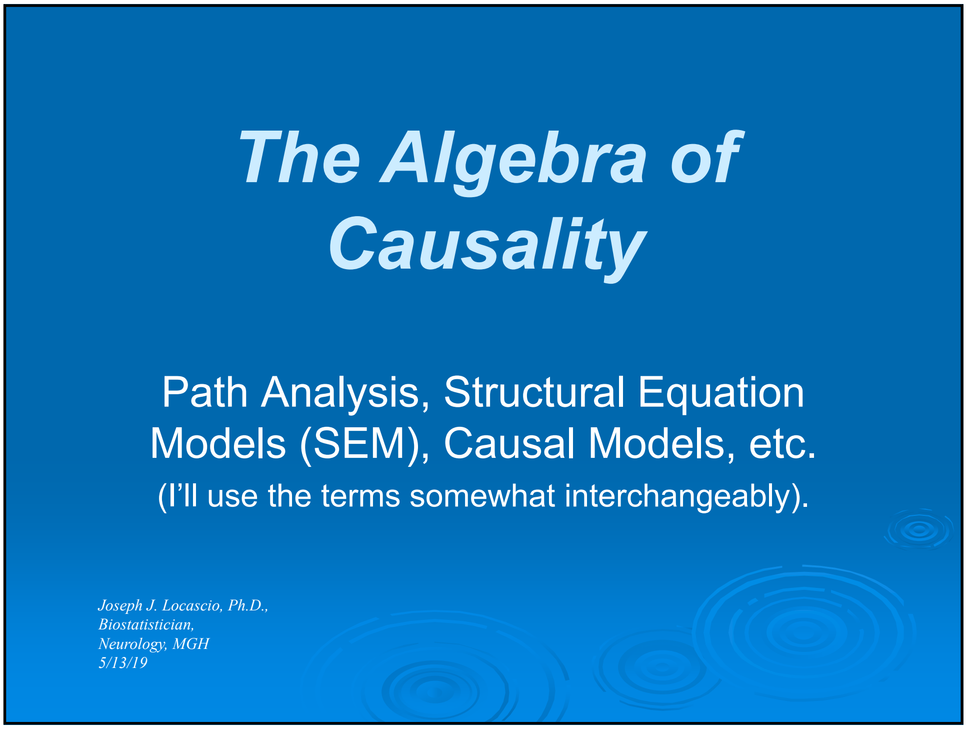
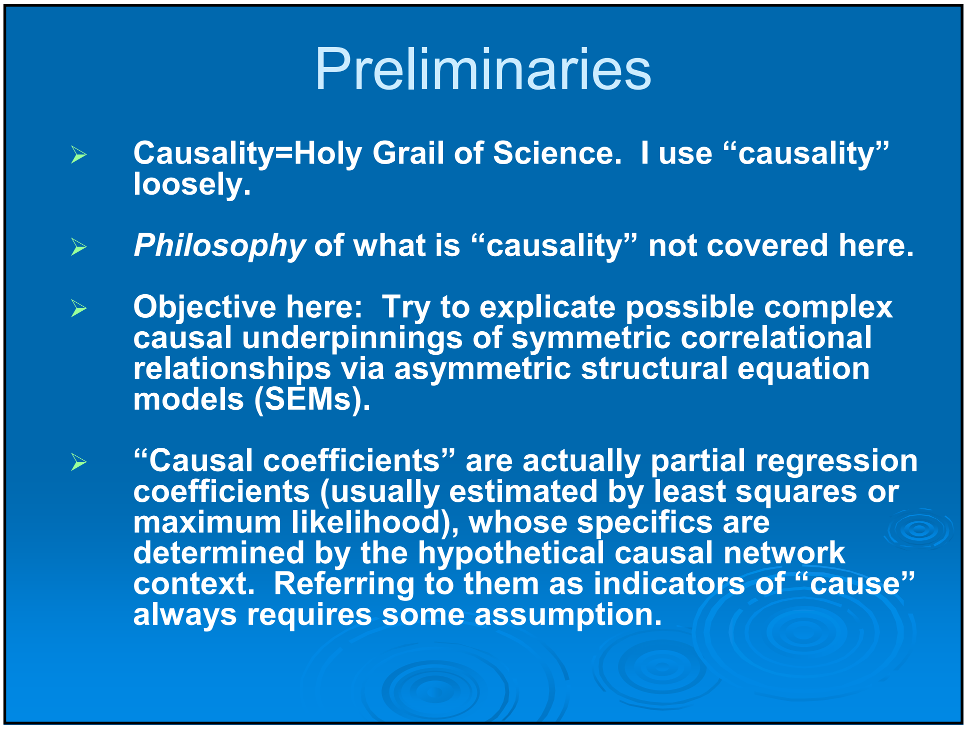


The Algebra of Causality

Path Analysis, Structural Equation Models (SEM), Causal Models, etc.

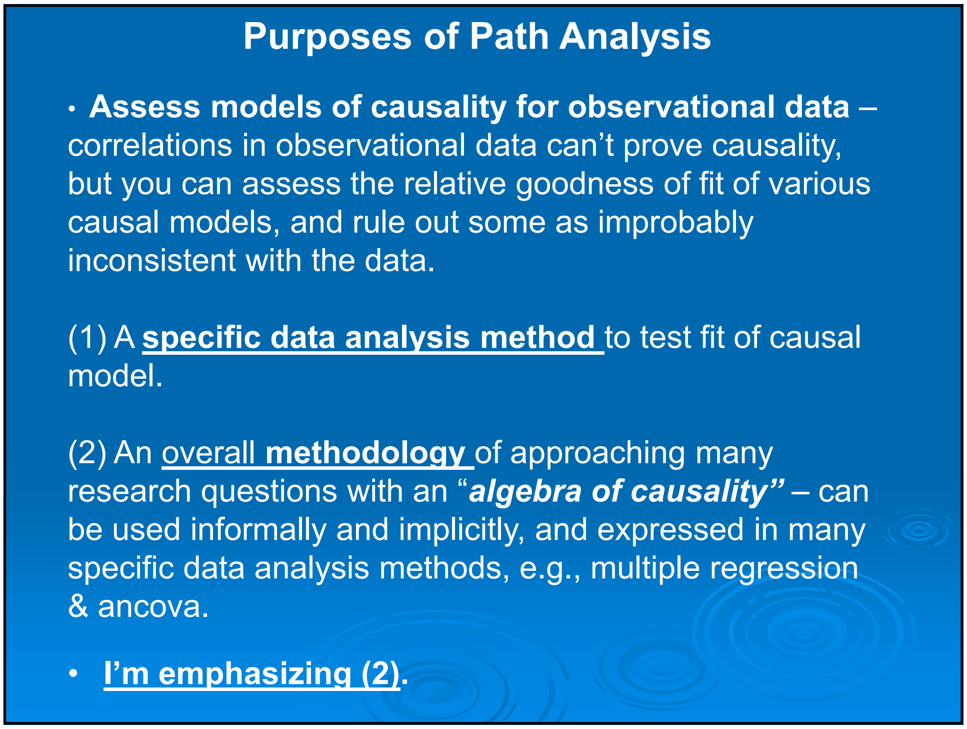
(I'll use the terms somewhat interchangeably).

Joseph J. Locascio, Ph.D.,
Biostatistician,
Neurology, MGH
5/13/19



Preliminaries

- **Causality=Holy Grail of Science. I use "causality" loosely.**
- ***Philosophy* of what is "causality" not covered here.**
- **Objective here: Try to explicate possible complex causal underpinnings of symmetric correlational relationships via asymmetric structural equation models (SEMs).**
- **"Causal coefficients" are actually partial regression coefficients (usually estimated by least squares or maximum likelihood), whose specifics are determined by the hypothetical causal network context. Referring to them as indicators of "cause" always requires some assumption.**



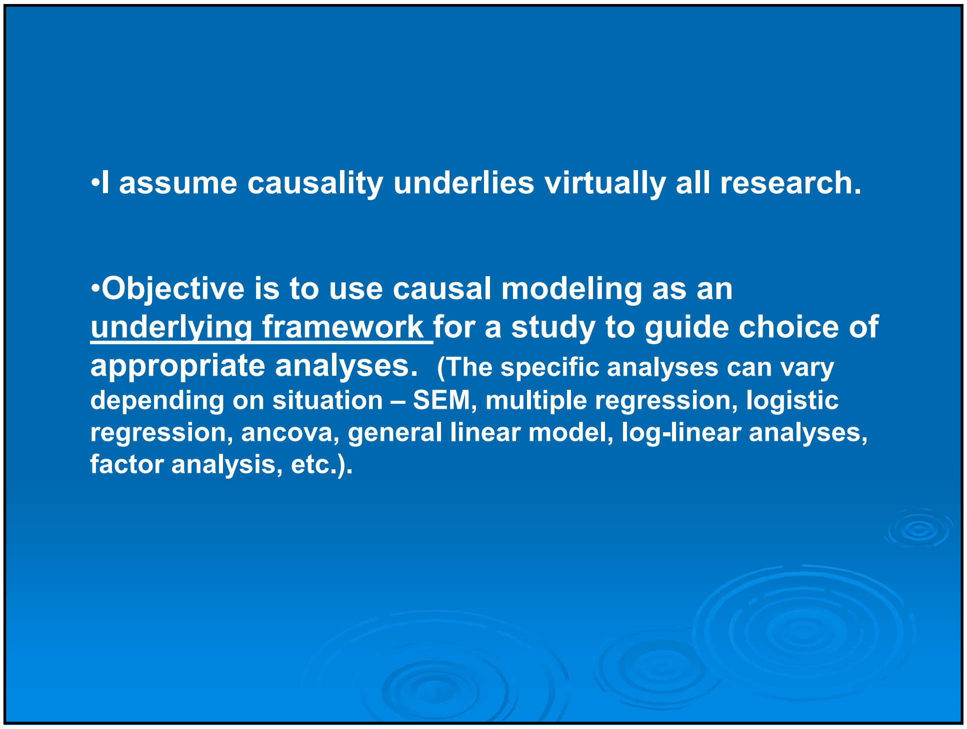
Purposes of Path Analysis

- **Assess models of causality for observational data** – correlations in observational data can’t prove causality, but you can assess the relative goodness of fit of various causal models, and rule out some as improbably inconsistent with the data.

(1) A **specific data analysis method** to test fit of causal model.

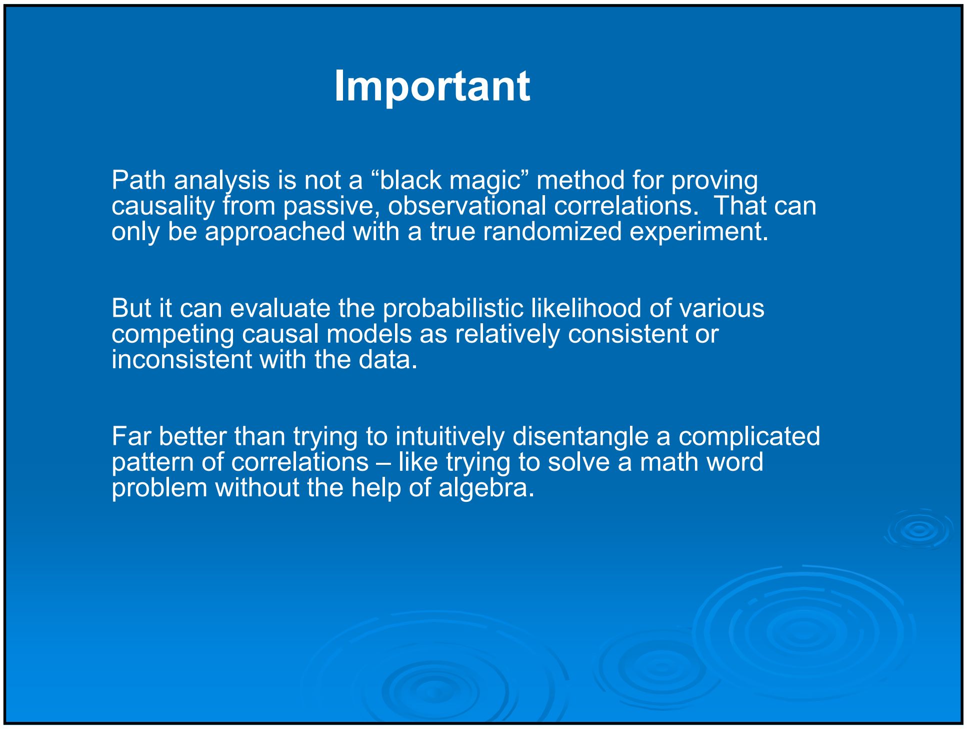
(2) An overall **methodology** of approaching many research questions with an “***algebra of causality”*** – can be used informally and implicitly, and expressed in many specific data analysis methods, e.g., multiple regression & ancova.

- **I’m emphasizing (2).**



•I assume causality underlies virtually all research.

•Objective is to use causal modeling as an underlying framework for a study to guide choice of appropriate analyses. (The specific analyses can vary depending on situation – SEM, multiple regression, logistic regression, ancova, general linear model, log-linear analyses, factor analysis, etc.).

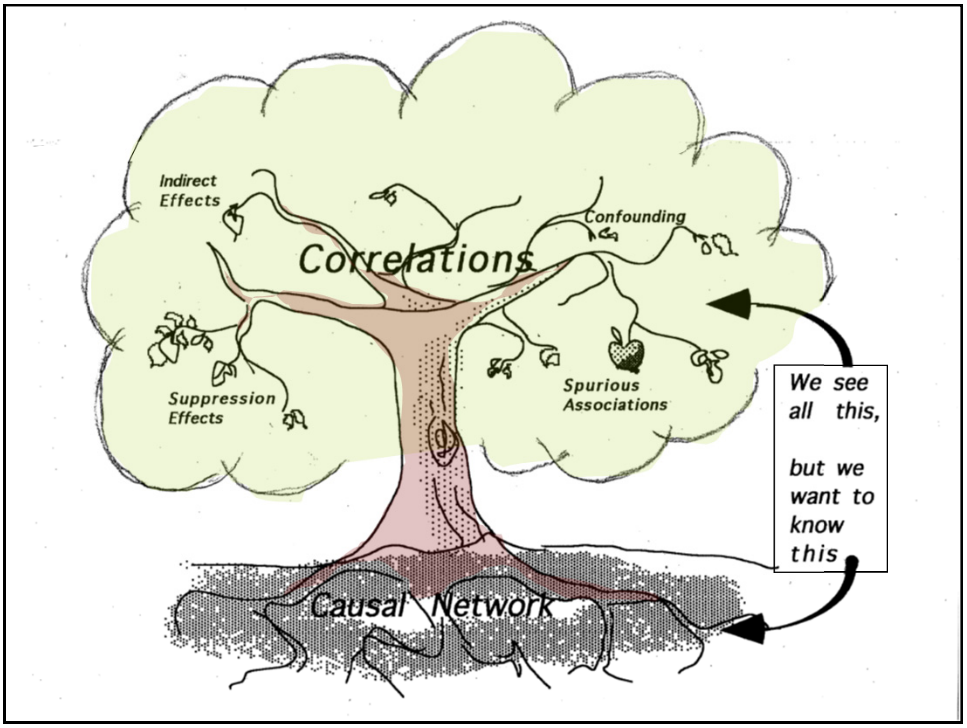


Important

Path analysis is not a “black magic” method for proving causality from passive, observational correlations. That can only be approached with a true randomized experiment.

But it can evaluate the probabilistic likelihood of various competing causal models as relatively consistent or inconsistent with the data.

Far better than trying to intuitively disentangle a complicated pattern of correlations – like trying to solve a math word problem without the help of algebra.



We see all this, but we want to know this

Uses of Path Analysis

Make sense of a complicated correlation matrix.

Provide information on:

✓ direct & indirect causal effects

✓ spurious relations & suppression effects

✓ relations among latent as well as observed variables

✓ measurement models

✓ reciprocal causality & feedback loops (nonrecursive, as opposed to recursive models)

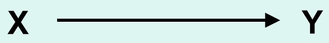
✓ used in both cross-sectional & longitudinal studies (I mostly discuss cross-sectional here)

Subsumes as specific cases: confirmatory factor analysis models, most standard parametric analyses like multiple regression, anova, ancova, general linear models, latent growth longitudinal models, etc.

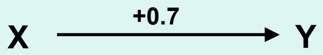
Path Analysis Diagrams

Path Diagram translates into algebraic formulas (simultaneous equations) & vice versa, but diagram easier to work with. (Directed Acyclic Graphs, “DAG”s, are a type of unidirectional, “recursive”, path diagram).

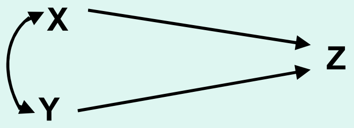
- An arrow indicates a causal effect in the direction of the arrow, e.g. variable X causes variable Y: (error terms omitted in diagrams for simplicity).



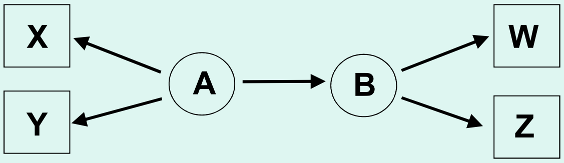
- A standardized **path coefficient** and its sign (generally -1 to +1 like a correlation coefficient) indicates strength and direction of the causal impact. E.g., a moderately strong positive causal effect of X on Y:

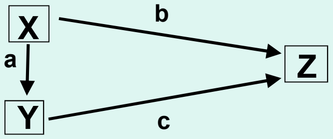


- A curved double headed arrow indicates a correlation among *exogenous* variables (variables at beginning of causal chain, as opposed to *endogenous*).



A **rectangle/square** = **observed** variable; **Ellipse/circle** = **latent** variable, e.g., latent variable “A” below causes observed variables “X”, “Y” (may be measures of “A”) and also causes latent variable “B” which in turn causes observed variables “W’ and “Z”.





As equations:

$Y = aX$

$Z = bX + cY$

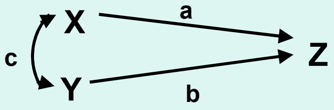
(for simplicity, I leave out circles and squares in some diagrams below)

Features of Path Analysis

- Causality of variables is assessed **holding other variables constant (partialed)**, as dictated by the model. Thus causality disentangled from correlation, confounding, spurious associations, suppression effects, and indirect versus direct effects assessed, etc.

- **Path coefficients are standardized**, like Pearson correlation coefficients, so relative impact of variables assessed. In a one arrow diagram, path coefficient = correlation coefficient. As models become more complex, they become variations of **standardized partial regression coefficients.** (Unstandardized coefficients sometimes used).

- For just identified models, **tracing rule** reproduces correlations, i.e., trace all paths between 2 variables multiplying coefficients along the way = correlation.



$r_{XY} = c$ **$r_{XZ} = a + cb$** **$r_{YZ} = b + ca$**

3 Basic "Junctions"

1. "Chain" = Indirect Effect or Mediation

$$X \xrightarrow{a} Y \xrightarrow{b} Z$$

r_{XZ} = ab

2. "Fork" = Confounder or "spurious correlation"

$$X \xleftarrow{a} Y \xrightarrow{b} Z$$

r_{XZ} = ab

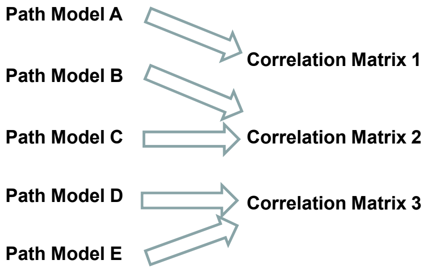
3. "Collider"

$$X \xrightarrow{a} Y \xleftarrow{b} Z$$

r_{XZ} = anything

For #3: r_{XZ} is not dictated by model. X and Z are exogenous and can have zero correlation or any correlation. Tracing rule does not apply to r_{XZ} here. And covarying for Y can induce a confound & spurious association that wasn't there before!

- **One path model produces one correlation matrix.**
- **But, one correlation matrix may be the manifestation of many possible path models.**



Path Model A
Path Model B
Path Model C
Path Model D
Path Model E

Correlation Matrix 1
Correlation Matrix 2
Correlation Matrix 3

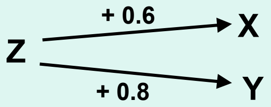
For example,

$$X \xrightarrow{a} Z \xrightarrow{b} Y \qquad X \xleftarrow{a} Z \xrightarrow{b} Y$$

both path models above produce the same correlations below:

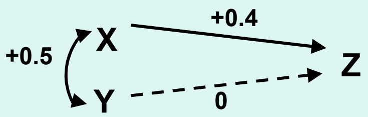
r_{xz} = a r_{zy} = b r_{XY} = ab

Example of **Spurious Association**. r_{XY} is spurious; X and Y not causally related, (e.g., white hair and senility both caused by age).



r_{XY} = (0.6)(0.8) = +0.48

or r_{YZ} is spurious (or only indirect)

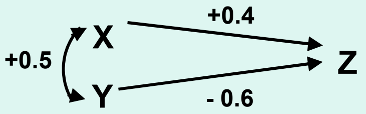


r_{YZ} = 0 + (0.5)(0.4) = +0.2

Many other possibilities and variations……..

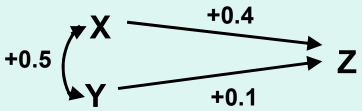
Note: a "spurious" association does <u>not</u> mean it is not real! It can still have e.g., clinical utility for prediction.

Example of **Suppression Effect**. (r_{XZ} and r_{YZ} suppressed in absolute value, e.g., if X and Y are compensatory biological processes).



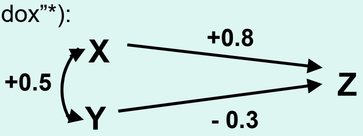
r_{XZ} = 0.4 + (0.5)(-0.6)= 0.4 - 0.3 = 0.1 **r_{YZ} = - 0.6 + (0.5)(0.4) = -0.6 + 0.2 = -0.4**

Or a weak underlying causal effect can have its correlation **enhanced**:



r_{YZ} = 0.1 + (0.5)(0.4)= 0.1 + 0.2 = 0.3

Or the sign of an underlying causal effect can be **reversed** (as in "Simpson's Paradox"*):

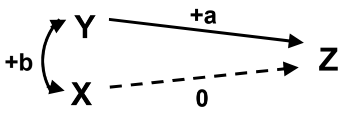


r_{YZ} = -0.3 + (0.5)(0.8)= -0.3 + 0.4 = +0.1

***see also "Lord's Paradox" (Pearl, 2018)**

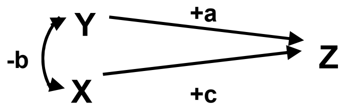
<u>Rules using diagrams </u>above can be:
(1) <u>proved with "covariance algebra</u>", or
(2) <u>demonstrated with graphical illustrations (e.g., 3D graphs)</u>.

Example: Spurious Association



$r_{XY} = b$ $r_{YZ} = a$ $r_{XZ} = ba$ (spurious association)

Example: Suppression Effect

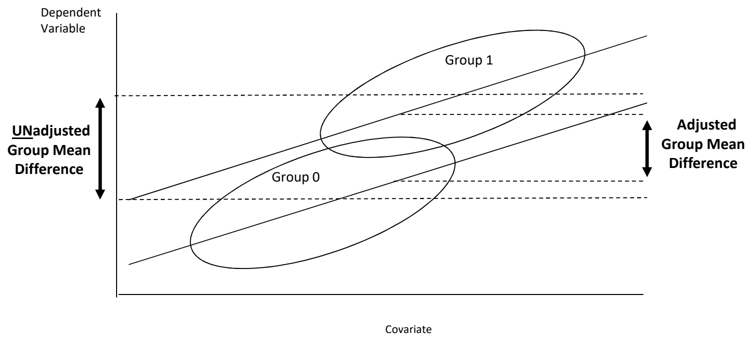


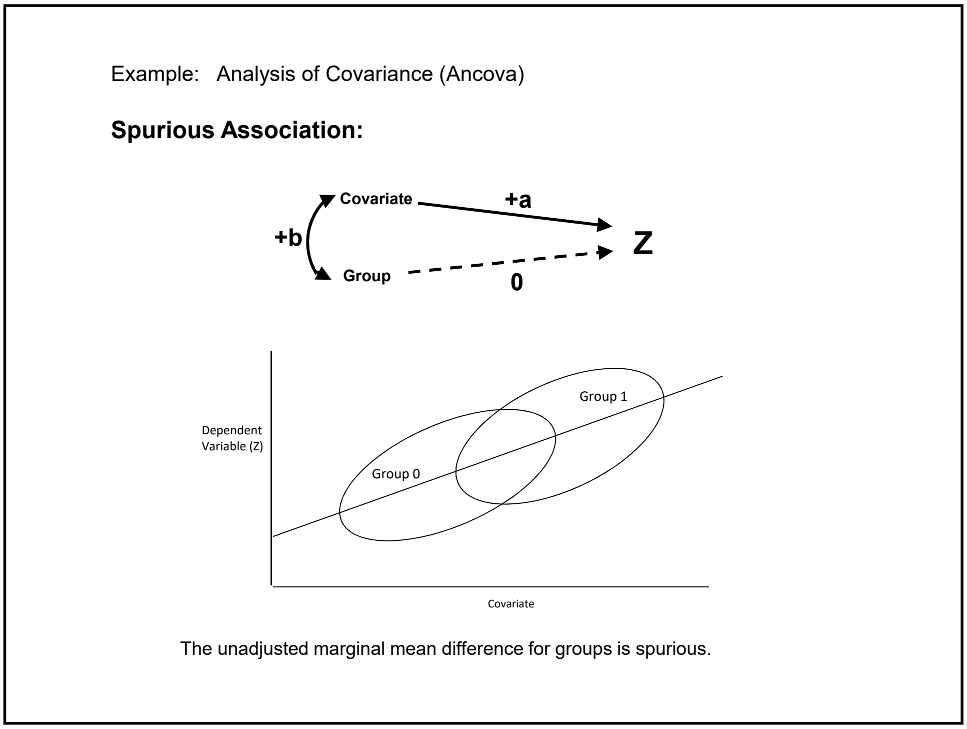
$r_{XY} = -b$ $r_{YZ} = a - bc$ (suppression) $r_{XZ} = c - ba$ (suppression)

Path analysis(causal modeling) <u>principles</u> are not specific to correlation/regression. Apply to many analysis methods.

Example: Analysis of Covariance (Ancova)

Graphical representation of Ancova with linear covariate:
(ellipses = point scatters per group; solid diagonal lines = regression lines per group).

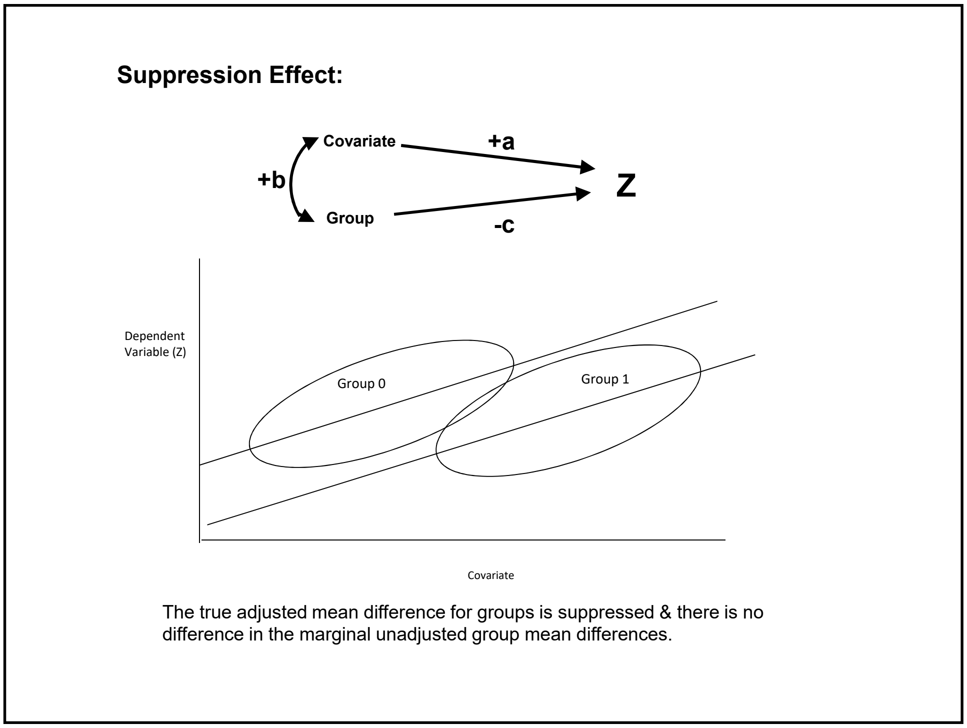

Dependent Variable
Group 1
UNadjusted Group Mean Difference
Adjusted Group Mean Difference
Group 0
Covariate



Example: Analysis of Covariance (Ancova)

Spurious Association:

The unadjusted marginal mean difference for groups is spurious.



Suppression Effect:

The true adjusted mean difference for groups is suppressed & there is no difference in the marginal unadjusted group mean differences.

Can do Piece-meal causal network modeling:

For example,

One part of the model is assessed with multiple regression,

another part just needs a correlation,

another portion needs logistic regression because the endogenous variable in that region is binary,

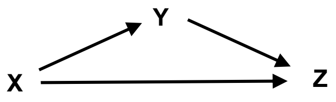
another factorial anova,

another ancova,

another portion is a multiple indicator measurement model & needs a confirmatory factor analysis,

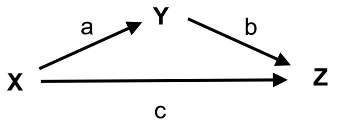
etc., etc.

Direct, Indirect, and Total Causal Effects assessed.



X has direct effect on Y, an indirect effect on Z via Y, and an additional direct effect on Z. The combined indirect and direct effects = total effect of X on Z. E.g., a toxin in brain (X) might directly kill neurons (Z), but also adversely affect vascular system (Y) and the vascular deficiency (Y) then also kills neurons (Z).

Mediation/Indirect Effects, such as above:



(a,b,c are path coefficients).

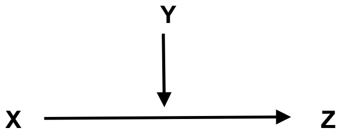
Y mediates the causal impact of X on Z. X may also have an additional direct effect on Z (via c).

Correlation of X & Y = a .
b and c are partial regression coefficients in multiple regression of Z on X & Y.
Direct effect of X on Z = c .
Mediated (indirect) effect of X on Z = a x b
Total Effect of X on Z = c + (a x b) . [by tracing rule]

Can rule out various effects as nonsignificant by correlating X vs Y to assess a, and running multiple regression of Z on X & Y to assess b and c.

There are many variations of “mediation effect analyses”, but all rest on above.

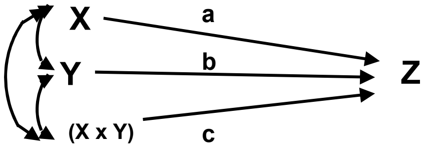
Moderated Effect as contrasted with Mediated Effect:



Here Y “moderates” or “modulates”, not “mediates” the causal impact of X on Z .

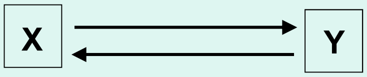
Y interacts with X in its effect on Z, i.e., the effect of X on Z varies depending on the level of Y.

Could view this as:



Steps in Doing Formal Path Analysis (or SEM)

1. Compute Correlations among Observed Variables – **Correlation Matrix**
2. Build a **Causal Model** (make a path diagram)
3. Ensure **Identification**. "Identification" roughly means you don't have too many arrows relative to the number of actual correlations to estimate them, e.g., reciprocal causal model below has 2 arrows but only 1 correlation, so it's under-identified.



4. Estimate **Path Coefficients** – depending on the model, using ordinary least squares regression, maximum likelihood, etc.
5. Assess **Fit** of Overall Model & Statistical Significance of each Coefficient
6. If fit is bad, possibly modify model (with "**modification indices**"), but don't overdo this and capitalize on chance.

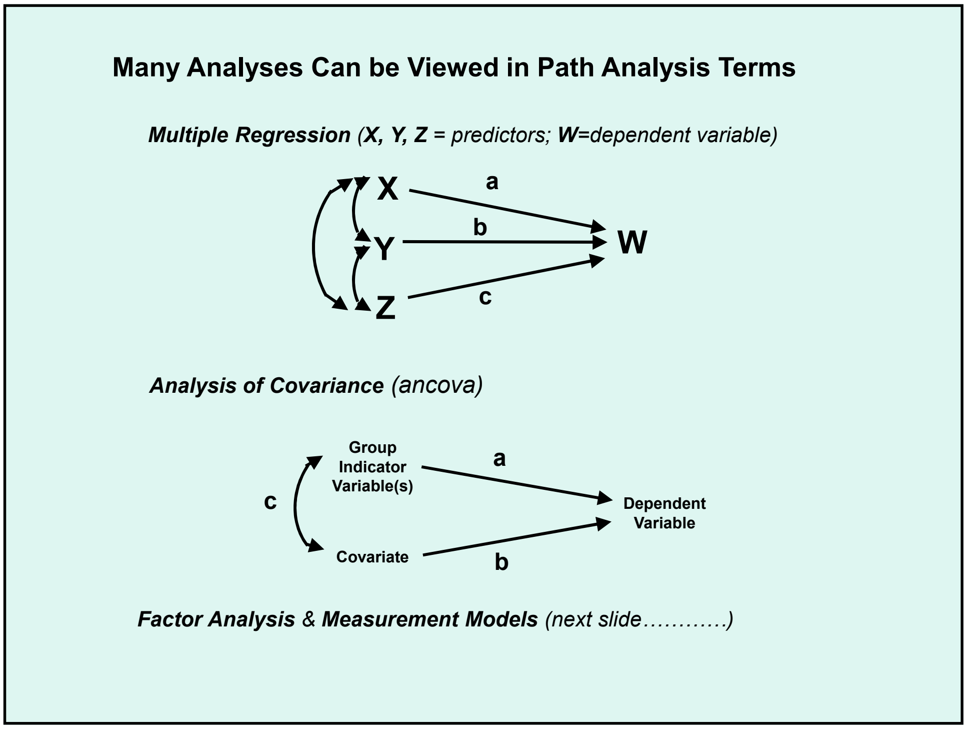
Identification of Model and Assessment of Fit

- **Unidentified/Under-Identified**: Not enough correlations relative to number of arrows. (Infinite number of path coefficient estimates fit model).
- **Just-Identified:** Number correlations just enough to estimate path coefficients.

  Perfect fit of model. Multiplying path coefficient estimates according to tracing rules reproduces correlation matrix exactly.
- **Over-Identified:** More than enough correlations to estimate path coefficients.

  Relative **fit of various causal models** to correlations can be assessed. Multiplying path coefficient estimates according to various rules generally do not reproduce correlation matrix exactly.

- **Fit of Overall Model** -- Assess fit of reproduced correlation (or covariance) matrix to actual matrix with chi-square test (want nonsig. result) and other fit indices that take into account model complexity, number of parameters, residuals, likelihood function value, etc. Significance of each path coefficient assessed.

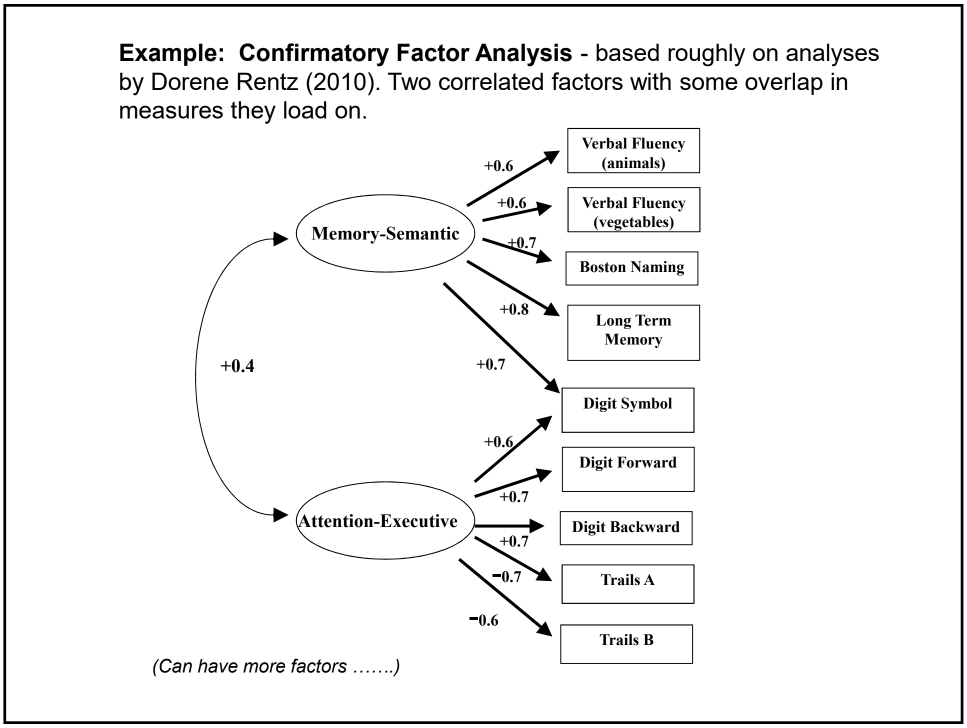


Many Analyses Can be Viewed in Path Analysis Terms

***Multiple Regression** (**X, Y, Z** = predictors; **W**=dependent variable)*

***Analysis of Covariance** (ancova)*

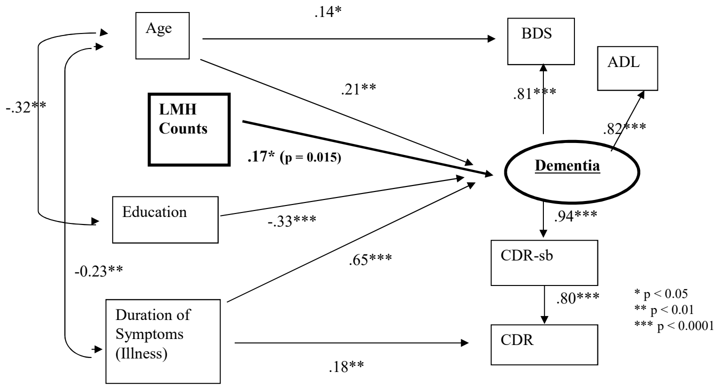
***Factor Analysis** & **Measurement Models** (next slide…………)*



Example: Confirmatory Factor Analysis - based roughly on analyses by Dorene Rentz (2010). Two correlated factors with some overlap in measures they load on.

(Can have more factors …….)

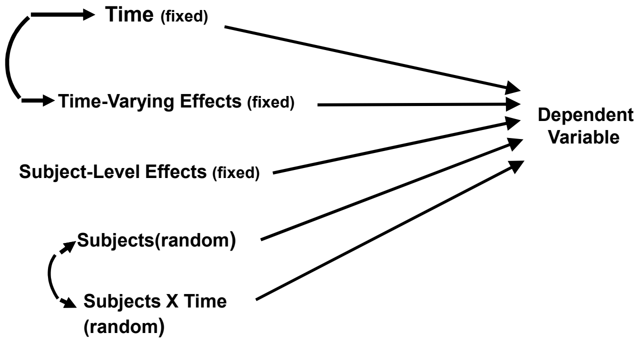
Example: Lobar Micro-hemorrhages Cause Dementia? (Atri et al., 2005)

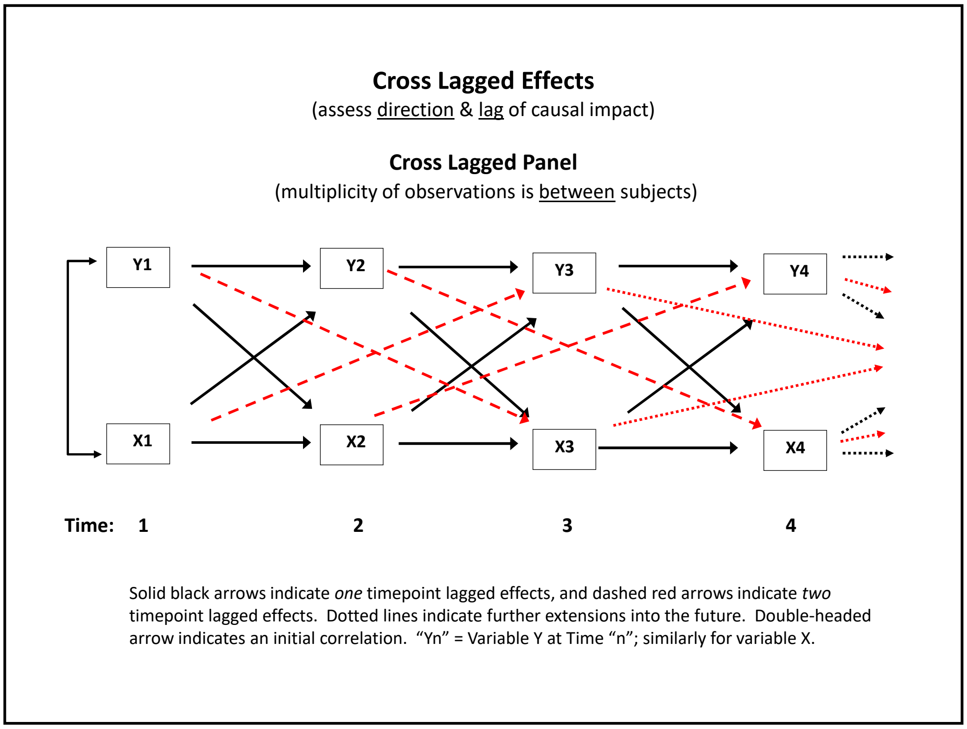


More Lobar Micro-hemorrhages (LMH) appear to increase severity of dementia, a latent variable measured by Blessed Dementia Scale (BDS), Activities of Daily Living (ADL) and Clinical Dementia Rating Scale (CDR), taking into account demographic & clinical covariates. Note: a measurement sub-model is part of the overall path model.

Some Longitudinal Examples

Mixed Effects Longitudinal Model





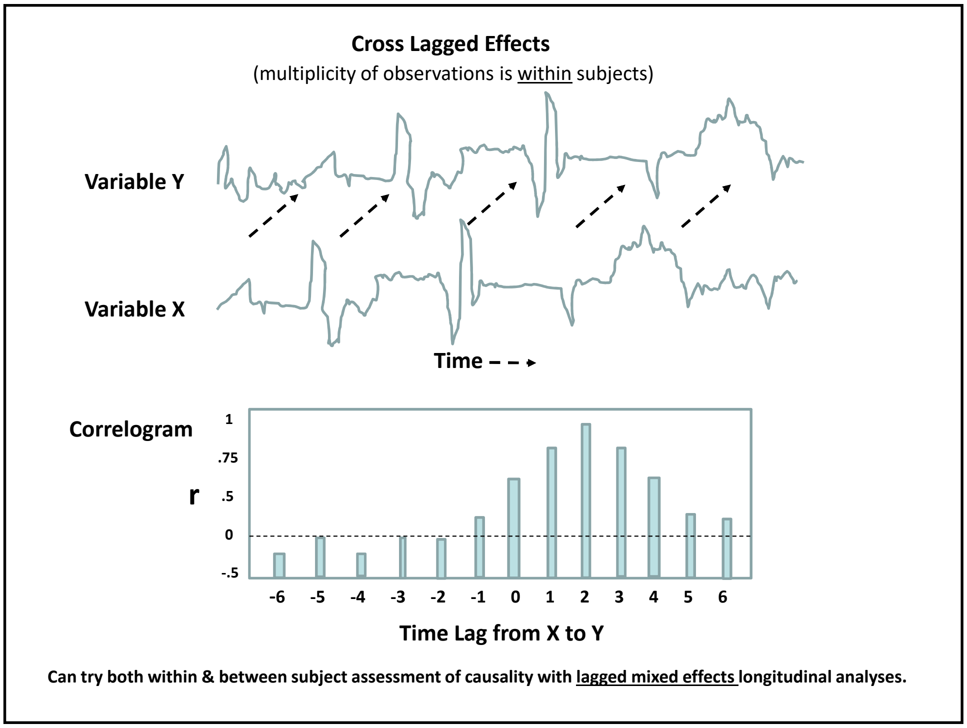
Cross Lagged Effects

(assess direction & lag of causal impact)

Cross Lagged Panel

(multiplicity of observations is between subjects)

Solid black arrows indicate *one* timepoint lagged effects, and dashed red arrows indicate *two* timepoint lagged effects. Dotted lines indicate further extensions into the future. Double-headed arrow indicates an initial correlation. "Yn" = Variable Y at Time "n"; similarly for variable X.

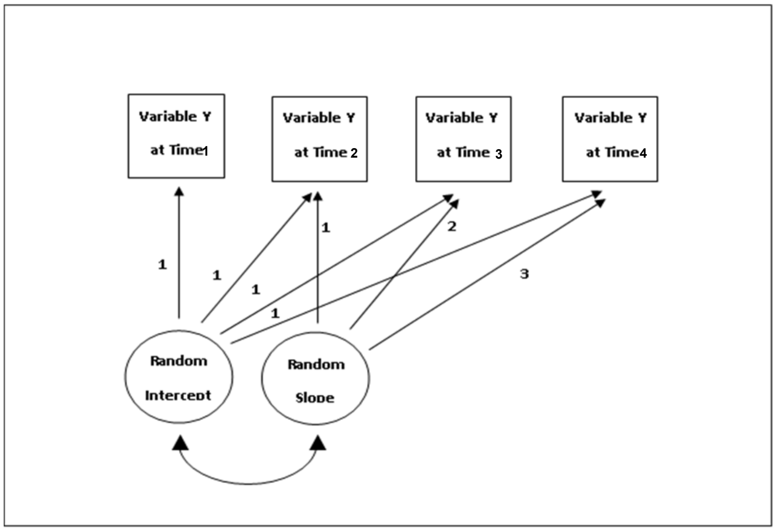


Cross Lagged Effects

(multiplicity of observations is within subjects)

Can try both within & between subject assessment of causality with lagged mixed effects longitudinal analyses.

Latent Growth Curve Model



Path Analysis Software

- **SAS Calis Procedure** (**Proc Calis**, Covariance Analysis of Linear Structures) or implicitly with Proc Reg for multiple regression, Proc GLM for general linear models & Ancova, Proc Factor for factor analysis, and others
- **JMP** interactive software (SAS affiliated) – you can draw path diagrams and have them estimated.
- **LISREL** (Linear Structural Relations; Karl Joreskog, Sweden)
- **MPlus** (Muthen et al.)
- **EQS** (originally from BMDP?)
- **Amos** (IBM-SPSS)
- I'm sure **R** has path analysis/SEM procedures.
- Others?

Some Criticisms/Limitations of Path Analysis

- Trying to prove causality from correlations & passive, observational data is “black magic”. Need a true randomized experiment.
  - Correct, but one can probabilistically rule out various competing causal models as relatively consistent or inconsistent with the data.
- Sometimes **abused** – refitting & over-fitting models via “modification indices” until a good fit is achieved by capitalization on chance, e.g., stipulating correlated residual terms and causal effects in parts of the model which don’t make substantive sense, in order to achieve an acceptable statistical fit.
  - Widely cautioned against. This is a general problem in statistical inference.
- **“Potential Outcomes”/“Counterfactuals”** or “Rubin’s Causal Model” – path models do not usually consider this explicitly.
  - Pearl (2009) – this are just specific instantiations of a SEM.

Some Useful References in Path Analysis and Related

SEM/Causal Modeling/Path Analysis/Finite Mixture Models and Latent Growth Curves

Asher, Herbert B. *Causal Modeling*. 2nd Edition. *Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1983.

Bollen, Kenneth A. *Structural Equations with Latent Variables*. NY, NY: John Wiley & Sons, 1989.

Cook T.D. and Campbell, D.T. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago, Illinois: Rand McNally College Publishing Company, 1979.

Davis, James A. *The Logic of Causal Order. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1985.

Duncan, Otis D. *Introduction to Structural Equation Models*. NY, NY: Academic Press, Inc., 1975.

Hatcher, Larry. *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute, Inc., 1994.

Kenny, David A. *Correlation and Causality*. NY, NY: John Wiley & Sons, 1979.

Locascio, Joseph J., Lee, Jarchi. and Meltzer, Herbert Y. The importance of adjusting for correlated concomitant variables in psychiatric research. *Psychiatry Research*, 1988, 23, 311-327.

Long J. Scott. *Covariance Structure Models: An Introduction to LISREL. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1983.

Lubke, G.H. & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21-39.

McCutcheon, Allan L. *Latent Class Analysis. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1987.

MPlus Software. Muthen, Bengt and Muthen, Linda. Version 5. Los Angeles, CA, 2009.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling. *Psychological Methods*, 8, 369-377.

Muthen, B. O. and Curran, P.J. General Longitudinal Modeling of Individual Differences in Experimental Designs: A Latent Variable Framework for Analysis and Power Estimation. *Psychological Methods*, 1997, Vol 2(4) 371-402.

Muthen, B. et al. General Growth Mixture Modeling for Randomized Preventive Interventions, *Biostatistics*, 2002, Vol 3(4), 459-475.

Pearl, J. Causal inference in statistics: an overview. *Statistics Surveys*, 2009, Vol 3, 96-146.

Pearl, J., *The Book of Why*, co-authored with Dana Mackenzie, Basic Books, NY, NY, 2018.

SAS/STAT Software: *SAS STAT User's Guide.* Version 9.3. Introduction to Structural Equation Models with Latent Variables. The CALIS Procedure. Cary, NC: SAS Institute, Inc., 2011.

Factor Analysis

Gorsuch, Richard L. *Factor Analysis*. Philadelphia, PA: W.B. Saunders Co., 1974.

Harman, Harry H. *Modern Factor analysis*. Chicago, Ill.: The University of Chicago Press, 1976.

Kim Jae-On and Mueller, Charles W. *Introduction to Factor Analysis. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1978.

Kim Jae-On and Mueller, Charles W. *Factor Analysis: Statistical Methods and Practical Issues. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1978.

Long J. Scott. *Confirmatory Factor Analysis. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1983.

SAS/STAT Software: *SAS STAT User's Guide.* Version 9.3. The Factor Procedure. Cary, NC: SAS Institute, Inc., 2011.

Cluster Analysis

Aldenderfer, Mark S. and Blashfield, Roger, K. *Cluster Analysis. Quantitative Applications in the Social Sciences*. Thousand Oaks, California: Sage Publications, Inc., 1984.

SAS/STAT Software: *SAS STAT User's Guide.* Version 9.3. Introduction to Clustering Procedures. The Cluster Procedure. The FastClus Procedure. The ModeClus Procedure. Cary, NC: SAS Institute, Inc., 2011.

Power Analysis

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Ed. Lawrence Erlbaum Associates: Hillsdale, NJ 1988.

Multiple Regression

Cohen, J. and Cohen, P. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd Ed. Lawrence Erlbaum Associates: Hillsdale, NJ 1983.