Confirmatory factor analysis for school data

The example discussed in this section is based on school data that were collected during a 1994 survey in South Africa.

A brief description of the **SA_Schools94.dat** data set is as follows: N = 136 schools were selected and the total number of children within schools $\sum_{i=1}^{N} n_i = 6047$, where n_i varies from 20 to 60. The data set contains 20 variables as shown in Table 2.

Table 2: Description of variables in SA_School94.dat

Variable	Name	Description	Number missing
1	Student	Level-1 identification	0
2	School	Level-2 identification	0
3	Constant	All values equal to 1	0
4	Grade	0 = Grade 2 1 = Grade 3 2 = Grade 4	0
5	Language	0 = White 1 = Black	0
6	Gender	1 = Male 2 = Female	1
7	Mothedu	Mother's level of education on a scale from 1 to 7	783
8	Fathedu	Father's level of education on a scale from 1 to 7	851
9 10 11 12	Read Speech Write Arithm	Teacher's evaluation on a scale from 1 to 5: 1 = Poor 5 = Excellent	482 470 467 451
13	Socio	Socio-economic status indicator, scale 0 to 5 on school level	0
14	Classif	Classification: total correct out of 30 items	23

15	Compar	Comparison: total correct out of 23 items	27
16	Verbal	Verbal Instructions: total correct out of 50 items	20
17	Figure	Figure Series: total correct out of 24 items	118
18	Pattcomp	Pattern Completion: total correct out of 24 items	109
19	Knowled	Knowledge: total correct out of 32 items	112
20	Numserie	Number Series: total correct out of 15 items	2305

The variables Language and Socio are school-level variables and their values do not vary within schools. Listwise deletion of missing cases results in a data set containing only 2691 of the original 6047 cases.

For this example, we use the variables Classif, Compar, Verbal, Figure, Pattcomp and Numserie from the schools data set discussed in the previous section. Two common factors are hypothesized: verbal and numerical ability. The first three variables are assumed to measure Verbfac and the last three to measure Numfac. A path diagram of the hypothesized factor model is shown in Figure 2.

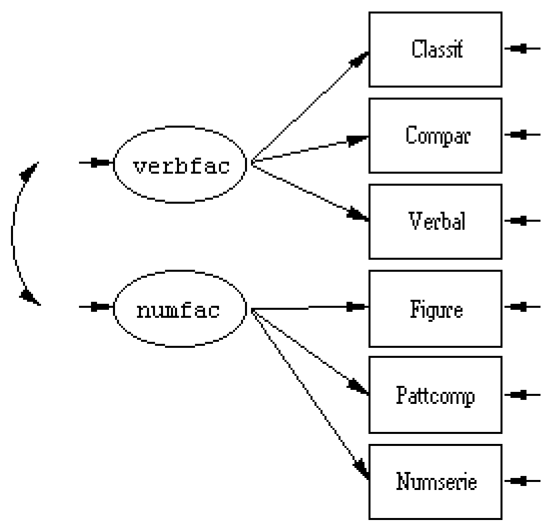

Figure 2: Confirmatory factor analysis model for 6 variables

The between- and within-school structural equations models are

$$\begin{aligned} \mathbf{\Sigma}_W &= \mathbf{\Lambda}_W \mathbf{\Psi}_W \mathbf{\Lambda}_W' + \mathbf{D}_W \\ \mathbf{\Sigma}_B &= \mathbf{\Lambda}_B \mathbf{\Psi}_B \mathbf{\Lambda}_B' + \mathbf{D}_B. \end{aligned} \qquad (20)$$

where

$$\mathbf{\Lambda}_W = \mathbf{\Lambda}_B = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix},$$

and where factor loadings are assumed to be equal on the between (schools) and within (children) levels. The 2 x 2 matrices $\mathbf{\Psi}_B$ and $\mathbf{\Psi}_W$ denote unconstrained factor covariance matrices. Diagonal elements of $\mathbf{D}_B$ and $\mathbf{D}_W$ are the unique (error) variances.

Gender and Grade differences were accounted for in the means part of the model,

$$E(y_{ijk}) = \beta_{k0} + \beta_{k0} Gender + \beta_{k2} Grade,$$

where the subscripts *i*, *j* and *k* denote schools, students and variables *k*, respectively.

From the description of the school data set, we note that the variable Numserie has 2505 missing values. An inspection of the data set reveals that the pattern of missingness can hardly be described as missing at random. To establish how well the proposed algorithm perform in terms of the handling of missing cases, we have decided to retain this variable in this example. The appropriate LISREL syntax file for this example is given below.

```
Group1: Between Schools HSRC School Project
DA NI=6 NO=0 NG=2 MA=CM MI=-9.0
LA
Classif Compar Verbal Figure Pattcomp Numserie
RA = SA_Schools94.dat
$CLuster= School
SE
1 2 3 4 5 6 /
MO NY=6 NE=2 LY=FU,FI PS=SY,FR TE=DI,FR
LE
verbfac  numfac
FR LY(2,1) LY(3,1) LY(5,2) LY(6,2)
VA 1.00 LY(1,1) LY(4,2)
PD
OU ME=ML
Group2: Within Schools HSRC School Project
LA
Classif Compar Verbal Figure Pattcomp Numserie
DA NI=6 NO=0 NG=2 MA=CM MI=-9.0
RA = SA_Schools94.dat
SE
1 2 3 4 5 6 /
MO NY=6 NE=2 LY=IN PS=IN TE=IN
```

```
LK
verbfac  numfac
FR  PS(1,1) PS(2,1) PS(2,2) TE(1,1) TE(2,2)
FR TE(3,3) TE(4,4) TE(5,5) TE(6,6)
! FR LY(2,1) LY(3,1) LY(5,2) LY(6,2)
OU
```

Table 3 shows the estimated between-schools covariance matrix $\hat{\boldsymbol{\Sigma}}_B$ when no restrictions are imposed on its elements, and the fitted covariance matrix $\boldsymbol{\Sigma}_B(\hat{\boldsymbol{\gamma}})$ where $\boldsymbol{\gamma}$ is the vector of parameters of the CFA models given in (20).

Table 3: Estimated between-schools covariance matrix, $\boldsymbol{\Sigma}_B$

(i) $\hat{\boldsymbol{\Sigma}}_B$ unrestricted

	Classif	Compar	Verbal	Figure	Pattcomp	Numserie
Classif	1.29					
Compar	1.27	2.66				
Verbal	2.83	3.54	10.42			
Figure	2.06	2.70	6.89	5.53		
Pattcomp	2.17	2.60	6.58	5.09	5.34	
Numserie	1.46	1.85	4.93	3.85	3.81	3.16

(ii) $\boldsymbol{\Sigma}_B(\hat{\boldsymbol{\gamma}})$ for the CFA model

	Classif	Compar	Verbal	Figure	Pattcomp	Numserie
Classif	1.61					
Compar	1.74	3.86				
Verbal	2.21	3.22	6.76			
Figure	2.52	3.67	4.66	5.82		
Pattcomp	2.29	3.33	4.22	5.03	4.93	
Numserie	1.79	2.60	3.30	3.93	3.56	3.11

Likewise, Table 4 shows $\hat{\boldsymbol{\Sigma}}_W$ for the unrestricted model and $\boldsymbol{\Sigma}_W(\hat{\boldsymbol{\gamma}})$ for the CFA model.

Table 4: Estimated within-schools covariance matrix, $\boldsymbol{\Sigma}_W$

(i) $\hat{\boldsymbol{\Sigma}}_W$ unrestricted

	Classif	Compar	Verbal	Figure	Pattcomp	Numserie
Classif	8.49					
Compar	4.59	18.77				
Verbal	5.52	7.64	17.26			
Figure	4.45	7.21	8.49	16.27		

Pattcomp	4.30	7.21	8.45	9.55	16.19	
Numserie	2.69	4.05	5.28	7.31	5.80	7.31

(ii) $\mathbf{\Sigma}_W(\hat{\gamma})$ for the CFA model

	Classif	Compar	Verbal	Figure	Pattcomp	Numserie
Classif	7.81					
Compar	3.31	17.12				
Verbal	4.19	6.11	15.07			
Figure	3.72	5.42	6.87	14.58		
Pattcomp	3.37	4.91	6.23	8.13	14.61	
Numserie	2.64	3.84	4.87	6.36	5.76	7.21

The goodness of fit statistics for the CFA model are shown in Table 5.

Table 5: Goodness-of-fit statistics

(6047 students, 136 schools)

χ^2 = 159.87	degrees of freedom = 20
RMSEA=0.061	

Parameter estimates and estimated standard errors are given in Table 6.

It is typical of SEM models to produce large χ^2-values when sample sizes are large, as in the present case. The RMSEA may be a more meaningful measure of goodness of fit and the value of 0.061 indicates that the assumption of equal factor loadings between and within schools is reasonable.

Table 6: Parameter estimates and standard errors

	Estimate	**Standard error**
Factor loadings		
λ_{11}	1.0	
λ_{21}	1.456	0.048
λ_{31}	1.846	0.054
λ_{42}	1.0	
λ_{52}	0.906	0.017
λ_{62}	0.708	0.014
Factor covariances (between schools)		
Ψ_{11}	1.196	0.185
Ψ_{21}	2.524	0.342

Ψ_{22}	5.546	0.729
Error variances (between schools)		
Classif	0.413	0.081
Compar	1.327	0.223
Verbal	2.673	0.388
Figure	0.279	0.090
Pattcomp	0.377	0.092
Numserie	0.325	0.069
Factor covariances (within schools)		
Ψ_{11}	2.271	0.114
Ψ_{21}	3.722	0.128
Ψ_{22}	8.976	0.272
Error variances (within schools)		
Classif	5.538	0.119
Compar	12.305	0.262
Verbal	7.328	0.222
Figure	5.606	0.169
Pattcomp	7.239	0.178
Numserie	2.710	0.098