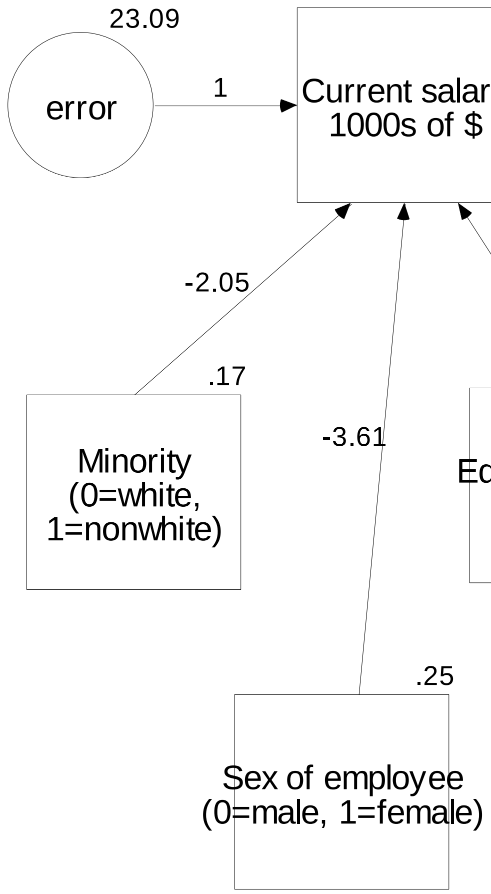
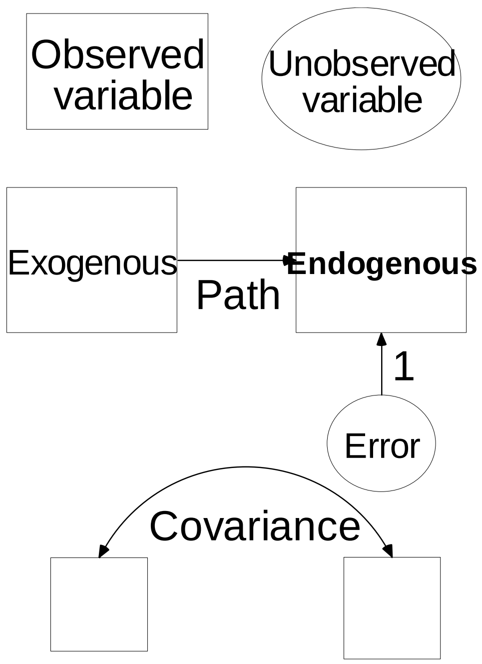
Examples and Notes on SEM

Example 1: Multiple Regression with Observed Variables





Notes: The coefficients are unstandardised.

The number of sample 'moments' and therefore the **total degrees of freedom** is equal to 4(4 + 1)/2 = 10, i.e., K (K+1) /2 where K is the number of observed variables.

Amos estimates path coefficients, the variance of errors, and the variance of exogenous variables. The **number of parameters to be estimated** therefore equals 3 + 1 + 3 = 7.

The **degrees of freedom of the model** are therefore 10 – 7 = 3.
(Dataset: *bank3.sav*)

Selected Output for Example 1

Sample Covariances

	SEX	EDLEVEL	MINORITY	SALNOW
SEX	.248			
EDLEVEL	-.511	8.305		
MINORITY	-.016	-.158	.171	
SALNOW	-1.529	12.988	-.501	46.554

The sample covariance is Σ {[X_i - mean(X)] [Y_i - mean(Y)]} / n -1.

Computation of degrees of freedom

Number of distinct sample moments:	10
Number of distinct parameters to be estimated:	7
Degrees of freedom (10 - 7):	3

Regression Weights:

	Estimate	S.E.	C.R.	P	Label
SALNOW <--- MINORITY	-2.047	.534	-3.834	***	
SALNOW <--- EDLEVEL	1.303	.077	16.994	***	
SALNOW <--- SEX	-3.609	.444	-8.135	***	

Implied (for all variables) Covariances

	SEX	EDLEVEL	MINORITY	SALNOW
SEX	.248			
EDLEVEL	.000	8.305		
MINORITY	.000	.000	.171	
SALNOW	-.895	10.820	-.351	41.134

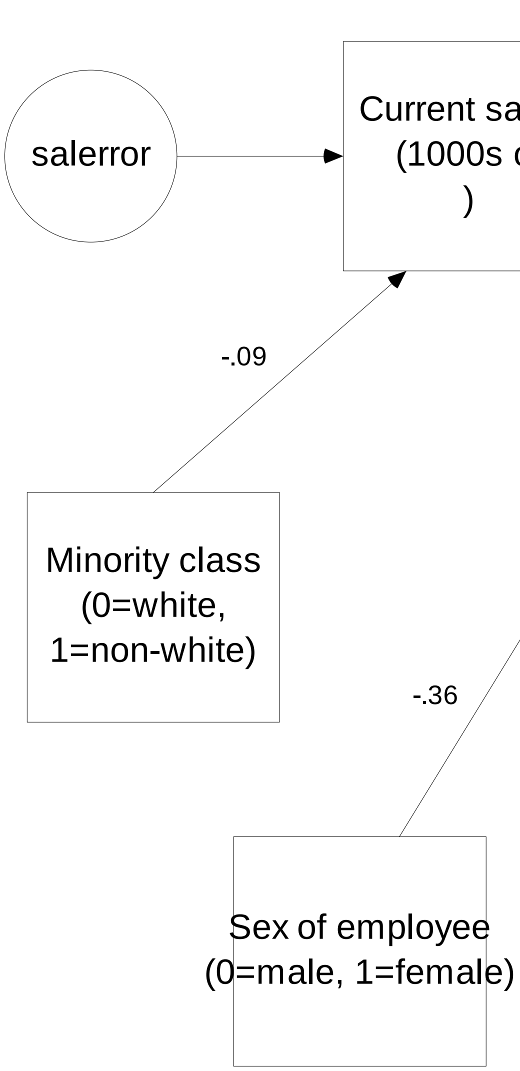
Note the differences between the observed and implied covariances.

CMIN – Chi-squared values

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	7	80.934	3	.000	26.978
Saturated model	10	.000	0		
Independence model	4	412.623	6	.000	68.770

Other indices of fit: χ^2 **/df** = 26.98 (ideally < 2)**, TLI** = .617 (ideally > .95), **CFI** = .808 (ideally > .95), **RMSEA** = .234 (ideally < .05).

Example 2: Indirect Effect



Note: The estimates are standardised. The numbers above the boxes are squared correlation coefficients. Dataset: *bank3.sav*.

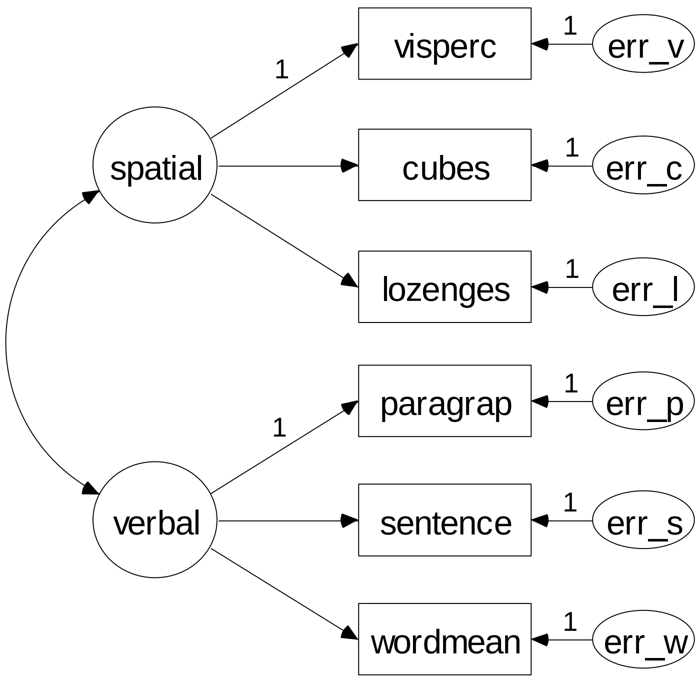
Selected Output for Example 2

Implied (for all variables) Covariances

	SEX	MINORITY	EDLEVEL	SALNOW
SEX	.248			
MINORITY	.000	.171		
EDLEVEL	-.511	.000	8.305	
SALNOW	-.784	-.257	12.750	45.823

Degrees of freedom: Total = 4(4+1)/2 = 10
Parameters to be estimated = 3 + 2 + 2 = 7
Model DF = 10 - 7

Example 3 Confirmatory factor Analysis



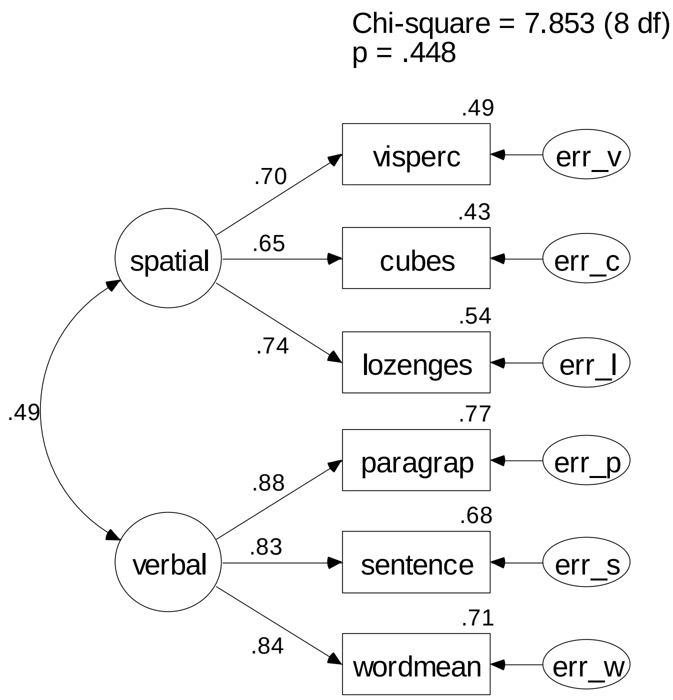
Example 3
Factor analysis: Girls' sample
Holzinger and Swineford (1939)
Model Specification

Dataset: *Grnt_fem.sav*.

Number of distinct sample moments:	21	6(6+1)/2 = 21
Number of distinct parameters to be estimated:	13	5 + 6 + 2 = 13
Degrees of freedom (21 - 13):	8	

Sample Covariances

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	68.260					
sentence	28.845	25.197				
paragrap	21.718	12.864	12.516			
lozenges	23.947	13.228	9.056	61.726		
cubes	6.840	4.036	3.356	17.416	20.265	
visperc	13.037	12.645	8.335	26.531	14.931	47.175



Example 3
Factor analysis: Girls' sample
Holzinger and Swineford (1939)
Standardized estimates

Implied Covariances

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	68.260					
sentence	28.845	25.197				
paragrap	21.718	12.864	12.516			
lozenges	.000	.000	.000	61.726		
cubes	.000	.000	.000	17.416	20.265	
visperc	.000	.000	.000	26.531	14.931	47.175

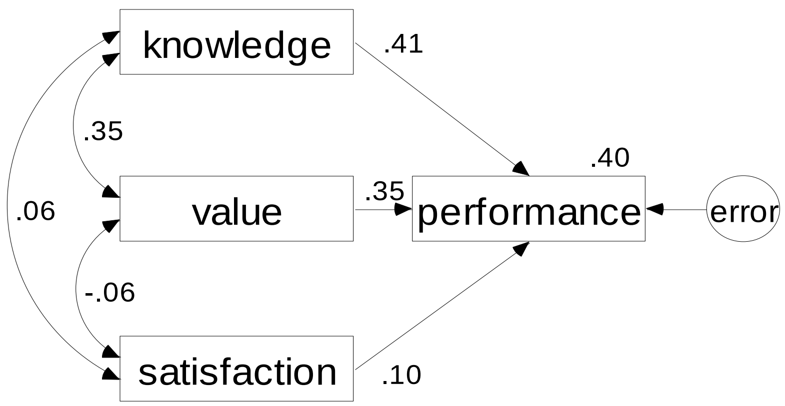
Example 4 Conventional Regression Analysis of Farm Manager Data

IVs: Knowledge: 26 items.

Value: (30) "tendency to rationally evaluate means to an economic end"

Satisfaction: (11) "gratification from the managerial role".

DV: Performance (24 items)



Example 4: Conventional linear regression
Job performance of farm managers
(Standardized estimates)

Variance of performance score accounted for: 40%. Dataset: *WARREN5V (XLS)* in *UserGuide.xls*.

Degrees of freedom: Total = 4(4+1)/2 = 10
Parameters to be estimated = 6 + 3 + 1 = 10
Model DF = 10 – 10 = 0

<u>Example 5 Regression Analysis with Latent Variables</u>

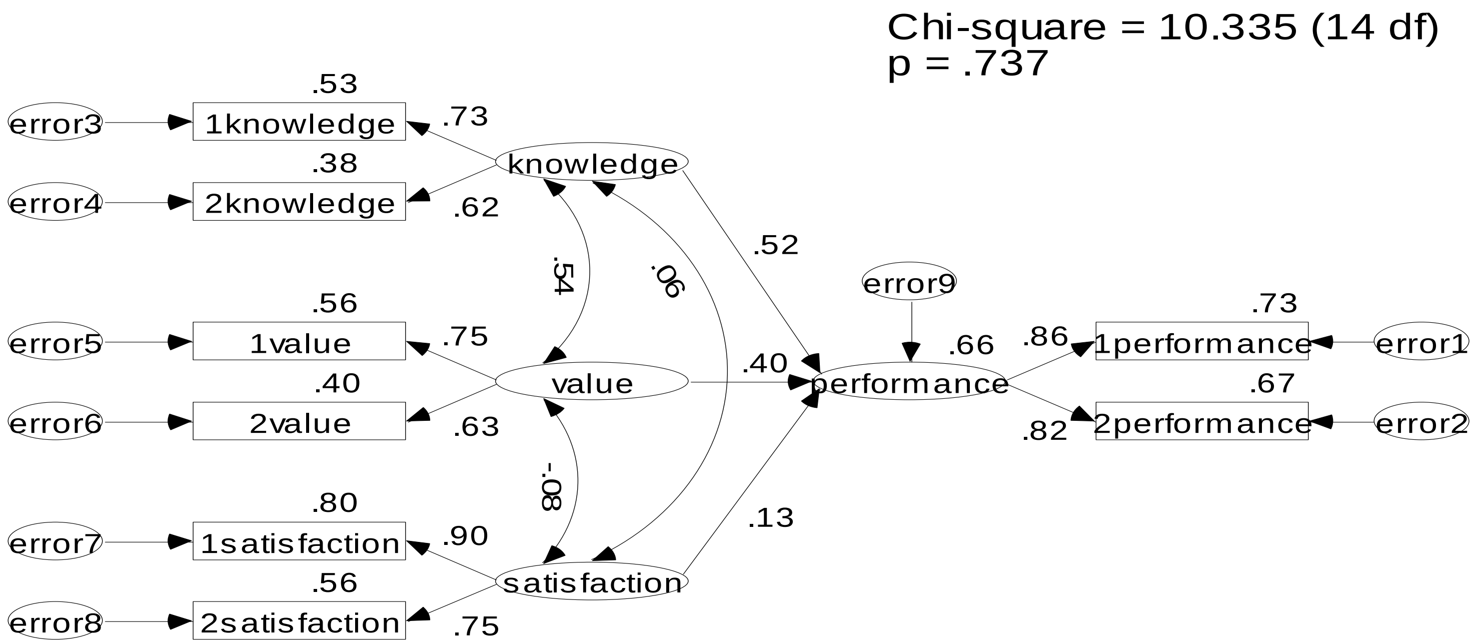
<u>Notes</u>: Use of item parcels

See, for example:

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, 63 (5), 729-757.

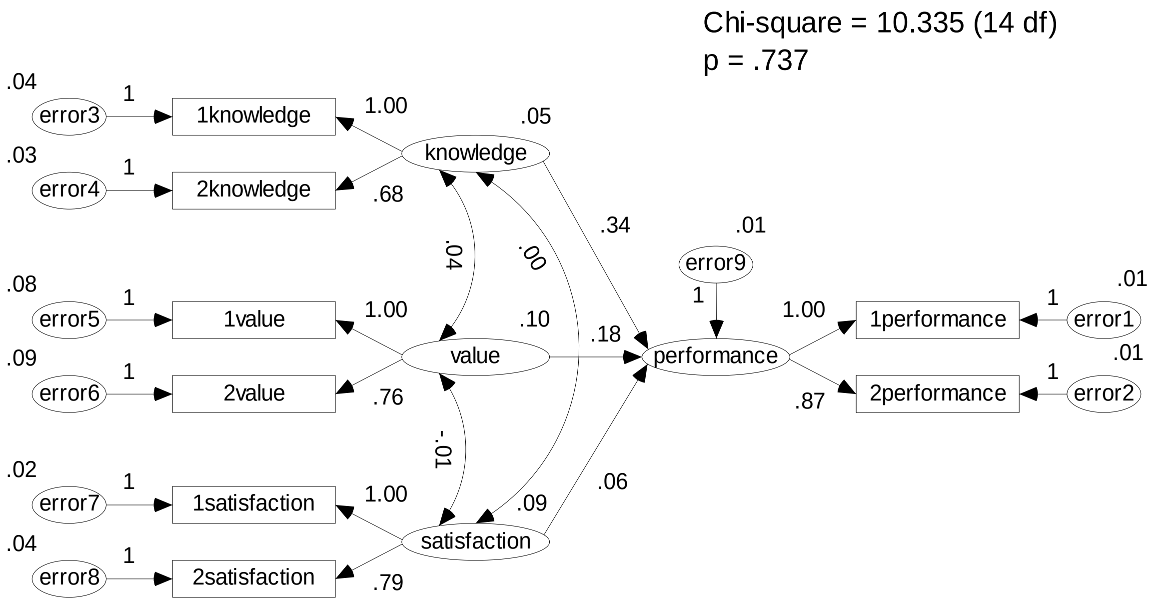
Hau, K.-T., & Marsh, H.W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology*, 57, 327-351.

Degrees of freedom: Total = 8(8+1)/2 = 36
Parameters to be estimated = 7 + 3 + 12 = 22
Model DF = 36 – 22 = 14



Example 5: Regression with unobserved variables
Job performance of farm managers
Warren, White and Fuller (1974)
Standardized estimates

Calculation of the degrees of freedom for Example 5



Example 5: Model A
Regression with unobserved variables
Job performance of farm managers
Warren, White and Fuller (1974)
Unstandardized estimates

Degrees of freedom: Total = 8(8+1)/2 = 36
Parameters to be estimated = 7 + 3 + 12 = 22
Model DF = 36 – 22 = 14

It's easier to work out the *df* for this model if you can see which paths were constrained, which isn't obvious from the standardised results.

Paths (7)

Because one of the two paths from each latent variable to its two observed variables is constrained to 1.0 to give the latent variable a scale, there are just four paths in the measurement part of the model.

There are then three paths, one from each latent variable IV to the latent variable DV, in the structural part of the model.

This gives 4 + 3 = 7 paths altogether.

Covariances (3)

The three latent variable IVs are allowed to covary, which gives three covariances.

Variances of exogenous variables (12)

There are nine error terms (which are exogenous variables), including one for the DV. The three latent variable IVs are exogenous, so this makes 12 variances to estimate altogether.

General Notes

What makes SEM different from carrying out a regression analysis?

(a) **Regression Analyses**

It is possible to carry out a regression analysis using a SEM program such as AMOS (Example 1). Note the conventions of representation of the observed variables, errors and causal paths.

(b) **More Complicated Models**

SEM can handle more complicated models (Example 2). This is an example of path analysis, and includes indirect (mediation) as well as direct effects. We could analyse Example 2 and get path coefficients using two multiple regression analyses.

(c) **Goodness of Fit (GOF)**

However, we would get something from using a SEM program which we wouldn't normally get from regression analysis. That is, we would get an idea of how well the model fitted the data. A SEM analysis starts with a matrix of variances and covariances (see the Sample Variance-Covariance Matrix for Examples 1 and 3 and Implied Covariances for Examples 1-3). GOF in a SEM is judged by how accurately the original variances and covariances are reproduced by the model.

Goodness of fit is not the same thing as variance accounted for – we can have a poorly fitting model which accounts for a large amount of the variance of an outcome variable, and vice versa.

It is possible to compare the fit of different models, and this is one of the most powerful aspects of SEM – we can make direct comparisons of models and decide that some provide implausible accounts of the data and that we can eliminate them from consideration.

Direct comparisons can be carried out by testing the significance of the difference between the χ^2 GOF statistics for two models, as long as one model is nested under the other. Model B is said to be nested under Model A if Model B is obtained by constraining one or more of the parameters in Model A. The most common constraint is that a path coefficient or covariance is set to zero, but other constraints are possible, such as constraining two estimated parameters to be equal, or setting a parameter to a value other than zero.

(d) **Unobserved Variables**

The main difference is that SEM can deal with unobserved variables.

(i) Factor Analysis

An obvious kind of model where we deal with unobserved variables is factor analysis (Example 3). The correlations among measured variables are seen to be explained by the fact that each is a measure or indicator of a latent, hypothesised or unobserved variable.

Here is a strength of SEM – instead of pretending that we meet the assumption that variables are measured with without error, we can acknowledge that our observed variables are IMPERFECT measures of a hypothetical underlying construct and it's these underlying variables, and the relationships between them, that we really want to study. The part of a model that consists of latent variables and their indicators is called the measurement model (as opposed to the structural model, which consists of the relationships between latent variables).

(ii) Regression with Unobserved Variables

In the farm manager example with unobserved variables (Example 5), the items of the questionnaires are divided up to make indicator variables – these are called item parcels or testlets. In general, indicators can be scores for whole tests at one extreme, and individual items at the other.

(e) **Fixing Parameters to a Particular Values**

As mentioned above, another thing which is possible in SEM but not in conventional regression models is that coefficients, variances and covariances can be set to zero or to particular values (or set to be equal for different paths or variances) in order to test hypotheses. This lends itself to testing reduced models, as described above.

Summary

What makes SEM different from conventional regression? We can

1. Deal more readily with complicated models
2. Test the GOF and compare the GOF for different models.
3. Allow unobserved variables, which in turn allows us to look at the relationship of pure versions of variables which we can normally measure only imperfectly
4. Set parameters to particular values or apply particular constraints for model testing.

I think the general aim of SEM is to fit models in which not all possible paths are included (i.e., the model is constrained) but which provide an acceptable account of the data.

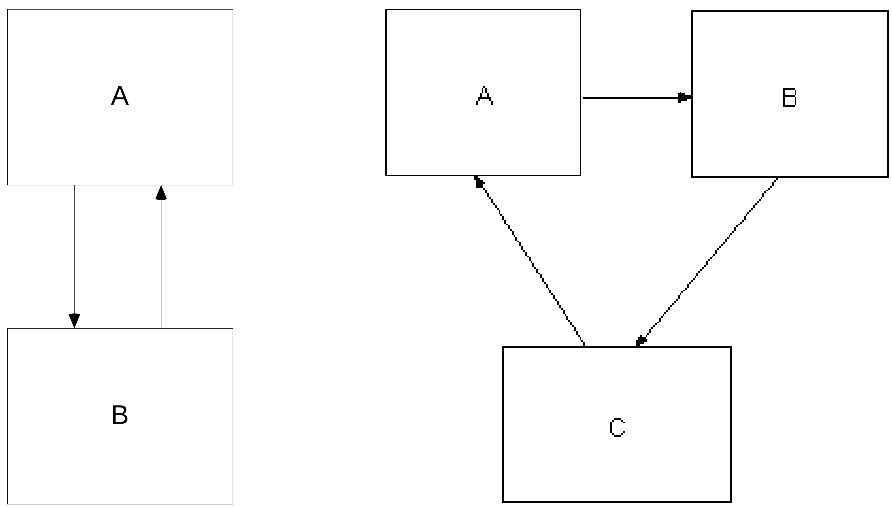
Some General Points & Issues Not Already Covered

Some of these will have been referred to in the examples which we've worked through.

1. Cause. SEM and path analysis are associated with explicit testing of causal paths. But actually, SEMs are no more able to 'prove' cause than a conventional nonexperimental regression analysis. SEM can provide a causal model which is compatible with data as against other models which are not compatible. The onus is still on the user to test sensible, possible, models. For example, the onus is on the user not to fit impossible models, such as ones in which A is the 'cause' of B when it couldn't possibly precede it.

2. Need for a Model. Because of the complexity, there's a need for models to be specified in some detail beforehand, otherwise the analyst can be overwhelmed by the possibilities. I think a reasonable way to proceed is to set up the most optimistic model, the most economical or parsimonious, say, but to have one or more fall-back models. If none of the models fit, then resort to seeing what does fit, but admit that such *post hoc* models are provisional only, and require testing on a new dataset another half of the original dataset). See MacCallum, Roznowski & Necowitz (1992) for a discussion of the dangers of making data-driven changes to models and retesting them on the original sample in order to achieve good fit.

3. Recursive and Non-recursive Models. In recursive models, a variable never causes itself. In non-recursive models, a path beginning at a variable can be traced back to that variable. This may happen, for example, when A causes B, which causes C which causes A (see the example on the right below), or through reciprocal causation (left below). It is possible to fit non-recursive models in AMOS, but we'll deal only with recursive models.



4. Identification. This relates to whether there is enough information available to compute a unique set of estimates of the path coefficients and variances. An example of under-identification would occur if you were given the information that 6 oranges and 3 apples cost $10, and asked what an orange costs. You simply would not have enough information to give a unique answer, but could provide a whole series of possibilities. It is necessary to include constraints in models which lead to the model being identified. A special case is the just-identified model, where the model must fit the data perfectly, because all possible paths are estimated, and no GOF test is possible. To state the aim of SEM in another way – we want to have over-identified models which fit the data well. Another way of putting it is to say that we can't get more information out than we put in. For example, with four variables there are $4(4+1)/2 = 10$ variances and covariances – we can't estimate more than 10 paths, etc.

5. Endogenous and Exogenous Variables. Endogenous variables are caused by variables included in the model – they have at least one single-headed arrow leading to them. Exogenous variables are not caused by any variable in the model – they may covary with other exogenous variables.

6. Assumptions. Any endogenous variable must be normally-distributed like a DV in a regression analysis. Exogenous variables don't have to be normally-distributed.

7. Number of subjects. Most sources advise using 100+ cases, at least 5-10 cases for each parameter estimated. This leads to item parcelling to reduce the number of paths and variances which have to be estimated. The following discussion comes from the University of Texas document, *An Introduction to Structural Equation Modeling using AMOS*, which is available on

http://www.psy.mq.edu.au/psystat/SPSSforWindows.html

A Reasonable Sample Size

Structural equation modeling is a flexible and powerful extension of the general linear model. Like any statistical method, it features a number of assumptions. These assumptions should be met or at least approximated to ensure trustworthy results.

According to James Stevens' Applied Multivariate Statistics for the Social Sciences, *a good rule of thumb is 15 cases per predictor in a standard ordinary least squares multiple regression analysis. Since SEM is closely related to multiple regression in some respects, 15 cases per measured variable in SEM is not unreasonable. Bentler and Chou (1987) note that researchers may go as low as five cases per parameter estimate in SEM analyses, but only if the data are perfectly well-behaved (i.e., normally distributed, no missing data or outlying cases, etc.). Notice that Bentler and Chou mention five cases per parameter estimate rather than per measured variable. Measured variables typically have at least one path coefficient associated with another variable in the analysis, plus a residual term or variance estimate, so it is important to recognize that the Bentler and Chou and Stevens recommendations dovetail at approximately 15 cases per measured variable, minimum.*

More generally, Loehlin (1992) reports the results of monte carlo simulation studies using confirmatory factor analysis models. After reviewing the literature, he concludes that for this class of model with two to four factors, the investigator should plan on collecting at least 100 cases, with 200 being better (if possible). Consequences of using smaller samples include more convergence failures (the software cannot reach a satisfactory solution), improper solutions (including negative error variance estimates for measured variables), and lowered accuracy of parameter estimates and, in particular, standard errors – SEM program standard errors are computed under the assumption of large sample sizes.

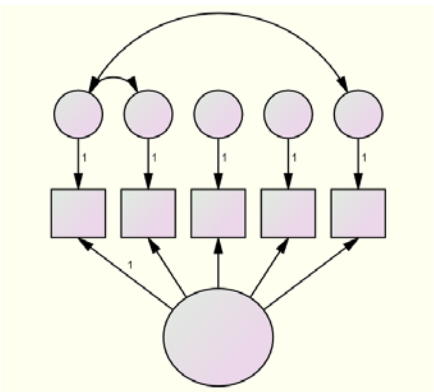
When data are not normally distributed or are otherwise flawed in some way (almost always the case), larger samples are required. It is difficult to make absolute recommendations as to what sample sizes are required when data are skewed, *kurtotic, incomplete, or otherwise less than perfect. The general recommendation is thus to obtain more data whenever possible.*

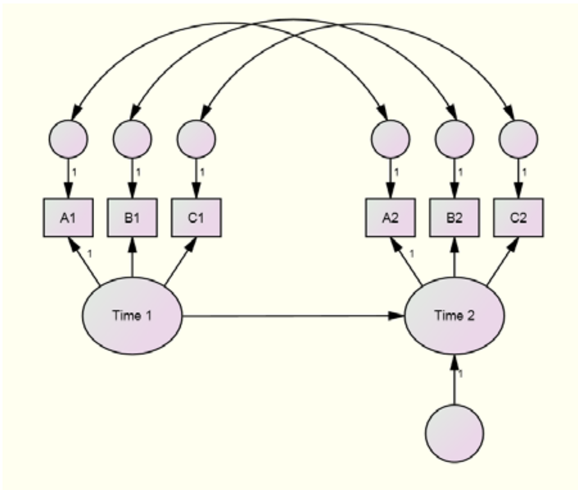
(See the document for references)

8. Goodness of Fit Indices. Problems with the χ^2 GOF index, which is affected by sample size, have led to a multitude of indices. See Thompson p. 269-271 in Grimm & Yarnold, 2000. For reasons given by Marsh, Bella & McDonald (1988) I use the Tucker-Lewis Index (TLI).

9. Equivalent Models. While we may find that our favoured model fits the data, the chances are other models, some substantively meaningful, which fit the data just as well. These are known as equivalent models, and they should be taken into account when you are discussing your results. MacCallum, Wegener, Uchino & Fabrigar (1993) consider this issue.

10. Correlated Residuals. Residuals show the variance of endogenous variables which is not accounted for by the variables in the model on which the endogenous variables are regressed. In some cases the unknown sources of influence from outside the model are common for two or more endogenous variables, and it may be appropriate to allow the residuals to be correlated. A well-known example is common 'method variance', which occurs when the method used to obtain data leads to a





Examples of correlated residuals

correlation between observed variables over and above that which would be expected because of their joint correlation with a common factor. For instance, responses to survey questions which are couched in negative terms tend to correlated regardless of the content of the questions. Also, responses to questions which share common key words (like 'my job') can also be correlated more than would be expected simply on the basis of their tapping a common underlying dimension (these can be modelled as in the first example above). Finally, when a variable is measured on more than one occasion, as in a longitudinal study, it is likely that sources of unexplained variance at one time point will be correlated with (i.e., be the same as) those at another time point, so that it is reasonable to allow the residuals at the different time points to be correlated (second example).

Allowing residuals to be correlated on an ad hoc basis to increase goodness of fit is, of course, frowned on, as the presence of correlated residuals may indicate that the model is inadequate; for example, there may be a need to introduce a further latent variable, or to allow observed variables to load on more than one factor.

These issues are discussed by Cole, Ciesla & Steiger (2007).

11. Testing Mediation Models. SEMs are sometimes used to test for mediation. In two articles Cole and Maxwell (Cole & Maxwell, 2003; Maxwell & Cole, 2007) provide a very helpful framework for testing mediation in latent variable longitudinal studies (the 2003 article) and cast considerable doubt on the value of studies which test for mediation in cross-sectional studies (both articles).

12. Missing Data. Amos can deal with missing data using what is called full information maximum likelihood, which involves maximising the likelihood for the data which have been obtained, and means that cases with some missing values can be included in an analysis. Arbuckle (1996) discusses this method. The other method of choice for dealing with missing data is multiple imputation. Amos has the facility to produce a set of datasets with imputed values, but currently is not able to combine the datasets to provide parameter estimates and standard errors. Mplus does have this latter ability and version 6 is also able to produce multiple imputed datasets. For information about multiple imputation and other strategies, see Sinharay, Stern & Russell (2001) and Schafer & Graham (2002).

References and Useful Reading

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. Chapter 9 in Marcoulides, G. A., & Schumacker, R. E. (Eds) *Advanced strutural equation modeling: Issue and techniques*. Mahwah, N.J.: Lawrence Erlbaum.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12, 381-398.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology,* 112, 381-398.

Hoyle, R (Ed.) (1995). *Structural equation modeling : concepts, issues, and applications*. Thousand Oaks : Sage Publications. [H61.25.S767]

Klem, L. Path analysis. In Grimm, L. G., & Yarnold, P. R. (Eds). *Reading and understanding multivariate statistics*. Washington: American Psychological Association, 1995. [QA278.R43]

Klem, L. Structural equation modeling. In Grimm, L. G., & Yarnold, P. R. (Eds). *Reading and understanding more multivariate statistics*. Washington: American Psychological Association, 2000. [QA278.R42]

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490-504.

MacCallum, R.C., Wegener, D.T., Uchino, B.N., & Fabrigar, L.R. (1993). The problem of equivalent models in applications of covariance structure analysis. Psychological Bulletin, 114, 185-199.

Marsh, H.W., Bella, J.R., & McDonald, R.P. (1988). Goodness-of-fit in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391-410.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23-44.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6, 317-329.

Thompson, B. Ten commandments of structural equation modeling. In Grimm, L. G., & Yarnold, P. R. (Eds). *Reading and understanding more multivariate statistics*. Washington: American Psychological Association, 2000. [QA278.R42]

Alan Taylor
Department of Psychology
Last updated April 15th 2010