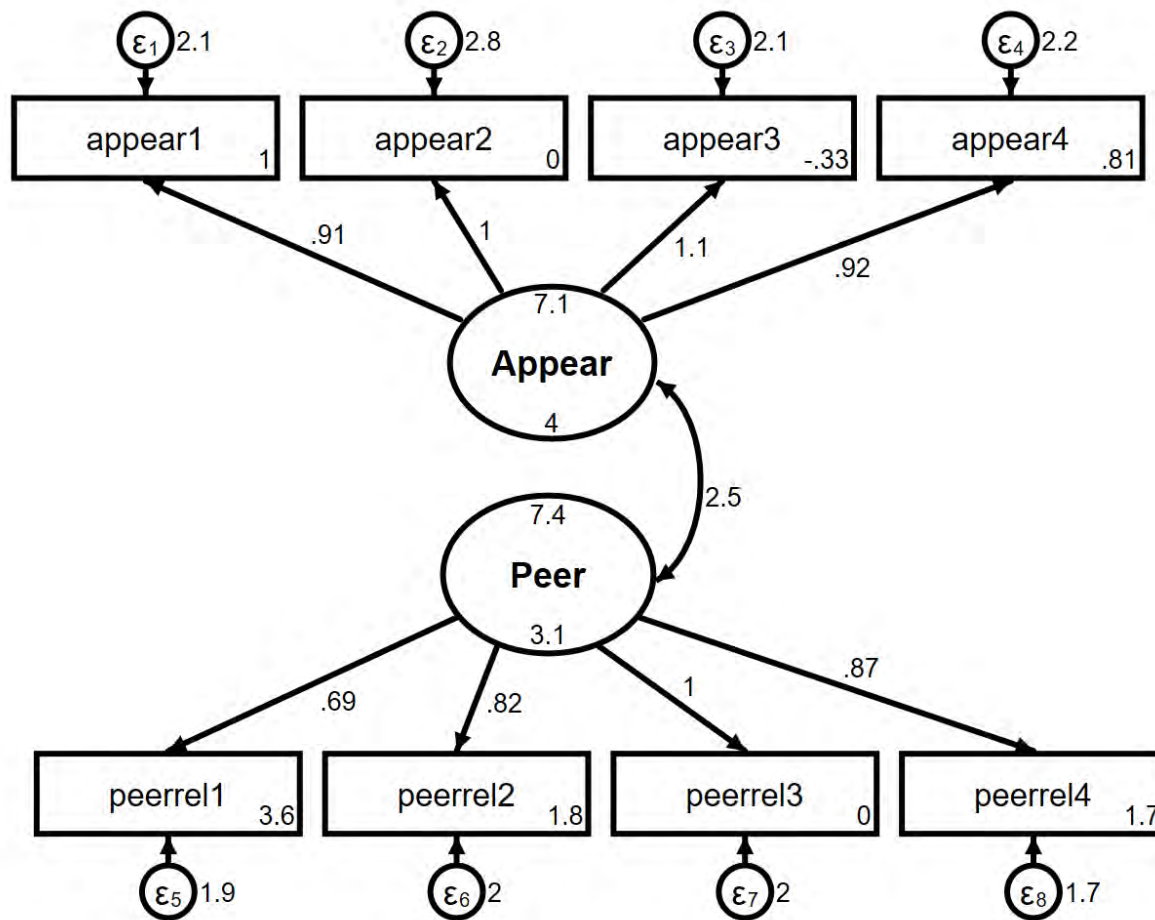
. sem (Peer -> peerrel1-peerrel4) (Appear -> appear1-appear4)

		OIM				
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Measurement						
peerrel1						
Peer	1	(constrained)				
_cons	8.681221	.0937197	92.63	0.000	8.497534	8.864908
peerrel2						
Peer	1.196995	.1043371	11.47	0.000	.9924981	1.401492
_cons	7.828623	.1037547	75.45	0.000	7.625268	8.031979
peerrel3						
Peer	1.451885	.1173682	12.37	0.000	1.221848	1.681922
_cons	7.359896	.1149905	64.00	0.000	7.134519	7.585273
peerrel4						
Peer	1.262761	.1036011	12.19	0.000	1.059706	1.465815
_cons	8.150779	.1023467	79.64	0.000	7.950183	8.351375

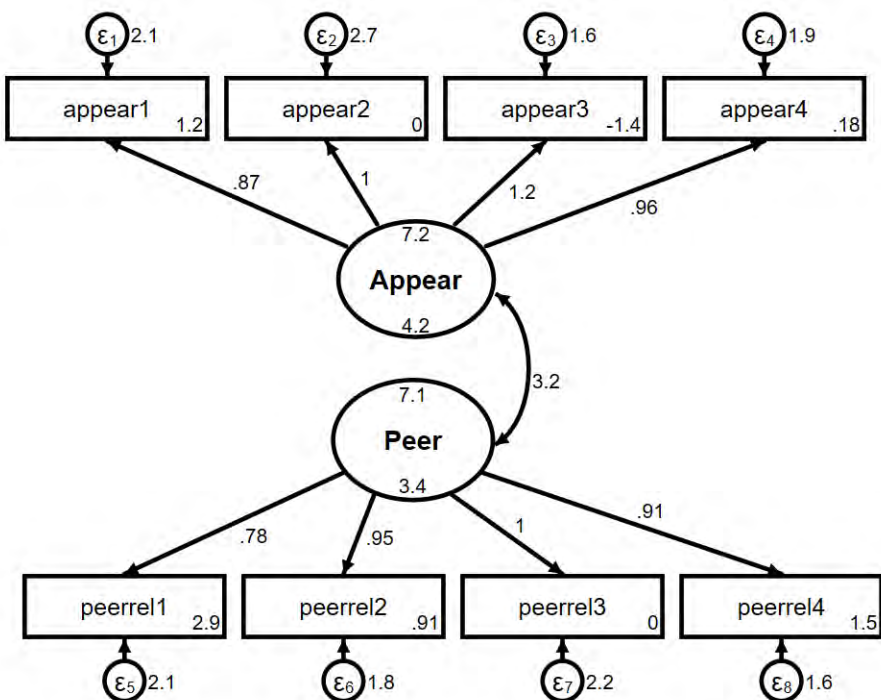
Overall model



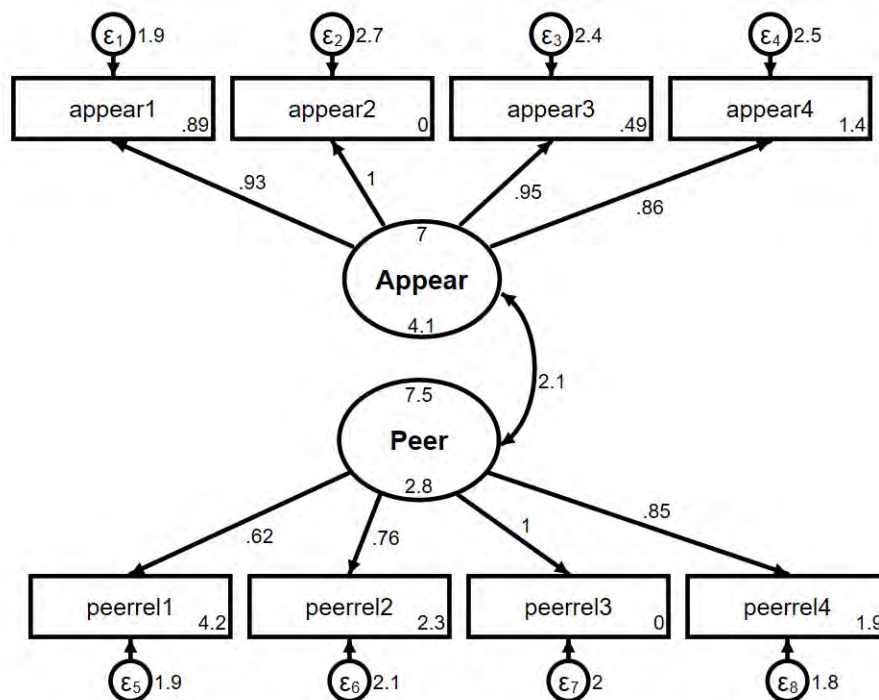
Overall model



Configural invariance



Grade 4



Grade 5

```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(38)	122.802	model vs. saturated
p > chi2	0.000	
chi2_bs(56)	1537.655	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.108	Root mean squared error of approximation
90% CI, lower bound	0.087	
upper bound	0.129	
Information criteria		
AIC	12350.418	Akaike's information criterion
BIC	12548.080	Bayesian information criterion
Baseline comparison		
CFI	0.943	Comparative fit index
TLI	0.916	Tucker-Lewis index
Size of residuals		
SRMR	0.051	Standardized root mean squared residual
CD	0.962	Coefficient of determination

. estat mindices

Modification indices

		MI	df	P>MI	EPC	Standard EPC
Measurement						
appear1	Peer 2	6.213	1	0.01	-.2640433	-.1885013
appear4	Peer 2	8.972	1	0.00	.3290841	.2349341
peerrel1	Appear 1	13.069	1	0.00	-.6260441	-.6288217
peerrel2	Appear 1	12.702	1	0.00	.6564658	.6088857
cov(e.appear1,e.appear2)	1	14.137	1	0.00	.9626506	.411002
	2	49.703	1	0.00	1.746381	.7620227
cov(e.appear1,e.appear3)	2	6.276	1	0.01	-.591031	-.2728838
cov(e.appear1,e.appear4)						

. estat mindices

Modification indices

		MI	df	P>MI	EPC	Standard EPC
Measurement						
appear1	Peer 2	6.213	1	0.01	-.2640433	-.1885013
appear4	Peer 2	8.972	1	0.00	.3290841	.2349341
peerrel1	Appear 1	13.069	1	0.00	-.6260441	-.6288217
peerrel2	Appear 1	12.702	1	0.00	.6564658	.6088857
cov(e.appear1,e.appear2)	1	14.137	1	0.00	.9626506	.411002
	2	49.703	1	0.00	1.746381	.7620227
cov(e.appear1,e.appear3)	2	6.276	1	0.01	-.591031	-.2728838
cov(e.appear1,e.appear4)						

SDQ items

FALSE

MOSTLY FALSE

SOMETIMES FALSE
SOMETIMES TRUE

MOSTLY TRUE

TRUE

Appearance

I am good looking

I'm an attractive person

I'm better looking than most of my friends

I like the way I look

Peer relationships

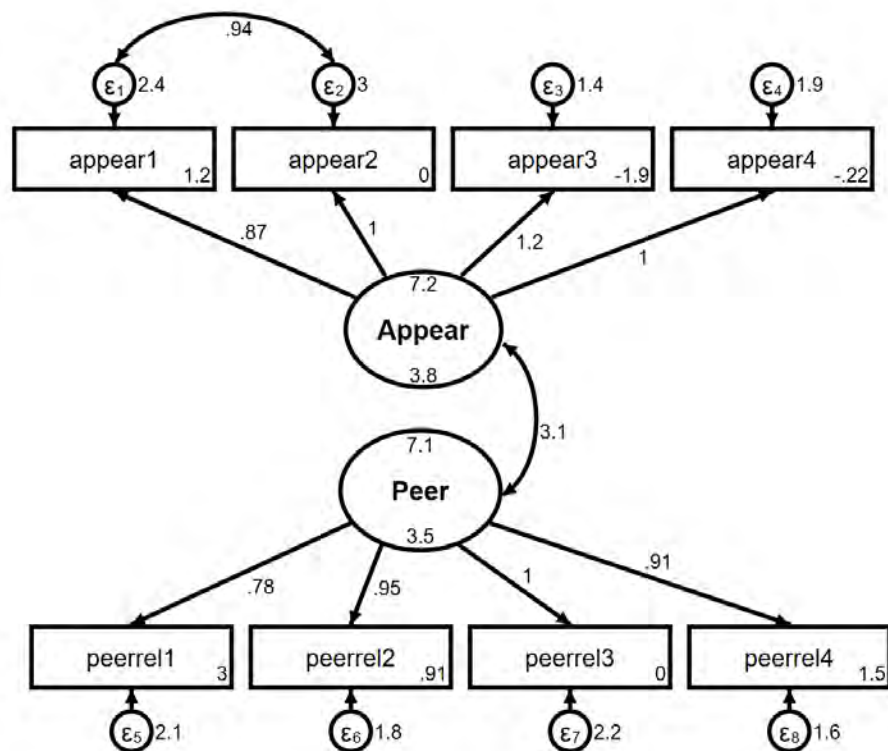
I make friends easily

I have lots of friends

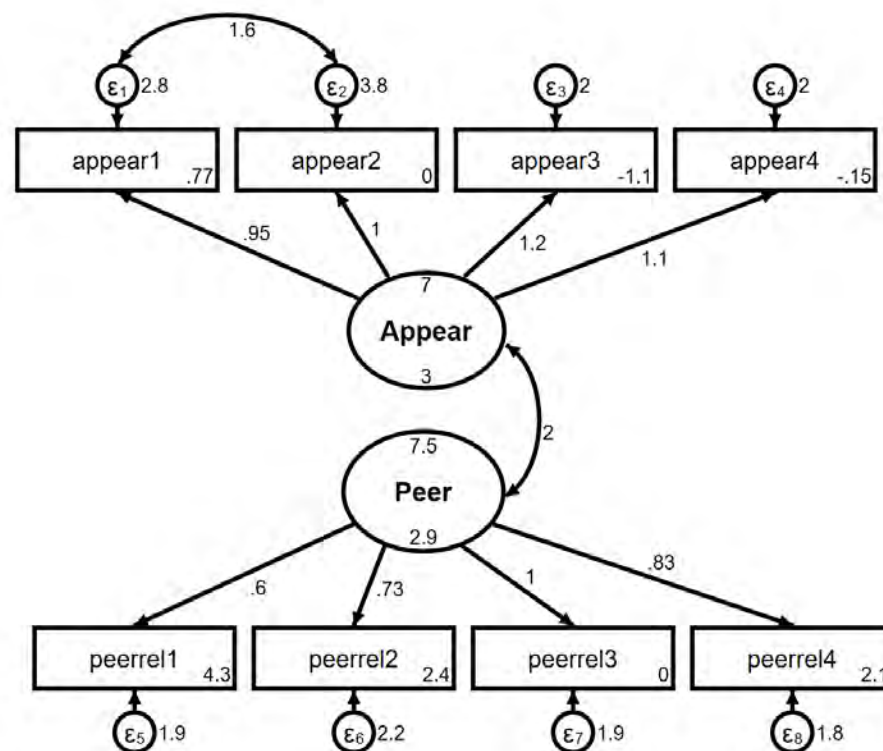
I am popular with kids my own age

I am easy to like

Configural invariance



Grade 4



Grade 5

```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(36)	66.634	model vs. saturated
p > chi2	0.001	
chi2_bs(56)	1537.655	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.067	Root mean squared error of approximation
90% CI, lower bound	0.041	
upper bound	0.091	
Information criteria		
AIC	12298.250	Akaike's information criterion
BIC	12503.819	Bayesian information criterion
Baseline comparison		
CFI	0.979	Comparative fit index
TLI	0.968	Tucker-Lewis index
Size of residuals		
SRMR	0.039	Standardized root mean squared residual
CD	0.955	Coefficient of determination

Configural invariance

```
. collect get r(chi2_ms) r(df_ms) r(p_ms) r(rmse) r(srmmr) ///  
> r(cfi) r(tli) r(aic) r(bic), tags(Model["Configural"])
```

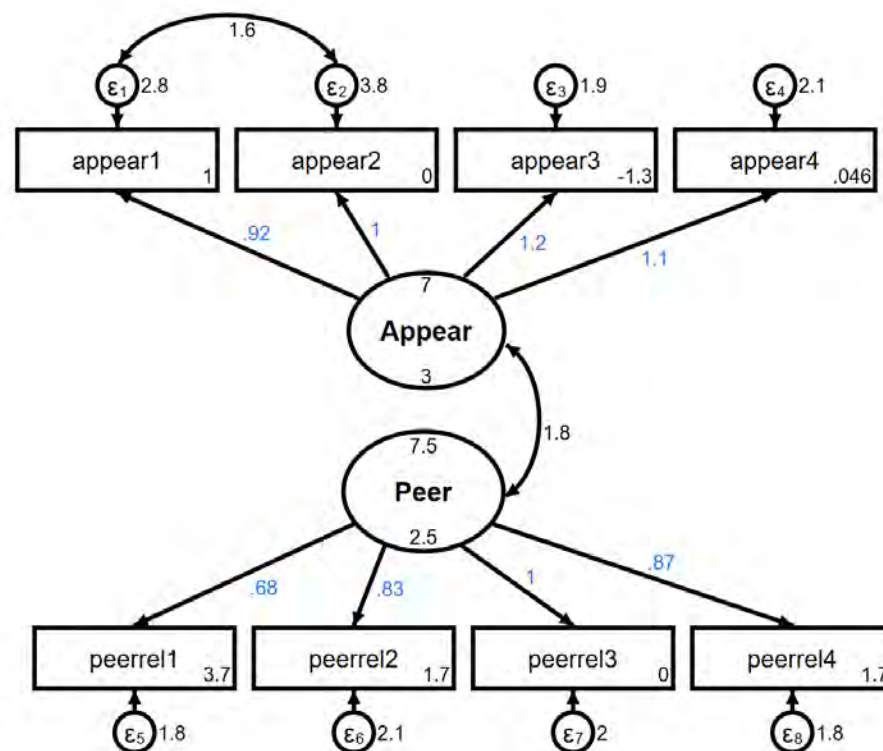
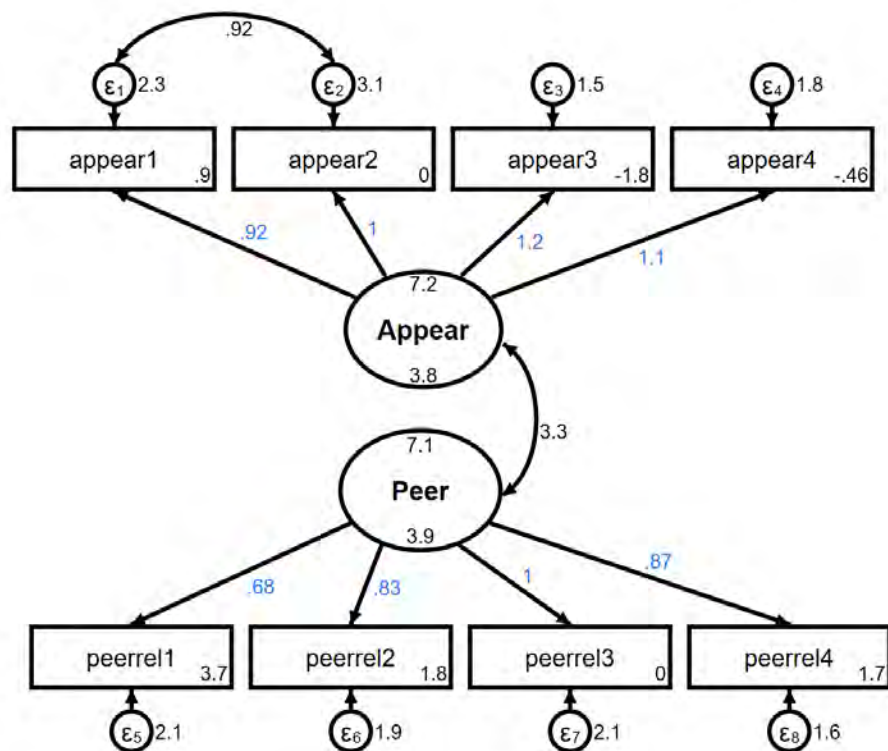
```
. estat ggof
```

Group-level fit statistics

	N	SRMR	CD	chi2	df	p>chi2
grade						
1	134	0.046	0.965	37.978	18	0.004
2	251	0.032	0.952	28.657	18	0.053

```
. estimates store config
```

Weak factor invariance



```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(42)	72.146	model vs. saturated
p > chi2	0.003	
chi2_bs(56)	1537.655	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.061	Root mean squared error of approximation
90% CI, lower bound	0.036	
upper bound	0.085	
Information criteria		
AIC	12291.762	Akaike's information criterion
BIC	12473.611	Bayesian information criterion
Baseline comparison		
CFI	0.980	Comparative fit index
TLI	0.973	Tucker-Lewis index
Size of residuals		
SRMR	0.050	Standardized root mean squared residual
CD	0.955	Coefficient of determination

Weak factor invariance

```
. collect get r(chi2_ms) r(df_ms) r(p_ms) r(rmse) r(srmr) ///  
> r(cfi) r(tli) r(aic) r(bic), tags(Model["Weak"])
```

```
. estimates store weak
```

```
. lrtest weak config
```

Likelihood-ratio test

Assumption: weak nested within config

```
LR chi2(6) =    5.51  
Prob > chi2 = 0.4800
```

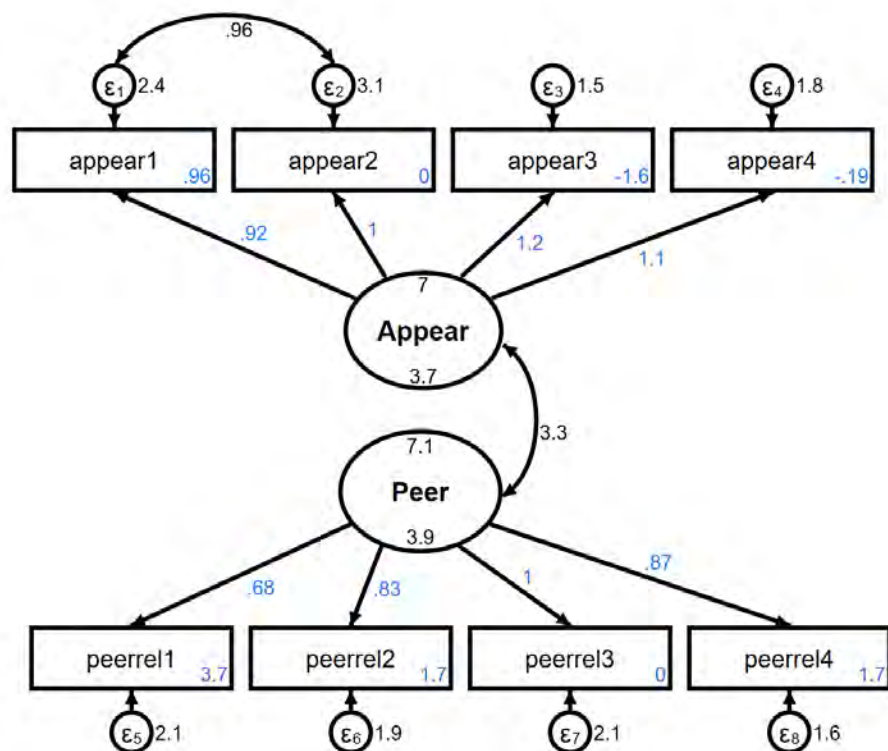
```
. collect get r(chi2) r(df) r(p), tags(Model["Weak"])
```

Weak factor invariance

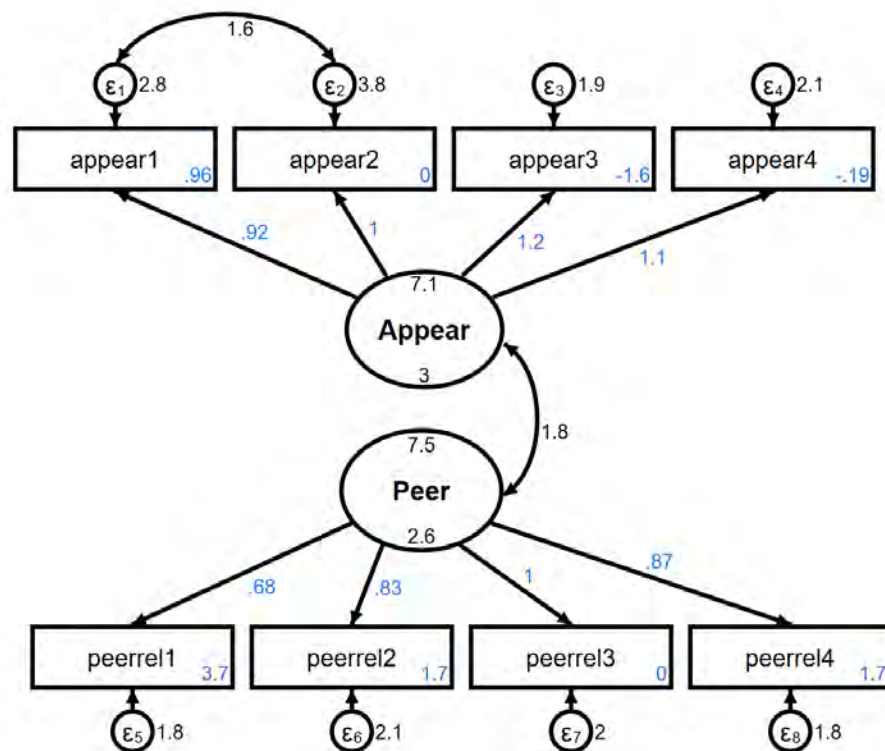
```
. collect style use minv_sty  
. collect label use minv_lab  
. collect layout (result) (Model)
```

	Configural	Weak
Chi-squared	66.6	72.1
df	36	42
p	0.001	0.003
RMSEA	0.067	0.061
SRMR	0.039	0.050
CFI	0.979	0.980
TLI	0.968	0.973
AIC	12298.3	12291.8
BIC	12503.8	12473.6
Δ Chi-squared		5.5
Δ df		6
Δ p		0.480

Strong factor invariance



Grade 4



Grade 5

```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(48)	76.652	model vs. saturated
p > chi2	0.005	
chi2_bs(56)	1537.655	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.056	Root mean squared error of approximation
90% CI, lower bound	0.031	
upper bound	0.078	
Information criteria		
AIC	12284.268	Akaike's information criterion
BIC	12442.397	Bayesian information criterion
Baseline comparison		
CFI	0.981	Comparative fit index
TLI	0.977	Tucker-Lewis index
Size of residuals		
SRMR	0.050	Standardized root mean squared residual
CD	0.955	Coefficient of determination

Strong factor invariance

```
. collect get r(chi2_ms) r(df_ms) r(p_ms) r(rmse) r(srmr) ///  
> r(cfi) r(tli) r(aic) r(bic), tags(Model["Strong"])
```

```
. estimates store strong
```

```
. lrtest weak strong
```

Likelihood-ratio test

Assumption: strong nested within weak

```
LR chi2(6) =    4.51  
Prob > chi2 = 0.6086
```

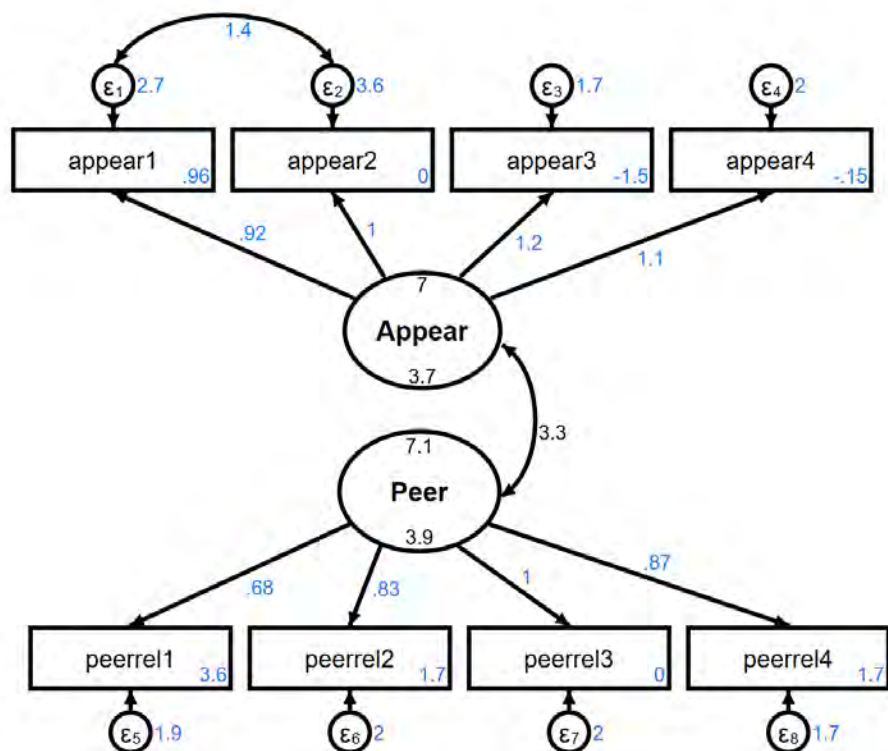
```
. collect get r(chi2) r(df) r(p), tags(Model["Strong"])
```

Strong factor invariance

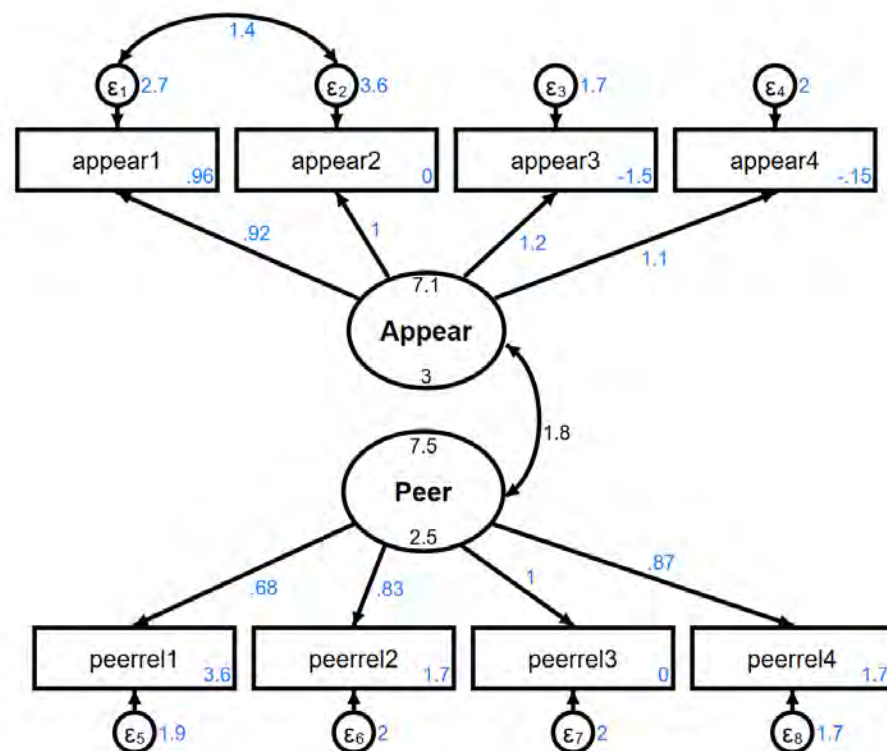
. collect preview

	Configural	Weak	Strong
Chi-squared	66.6	72.1	76.7
df	36	42	48
p	0.001	0.003	0.005
RMSEA	0.067	0.061	0.056
SRMR	0.039	0.050	0.050
CFI	0.979	0.980	0.981
TLI	0.968	0.973	0.977
AIC	12298.3	12291.8	12284.3
BIC	12503.8	12473.6	12442.4
Δ Chi-squared		5.5	4.5
Δ df		6	6
Δ p		0.480	0.609

Strict factor invariance



Grade 4



Grade 5

```
. estat gof, stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(57)	82.421	model vs. saturated
p > chi2	0.015	
chi2_bs(56)	1537.655	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.048	Root mean squared error of approximation
90% CI, lower bound	0.022	
upper bound	0.070	
Information criteria		
AIC	12272.037	Akaike's information criterion
BIC	12394.588	Bayesian information criterion
Baseline comparison		
CFI	0.983	Comparative fit index
TLI	0.983	Tucker-Lewis index
Size of residuals		
SRMR	0.054	Standardized root mean squared residual
CD	0.957	Coefficient of determination

Strict factor invariance

```
. collect get r(chi2_ms) r(df_ms) r(p_ms) r(rmse) r(srmr) ///  
> r(cfi) r(tli) r(aic) r(bic), tags(Model["Strict"])
```

```
. estimates store strict
```

```
. lrtest strict strong
```

Likelihood-ratio test

Assumption: strict nested within strong

```
LR chi2(9) =    5.77  
Prob > chi2 = 0.7628
```

```
. collect get r(chi2) r(df) r(p), tags(Model["Strict"])
```

Strict factor invariance

. collect preview

	Configural	Weak	Strong	Strict
Chi-squared	66.6	72.1	76.7	82.4
df	36	42	48	57
p	0.001	0.003	0.005	0.015
RMSEA	0.067	0.061	0.056	0.048
SRMR	0.039	0.050	0.050	0.054
CFI	0.979	0.980	0.981	0.983
TLI	0.968	0.973	0.977	0.983
AIC	12298.3	12291.8	12284.3	12272.0
BIC	12503.8	12473.6	12442.4	12394.6
Δ Chi-squared		5.5	4.5	5.8
Δ df		6	6	9
Δ p		0.480	0.609	0.763

Strict factor invariance

. collect preview

	Configural	Weak	Strong	Strict
Chi-squared	66.6	72.1	76.7	82.4
df	36	42	48	57
p	0.001	0.003	0.005	0.015
RMSEA	0.067	0.061	0.056	0.048
SRMR	0.039	0.050	0.050	0.054
CFI	0.979	0.980	0.981	0.983
TLI	0.968	0.973	0.977	0.983
AIC	12298.3	12291.8	12284.3	12272.0
BIC	12503.8	12473.6	12442.4	12394.6
Δ Chi-squared		5.5	4.5	5.8
Δ df		6	6	9
Δ p		0.480	0.609	0.763





Research question (1)

Are relationships with peers stronger in 5th grade than in 4th grade?

Research question (1)

Are relationships with peers stronger in 5th grade than in 4th grade?

```
. estat ginvariant
```

Tests for group invariance of parameters

	Wald test			Score test		
	chi2	df	P>chi2	chi2	df	P>chi2
mean(Peer)	4.392	1	0.0361	.	.	.
mean(Appear)	0.203	1	0.6525	.	.	.

Research question (1)

Are relationships with peers stronger in 5th grade than in 4th grade?

```
. estat ginvariant
```

Tests for group invariance of parameters

	Wald test			Score test		
	chi2	df	P>chi2	chi2	df	P>chi2
mean(Peer)	4.392	1	0.0361	.	.	.
mean(Appear)	0.203	1	0.6525	.	.	.



Yes, relationships with peers in 5th grade (M= 7.52 [7.27,7.77]) were higher than in 4th grade (M= 7.06 [6.68,7.44])

Research question (2)

Are physical appearance ratings higher in 5th grade than in 4th grade?

```
. estat ginvariant
```

Tests for group invariance of parameters

	Wald test			Score test		
	chi2	df	P>chi2	chi2	df	P>chi2
mean(Peer)	4.392	1	0.0361	.	.	.
mean(Appear)	0.203	1	0.6525	.	.	.



No, physical appearance in 5th grade (M= 7.11 [6.82,7.40]) was not different than in 4th grade (M= 7.01 [6.62,7.40])

Research question (3)

Is physical appearance more strongly associated with peer relationships in 5th grade than in 4th grade?

```
. estat ginvariant
```

Tests for group invariance of parameters

	Wald test			Score test		
	chi2	df	P>chi2	chi2	df	P>chi2
cov(Peer, Appear)	6.821	1	0.0090	.	.	.

Research question (3)

Is physical appearance more strongly associated with peer relationships in 5th grade than in 4th grade?

```
. estat ginvariant
```

Tests for group invariance of parameters

	Wald test			Score test		
	chi2	df	P>chi2	chi2	df	P>chi2
cov(Peer, Appear)	6.821	1	0.0090	.	.	.



No, physical appearance is more strongly associated with peer relationships in 4th grade ($\sigma_{12}=3.25$ [2.23,4.27]) than in 5th grade ($\sigma_{12}=1.84$ [1.30,2.38])



What if we didn't have MI/E?

- Modify the scale (take out items)
- Release constraints on items (partial invariance)
 - A factor can be considered partially invariant if the majority of items on the factor are invariant (Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000).
 - As the proportion of noninvariant items on a factor increased, so did the bias (Chen, 2008)
- Modify the comparison groups
- Take a step back and reconsider your assessment



References

- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal*, 14(3), 464-504.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, 95(5), 1005.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, 9(2), 233-255.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, 18(3), 117-144.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41, 71-90.
- Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, 25(1), 78-90.
- Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in psychology*, 4, 770.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1), 4-70.
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). American Psychological Association.
- Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological methods*, 21(3), 405.

Thank you!

Questions?

mcain@stata.com