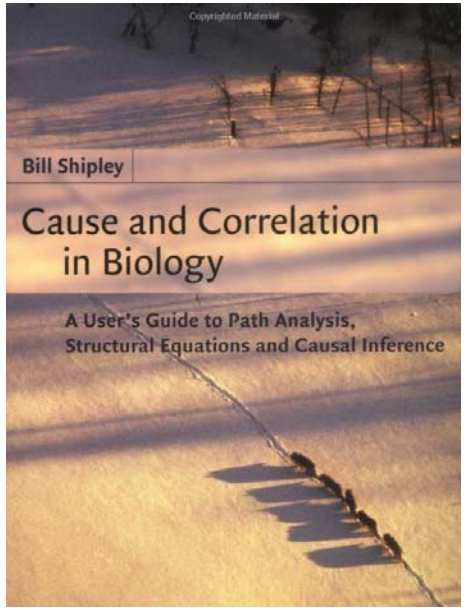
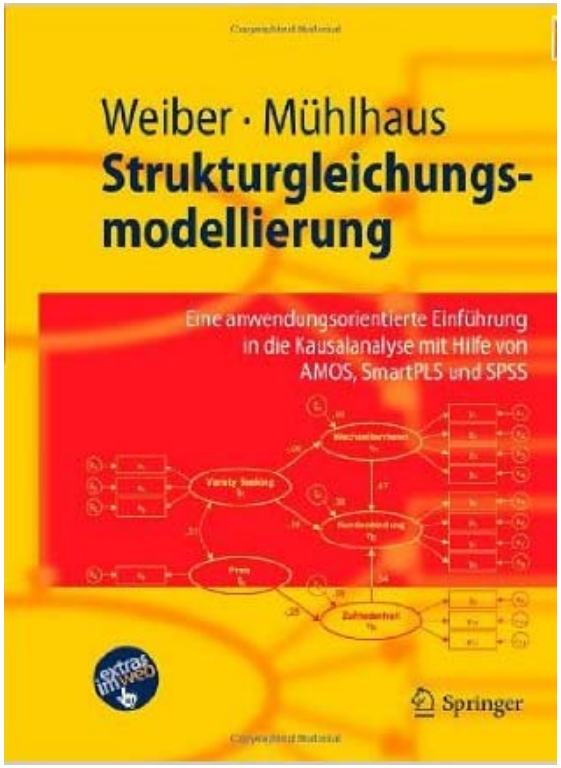
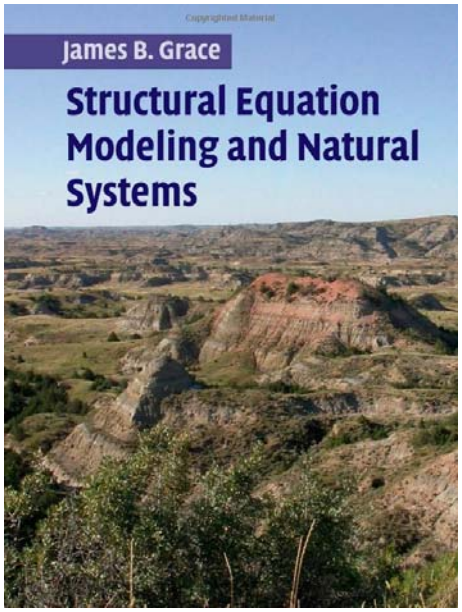
Structural Equation modelling (SEM)

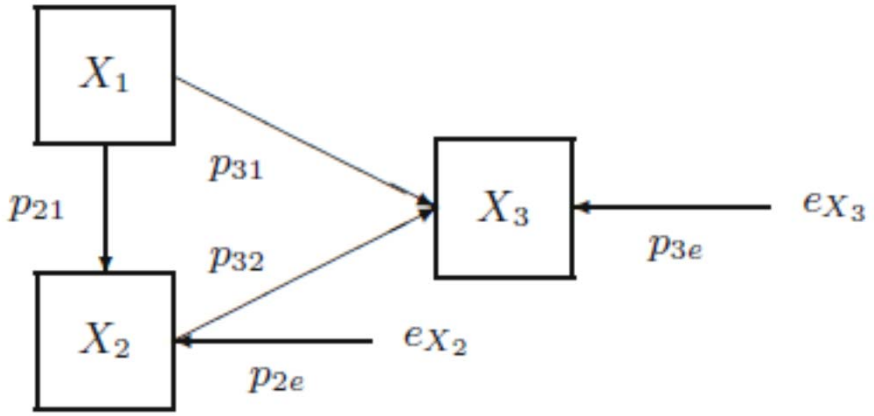
- Test of causal relationships
- Tests multivariate relationships
- Test of *a priori* hypotheses
- inferential statistic

Grace: „Understanding [ecological] systems requires the capacity to examine simultaneous influences and responses“

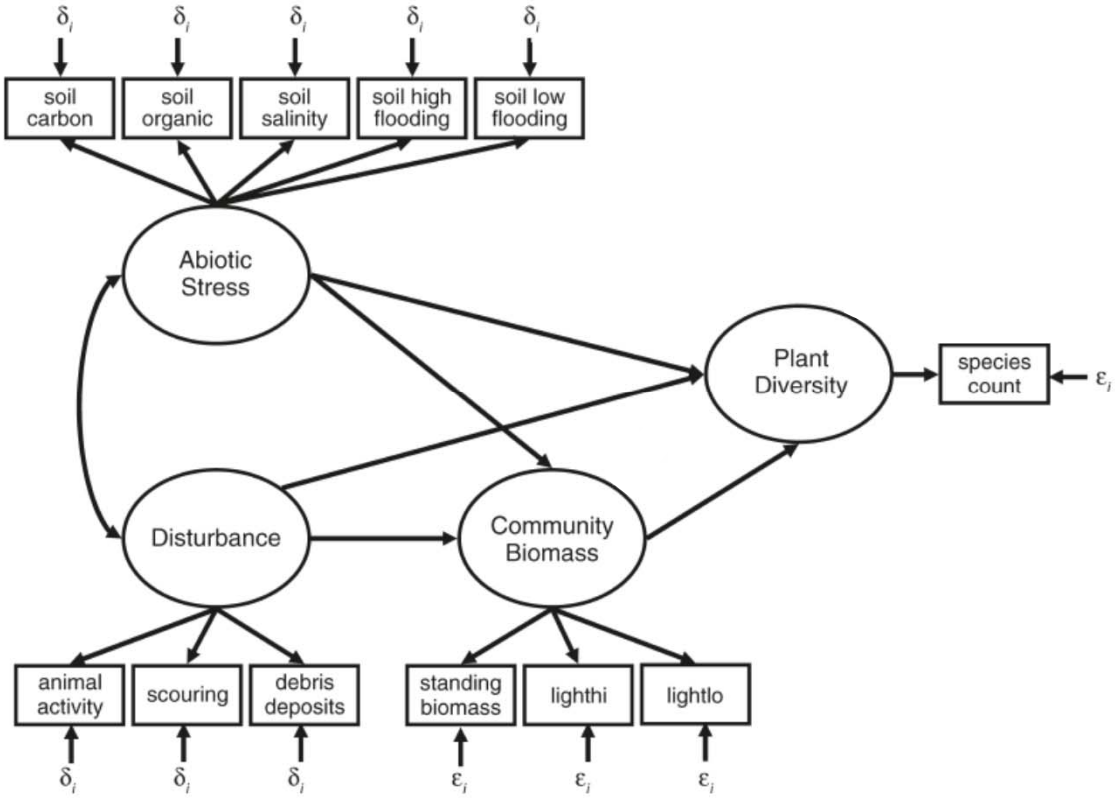

Bill Shipley
Cause and Correlation in Biology
A User's Guide to Path Analysis, Structural Equations and Causal Inference


Weiber · Mühlhaus
Strukturgleichungs-modellierung
Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS
Springer


James B. Grace
Structural Equation Modeling and Natural Systems



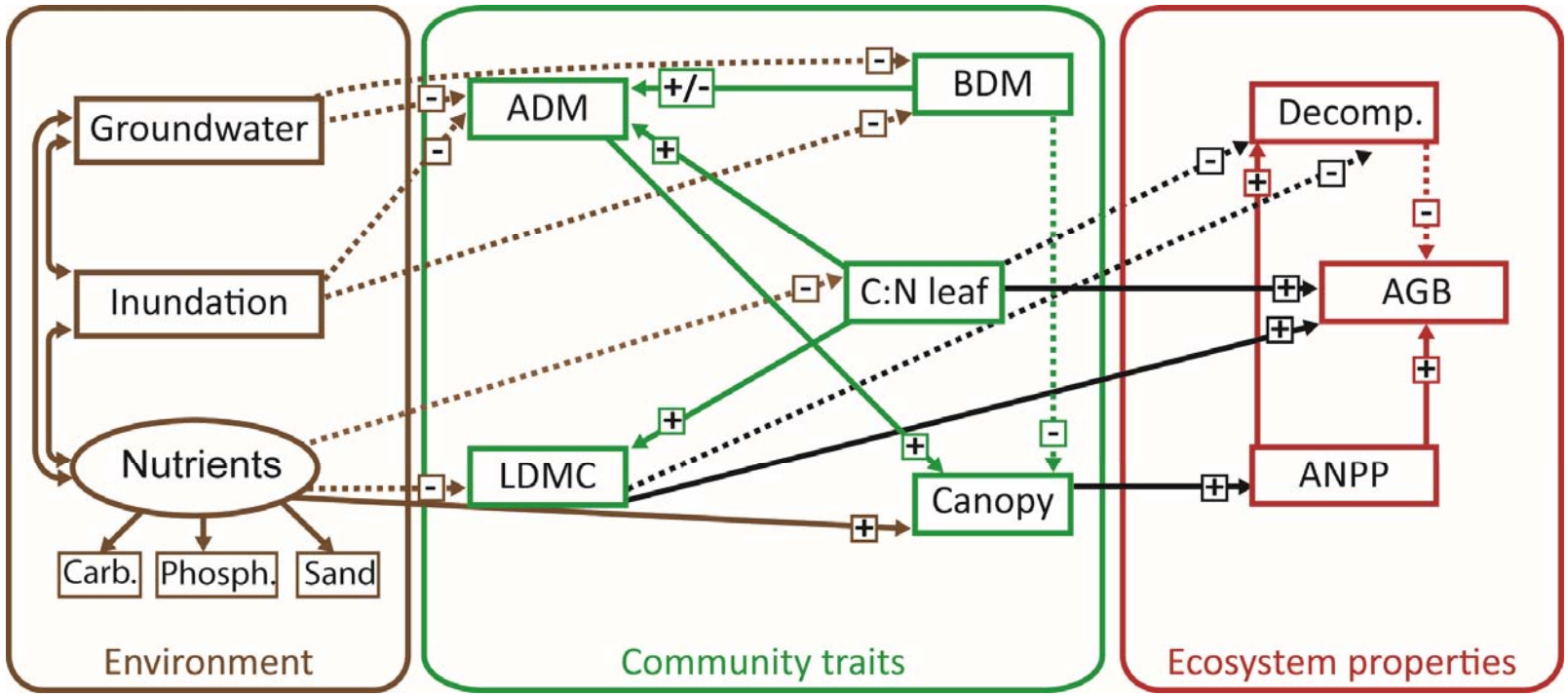
Path analysis



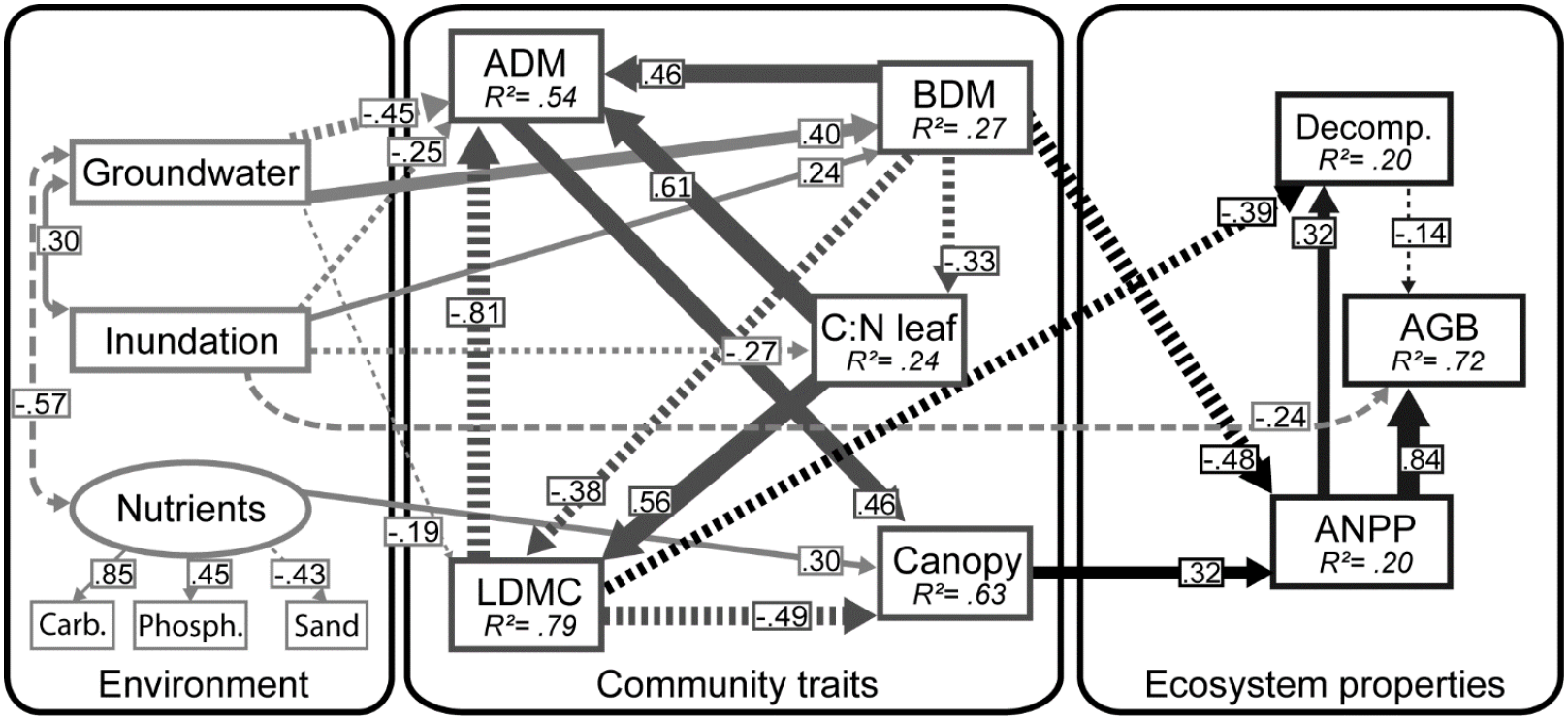
Structural Equation Modelling

(from Grace (2010) Ecol. Monogr. 80: 67-87)

Grace: „Structural equation modeling typically begins with an initial model or selection of competing models which are formulated based on a-priori information“



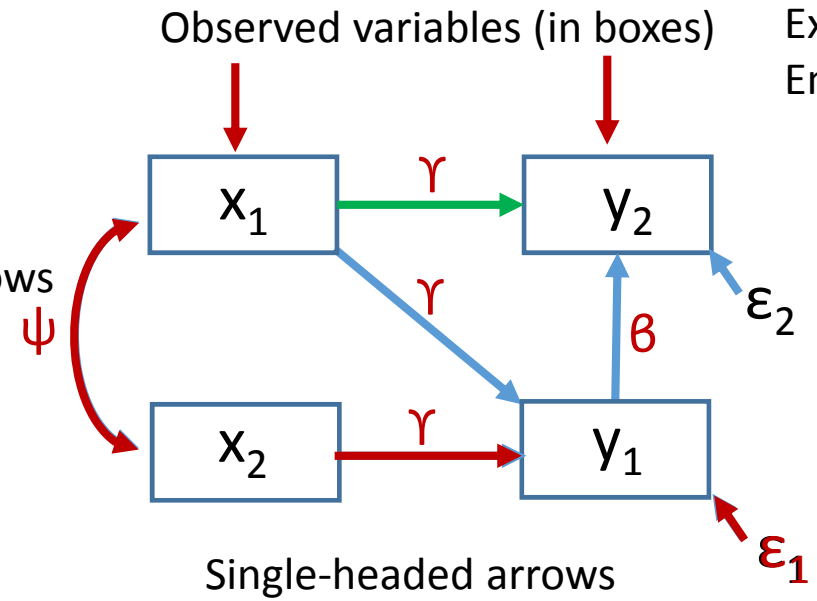
Minden & Kleyer (in press) JVS



Minden & Kleyer (in press) JVS

Grace: „…it is not the statistical result per se that demonstrate causation. Rather, the case for making a causal interpretation depends primarily on prior experience and substantive knowledge“

The anatomy of observed variable models (path models)



Exogenous variable: independent variable (x)
Endogenous variable: dependent variable (y)

γ: path coefficient between exogenous and endogenous variables

β: path coefficient between two endogenous variables

Double-headed curved arrows (undirected paths, unresolved correlations)

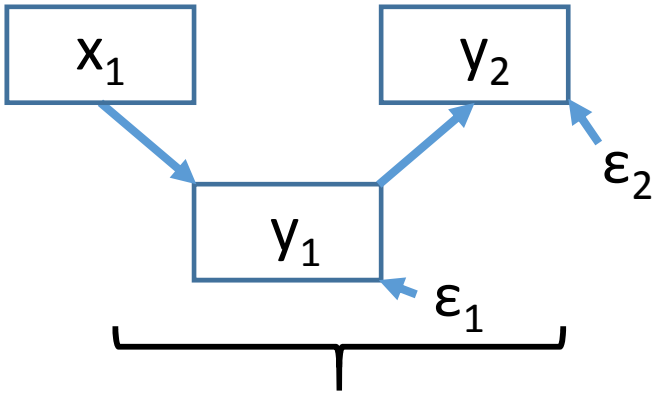
Single-headed arrows (directed relationships)

Residual variances (not in boxes) =error variable (represents all unmodelled causes of the variable it points in)

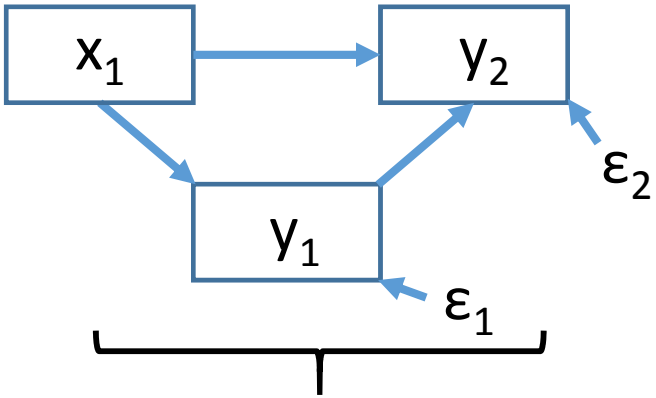
Connection of variables: x_1 and y_2 are connected indirectly (through y_1)
(x_1 has an indirect effect on y_2)

Now: indirect and direct path between x_1 and y_2,
direct path: relationship between two variables that cannot be explained through any other relationships

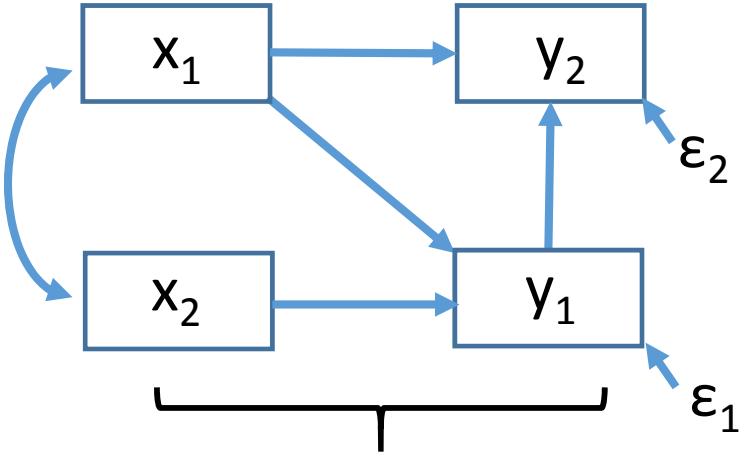
The anatomy of observed variable models (path models)



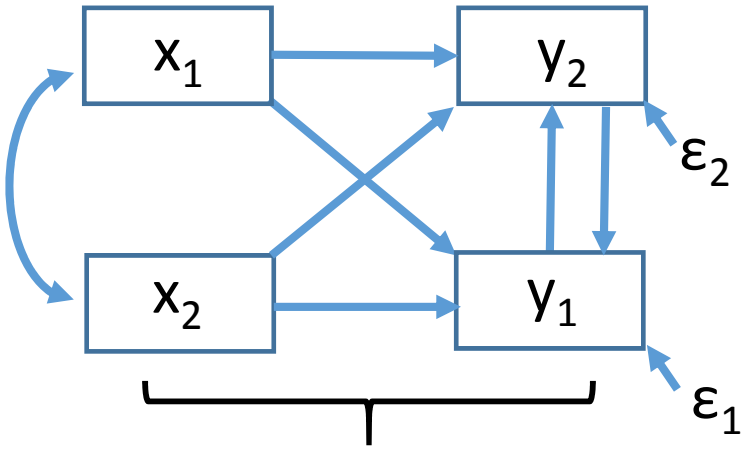
Unsaturated model



Saturated model
(all possible interconnections are specified)

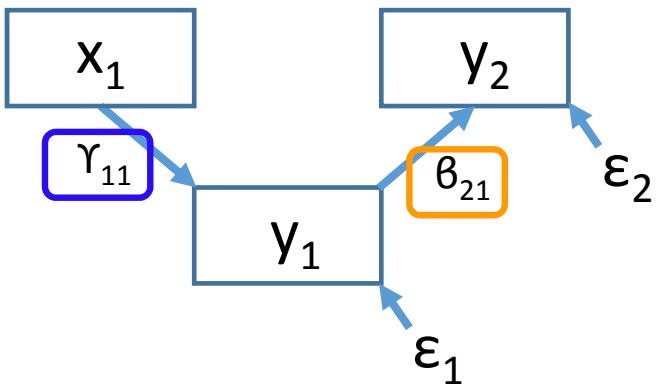


Unsaturated, recursive model,
no reciprocal interactions



Saturated, non-recursive model, i.e.
with reciprocal interactions (between y_1 and y_2)

Correlations are presumed to be caused by some common influence which was not measured



Correlation-matrix			
	x_1	y_1	y_2
x_1	1		
y_1	0.5	1	
y_2	0.3	0.6	1

2 direct paths: x_1 to y_1 and y_1 to y_2
1 indirect path: x_1 to y_2

→ Three path-coefficients

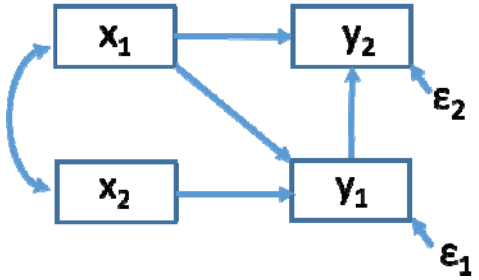
For direct paths: regression coefficient (y = ax + b);
if coefficients are standardized that equals correlation coefficient

For indirect paths (x1 → y2): product of direct paths
= 0.5 * 0.6 = 0.3 (here: correlation coefficient)

Υ: effects of exogenous variables on endogenous variables
β: effects of endogenous variables on other endogenous variables

Υ_{11}: Effect of x_1 on y_1
β_{21}: Effect on y_2 of y_1

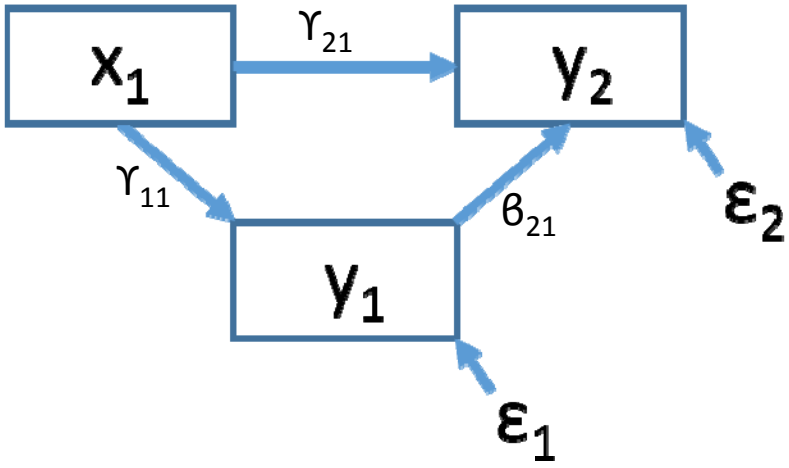
1. Rule of path coefficients: The path coefficient for an unanalyzed relationship (i.e. correlation) between exogenous variables is simply the bivariate correlation (standardized form) or covariance (unstandardized) between the two variables



2. Rule of path coefficients: when two variables are connected by a single causal path, the coefficient for a directed path connecting the variables is the (standardized or unstandardized) regression coefficient

3. Rule of path coefficients: the mathematical product of path coefficients along a compound path (one that includes multiple links) yields the strength of that compound (multi-linkage) path

4. Rule of path coefficients: when two variables are connected by more than one causal pathway, the calculation of partial regression coefficient becomes involved

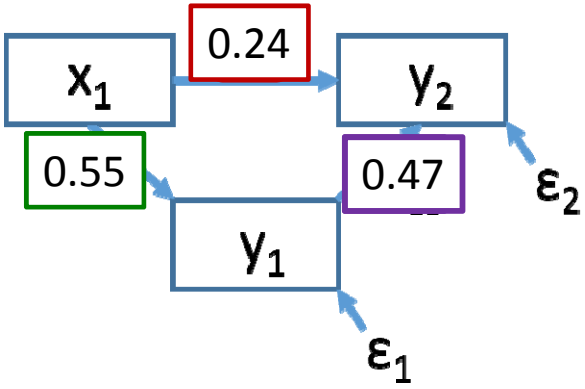


Indirect pathway between x_1 and y_2 is 0.3,
but correlation coefficient is not (it is 0.5)

Correlation-matrix			
	x_1	y_1	y_2
x_1	1		
y_1	0.55	1	
y_2	0.5	0.6	1

Example path between x_1 and y_2:
One causal pathway between x_1 and y_2 (through y_1)
One causal pathway between x_1 and y_2 (direct path)

Partial regression or partial correlation between two variables is one that accounts for the influences of additional variables that affect or correlate with those variables



Partial regression

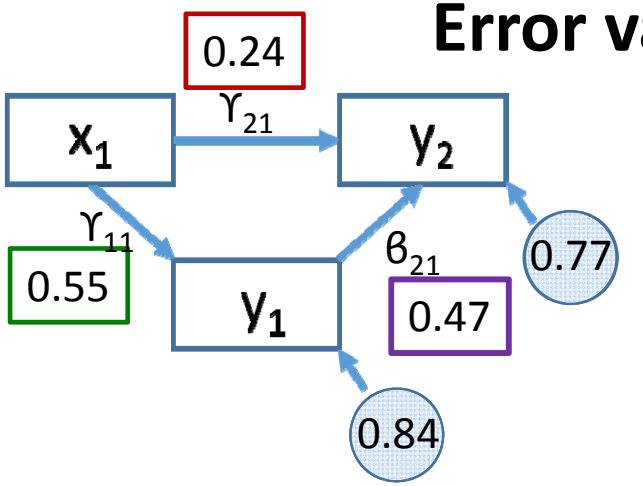
Correlation-matrix			
	x_1	y_1	y_2
x_1	1		
y_1	0.55	1	
y_2	0.5	0.6	1

$$\Upsilon_{21} = \frac{r_{x1y2} - (r_{x1y1} * r_{y1y2})}{1 - r^2_{x1y1}} = \frac{0.5 - (0.55 * 0.6)}{1 - 0.55^2} = \frac{0.17}{0.6975} = 0.24$$

y_1 and y_2 are both connected through a common cause (effect of x_1), they are causally connected through a second path that spans between them through x_1

$$\beta_{21} = \frac{r_{y1y2} - (r_{x1y1} * r_{x1y2})}{1 - r^2_{x1y1}} = \frac{0.6 - (0.55 * 0.5)}{1 - 0.55^2} = \frac{0.325}{0.6975} = 0.47$$

Error variables and their path coefficients



There is no error associated with the exogenous (predictor) variables

Coefficients representing the errors are regression coefficients representing the unexplained (residual) variance

Alternative:
- present standardized values of the error itself:

$\varepsilon_1 = 1\text{-}R^2_{y1} = 1\text{-}0.55^2 = 0.70$

- present the R^2 itself ($R^2_{y1} = 0.55$)

$\varepsilon_1 = \sqrt{1\text{-}R^2_{y1}} = \sqrt{1\text{-}0.55^2} = 0.84$

$\varepsilon_2 = \sqrt{1\text{-}R^2_{y2}} = \sqrt{1\text{-}0.40^2} = 0.77$

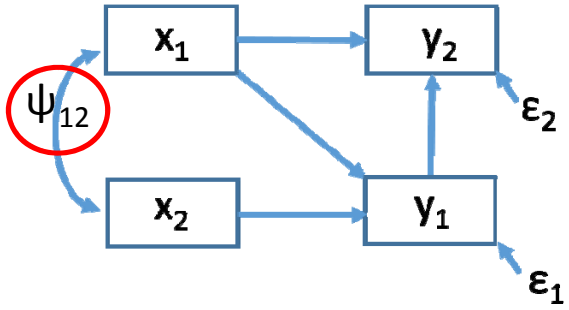
$R^2_{y2} = \gamma_{21}(r_{x1y2}) + \beta_{21}(r_{y1y2})$

$= 0.24 * 0.5 + 0.47 * 0.6 = 0.4$

Correlation-matrix			
	x_1	y_1	y_2
x_1	1		
y_1	0.55	1	
y_2	0.5	0.6	1

5. Rule of path coefficients: Coefficients associated with paths from error variables are correlations or covariances relating the effects of error variables

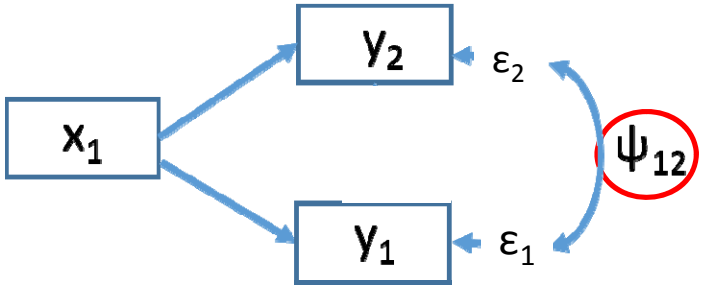
1. ***Rule of path coefficients:*** The path coefficient for an unanalyzed relationship (i.e. correlation) between *exogenous* variables is simply the *bivariate correlation* (standardized form) or covariance (unstandardized) between the two variables



Correlation-matrix				
	x_1	x_2	Y_1	y_2
x_1	1			
x_2	0.80	1		
y_1	0.55	0.40	1	
y_2	0.30	0.23	0.35	1

Partial correlation

6. Rule of path coefficients: Unanalyzed correlations between *endogenous* variables represent *partial correlations* or partial covariances

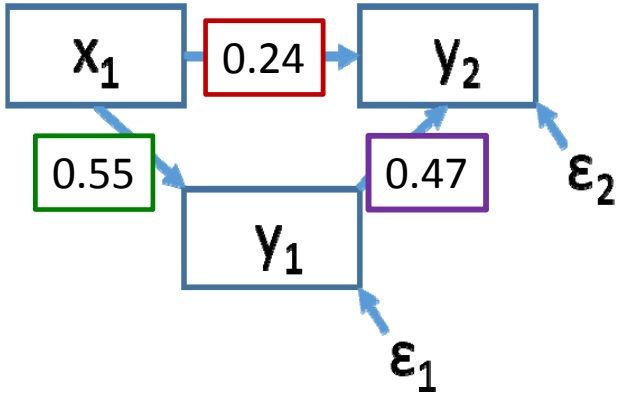


$$\Psi_{12} = \frac{r_{y1y2} - (r_{x1y1} * r_{x1y2})}{\sqrt{(1 - r^2_{x1y1})\,(1 - r^2_{x1y2})}}$$

$$\Psi_{12} = \frac{0.6 - (0.55 * 0.5)}{\sqrt{(1 - 0.55^2)\,(1 - 0.5^2)}} = 0.45$$

Correlation-matrix			
	x_1	y_1	y_2
x_1	1		
y_1	0.55	1	
y_2	0.50	0.6	1

Direct, indirect, total effects



2. Rule of path coefficients: when two variables are connected by a single causal path, the coefficient for a directed path connecting the variables is the (standardized or unstandardized) regression coefficient

DIRECT EFFECT of x_1 → y_1: 0.55

3. Rule of path coefficients: the mathematical product of path coefficients along a compound path (one that includes multiple links) yields the strength of that compound (multi-linkage) path

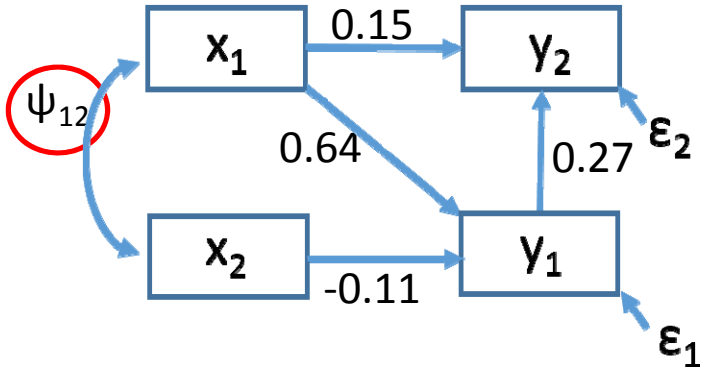
INDIRECT EFFECT of x_1 → y_2: 0.55 * 0.47 = 0.26

7. Rule of path coefficients: The total effect one variable has on another equals the sum of its direct and indirect effects

TOTAL EFFECT of x_1 on y_2 : 0.24 + 0.26 = 0.5

Total correlation

8. Rule of path coefficients: The sum of all pathways connecting two variables, including both causal and non-causal paths, adds up to the value of the bivariate or total correlation between those two variables



All directed paths involve partial regression coefficients,
only coefficient for undirected path (x_1 and x_2) equals correlation coefficient

Correlation-matrix				
	x_1	x_2	y_1	y_2
x_1	1			
x_2	0.80	1		
y_1	0.55	0.40	1	
y_2	0.30	0.23	0.35	1

Total effect of x_1 on y_1 is 0.64; meaning that if x_1 increases by 1 standard deviation (while holding x_2 constant), y_1 would increase by 0.64 its standard deviation

Total effect of x_1 on y_2 is 0.15 + (0.64*0.27) = 0.32; meaning that if x_1 increases by 1 standard deviation (while holding x_2 constant), y_2 would increase by 0.32 its standard deviation, however, y_1 would covary in this process

Total correlation: sum of total effects and undirected relations between variables

Total correlation between x_1 and y_1 = 0.64 + (0.8 * -0.11) = 0.55 (correlation between x_1 and y_1)

Total correlation between x_2 and y_2 = (-0.11 * 0.27) + (0.8 * 0.15) + (0.8 * 0.64 * 0.27) = 0.23

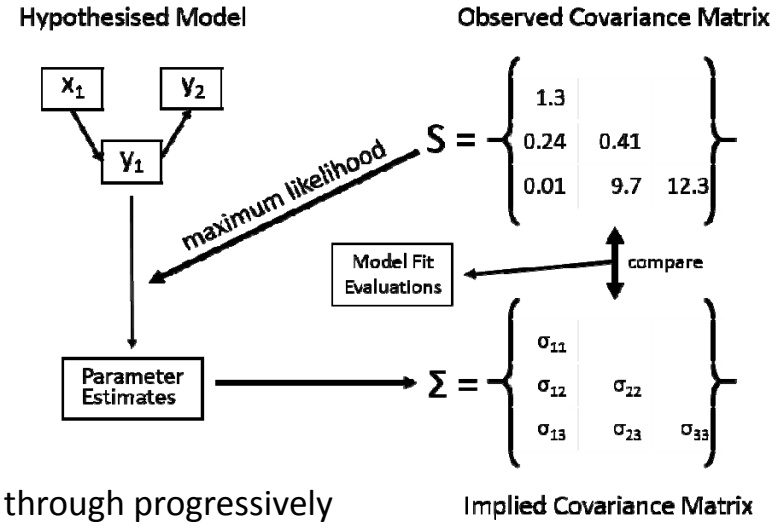
Testing path models

Path diagrams/structural equation models contain three types of variables:

- Manifest variables→ directly observed and measured
- Latent variables → hypothesized to play causal role, but not directly observed or measured
- Error variables (ε) → represents all other unmodelled causes of the variable into which it points
  generally normally distributed and has mean of 0 and variance of 1

Exogenous variable: without ‚causal parent'
Endogenous variable: is caused by some other variable in the model

1. Compute an observed covariance matrix (S) from the variables at hand
2. Compare this to an implied covariance matrix Σ
3. Generate values for the structural coefficients in Θ
   (room of all possible parameter values) such that S is as close to Σ as possible



Grace: „The process of maximum likelihood estimation is one that iteratively searches through progressively refined estimates of parameter values until a set is found that maximizes the likelihood that differences between S and Σ are only those resulting from normal sampling error"

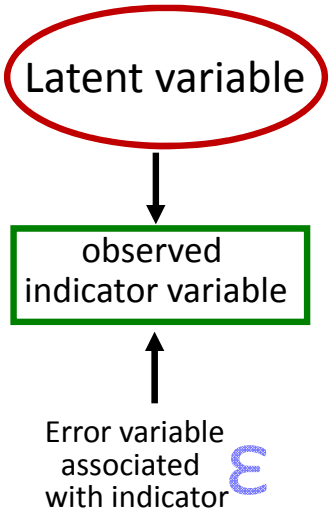
Grace, Fig. 5.4

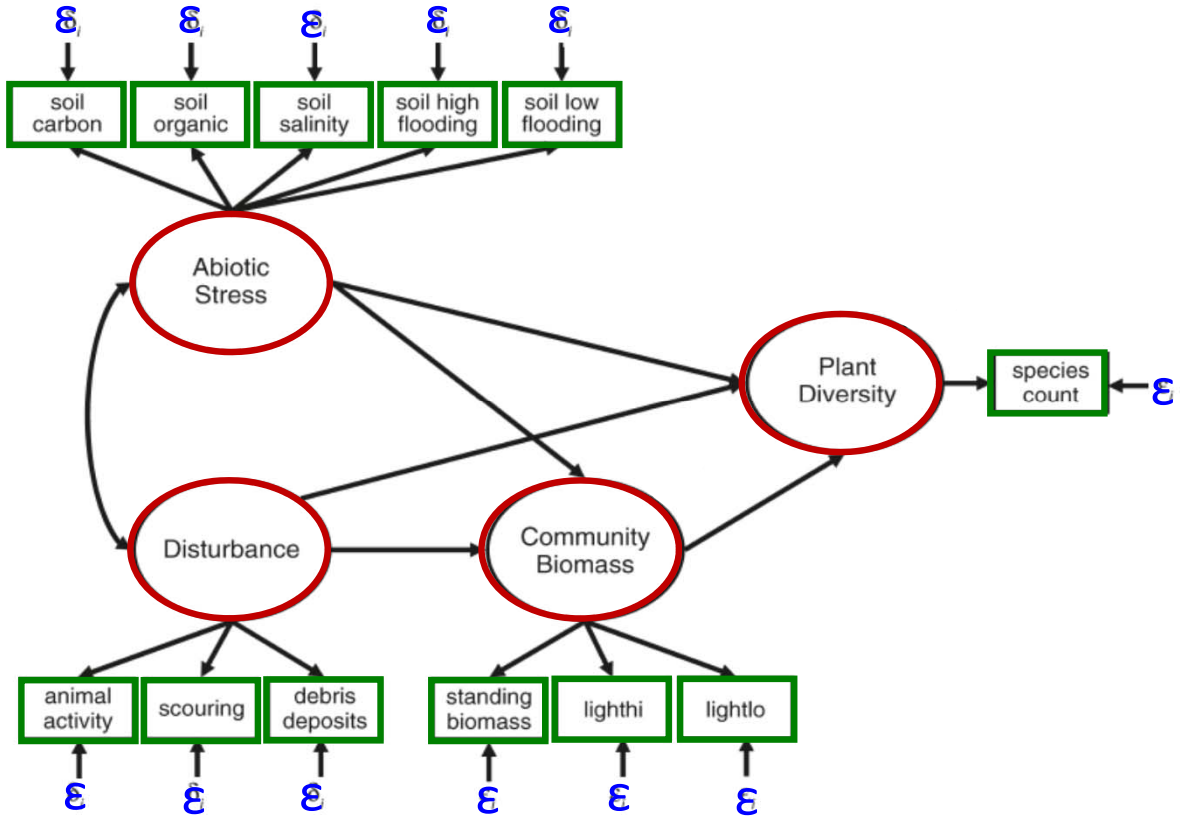
Standardized vs unstandardized path coefficients

Standardized coefficients:
- Use of correlations
- Scale is the same (standard units) for different relationships
- Allows direct comparisons between relationships that are measured on different scales

Unstandardized coefficients:
- Use of covariances
- When one is particularly interested in the differences between the scales
- Comparisons of separate SEMs for different groups (multigroup analysis)
- Repeated measurements at different points in time

Structural equation modelling with latent variables ←→ path analysis with observed (manifest) variables



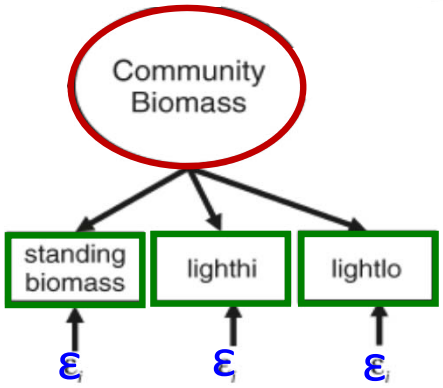


Latent variable:

- „one that is hypothesized to exist, but has not been measured directly“
- „not directly observed but are rather inferred

(through a mathematical model) from other variables that are observed (directly measured)”

Latent variables



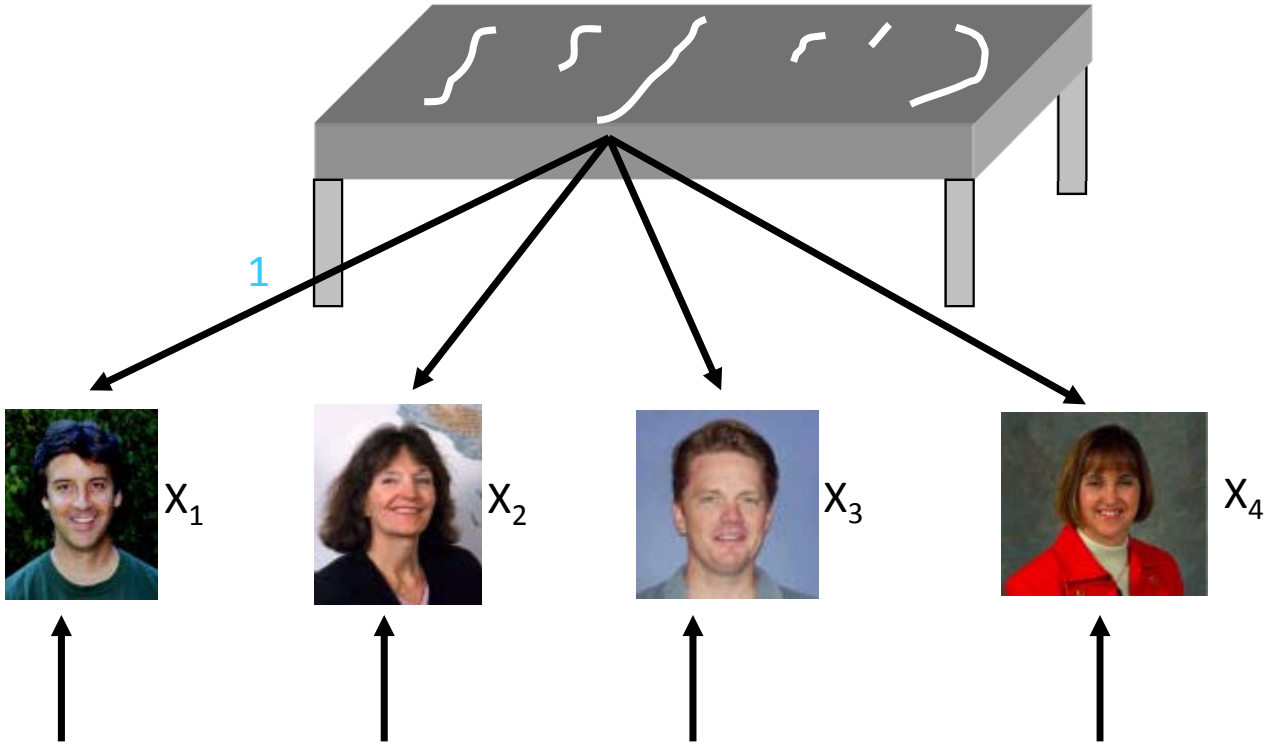
These variable has not actually been measured

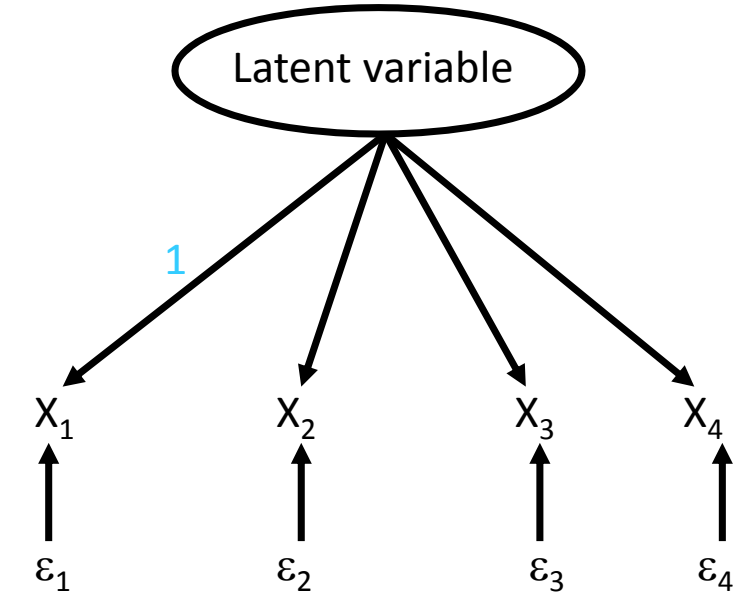
These variables have been measured

Latent variables / observed indicators: tightly correlated but not really the same

It is recommended to use at least three or more observed variables to construct one latent variable, or if two observed variables are used, their correlation should be >0.7 (Grace 2006)

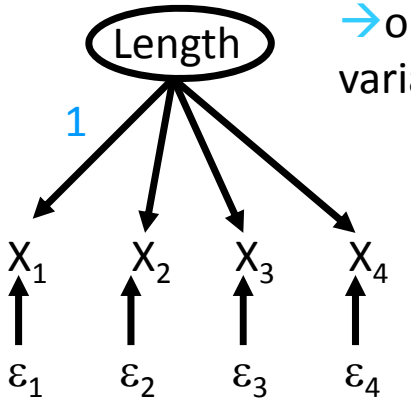
True length of strings (latent)





Which unit does the latent variable have??

Because the observed variables were measured on different scales, one of the path coefficients needs to be fixed
The unit of the scale is transferred from the observed to the latent variable (here cm)



→one unit increase in the latent variable is caused by one unit increase in X_1

$L=N(0,\sigma)$ $\varepsilon_1=N(0,\sigma)$ $\varepsilon_2=N(0,\sigma)$ $\varepsilon_3=N(0,\sigma)$ $\varepsilon_4=N(0,\sigma)$

$X_1=1L+\varepsilon_1$

$X_2=a_2L+\varepsilon_2$

$X_3=a_3L+\varepsilon_3$

$X_4=a_4L+\varepsilon_4$

$Cov(\varepsilon_1,\varepsilon_2)=Cov(\varepsilon_1,\varepsilon_3)=Cov(\varepsilon_1,\varepsilon_4)=Cov(\varepsilon_2,\varepsilon_3)=Cov(\varepsilon_2,\varepsilon_4)=Cov(\varepsilon_3,\varepsilon_4)=0$

Path coefficient that needs to be estimated

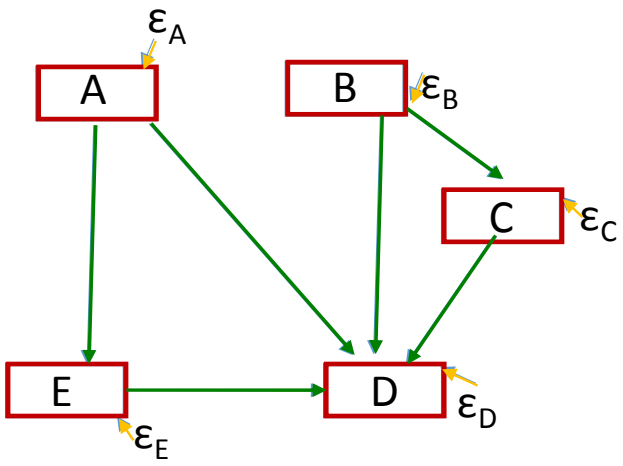
Sometimes it makes more sense to fix the latent variable, if e.g. the observed variables were measured on ‚strange' scales (IQ-Test)

From Shipley's workshop

Degrees of freedom

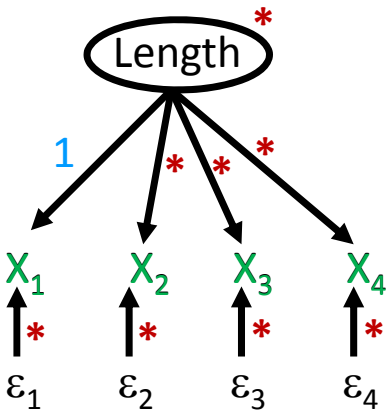
$\frac{V(V+1)}{2}$ - Parameters that have to be estimated = df

V: no. of Variables



$\frac{5(5+1)}{2} - 6 - 5 = 4$

Degrees of freedom = V(V+1)/2 - Parameters that have to be estimated



L=N(0,σ) $X_1=1L + \varepsilon_1$

ε_1=N(0, σ) $X_2=a_2L + \varepsilon_2$

ε_2=N(0, σ)

ε_3=N(0, σ) $X_3=a_3L + \varepsilon_3$

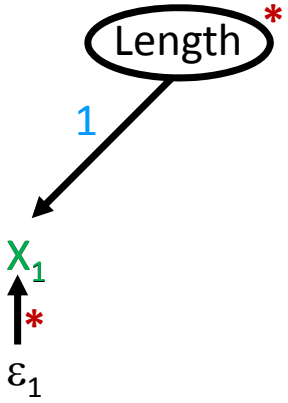
ε_4=N(0, σ)

$X_4=a_4L + \varepsilon_4$

$$df = \frac{4(4+1)}{2} - 1 - 3 - 4$$

= 10-8 = 2

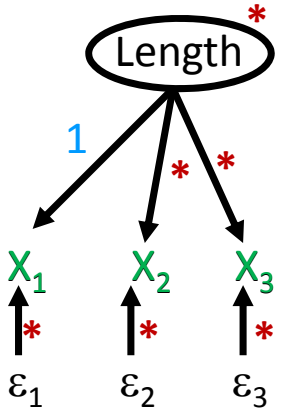
Model overidentified



$$\frac{1(1+1)}{2} - 1 - 1 = 1 - 2 = -1$$

Model underidentified

Not enough information to achieve unique solutions,
like a + b = 8,
possible solutions: 1+7, 2+6, 3+5 etc.
but no *unique* solution

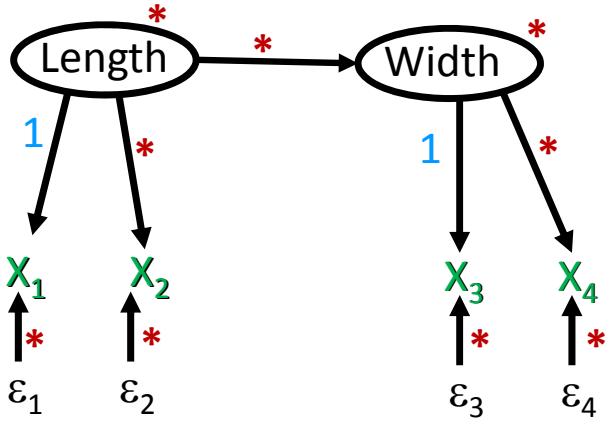


$$\frac{3(3+1)}{2} - 1 - 2 - 3 = 6 - 6 = 0$$

Model just-identified

No. of free parameters exactly equals the number of known values

Modell is fitted to the data
→ Not good



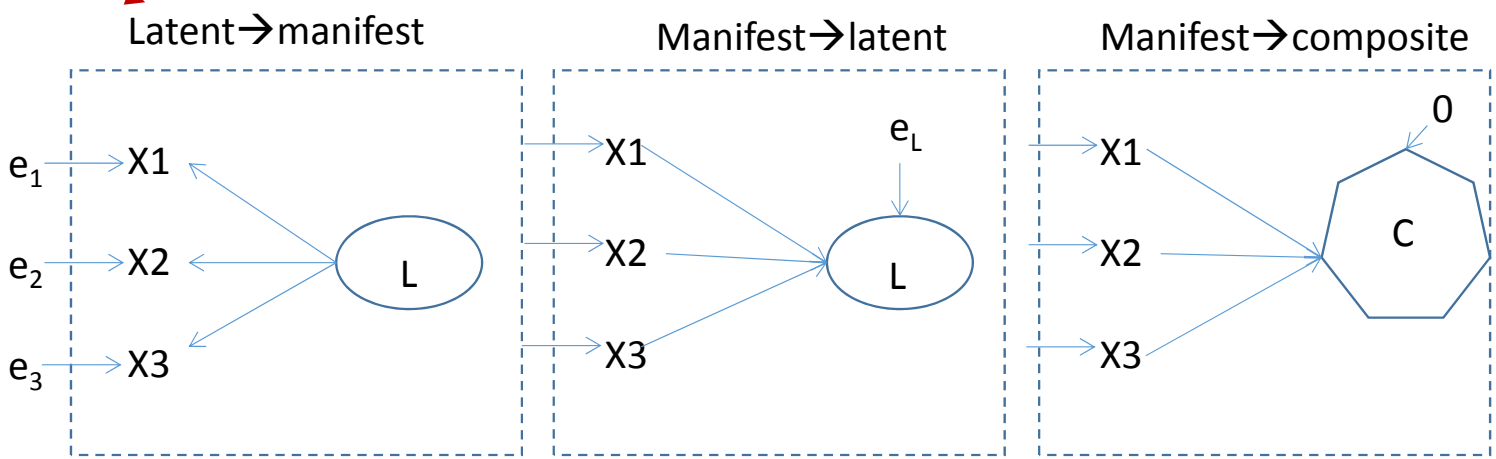
$$\frac{4(4+1)}{2} - 2 - 1 - 2 - 4 = 10 - 9 = 1$$

Model overidentified

All parameters are identified, more knowns than unknows

translates into:
latent *causes* manifest

Three basic types of latent/composite blocks



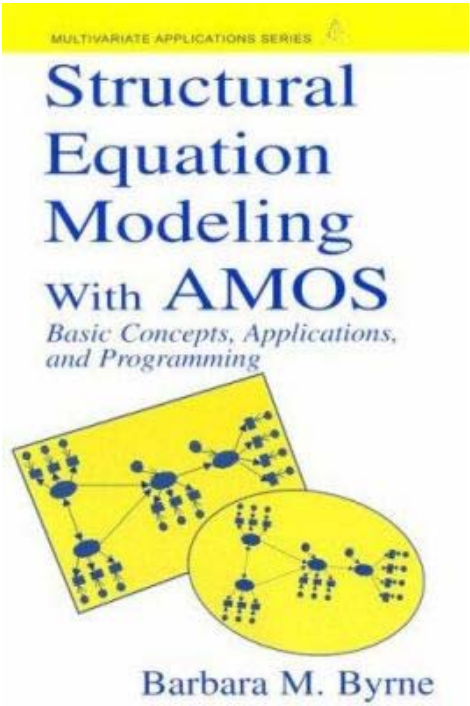
The observed variables are *necessarily* correlated because the are all caused by a single unmeasured variable.
The latent may have an error variance, depending on the rest of the model.

The observed variables are not *necessarily* correlated (but they can be) because they jointly cause the single unmeasured variable.
The latent will always have an error variance because it is not completely caused by the observed Variables.

The observed variables are not *necessarily* correlated (but they can be) because they jointly cause the single unmeasured variable.
The composite variable will always have a **zero** error variance because it is completely caused by the observed variables; the composite variable is the direct product of some set of causes.

From Shipley's workshop, see also Grace (2010) Ecol. Monogr.

Measurements of model fitness



CFI: Comparative Fit Index

ranges from 1 to 0
compares the hypothezised model with an independent model
the closer to 1 the better (CFI of 0.95 → acceptable model)

RMSEA: Root mean error of approximation

„How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?“ (Browne & Cudeck, 1993)

should be <0.05, value of 0 means the model fits perfectly

In text: χ^2, df, p-value, Fit-Indices

Rules of thump:

For latent variables: It is best to have observed variables close to each other in terms of differences in the units
(use km for all instead of m and km)

The more observed variables, the better the latent variables are explained

Negative variances: causal structure of model is wrong
Options: get units closer together (all in meters)
use log-scale for all variables

Rules of thump:

Sample sizes:

Problem: the probabilities calculated using maximum likelihood methods are only ***asymptotically*** unbiased. What does this mean?

The calculated probabilities are only exact as sample size reaches infinity!!!

In practice, the minimum sample size needed for good probability estimates depends on the ratio of observations to free parameters that have to be estimated.

If the data are normally distributed, then 5 times as many observations as free parameters is fine...

As the data become increasingly non-normal, the minimum sample size must increase accordingly.

If the sample size is not large enough, then the estimated probabilities will usually be smaller than they should be - so you reject models more often than you should.

From Shipley's workshop

Rules of thump:

Statistical power:

Null hypothesis in testing a structural equation model:

The data were generated by the structural equations given in my model

- there are no causal relationships in my model that are wrong
- there are no causal relationships that are missing in my model
- the relationships between the variables are exactly linear
- the data are normally distributed or else there are lots of them...

As sample size increases, statistical power increases.
Therefore, **even very small errors in the above assumptions will cause the model to be rejected.**

From Shipley's workshop

- there are no causal relationships that are missing in my model

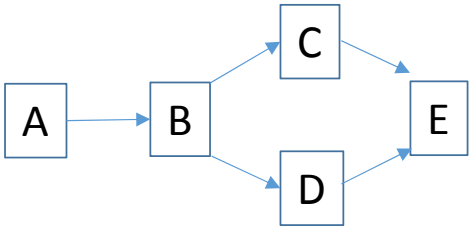
Can be tested in R, for that program your model (library ggm):

```
dag1<-DAG(B~A, C~B, D~B, E~C+D)
```
→ DAG(dependent1~parent1, dependent2~parent2…)

```
dag1
```
dependent

parent	B	A	C	D	E
B	0	0	1	1	0
A	1	0	0	0	0
C	0	0	0	0	1
D	0	0	0	0	1
E	0	0	0	0	0

Translates into: B is dependent on A,
C is dependent on B,
D is dependent on B
E is dependent on C and D



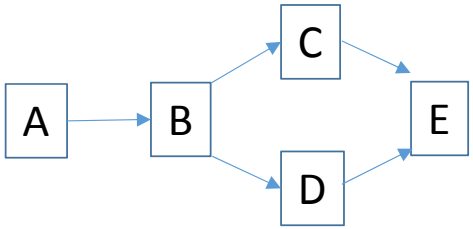

```
shipley.test(dag1, var(my.dat), 500)
$ctest
[1] 13.94868

$df
[1] 10

$pvalue
[1] 0.1753452
```

Computes probability values for each independent relationship, adds them:
$C=-2 \sum \ln P_i$ should not be significant according to χ^2 distribution

- Are the variabels in the model independent from each other…



d-separation

Two variables are d-separated, they are independent
Two variabels are not d-separated, they are not independent

For model-development, it is important to know which variables depend on each other or which do not…

1) List all variables that don‘t have a direct path between them
2) Write down *causal children* and *causal parents*

Basis set		Conditioning set		
A	C	Ø	B	A_||_C|{B}
A	E	Ø	C,D	A_||_E|{C, D}
A	D	Ø	B	A_||_D|{B}
B	E	A	C, D	B_||_E|{A, C, D}
C	D	B	B	C_||_D|{B}

Parent of this

Parent/s of this

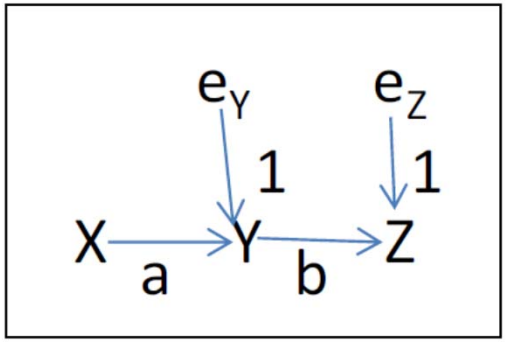
Translates into: A and C are d-separated given B,
therefore A and C will be independent conditional on A

```
dag1<-DAG(B~A, C~B, D~B, E~C+D)
dSep(dag1, "A", "C", "B")
[1] TRUE
```

Translates into: A_||_C|{B}

For more on d-separation see Shiply‘s book

AMOS



If we translate this into structural equations we get:

$$X = N(0, \sigma_X); e_Y = N(0, \sigma_{e_Y}); e_Z = N(0, \sigma_{e_Z})$$

$$Y = aX + e_Y$$

$$Z = bY + e_Z = b(aX + e_Y) + e_Z = abX + be_Y + e_Z$$

Definition of a populational (co)variance:

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}; Cov(x,x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N} = Var(x) = \sigma_x^2$$

However, in SEM we work with centered variables; i.e. the means are always zero, so:

$$Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i)(y_i)}{N}; Cov(x,x) = \frac{\sum_{i=1}^{N}(x_i)(x_i)}{N} = Var(x) = \sigma_x^2$$

Since we are working with populational values (the reason why we are dividing by N rather than N-1), we will use the mathematical notation for "expectation" (i.e. the mean in the population):

$$Cov(x,y) = E[xy] = \sigma_{xy}; Cov(x,x) = E[x^2] = Var(x) = \sigma_x^2$$

Our goal: we want to express the model covariance matrix of X, Y, Z in terms of the parameters of the structural equations (i.e. variances of the exogenous variables and path coefficients):

	X	Y	Z
X	s^2_X	s_{XY}	s_{XZ}
Y	s_{XY}	s^2_Y	s_{YZ}
Z	s_{XZ}	s_{YZ}	s^2_Z

This is how we express the covariance between X and Y in terms of the reduced form parameters:

$$Cov(X,Y) = E[XY] = E[X(aX + e_Y)] = E[aX^2] + E[Xe_Y].$$

Now, X and e_Y are independent (remember d-separation) and so their covariance must be zero given this model, so:

$$Cov(X,Y) = E[aX^2] = aE[X^2] = a\sigma_X^2$$

For example

$$Cov(X,Y) = E\left[aX^2\right] = aE\left[X^2\right] = a\sigma_X^2$$

If we do the same thing for the covariance between X and Z then we get:

$$Cov(X,Z) = E\left[X(abX + be_Y + e_Z)\right] = E\left[abX^2\right] + E\left[bXe_Y\right] + E\left[Xe_Z\right] = \sigma_{XZ}$$

$$\sigma_{XZ} = abE\left[X^2\right] + 0 + 0 = ab\sigma_X^2$$

The variance of Y:

$$Cov(Y,Y) = \sigma_Y^2 = E\left[(aX + e_Y)(aX + e_Y)\right] = E\left[a^2X^2\right] + E\left[e_Y^2\right] = a^2\sigma_X^2 + \sigma_{e_Y}^2$$

The covariance between Y and Z:

$$Cov(Y,Z) = \sigma_{YZ} = E\left[(aX + e_Y)(abX + be_Y + e_Z)\right] = E\left[a^2bX^2\right] = a^2b\sigma_X^2$$

The variance of Z:

$$Cov(Z,Z) = \sigma_Z^2 = E\left[(abX + be_Y + e_Z)(abX + be_Y + e_Z)\right]$$

$$\sigma_Z^2 = E\left[a^2b^2X^2\right] + E\left[b^2e_Y^2\right] + E\left[e_Z^2\right] = a^2b^2\sigma_X^2 + b^2\sigma_{e_Y}^2 + \sigma_{e_Y}^2$$

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & a\sigma_X^2 & ab\sigma_X^2 \\ a\sigma_X^2 & a^2\sigma_X^2 + \sigma_{e_Y}^2 & a^2b\sigma_X^2 \\ ab\sigma_X^2 & a^2b\sigma_X^2 & a^2b^2\sigma_X^2 + b^2\sigma_{e_Y}^2 + \sigma_{e_Y}^2 \end{bmatrix}$$

Observed Covariance Matrix

Implied/Population Covariance Matrix