Session 4

Measurement Models

Session 4: Measurement Models

Introduction

The factor analysis and principal components techniques considered in previous sessions are examples of *exploratory factor analysis (EFA)*. We explore our data to discover what latent variables/factors/principal components exist, and how they are related to the observed variables.

Either by using EFA or having other knowledge of the proposed relationships between the latent and observed variables (replicating a previous study, perhaps, or designing a battery of tests with certain latent variables in mind), we develop *hypothesised models*, or a set of *hypothesised relationships*.

Using multiple regression, we then confirm or reject these hypotheses – this is known as *confirmatory factor analysis (CFA)*, and is a specific type of *structural equation modelling (SEM)*.

Note: modifying the hypothesised models as a result of CFA can lead back to the area of EFA – however, EFA and CFA should just be regarded as useful labels, and not as mutually exclusive techniques that cannot be mixed.

Other phrases sometimes used to refer to SEM:

- causal modelling or causal analysis
- simultaneous equation modelling
- analysis of covariance structures

At the simplest level of SEM, the hypothesised relationship will be between one dependent measured variable and one or more independent measured variables – in other words, just multiple regression. At the other extreme are models with several latent variables which not only have causal relationships with the observed variables, but also have causal relationships with each other.

Remember that the observed variables are often referred to as **indicators** – that is, indicators of the latent variable. For example, your score on a visual-spatial test (a measured variable) can be an *indicator* of your visual-spatial ability (an unmeasured latent variable).

Path diagrams

Path diagrams are extremely useful for SEM – the structure of a complex model can be quite difficult to grasp just through a series of equations.

We now use the convention of denoting measured variables by **V** (V1, V2, etc.) and latent variables (or factors) by **F** (F1, F2, etc.); we use **E** for the errors of measured variables, and **D** for the errors (or *disturbances*) of latent variables.

Recall that in path diagrams, measured variables are enclosed in rectangles, and latent variables (both factors and errors) are enclosed in circles, although in some texts and software packages, errors are not enclosed.

Single-headed arrows denote a causal relationship, but double-headed arrows denote covariance or correlation *without an implied direction of causality*.

Example: two factor model

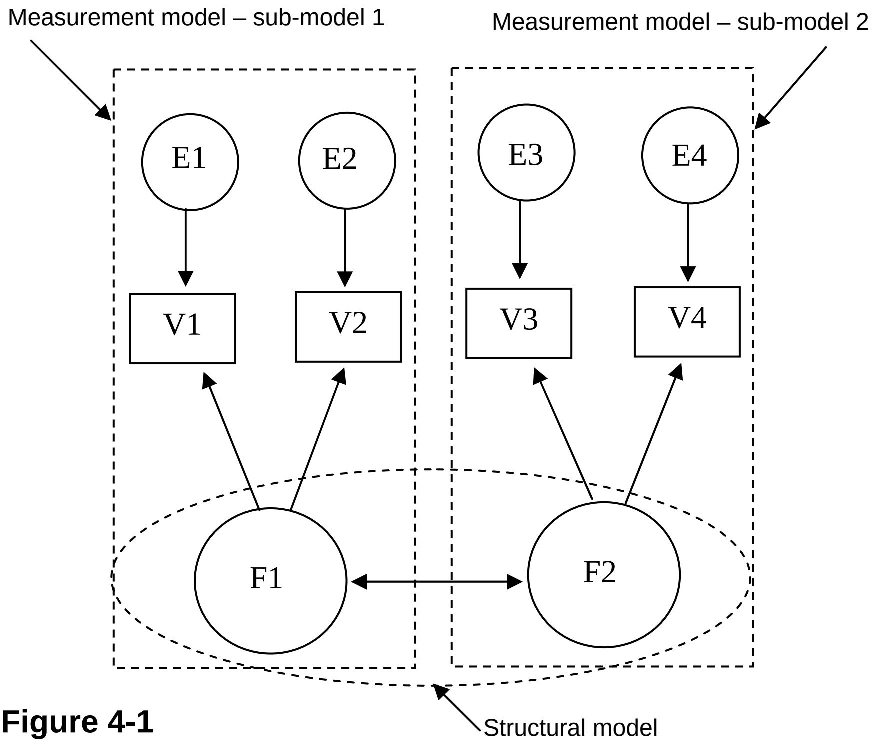


Figure 4-1

The set of connections between the observed (measured) and unobserved (latent) variables is the **measurement model**. In Figure 4-1 there are 2 measurement sub-models.

The connections between the latent variables is usually referred to as the **structural model**. Here it is a simple covariance between the 2 factors, but later we will consider more complex structures.

Independent variables (IVs) do *not* have single-headed arrows pointing to them. Dependent variables (DVs) *do* have single-headed arrows pointing to them.

In the above diagram:

IVs	F1	F2	E1	E2	E3	E4
DVs	V1	V2	V3	V4		

Notice that all DVs have errors (Es) pointing to them.

It is hypothesised that F1 and F2 are correlated, but not that one 'causes' the other.

Regression using covariances

If we have two measured variables, V1 and V2, which have a common factor F1, we can write the equations:

$$V1 = a + b\,F1 + E1$$

$$V2 = c + d\,F1 + E2$$

notice there are no parameter estimates for the errors – they are set at 1

Usually, we are not interested in the intercept terms, *a* and *c*, but only in the slope parameters, *b* and *d* (the single-headed arrows on path diagrams).

Recall that when investigating the existence of latent variables, we are concerned with explaining the correlation between variables, by using the *covariances*.

We assume that any two measurement errors, E*i* and E*j*, are uncorrelated, and that each E*i* is also uncorrelated with any factor F*k*. So:

$$\text{Cov}(E_i, E_j) = 0 \qquad\qquad \text{Cov}(F_k, E_i) = 0$$

Using these assumptions and the two simple regression equations above, we can therefore write:

$$\text{Var}(V1) = b^2\,\text{Var}(F1) + \text{Var}(E1)$$

$$\text{Var}(V2) = d^2\,\text{Var}(F1) + \text{Var}(E2)$$

(ignoring the intercept terms *a* and *c*).

In addition, we can write

$$\text{Cov}(V1, V2) = b\,d\,\text{Var}(F1).$$

From our observed data, we have all the variances and covariances of the *measured* variables (the Vs), but we do not have variances and covariances relating to the *latent* variables (the Fs and Es). In this example, we will therefore need estimates of Var (F1), Var (E1) and Var (E2) from the modelling process, as well as estimates of the slope parameters *b* and *d*.

In general, we use the observed covariance matrix to obtain estimates of the slope parameters, the variances of the factors, and the variances of the errors (plus any factor covariances as necessary).

Model identification

A model is said to be *identified* if there is a unique numerical solution for each of the parameters in the model.

The first issue regarding model identification is **establishing the scale of the factors**.

We return to the simple 2-DV, one factor example for illustration, but this concept applies in the general case. We have equations:

$$\text{Var (V1)} = b^2 \text{ Var (F1)} + \text{Var (E1)}$$

$$\text{Var (V2)} = d^2 \text{ Var (F1)} + \text{Var (E2)}$$

$$\text{Cov (V1, V2)} = b\, d \text{ Var (F1)}$$

for which there are *no unique numerical solutions*.

This is because the estimated value of Var (F1) can be divided by a constant *m* and both *b* and *d* multiplied by $\sqrt{m}$, without altering the predicted variances and covariances of V1 and V2.

For example:

Var (F1) = 4, *b* = 2, *d* = 3	Var (V1) = 2^2.4 + Var (E1) = 16 + Var (E1) Var (V2) = 3^2.4 + Var (E2) = 36 + Var (E2) Cov (V1, V2) = 2 . 3 . 4 = 24

Divide Var (F1) by 4, multiply *b* by 2 and *d* by 2:

Var (F1) = 1, *b* = 4, *d* = 6	Var (V1) = 4^2.1 + Var (E1) = 16 + Var (E1) Var (V2) = 6^2.1 + Var (E2) = 36 + Var (E2) Cov (V1, V2) = 4 . 6 . 1 = 24

The solution is to fix one of the parameters involved – either Var (F1), or one of the regression coefficients (*b* or *d*) is fixed to 1 (or any other constant) before modelling commences.

If the factor is an IV, either alternative is acceptable.

When the factor is a DV, it is more common to fix a regression coefficient.

The scale of ***each latent variable in the model*** should be set in this way. If there are several Fs, a parameter must be fixed for each of them.

Remember that ***errors are also latent variables***. Therefore, we need to set the scale of each of these too. This is almost always done by setting the slope parameter to 1, and is usually done automatically by software packages. Setting the scale of the error latent variables is why we write our equations e.g.:

V1 = *b* F1 + E1 *i.e. with no parameter estimate for the error*

The second identification issue is concerned with the **number of observed moments** (the number of variances and covariances between the observed variables), or 'data points' and **number of parameters**.

Note that Cov (V1, V2) = Cov (V2, V1) and counts only once!

To calculate the moments for *p* measured variables, we use *p(p+1) / 2*.

The number of parameters is obtained from the total of the regression coefficients (or slope parameters), the factor variances, the error variances (and any factor covariances) which are to be ***estimated*** – that is, not including any parameters you fix in order to set the scale of the factors.

Then:

1. If there are more moments (data points) than parameters, the model is 'over-identified', which is a necessary condition for proceeding with the analysis.

2. If there are the same number of data points and parameters, the model is 'just-identified'. Analysis can be performed, but the parameter estimates simply reproduce the sample covariance matrix and hypotheses cannot be tested.

3. If there are fewer data points than there are parameters, the model is 'under-identified' and we cannot proceed.

The easiest way to improve an under-identified model is to simply delete some of the parameters.

If this is not sensible, a parameter may be fixed to a specific value or constraints can be introduced – setting one parameter equal to another.

In the two-DV, one factor example shown Figure 4-2 below, we have:

3 data points – Var (V1), Var (V2), Cov (V1, V2)

4 parameters – *b*, *d*, Var (E1), Var (E2)

Var (F1), and the slope parameters for E1 and E2 are fixed to 1

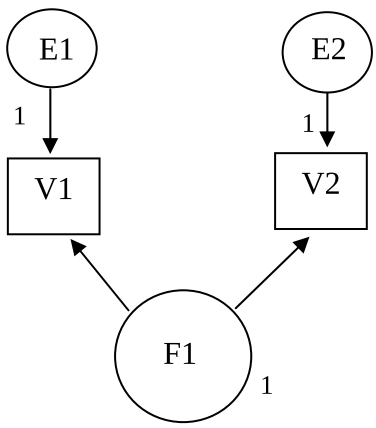


Figure 4-2

This model is under-identified, but in such a simple model, there is little chance of deleting unnecessary paths or placing sensible constraints on the parameter estimates.

However, in a four-DV, one factor model:

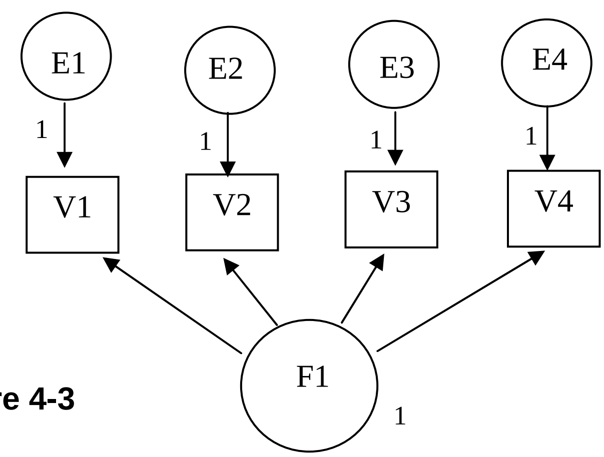


Figure 4-3

10 data points – 4 observed variances, 6 observed covariances

8 parameters – 4 regression coefficients, 4 error variances
Var (F1), and the slope parameters for E1 to E4 are all fixed to 1

This model is over-identified.

Notice that establishing the scale of a factor reduces the number of parameters to be estimated by 1, thus assisting in achieving an over-identified model.

Using EFA results to form hypothesised models

EFA using maximum likelihood factor analysis gives *factor loadings* for the variables, which are interpreted as correlations between the factor and the variables.

Every variable in the analysis has a factor loading, but some are much higher than others. The greater the loading, the more the variable is a pure measure of the factor.

In a hypothesised model, we form pathways between factors and variables only where the loading is sufficiently high, and omit pathways between low correlated factors and variables – these variables do not 'belong' with the factor.

'Sufficiently high' means high **absolute** values, that is, large **positive or negative** loadings, although, admittedly, negative loadings are not always easy to understand in terms of latent variable descriptions. As an example, consider a latent variable 'dislike of school'; this might be related to measures of academic success such as test scores. It could be hypothesised that a child who has a great dislike of school will perform less well on tests – this is a negative relationship.

As a rule of thumb, variables with loadings of 0.32 and larger are collected into a factor, the rest are omitted. However, this criterion is a matter of choice, and sometimes the cut-off point is chosen differently because one can interpret factors formed using one cut-off but not when using another.

Comrey and Lee (1992)[1] suggest the following be used to help in the decision:

Factor loadings above:	Overlapping variance	
0.71	50%	Excellent
0.63	40%	Very good
0.55	30%	Good
0.45	20%	Fair
0.32	10%	Poor

Sometimes there is a gap in the loadings, and if the chosen criterion falls in this gap it is easy to specify which variables should be included and which not. Otherwise, it may be necessary to modify the structure of the factors after the first stages of the modelling process.

1 Comrey, A. L. and Lee, H. B. (1992) *A first course in factor analysis* (2nd Ed.) Hillsdale, NJ: Erlbaum

Measurement models in Amos

To bring all the above theory together, we will now calculate the estimates for a simple one factor measurement model in Amos.

Attig (1983) showed 40 subjects a booklet containing several pages of advertisements, and then gave each subject 3 memory performance tests:

Test	**Requirement**
Recall	To recall as many advertisements as possible
Cued	To recall as many advertisements as possible after receiving cues
Place	To recall the place in the booklet of given advertisements

After the tests were carried out, a training exercise intended to improve memory performance was given, and then the 3 tests were administered again. In the data set, the scores for the first set of tests are in the variables *recall1, cued1, place1*, and the scores for the second set of tests are in the variables *recall2, cued2, place2*. A variety of further variables are also recorded (e.g. sex, age). The data is in the file *attigall.sav* (there are two further versions of the data set which contain only young or old subjects).

For the purposes of this example, we hypothesise that a single latent variable, 'General memory' (F1), has an effect on the 2 recall and 2 place variables:

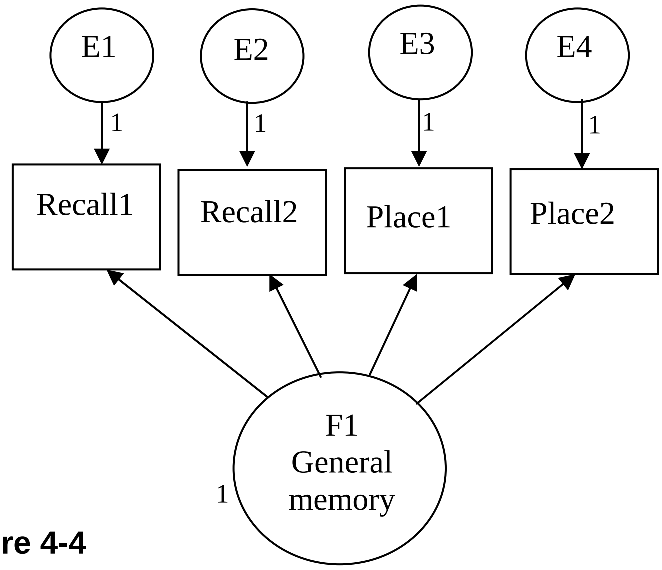


Figure 4-4

The regression equations are:

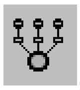
Recall1 = *a* F1 + E1 Place1 = *c* F1 + E3
Recall2 = *b* F1 + E2 Place2 = *d* F1 + E4

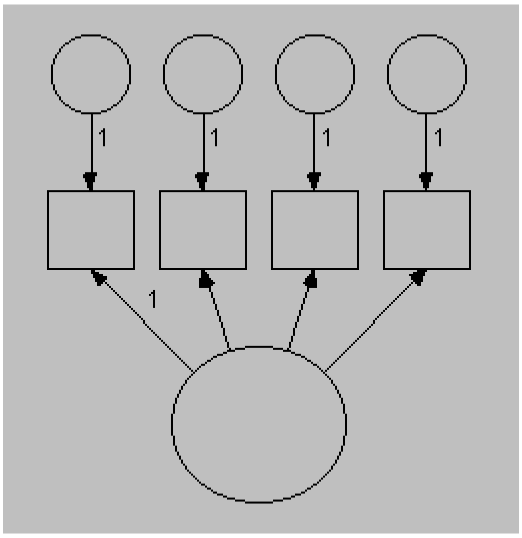
As explained previously, in order to set the scale of the factor, either the variance of the factor or one of the regression weights must be fixed at 1. In this case we have chosen to set the variance of F1 to be 1, and this is marked in Figure 4-4 as the number 1 next to the latent variable.

Notice that the regression weights of the errors are assumed to be 1 (marked as fixed to 1 on the path diagram in Figure 5-4). The errors in these equations are also latent variables, and therefore must have the scale set. It is usual to set the regression weight to 1 (as here), and estimate the error variance.

In this example, there are 10 data points (the 4 variances and 6 covariances between the 4 observed variables) and 8 parameters to be estimated – the 4 regression coefficients between F1 and the observed variables, and the 4 error variances. This is therefore an over-identified model.

In Amos (Graphics), we attach the file *attigall.sav*, and then draw the path diagram (Figure 4-4) on the ‘page’, choosing the appropriate tools to draw the observed variables (rectangles), the factor (a latent variable in an ellipse), the errors (latent variables in ellipses), and the single-headed arrows.

When drawing path diagrams of this nature, a short cut is available. Choosing the ‘Draw a latent variable or add an indicator…’ tool , we can draw the factor ellipse (left click and drag out the shape before releasing), and then left-click in the ellipse to attach (with arrows) observed variable rectangles with connected error ellipses – one click for each observed rectangle required. Doing this gives a diagram as follows:

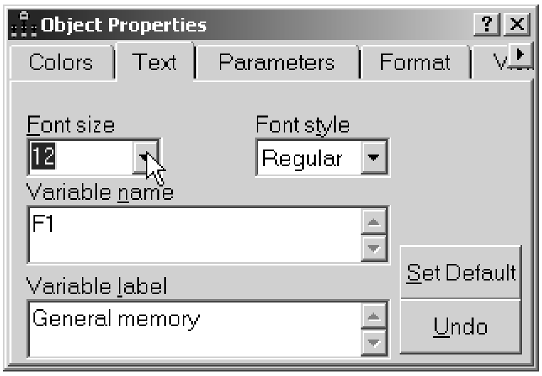


Notice how this has automatically fixed at 1 the error regression weights, and also the regression coefficient between the factor and the first variable. We want to change this, but first we add the variable names.

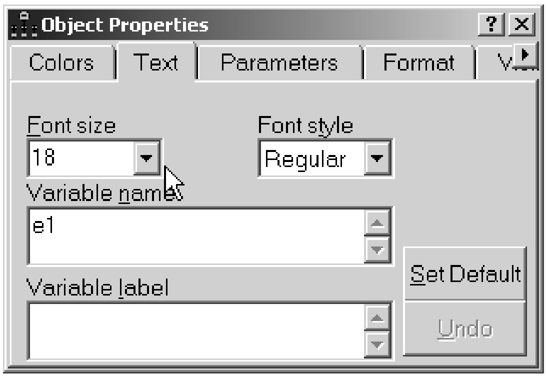
As before, we can drag the observed variables to the appropriate rectangle from the list of variables in the data set. When drawn using the shortcut tool, the rectangles may be too small to fit the name of the variable inside – we will deal with this in a moment.

For the errors and the factor, we should double click on one of these to launch the ‘Object properties’ pop-up window, and enter these variable names on the ‘Text’ tab. Labels can also be added, such as ‘General memory’ for F1.

Also on the ‘Text’ tab is a font size option – we can reduce the size of the variable name text to fit into the shapes instead of increasing the size of the shape itself. Try reducing the font from 18pt to 12pt or 11pt if necessary.



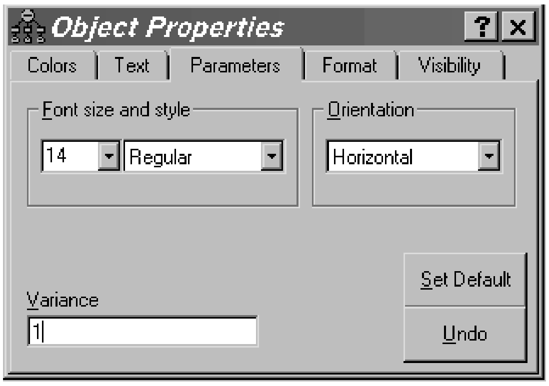
Entering factor details



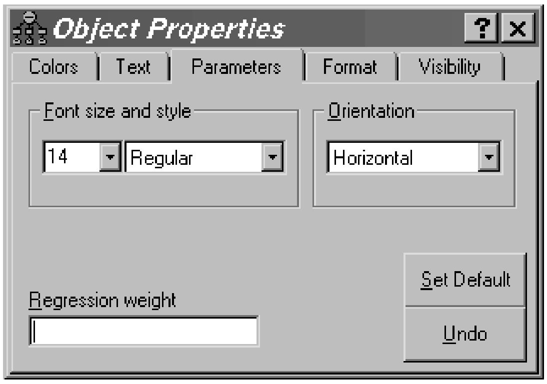
Entering error details

While this pop-up window is still open, select the factor, and change to the ‘Parameters’ tab. Here we can set the variance to 1.

Similarly, the arrows can be selected, and the regression weight set or unset on the ‘Parameters’ tab. Select the arrow where Amos automatically set the weight to 1 and delete the number in the regression weight box.



When a latent variable is selected



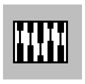
When a path (arrow) is selected

After doing this, the path diagram in Amos should look like Figure 4-4. If the fixed parameters are hidden by arrows, or are too close to other elements, they can be moved by choosing the ‘Move parameter values’ tool  and dragging them to a better placement.

Before analysis, we need to save the diagram file – *sess4a.amw*.

Running the analysis now will provide a standard amount of output – however, as in Exercise 1, we may wish to request additional output. From the

‘View/set’ menu, choose ‘Analysis properties’ or click the  button on the toolbar. On the ‘output’ tab choose the ‘standardized estimates’ and ‘Squared multiple correlations’ for this example, and close this window.

Run the analysis, choosing either ‘Calculate estimates’ from the ‘Model-Fit’ menu, or the abacus-shaped tool  on the toolbar.

Amos output

When the estimates have been calculated, ‘Finished’ will appear in the ‘Computation summary’ window. Open the output by choosing ‘Text output’ from the ‘View/set’ menu, or clicking on  on the toolbar. We will now begin to go through the output page by page – select the pages from the scroll menu at the top left of the output window. It is simple to print individual pages from this output, or the entire file, just by clicking on the print icon and choosing a page range or ‘All’ as required.

The ‘Analysis Summary’ page just gives the date and time the analysis was run, along with the file name and any descriptive title you may have added.

The ‘Notes for group’ output is simple for this model – it identifies the sample size and whether the model is ‘recursive’ or ‘non-recursive’. Briefly, a recursive model is one in which we can step back through the single-headed arrows from a DV to the IVs without encountering a loop. Non-recursive models contain loops.

Notes for group

The model is recursive.

Sample size = 80

‘Variable summary’ lists out the variables in the model, along with their description (observed, unobserved etc.), and a summary of the number of each type.

Variable Summary (Group number 1)

Your model contains the following variables (Group number 1)

Observed, endogenous variables
RECALL1
RECALL2
PLACE1
PLACE2

Unobserved, exogenous variables
F1
E1
E2
E3
E4

Variable counts (Group number 1)

Number of variables in your model:	9
Number of observed variables:	4
Number of unobserved variables:	5
Number of exogenous variables:	5
Number of endogenous variables:	4

In a similar manner, 'Parameter summary' shows the number of the various types of parameters – regression weights, variances and covariances, means and intercepts. Each of these types is further split into 'fixed' parameters (such as the fixed factor variances in order to set the scale of the factor, or the fixed regression weights of the errors), and 'Labelled' or 'Unlabeled' parameters – these are the parameters to be estimated, and will usually be of the 'Unlabeled' type in this course. We will discuss the 'Labelled' type later when we introduce constraints on parameters.

Parameter summary (Group number 1)

	Weights	Covariances	Variances	Means	Intercepts	Total
Fixed	4	0	1	0	0	5
Labeled	0	0	0	0	0	0
Unlabeled	4	0	4	0	0	8
Total	8	0	5	0	0	13

The 'Notes for model' page shows the number of degrees of freedom – the difference between the number of 'sample moments' (data points, or number of observed variances and covariances) and the number of parameters to be estimated. This difference is what we calculated to ensure that our model was over-identified.

Notes for Model (Default model)

Computation of degrees of freedom (Default model)

Number of distinct sample moments:	10
Number of distinct parameters to be estimated:	8
Degrees of freedom (10 - 8):	2

Result (Default model)

Minimum was achieved Chi-square = 13.80639302 Degrees of freedom = 2 Probability level = .00100457

A note on the minimisation process is made – this is concerned with the computation process by which the parameters are estimated (a later section of output, 'Minimization history', shows the details of this).

The last few lines of this section of output show how well the model fits the data, using the chi-squared test. Alternatively, if there are problems with the estimation process (such as omitting a fixed parameter to enable the scale of the factor to be set), an appropriate message will be printed here to alert you.

The null hypothesis for models of this kind is that the model fits the data perfectly. The chi-squared statistic is therefore a measure of discrepancy – i.e. how far the values predicted by our current model differ from the observed values. What we are aiming for is a model that, given the number of degrees of freedom, shows only a small discrepancy from the perfect fit – choosing the 5% significance level, this would be a chi-squared statistic that has a probability of 0.05 or more of occurring by chance when the model fits the data perfectly *in the population*. If the p-value (probability level) shown here is less than 0.05, there is sufficient evidence at the 5% level to reject the null hypothesis that the model fits the data – the alternative hypothesis is that the model does *not* fit the data well.

Here, with the chi-squared statistic of 13.806 on 2 df, giving a p-value of 0.001, we see that this one-factor model does not fit the data well.

The 'Estimates' page provides not only the parameter estimates, but also standard errors and statistical tests for the estimates (which we will discuss further in the next session). If any parameters are 'Labelled', (see the introduction of constraints later in the course), this will be marked here.

Scalar Estimates (Group number 1 - Default model)

Maximum Likelihood Estimates

Regression Weights: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
RECALL1 <--- F1	1.21396305	.28842585	4.20892592	***	
RECALL2 <--- F1	1.83690571	.40187546	4.57083321	***	
PLACE1 <--- F1	4.47605781	.62633743	7.14640003	***	
PLACE2 <--- F1	5.39150832	.65668754	8.21015777	***	

Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
RECALL1	<---	F1	.48370845
RECALL2	<---	F1	.52028431
PLACE1	<---	F1	.77753592
PLACE2	<---	F1	.88078936

Variances: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
F1	1.00000000				
E1	4.82488728	.81531381	5.91782848	***	
E2	9.09077693	1.55905154	5.83096625	***	
E3	13.10474976	3.51149440	3.73195804	***	
E4	8.40101251	4.26285938	1.97074587	.04875295	

The regression weights in the first part of this output allow us to write the equations:

Recall1 = 1.214 F1 + E1 Place1 = 4.476 F1 + E3
Recall2 = 1.837 F1 + E2 Place2 = 5.492 F1 + E4

These equations are interpreted in the same manner as the usual multiple linear regression equations. For example, for every 1 point increase in the latent variable ‘General memory’ (F1), the score on the first ‘recall’ test increases by 1.214 points, and the score on the first ‘place’ test increases by 4.476. Similarly, we can see how much change on the ‘recall2’ and ‘place2’ scores is obtained from a one point increase in ‘General memory’.

We requested the standardised estimates before running the analysis. These standardised regression weights are calculated after the variables have been rescaled to have unit variance, and we can therefore compare the regression weights to each other. For example, the standardised regression weight for F1 on ‘place2’ is the largest of the four, indicating that ‘General memory’ has the greatest effect on this score; the smallest effect is on ‘recall1’, followed by ‘recall2’.

The next section of this page of output gives the estimates of the variances. Notice that the variance of F1, which was fixed to 1, is shown as having an ‘estimate’ of 1, but the lack of a standard error (S.E.) and statistical test shows it to be a fixed parameter.

The estimates of the variances of the errors can be thought of as the precision with which the observed (dependent) variables are measuring the latent variable (or variables, as we will see later). Large error variances indicate that there is a lot of variability in the observed variable that cannot be explained by the factor(s). Small error variances occur when the variance of the observed variable is close to that of the factor, i.e. when the observed variable is a fairly ‘pure’ measure of the latent variable. A completely ‘pure’ measure of the latent variable is when the variance of the observed variable equals the variance of the factor(s) – when the variance of the error is zero. We will discuss in the next session how we consider whether the parameter estimates given in this part of the output are significantly different from zero.

The size of the error variances must be related to the original scale of the dependent variable. In this example, the variances for *recall1* and *recall2* are much smaller than those for *place1* and *place2* (6.378 and 12.623 for the two ‘recall’ variables respectively; 33.559 and 37.994 for the two ‘place’ variables respectively).

Now, recall, for example:

$$\text{Var}(\textit{recall1}) = b^2\ \text{Var(F1)} + \text{Var(E1)}$$

We now have each of these items, so could (if we wished) calculate how much of the variance of the measured, dependent variable has to be accounted for by error. However, by requesting the ‘Squared multiple correlations’, AMOS provides the complement of this – the proportion of the variance of each measured, dependent variable which is accounted for by the model:

Squared Multiple Correlations: (Group number 1 - Default model)

	Estimate
PLACE2	.77578989
PLACE1	.60456211
RECALL2	.27069577
RECALL1	.23397387

Notice the order of the variables is not necessarily the same as before

The dependent variable with the highest proportion of variance explained by the model is *place1*, with 77.6%; *place2* has the next highest proportion, followed by *recall2* at 27.1% and *recall1* at just 23.4%.

Therefore, taking into account the scale of the variables, *recall1* has the largest error variance associated with it, *recall2* the next, *place1* the next, and *place2* has the smallest associated error variance.

Of these four indicators, *place2* is the purest measure of ‘General memory’. The two ‘recall’ variables are much less pure measures, with a large proportion of their variances being accounted for by error.

Exercise 4

1. Track records at the 1984 Olympics

Recall the *track.sav* data from session 3 – this consists of the national records for track events ranging from 100 metres to the marathon for a number of countries. The reciprocals for each variable have been calculated to give more 'natural' measures – larger values equate to faster speeds.

In Amos, attach this data set, and draw an input diagram where there is a single common factor F1 for the 5 observed variables 100 metres, 200 metres, 400 metres, 800 metres, 1500 metres (r100, r200, etc.)

Name the 5 errors as 'E100', 'E200' etc. Set the scale of the factor by fixing the factor variance to 1 (instead of fixing a regression weight). Remember to ensure the error regression weights are fixed at 1.

a) Request the standardised estimates and 'Squared multiple correlations' in your output. Save the diagram as *exer4-1.amw* and run the analysis.

b) Does the fitted model fit the data well? *(Look at 'Notes for Model')*

c) Consider the standardised results – does the factor have a noticeably larger effect on any of the events?

d) To view the error variances, you may need to increase the number of decimal places displayed for the estimates. There is a drop-down list on the toolbar of the 'Estimates' window – ensure this is at least 8 d.p. so that the error variances no longer just show as '0.000'.

e) According to the 'Squared multiple correlations', which observed variables are measuring the factor most precisely? Compare your conclusions with the size of the (raw) error variances, and make sure you understand the effect of the scale of the original variables.

f) Given the positive direction of the regression weights, and the use of *reciprocals* for the observed variables, can you think of a suitable description for the common factor?

2. Holzinger and Swineford psychological tests

Recall the data on 6 psychological tests for 145 students in Exercise 2 and Exercise 3. The scores for the 6 tests and a gender variable are held in the data file *Holz-Swin.sav*.

Create an Amos input file for this data with a single common factor F1 for the 6 tests. Fix parameters as appropriate.

a) Request the standardised estimates and 'Squared multiple correlations' in your output. Save the diagram as *exer4-2.amw* and run the analysis.

b) Does the fitted model fit the data well?

c) Consider the standardised results – does the factor have a noticeably larger effect on any of the tests?

d) How do changes in the factor affect the observed variables? According to the 'Squared multiple correlations' and error variances, which observed variables are measuring the factor most precisely?