

### **3 Confirmatory Factor Analysis**

#### **References:**

- Beaujean (2014). Chapter 1, Appendix A.
- Maruyama (1997). Chapter 10.
- Rosseel (2012).

### 3.1. Introduction

- CFA versus EFA

<u>EFA</u>	<u>CFA</u>
theory development	theory testing
no. of factors not fixed	fixed no. of factors
orthogonal factors	usually correlated
rotation	not necessary
variables load on all factors	load on specific factors
uncorrelated errors	correlated errors if needed
correlations as input	covariances as input
simple extraction methods	sophisticated estimation

- EFA as a preliminary step before CFA
- Use CFA for
  - testing single model (strictly confirmatory)
  - comparing alternative models
- Tremblay & Gardner's (1996) study

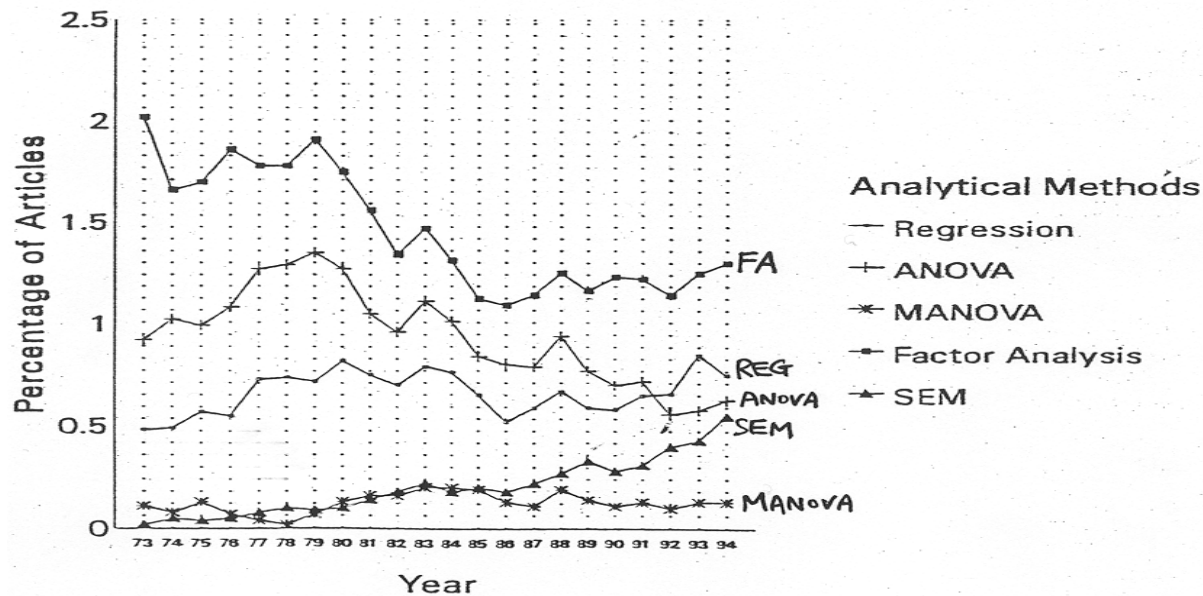
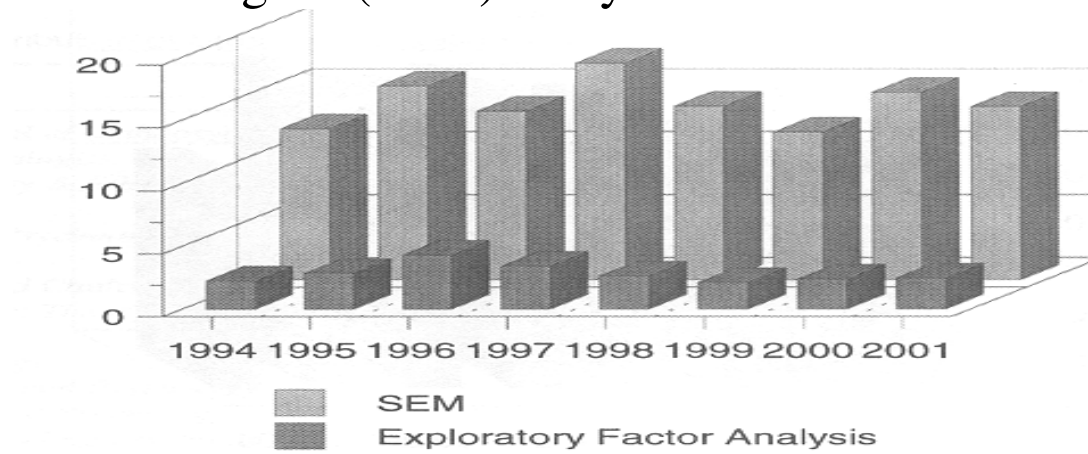


FIGURE 3 Percentage of articles using different analytical methods (based on total number of articles) by year.

- Hershberger's (2003) study



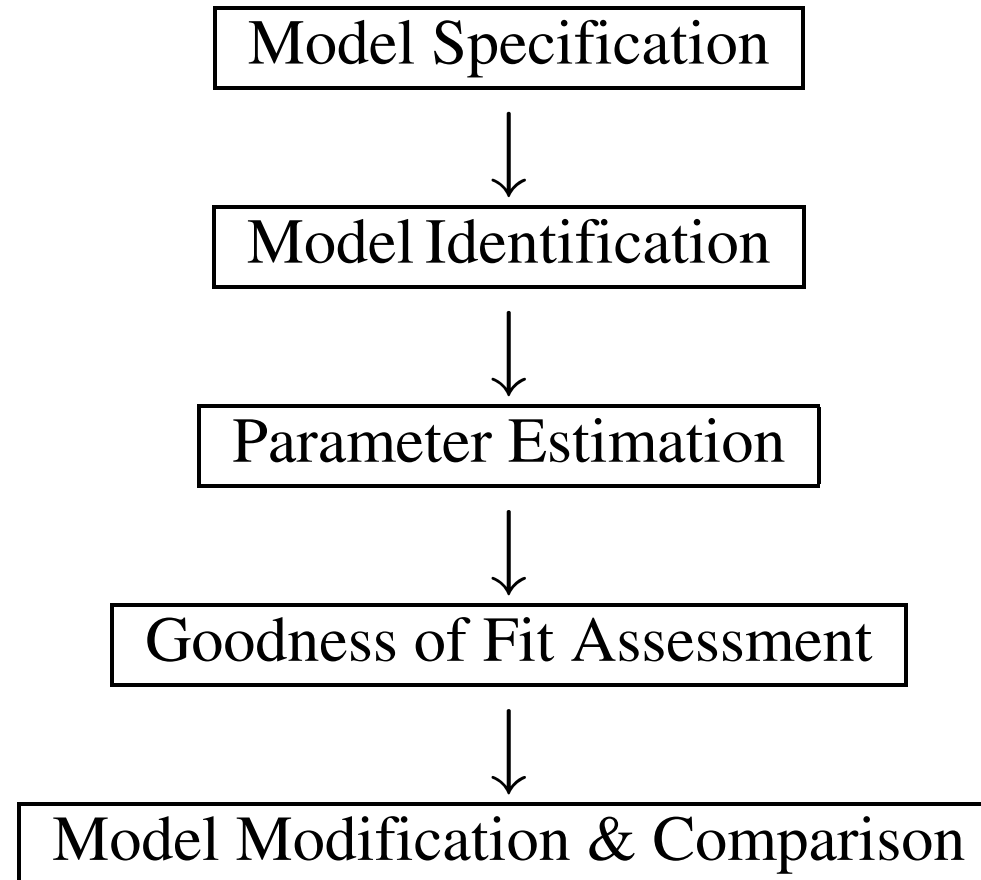
### 3.2. An Example: Subjective Well Being (SWB) Model

- To examine the hypothesis that subjective well being is a multidimensional construct composed of general life satisfaction (GLS) and work-related satisfaction (WS)
- Data: 5 variables were measured in a sample of size 500

covariance matrix (filename: *swb.cov*)

	V1	V2	V3	V4	V5
V1 (gls1)	198				
V2 (gls2)	82	86			
V3 (gls3)	54	28	24		
V4 (work1)	52	30	18	151	
V5 (work2)	16	10	7	44	28

### 3.3. Five Steps in CFA

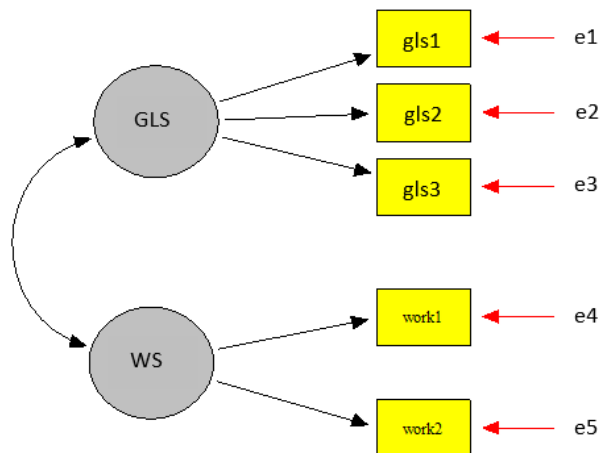


### 3.4. Model Specification



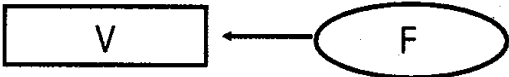
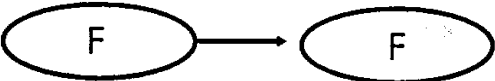
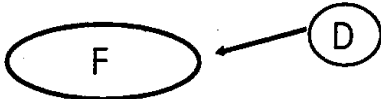

- First, we have  $p$  variables ( $V_1, V_2, \dots, V_p$ ) that are supposed to load on  $k$  factors ( $F_1, \dots, F_k$ )
- A CFA model is then specified by a model diagram, a set of model equations, or a matrix equation.

#### 3.4.1. Path diagrams

- For the SWB example:



- Some rules:

Symbol	Representation
	• Unobserved (latent) factor
	• Observed variable
	• Path coefficient for regression of observed variable onto unobserved factor
	• Path coefficient for regression of one factor onto another
	• Residual error (disturbance) in prediction of unobserved factor
	• Measurement error associated with observed variable

### The Greek Alphabet

$\alpha$	A	alpha
$\beta$	B	beta
$\gamma$	$\Gamma$	gamma
$\delta$	$\Delta$	delta
$\epsilon$	E	epsilon
$\zeta$	Z	zeta
$\eta$	H	eta
$\theta$	$\Theta$	theta
$\iota$	I	iota
$\kappa$	K	kappa
$\lambda$	$\Lambda$	lambda
$\mu$	M	mu
$\nu$	N	nu
$\xi$	$\Xi$	xi, ksi
$\omicron$	O	omicron
$\pi$	$\Pi$	pi
$\rho$	P	rho
$\sigma$	$\Sigma$	sigma
$\tau$	T	tau
$\upsilon$	$\Upsilon$	upsilon
$\phi$	$\Phi$	phi
$\chi$	X	chi
$\psi$	$\Psi$	psi
$\omega$	$\Omega$	omega

(source: Byrne, 1994)



### 3.4.2. The measurement model equations

- The model diagram can be converted into equations ( $F_1$ =GLS,  $F_2$ =WS)

$$\text{gls1} = V_1 = \mu_1 + \lambda_{11}F_1 + e_1$$

$$\text{gls2} = V_2 = \mu_2 + \lambda_{21}F_1 + e_2$$

$$\text{gls3} = V_3 = \mu_3 + \lambda_{31}F_1 + e_3$$

$$\text{work1} = V_4 = \mu_4 + \lambda_{42}F_2 + e_4$$

$$\text{work2} = V_5 = \mu_5 + \lambda_{52}F_2 + e_5$$

- Using matrix,

$$v = \mu + \Lambda f + e$$

$v$  is  $p \times 1$  vector of observed variables

$\mu$  is  $p \times 1$  vector of intercepts (means of  $v$ )

$\Lambda$  is  $p \times k$  factor loading matrix

$f$  is  $k \times 1$  vector of latent factors

$e$  is  $p \times 1$  vector of measurement errors

- Assumptions:

1. Latent variables are measured as deviations from their means

$$E(f) = 0, \quad E(e) = 0, \quad E(v) = \mu$$

2. Common factors ( $f$ ) and measurement errors ( $e$ ) are uncorrelated

$$E(fe') = 0$$

- Under these assumptions, the covariance matrix of  $v$  is

$$\Sigma = E[(v-\mu)(v-\mu)'] = \Lambda\Psi\Lambda' + \Theta = \Sigma(\theta)$$

- Hence,  $H_0 : \Sigma = \Sigma(\theta)$  is the hypothesized model structure in CFA. It means that the covariance matrix of the observed variables is structured as a function of a smaller set of parameters,  $\theta$ .

- In CFA, the variables are:

Name	Type	Cause/Effect	dimension
$v$	observed	DV	$p \times 1$
$f$	latent	IV	$k \times 1$
$e$	latent	IV	$p \times 1$

- And the parameter matrices are:

Parameter matrix	Symbol	Name	dimension
factor loading	$\Lambda$	lambda	$p \times k$
variance-covariance matrix of the factors	$\Psi$	psi	$k \times k$
variance-covariance matrix of errors	$\Theta$	theta	$p \times p$

- In CFA, our task is to
  - (1) estimate and test  $\theta$  (the unknown parameters in  $\Lambda$ ,  $\Psi$ , and  $\Theta$ ),
  - (2) evaluate the overall goodness of fit of the proposed model

### 3.5. Model Identification

- The CFA model is not identified because the scale of the latent factors is arbitrary:

Suppose  $\Sigma = \Lambda \Psi \Lambda' + \Theta = \Sigma(\theta)$

Let  $\Lambda^* = \Lambda D$  ( $D$  is an arbitrary  $k \times k$  square matrix such that  $DD^{-1} = I$ )  
 $\Psi^* = D^{-1} \Psi D^{-1}$   
 $\Theta^* = \Theta$

Then  $\Lambda^* \Psi^* \Lambda^{*'} + \Theta^* = \Lambda \Psi \Lambda' + \Theta = \Sigma$   
 $\Sigma(\theta^*) = \Sigma(\theta) = \Sigma$

- The parameters cannot be *uniquely* determined even  $\Sigma$  is known. This is the *identification problem or factor indeterminacy problem* in CFA
- A model is *identified* if there are no vectors  $\theta^*$  and  $\theta$  such that  $\Sigma(\theta^*) = \Sigma(\theta)$  unless  $\theta^* = \theta$  (unique solutions)

- That means every model parameter has to be uniquely solved in terms of the population variances and covariances of the observed variables
- This can be achieved by setting the metric of each latent variable with:
  1. unit loading identification (ULI): fix a path from a latent variable to another variable at a given value (usually 1.0), or
  2. unit variance identification (UVI): fix the factor variance at a given value (usually 1.0) (for independent latent variable only)

### 3.5.1. The *t*-rule

- Let  $p^* = \frac{1}{2}p(p+1)$  be the no. of available pieces of information (variances and covariances)
- Let  $q$  be the no. of free/unknown parameters in  $\Lambda$ ,  $\Psi$ , and  $\Theta$  (dimension of  $\theta$ )
- Define degrees of freedom of a model as,  $df = p^* - q$
- A ***necessary*** condition for the identification of the CFA model is that  $p^* \geq q$  ( $df \geq 0$ )

- ***Overidentification***: the pieces of information is more than the no. of unknown parameters ( $p^* > q$ ;  $df > 0$ )
- ***Underidentification***: the pieces of information is less than the no. of unknown parameters ( $p^* < q$ ;  $df < 0$ )
- ***Empirical underidentification***: the model  $df > 0$ , but the model is not identified because one of the model parameters is 0, or very close to 0
- ***Just identification***: the pieces of information is equal to the no. of unknown parameters ( $p^* = q$ ;  $df = 0$ )



### **3.5.2. The two-indicator rule**

- If the following conditions are satisfied, then the CFA model is identifiable:
  - at least two factors
  - factor correlations are free
  - two or more indicators per factor
  - each indicator loads on one factor
  - errors are uncorrelated
- Sufficient condition for model identification

### 3.6. Parameter Estimation

Theory World:  $\Sigma = \Lambda\Psi\Lambda' + \Theta = \Sigma(\theta)$

Data World:  $S \simeq \hat{\Lambda}\hat{\Psi}\hat{\Lambda}' + \hat{\Theta} = \hat{\Sigma} = \Sigma(\hat{\theta})$

- $\hat{\Sigma} = \Sigma(\hat{\theta})$  is called the *implied, reproduced, or fitted* covariance matrix
- Want to find  $\hat{\Lambda}$ ,  $\hat{\Psi}$ , and  $\hat{\Theta}$  (estimates of  $\Lambda$ ,  $\Psi$ ,  $\Theta$ ) such that  $\hat{\Sigma}$  and  $S$  are *as close as possible*
- Different estimation methods use different ways to define the closeness between  $S$  and  $\Sigma(\theta)$
- Normal-theory methods such as *maximum likelihood* estimation (ML) or *generalized least squares* estimation (GLS) assume that the data follow a multivariate normal distribution

- ML estimator  $\hat{\theta}_{\text{ML}}$  is obtained by minimizing

$$F_{\text{ML}}(S, \Sigma(\theta)) = \text{tr}(S\Sigma(\theta)^{-1}) + \ln |\Sigma(\theta)| - \ln |S| - p$$

over admissible choices of  $\theta$

- GLS estimator  $\hat{\theta}_{\text{GLS}}$  is obtained by minimizing

$$F_{\text{GLS}}(S, \Sigma(\theta)) = \text{tr}[(S - \Sigma(\theta))S^{-1}]^2$$

over admissible choices of  $\theta$

- When sample size is large (*asymptotically*), the ML or GLS estimators are *consistent, efficient and jointly normal*

- The *asymptotically distribution free* (ADF) estimation (Browne, 1984)

- does not require multivariate normality
- raw data are required
- a huge sample for reliable results (e.g., Hu, Bentler, & Kano, 1992)

- ADF estimator  $\hat{\theta}_{\text{ADF}}$  is obtained by minimizing

$$F_{\text{ADF}}(s, \sigma(\theta)) = (s - \sigma(\theta))' W^{-1} (s - \sigma(\theta))$$

over admissible choices of  $\theta$ , where  $W$  is called the *weight* matrix

- Both the ML, GLS, or ADF estimates are found by numerically searching methods with repeated iterations. To begin the search, *starting values* are needed.

### 3.7. Goodness of Fit Assessment

#### 3.7.1. Chi-square goodness of fit test

$$T_{\text{ML}} = (N - 1)F_{\text{ML}}(S, \Sigma(\hat{\theta}_{\text{ML}}))$$

$$T_{\text{GLS}} = (N - 1)F_{\text{GLS}}(S, \Sigma(\hat{\theta}_{\text{GLS}}))$$

$$T_{\text{ADF}} = (N - 1)F_{\text{ADF}}(s, \sigma(\hat{\theta}_{\text{ADF}}))$$

- Under  $H_0: \Sigma = \Sigma(\theta)$ , both  $T_{\text{ML}}$ ,  $T_{\text{GLS}}$ , and  $T_{\text{ADF}}$  are asymptotically distributed as a  $\chi^2$  variate with  $df = p^* - q$
- Reject  $H_0$  at  $\alpha$  level of significance if  $T > \chi_{\alpha}^2(df = p^* - q)$

- In practice, it is very difficult to get a nonsignificant  $T$  because
  - (1) underlying assumptions may not hold (Hu et al., 1992)
  - (2) with large  $N$ , trivial discrepancy between  $S$  and  $\hat{\Sigma}$  will lead to rejection of  $H_o$

### 3.7.2. Goodness of fit indices

#### 3.7.2.1. *Absolute fit indices*

- Measure how well a model reproduces the sample covariance matrix

##### 1. Goodness of fit index (*GFI*)

$$GFI = 1 - \frac{(s - \hat{\sigma})' W^{-1} (s - \hat{\sigma})}{s' W^{-1} s} \quad (\text{Tanaka \& Huba, 1984})$$

$$GFI = 1 - \frac{\text{tr}[(\hat{\Sigma}^{-1} S - I)^2]}{\text{tr}[(\hat{\Sigma}^{-1} S)^2]} \quad (\text{for ML methods, Joreskog \& Sorbom, 1984})$$

- *GFI* measures the relative amount of the variance and covariance in  $S$  that are predicted by  $\hat{\Sigma}$  (similar to  $R^2$  in regression)

## 2. Adjusted $GFI$ ( $AGFI$ )

$$AGFI = 1 - \frac{p(p+1)}{2df}(1 - GFI)$$

- $AGFI$  adjusts for the  $df$  of a model, it rewards simpler models with fewer parameters
- When  $S = \hat{\Sigma}$ ,  $GFI = AGFI = 1.00$
- Usually between zero and one, though it is possible for them to be negative



### 3.7.2.2. *Incremental fit indices*

- Measure the relative improvement in fit by comparing a target model with a baseline (independence) model

#### 1. Normed fit index (Bentler & Bonett, 1980)

$$NFI = \frac{T_B - T_T}{T_B}$$

- Measure the proportionate reduction in the chi-square values when moving from baseline to hypothesized model
- $NFI = 1.0$  when  $T_T = 0$
- Usually  $NFI > 0$  because  $T_B > T_T > 0$  (normed)
- No control for degrees of freedom

## 2. Relative fit index (Bollen, 1986)

$$RFI = \frac{T_B/df_B - T_T/df_T}{T_B/df_B}$$

- Modified *NFI* by comparing the fit per degrees of freedom
- Usually in 0-1 range, though it can be negative
- Usually,  $RFI < NFI$  because  $df_B > df_T$

## 3. Non-normed fit index (Bentler & Bonett, 1980)

$$NNFI = \frac{T_B/df_B - T_T/df_T}{T_B/df_B - 1}$$

- Outside 0-1 range (non-normed)
- Also known as Tucker-Lewis index (TLI)
- $NNFI > RFI$

#### 4. Incremental fit index (Bollen, 1989)

$$IFI = \frac{T_B - T_T}{T_B - df_T}$$

- Outside 0-1 range
- $IFI > NFI$

#### 5. Comparative fit index (Bentler, 1990)

$$CFI = 1 - \frac{\max[(T_T - df_T), 0]}{\max[(T_T - df_T), (T_B - df_B), 0]}$$

- Based on the concept of noncentrality parameter ( $\lambda$ ), an index measuring the discrepancy between  $\Sigma$  and  $\Sigma(\theta)$
- For the target model,  $\lambda$  is estimated by  $T_T - df_T$
- $CFI$  lies within 0-1 range

### ***3.7.2.3. How to interpret these indices?***

- Depending on the choice of the baseline model, the standards set by prior work, and the selection of the fitting function
- Old rules:  $> 0.90$  is required

### ***3.7.2.4. Expected cross-validation index (Browne & Cudeck, 1989)***

$$ECVI = \frac{T_T + 2q}{N - 1}$$

- A measure of the discrepancy between the fitted covariance matrix in the analyzed sample and the expected covariance matrix that would be obtained in another sample of the same size
- Small values indicate a greater chance of cross-validity

### 3.7.2.5. *Standardized root mean squared residual*

$$SRMR = \sqrt{\sum_{i=1}^p \sum_{j=1}^i \left( \frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii} s_{jj}}} \right)^2 / \left( \frac{1}{2} p(p+1) \right)}$$

- A large *SRMR* indicates a poor model fit
- Sensitive to model misspecification
- Old rules:  $< .05$  suggests a good model fit (Byrne, 1998)

### 3.7.2.6. Root mean square error of approximation (Steiger & Lind, 1980)

$$RMSEA = \sqrt{\hat{F}_0 / df_T}$$

where  $\hat{F}_0 = \max[(T_T - df_T)/(N - 1), 0]$  is an estimated value of the population discrepancy function.

- *RMSEA* lies in 0-1 range
- Old rules (Browne & Cudeck, 1993):

$< 0.05$	close fit
$0.05 - 0.08$	reasonable fit
$> 0.1$	inadequate fit

### 3.7.2.7. Fit indices and their thresholds: A more contemporary view (Hooper, Coughlan, & Mullen, 2008)

Fit Index	Acceptable Threshold Levels	Description
<i>Absolute Fit Indices</i>		
Chi-Square $\chi^2$	Low $\chi^2$ relative to degrees of freedom with an insignificant $p$ value ( $p > 0.05$ )	
Relative $\chi^2$ ( $\chi^2/\text{df}$ )	2:1 (Tabachnik and Fidell, 2007) 3:1 (Kline, 2005)	Adjusts for sample size.
Root Mean Square Error of Approximation (RMSEA)	Values less than 0.07 (Steiger, 2007)	Has a known distribution. Favours parsimony. Values less than 0.03 represent excellent fit.
GFI	Values greater than 0.95	Scaled between 0 and 1, with higher values indicating better model fit. This statistic should be used with caution.
AGFI	Values greater than 0.95	Adjusts the GFI based on the number of parameters in the model. Values can fall outside the 0-1.0 range.
RMR	Good models have small RMR (Tabachnik and Fidell, 2007)	Residual based. The average squared differences between the residuals of the sample covariances and the residuals of the estimated covariances.
SRMR	SRMR less than 0.08 (Hu and Bentler, 1999)	Unstandardised. Standardised version of the RMR. Easier to interpret due to its standardised nature.
<i>Incremental Fit Indices</i>		
NFI	Values greater than 0.95	Assesses fit relative to a baseline model which assumes no covariances between the observed variables. Has a tendency to overestimate fit in small samples.
NNFI (TLI)	Values greater than 0.95	Non-normed, values can fall outside the 0-1 range. Favours parsimony. Performs well in simulation studies (Sharma et al, 2005; McDonald and Marsh, 1990)
CFI	Values greater than 0.95	Normed, 0-1 range.

### 3.7.2.8. Summary of fit indices (Maruyama, 1998; p.240-241)

TABLE 10.1 Fit Indexes Summary

*Types (classes) of fit indexes*

1. *Absolute*: Is the residual (unexplained) variance appreciable? (e.g., chi-square, chi-square/df, RMR, GFI)
2. *Relative*: How well does the model do compared with (a range of) other possible models with the same data? (e.g., Type 1: NFI; Type 2: TLI, IFI; Type 3: BFI or RNI)
3. *Adjusted*: How does the model combine fit and parsimony? (e.g., LISREL's PGFI, PNFI, TLI)

*Specific indexes*

In the following,  $F$  stands for the function that is minimized, and  $\chi^2 = (N - 1)F$ . The subscript key is as follows:  $t$  = theoretical;  $n$  = null;  $s$  = saturated;  $a$  and  $b$  = alternative models. The symbol  $k$  = number of measures in the model.

**Root mean residual**

This statistic is simply the square root of the mean of the squared discrepancies between all the elements of the predicted ( $\Sigma$ ) and observed ( $S$ ) matrices.

**LISREL goodness of fit index:**

$$GFI = 1 - [\text{tr}(\Sigma^{-1}S - I)^2 / \text{tr}(\Sigma^{-1}S)^2]$$

The GFI measures the proportion of weighted information in  $S$  that fits weighted information in  $\Sigma$ , such as the coefficient of determination. The ratio part of the formula is like a ratio of residual to total variance.

**LISREL adjusted goodness of fit index:**

$$AGFI = 1 - [k(k + 1) / 2 df_a] \times (1 - GFI)$$

The AGFI is not recommended (see Mulaik et al., 1989).

**Bentler and Bonett's (1980) normed fit index:**

$$NFI = (F_a - F_b) / F_n, \text{ or } (\chi^2_a - \chi^2_b) / \chi^2_n$$

Implicitly, the denominator is  $F_n - F_s$ , but  $F_s = 0$ .

**Tucker-Lewis index (Tucker & Lewis, 1973):**

$$TLI = \frac{(\chi^2_n / df_n - \chi^2_t / df_t)}{(\chi^2_n / df_n - 1)} = \frac{((F_n / df_n) - (F_t / df_t))}{(F_n / df_n) - (1 / (N - 1))}$$

where  $N$  is the sample size. This also is the Bentler and Bonett non-normed fit index formula.

TABLE 10.1 Continued

**Bollen's (1989) incremental fit index:**

$$IFI = (\chi^2_n - \chi^2_t) / (\chi^2_n - df_t)$$

Note that  $df_t$  = expected value of  $\chi^2$  with  $t$  degrees of freedom.

**Bentler's (1990) and McDonald and Marsh's (1990) relative noncentrality index:**

$$RNI \text{ or BFI} = [(\chi^2_n - df_n) - (\chi^2_t - df_t)] / (\chi^2_n - df_n)$$

**James, Mulaik, and Brett's (1982) parsimonious fit index:**

$$PGFI = \{df_t / [k(k + 1) / 2]\} GFI,$$

where  $df_t$  is degrees of freedom of the model,  $k$  = size of input matrix,  $k(k + 1) / 2$  = total possible degrees of freedom, and GFI is the index defined above.

**Mulaik et al.'s (1989) parsimonious normed fit index:**

$$PNFI = (df_t / df_n) NFI \text{ or } \{df_t / [k(k - 1) / 2]\} NFI$$

**Mulaik et al.'s parsimonious normed fit index, Type 2:**

$$PNFI2 = (df_t / df_n) IFI \text{ or } \{df_t / [k(k - 1) / 2]\} IFI$$

**Akaike information criteria (Akaike, 1987):**

$$AIC \text{ (Joreskog)} = \chi^2_t - 2df_t$$

$$AIC \text{ (Tanaka)} = \chi^2_t + 2(\text{number of free parameters})$$

**Bozdogan's (1987) modified AIC:**

$$CAIC = \chi^2_t - (1 + \ln N)df_t$$

**Browne and Cudeck's (1989) expected cross-validation index:**

$$ECVI =$$

$$[\chi^2_t / (N - 1)] + 2[\text{number of free parameters} / (N - 1)]$$

**Steiger's (1990) root mean square error of approximation:**

$$RMSEA = \text{SQRT}(F_t / df_t)$$

NOTE: RMR = root mean residual; GFI = goodness of fit index; NFI = normed fit index; TLI = Tucker-Lewis index; IFI = incremental fit index; BFI or RNI = relative noncentrality index; CFI = comparative fit index; PGFI = parsimonious GFI; PNFI = parsimonious NFI; AGFI = adjusted GFI; AIC = Akaike information criteria; CAIC = modified AIC; ECVI = expected cross-validation index; RMSEA = root mean square error of approximation.



### 3.8. Model Modification

- Adding or deleting parameters in a post-hoc fashion
- Theoretical justification is important
- Cross validation (Cudeck & Browne, 1983)

#### 3.8.1. Adding new parameters

- Test whether restrictions can be removed from the current model in order to improve the fit
- *Lagrange Multiplier* test, also known as *modification indices* in other softwares such as *lavaan* and *LISREL*

### 3.8.2. Deleting existing parameters

- Test whether restrictions can be imposed on the current model without affecting the fit (individual test for each parameter)
- *Wald* test

### 3.9. Model Comparison

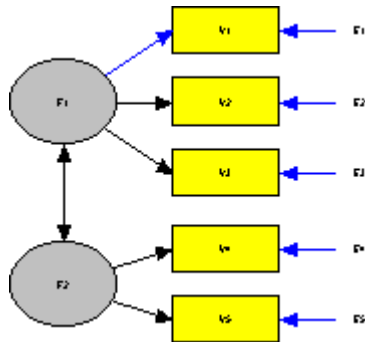
- *Nested models* : any model which requires that some function of its free parameters equals another free parameter or equals a constant is nested in the identical model that has no such restriction

- Define

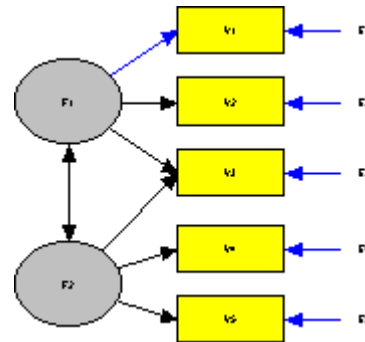
$M_0$ : A restricted model with chi-square value  $T_0$  and  $df_0$

$M_1$ : A less restricted model, in which  $M_0$  is nested within  $M_1$ , with chi-square value  $T_1 (< T_0)$  and  $df_1 (< df_0)$

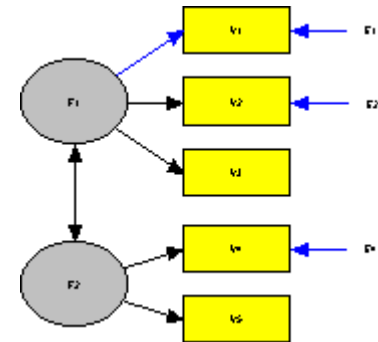
Model 1



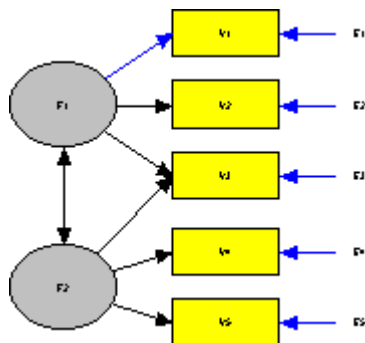
Model 2



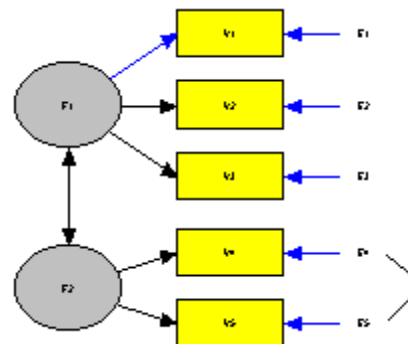
Model 3



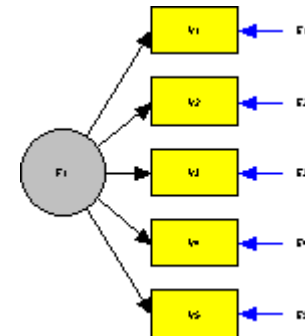
Model 4



Model 5



Model 6



### 3.9.1. Likelihood-ratio (*LR*) or chi-square difference test

- *LR* test evaluates the decrease in the goodness-of-fit when a less restricted model  $M_1$  is further restricted
- The greater the reduction, the less likely that the restricted model  $M_0$  is true
- $H_0$ : no difference between  $M_0$  and  $M_1$  in terms of goodness of fit

$$\Delta\chi^2 = T_0 - T_1 \sim \chi^2(df_0 - df_1)$$

- Reject  $H_0$  if  $\Delta\chi^2$  is significant at  $\alpha = 0.05$
- A computational disadvantage of the *LR* test is that 2 models must be estimated for every pair of model comparisons
- Only useful for comparing nested models

### 3.9.2. Goodness of fit indices

- Use those indices that account for degrees of freedom because models with more parameters can fit the data better in general
- *AGFI, NNFI, RFI*

### 3.9.3. Akaike information criterion (*AIC*) and modified *AIC* (*CAIC*)

- *AIC* (Akaike, 1987) and *CAIC* (Bozdogan, 1987) can be used for comparing non-nested models

$$AIC = T_u + 2q$$

$$CAIC = T_u + (1 + \ln N)q$$

- Choose model with the smallest *AIC* or *CAIC*

### 3.10. How to Present CFA Results? (see, Boomsma, 2000; Tabachnick & Fidell, 2007)

#### Results

##### *The Hypothesized Model*

A confirmatory factor analysis, based on data from learning-disabled children, was performed through LISREL on the eleven subtests of the WISC-R. The hypothesized model is presented in Figure 14.8 where circles represent latent variables, and rectangles represent measured variables. Absence of a line connecting variables implies no hypothesized direct effect. A two factor model of IQ, Verbal and Performance, is hypothesized. The information, comprehension, arithmetic, similarities, vocabulary, and digit span subtests serve as indicators of the Verbal IQ factor. The picture comprehension, picture arrangement, block design, object assembly, and coding subtests serve as indicators of the Performance IQ factor. The two factors are hypothesized to covary with one another.

##### *Assumptions*

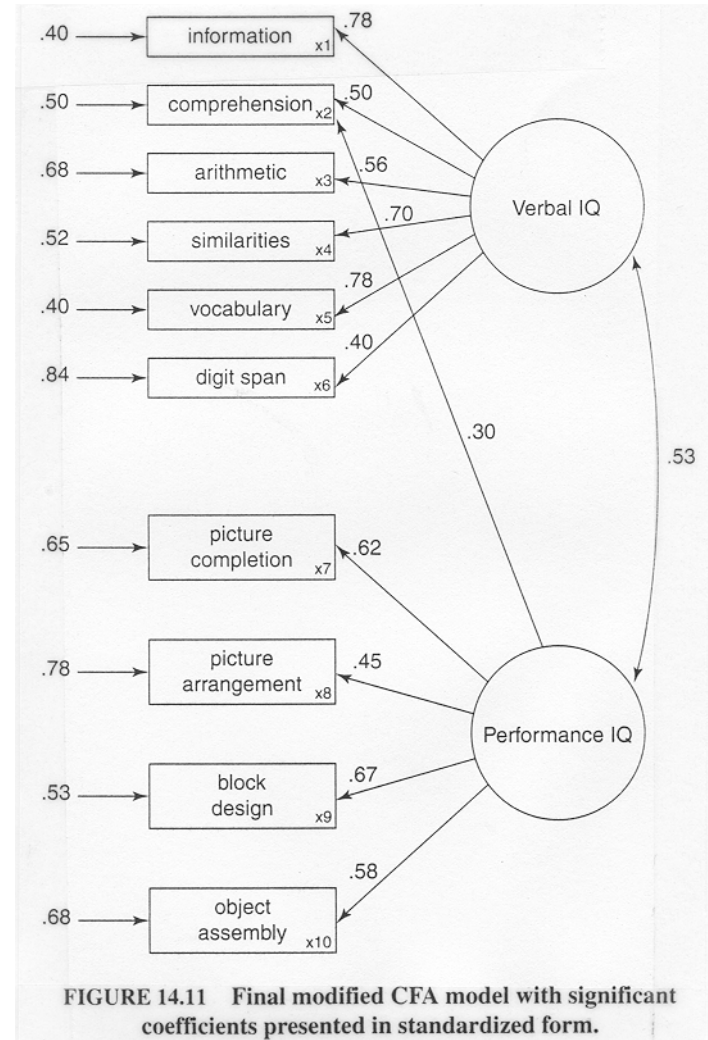
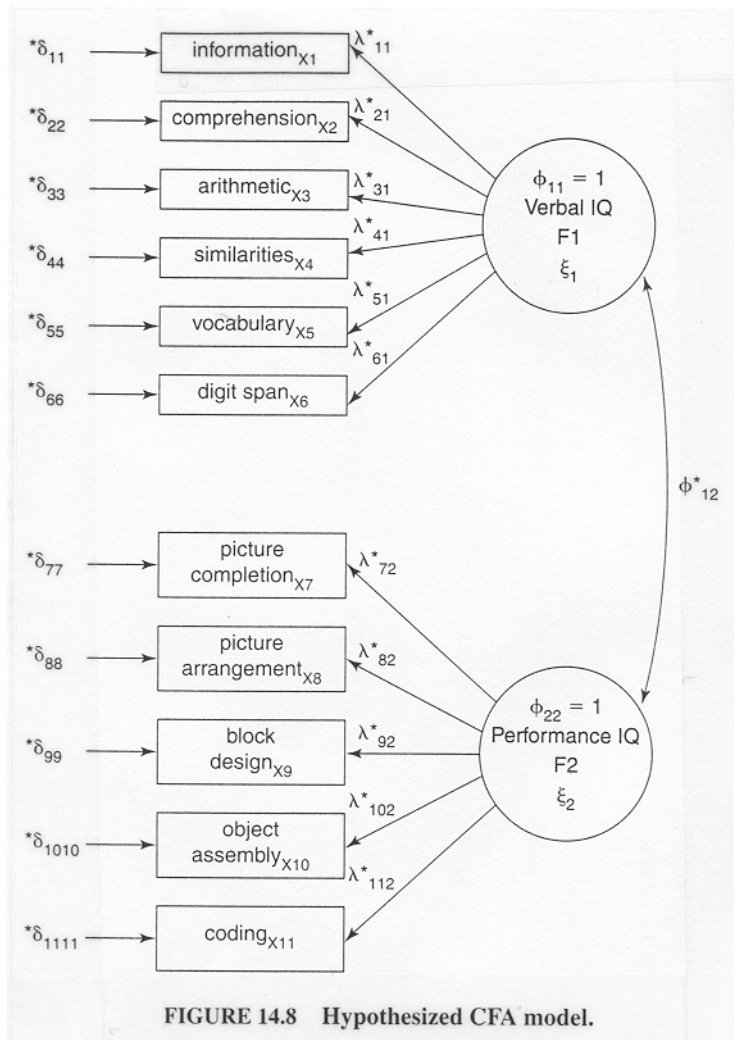
The assumptions of multivariate normality and linearity were evaluated through SPSS. One child had an extremely high score on the arithmetic subtest (19,  $z = 4.11$ ,  $p < .01$ ) and his data were deleted from the analysis. Using Mahalanobis distance, another child was a multivariate outlier,  $p < .001$ , and the data from this child were also deleted. This child had an extremely low comprehension subtest score and an extremely high arithmetic score. Structural equation modeling (SEM) analyses were performed using data from 175 children. There were no missing data.

#### Model Estimation

Maximum likelihood estimation was employed to estimate all models. The independence model that tests the hypothesis that all variables are uncorrelated was easily rejectable,  $\chi^2(55, N = 175) = 516.24$ ,  $p < .01$ . The hypothesized model was tested next and support was found for the hypothesized model,  $\chi^2(43, N = 175) = 70.24$ ,  $p = .005$ , comparative fit index (CFI) = .94. A chi-square difference test indicated a significant improvement in fit between the independence model and the hypothesized model.

Post hoc model modifications were performed in an attempt to develop a better fitting and possibly more parsimonious model. On the basis of the Lagrange multiplier test, a path predicting the comprehension subtest from the Performance factor was added,  $\chi^2(42, N = 172) = 60.29$ ,  $p = .03$ , CFI = .96, CAIC = 108.25, AIC = 108.295. A chi square difference test indicated that the model was significantly improved by addition of this path,  $\chi^2_{diff}(1, N = 172) = 9.941$ ,  $p < .01$ . Second, because the coefficient predicting the coding subscale from the Performance factor (.072) was not significant, SMC = .005, this variable was dropped and the model re-estimated,  $\chi^2(33, N = 172) = 45.018$ ,  $p = .08$ , CFI = .974, CAIC = 180.643, AIC = 89.018. Both the CAIC and AIC indicated a better fitting, more parsimonious model after the coding subtest is dropped.

Because post hoc model modifications were performed, a correlation was calculated between the hypothesized model parameter estimates and the parameter estimates from the final model,  $r(18) = .95$ ,  $p < .01$ ; this indicates that parameter estimates were hardly changed despite modification of the model. The final model, including significant coefficients in standardized form, is illustrated in Figure 14.11.





### 3.11. Using *lavaan* (Rosseel, 2012)

- Steps:

1. Install R in your computer, <http://cran.r-project.org/>

- 1.1. Try RStudio, a free and user-friendly environment for doing R programming, <https://www.rstudio.com/>

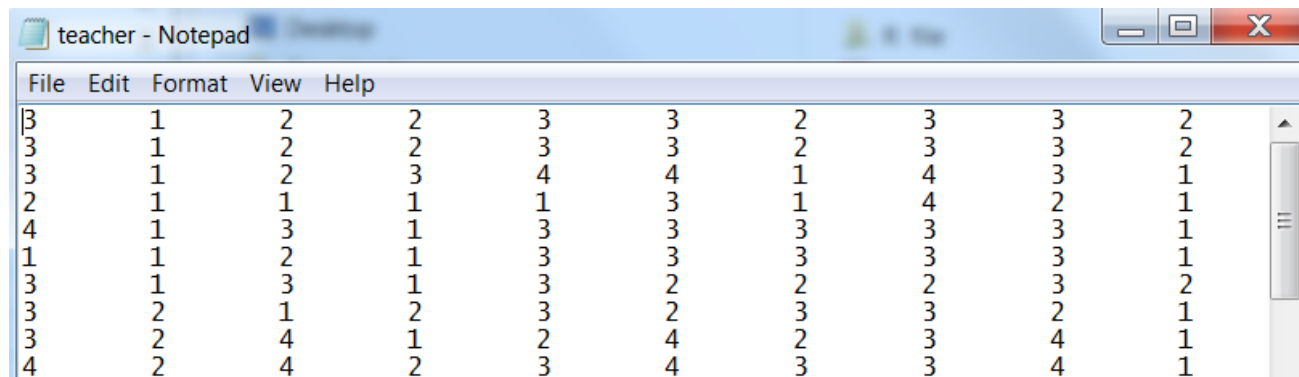
2. Download the lavaan package in R  
`install.packages("lavaan", dependencies=TRUE)`

3. Learn how to enter or import the data into R

4. Learn lavaan syntax (e.g., the lavaan tutorial (Rosseel, 2017))

### 3.11.1. Input raw data

(a) Tab delimited file: *teacher.dat* (10 variables,  $N=89$ )



3	1	2	2	3	3	2	3	3	2
3	1	2	2	3	3	2	3	3	2
3	1	2	3	4	4	1	4	3	1
2	1	1	1	1	3	1	4	2	1
4	1	3	1	3	3	3	3	3	1
1	1	2	1	3	3	3	3	3	1
3	1	3	1	3	2	2	2	3	2
3	2	1	2	3	2	3	3	2	1
3	2	4	1	2	4	2	3	4	1
4	2	4	2	3	4	3	3	4	1

```
#set work directory
setwd("C:/Users/wchan/Google Drive/stat6108/data")
```

```
# Method 1: input tab delimited file
data1 <- read.table("teacher.dat", header=FALSE)
```

(b) SPSS data file: *teacher.sav*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	3	1	2	2	3	3	2	3	3	2
2	3	1	2	2	3	3	2	3	3	2
3	3	1	2	3	4	4	1	4	3	1
4	2	1	1	1	1	3	1	4	2	1
5	4	1	3	1	3	3	3	3	3	1

#set work directory

```
setwd("C:/Users/wchan/Google Drive/stat6108/data")
```

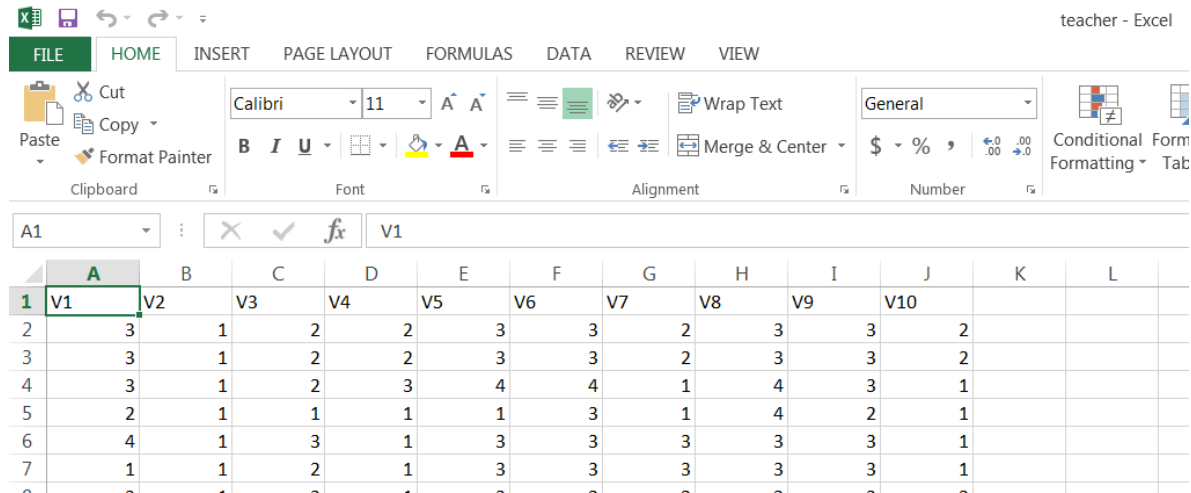
# Method 2: input SPSS data file

# load the foreign package

```
library(foreign)
```

```
data2 <- read.spss("teacher.sav", to.data.frame=TRUE)
```

(c) Excel file: *teacher.xlsx*



	A	B	C	D	E	F	G	H	I	J	K	L
1	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10		
2	3	1	2	2	3	3	2	3	3	2		
3	3	1	2	2	3	3	2	3	3	2		
4	3	1	2	3	4	4	1	4	3	1		
5	2	1	1	1	1	3	1	4	2	1		
6	4	1	3	1	3	3	3	3	3	1		
7	1	1	2	1	3	3	3	3	3	1		

```
#set work directory
```

```
setwd("C:/Users/wchan/Google Drive/stat6108/data")
```

```
# Method 3: input Excel data file
```

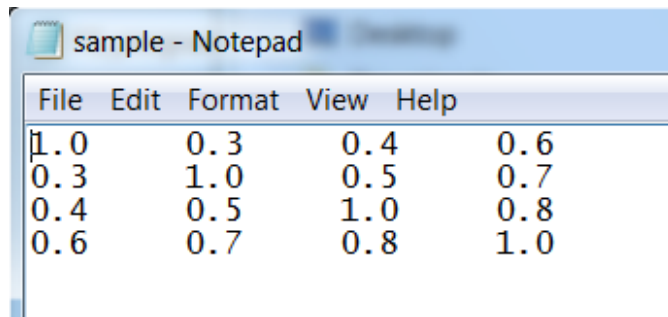
```
# load the xlsx package
```

```
library(xlsx)
```

```
data3 <- read.xlsx("teacher.xlsx", 1)
```

### 3.11.2. Import sample covariance

(a) full matrix (sample.cor)

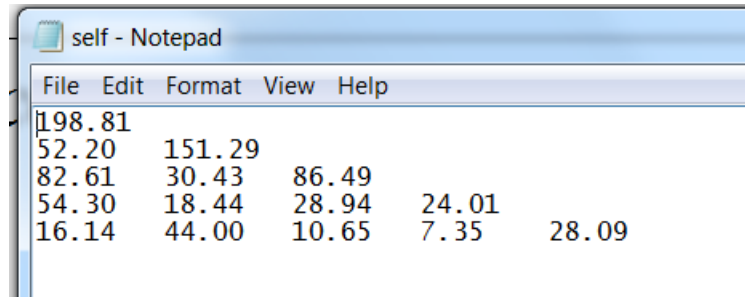


The screenshot shows a Notepad window with the title 'sample - Notepad'. The menu bar includes 'File', 'Edit', 'Format', 'View', and 'Help'. The text content is a 4x4 matrix of numerical values representing a covariance matrix.

1.0	0.3	0.4	0.6
0.3	1.0	0.5	0.7
0.4	0.5	1.0	0.8
0.6	0.7	0.8	1.0

```
> #set work directory
> setwd("C:/Users/wchan/Google Drive/stat6108/data")
>
> # Read full sample covariance matrix
> full <- read.table("sample.cor")
> rownames(full)=colnames(full)
> full
      v1 v2 v3 v4
v1 1.0 0.3 0.4 0.6
v2 0.3 1.0 0.5 0.7
v3 0.4 0.5 1.0 0.8
v4 0.6 0.7 0.8 1.0
```

(b) lower triangular matrix (*swb.cov*,  $N=500$ )



```
> # Read sample covariance matrix in lower triangular form
> # load the lavaan package
> library(lavaan)
This is lavaan 0.5-23.1097
lavaan is BETA software! Please report any bugs.
> mycov <- scan("swb.cov")
Read 15 items
> swb_cov <- getCov(mycov, lower=TRUE,
+ names=c("gls1", "gls2", "gls3", "work1", "work2"))
> swb_cov
```

	gls1	gls2	gls3	work1	work2
gls1	198	82	54	52	16
gls2	82	86	28	30	10
gls3	54	28	24	18	7
work1	52	30	18	151	44
work2	16	10	7	44	28

```
> |
```

### 3.11.3. *lavaan* syntax

- The 4 equation types and operators

Formula type	Operator	Mnemonic
Latent variable	=~	is manifested by
Regression	~	is regressed on
(Residual) (co)variance	~~	is correlated with
Intercept	~ 1	intercept
Defined parameter	:=	is defined as
Equality constraint	==	is equal to
Inequality constraint	<	is smaller than
Inequality constraint	>	is larger than

1. regression:  $y1 \sim x1+x2+x3$
2. measurement model:  $f1 =\sim v1+v2+v3$
3. variance and covariance:  $f1 \sim\sim f2$
4. intercept:  $y1 \sim 1$

- Fixing, labelling, and assigning starting values to model parameters

1. fixing parameters: pre-multiply a parameter with a numeric value

$$f1 \sim 1 * f1$$

2. labelling parameters: pre-multiply a parameter with a character string (label), and model parameters with the same label are considered to be equal

$$y \sim \text{label1} * x1 + b * x2 + b * x3$$

3. starting values: pre-multiply a parameter with `start(sv)`

$$f1 = \sim x1 + \text{start}(0.8) * x2 + \text{start}(1.2) * x3$$



- Defining additional parameters (:=)

$$Y \sim \text{direct} * X + b * M$$

$$M \sim a * X$$

$$\text{indirect} := a * b$$

$$\text{total} := \text{direct} + \text{indirect}$$

- Constraining model parameters (==)

$$\text{direct} == a * b$$