

Structural Equation Modeling/Path Analysis

Introduction:

Path Analysis is the statistical technique used to examine causal relationships between two or more variables. It is based upon a linear equation system and was first developed by Sewall Wright in the 1930s for use in phylogenetic studies. Path Analysis was adopted by the social sciences in the 1960s and has been used with increasing frequency in the ecological literature since the 1970s. In ecological studies, path analysis is used mainly in the attempt to understand comparative strengths of direct and indirect relationships among a set of variables. In this way, path analysis is unique from other linear equation models: In path analysis mediated pathways (those acting through a mediating variable, i.e., “Y,” in the pathway $X \rightarrow Y \rightarrow Z$) can be examined. Pathways in path models represent hypotheses of researchers, and can never be statistically tested for directionality. Numerous articles deal with the use of path analysis in ecological studies (See Shipley 1997, 1999 and Everitt and Dunn 1991 [Section 14.8] for discussion of ecological applications and misuses of the technique).

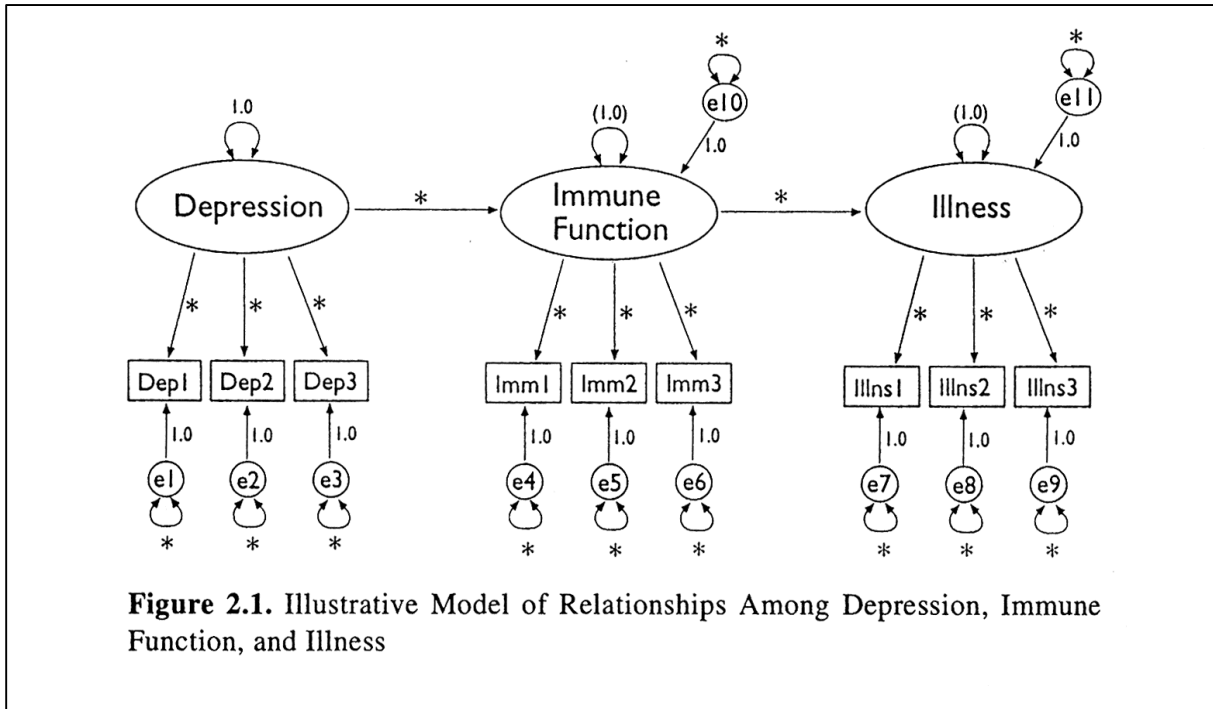
Path analysis is a subset of Structural Equation Modeling (SEM), the multivariate procedure that, as defined by Ullman (1996), “allows examination of a set of relationships between one or more independent variables, either continuous or discrete, and one or more dependent variables, either continuous or discrete.” SEM deals with measured and latent variables. A *measured variable* is a variable that can be observed directly and is measurable. Measured variables are also known as observed variables, indicators or manifest variables. A *latent variable* is a variable that cannot be observed directly and must be inferred from measured variables. Latent variables are implied by the covariances among two or more measured variables. They are also known as factors (i.e., factor analysis), constructs or unobserved variables. SEM is a combination of multiple regression and factor analysis. Path analysis deals only with measured variables.

Components of a Structural Equation Model:

Structural Equation Models are divided into two parts: a measurement model and a structural model. The measurement model deals with the relationships between measured variables and latent variables. The structural model deals with the relationships between latent variables only. One of the advantages to SEM, is that latent variables are free of random error. This is because error has been estimated and removed, leaving only a common variance.

The diagram below shows an example of a Structural Equation Model (taken from Hoyle 1995, p. 26).

Example (taken from Hoyle 1995, p. 26):



In SEM, measured variables are indicated by rectangles or squares (i.e., Dep 1, Imm 2, Illns 3, etc., in the diagram above) and latent variables are indicated by ellipses or circles (i.e., Depression, Immune Function and Illness in the diagram above). Error terms (“disturbances” for latent variables) are included in the SEM diagram, represented by “E’s” for measured variables and “D’s” for latent variables. The error terms represent residual variances within variables not accounted for by pathways hypothesized in the model.

The parameters of a SEM are the variances, regression coefficients and covariances among variables. A variance can be indicated by a two-headed arrow, both ends of which point at the same variable, or, more simply by a number within the variable’s drawn box or circle. Regression coefficients are represented along single-headed arrows that indicate a hypothesized pathway between two variables (These are the weights applied to variables in linear regression equations). Covariances are associated with double-headed, curved arrows between two variables or error terms and indicate no directionality. The data for a SEM are the sample variances and covariances taken from a population (held in **S**, the observed sample variance and covariance matrix).

A study by Bart and Earnst (1999) provides an example of the ecological application of path analysis. In this example, the relative importance of male traits and territory quality are examined in reference to number of females paired with each male. The application uses territory quality, male quality, and male pairing success (# of females/male) as the variables in question. Male pairing success is the dependent variable. An indirect effect of male quality on territory quality (male quality affecting a male’s ability to obtain territory of high quality) is seen to affect male

pairing success. A direct effect of male quality on pairing success is seen as well. The analysis is not conducted in an SEM framework; however, it would be possible to envision male quality and territory quality as latent variables affecting an observed variable, male pairing success. In the study, comb size, tarsus length and wing chord are used as measures of male quality. These variables could be viewed as the observed variables indicating the latent variable male quality. Proportion of territory covered by dunes, willow density and territory size are used as measures of territory quality. These could be viewed as the observed variables indicating the latent variable territory quality.

Structural Equation Model Construction:

The goal in building a path diagram or other structural equation model, is to find a model that fits the data (**S**) well enough to serve as a useful representation of reality and a parsimonious explanation of the data.

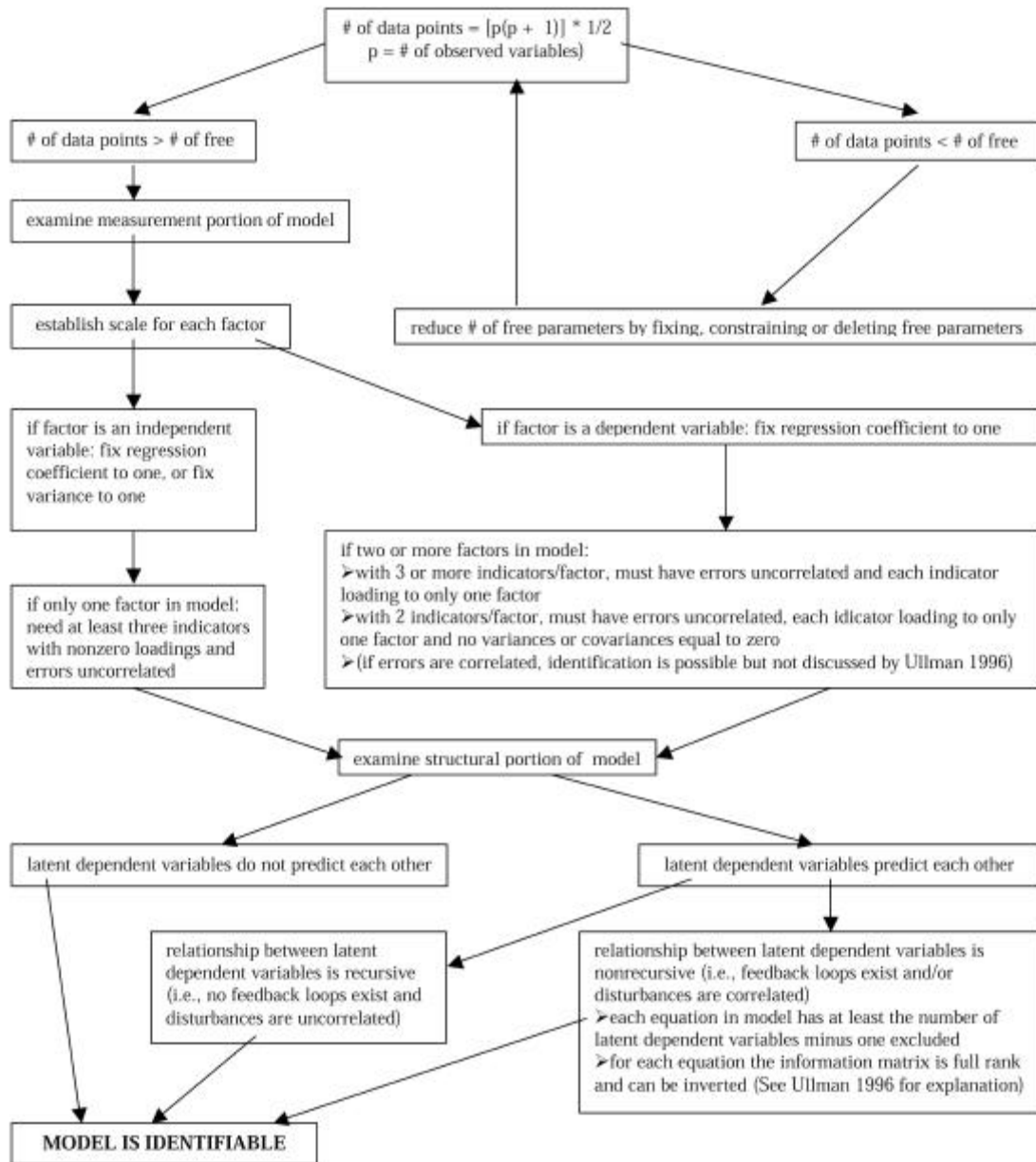
There are five steps involved in SEM construction:

1. Model Specification
2. Model Identification (Some authors include this step under specification or estimation)
3. Model Estimation
4. Testing Model Fit
5. Model Manipulation

Model Specification is the exercise of formally stating a model. It is the step in which parameters are determined to be fixed or free. *Fixed parameters* are not estimated from the data and are typically fixed at zero (indicating no relationship between variables). The paths of fixed parameters are labeled numerically (unless assigned a value of zero, in which case no path is drawn) in a SEM diagram. *Free parameters* are estimated from the observed data and are believed by the investigator to be non-zero. Asterisks in the SEM diagram label the paths of free parameters. Determining which parameters are fixed and which are free in a SEM is extremely important because it determines which parameters will be used to compare the hypothesized diagram with the sample population variance and covariance matrix in testing the fit of the model (Step 4). The choice of which parameters are free and which are fixed in a model is up to the researcher. This choice represents the researcher's *a priori* hypothesis about which pathways in a system are important in the generation of the observed system's relational structure (e.g., the observed sample variance and covariance matrix).

Model Identification concerns whether a unique value for each and every free parameter can be obtained from the observed data. It depends on the model choice and the specification of fixed, constrained and free parameters. A parameter is constrained when it is set equal to another parameter. Models need to be overidentified in order to be estimated (Step 3 in SEM construction) and in order to test hypotheses about relationships among variables (See Ullman 1996 for a more detailed explanation of the levels of model identification). A necessary condition for overidentification is that the number of data points (number of variances and covariances) is less than the number of observed variables in the model. A flow-chart schematic (extracted from Ullman 1996) summarizes the procedures of model identification:

Flowchart to Determine Model Identifiability:



The above flowchart offers brief guidelines to determining if a model is identifiable. However, even with these guidelines, it is possible that the model may not be identifiable. Characteristics

of the data or errors made by the researcher can lead to problems with model identification (See Hoyle 1995 and Ullman 1996 for more details).

Estimation: In this step, start values of the free parameters are chosen in order to generate an estimated population covariance matrix, $S(q)$, from the model. Start values can be chosen by the researcher from prior information, by computer programs used to build SEMs (see “References” section at end of this web page), or from multiple regression analysis (See Ullman 1996 and Hoyle 1995 for more start value choices and further discussion). The goal of estimation is to produce a $S(q)$ that converges upon the observed population covariance matrix, S , with the residual matrix (the difference between $S(q)$ and S) being minimized. Various methods can be used to generate $S(q)$. Choice of method is guided by characteristics of the data including sample size and distribution. Most processes used are iterative. The general form of the minimization function is:

$$Q = (s - s(q))'W(s - s(q))$$

where,

s = vector containing the variances and covariances of the observed variables

$s(q)$ = vector containing corresponding variances and covariances as predicted by the model

W = weight matrix.

(Some authors refer to Q as F).

The weight matrix, W , in the function above, corresponds to the estimation method chosen. W is chosen to minimize Q , and $Q(N-1)$ gives the fitting function, in most cases a X^2 -distributed statistic. The performance of the X^2 is affected by sample size, error distribution, factor distribution, and the assumption that factors and errors are independent (Ullman 1996). Some of the most commonly used estimation methods are:

Generalized Least Squares (GLS)

$$F_{GLS} = 1/2 \text{tr}[(S - S(q))W^{-1}]^2$$

where,

tr = trace operator, takes sum of elements on main diagonal of matrix

W^{-1} = optimal weight matrix, must be selected by researcher (most common choice is S^{-1})

Maximum Likelihood (ML)

$$F_{ML} = \log|S| - \log|S| + \text{tr}(SS^{-1}) - p$$

in this case, $W = S^{-1}$ and p = number of measured variables

Asymptotically Distribution Free (ADF) Estimator

$$F_{ADF} = [S - s(q)]'W^{-1}[S - s(q)]$$

W , in this function, contains elements that take into account kurtosis.

Ullman (1996) and Hoyle (1995) discuss the advantages and limitations of the above estimators. ML and GLS are useful for normally distributed data when factors and errors are independent. ADF is useful for nonnormally distributed data, but is shown only to work well with sample sizes above 2,500. Ullman indicates that the best estimator for nonnormally distributed data and/or dependence among factors and errors is the Scaled ML (See Ullman 1996 for further discussion). Whatever function is chosen, the desired result of the estimation process is to obtain a fitting function that is close to 0. A fitting function score of 0 implies that the model's estimated covariance matrix and the original sample covariance matrix are equal.

Assessing Fit of the Model: As stated in the last section, a fitting function value of close to 0 is desired for good model fit. However, in general, if the ratio between X^2 and degrees of freedom is less than two, the model is a good fit (Ullman 1996).

To have confidence in the goodness of fit test, a sample size of 100 to 200 is recommended (Hoyle 1995). In general a model should contain 10 to 20 times as many observations as variables (Mitchell 1993).

Ullman (1996) discusses a variety of non- X^2 -distributed fitting functions, which he calls "comparative fit indices." Hoyle (1995) refers to these as "adjunct fit indices." Basically, these approaches compare the fit of an independence model (a model which asserts no relationships between variables) to the fit of the estimated model. The result of this comparison is usually a number between 0 and 1, with 0.90 or greater accepted as values that indicate good fit. Both Hoyle and Ullman suggest use of multiple indices when determining model fitness.

Model Modification: If the covariance/variance matrix estimated by the model does not adequately reproduce the sample covariance/variance matrix, hypotheses can be adjusted and the model retested. To adjust a model, new pathways are added or original ones are removed. In other words, parameters are changed from fixed to free or from free to fixed. It is important to remember, as in other statistical procedures, that adjusting a model after initial testing increases the chance of making a Type I error.

The common procedures used for model modification are the **Lagrange Multiplier Index (LM)** and the **Wald test**. Both of these tests report the change in X^2 value when pathways are adjusted. The LM asks whether addition of free parameters increases model fitness. This test uses the same logic as forward stepwise regression. The Wald test asks whether deletion of free parameters increases model fitness. The Wald test follows the logic of backward stepwise regression.

To adjust for increased type one error rates, Ullman (1996) recommends using a low probability value ($p < 0.01$) when adding or removing parameters. Ullman also recommends cross-validation

with other samples. Because the order in which parameters are freed can affect the choice of remaining parameters, LM should be applied before the Wald test (i.e., add all parameters before beginning to delete them) (MacCullum 1986, cited in Ullman 1996). Refer to Ullman (1996) and Hoyle (1995) for further description of these and other model modification techniques.

Final Presentation of Model: Once the model has attained an acceptable fit, individual estimates of free parameters are assessed. Free parameters are compared to a null value, using a z-distributed statistic. The z statistic is obtained by dividing the parameter estimate by the standard error of that estimate. The ratio of this test must exceed +/-1.96 in order for the relationship to be significant. After the individual relationships within the model are assessed, parameter estimates are standardized for final model presentation. When parameter estimates are standardized, they can be interpreted with reference to other parameters in the model and relative strength of pathways within the model can be compared.

Limitations and Advantages of SEM:

Once again, SEM cannot test directionality in relationships. The directions of arrows in a structural equation model represent the researcher's hypotheses of causality within a system. The researcher's choice of variables and pathways represented will limit the structural equation model's ability to recreate the sample covariance and variance patterns that have been observed in nature. Because of this, there may be several models that fit the data equally well. In spite of this, the SEM approach remains useful in understanding relational data in multivariate systems. The abilities of SEM to distinguish between indirect and direct relationships between variables and to analyze relationships between latent variables without random error differentiate SEM from other simpler, relational modeling processes.

Structural Equation Modeling Programs:

There are many web-accessible programs designed to do Structural Equation Modeling. A good site to access these from is:

<http://www.gsu.edu/~mkteer/software.html>

Four programs that have free demonstration versions accessible on the web are LISREL, AMOS, EQS and PISTE. The program that has been most widely used is LISREL This program has been around since the 1970's. There are many texts written about the use of LISREL and they are necessary to begin to understand this program (A listing of these texts is available on the LISREL website). A free downloadable student version can be accessed at <<http://www.ssicentral.com/lisrel/mainlis.htm>>. This program is normally priced at \$575 for Windows, and \$475 for PowerMac.

Another free downloadable program for students is AMOS. This demonstration version is limited to estimation of problems with 8 or fewer indicators and no more than 54 free parameters. It operates with Windows software. The literature on AMOS is not nearly as

extensive as that for LISREL. It can be downloaded at <http://www.spss.com/software/spss/base/amos/>. The non-student pricing for AMOS is \$395.

The third program is EQS. This program is much more user friendly than the others. The program allows the user to draw a path diagram and proceed from there in the development and testing of the model. The literature for EQS is also not as extensive as that for LISREL, but it isn't necessary to the same degree. A demonstration version EQS can be downloaded at <http://www.mvsoft.com>. The demonstration is limited to 14 parameters (25 for multiple groups) and it does not allow information to be saved, printed or copied into another application. EQS5 for Windows is available \$595 (academic discount), and for Dos \$495.

PISTE is a Macintosh-based program available from the University of Montreal. This program is for path analysis and is downloadable at ftp://biol10.biol.umontreal.ca/public_ftp/labo/R/V3/Mac/piste-en.hqx.

References:

- Everitt, B.S. and G. Dunn. 1991. *Applied Multivariate Data Analysis*. Halsted Press. New York, NY. pp. 257-275. This chapter lays out two mathematical approaches to SEM construction, and includes a very useful discussion on the limitations of path analysis at its end.
- Hoyle, R.H. (ed.) 1995. *Structural Equation Modeling*. SAGE Publications, Inc. Thousand Oaks, CA. This book provides a very readable, broken-down introduction to SEM. It discusses SEM in relation to AMOS software.
- Johnson, R.A., and D.W. Wichern. 1982. *Applied Multivariate Statistical Analysis*. Prentice Hall, Inc. Englewood Cliffs, NJ. pp. 326-333.
- Kelloway, E.K. 1998. *Using LISREL for Structural Equation Modeling*. SAGE Publications, Inc. Thousand Oaks, CA. Ch 6, Ch 7.
- Loehlin, J.C. 1987. *Latent Variable Models*. Lawrence Erlbaum Associates, Inc. Hillsdale, NJ.
- Maruyama, G.M. 1998. *Basics of Structural Equation Modeling*. SAGE Publications, Inc. Thousand Oaks, CA
- Mitchell, R.J. 1993. Path analysis: pollination. In: *Design and Analysis of Ecological Experiments* (Scheiner, S.M. and Gurevitch, J., Eds.). Chapman and Hall, Inc. New York, NY. pp. 211-231. This short and easy-to-follow chapter uses an ecological example to illustrate and discuss path analysis.
- Schumacker, R.E. and R.G. Lomax. 1996. *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates, Inc. Mahwah, NJ. This book seemed slightly less readable than Hoyle's, but it provides a nice introduction to SEM in the beginning chapters. LISREL8-SIMPLIS and EQS computer applications are discussed.

Ullman, J.B. 1996. Structural equation modeling (In: *Using Multivariate Statistics*, Third Edition, B.G. Tabachnick and L.S. Fidell, Eds.). HarperCollins College Publishers. New York, NY. pp. 709-819. This general introduction to SEM runs through the matrix-based approach to structural equation modeling and discusses all steps involved in the process. It compares various computer program outputs using the same data set to illustrate differences among the programs. Because of the jump into mathematical explanation, this is at first less readable than Hoyle's or Schumacker and Lomax's introductions.

Articles Using or Addressing Issues in Path Analysis in Biology:

Abell, A.J. 1999. Variation in clutch size and offspring size relative to environmental conditions in lizard *Sceloporus virgatus*. *Journal of Herpetology* 33(2): 173-180.

Bart, J. and S.L. Earnst. 1998. Relative importance of male and territory quality in pairing success of male rock ptarmigan (*Lagopus mutus*). *Behavioral Ecology and Sociobiology* 45: 355-359.

Magura, S. and A. Rosenblum. 2000. Modulating effect of alcohol use on cocaine use. *Addictive Behaviors* 25(1): 117-122.

Sanchez-Pinero, F. and G.A. Polis. 2000. Bottom-up dynamics of allochthonous input: direct and indirect effects of seabirds on islands. *Ecology* 81: 3117-3132.

Shine, R. 1996. Life-history evolution in Australian snakes: a path analysis. *Oecologia* 107: 484-489.

Shipley, B. 1997. Exploratory path analysis with applications in ecology and evolution. *The American Naturalist* 149(6): 1113-1138.

Shipley, B. 1999. Testing causal explanations in organismal biology: causation, correlation and structural equation modeling. *Oikos* 86(2): 374-382.

Sinervo, B. 1998. Adaptation of maternal effects in the wild: path analysis of natural variation and experimental tests of causation (In: *Maternal Effects of Adaptations*, T.A. Mousseau and C.W. Fox, Eds.). Oxford University Press. New York, NY. pp. 288-306.

Links to other sites on SEM and Path Analysis:

<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>

This is an excellent web page. It covers SEM in depth, mostly focusing on goodness of fit tests and assumptions of SEM. It is very thorough and therefore lengthy. Discussions are related to AMOS, LISREL and EQS computer applications, especially in terms of capabilities in goodness of fit tests. The discussions of identification, handling of missing data, and methods for

estimating path coefficients are all very helpful. The end of the page gives information in a “Frequently Asked Questions” format and also provides links to a number of SEM packages (under the question heading: Does it matter which statistical package you use for structural equation modeling?)

<http://www2.chass.ncsu.edu/garson/pa765/path.htm>

This web page from the above author focuses on path analysis. It is also a very useful site, with discussion of key concepts and terms, assumptions, and a “Frequently Asked Questions” section.

<http://www.statsoft.com/textbook/stathome.html>

This site provides a short explanation of SEM.

<http://www.maths.ex.ac.uk/~jph/psy6003/pathanal.html>

This site from the University of Exeter provides a brief introduction and explanation of path analysis. It is part of a set of class notes.

<http://www.uic.edu/classes/idsc/ids570/ntspath.htm>

This site from the University of Chicago also provides class notes on path analysis and SEM. It answers a general list of questions about SEM: why to use it, when to use it and how, etc.

<http://www.geocities.com/CollegePark/Campus/caveat.htm>

This short paragraph discusses the uncertainty of causal direction in SEM pathways and the problems caused by *post hoc* adjustment of models. It provides references to more in depth discussion of these topics.