SEM: Simultaneous Equation Models – Introduction and Overview

Harry B.G. Ganzeboom
VU GSSS
December 18 2014

SEM

- SEM can mean:
  - Structural Equation Modelling
  - Simultaneous Equation Modelling
- Brings together:
  - Structural (causal ‘path’) model with multiple dependent (endogenous) variables.
  - Measurement (i.c. common latent factor) model.
- The mathematics of multiple regression analysis applies to both parts.

Advantages

- SEM makes you think about how the world works: causal effects are everywhere.
- SEM makes you aware of the biases that occur through (random and systematic) measurement error and provides tools to diagnose and repair these.
- Also:
  - ML estimation of data with random missings.
  - Constrained estimation: SEM's can fix coefficient to some specific values, make them equal or restrict them to a certain range.

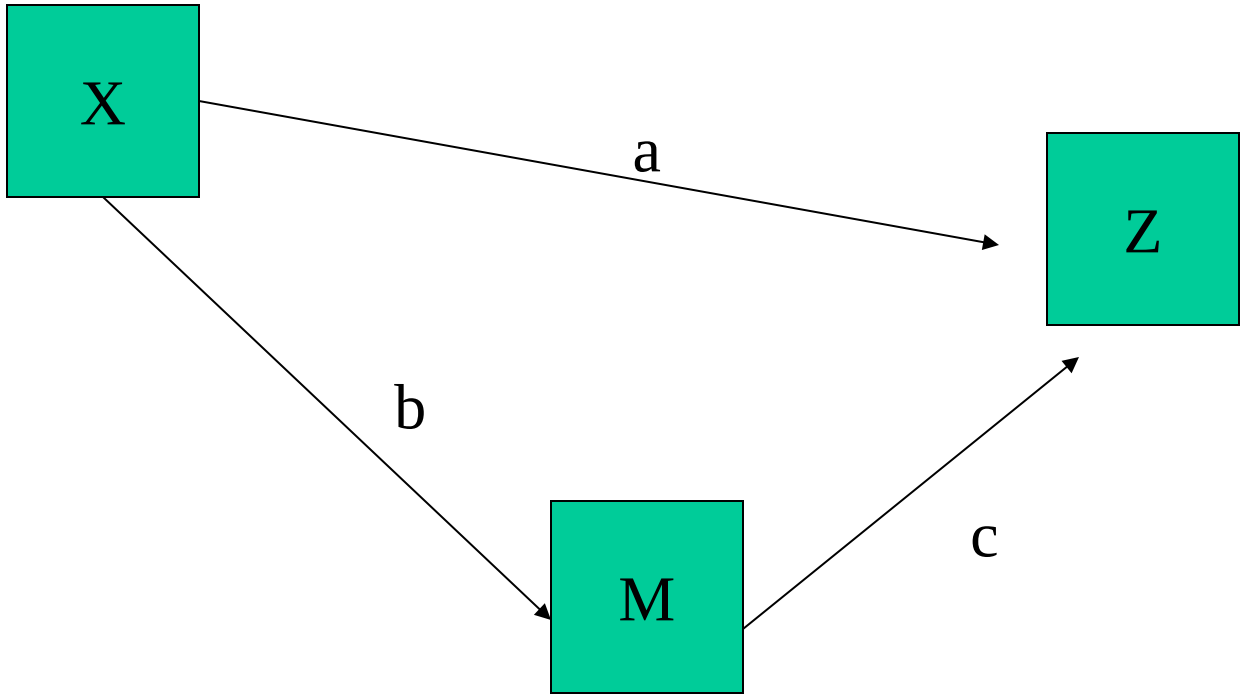
Software

- LISREL (Jöreskog & Sörbom) – SEM models are often referred to as “lisrel-models”, users sometimes as “lisrelites”.
- AMOS
- Mplus
- Stata12
- Stata13
- I use LISREL and Stata12, but my impression is that Mplus is a bit more powerful.

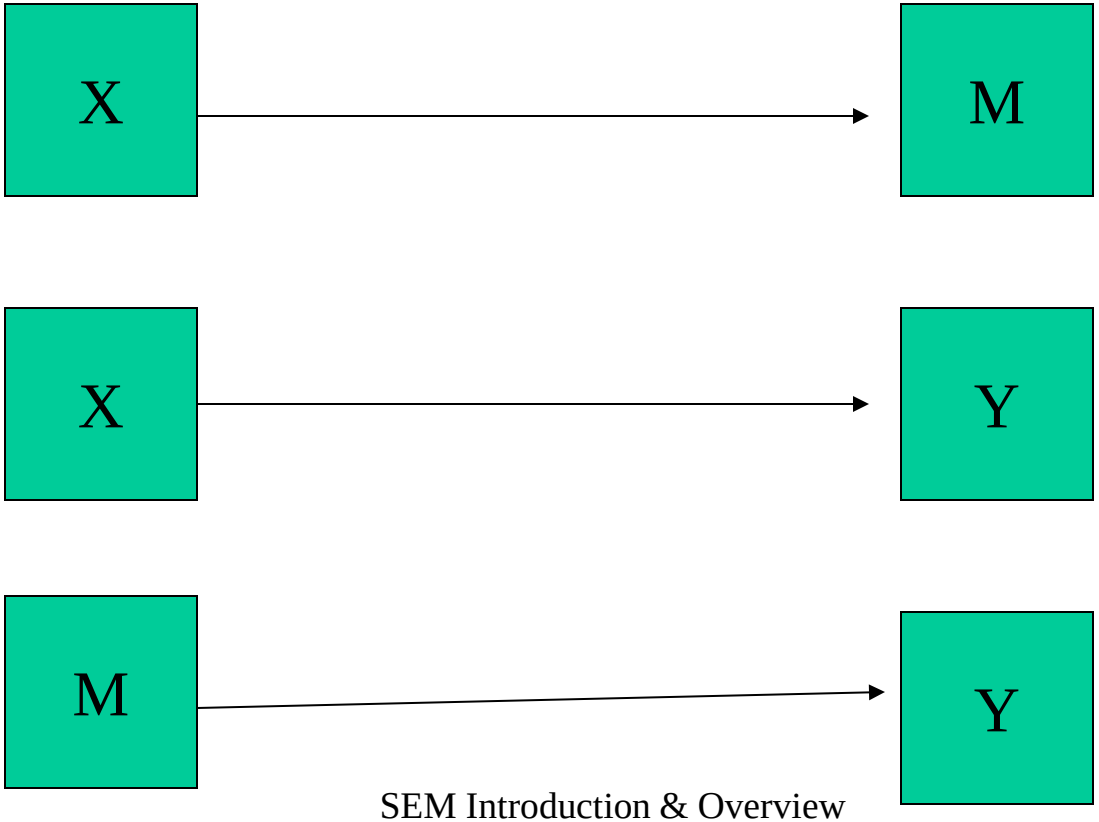
The world as a correlation matrix

- Although SEM's are not restricted to covariances and correlations, the classical models and applications are.
- In an SEM state of mind, one sees the world summarized in a covariance matrix; the researcher's task is to invent a set of (simple) mathematical equations that will reproduce this matrix.
- I find it more convenient to think in terms of correlations (standardized) in stead of covariances.
- The algebra of correlations and causal effects is fairly simple. You do not need regression analysis to estimate path coefficients, the path-analytic decomposition will do!

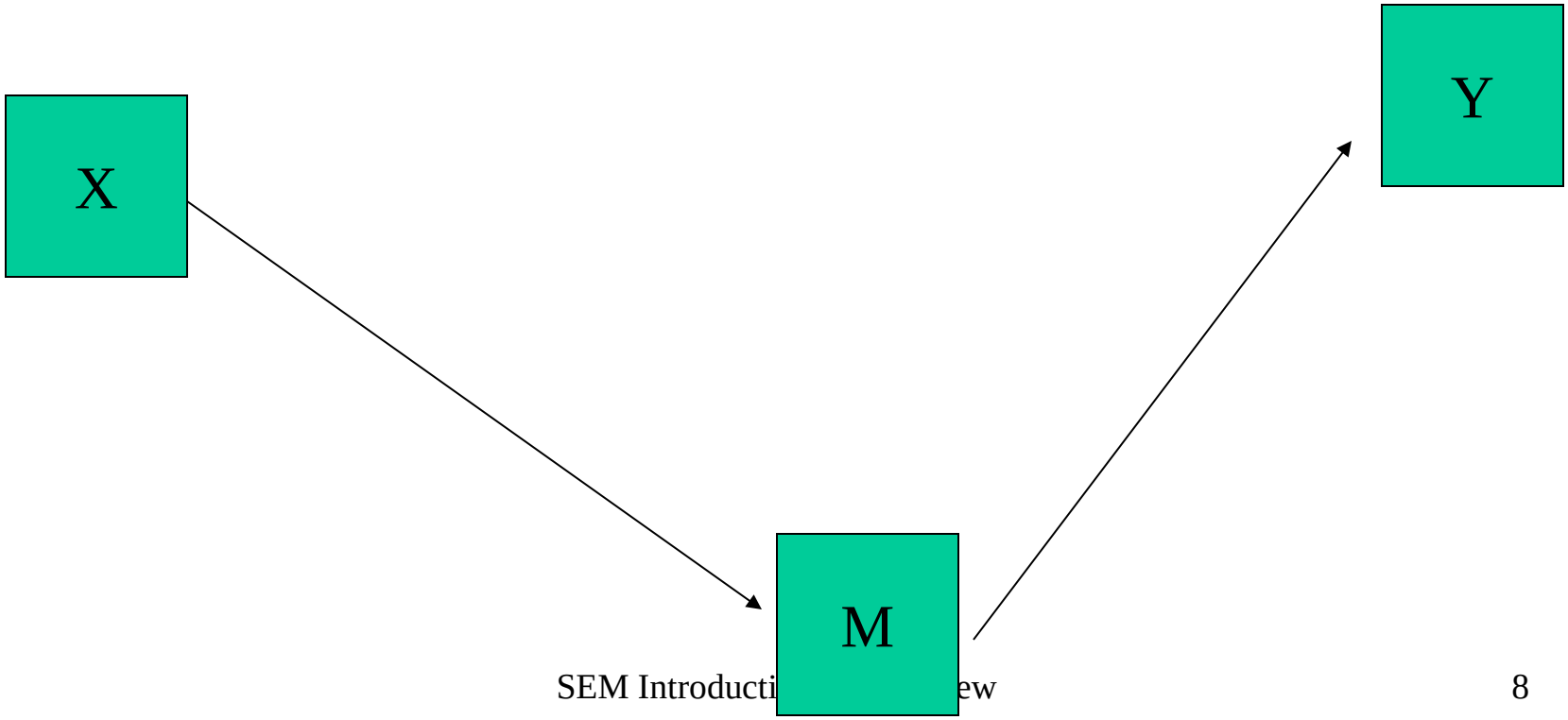
The elementary causal model



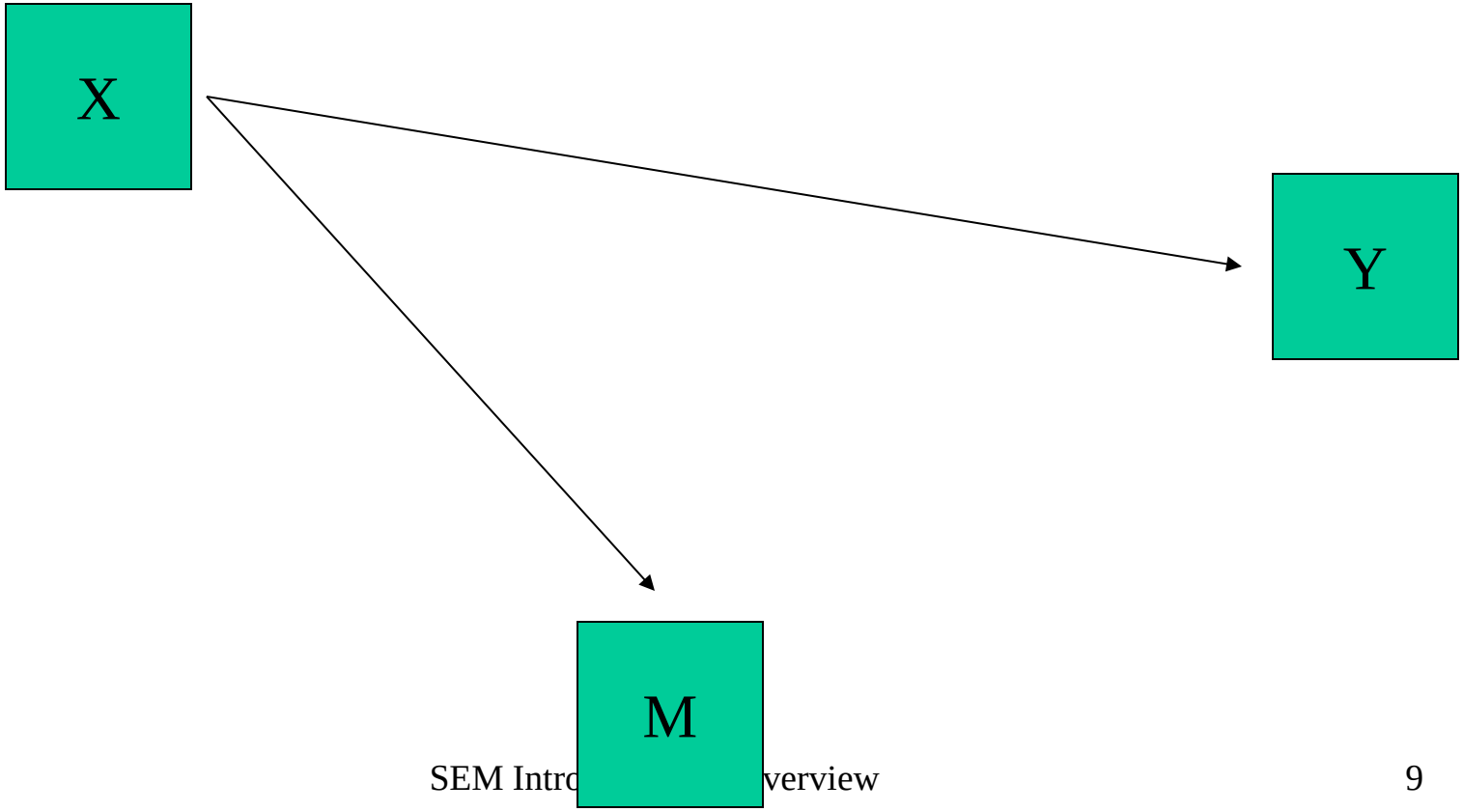
Direct effects



Indirect effect



Confounding effect



The path-analytic theorem

- Total correlation =
  - ***Direct effect + indirect effects + confounding effects.***
- Indirect effects are the multiplication of the two direct effects (in a chain).
- Confounding effects are the multiplication of the two direct effects (in a fork).
- Notice that while the definition and calculation of confounding and indirect effects is fairly similar, their causal interpretation is radically different:
  - Indirect effects inform you how (via which mechanism) X causes Y;
  - Confounding effects inform to what extent the correlation between X and Y is NOT causal (but spurious).

The calculations

- $r_{XM} = b$
- $r_{XY} = a + b*c$
- $r_{MY} = a*b + c$
- Three equations with three unknowns: exactly identified.
- Notice that you can find regression coefficients without doing regression analysis, just by using the correlation matrix!

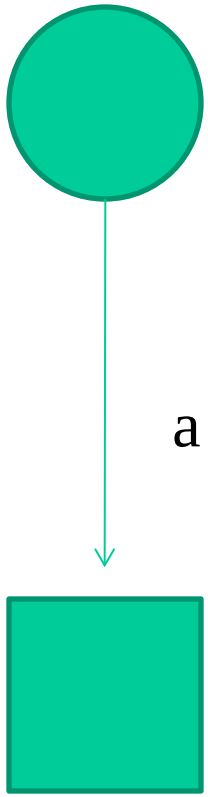
Identification

- In non-recursive models it is generally true that we have as many unknowns as equations, and with some work, can solve for the unknowns.
- In many applications have more equations (correlations) than unknowns (coefficients) → overidentification.
- Overidentified systems as solved by mnimizing the distance between the empirical covariance matrix and the expected matrix.

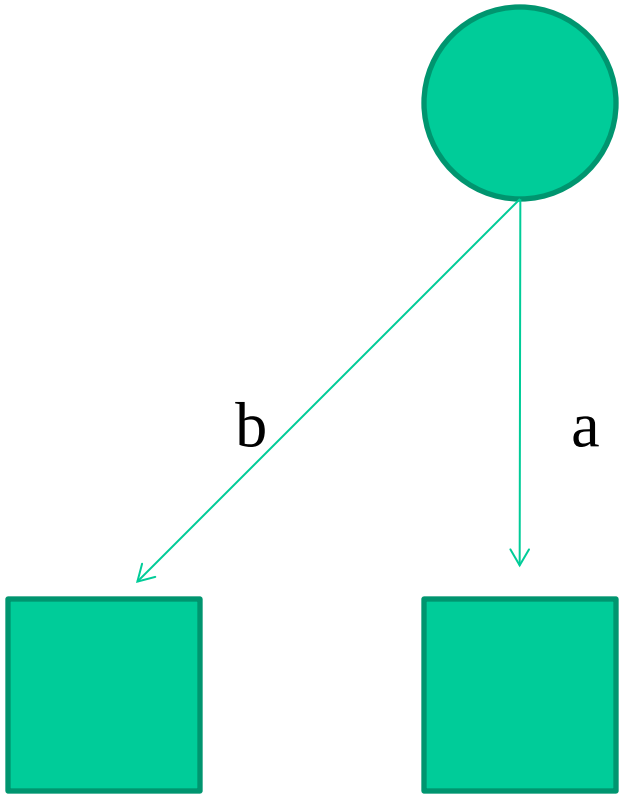
Measurement models

- Measurement can be interpreted as a causal model in which a latent variable causes the response on an observed variable.
- → We see the reality in our observed variables always with some measurement error (Plato).
- We can estimate the measurement error when we repeat the measurement and compare the indicators.
- If we do not have repeat measures, we cannot know the amount of measurement error, but it is still there.
- Measurement error in a model with two measures is not identified as such (but see below), but a latent variable model with three indicators is exactly identified, much in the same way as we can solve for the coefficients in the elementary causal model (three correlations and three unknown coefficients).

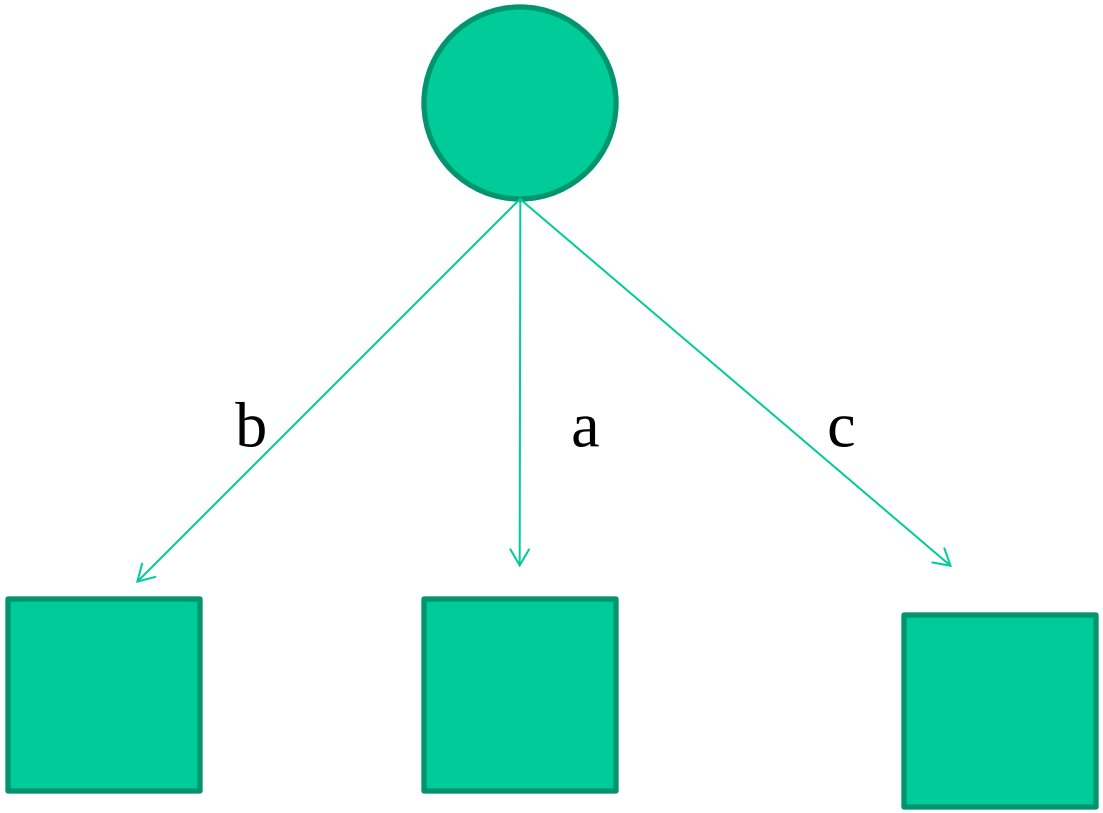
Elementary measurement model



Elementary measurement model



Elementary measurement model



Elementary measurement model

- The measurement model with three indicators is exactly identified: we can find out the quality of each indicator separately (with respect to random measurement error.)
- However, when estimated in a larger model with auxiliary variables – two indicators will often do.

Putting it together

- The elementary causal model and the measurement model are both SEM's, but the real SEM arises when we combine them in a single model.
- Note that if we combine measurement model and causal model, it is (mostly) not necessary anymore to have three indicators for each latent variable: two is enough for identification.

Random measurement error

- Random measurement error (or: unreliability) arises as if by a random process: it is unpredictable when and how much deviation from the true score will arise for each individual.
- Random error makes measures unreliable (or: unstable): it leads to different answers all of the time.
- With a SEM common factor model we can estimate *how much* error occurs, but also diagnose *where* it occurs.

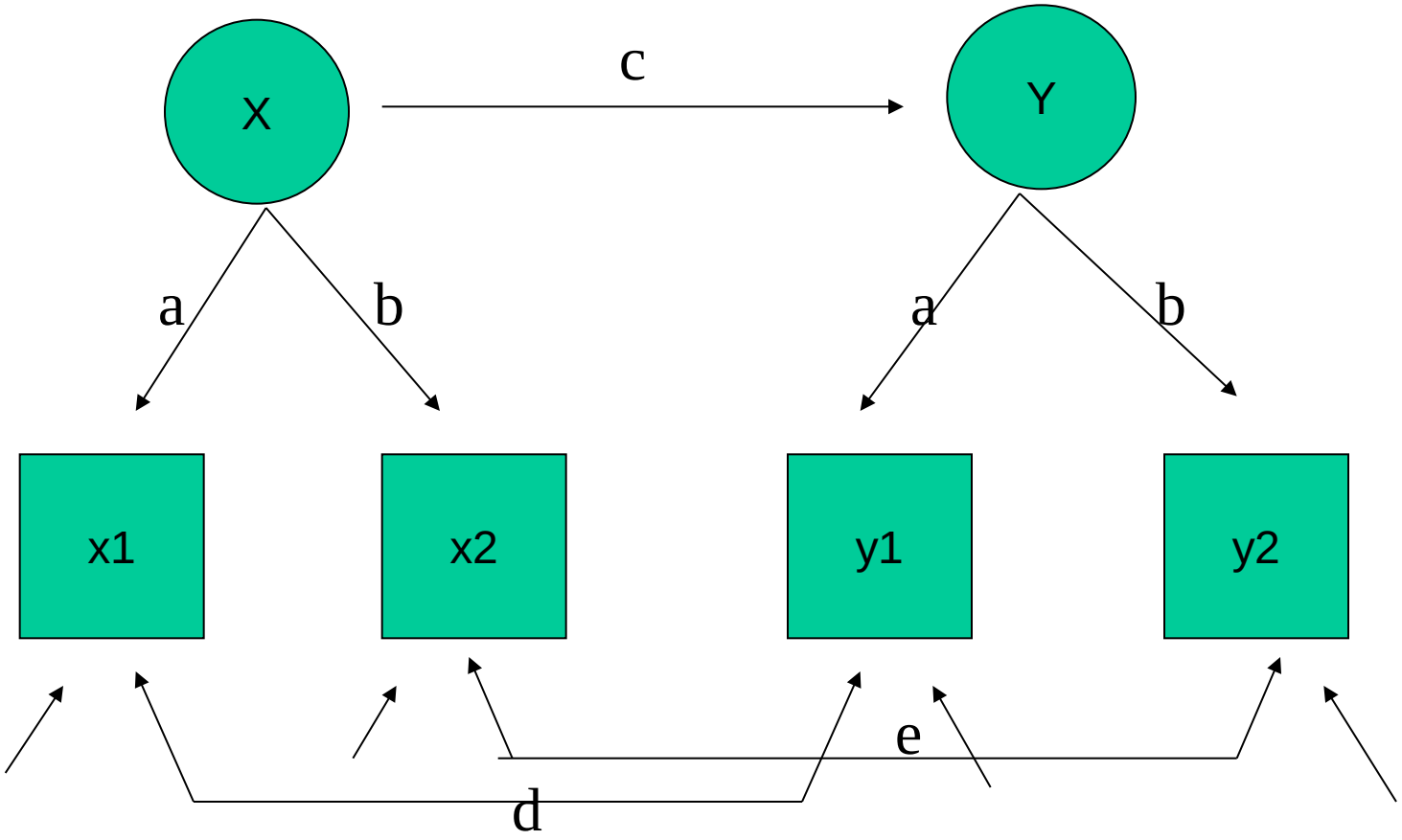
Systematic measurement error

- Some kind of measurement error arise systematically, the deviation from the true score has some consistency (within persons, between measures).
- Systematic measurement error is also known as invalidity or bias.
- We can trace systematic errors by repeating the error:
  - Random error: repeat the measurement
  - Systematic error: repeat the error.

Correlated error

- If we have two measures that have the same (=systematic) error, this arises as correlation between the measures (even if the two measures do not have a true score in common).
- Systematic measurement modeling is just a variety of (multiple) common factor analysis.
- MTMM models: Multiple Traits, Multiple Methods – is a traditional name for separating random error from systematic (‘method’) error.

SEM / MTMM model



SEM and factor analysis

- Another perspective on SEM is that it adds a causal structure to the correlations among the common factors.
- As a matter of fact, it is useful to run any SEM models with multiple indicators as such a common factor model, either exploratory (SPSS) or confirmatory.
- This model informs about the measurement part, without being confounded by problems in the causal part.

How to do it -- outline

- Generate the correlation / covariance matrix to analyze. You can also analyze individual data.
- Write up your causal / measurement model as a diagram.
- Write up your causal / measurement model as a set of linear equations (one for each dependent variable).
- Program the equations in a SEM program.
- Assess how the reproduced correlations fit the observed correlations.
- Adjust (‘modify’) the model by allowing more effects or trim superfluous ones.

Application: multiple indicator measurement in social mobility

- SEM models are most often applied on attitudinal data, not on demographic data such as age, gender, education, occupation.
- However, this is exactly my concern: I analyze comparative patterns of intergenerational occupational mobility (mostly father – son) and fear that my conclusions are biased by measurement error.
- So I want to use SEM to diagnose and correct measurement error.

Multiple indicators for occupations

- It is not easy to ask demographic indicators in a multiple indicator design. It irritates the respondents to ask the same questions again.
- However, one solution is to ask occupation questions both in a crude (precoded) and detailed (open format). With the proper intro, this looks like two different questions, much like attitude items.
- I have first encountered this idea in ISSP 1987. I have applied in my own data-collections for ISSP-NL since 1996.
- By sheer coindence the design has also been applied in the ESS, but only for father's and mother's occupation.

ESS showcard

1. Traditional professionals
2. Modern professionals
3. Clerical and intermediate
4. Senior manager and administrator
5. Technical and craft
6. Semi-routine manual and service
7. Routine manual and service
8. Middle and junior managers

ESS showcard

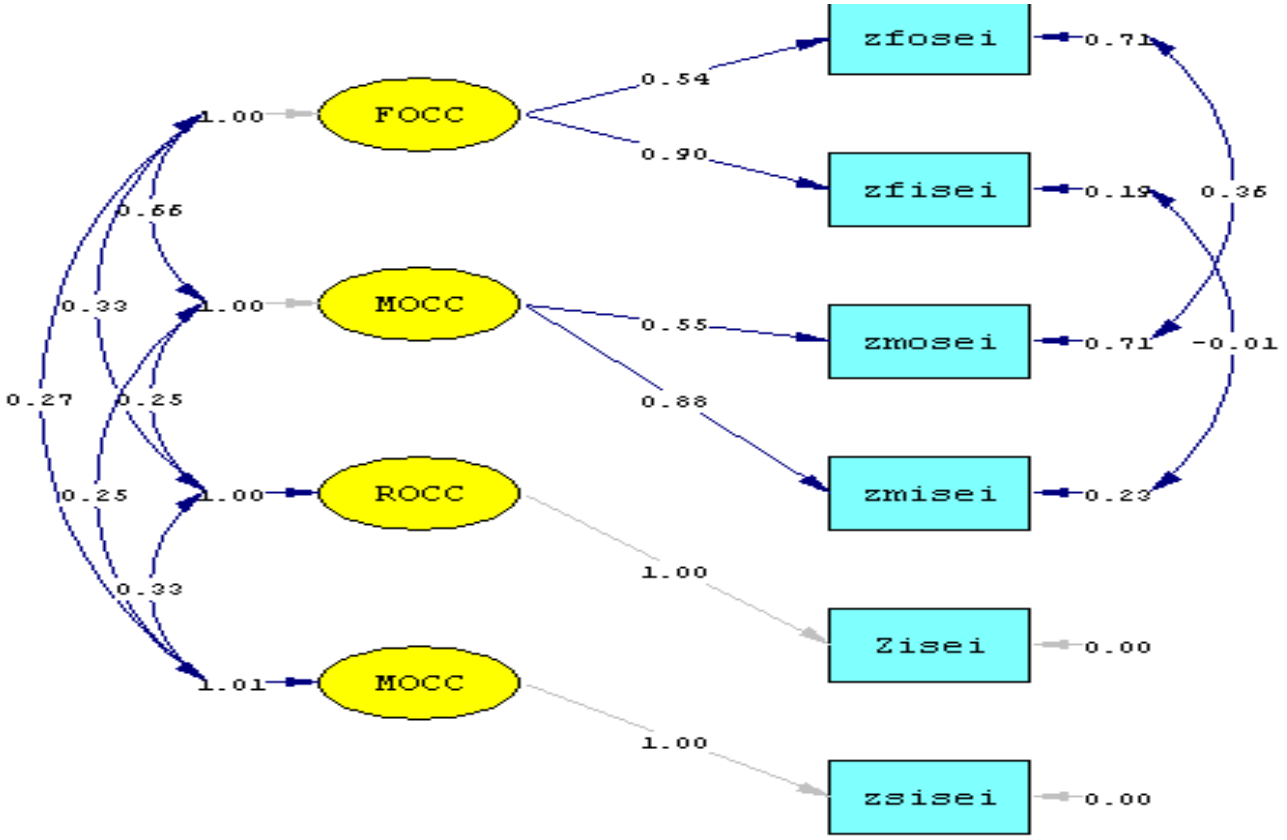
- This is a very bad way to ask for an occupation:
  - Unclear labels
  - Categories out of order.
  - No farmers (this is the most frequently occurring parental occupation!) – where would it fit?
- But it is still a second, independently collected indicator, that allows us to model measurement error.

The observed correlation matrix

Correlations

	zfisei Zscore(fisei)	zfosei Zscore(fosei)	zmisei Zscore(misei)	zmosei Zscore(mosei)
zfisei Zscore(fisei)	1	.488	.503	.316
		0.000	.000	.000
	8648	8381	2407	2408
zfosei Zscore(fosei)	.488	1	.320	.542
	0.000		.000	.000
	8381	8538	2371	2405
zmisei Zscore(misei)	.503	.320	1	.484
	.000	.000		.000
	2407	2371	2678	2607
zmosei Zscore(mosei)	.316	.542	.484	1
	.000	.000	.000	
	2408	2405	2607	2692

The SEM model



hi-Square=3.09, df=3, P-value=0.37806, RMSEA=0.002

Observed correlation matrix

- Note that the correlation between father and mother occupation is stronger using the crude measure (zfosei * zmosei) than in detailed measure (zfisei * zmisei).
- Also note that the cross-indicator correlations (zfosei*zmisei, zfisei*zmosei) are much lower.
- Also that the two indicators (e.g. zfisei*zfosei) correlate quite weakly.

The SEM model

- We use two auxiliary variables (ROCC SOCC) to make all coefficients identified.
- The crude measures have very low coefficients (around 0.54) relative to the detailed measures (around 0.89).
- Crude measures suffer from systematic (correlated) measurement error, the detailed measures not.
- Interpretation: the respondents did not really know how to interpret the showcard, but when they answered, they chose similar answers for father and mother!

Which data to analyze?

- Main choices are:
  - Correlation matrix
  - Unstandized correlation matrix (covariance matrix).
  - Unit data, either unstandardized or standardized.
- I strongly prefer to analyze standardized (correlational) data.
  - Much more stable in computations
  - Easier to interpret
  - Unstandardized effects are often intrinsically uninteresting, but this depends on context.
- Notice that SPSS makes the same choice for factor analysis.

Unit data / missing values

- The crucial advantage of modeling unit data is that this allows for FIML (Full Information Maximum Likelihood) estimation in data with missing values.
- Effectively, this is an advanced and proper way to apply “pairwise deletion of missing values”.
- You have to assume that the missings are Missing At Random (MAR).
- What you get is adjustment of the standard errors to the amount of valid data involved in the estimation of the effects.

Standardization

- Standardization is a difficult problem in latent variable models.
- Thus is so, because latent variables do not have a natural unit of measurement.
- Basic choices are:
  - Make the unit of measurement equal to one of the observed indicators. This is called the ‘reference indicator’.
  - Standardize the latent variables (Z-unit).
- The two choices are equivalent. Lisrel provided the second option as the “standardized solution”. If also the observed variables are standardized it is called the “completely standardized solution”.
- Notice that standardizing the observed variables (correlations as input) is not the same as standardizing the latent variables – but it makes the two standardized solutions identical.

Fit

- SEM models come with an overload of fit statistics that all inform about how well the model reproduces the observed correlation/covariance matrix.
- Most popular are:
  - LR chi square
  - RMSEA with a test that assesses whether the mean standardized residual is above .05 (an arbitrary cut-off point).
- Lisrelites worry quite a bit about fit – a concern that most other practitioners are simply unaware of.
- The fit statistics in SEM are known to be very powerful – even too powerful – they reject the H0 all the time.

Latent means and dispersions

- In a completely standardized model, we do not address latent means and dispersions, we simply fix these as Z-scores.
- However, when comparing groups (or time periods), our interest may be how latent means (or even latent dispersions) are different.
- In this case we need to equalize the latent unit of measurement with one of the observed indicators (reference effect).
- LISREL is quite complicated in this respect, Mplus and Stata12 are more natural.

Confirmatory or exploratory?

- SEM are often cast as “confirmatory” as opposed to “exploratory”. We have to choose the specification, and do not leave it to the program.
- The reference here is in particular to factor analysis, that is a truly exploratory technique.
- However, SEM’s have their exploratory features too, in particular when you adjust of “modify” your model.
- In fact, LISREL includes a mode, in which you can ask the program to specify your model automatically – do not use this.

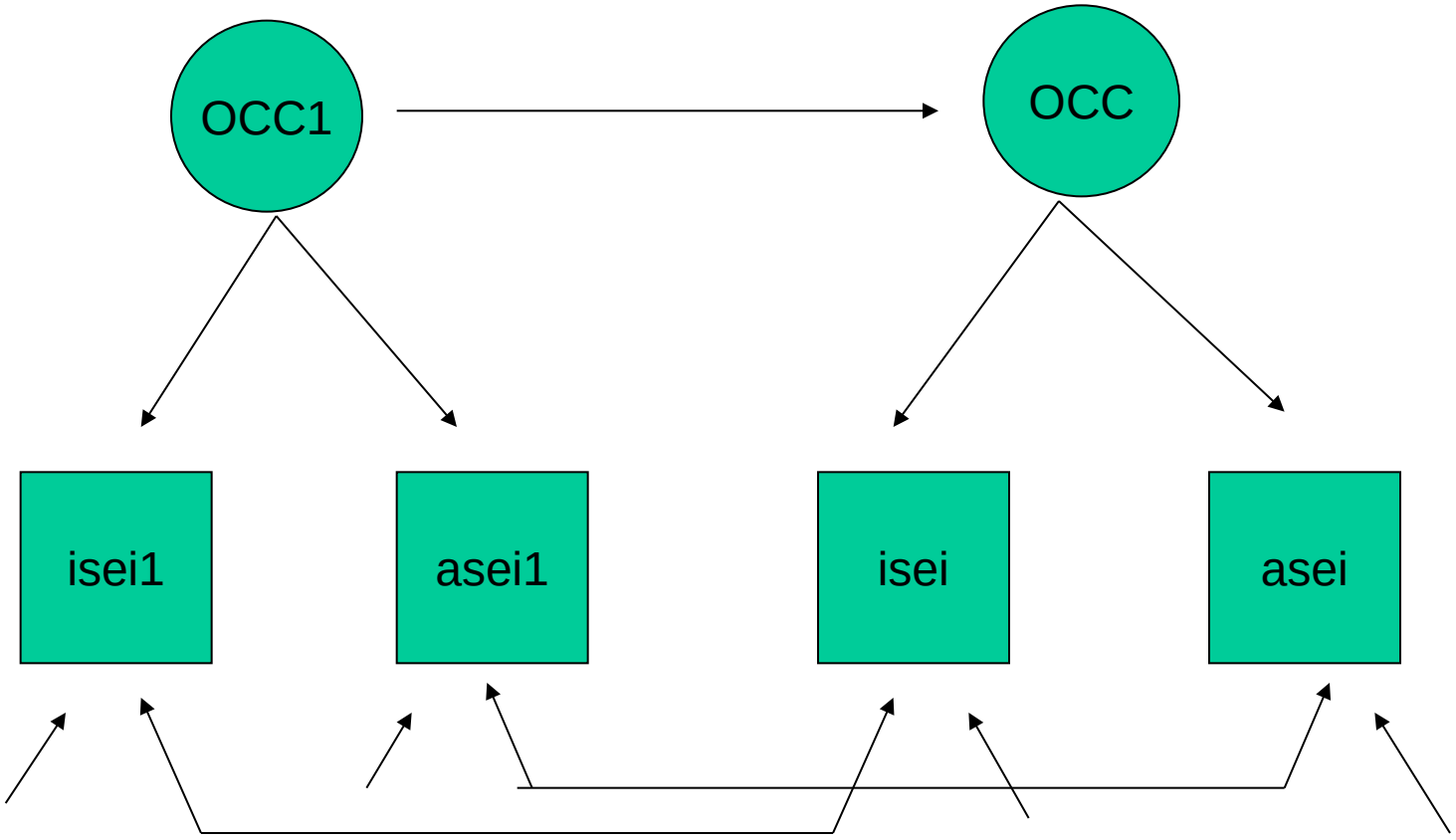
Example: status attainment models with multiple indicators

- SEM's with multiple indicators are most often used in modeling attitudinal data.
- Most empirical work on social structural variables effectively assumes that social demographic measures (education, occupation, income) do not contain measurement error.
- This is of course not true, but we need multiple indicators to find out.

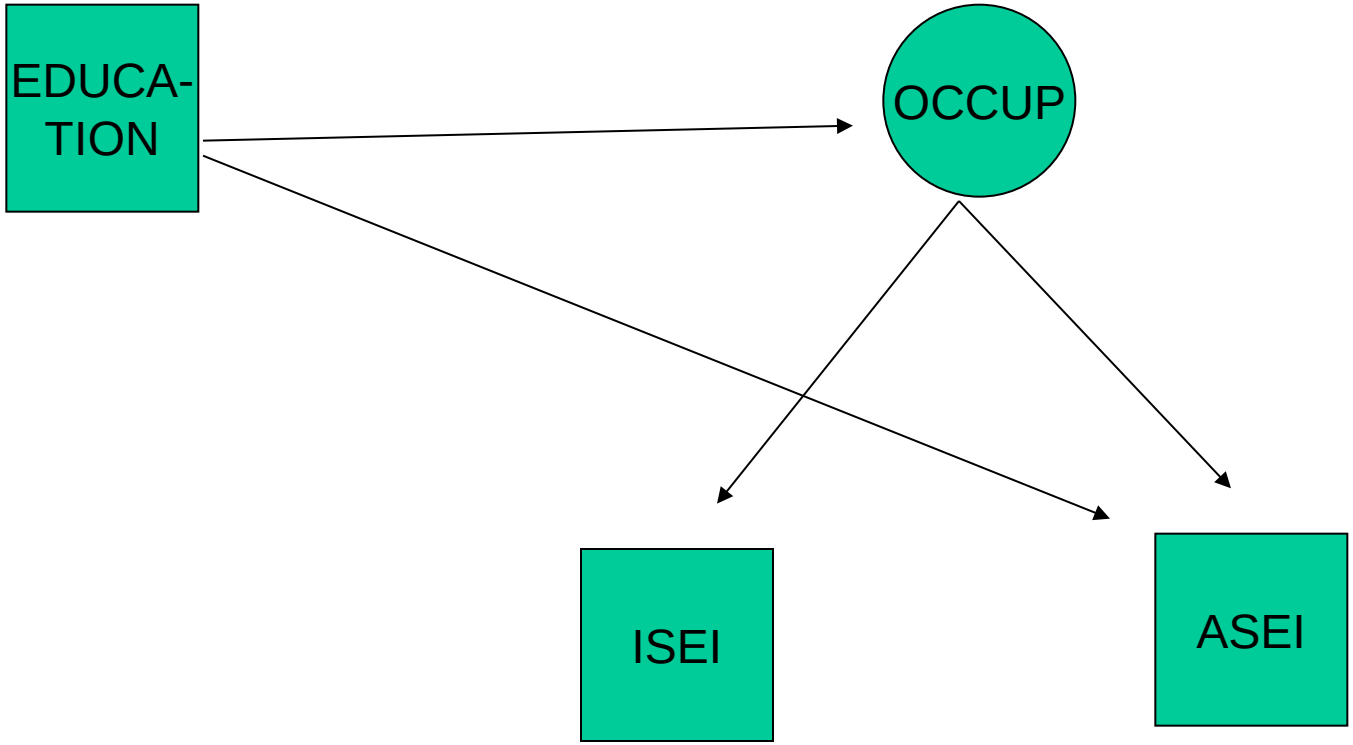
De Vries & Ganzeboom (2008): Multiple indicators for occupational status

- Asking a simple question for occupational status is already hard, asking a second, parallel one, is even harder.
- Solution proposed by Ganzeboom (2005): ask a detailed and a crude questions at the same time (‘alternate form’).
- If we repeat this across multiple occupations (father, mother, first, current), we obtain a MTMM design.
- Schröder & Ganzeboom (2010) show that the same trick can be played for education: you can measure it using (scaled) qualifications and duration as multiple indicators.

SEM / MTMM model



SEM: Education bias



Results ISSP-NL

- Detailed occupation: 0.84 / 0.77
- Crude occupation: 0.87 / 0.82
- Residual (crude) 0.10 / 0.06
- Residual (detailed) 0.05 / 0.03
- Education bias 0.03

ISSP-NL: findings

- Table 1-2-3 report on the status attainment correlations and on the LISREL model.
- With respect to random error, crude occupation measures are as good as or better than detailed measures.
- The amount of correlated error (‘echo’) is limited and does not bias causal coefficients; it is almost the same for crude and detailed measures!
- The amount of education bias in the crude measure is statistically significant (!) but substantially negligible.

Multiple indicators for demographic variables

- When faced with two indicators for the same concept, the natural response is to ask which is the best.
- However, this is NOT the idea of a SEM measurement model. While SEM can identify a best indicator, “best” does not imply it is “perfect”.
- Perfect measurement we can only obtain in a SEM models that allows us to correct measurement error.
- By choosing the best indicator we disattenuate by 5%, if we include multiple indicators, we disattenuate by 15%.