Ten reasons why you should be doing SEM for the rest of your life

Introduction for Quantlab Seminar, University of Melbourne

October 28, 2024

Harry BG Ganzeboom

Department of Sociology, VU University Amsterdam

Ten reasons

- Reason #1: SEM makes you think causally - use directed arrows.
- Reason #2: SEM makes you think processually – mediation and confounding.
- Reason #3: We observe covariations – the task of science is to theorize and test the explanation of these covariations – this is what SEM does.
- Reason #4: SEM makes you distinguish between the real world and data – the difference is measurement.
- Reason #5: SEM allows to diagnose and correct attenuation by random measurement error [unreliability] – using multiple indicator measurement.
- Reason #6: SEM allows to diagnose and correct systematic measurement error [invalidity] – using multi-trait multi-methods modelling.
- Reason #7: SEM allows you to test measurement invariance between groups and/or constructs.
- Reason #8: SEM lets you boost statistical power, by using constrained estimation.
- Reason #9: SEM lets you use all available data – which lets you boost statistical power and repair.
- Reason #10: with SEM you can estimate causally smart designs, such as simplex (markow) models. instrumental variables and reciprocal effects.

Acknowledgements

- Please cite these materials as: Ganzeboom, Harry BG (2024). “Ten reasons why you should be doing SEM for the rest of your life.” Presentation at the Quantlab Seminar, University of Melbourne [AU], October 28, 2024.
- This presentation is a summary of materials on SEM modelling that I assembled during PhD courses at the University of Melbourne, in 2015 and 2017, and subsequently updated in MA tutorials at VU University Amsterdam.
- These materials are available at http://www.harryganzeboom.nl/Teaching/SEM/index.htm.
- I have added a document with Stata syntax and comments as worked examples of some of my statements.

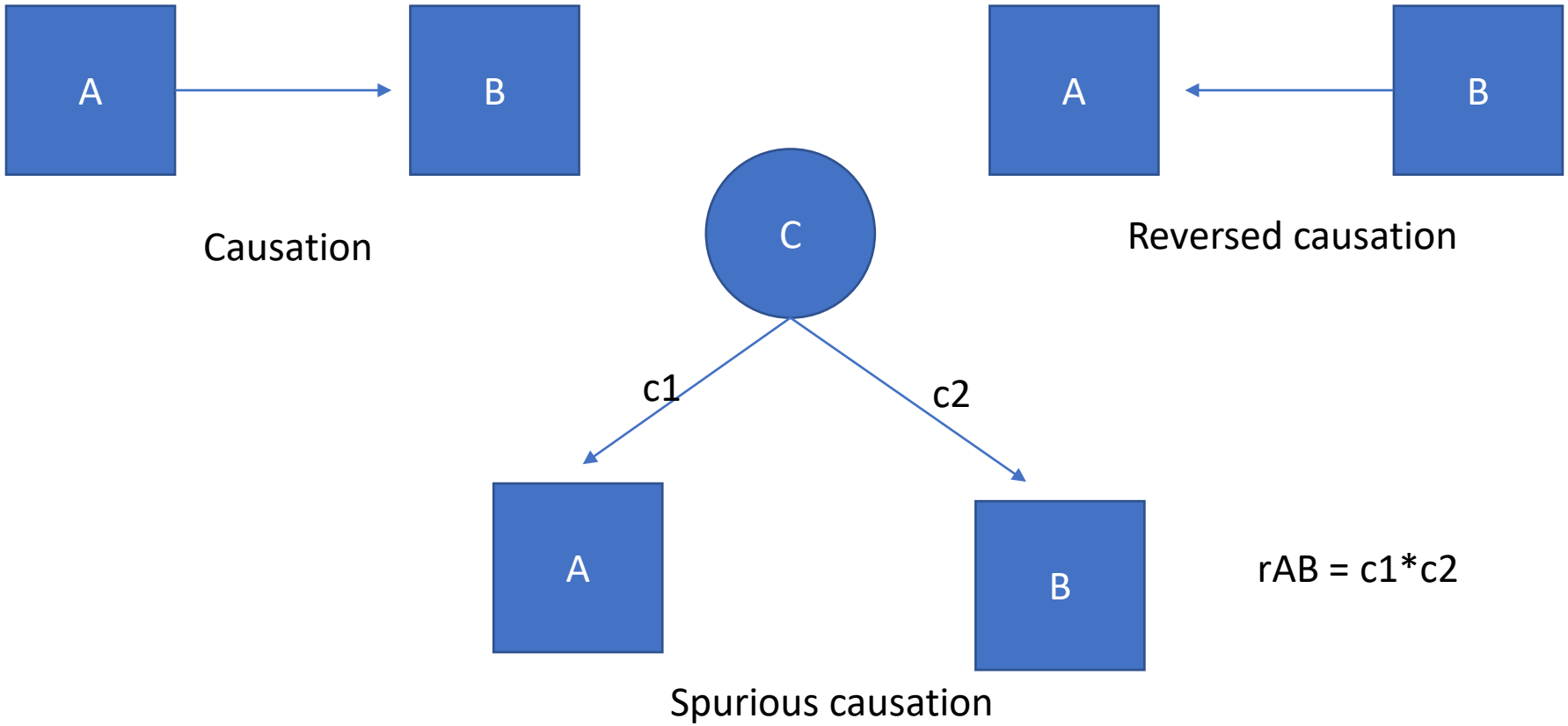
Good readings

- StataCorp. 2011. *Stata Structural Equation Modeling Reference Manual Release 12. [Software Manual]*.
- Bollen, Kenneth A., and Judea Pearl. 2013. "Eight Myths About Causality and Structural Equation Models." *Handbooks of Sociology and Social Research*, no. January: 301–28. https://doi.org/10.1007/978-94-007-6094-3_15.
- Pearl, Judea. 2014. "The Causal Foundations of Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by Rick H Hoyle, 68–91. New York: Guilford.

Let us look at a single correlation

- There are three possible ways how a correlation between A and B can come about:
  - Causation: A → B
  - Reversed causation: B → A
  - Spurious causation B ← C → A (fork)

Three possible explanations of a correlation



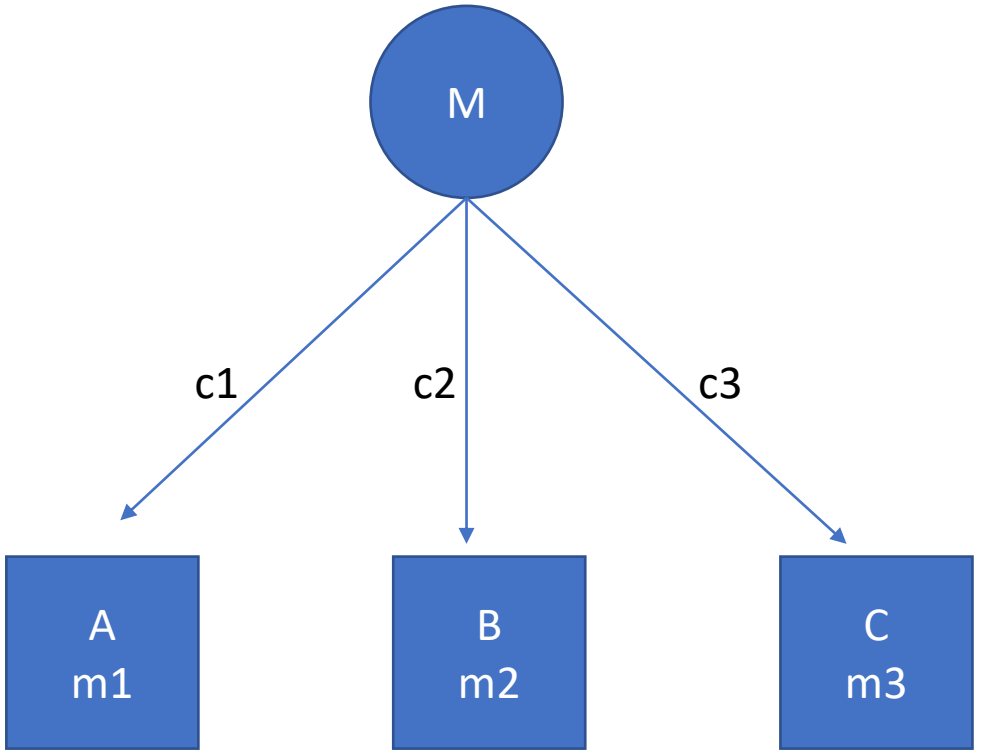
Fundamental theorem of path analysis

- ***Correlation = direct effect + indirect effects + spurious effects***
- ***Correlation = direct effect + chains + forks***
- Original:
  - Wright, Sewall. 1921. “Correlation and Causation.” *Journal of Agricultural Research* 20 (7): 557–85.
  - Wright, Sewall. 1934. “The Method Of Path Coefficients.” *The Annals of Mathematical Statistics* 5 (3): 161–215. https://doi.org/10.1214/aoms/1177732676.
- In sociology:
  - Duncan, Otis Dudley, and Robert W Hodge. 1963. “Education and Occupational Mobility a Regression Analysis.” *American Journal of Sociology* 68 (6): 629–44. https://doi.org/10.1086/223461.
  - Duncan, Otis Dudley. 1967. “The Process of Stratification.” In *Blau, Peter M.; Duncan, O. Dudley, The American Occupational Structure. Wiley, New York*, 163–205.
  - Alwin, Duane F, and Robert M Hauser. 1975. “The Decomposition of Effects in Path Analysis.” *American Sociological Review* 40 (1): 37–47. https://doi.org/10.2307/2094445.

Factor analysis = algebra of the fork

- We cannot solve: rAB = c1*c2.
- (unless assuming c1 = c2).
- But with a third indicator the system of equations is exactly identified.

Factor analysis – latent variable measurement model



rAB = c1*c2
rAC = c1*c3
rBC = c2* c3

Three equations with three unknowns == Identified

c1 c2 c3 are factor loadings = measurement coefficients

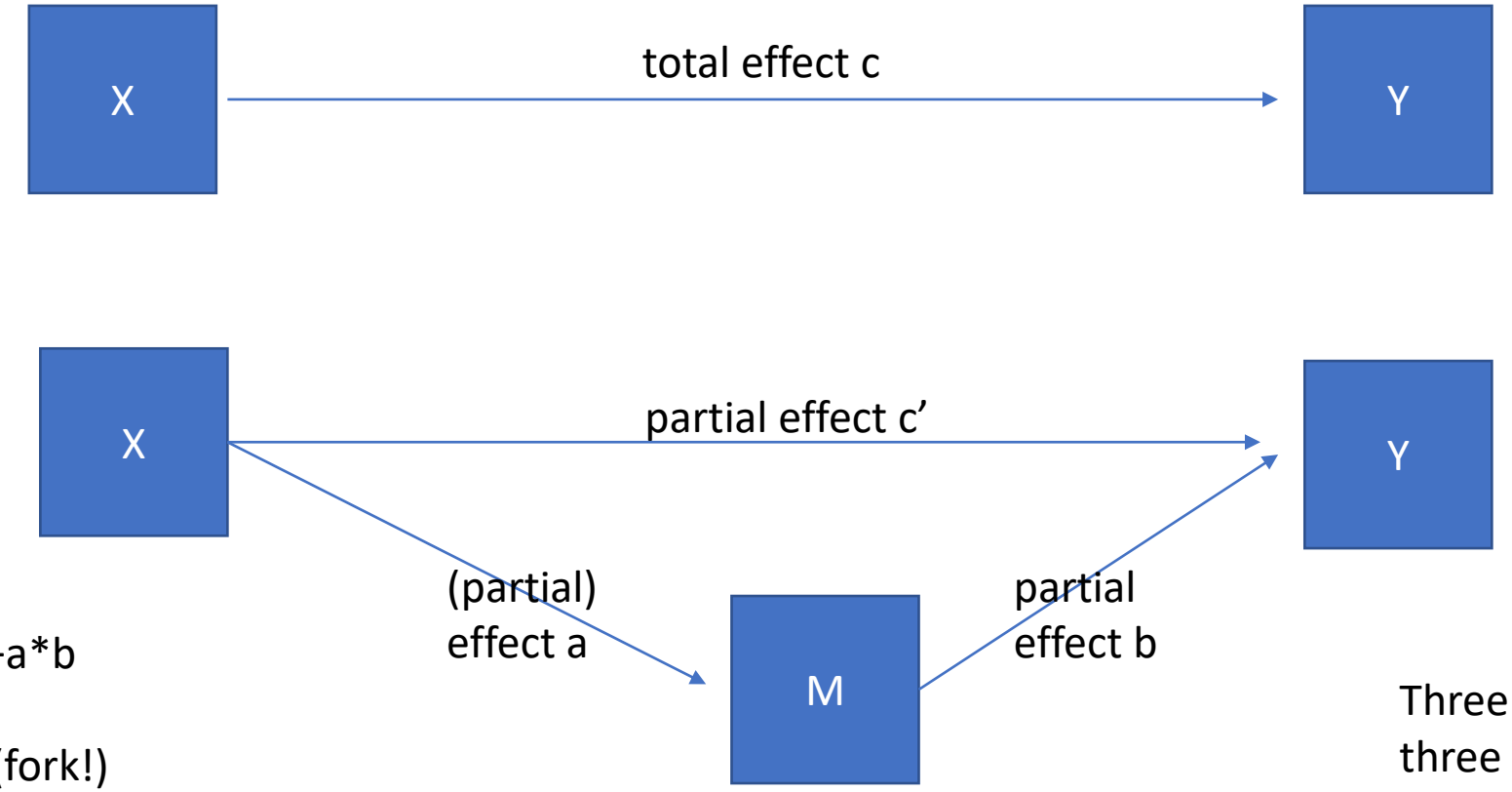
What have you learned untill now (I)?

- The world is a correlation (covariance) matrix.
- Correlation implies causation.
- Measurement is a causal process: a latent variable causes the observed indicators.
- Causality is symbolized by arrows: these arrows have a **direction** and a **strength**.
- The structure (direction) of the causal effects comes from theory (not from the data). Only the strength of effects comes from the data.
- Algebra of forks: confounding effect = multiplication of direct effects.
- Causal model generates expected correlations: the difference between expected and observed correlation == model fit.
- Also (but NOT TRUE): for proper measurement you need THREE indicators.

Fit statistics

- SEM models imply expected correlations: c1*c2 = r(AB).
- Expected correlations are compared to observed correlations in fit statistics.
- Most relevant:
  - L2(df): a Chi2 statistic that tests significance of the differences between expected and observed correlations. Strongly sensitive to sample size.
  - RMSEA: Root Mean Square Error of Approximation. Tests whether misfit is within acceptable boundaries. Comes also with a significance test.
- The SEM literature is littered with (other) fit statistics. Not useful. Rather concentrate on parameters and changes between models.

Mediation analysis = algebra of the chain



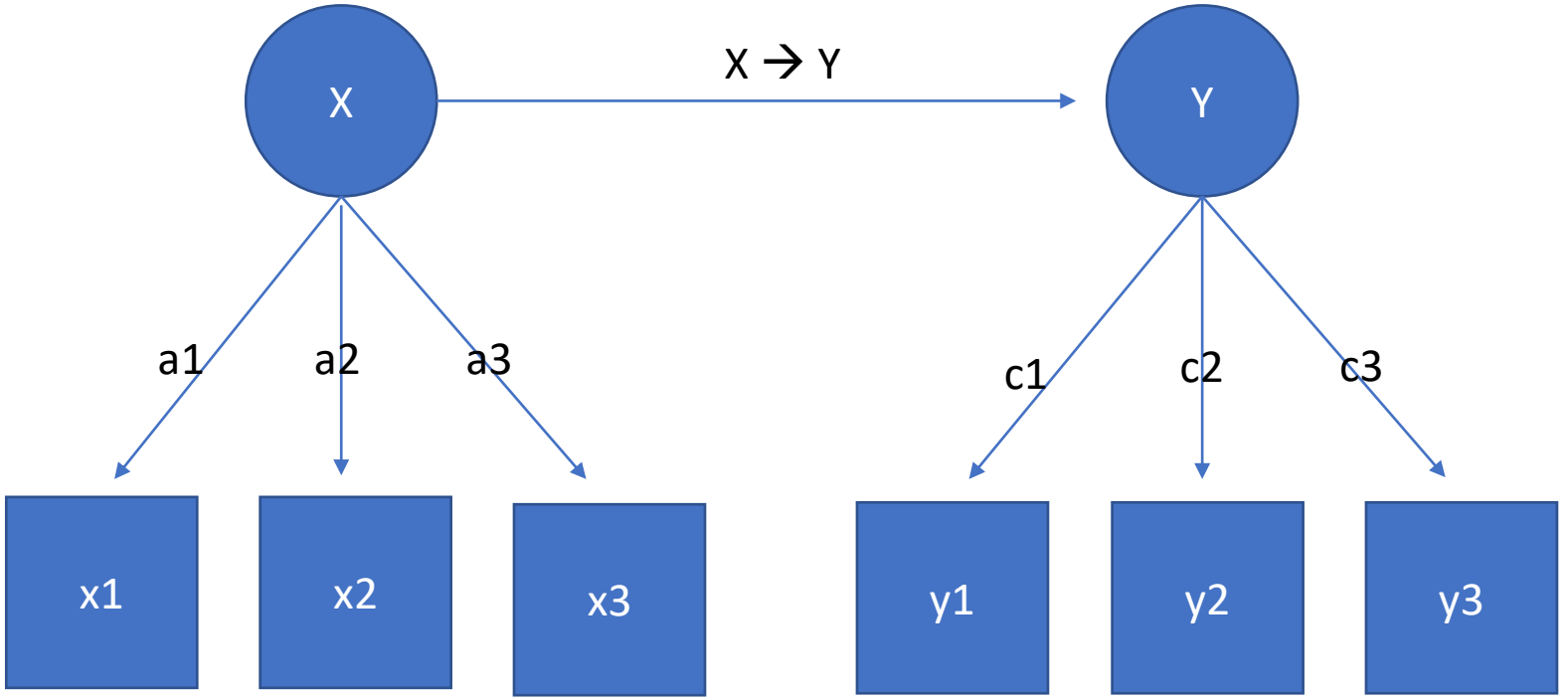
rXY = c = c'+a*b
rXM = a
rMY = a*c' (fork!)
a*b =I ndirect effect (chain)
a*c = confounding effect (fork)

Three equations with three unknonws == identified

SEM = Structural (or: Simultaneous) Equation Modelling

- SEM merges measurement models (factor analysis) with structural (mediation) models.
  - Blalock, Hubert M. 1964. *Causal Inferences in Nonexperimental Research.* New York: Norton.
  - Jöreskog, K. G. 1970. "A General Method for Analysis of Covariance Structures." *Biometrika* 57 (2): 239–51. https://doi.org/10.1093/biomet/57.2.239.
  - Hauser, Robert M, and Arthur S Goldberger. 1971. "The Treatment of Unobservable Variables in Path Analysis." *Sociological Methodology* 3: 81. https://doi.org/10.2307/270819.
- LISREL, AMOS, EQS, MPLUS, SEM in Stata, Lavaan in R.

The total (=bivariate) effect X → Y in SEM



r(x1,x2) = a1*a2
r(x1,x3) = a1*a3
r(x2,x3) = a2*a3

r(y1,y2) = c1*c2
r(y1,y3) = c1*c3
r(y2,y3) = c2*c3

r(x1,y1) = a1*XY*c1
r(x1,y2) = a1*XY*c2
r(x1,y3) = a1*XY*c3

r(x2,y1) = a2*XY*c1
r(x2,y2) = a2*XY*c2
r(x2,y3) = a2*XY*c3

r(x3,y1) = a3*XY*c1
r(x3,y2) = a3*XY*c2
r(x3,y3) = a3*XY*c3

Example from fake.dta

- What happens if we leave out indicators from the measurement model?
- Using constrained estimation: forcing equality between coefficients.
- Model fit, residuals.
- Correction for attenuation with known (un)reliability.

How many indicators?

- If you embed a measurement model in a structural model, TWO indicators suffice for disattenuating structural relationships.
- More indicators and better (stronger correlated) indicators (should) have no consequences for the structural relationships, but make SE smaller. Better measurement is like having a larger sample.
- However, with more than three indicators the measurement model is driven by internal consistency, with two indicators the measurement model is driven by external variables.
- If you prefer two indicators and have more, consider random split-half parcelling: assign indicators to two parts by random lottery.

What if you have only one indicator?

- If you have only a single indicator, you cannot estimate a measurement model.
- However, this does not mean that your single measure is perfectly reliable…
- SEM allows you to correct attenuation when you can assume a level of reliability, such as Cronbach's alpha or McDonalds Omega.
- This may get you the correct point estimate of the structural relationships, but I have my doubt about the SE.

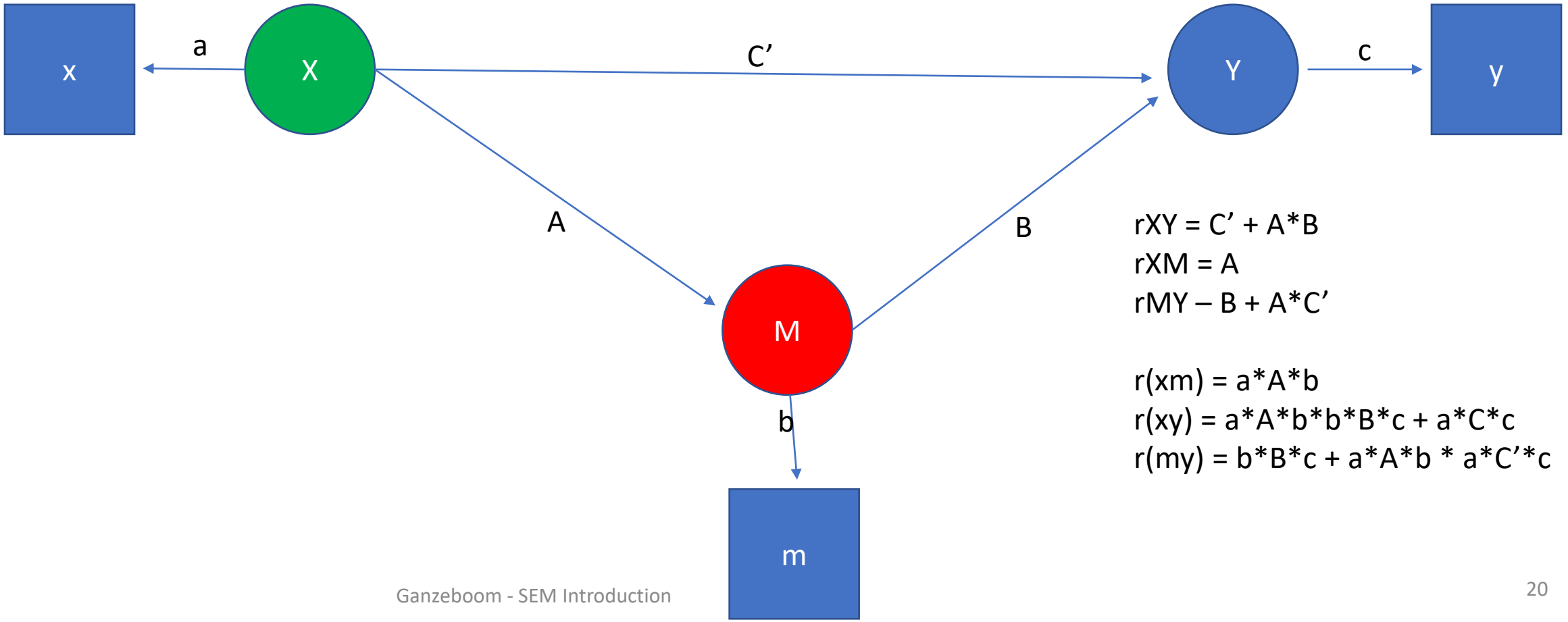
What have your learned now (II)?

- Measurement: observed variable = true score + measurement error.
- Random measurement error = unreliability.
- Unreliability always weakens (‘attenuates’) correlations, it never makes them stronger.
- A measurement model corrects this attenuation: disattenuation.
- If you embed a measurement model in a structural model, TWO indicators are enough for disattenuation.
- More indicators and less random error in each does not change disattenuation (much), but it does lead to smaller SE (sampling fluctuations).
- A latent variable is the real thing. Data exist only on your computer.
- Averaging multiple indicators into an index gets you closer to the true score, but some difference remains. The difference is measured by reliability coefficients like Cronbach’s alpha or McDonald’s Omega.

Mediation (indirect effects)

- In mediation problems the algebra of chains applies: an indirect effect is a multiplication of the two direct effects.
- This algebra of a chain is similar, but not the same as the algebra of a fork.
- In a mediation model you can see both a chain (X → M → Y) and a fork (Y ← X → M).
- Confounding and mediation are very much related, but they have radically different interpretations:
  - Confounding: no causal effect X → Y
  - Mediation: causal effect X → Y is explained
- Both confounding and mediating variables are called control variables. The use of the word “control variable” is rather confusing and should be banned.

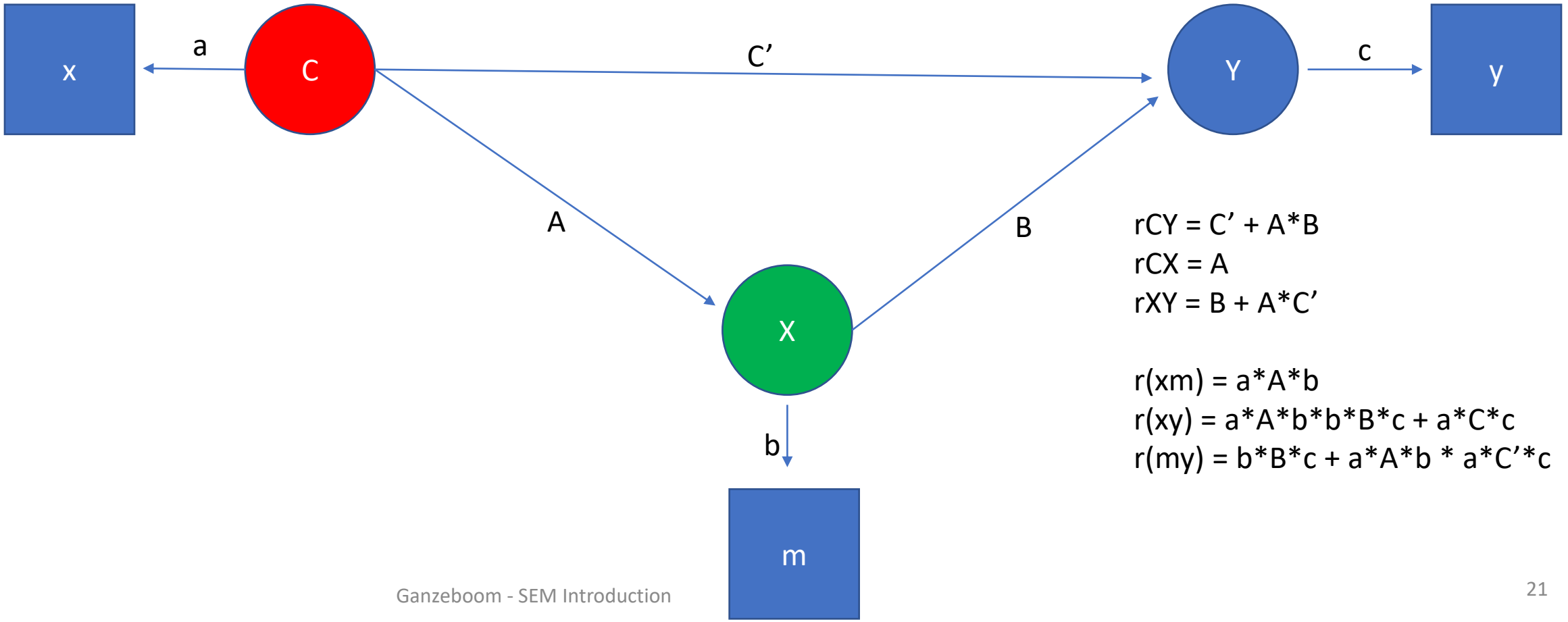
Mediation with measurement effects



rXY = C' + A*B
rXM = A
rMY – B + A*C'

r(xm) = a*A*b
r(xy) = a*A*b*b*B*c + a*C*c
r(my) = b*B*c + a*A*b * a*C'*c

Confounding with measurement effects



rCY = C' + A*B
rCX = A
rXY = B + A*C'

r(xm) = a*A*b
r(xy) = a*A*b*b*B*c + a*C*c
r(my) = b*B*c + a*A*b * a*C'*c

What have you learned by now (III)?

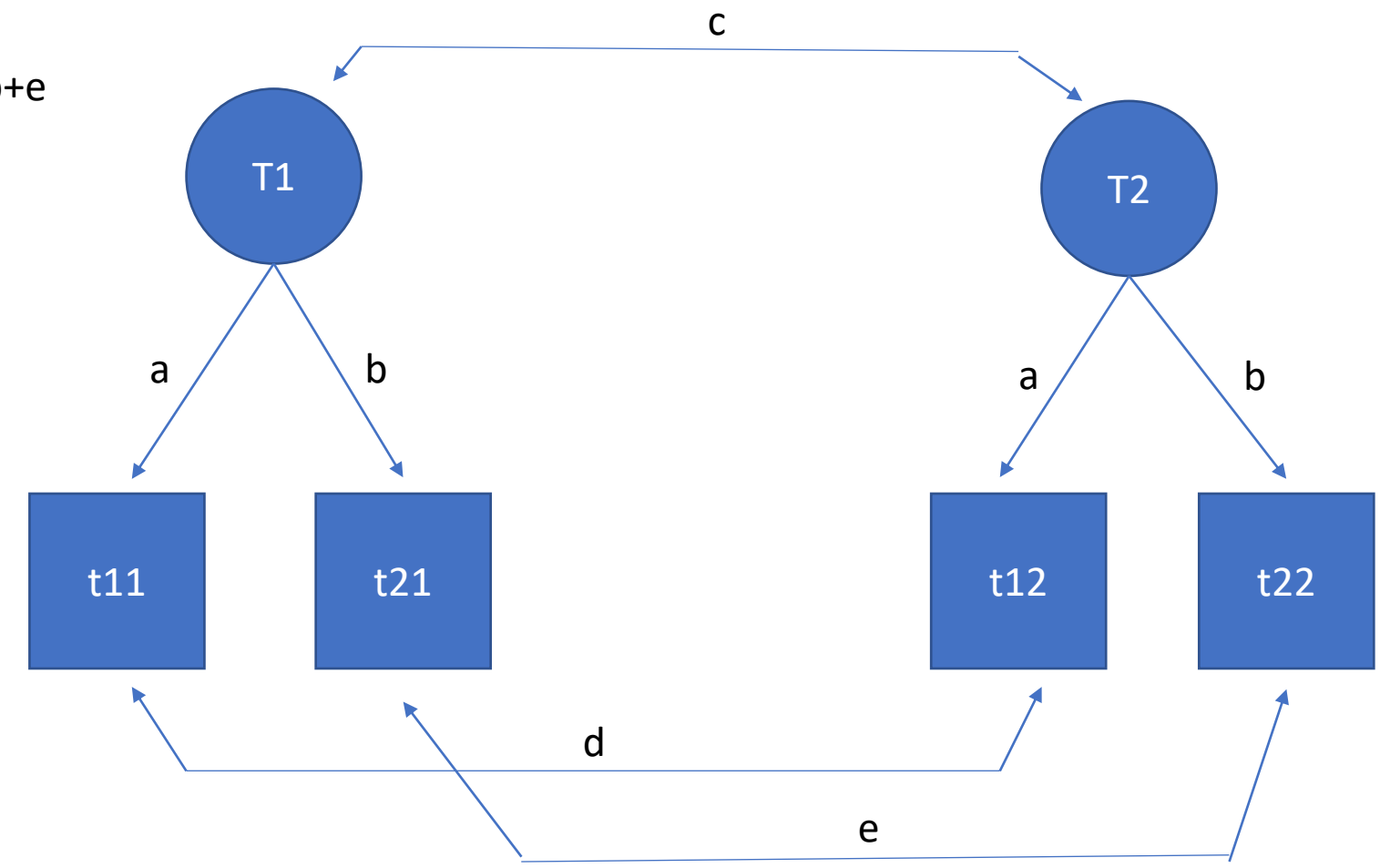
- Confounding by prior variables (spurious causation) and mediation by an intervening variable (indirect causation) follow the same algebra: confounding is a fork, mediation is a chain, but in both cases the measurement coefficients kick in twice.
- In fact, to obtain valid conclusions about mediation or confounding, the mediating or confounding variables must be perfectly measured.
- Having perfect measurement of mediation and confounding variables (M and C, both often called “control variables”) is more important to your conclusions than perfect measurement of X and Y.
- Perfect measurement can be achieved with a (SEM) measurement model.
- If you use a SEM measurement model or other forms of correction for attenuation of random measurement error, the size of indirect and spurious effects will increase, and the direct effect will be decrease.

Systematic measurement errors (invalidity)

- Measurement does not only contain random measurement error (**unreliability**), but may also have systematic measurement error (**invalidity** or bias).
- Random measurement error can be traced by **repeating the measurement**, systematics measurement error can be traced by **repeating the measurement error**.
- This is the idea behind MTMM models: multiple-trait multiple-methods.
- Sources:
  - Campbell, Donald T, and Donald W Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56 (2): 81–105. https://doi.org/10.1037/h0046016.
  - Andrews, Frank M. 1984. “Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach.” *The Public Opinion Quarterly* 48 (2): 409–42. https://doi.org/10.1086/268840.
  - Saris, Willem E, and Frank M Andrews. 1991. “Evaluation of Measurement Instruments Using a Structural Modeling Approach.” In *Measurement Errors in Surveys*, edited by Paul Biemer, Robert M Groves, Lars E Lyberg, Nancy E Mathiowetz, and Seymour Sudman, 575–97.

r(t11,t21) = a*b
r(t12,t22) = a*b
r(t11,t12) = a*c*a+d
r(t11,t22) = a*c*b
r(t21,t22) = b*c*b+e
r(t11,t12) = b*c*a

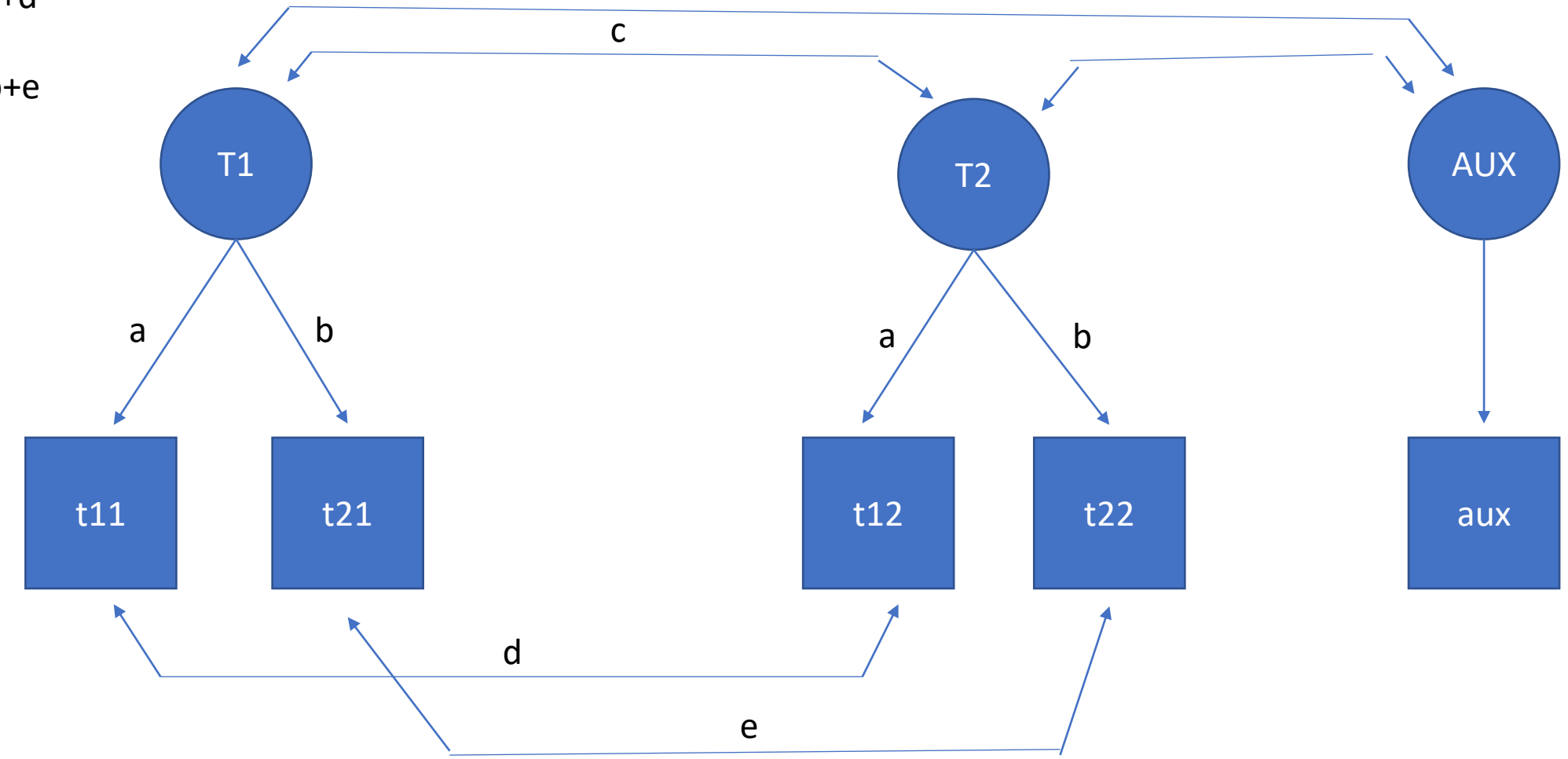
MTMM in reduced form



Model is NOT identified, but can be made identified with auxiliary variables

r(t11,t21) = a*b
r(t12,t22) = a*b
r(t11,t12) = a*c*a+d
r(t11,t22) = a*c*b
r(t21,t22) = b*c*b+e
r(t11,t12) = b*c*a

MTMM with auxiliary variables



What you should have learned by now (IV) …

- SEM MTMM allows you to separate (un)reliability and (in)validity.
- While “systematic error” sounds more threatening than “random error”, in practice random error is more problematic than systematic error.
  - Systematic error may occur, random error is always present.
  - Even if your measurement is biased, you are measuring something. If you measurement in unreliable, you are measuring nothing/

Incomplete data

- SEM can use all available data (=pairwise correlation matrix) with Full Information Maximum Likelihood (FIML, MLMV).
- FIML is much more appealing than Multiple Imputation (to which is is asymptotically equivalent).
- The two problems of complete cases analysis (=using listwise correlations):
  - Inefficiency: you have fewer cases than that you have collected.
  - Bias: the complete cases may be different from the population.
- MI makes people think about the cases that are missing. FIML concentrates on the information that you do have, but would become unused by complete cases selection.

FIML: how does it work?

- Think about dividing your data in as many groups as you have missing value pattern.
- Then estimate the model with equality constraint over all groups.
- SEM programs have standard options for this. In Stata: method(mlmv).
- You will see that the SE are wider for the parts of the correlations where you have fewer cases.
- Example: effects of father's and mother's occupation on occupation.

What you should have learned (V)…

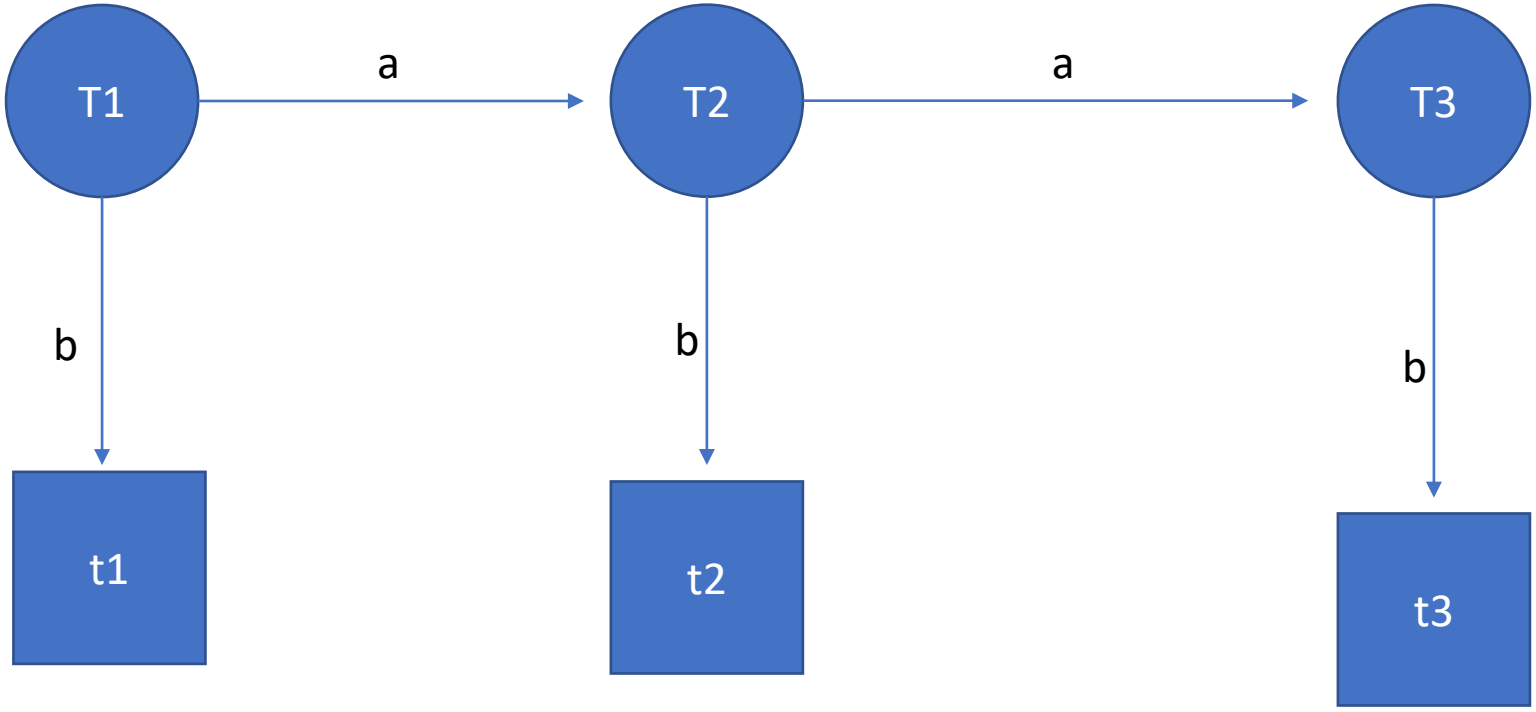
- Incomplete data are everywhere
- Using complete data always causes inefficiency and may cause bias.
- FIML is a easy and conceptually appealing way to analyse data with missing values.

Simplex (markov) model

This model can separate true change (a) from unreliability (b) in three wave panel data.

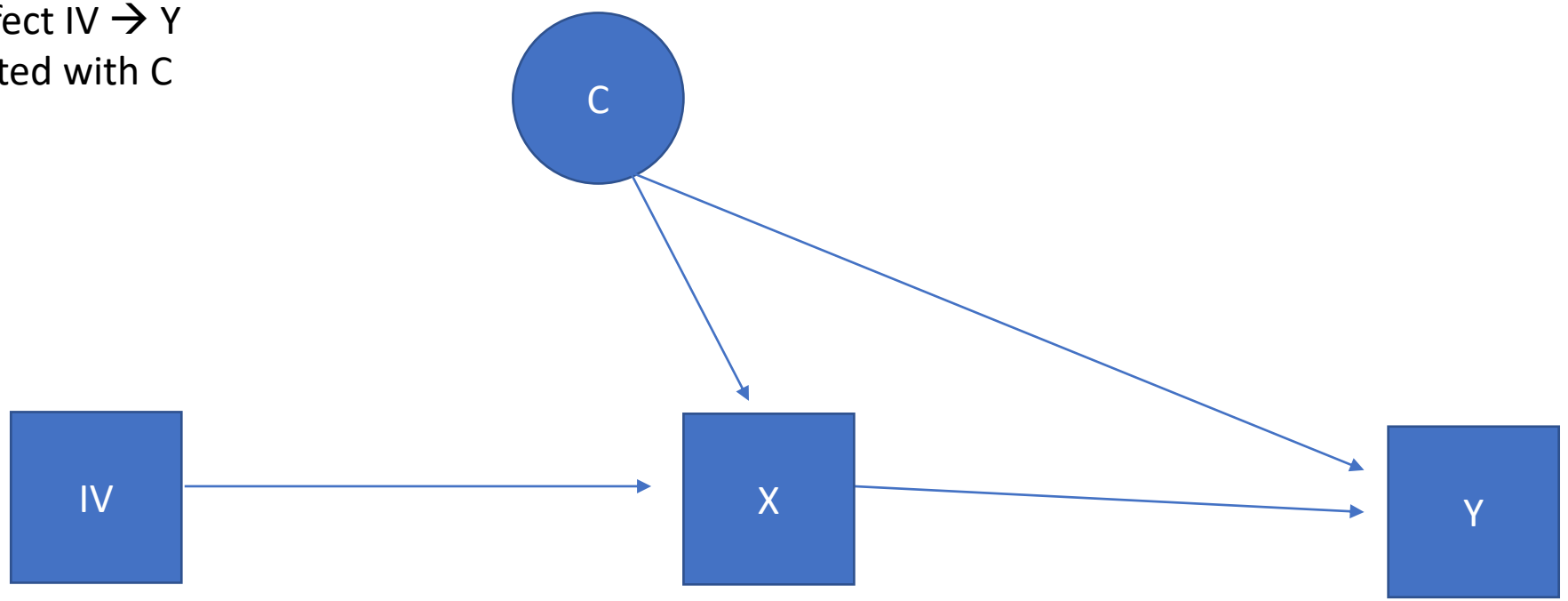
Can be used to obtain reliability estimates from single measure attributes

Assumption:
No direct effect
T1 → T3

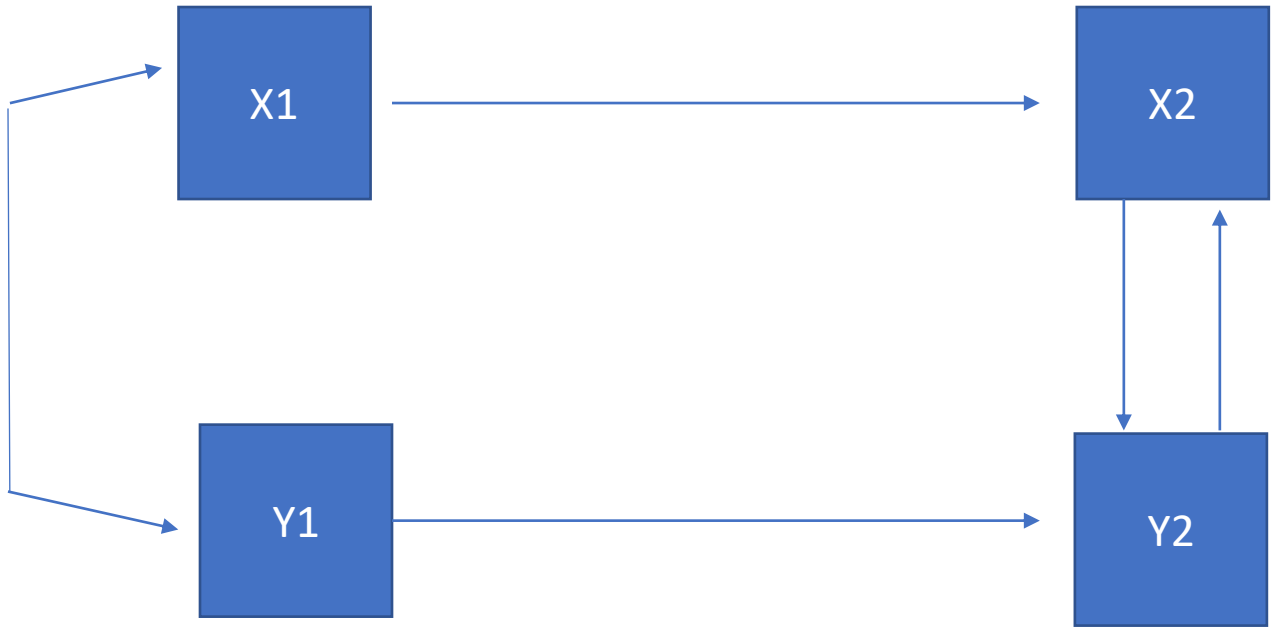


Instrumental variables model in SEM

Constraints:
No direct effect IV → Y
IV uncorrelated with C



Reciprocal effects in 2-wave panel



Alternative to cross-lagged panel-model with residual Correlation between X2 and Y2.

Restrictions & extensions

- SEM works best with continuous data; it then assumes a multivariate normal distribution. Bootstrapped and other (computer-intensive) SE estimates are available.
- Extensions to categorical data: GSEM in Stats, MPLUS → latent class analysis.
- Also available in GSEM: multi-level estimates.