Introduction to Structural Equation Modeling

Notes Prepared by:

Lisa Lix, PhD

Manitoba Centre for Health Policy

Topics

- Section I: Introduction
- Section II: Review of Statistical Concepts and Regression
- Section III: Path Analysis
- Section IV: Confirmatory Factor Analysis
- Section V: Structural Regression
- Section VI: Wrap-Up

Topics Not Discussed

- Data preparation:
  - Missing data
  - Multivariate normality
  - Linear associations among variables
  - Collinearity
- Specialized SEM analyses such as:
  - Multi-group comparisons
  - Latent growth curve models
  - Analyses involving categorical and/or dichotomous variables

Learning Outcomes

- Workshop participants will be able to:
  - Understand the theory underlying SEM;
  - Describe the differences between path analysis, confirmatory factor analysis, and structural regression analysis;
  - Describe potential applications of each technique in the health and behavioral sciences;
  - Work through the steps of structural equation modeling to analyze a covariance or correlation matrix using LISREL.

Section I: Introduction

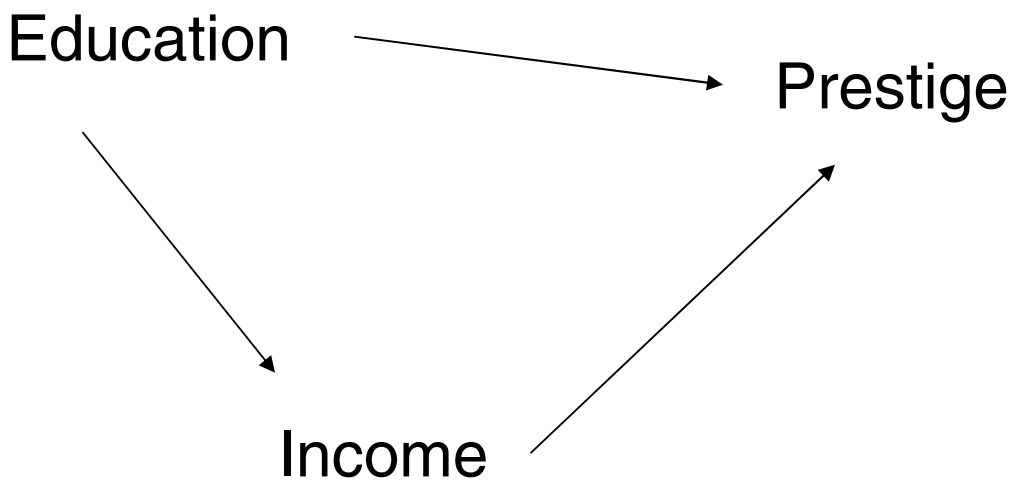
- What is SEM?
- Other names for SEM
- Why is SEM increasing in popularity?
- Key terms and notation
- The SEM core family
- Steps in SEM
- Software choices
- Resources

What is SEM?

- A statistical model that allows researchers to simultaneously investigate multiple dependent variables and examine both the direct and indirect effects of explanatory variables.

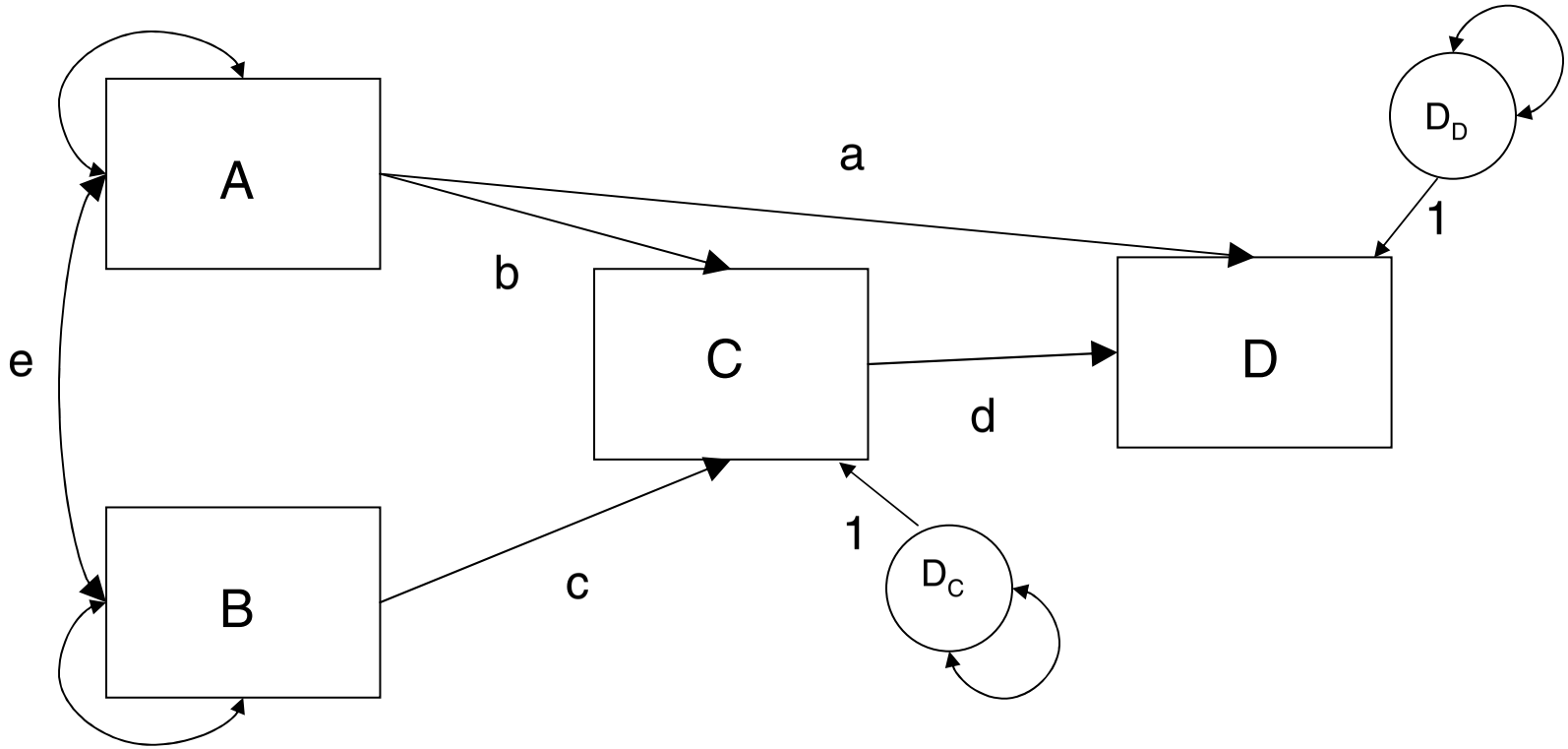
- Can be visually represented using a path diagram.
- Important note: Requires large sample size!

Example



Source: Fox, 1997

Path Diagram



What is SEM? (Hancock, 2005)

- A process that allows for the testing of competing theories that are hypothesized *a priori* to explain the correlations (or variances and covariances) among observed variables.

Other Names for SEM

- Latent variable analysis
- LISREL modeling
- Simultaneous equation modeling
- Covariance structure modeling
- Causal modeling

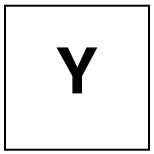
Why is SEM Increasing in Popularity?

- It can be used to analyze complex datasets containing multiple independent and dependent variables
- It has been extended to longitudinal/repeated measures data
- The diversity of applications is increasing
  - Clinical research
  - Genetics
  - Quality of life
  - Health services

Key Terms

- Exogenous and endogenous variables: Causes of exogenous variables are unknown. Endogenous variables are caused by other variables.
- Manifest and latent variables: A manifest variable is another name for a variable that is observed or directly measured. A latent variable is another name for a variable that cannot be directly observed or measured
- Direct and indirect effects: Direct effects are the effects of one variable on another variable. Indirect effects involve one or more intervening variables presumed to "transmit" some of the causal effects of prior variables onto subsequent variables.

Notation



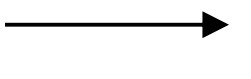
Observed/Manifest Variable

Unobserved/Latent Variable

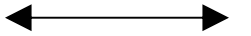
E Error

D Disturbance

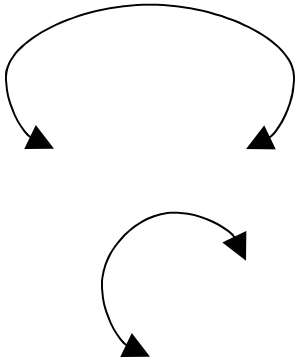
Notation



Uni-directional: Direct/causal effect



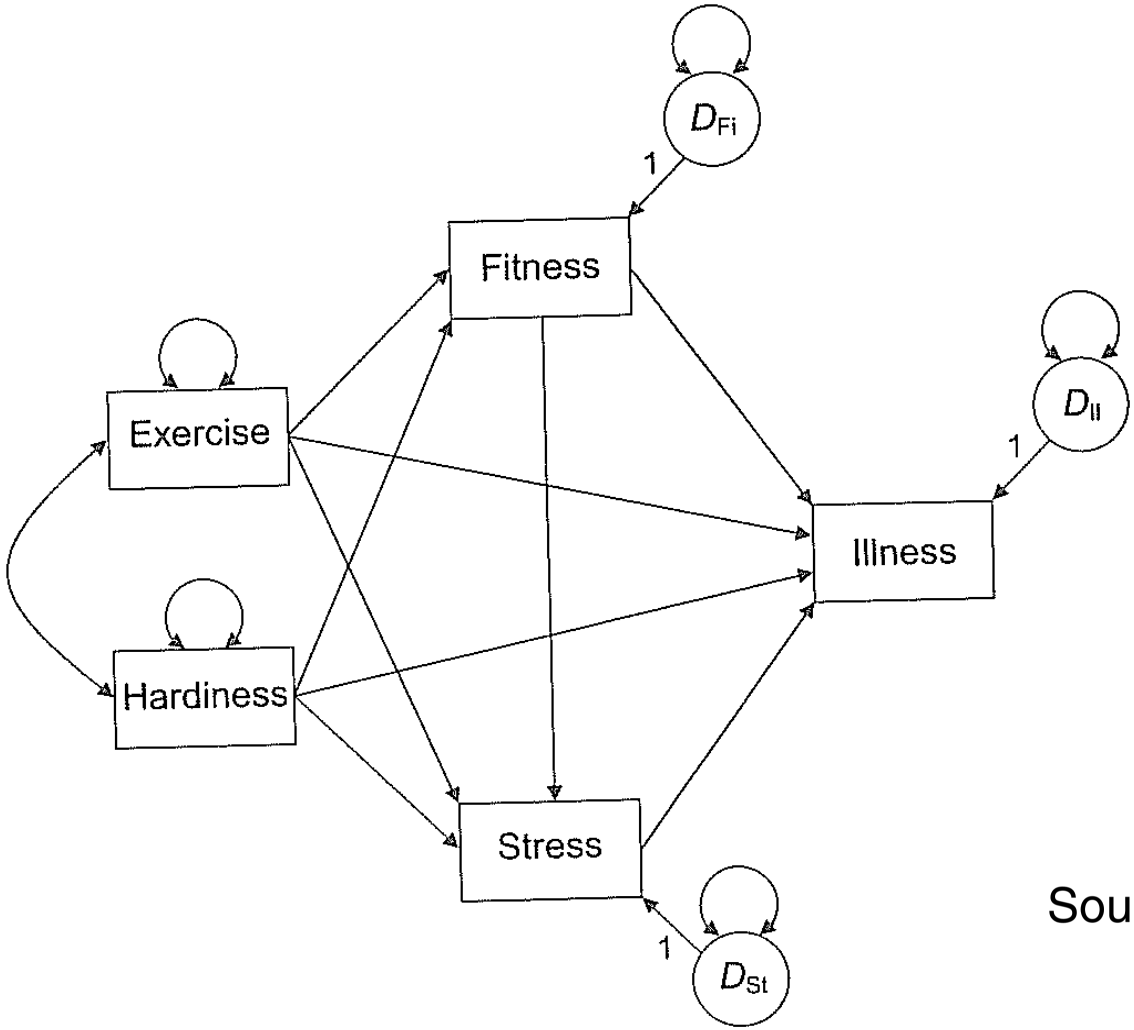
Bi-directional: Association/Non-causal effect/Covariation or Correlation



The SEM Core Family

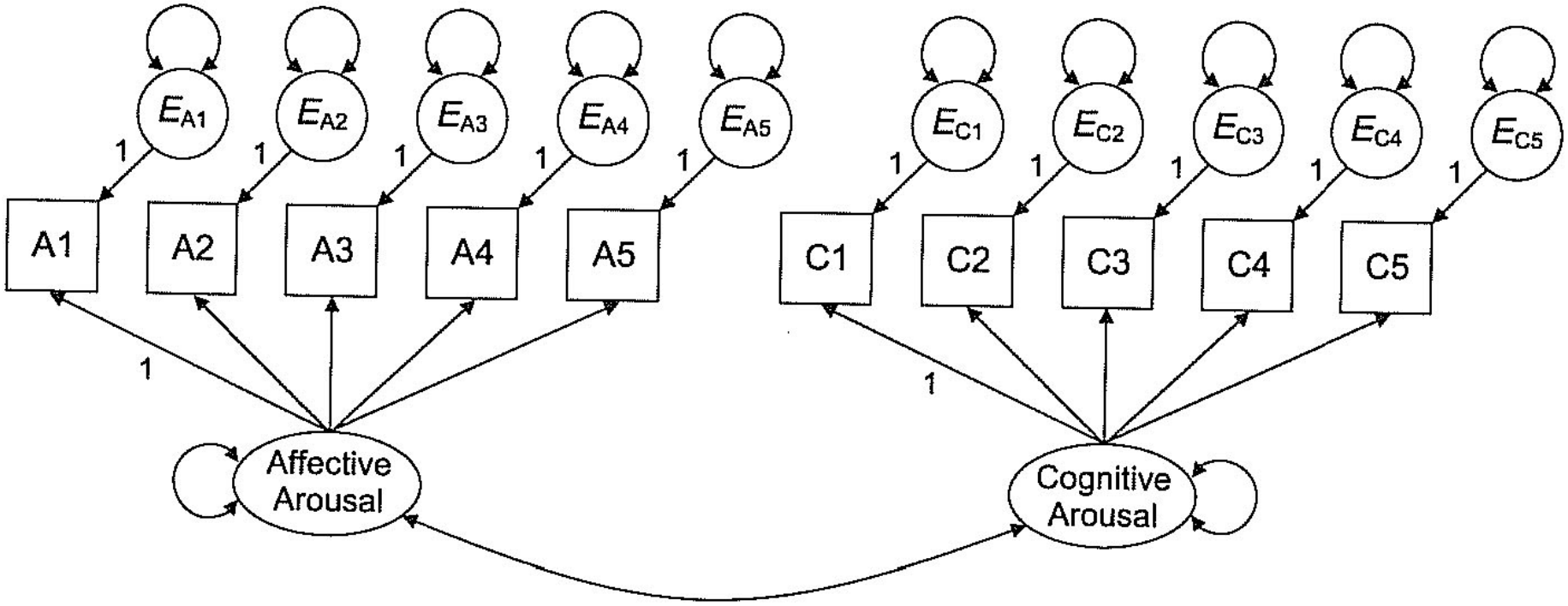
- Path analysis
- Confirmatory factor analysis
- Structural regression analysis

Example: Path Analysis



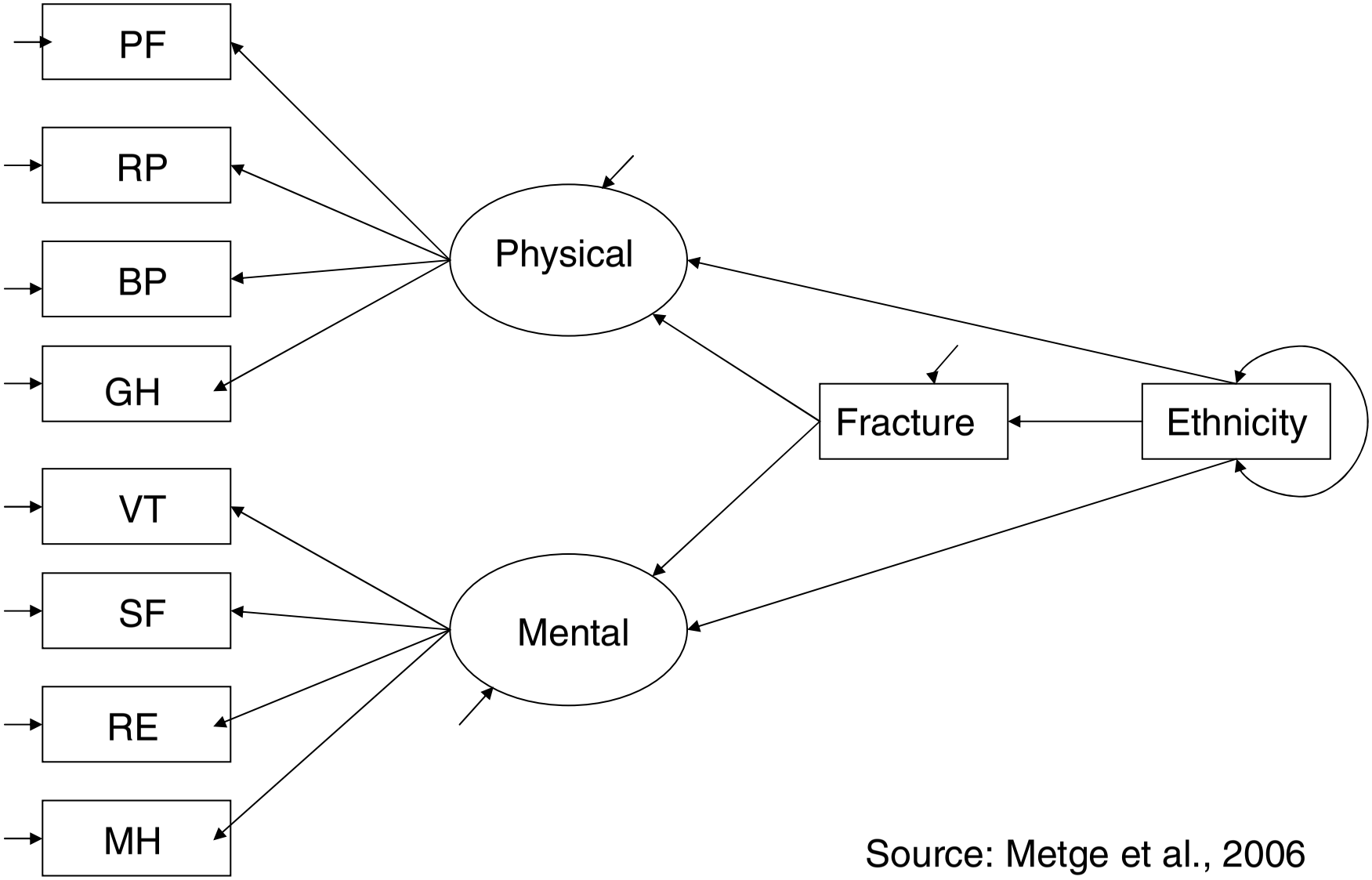
Source: Kline, 2005

Example: CFA



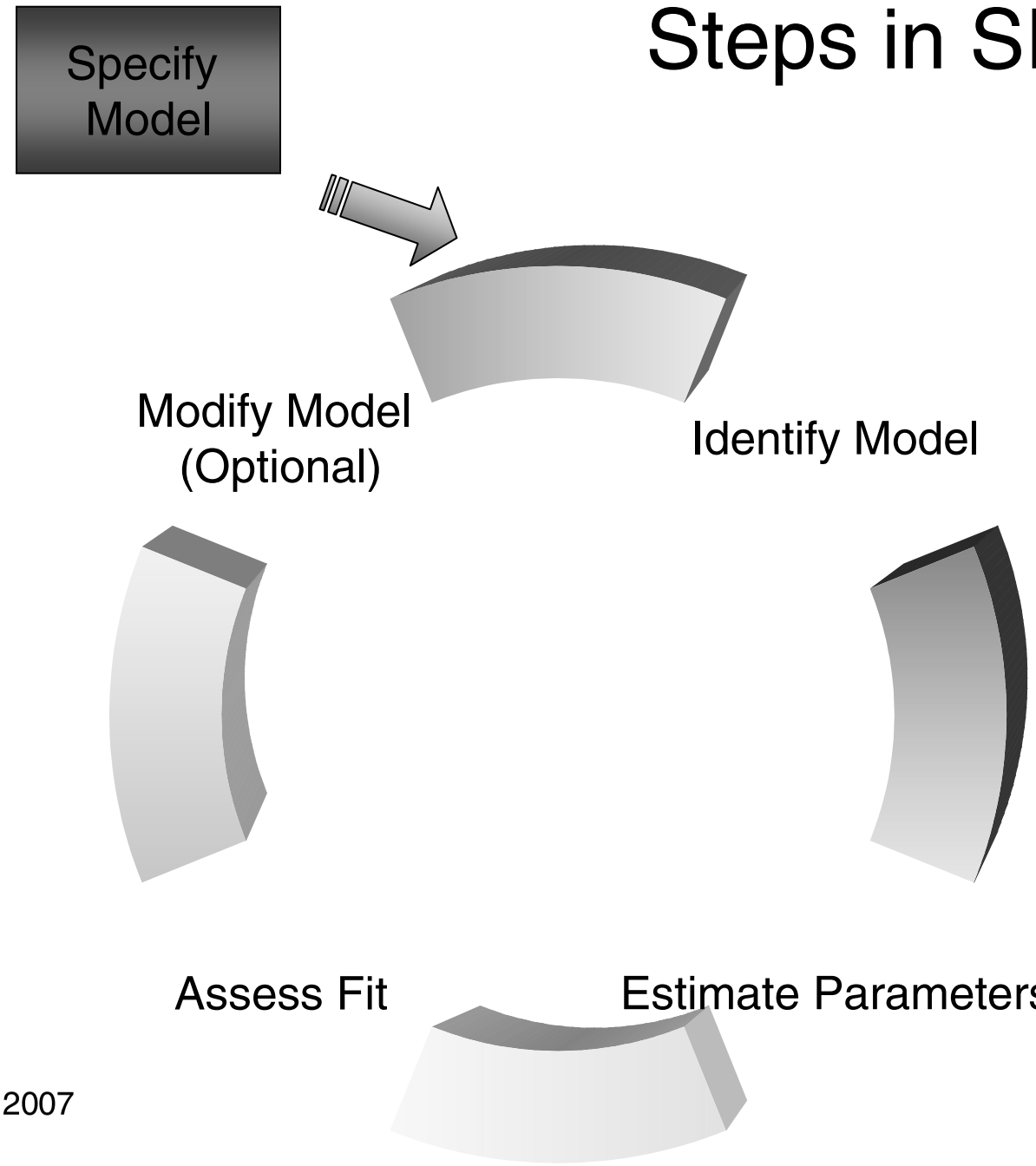
Source: Kline, 2005

Example: Structural Regression Model



Source: Metge et al., 2006

Steps in SEM


Specify Model
Identify Model
Estimate Parameters
Assess Fit
Modify Model (Optional)

Software Choices

- LISREL
- EQS
- AMOS
- Mplus
- PROC CALIS

Resources

- Books
- Articles
- Websites

Section II: Review of Statistical Concepts and Regression

- Standardized variables
- Correlation
- Covariance
- Standardized and unstandardized regression coefficients
- Partial regression coefficients
- Multiple regression

Example Dataset

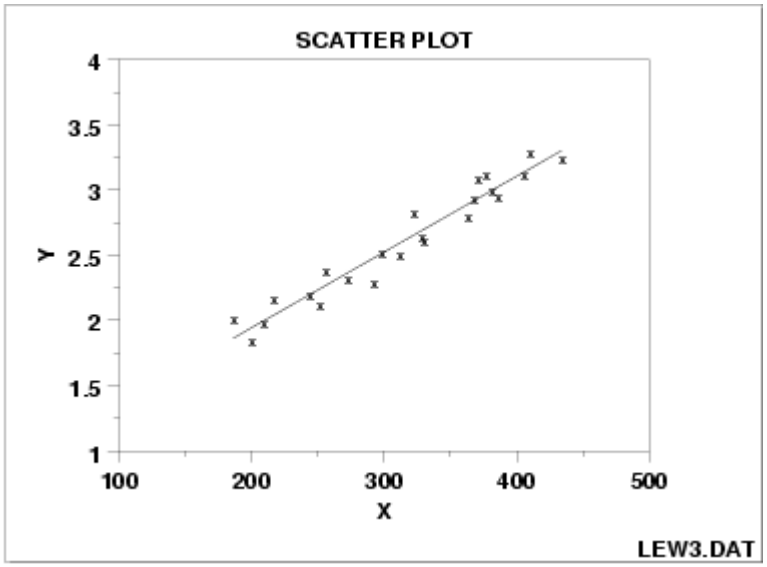
ID	X	Y
1	19.0	15.0
2	17.0	20.2
3	16.4	18.1
4	13.0	16.2
5	12.0	18.4
6	11.1	21.4
…	…	…

Standardized Variables

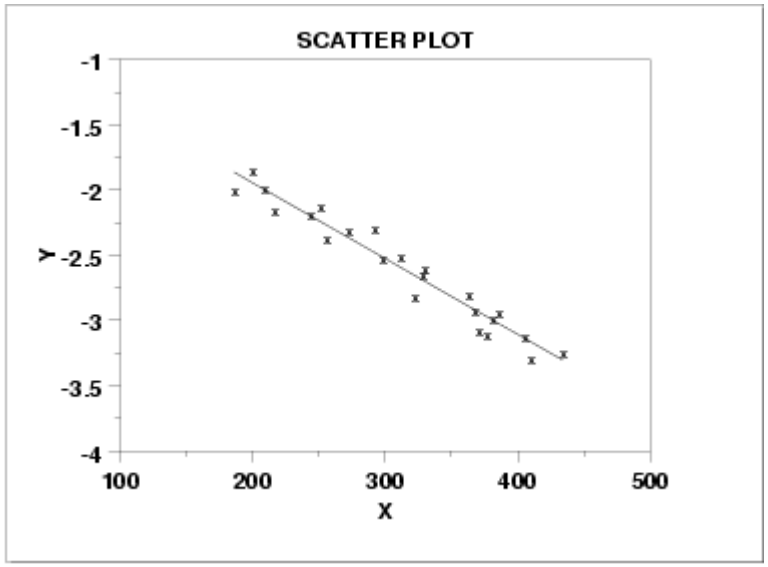
- Transformed to have a mean of 0 and a standard deviation of 1.0

$$Z_1 = \frac{X_1 - \bar{X}}{s_x}$$

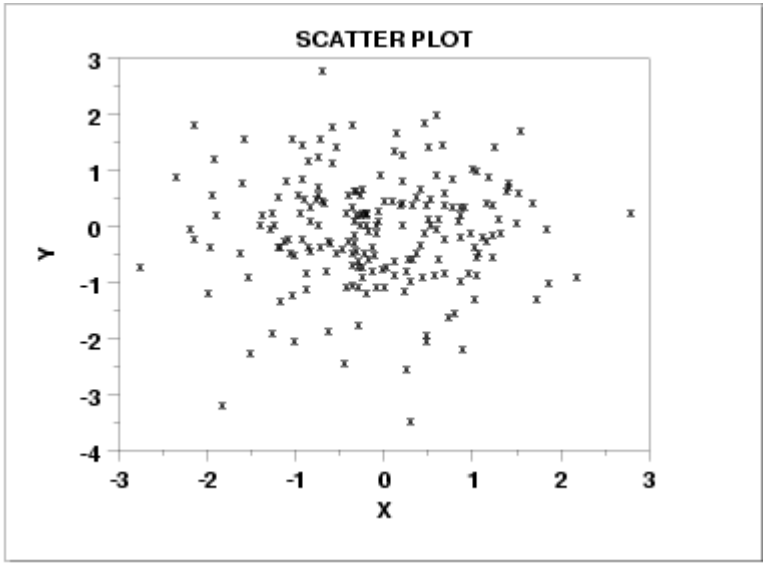
Correlation



r is a +ve number



r is a -ve number



r is close to zero

Correlation

Subject	*X*	*Y*
1	19.0	15.0
2	17.0	20.2
3	16.4	18.1
4	13.0	16.2
5	12.0	18.4
6	11.1	21.4
…	…	…

$r_{XY} = -.48$

$$\mathbf{R} = \begin{matrix} 1 & -.68 & .55 & .40 \\ & 1 & .30 & -.50 \\ & & 1 & .42 \\ & & & 1 \end{matrix}$$

$$r_{XY} = \frac{Z_X Z_Y}{N-1}$$

$$\mathbf{S} = \begin{matrix} 4.5 & -3.3 & 2.1 & -4.6 \\ & 7.1 & 5.5 & 2.0 \\ & & 5.2 & 3.9 \\ & & & 6.1 \end{matrix}$$

$$\mathrm{cov}_{XY} = r_{XY}(s_X s_Y)$$

Standardized & Unstandardized Regression Coefficients

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = r_{XY} \frac{s_X}{s_Y}$$

For two standardized variables, the correlation between them is the standardized regression coefficient (i.e., $\hat{b}_1$)

Partial Regression Coefficients

- *W* can influence the observed correlation between *X* and *Y*.
- Partial correlation removes the effect of the third variable, *W*, from both *X* and *Y*.
- "Controlling for *W*" or "Holding *W* constant"

$$r_{XY \cdot W} = \frac{r_{XY} - r_{XW} r_{YW}}{\sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)}}$$

Multiple Regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$\hat{b}_1$ and $\hat{b}_2$ denote standardized multiple regression coefficients, also known as beta weights

$$\hat{b}_1 = \frac{r_{Y1} - r_{Y2} r_{12}}{1 - r_{12}^2}$$

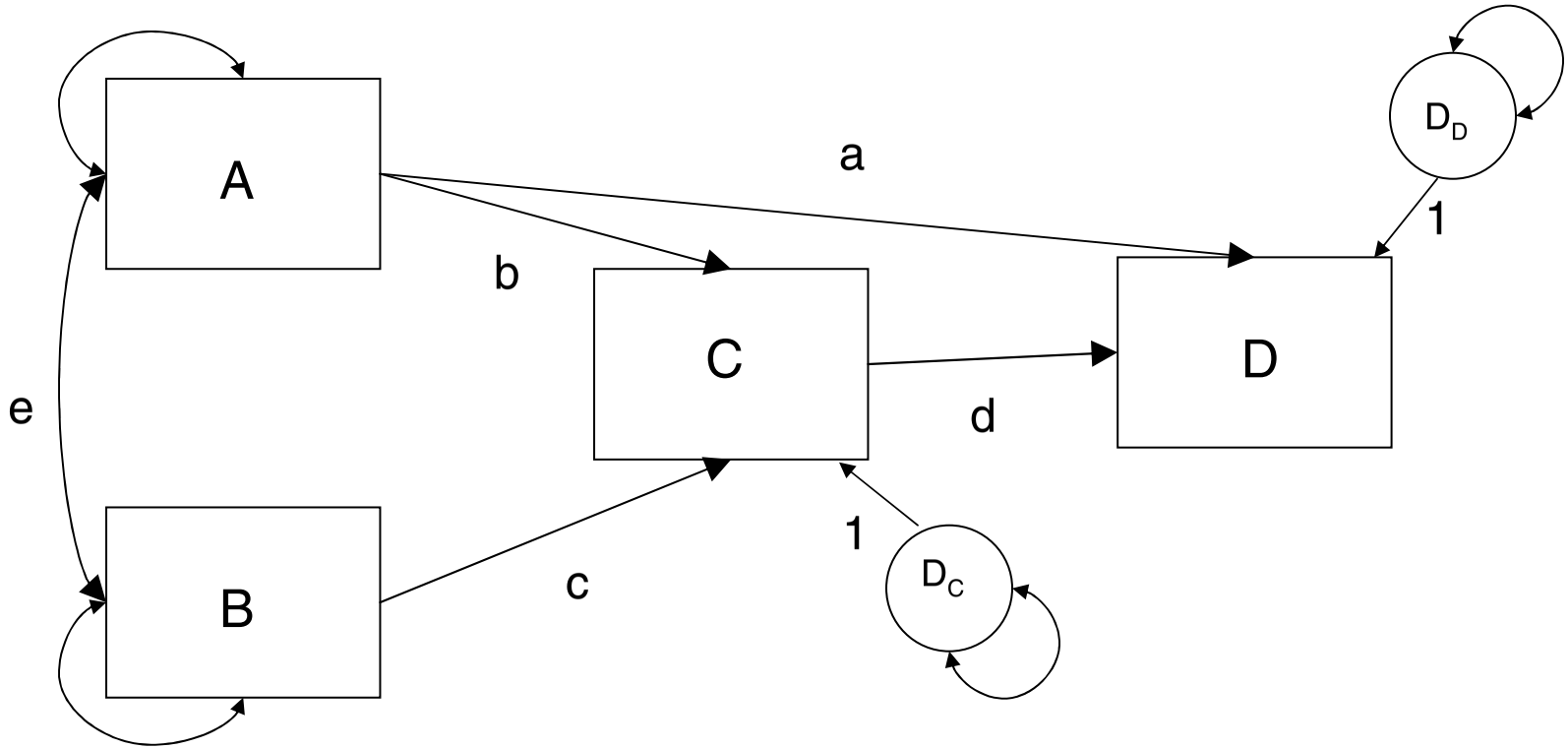
Section III: Path Analysis

- Path analysis or multiple regression?
- Review of model building blocks
- Model characteristics
- Model identification
- Estimation
- Assessing model fit
- Modifying and comparing models
- Equivalent models

Path Analysis or Multiple Regression?

- Multiple regression assumes that the predictors (i.e., exogenous variables) and residuals (i.e., disturbances) are uncorrelated.
- BUT, path analysis assumes that directionality of effects can be specified. What if they can't?
  - Forgo path analysis in favour of multiple regression
  - Specify and test alternate path models, each with different directional relationships
  - Include reciprocal effects (i.e., feedback loops)

Review of Model Building Blocks



Model Characteristics

- Recursive vs. non-recursive
  - Recursive: disturbances are uncorrelated and causal effects are unidirectional
  - Non-recursive: disturbances are correlated and contain feedback loops (not discussed in workshop)
  - Partially recursive: disturbances are correlated and causal effects are unidirectional (not discussed in workshop)

Model Characteristics

- Number of observations
  - Number of observations = number of variances and covariances among the observed variables (for unstandardized variables)
  - If v = number of observed variables then $v(v+1)/2$ = number of observations

Model Characteristics

- Model parameters
  - Number of model parameters = direct effects on endogenous variables from other observed variables and number of unanalyzed associations that are either observed or unobserved (i.e., disturbances)
  - Status of model parameters
    - **Free** to be estimated
    - **Fixed** to be a constant
    - **Constrained**: estimated with some restrictions, but not fixed to equal a constant; for example, a cross-group equality constraint means that a parameter is constrained to have the same estimate in two or more independent groups

Model Degrees of Freedom

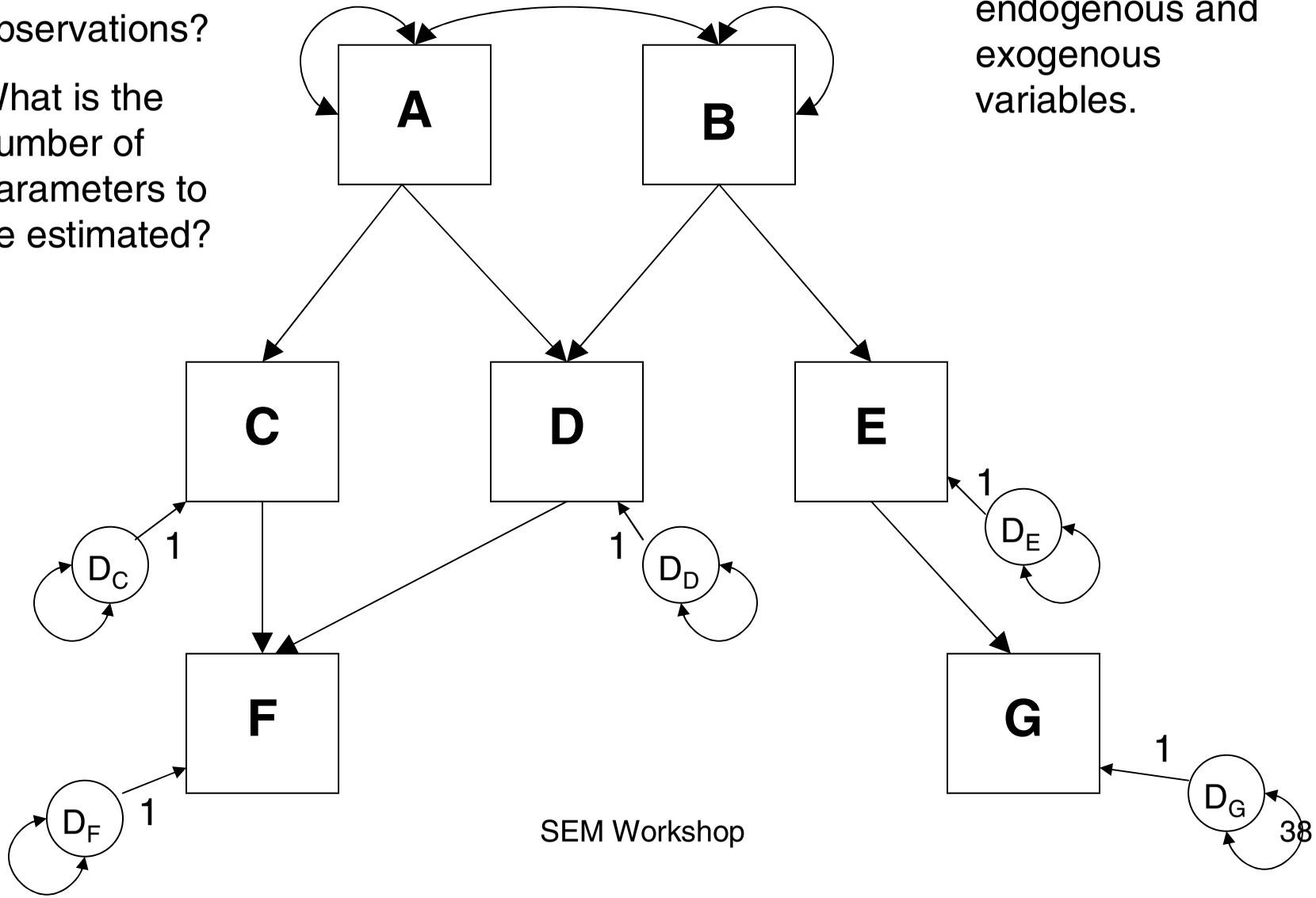
- Model degrees of freedom = # of observations minus # of parameters to be estimated
  - $df_M = v(v+1)/2 \quad - p$

Pop Quiz

What is the number of observations?

What is the number of parameters to be estimated?

Identify the endogenous and exogenous variables.



Model Identification

- A model is identified if it is theoretically possible to derive a unique estimate for each parameter.
- Requirements for identification:
  - $df_M \geq 0$
  - Every latent variable must be assigned a scale (i.e., metric)

Model Identification

- Under-identified models
  - Models that do not meet the requirements of identification
  - Empirically under-identified models may arise when there is a high degree of colinearity among variables

Model Identification

- Just-identified models
  - $df_M = 0$
  - Not very interesting!
- Over-identified models
  - $df_M > 0$
  - The goal in modeling!

Estimation

- Most SEM programs use maximum likelihood estimation
  - Parameter estimates maximize the likelihood that the data were drawn from the population
  - Normal theory method
  - An iterative process is used to solve the simultaneous set of equations implied by a path diagram

Assessing Model Fit

- Need to assess the fit of over-identified models
- Dozens of fit indices for SEM models have been developed
- The availability of so many indices is a problem – difficult to decide which ones to report

Model Fit Indices

- Recommended minimum set:
  - Model χ^2
  - Root mean square error of approximation (RMSEA) and 90% confidence intervals
  - Comparative fit index (CFI)
  - Standardized root mean square residual (SRMR)

Model Fit Indices

- Model χ^2
  - For a just-identified model: $\chi^2 = 0$, indicating that the model fits the data perfectly
  - For an over-identified model: $\chi^2 > 0$; it tests the null hypothesis that the model is correct (i.e., that is has perfect fit in the population)
  - Failure to reject the null hypothesis supports the researcher's model
  - Problems with $\chi^2 = 0$
    - Sensitive to sample size and magnitude of correlations
    - Tests an unlikely hypothesis

Model Fit Indices

- RMSEA
  - Parsimony-adjusted index, meaning it accounts for the complexity of the model
  - Measures error of approximation, the lack of fit of the researcher's model to the population covariance matrix, instead of error of estimation, the difference between the fit of of the model to the sample and population covariance matrix
  - Rule of thumb:
    - RMSEA ≤ .05: close approximate fit
    - .05 < RMSEA ≤ .08: reasonable error of approximation
    - RMSEA ≥ .10: poor error of approximation

Model Fit Indices

- RMSEA
  - Also look at lower and upper bounds of 90% confidence interval
  - Ideally lower bound is below .05 and upper bound is below .10

Model Fit Indices

- CFI
  - Incremental fit index, which means it quantifies the relative improvement in fit of the specified model over the baseline model
  - Baseline model = independence model (i.e., zero population covariances among observed variables)
  - Rule of thumb: CFI > .90 indicates reasonable fit of researcher's model

Model Fit Indices

- SRMR
  - Based on covariance residuals, which are the differences between the observed and predicted covariance values
  - Mean absolute value of covariance residuals
  - Rule of thumb: SRMR < .10 indicates a good fitting model

Modifying and Comparing Models

- Model trimming
  - Start with a just-identified model and eliminate paths
- Model building
  - Start with a basic over-identified model and add paths

Modifying and Comparing Models

- Model trimming or building can be conducted using theory as a guide, or empirical measures called **modification indices**.
- Modification indices: expressed as χ^2 statistics with one degree of freedom. Estimates the amount that the model χ^2 would decrease if a path were freely estimated. The bigger the modification index, the greater the improvement in model fit if the path were added to the model.

Modifying and Comparing Models

- For nested or hierarchical models, can compute a χ^2 likelihood ratio test (LRT) or difference test (i.e., χ^2_D)
- Two models are nested if one is a special case of the other; for example, trimming a model by removing one path results in nested models

Modifying and Comparing Models

- χ^2_D is the difference between the χ^2 values of two nested models
- It's degrees of freedom equals the difference in degrees of freedom for the two nested models
- The LRT tests the null hypothesis of identical fit of the two nested models in the population

Equivalent Models

- Two or more models that have the same numeric values for model fit indices, and same value of df_M but have a different configuration of paths among the same set of observed variables.
- A researcher needs to be able to justify why his/her model should be preferred over mathematically equivalent models.

Summary of Path Analysis

- Path analysis allows for the testing of causal theories about manifest variables.
- Covariances among observed variables can be decomposed into direct, indirect, and non-structural elements.
- A path model must be identified to estimate model parameters.

Summary of Path Analysis

- When a model is over-identified, model fit can be assessed using one of several fit indices
- Differences in fit can be tested for competing models

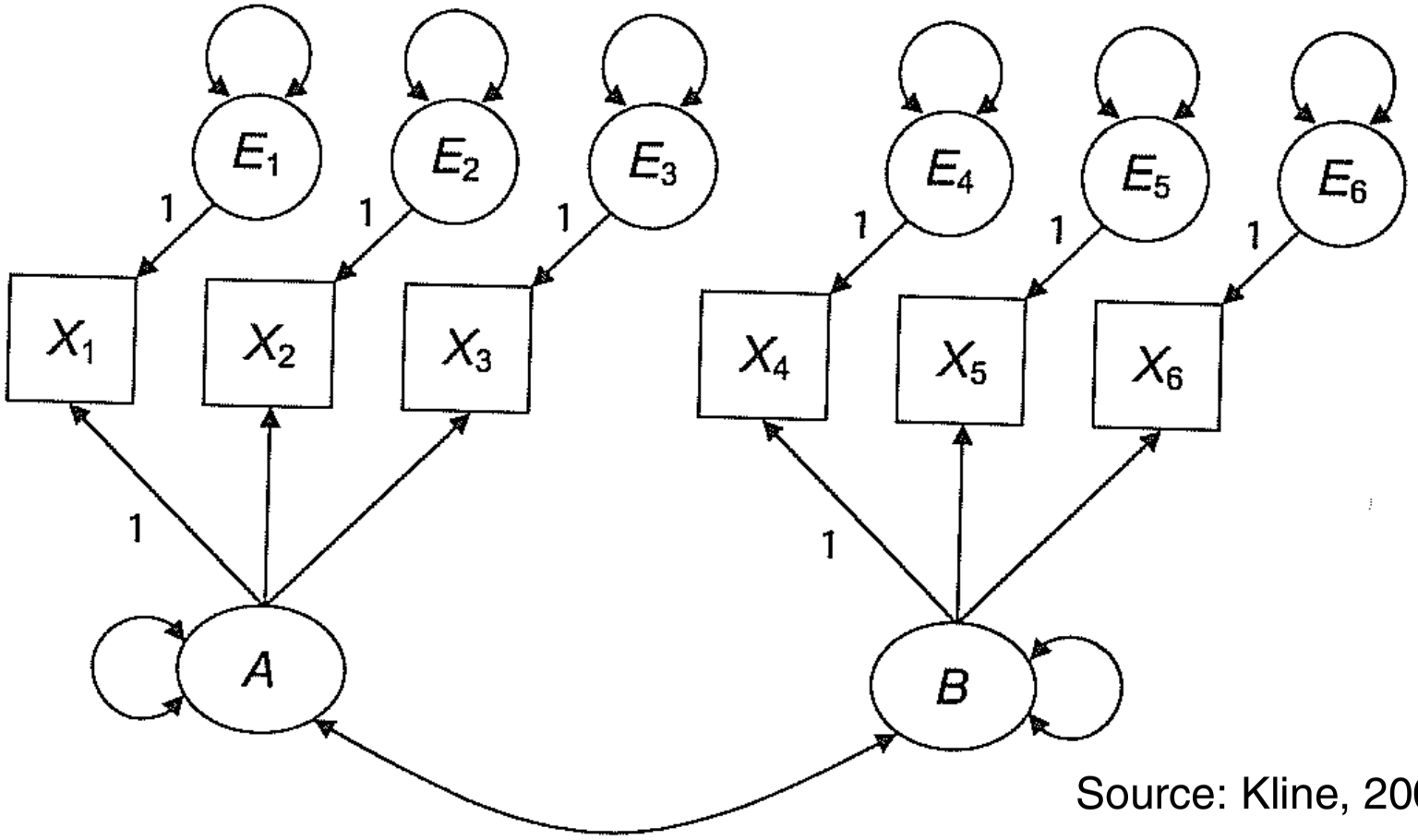
Section IV: Confirmatory Factor Analysis

- Exploratory vs. confirmatory factor analysis
- Review of model building blocks
- Model characteristics
- Model identification
- Assessing model fit
- Comparing models
- Modifying models

Exploratory vs. Confirmatory Factor Analysis

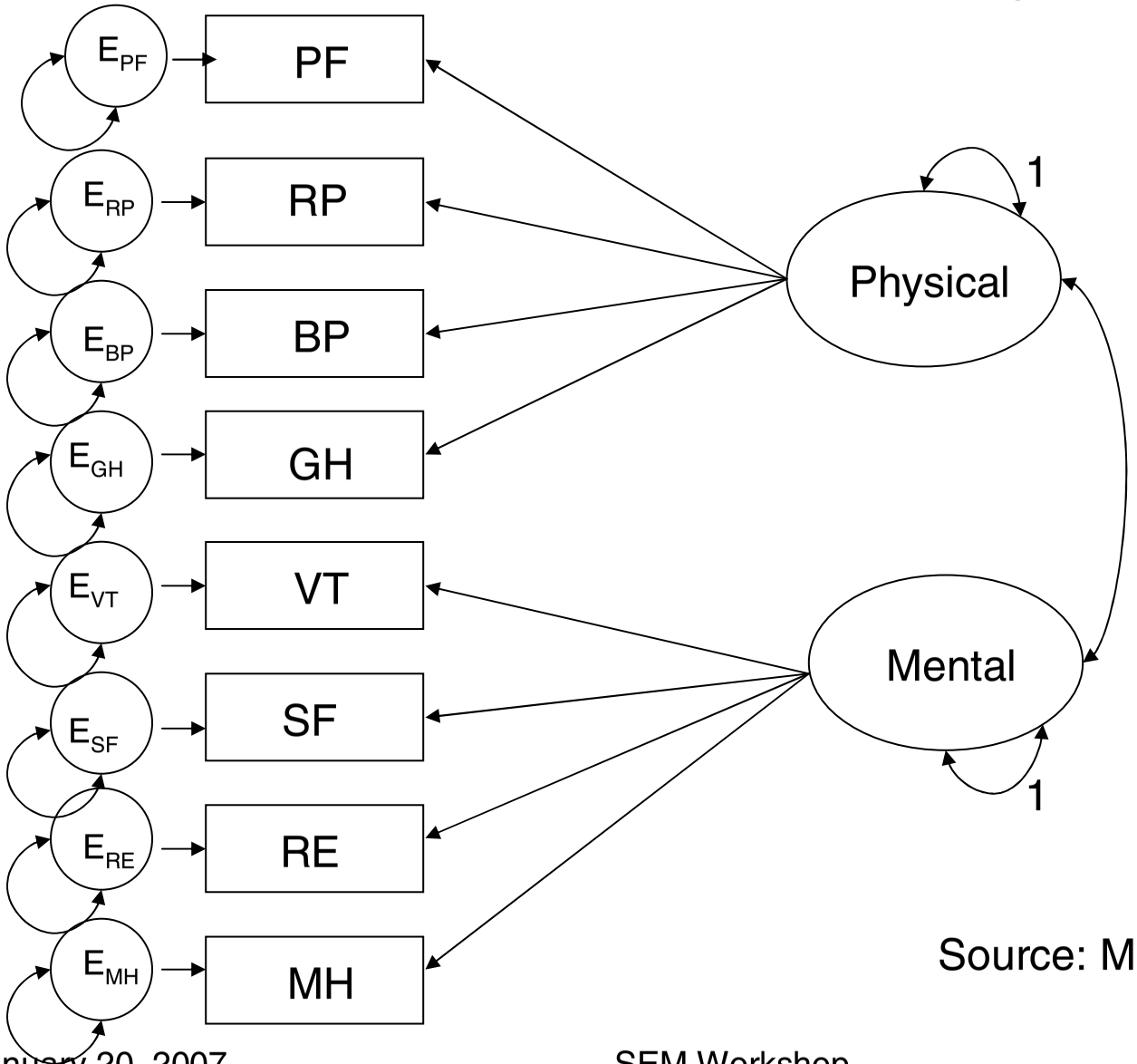
- EFA
  - Does not begin with an *a priori* model
  - Goal is to find a model that fits the data, and then to estimate the values of its paths and correlations
- CFA
  - Begins with an *a priori* theoretically derived model
  - Model fit is assessed to decide if the theory should be retained or rejected

Review of Model Building Blocks



Source: Kline, 2005

Review of Model Building Blocks



Source: Metge et al., 2006

Model Characteristics

- Unidimensional model
  - Each indicator depends on just one latent variable and error terms are independent
- Multidimensional model
  - Indicators can load on more than one latent variable or the error term of an indicator covaries with the error term of another indicator

Model Characteristics

- Model constraints
  - Measurement errors are almost always assigned a scale through a unit loading identification (ULI) constraint
  - Latent variables may be assigned a scale in one of two ways:
    - Unit load identification (ULI) constraint
    - Unit variance identification (UVI) constraint)
  - Both methods of scaling latent variables usually result in the same model fit; if the fit is different, the choice of a scale depends on goals of analysis; UVI constraint is easier to interpret, but generally is only appropriate when the latent variable is exogenous

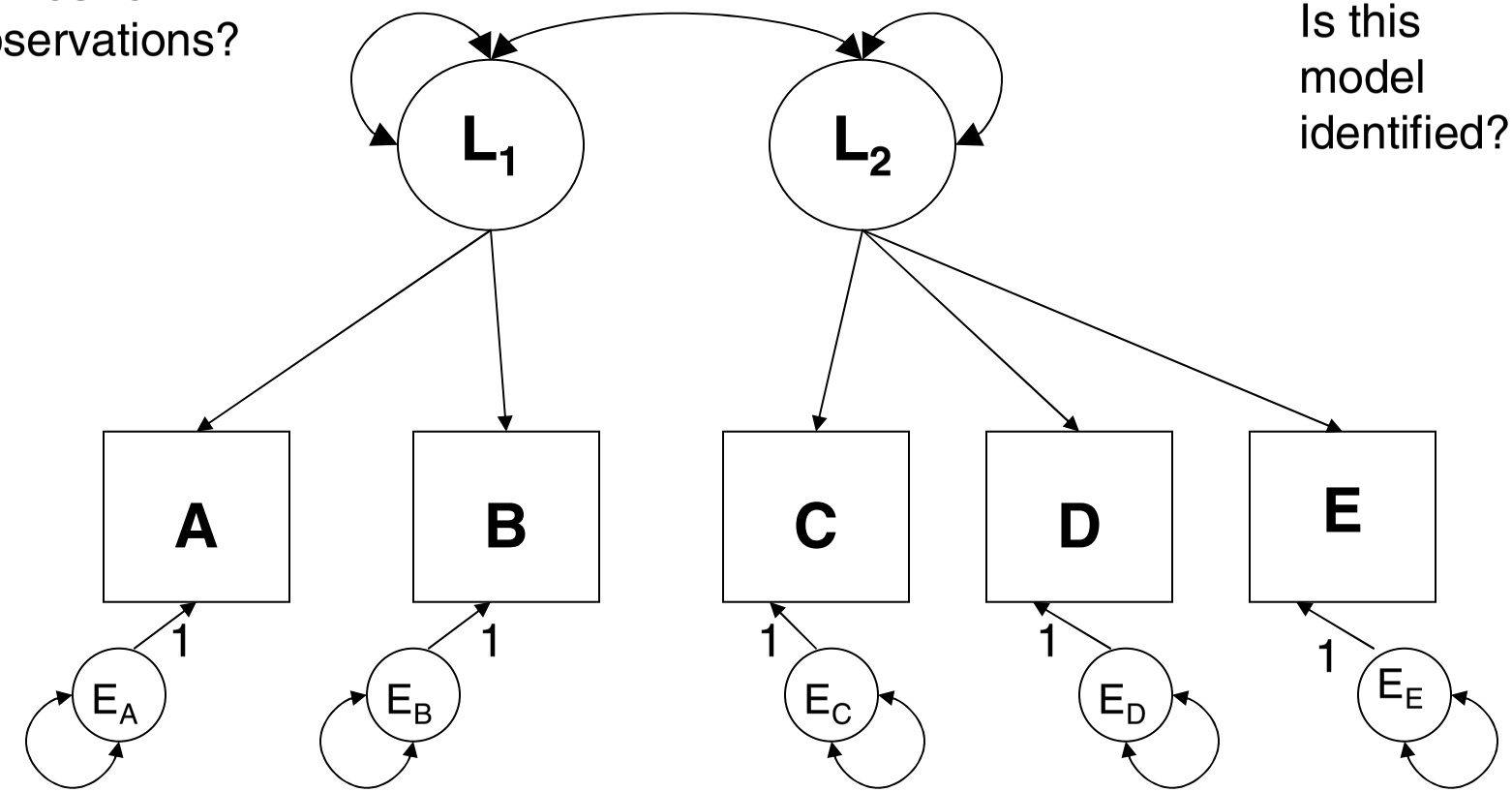
Model Identification

- A CFA model is identified if the following necessary conditions are met:
  - $df_M \geq 0$
  - Every latent variable, including measurement errors, must be assigned a scale (i.e., metric)
- A sufficient condition for identification is:
  - If a unidimensional CFA model with a single latent variable has at least three indicators, the model is identified. If a unidimensional model with two or more latent variables has at least two indicators per latent variable, the model is identified

Pop Quiz

What is the number of observations?

Is this model identified?



Assessing Model Fit

- All of the fit indices described previously for path analysis can be used to assess the fit of confirmatory factor analysis models

Comparing Models

- If theory is not specific about the number of latent variables, a first step may be to test the fit of a single-factor model
  - If the χ^2 test of this model is not rejected, it indicates that the observed variables do not show discriminant validity
- If the χ^2 test is rejected, a unidimensional model with two or more latent variables can then be fit to the data.

Comparing Models

- Test of discriminant validity: A LRT for nested models
  - A one-factor model is nested within the two-factor model because the former implies that the correlation between the two latent variables to 1.0

$$\mathrm{LRT} = \chi^2_{1\mathrm{F}} - \chi^2_{2\mathrm{F}}$$

$$\mathrm{LRT} \sim \chi^2[(1-\alpha); \mathrm{df}_{\mathrm{M}_1} - \mathrm{df}_{\mathrm{M}_2}]$$

Modifying Models

- Indicators may have low loadings (i.e., < .40) on the latent variable to which they were originally assigned.
  - Solution #1: specify that the indicator load on a different latent variable
  - Solution #2: specify that the measurement errors of two indicators covary
- Wrong number of factors may be specified
  - If two latent variables are highly correlated, this indicates that the model has too many factors

Modifying Models

- Examine correlation residuals and modification indices to determine ways to respecify a model

Summary of CFA

- CFA allows you to test an *a priori* model related to the measurement of latent variables
- Observed variables are imperfect measures of a latent construct
- Latent variables are given a scale either by assigning a reference variable or by standardizing the variance of the latent variable

Section V: Structural Regression Analysis

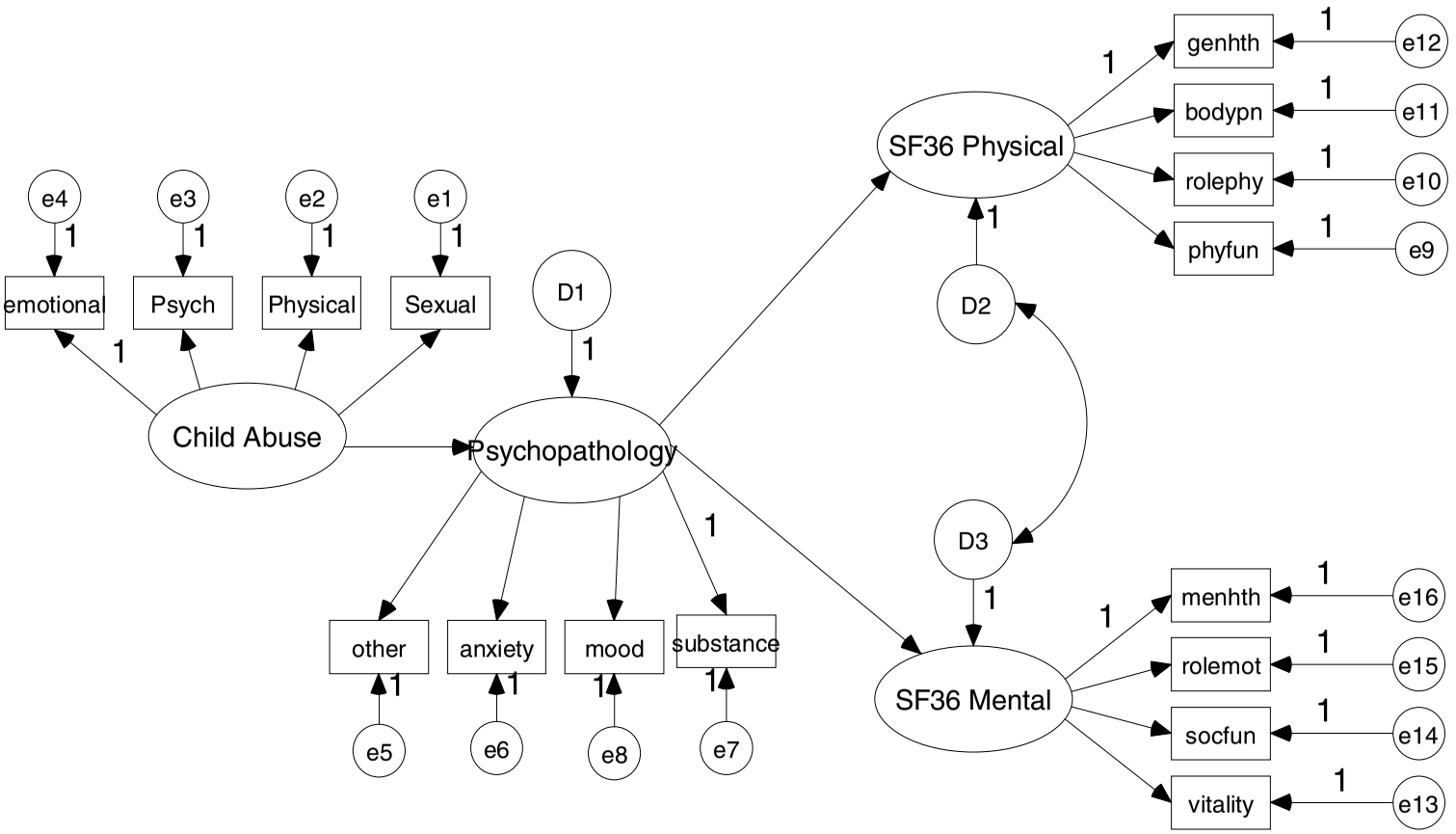
- Structural regression analysis vs. path analysis
- Review of model building blocks
- Identification of structural regression models
- Sequence of model testing
- Assessing model fit
- Equivalent models

Structural Regression Analysis vs. Path Analysis

- Path analysis: directional/causal relationships among observed variables

- Structural regression analysis: directional/causal relationships among latent variables (fully latent) or a mixture of observed and latent variables (partially latent)

Review of Model Building Blocks



Identification of Structural Regression Models

- Necessary conditions for identification:
  - $df_M \geq 0$
  - Each latent variable must be assigned a scale
- Additional requirement:
  - In order for the structural component to be identified, the measurement component must be identified

Identification of Structural Regression Models

- Model parameters include:
  - Variances and covariances (i.e., unanalyzed associations) of exogeneous variables (measurement errors, disturbances, exogenous latent variables)
  - Direct effects on endogeneous variables (factor loadings of indicators, direct effects on endogenous latent variables from other latent variables)

Sequence of Model Testing

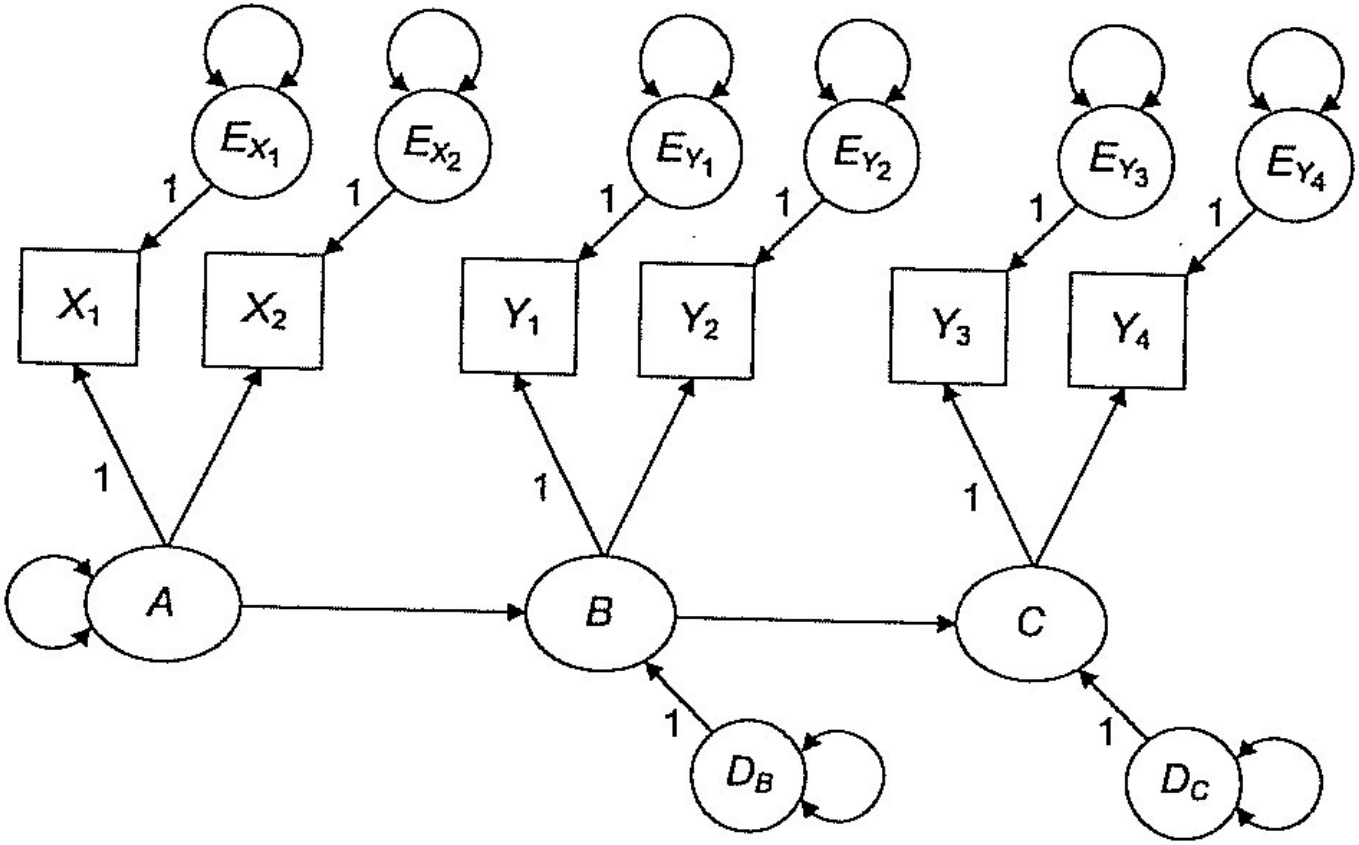
- Analysis is decomposed into two parts:
  - Measurement component: treat like CFA
  - Structural component: treat like path analysis

- A valid measurement model must be obtained before the structural component of a structural regression model can be tested.

Sequence of Model Testing

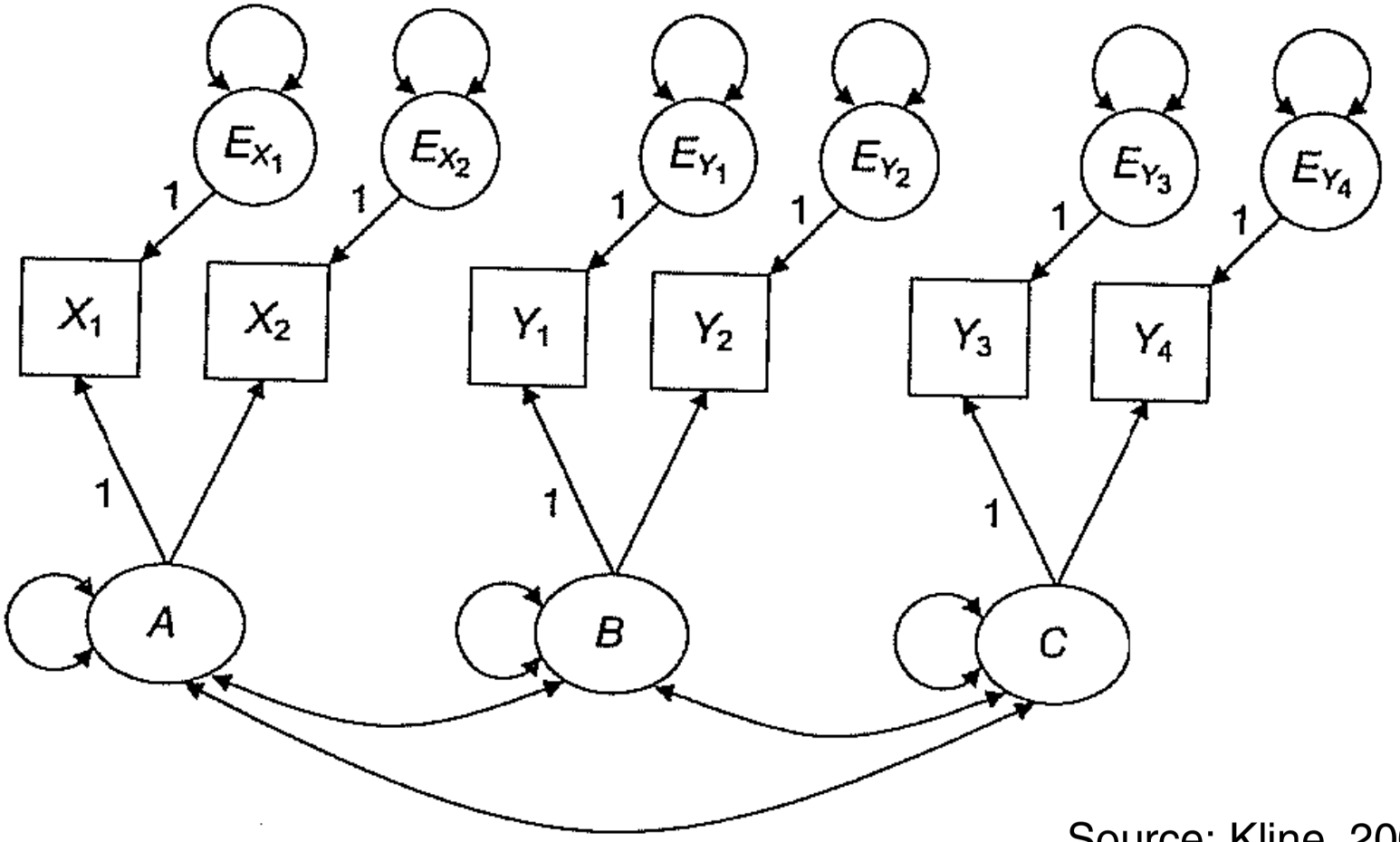
- Two-step modeling (Anderson & Gerbing, 1988)
  - Step #1: Find an acceptable-fitting CFA model
    - Structural regression model is first specified as a CFA model with unanalyzed associations among the latent variables.
  - Step #2: Assuming the measurement model provides an acceptable fit, specify the structural regression component and compare alternative structural models.

Sequence of of Model Testing



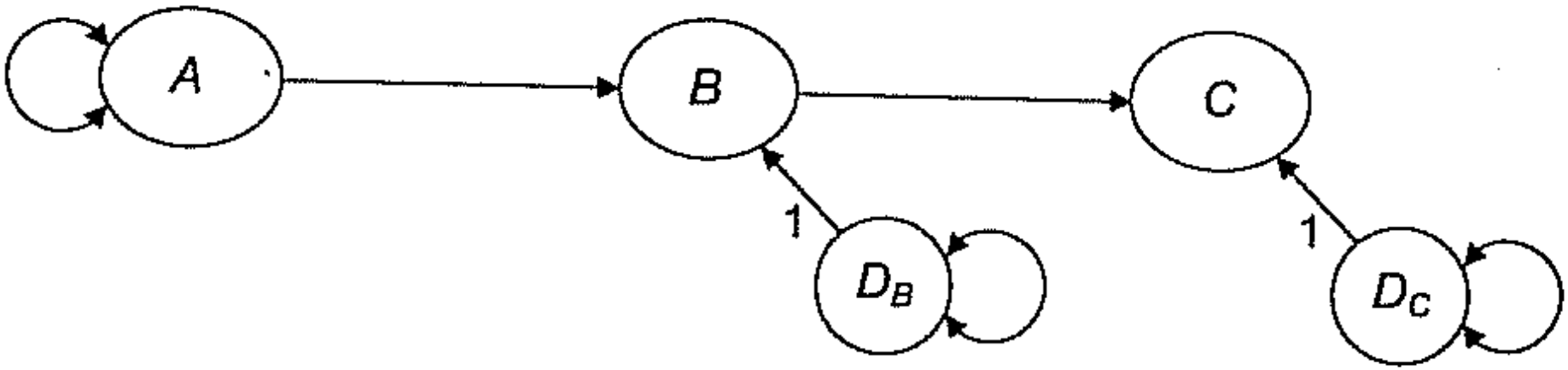
Source: Kline, 2005

Sequence of Model Testing



Source: Kline, 2005

Sequence of Model Testing



Source: Kline, 2005

Assessing Model Fit

- All of the fit indices described previously for path analysis can be used to assess the fit of structural regression models

Equivalent Models

- Hold the measurement model constant, and consider equivalent structural regression models
- Hold the structural regression model constant, and consider equivalent versions of the measurement model

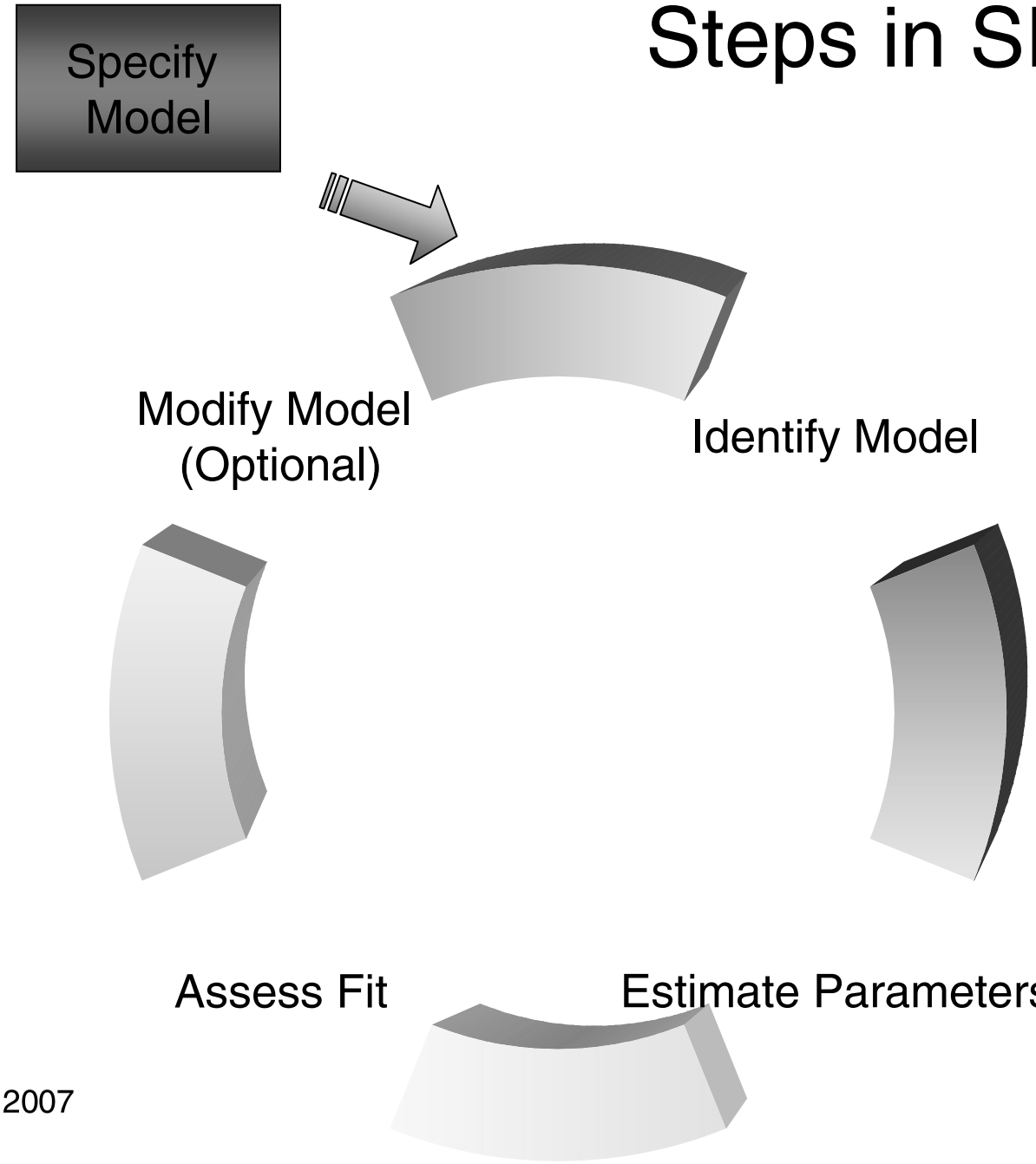
Summary of Structural Regression Analysis

- A structural regression model is similar to a simultaneous path analysis and CFA
- Analysis of a structural regresion model can be separated into two stages: (a) measurement model, and (b) structural model
- Identification of the measurement component of the model ensures identification of the structural component if the model is recursive

Section VI: Wrap-Up

- Review of the SEM process
- Other issues
- Limitations of SEM
- Strengths of SEM
- Next steps for learning

Steps in SEM


Specify Model
Identify Model
Estimate Parameters
Assess Fit
Modify Model (Optional)

Other issues

- Means
  - All analyses in this workshop focused on modeling variances and covariances
  - Differences in means are of particular interest in longitudinal data analysis
    - investigate changes in both means (i.e., intercepts) and slopes over time
- Interactions
  - Interactions are an important part of any model that is applied to data in the health and social sciences
  - Interactions are less commonly seen in SEM models because they increase model complexity; there is more than one way to include interaction terms in the model

Other issues

- Omitted variables
  - Failing to measure important variables can provide a misleading picture of measurement and/or causal structures, and can result in biased parameter estimates and inaccurate estimates of standard errors

Limitations of SEM

- Correlation does not imply causation
- Failure to consider equivalent models implies that the researcher's model is the "right" one

Strengths of SEM

- Requires the researcher to develop an *a priori* model for testing
- Can accommodate measurement error in observed variables
- Provides global assessments of fit – summary evaluations are available for even the most complex models
- Gives the researcher a broad set of tools for the analysis of multivariate data

Next Steps for Learning

- Review data preparation techniques
  - Missing data
  - Multivariate normality
  - Linear associations among variables
  - Collinearity
- Extend knowledge to include specialized SEM analyses
  - Multi-group comparisons
  - Latent growth curve models
  - Analyses involving categorical and/or dichotomous variables