Confirmatory Factor Analysis

Professor Timothy Bates

tim.bates@ed.ac.uk

http://timbates.wikidot.com/mv-stats

Outline

- Week 1: Factor analysis
  - What is a factor analysis?
  - Factor extraction
  - Factor rotation
  - Factor interpretation
  - Factor Scores
- **Week 2: Confirmatory Factor Analysis**
- Week 3: Path Analysis and SEM
- Week 4: Complex causal modelling
  - Twins, multiple groups…

Last week Summary

- What is factor analysis?
  - Statistical method compactly accounting for variance in observed traits
    - ("observed random variables")
  - In terms of a smaller number of factors
    - ("unobserved random variables")
  - Allows recovery of values for a subject from a linear combination of the extracted factors.
    - (with some error)
- Can think of the factors as independent variables and items as dependent variables

Summary cont.

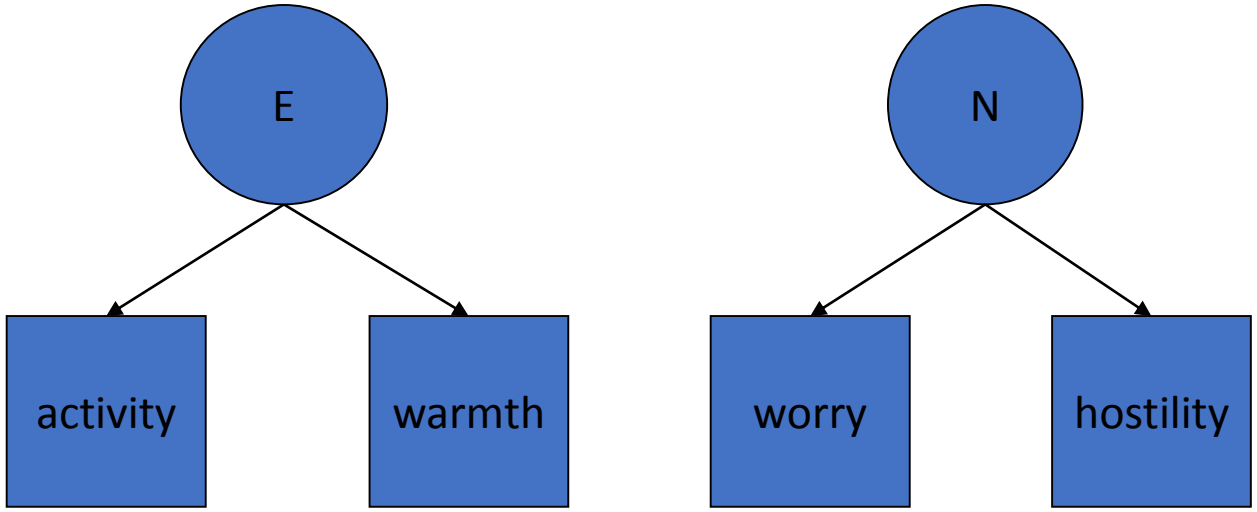
- What is a scree plot?
- What is an identity matrix?
- What are communalities?
- What is a factor loading?
- What is a factor score?
- Bartlett’s test of sphericity?
- KMO?
- “good” number of subjects?
- Why do we rotate factors?
- What rotations are there?
- How can we model and test causes and (model latent structure?)

2nd Week: Structural equation modeling I

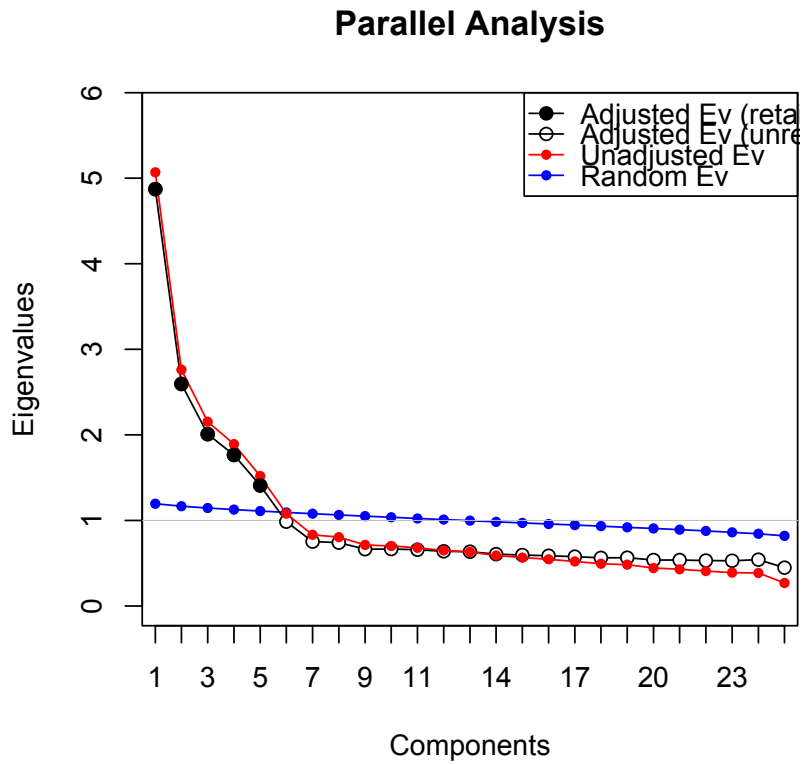
- Placing factor analysis into a confirmatory framework
  - CFA
  - Latent variables
  - Model fit

Structural Equation Modeling & Factor Analysis

- SEM incorporates
  - Measurement model (FA) + Structural model
  - CFA is a measurement model in which each factor (latent variable) has multiple indicators no direct effects (arrows connecting the observed variables)



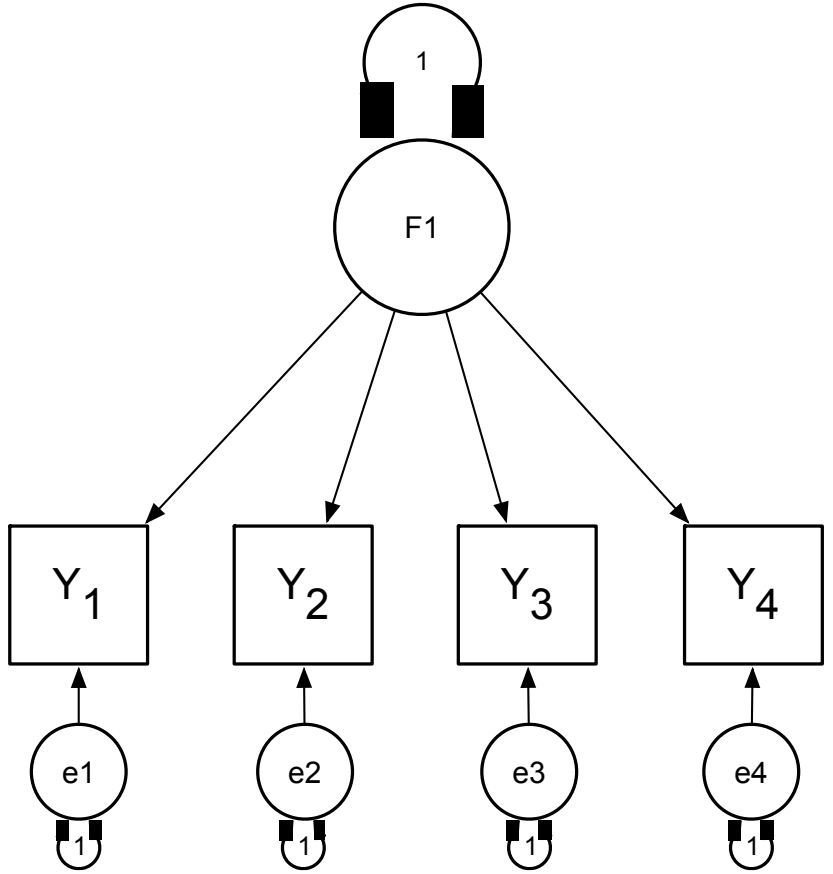
Component	Adjusted-Eigenvalue	Unadjusted-Eigenvalue	Estimated-Bias
1	4.873258	5.068516	0.195257
2	2.595945	2.762479	0.166533
3	2.007994	2.152622	0.144628
4	1.765957	1.892332	0.126375
5	1.408120	1.517532	0.109412



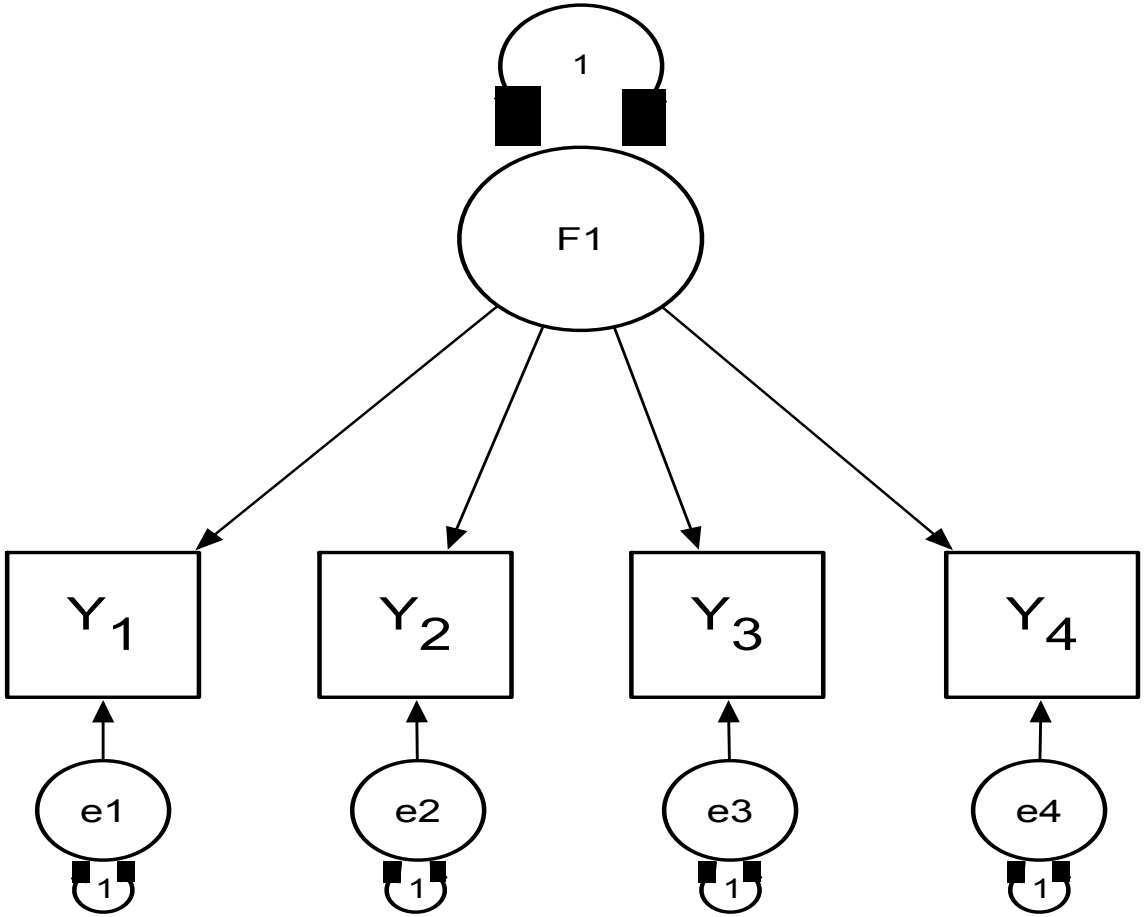
And this

item	Factor1	Factor2	Factor3	Factor4	Factor5
A1	-	-	-	-0.375	-
A2	-	0.195	0.143	0.579	-
A3	-	0.280	0.113	0.649	-
A4	-	0.172	0.226	0.453	-0.132
A5	-0.118	0.337	-	0.581	-
C1	-	-	0.528	0.215	
C2	-	-	0.617	0.137	0.125
C3	-	-	0.556	0.120	-
C4	0.222	-	-0.647	-	-
C5	0.266	-0.193	-0.572	-	-
E1	-0.578	-0.139	-		
E2	0.227	-0.675	-0.100	-0.157	-
E3	0.498	-	0.326	0.311	
E4	-0.123	0.602	-	0.390	-
E5	0.498	0.314	0.128	0.224	
N1	0.814	-	-	-0.208	-
N2	0.783	-	-	-0.203	-
N3	0.717	-	-	-	-
N4	0.563	-0.374	-0.191	-	-
N5	0.521	-0.183	0.109	-0.150	
O1	-	0.176	0.112	-	0.523
O2	0.173	-	-0.115	0.119	-0.467
O3	-	0.273	-	0.149	0.619
O4	0.211	-0.221	-	0.130	0.360
O5	-	-	-	-	-0.524

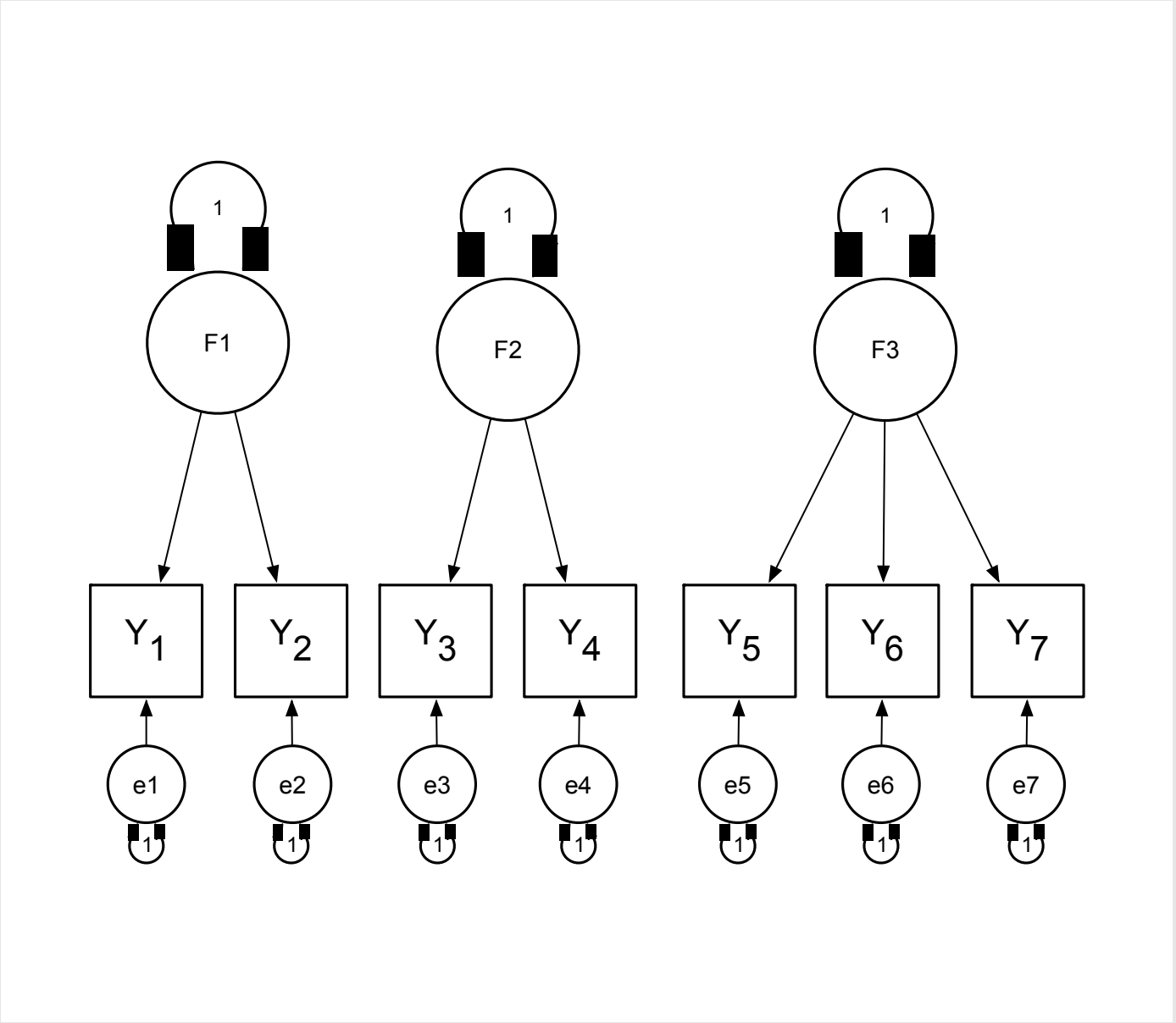
From EFA to CFA
Common Factor as a latent variable



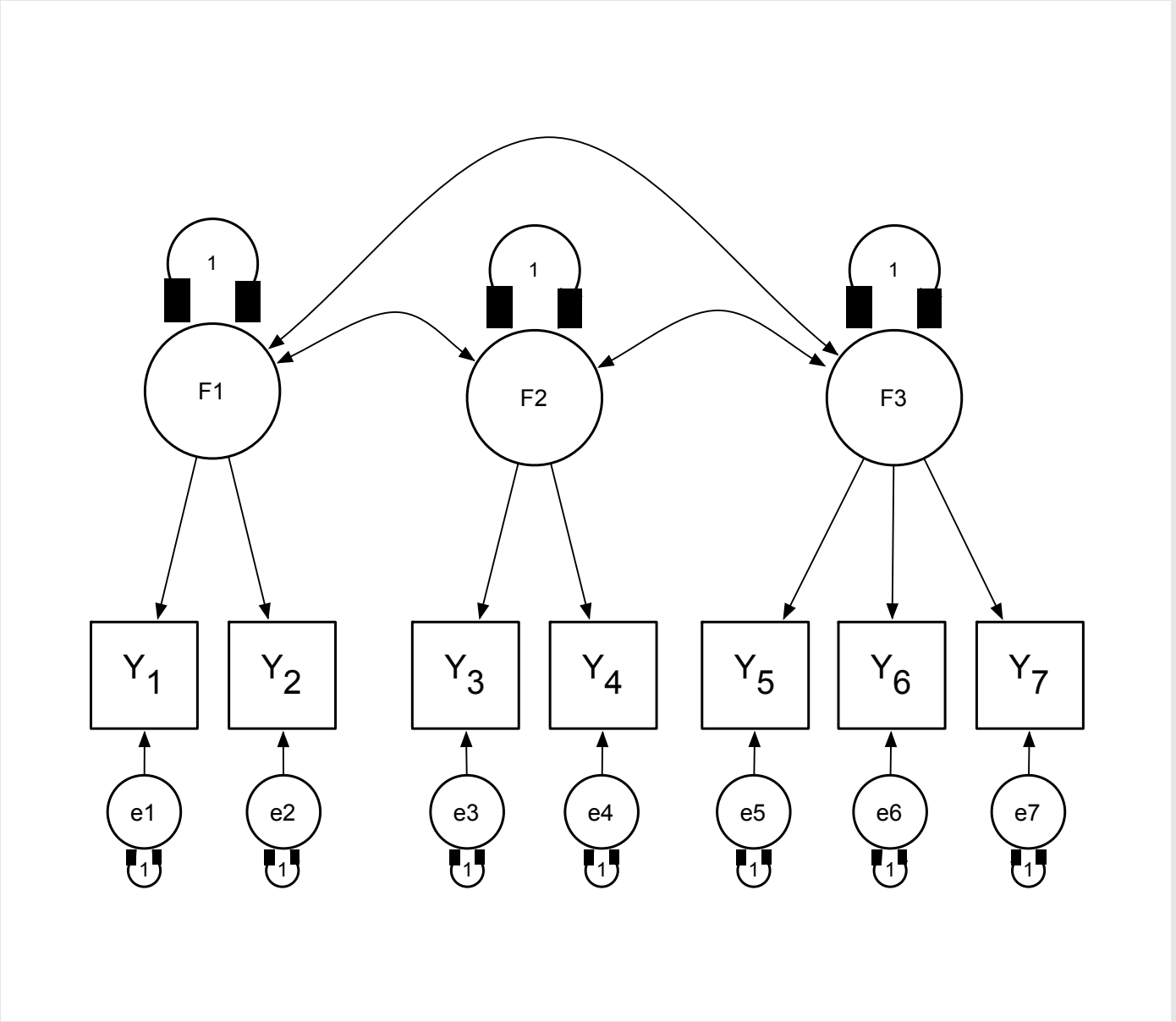
Note, we need to fix the variance of the latent variable (or b1 path)



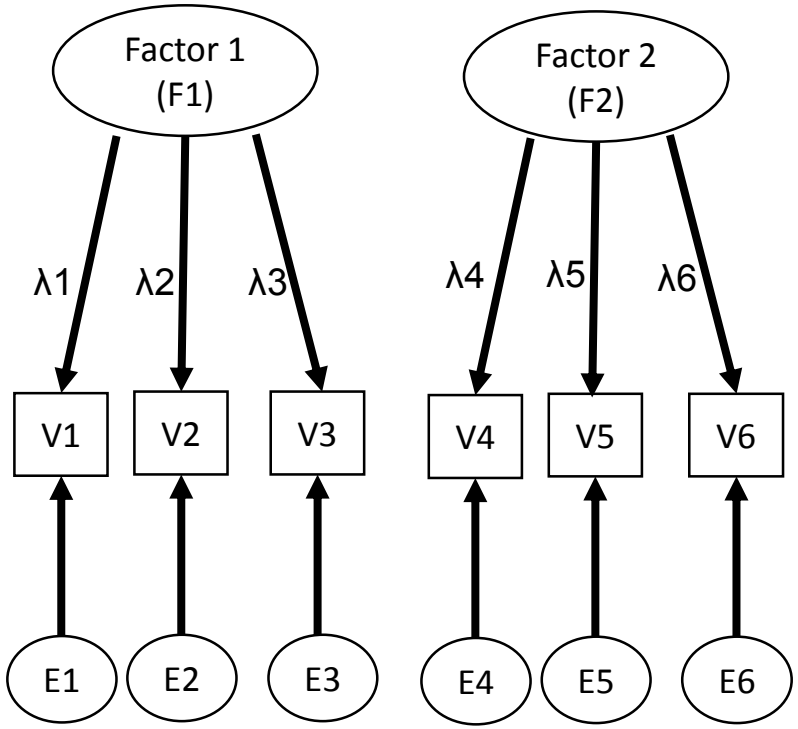
Equivalent of simple structure in an orthogonal rotation



Simple structure, oblique



Structural equations



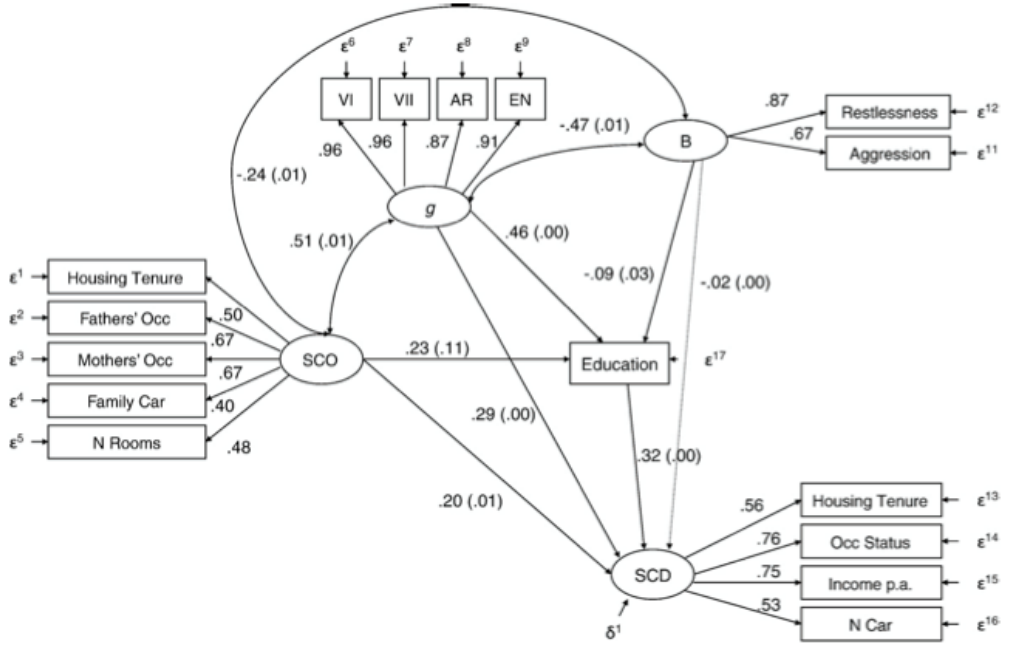
$$V1 = \lambda1 * F1 + 0 * F2 + E1$$
$$V2 = \lambda2 * F1 + 0 * F2 + E2$$
$$V3 = \lambda3 * F1 + 0 * F2 + E3$$
$$V4 = 0 * F1 + \lambda4 * F2 + E4$$
$$V5 = 0 * F1 + \lambda5 * F2 + E5$$
$$V6 = 0 * F1 + \lambda6 * F2 + E6$$

Measurement model

- **Indicator** variables = squares to which the latent variable points
  - Manifest variables
- **Loading matrix** = path values b1-b4 pointing away from the latent variable
- Jointly form the measurement model
- The meaning of the latent variable lies in the fact that it accounts for what is common to all the variables to which its out-going paths point.

Factor Analysis & Path Analysis

- SEM can be extended beyond factor analysis, where measurement instruments are modeled, to path analysis, where causal inferences are made, and models with both latent variables (factors) and paths specified connecting the latent variables



Putting the C in CFA

- When we run a CFA, we get back various measures of “fit”
  - RMSEA
  - CFI, TLI
  - χ^2
- Critically, we can compare the fit of models to each other
  - p-values on model fit
- A bad model will be disconfirmed
- A better model will fit significantly better

Model evaluation

- Goodness of fit of the estimated to the observed covariance matrix
- Many fit indices (need to consider several):
  - X^2-test should yield $p > 0.05$
    - Not a helful index: almost never true in large samples
  - Root Mean Square Error of Approximation (RMSEA) ≤ 0.06
  - Comparative Fit Index (CFI) ≥ 0.97
  - Tucker-Lewis Index TLI) ≥ 0.95
  - Akaike's Information Criterion (AIC) (smaller is better)
  - Bayes Information Criterion (BIC) (should be small)

Model identification

- Observed information:
  Variances and covariances of the variables
  = # variables * (# variables + 1) / 2
- *Degrees of freedom (df)*
  = observed information – estimated parameters
  - df < 0: *under-identified* model, cannot be estimated
  - df = 0: *just identified (saturated)* model, parameters can be estimated, but not tested for significance
  - df > 1: *over-identified* model, parameters can be estimated and the overall fit of the model can be tested
- Parameters can be *fixed* to a value or be dropped to increase dfs
  - Loadings of errors are fixed to 1
  - Either the variance of the factors or the loading of one variable per factor is usually fixed to 1

Model comparisons

- *Nested models* (i.e., simplified subset models with fixed or dropped parameters) can be compared against the original using the X^2 difference test and relative fit indices like AIC or BIC
- *Non-nested models* can only be compared using relative fit indices (AIC, BIC)
- *Multigroup comparisons* test if the same parameter estimates generalize across multiple groups (genders, time points, independent samples etc.)
  - Individual parameters can be set to vary between groups

How is this actually done? Parameter estimation

- Usually maximum-likelihood estimation
- *Iterative* estimation of observed covariance matrix from structural equations
- Structural equations require starting values, which might be adjusted to help convergence
- Convergence can fail due to too small sample, lack of multivariate normal distribution, collinearity, misspecification of model etc.

Ordinal data: Questionnaire items

- Pearson’s correlations assume continuous variables
- Items are likely best thought of as ordinal
- We can model this explicitly

Confirmatory Factor Analysis can test theory

- Are there 5 independent domains of personality?
- Do domains have facets?
- Is Locus of control the same as primary control?
- Is prosociality one thing , or three?
- Is Grit independent of Conscientiousness?

Buss & Perry (1992)

- 4 factors
- 29 items

Table 1
Four Aggression Factors

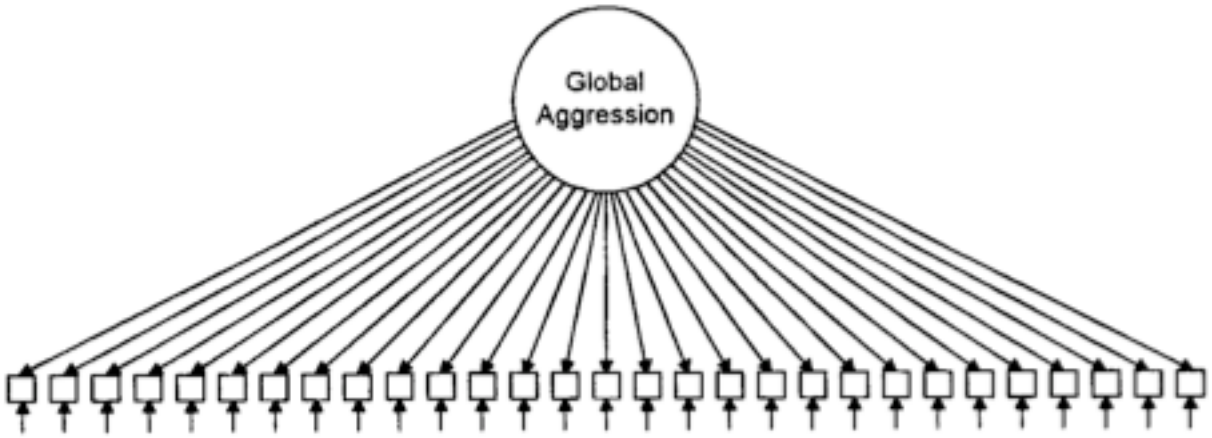
Factor	Factor loadings
Physical Aggression	
1. Once in a while I can't control the urge to strike another person.	.66, .55, .62
2. Given enough provocation, I may hit another person.	.79, .84, .80
3. If somebody hits me, I hit back.	.60, .65, .60
4. I get into fights a little more than the average person.	.44, .52, .58
5. If I have to resort to violence to protect my rights, I will.	.63, .68, .58
6. There are people who pushed me so far that we came to blows.	.60, .62, .65
7. I can think of no good reason for ever hitting a person.[a]	.47, .53, .51
8. I have threatened people I know.	.45, .48, .65
9. I have become so mad that I have broken things.	.47, .57, .47
Verbal Aggression	
1. I tell my friends openly when I disagree with them.	.41, .41, .48
2. I often find myself disagreeing with people.	.38, .49, .35
3. When people annoy me, I may tell them what I think of them.	.45, .45, .40
4. I can't help getting into arguments when people disagree with me.	.38, .41, .36
5. My friends say that I'm somewhat argumentative.	.37, .56, .46
Anger	
1. I flare up quickly but get over it quickly.	.53, .49, .49
2. When frustrated, I let my irritation show.	.47, .45, .37
3. I sometimes feel like a powder keg ready to explode.	.60, .35, .35
4. I am an even-tempered person.[a]	.64, .62, .69
5. Some of my friends think I'm a hothead.	.63, .51, .64
6. Sometimes I fly off the handle for no good reason.	.75, .64, .70
7. I have trouble controlling my temper.	.74, .66, .69
Hostility	
1. I am sometimes eaten up with jealousy.	.41, .43, .49
2. At times I feel I have gotten a raw deal out of life.	.61, .58, .52
3. Other people always seem to get the breaks.	.65, .65, .63
4. I wonder why sometimes I feel so bitter about things.	.48, .45, .59
5. I know that "friends" talk about me behind my back.	.55, .37, .47
6. I am suspicious of overly friendly strangers.	.42, .35, .43
7. I sometimes feel that people are laughing at me behind my back.	.66, .64, .70
8. When people are especially nice, I wonder what they want.	.55, .50, .47

[a] The scoring of these items is reversed.

Measurement invariance

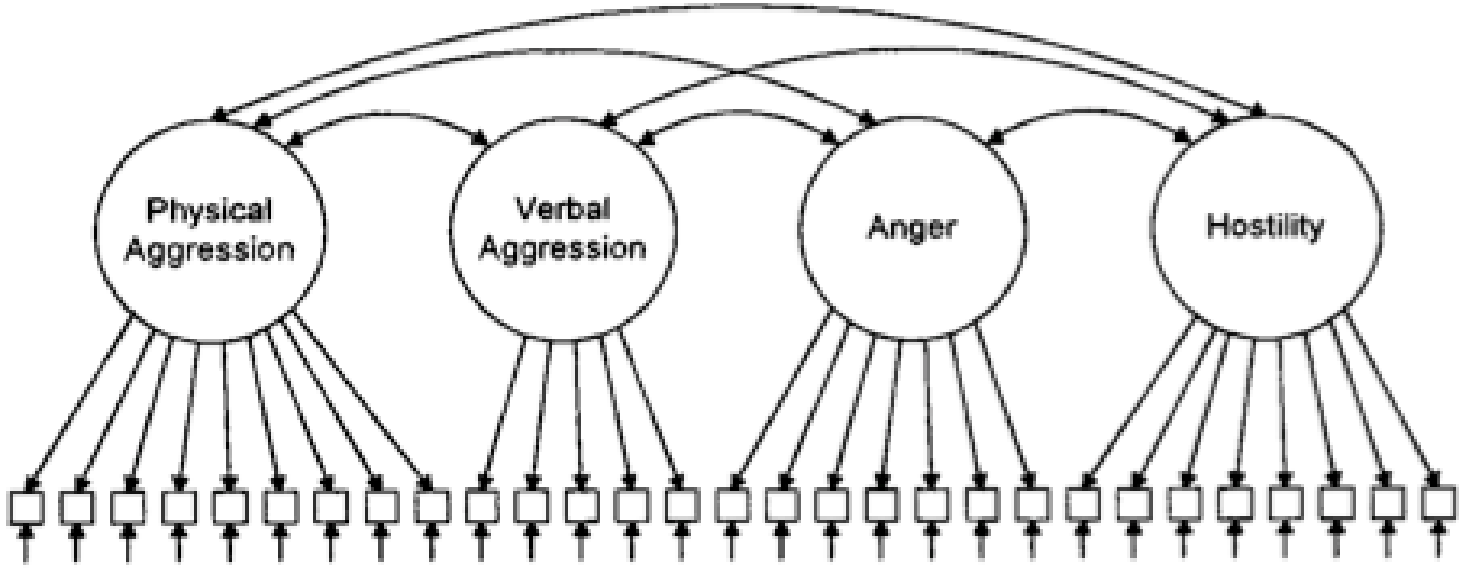
- Psychometric measures should assess the same latent variables across populations and time: measurement invariance
- Can be tested with multi-group comparisons
- Four levels of measurement invariance:
  - Configural invariance: The factor structure (pattern of fixed and free parameters) is the same
  - Weak factorial invariance: The standardized factor loadings are also equal across groups
  - Strong factorial invariance: The variable means are also equal across groups
  - Strict factorial invariance: The error terms are also exactly equal across groups (not to be expected!)

Bryant & Smith (2001)



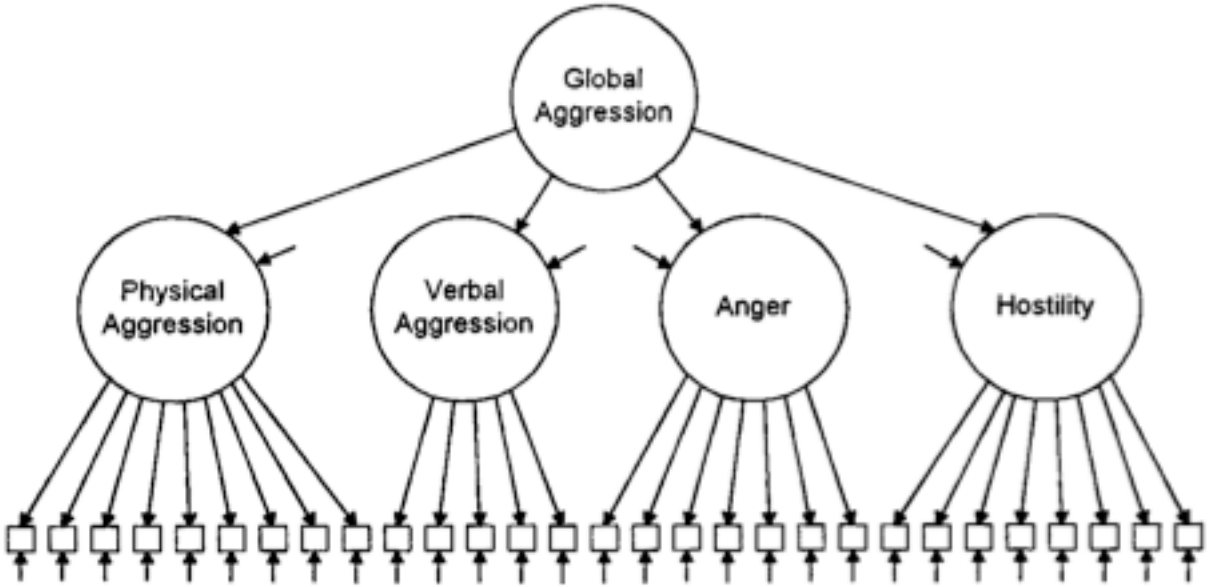
Bryant & Smith (2001)

- Oblique factors = non-zero two-headed arrows between circles
- Factor analysis = residual variance on manifest variables



Bryant & Smith (2001)

- Hierarchical factor analysis: factors only correlate via links to a common, higher-level, more general factor



Bryant & Smith (2001)

- 12 items, with better fit and cross-cultural reliability

TABLE 3
CFA Factor Loadings for the Refined 12-Item, Four-Factor Measurement Model of the AQ

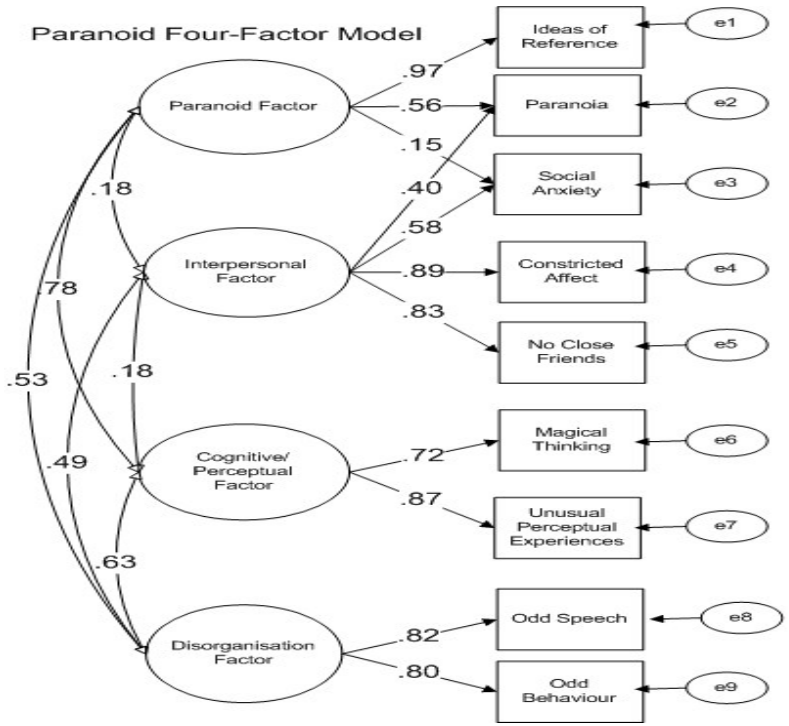
	PA sample			VA sample			ANG sample			HO sample		
AQ items	1	2	3	1	2	3	1	2	3	1	2	3
2. Given enough provocation, I may hit another person.	76	70	58									
6. There are people who pushed me so far that we came to blows.	72	73	65									
8. I have threatened people I know.	80	82	68									
11. I often find myself disagreeing with people.				80	75	70						
13. I can't help getting into arguments when people disagree with me.				82	71	68						
14. My friends say that I'm somewhat argumentative.				58	61	76						
15. I flare up quickly but get over it quickly.							50	62	69			
20. Sometimes I fly off the handle for no good reason							81	83	57			
21. I have trouble countrolling my temper.							71	71	34			
23. At times I feel I have gotten a raw deal out of life.										65	76	45
24. Other people always seem to get the breaks.										77	75	64
25. I wonder why sometimes I feel so bitter about things.										68	68	52

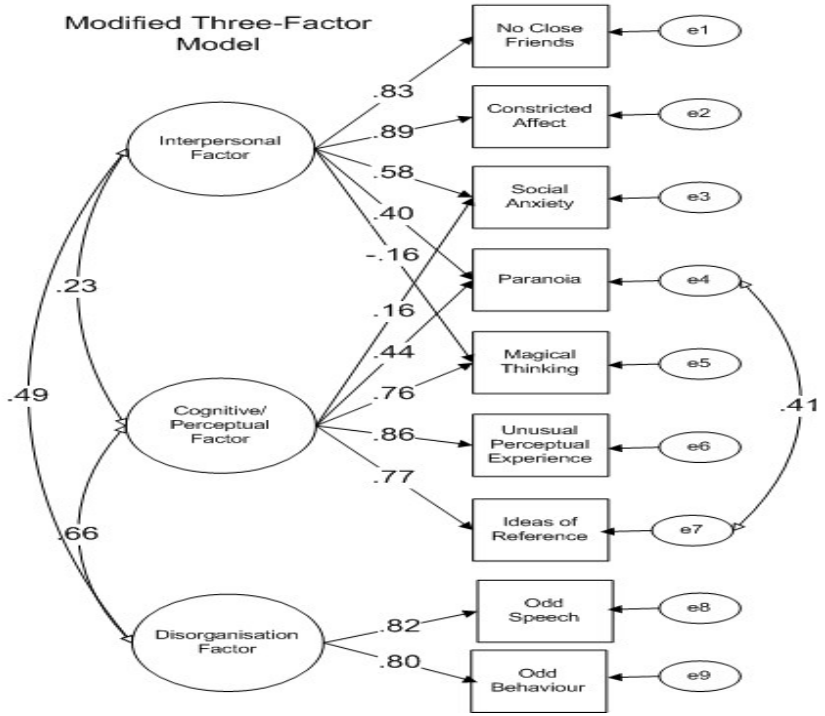
Bryant & Smith (2001)

TABLE 1
Goodness-of-Fit Statistics for Various Measurement Models of the AQ Imposed on Samples 1–3

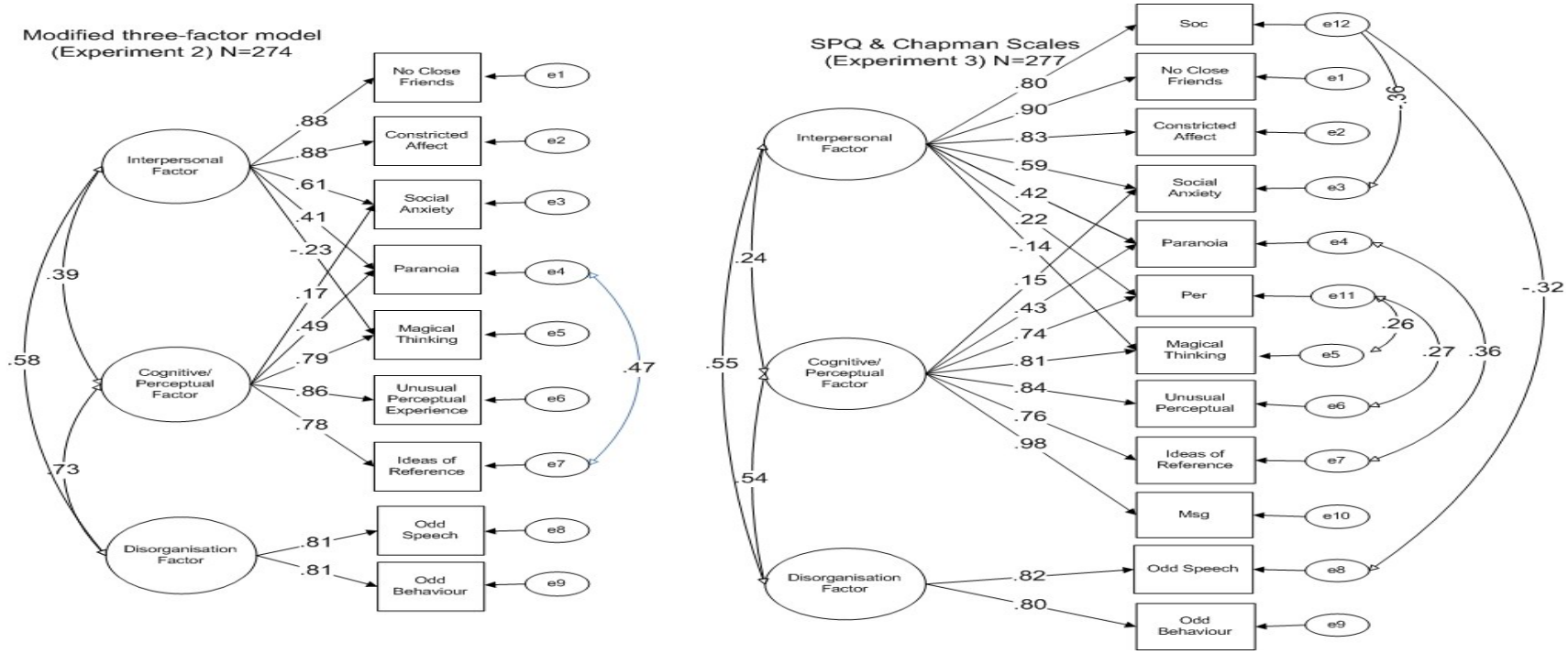
Model	No. items	Sample	Absolute fit measures					Relative fit measures	
			χ^2	*df*	χ^2/df	GFI	RMSEA	CFI	NNFI
One-factor	29	1	1567.9	377	4.2	.70	.102	.66	.64
(total score)		2	1267.8	377	3.4	.65	.109	.62	.59
		3	1469.1	377	3.9	.68	.098	.66	.63
Buss & Perry's	29	1	1042.8	371	2.8	.81	.077	.81	.79
four factors:		2	886.4	371	2.4	.76	.084	.78	.76
PA, VA, ANG, HO		3	950.3	371	2.6	.81	.072	.82	.80
Buss & Perry's	29	1	1046.4	373	2.8	.81	.077	.81	.79
hierarchical model:		2	888.5	373	2.4	.76	.083	.78	.76
one second-order factor		3	969.6	373	2.6	.81	.072	.81	.80
Buss & Perry's	27	1	881.9	318	2.9	.82	.076	.83	.82
PA, VA, ANG, &		2	734.2	318	2.3	.78	.081	.81	.79
Harris's HO factor		3	806.6	318	2.5	.83	.071	.83	.81
Four refined factors:	12	1	105.7	48	2.2	.94	.063	.96	.94
PA, VA, ANG, HO		2	92.4	48	1.9	.93	.068	.95	.93
		3	121.7	48	2.5	.94	.071	.91	.87
Refined hierarchical	12	1	108.5	50	2.2	.94	.062	.96	.94
model:		2	94.4	50	1.9	.93	.067	.95	.93
one second-order factor		3	133.6	50	2.7	.93	.074	.90	.86

Schizotypal Personality (Wuthrich & Bates, 2006)





SPQ Wuthrich & Bates (2006)



CFA
Practical

Using OpenMx & R

(CFA) Practical Session

Timothy C. Bates

Lets do some CFA…

- Lewis, Gary J., & Bates, Timothy C. (2011). A common heritable factor influences prosocial obligations across multiple domains. *Biology Letters*, 10.1098/rsbl.2010.1187. doi: 10.1098/rsbl.2010.1187

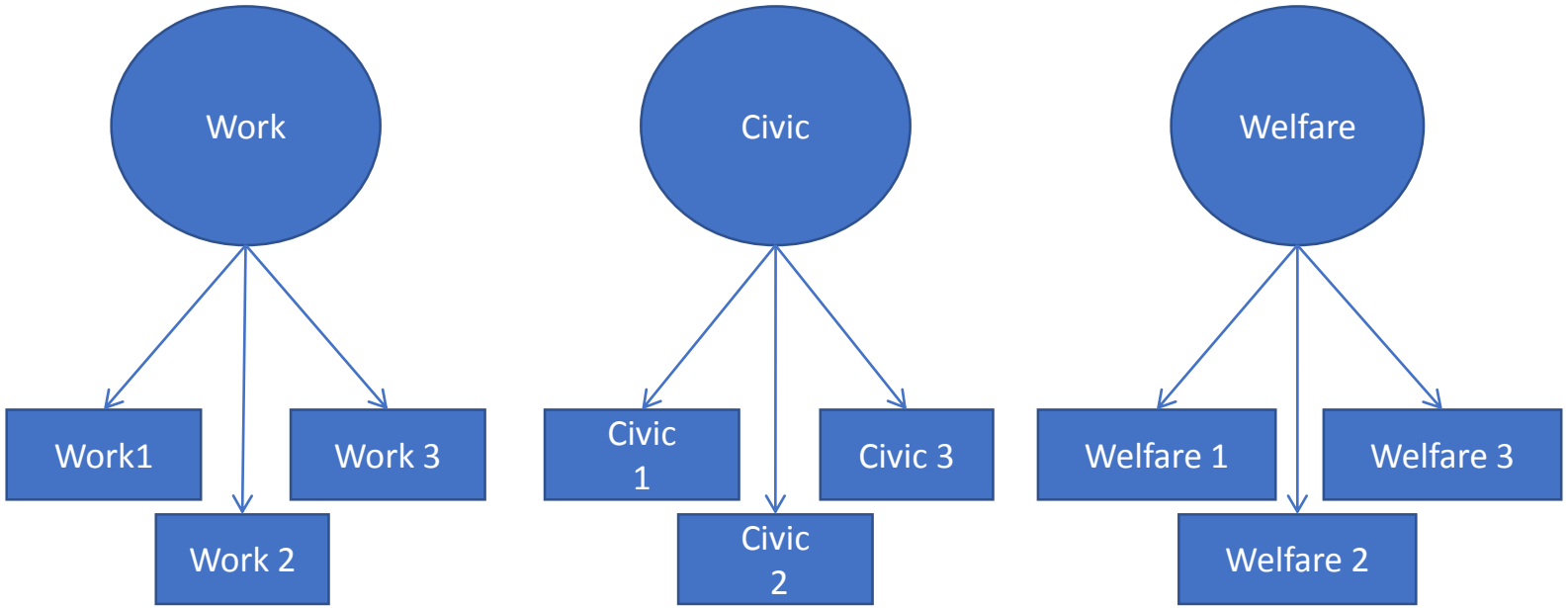
Pro-social Obligations

- Prosociality covers several domains, e.g.
  - Workplace: voluntary overtime
  - Civic life: Giving evidence in court
  - Welfare of others: paying for other's healthcare

- CFA can test the fit of theories
  - Which hypothesised model fits best?
    - 3 indicators of one factor?
    - 3 correlated factors?
    - 3 independent factors?

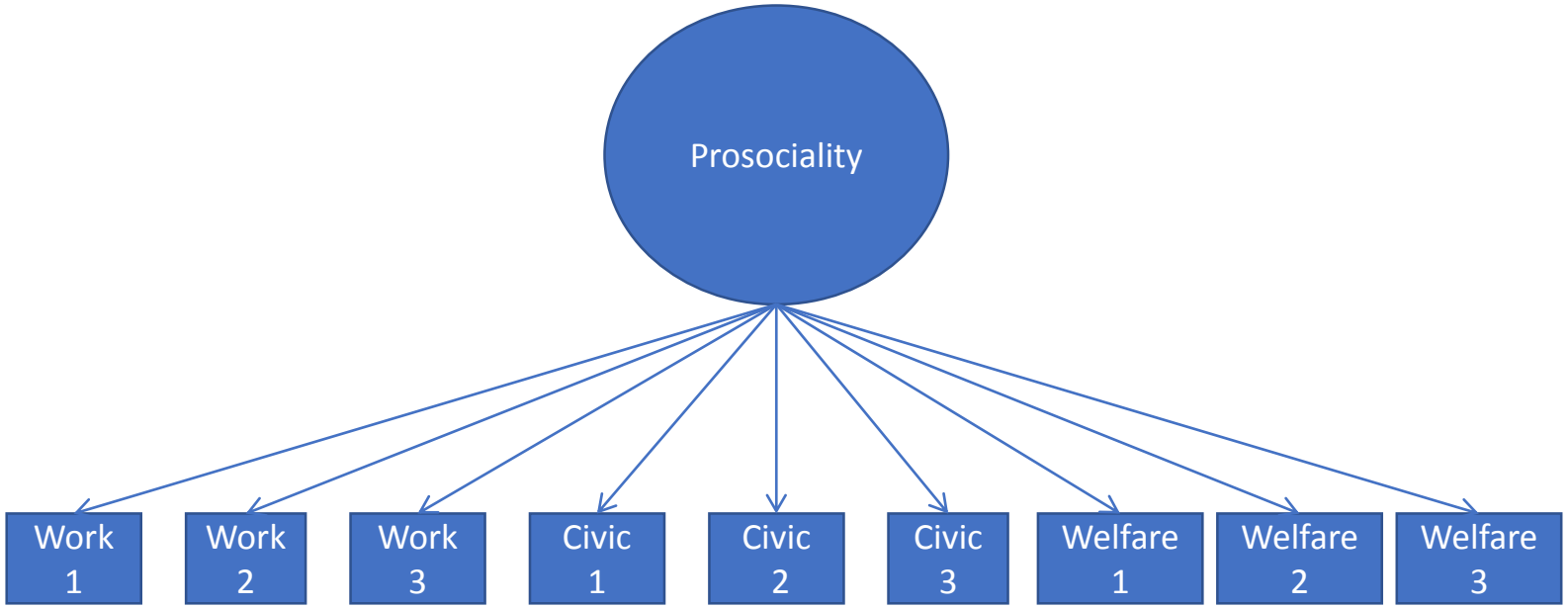
Factor structure of Prosociality

- Are there Distinct domains of pro-social behavior?



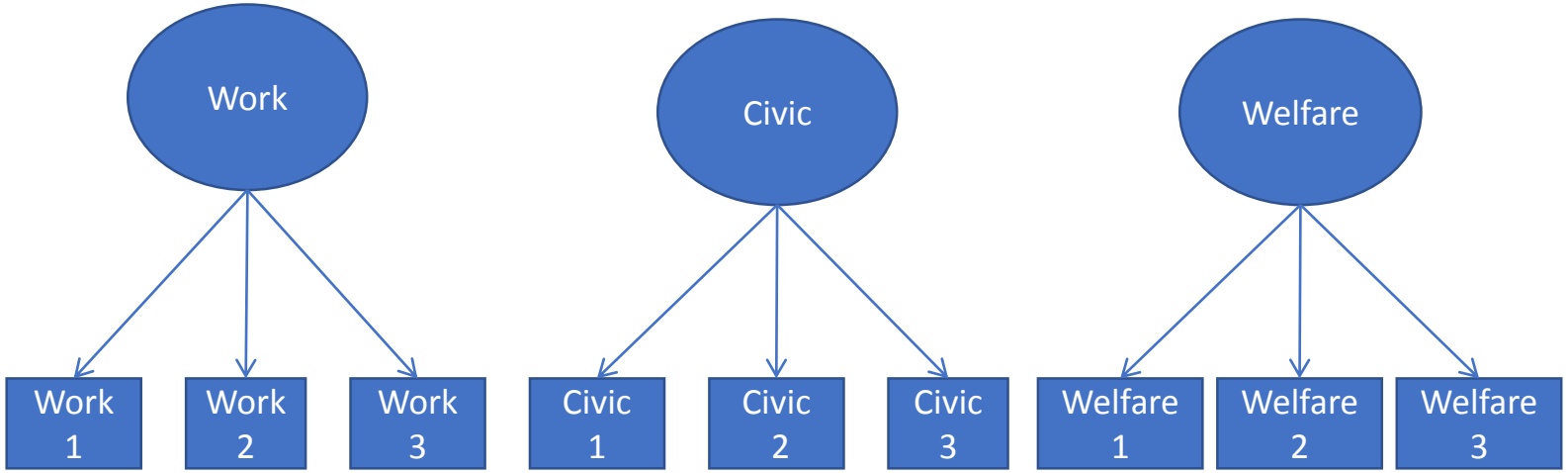
Factor structure of Prosociality

- One underlying prosociality factor?



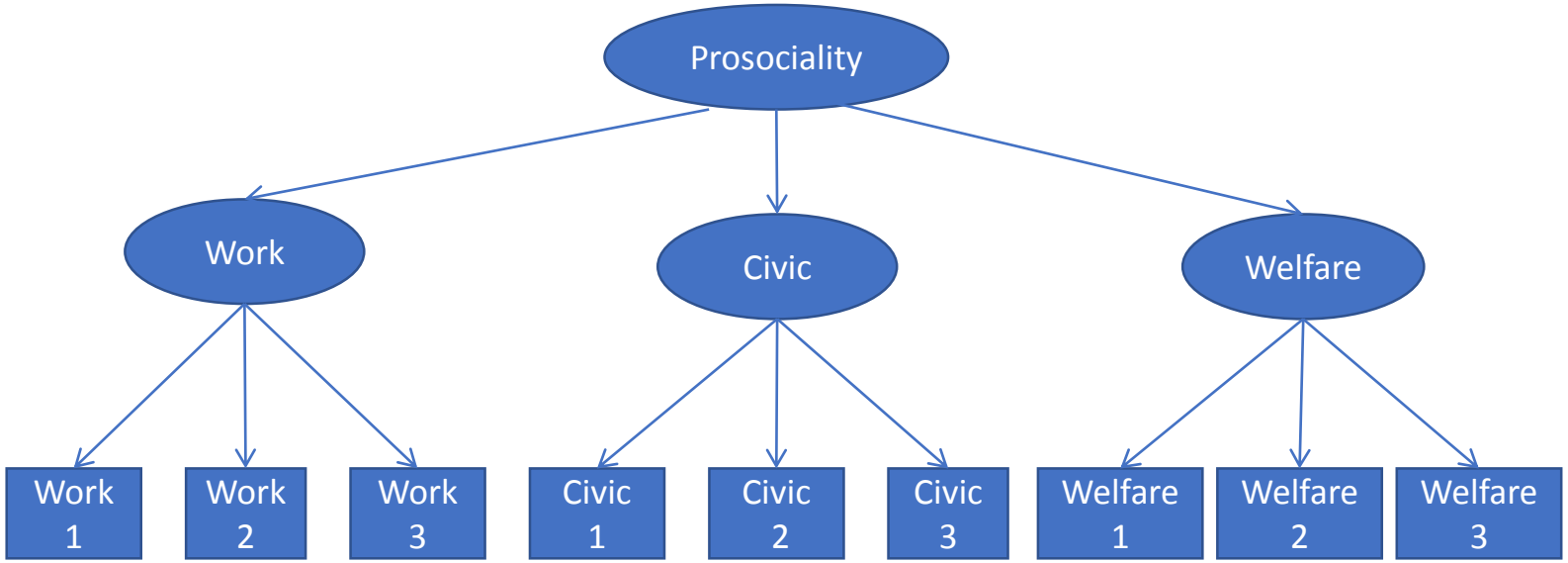
Factor structure of Prosociality

- Independent factors?

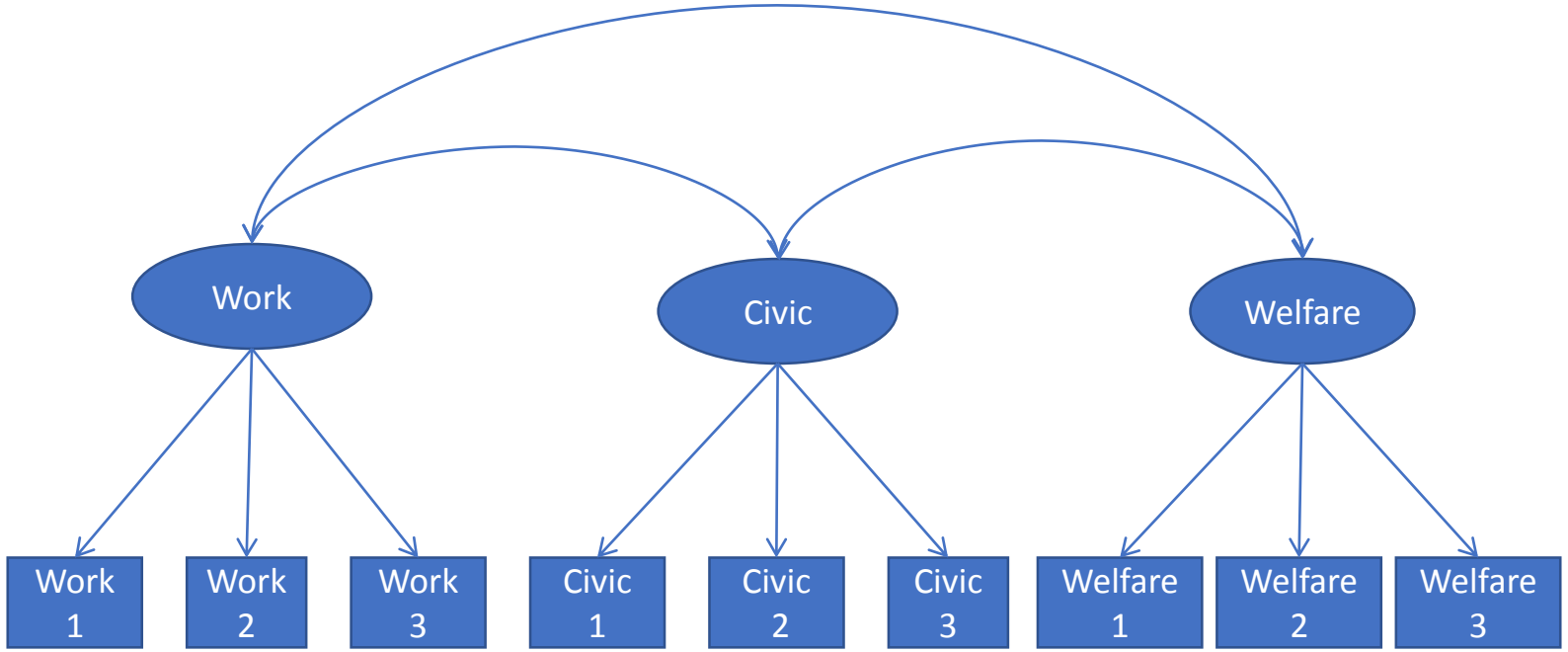


Factor structure of Prosociality

- Hierarchical structure: Common and distinct factors?



Note: These last two models are "*equivalent*" *



Advanced marks: does "*equivalent*" here mean "*the same*"?

CFA with R & OpenMx

- http://timbates.wikidot.com/mv-stats
- library(OpenMx)
- What is OpenMx?
  - "**OpenMx** is free and open source software for use with **R** that allows estimation of a wide variety of advanced multivariate statistical models."
    - Homepage: http://openmx.psyc.virginia.edu/

- *# library(sem)*

OpenMx Concepts

- Two approaches to modeling:
  - Path based
  - Matrix based
- We will use Path based here today

Path Based models in OpenMx

- Built using **mxModel**()
- Models *contain*
  - *lists of manifestVars and latentVars*
  - *paths*
  - *data*
  - *other neat stuff, like algebras, and constraints*

```
oneFactor <- mxModel("One Factor", type="RAM",
    # We have to specify the manifest and latent variables in the model
    manifestVars = vars, # List of 9 manifest variables
    latentVars   = "F1", # And one latent
    mxPath(from = "F1", to = vars), # factor loadings
    # latent variance
    mxPath(from = "F1", arrows = 2, free = F, values = 1, labels = "varF1"),
    # Residual variances
    mxPath(from = manifests, arrows = 2),
    mxData(cov(myData), type="cov", numObs = nrow(myData))
)
```

Core OpenMx functions

- **mxModel()**
  - this will contain all the objects in our model
  - **mxPath()**: each path in our model
  - **mxData()**: the data for the model
- **mxRun()**
  - Estimates parameters by sending the model to an optimizer
  - Returns a fitted model:
    - m1 = mxRun(m1)
    - summary(m1)

Our data

- Dataset: Midlife Study of Well-Being in the US (Midus)
- Prosocial obligations - c. 1000 individuals
  - *How much obligation do you feel…*
    - *to testify in court about an accident you witnessed?*’ [civic]
    - *to do more than most people would do on your kind of job?*’ [work]
    - *to pay more for your healthcare so that everyone had access to healthcare? [welfare]*

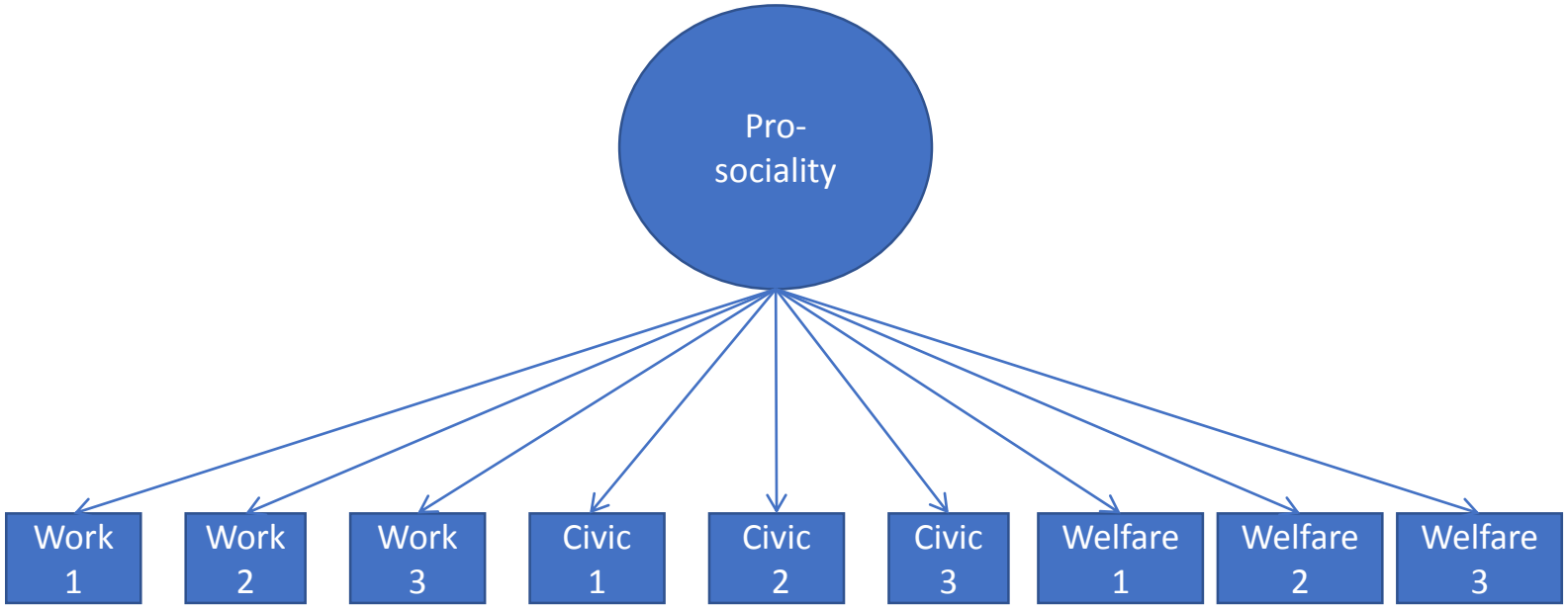
Preparatory code

- First we need to load OpenMx
  - require(OpenMx)
- Now lets get the data (R can read data off the web)
  - ? What are the names in the data?
- summary(myData) # We have some NAs
  - myData= na.omit(myData) # imputation is another solution
  - summary(myData)

CFA on our competing models

- Three competing models to test today:
  1. One general factor
  2. Three uncorrelated factors
  3. Three correlated factors
- In general, set out a preferred model, and then perhaps a most-complex and least complex alternative to judge it against.
- Then place constraints on the model
  - Can you set paths to zero?
  - Do you need additional paths?

Model 1: One general factor



Hands on

- A bit more code than you have used before:
  - We’ll go through it step-by-step

Model 1 – Get prepared

```
# Get set up
manifests = names(myData)
observedCov = cov(myData)
numSubjects = nrow(myData)
# Create a model using the mxModel function
Model1 <- mxModel(”My_first_CFA”, type="RAM”,
```

Model 1: What variable are in the model

Here we set the measured and latent variables

manifestVars = manifests,

latentVars = latents,

Step 3: Adding the paths

```
# Create residual variance for manifest variables using mxPath
mxPath( from=manifests, arrows=2),
# using mxPath, fix the latent factor variance to 1
mxPath( from = "F1", arrows = 2, free = F,  values = 1, labels = "varF1”),
# Using mxPath, specify the factor loadings
mxPath(from = "F1”,  to = manifests, arrows = 1, free = T, values = 1),
```

Set the data for the model

```
    # Give mxData the covariance matrix of ‘data2’ for analysis
    mxData(observedCov , type="cov", numObs=numSubjects )
    # make sure your last statement DOESN’T have a comma after it !!!
) # Close model
```

Run the model

```
# mxRun() fits the model
m1<- mxRun(m1)
# Lets see the summary output
Summary(m1)
# Exploring: What is in a model?
slotNames(m1@output)
names(m1@output)
```

Summary Output

 # Ideally, chi-square should be non-significant

chi-square: 564.41; p: < .001

 # Lower is better for AIC and BIC

AIC (Mx): 510.41

BIC (Mx): 191.47

 # < .06 is good fit: This model is a bad fit

RMSEA: 0.15

 # TLI > .95 is good fit: This model is a bad fit

TLI: 0.67

Summary: Model has poor fit

Ideally, chi-square should be non-significant (overly sensitive test)

chi-square: 576.50; p: <.0001

Lower is better for AIC (Akaike Information Criterion) and BIC

AIC (Mx): 522.50 ; BIC (Mx): 197.51

<.06 is a good for RMSEA and .96 for Tucker Lewis Index

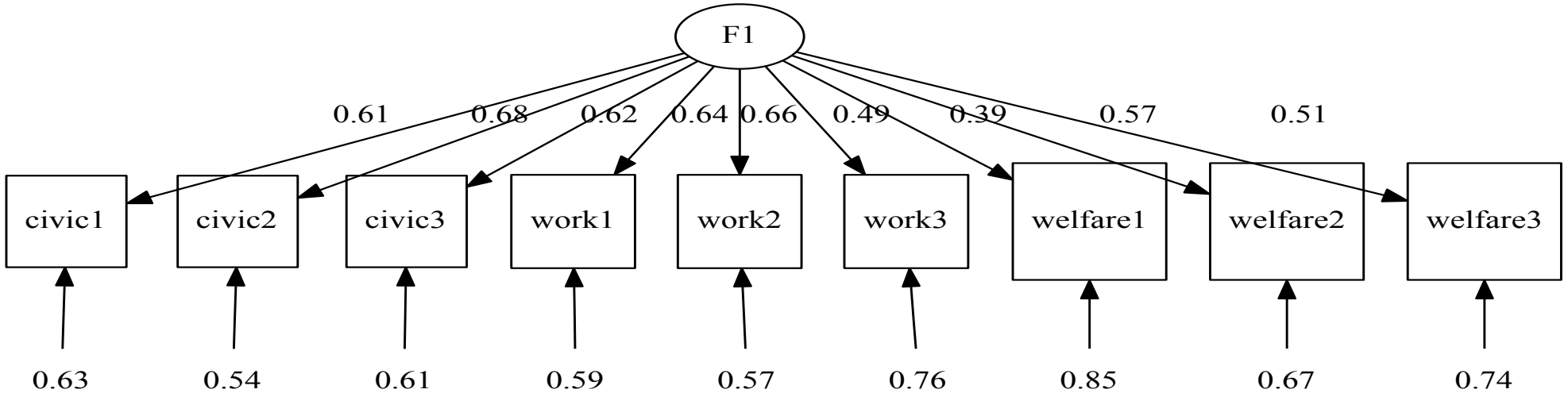
RMSEA: 0.16 ; **TLI**: .677

Story so far…

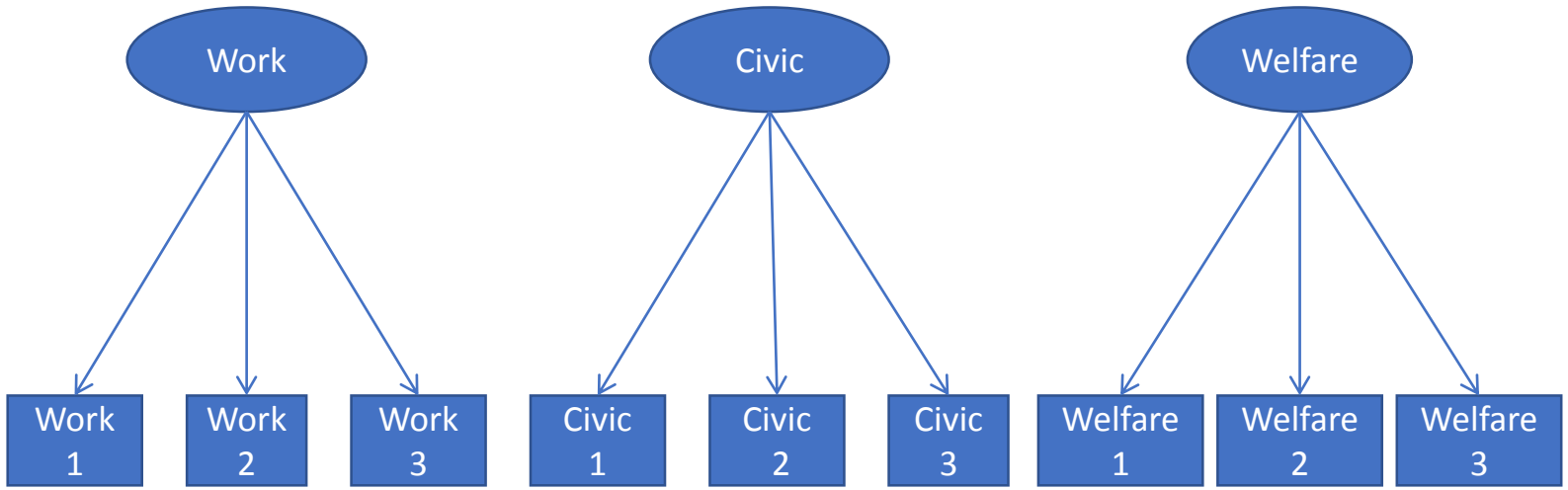
- Model 1 (one common factor) is a poor fit to the data)
  - Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Coventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.
  - Yu, C.Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. url

- Now you all do it for real…

All done and back together?



Model 2


Work
Civic
Welfare
Work 1
Work 2
Work 3
Civic 1
Civic 2
Civic 3
Welfare 1
Welfare 2
Welfare 3

Model 2: Alter # of latents and add their variance to the model

```
# We'll change the latent variables
   latentVars = c("F1","F2", "F3")
# Specify latent factor variances
 mxPath(from = c("F1","F2", "F3"), arrows = 2, free = F, values = 1)
```

What is a missing path?

- In SEM, silence speaks as loud as words
- If we don’t add a path between two places, what does that say?

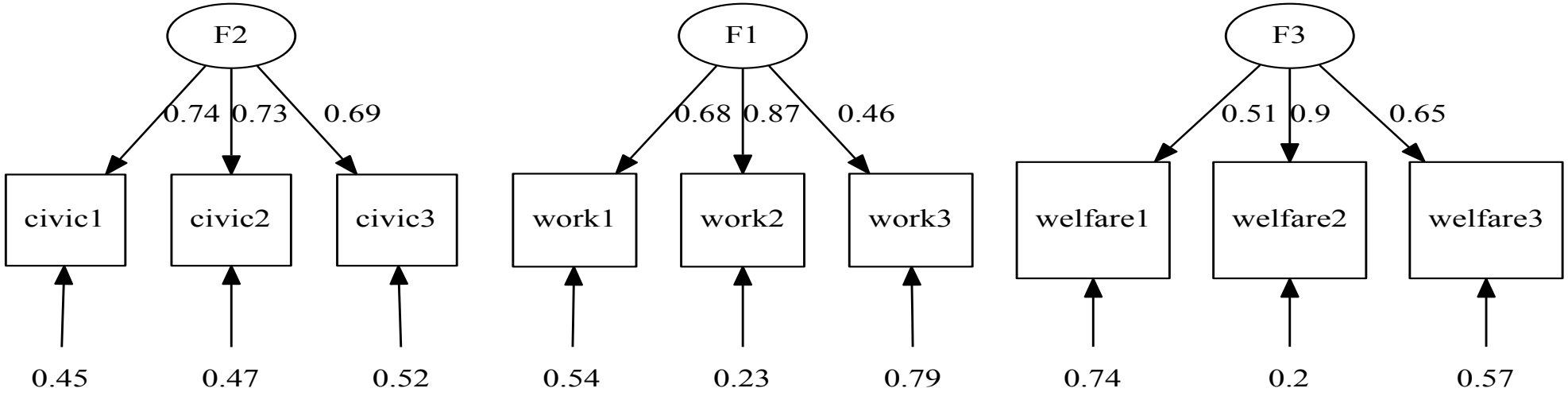
Model 2: factor loadings

```
# Specify the factor loadings
# i.e. F1 loads on work1,  2, and 3
mxPath(from="F1”,  to = c("work1","work2","work3”)),
# and F2 on civic…
mxPath(from = "F2", to = c("civic1","civic2", "civic3”))
mxPath(from = "F3", to = c("welfare1", "welfare2", "welfare3”))
```

Model 2 – Run the model

- Run the model
  - m1 <- **mxRun**(m2)
- *# Lets see the output*
  - **umxReportFit**(m1)
- Can also compare fits!
  - **mxCompare**(m1, c(m2))

Independent factors



Summary: Model has poor fit

Ideally, chi-square should be non-significant (overly sensitive test)

chi-square: 576.50; p: <.0001

Lower is better for AIC (Akaike Information Criterion) and BIC

AIC (Mx): 522.50 ; BIC (Mx): 197.51

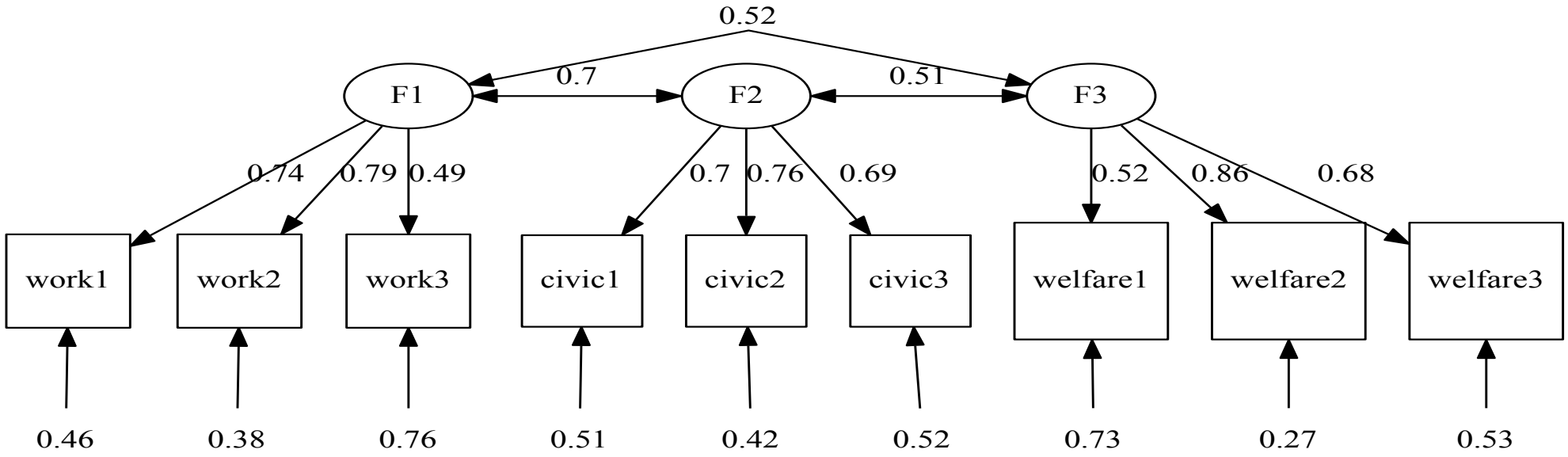
<.06 is a good for RMSEA and .96 for Tucker Lewis Index

RMSEA: 0.16 ; **TLI**: .677

Great resource for understanding fit

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

Allow the factors to correlate and test this model



Summary

- Factor analysis is EFA
- EFA is part of SEM
- EFA allows us to
  - Determine if a model has adequate fit
  - Compare models and objectively determine the best model
- Introduces us to ideas of maximum likelihood, parameter estimation, fit indices, latent variables, manifest variables, plots for these based on one-headed paths, two-headed paths, boxes, circles…
- In the tutorial we will begin building some simple models in *umx*
- *Next week:* Much more detail on SEM