# Structural equation modeling

Rosenthal, Sonny

2017

Rosenthal, S. (2017). Structural equation modeling. M. Allen (Eds.), The SAGE Encyclopedia of Communication Research Methods (pp. 1683-1687). Sage Publications. https://hdl.handle.net/10356/165157

https://hdl.handle.net/10356/165157

https://doi.org/10.4135/9781483381411

Cite as:

Rosenthal, S. (2017). Structural equation modeling. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 1683–1687). SAGE. https://doi.org/10.4135/9781483381411.n594

# Structural Equation Modeling

Structural equation modeling (SEM) is a versatile tool for conducting a wide range of multivariate statistical analyses, including multiple regression, mediation analysis, moderation analysis, and analyses of variance and covariance. Two specialized uses of SEM that appear frequently in communication research are confirmatory factor analysis (CFA) and path analysis. Confirmatory factor analysis specifies one or more unobserved constructs, or *latent factors*, that a number of *observed indicators* define. This analysis is useful for validating the composition of multiple-item indexes or scales. The other common use of structural equation modeling, path analysis, estimates correlation and regression paths among structural nodes, which may include both observed variables and latent factors. When path analysis includes latent factors, the definition of those factors is equivalent to CFA and is the basis of a *measurement model*. The specification of paths among latent factors and observed variables constitutes a *structural model*. Whatever the intended use of SEM in communication research, it should be based on careful theoretical considerations.

## General Approach

A set of variables have an observed covariance matrix, which accounts for all the relationships among the variables. A structural regression model that estimates all covariance paths among the variables will reproduce the covariance matrix exactly but is unlikely to resolve a theoretical understanding of how the variables relate. The aim of SEM is to define a parsimonious regression model that specifies theoretically consistent paths among variables. To the extent that model paths reflect established theory, the model will generally have good external validity.

Based on the specified regression model, a software algorithm will estimate a model-implied covariance matrix. This estimation can be done by hand following rules of path tracing for unstandardized or standardized parameter estimates. The estimation of model parameters includes correlation and regression coefficients for the specified paths, variances of exogenous variables, and residual variances of endogenous variables. Non-specified paths are constrained, usually by default, to a value of 0, but the modeler may constrain paths to any value. With most real-world data sets, as the number of model constraints increases, there is increasing deviation of the implied covariance matrix from the observed covariance matrix. The residual covariance matrix shows this deviation. Thus, a further aim of SEM is to define a parsimonious model that reproduces the observed covariance matrix with minimal residuals. To the extent that residuals are minimized, the model will generally have statistical validity, also described as good *model fit*.

The results of SEM should thus be grounded solidly in theory and satisfy certain statistical requirements before they are interpreted. Assuming that study design, sampling, and data collection further abide rigorous standards, thoughtful interpretation of SEM results can contribute meaningfully to scholarship.

*Tests of Model Fit*

There are many tests of a model's goodness of fit that provide information about its statistical validity. Each different test gives an indication of how well (or poorly) the model-implied covariance matrix reproduces the observed matrix. Some tests account for sample size and model parsimony, and combinations of tests can indicate model fit that balances type I and type II error. The tests below appear commonly in communication research and are adequate for reporting results of most SEM analyses.

*Chi-squared*. A chi-squared ($\chi^2$) goodness of fit test evaluates the null hypothesis that a model-implied covariance matrix is not significantly different from the observed matrix. In the case of SEM, a non-significant $\chi^2$ fails to reject the null hypothesis, which indicates good model fit. Thus, a modeler wants to have a non-significant $\chi^2$ estimate. This test is sensitive to sample size, and for samples where $n > 400$, the $\chi^2$ test will often be significant and not give an accurate indication of model fit.

There are at least two alternative approaches to using $\chi^2$ when the sample size is large. First, the modeler can analyze the covariance matrix (as opposed to raw data) and specify a smaller sample size, regardless of the actual sample size. Second, the modeler can use $\chi^2/df$, which gives the ratio of the $\chi^2$ estimate to the degrees of freedom associated with the $\chi^2$ test, where a value of 0 indicates perfect model fit. The problem with both approaches is that the pseudo sample size for the former and the cutoff value for the latter are somewhat arbitrary and have limited empirical support.

Fortunately, most SEM software packages provide additional tests of model fit that can supplement the $\chi^2$ test. These fit indices have empirically validated cutoff criteria that suggest a model's goodness of fit.

*RMSEA*. The root-mean squared error of approximation (RMSEA) is based on $\chi^2$ and *df* but includes the sample size in the denominator. Thus, this estimate is not sensitive to sample size. Typically, RMSEA values of .06 and smaller suggest good fit. In addition, the 90% confidence interval of the estimate should not exceed .10.

*CFI*. The comparative fit index (CFI) compares the fit of the model-implied covariance matrix to the fit of the matrix implied by the baseline model. The baseline model treats all variables as uncorrelated; thus, the baseline-implied matrix will generally have poor fit and a high $\chi^2$ test value relative to the implied matrix. A larger CFI value indicates better relative fit of the implied matrix, and CFI values larger than or equal to .95 generally suggest good fit.

*SRMR*. The standardized root mean square residual (SRMR) directly evaluates the residual covariance matrix. As the name implies, SRMR gives a standardized estimate of the average absolute value of the residuals; the smaller the residuals, the smaller the value of SRMR. Typically, SRMR values of .08 and smaller indicate good fit.

*Combination rules*. When used in combination, fit indices can provide a more robust indication of model fit. Type II error is minimized with an acceptable amount of type I error inflation when CFI is around .96 and SRMR is around .09. Another combination that balances type I and type II error is when RMSEA is around .06 and SRMR is around .09.

*Comparative fit*. Whereas the above fit indices are used to evaluate a single model, goodness of fit can be used to compare fit among different models, which can be nested or non-nested (see below for a discussion of model nesting). When two models are nested, the change in $\chi^2$ between the models can be evaluated at the corresponding change in *df* and respective *p*-value. The restricted model (i.e., the one with fewer degrees of freedom) has better fit if it has a significantly lower $\chi^2$ value.

For non-nested models, evaluating change in $\chi^2$ is not appropriate. Rather, a fit index such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), or sample-size adjusted BIC (SSABIC) should be used. These comparative indices account for the number of variables in the different models, which allows for a comparison of non-nested models. These comparative indices can also supplement comparisons of nested models. When comparing nested or non-nested models, the model with the lowest AIC, BIC, or SSABIC value has the best statistical fit.

*Modification Indices*

Most SEM software packages include an option to estimate modification indices, which provide information about the change in $\chi^2$ associated with freely estimating constrained parameters. The critical $\chi^2$ value for 1 degree of freedom at the $p$ = .05

significance level is 3.84. Thus, if the change in $\chi^2$ value for freely estimating one parameter is larger than 3.84, then its estimation will result in better model fit. A modeler should limit modifications to those that are consistent with theory. When a modeler includes new paths based on modification indices, this reduced model has fewer degrees of freedom and is nested within the full model.

*Nested and Non-Nested Models*

A modeler can use SEM to compare the fit of alternative models. The appropriate method of comparison depends on whether the models are nested or non-nested. An easy test to determine whether one model is nested within another is if the second model can be achieved by constraining some paths in the first model to 0.

To give an example, figure 1 shows three alternative structural models in which two independent variables and a mediator variable predict a dependent variable. Models A and B cannot be equated by constraining paths in one of the models to 0; thus, they are non-nested models. In order to evaluate their relative goodness of fit, a modeler should use a comparative index, such as the AIC. On the other hand, model C is nested within both model A and model B. It is identical to model A when path *a* equals 0; it is identical to model B when path *b* equals 0. Model C has one fewer degrees of freedom than models A and B, and is thus a restricted model, whose change in $\chi^2$ indicates fit relative to the other two models.

*Model Identification*

The SEM estimation requires that a model be over-identified ($df > 0$) or just identified ($df = 0$). If a model is under-identified ($df < 0$), either the algorithm will not proceed or the results will be invalid. A model's degrees of freedom is the difference between the number of observed parameters and the number of estimated parameters. In an unstandardized model, the observed parameters are contained in the covariance matrix and their number is equal to $[k*(k + 1)]/2$, where $k$ is the number of variables in the model. The estimated parameters include model-specified covariances and regression paths, variances of exogenous variables, and residual variances of endogenous variables.

For example, all three models in figure 1 have 10 observed parameters: 4 variances and 6 covariances. Models A and B estimate 7 parameters each: 3 regression paths, 2 variances, and 2 residual variances. Thus, these two models have 10 – 7 = 3 degrees of freedom. Model C includes one additional regression path, and thus has 10 – 8 = 2 degrees of freedom. All three models are over-identified.

One instance of under-identification occurs when two observed items ($k = 2$) indicate a latent factor. While there are 3 observed parameters, there are 4 estimated parameters ($df = -1$), and any results of this model are invalid. However, if two correlated latent factors each have two indicators ($k = 4$), then $df = 1$ and the model is over-identified. Thus, a two-item factor can be analyzed in SEM if it has non-zero covariance with at least one other variable, whether the other variable is latent or observed.

*Sample Size*

In determining sample size a priori, a basic rule of thumb recommends a minimum 10 observations per measurement item. Another recommendation is a minimum of 5 observations per free parameter. A more stringent a priori estimate of sample size accounts for desired statistical power and the nature of the model fit test. For example, $n = 200$ is recommended for a close fit test (e.g., RMSEA) of a model with 55 degrees of freedom and a desired power of .80. If the model has only 10 degrees of freedom, the recommended sample size jumps to 782. Modelers should make a priori sample size determinations based on the specific characteristics of their models, and there are calculators online and as parts of statistical software packages (e.g., R and G*Power) that can aid in this determination.

## Common Uses of SEM

### *Confirmatory Factor Analysis*

The measurement of theoretical constructs in quantitative research often involves multiple-item indexes or scales. Such measurement assumes that constituent items share common variance, and that this common variance indicates the construct of interest. Confirmatory factor analysis is a means of assessing this common variance by defining one or more latent factors via paths to observed items, also called factor *indicators*. After explicating a theoretical construct, a modeler can use exploratory factor analysis as a starting point for defining the factor structure, which the modeler can confirm using CFA.

Exploratory factor analysis statistically defines latent factors post hoc, where the underlying factor structure reflects common variance among observed items. This analysis is usually one of the first steps in validating a construct's operationalization if it involves a multiple-item index or scale. If the results of this analysis are consistent with theory, then the next step is to confirm the factor structure using CFA.

The CFA model specifies paths from latent constructs to indicators. These paths correspond to factor loadings, and the squared standardized loading is equivalent to $R^2$ for the indicator, just as the loadings are interpreted in exploratory factor analysis. If there are multiple factors, typically items will have loadings on only a single factor but may indicate more than one factor depending on theoretical and statistical needs. As well, there may be reasons to specify covariance among indicator residuals within or between factors. For example, if indicators of two different factors have very similar wording, then they may have residual variance in common that should covary. Proper accounting of residual variance can clarify the factor structure and improve model fit.

In practice, both exploratory and confirmatory factor analysis can use a single data set, given an adequate sample size. The modeler can conduct exploratory factor analysis on a random half of the cases in the sample and then conduct CFA on the remaining cases.

It is common to depict a CFA model using a path diagram, which uses lower-case Greek notation to define model components. Figure 2 depicts a basic CFA model with 3 indicators ($x_1$, $x_2$, $x_3$; $k = 3$) of a single factor ($\xi_1$). The model is just identified, with 6 observed parameters—three variances and three covariances—and six estimated parameters—two factor loadings ($\lambda$), three residual variances of $x$ indicators ($\delta$), and one variance of an exogenous latent factor ($\varphi$). Note that the factor loading associated with the marker indicator—i.e., for each factor, the indicator with the strongest loading—is constrained to a value of 1, which is why only two loadings are estimated.

*Structural Equation Modeling*

An advantage of SEM over other statistical analyses is that it can analyze simultaneous regression equations, which may include paths among latent factors. For illustrative purposes, figure 3 depicts an SEM with direct and mediated effects of two exogenous latent factors ($\xi_1$, $\xi_2$) and one endogenous latent factor ($\eta_1$) on an observed dependent variable ($y_3$). There are three indicators for each exogenous latent factor ($x_1$, $x_2$, $x_3$; $x_4$, $x_5$, $x_6$) and two indicators for the endogenous latent factor ($y_1$, $y_2$).

*Measurement model.* Prior to evaluating the structural paths—i.e., paths among the variables of interest—a modeler should test the measurement model. The measurement model includes the factor structure in which latent factors predict indicators, but freely estimates covariances among the latent factors and observed variables. In figure 3, the measurement model has 22 degrees of freedom, which reflect 45 observed parameters—9 variances and 36 covariances—and 23 estimated parameters—5 factor loadings, 8 error variances, 4 variances, and 6 covariances. If the measurement model has poor fit, then the modeler can specify theoretically consistent modifications, which may include removing indicators, adding factor cross-loadings, and covarying errors. If the measurement model has good fit, then the modeler may proceed to evaluate the structural model.

*Structural model.* The structural model includes the factor structure and theoretically consistent paths among model variables. In figure 3, paths among the latent factors and the observed dependent variable constitute the structural model. Whereas the measurement model included 6 covariances, the structural model specifies 4 regression paths, while constraining to 0 the remaining two covariances ($\xi_1$ with $\xi_2$ and $\xi_1$ with $y_3$). Given the two additional constraints, the structural model has 24 degrees of freedom. If the structural model has good fit, then the modeler can use inferential statistics to make conclusions about the modeled relationships.

*Path Analysis*

Communication researchers often use well established indexes or scales to measure constructs of interest and focus their analyses on the relationships among constructs rather than the composition and validity of the latent factors they represent. In such instances, a researcher can simply analyze a structural model using composite variables (e.g., averaged or summated scores). This model specifies only paths among variables of interest and does not include a measurement model. Without using SEM, a modeler can estimate a path model by conducting a series of linear regression analyses. Given the model paths, the modeler can use path tracing rules to estimate the implied covariance matrix and calculate tests of model fit. A simpler approach is to use SEM to analyze the structural model in a single equation. The results will include estimated parameters and tests of model fit.

*Extended Uses of SEM*

With an understanding of CFA and path analysis, a modeler can use SEM to conduct more complex analyses. The analyses below are not exhaustive of those available but offer solutions for a wide range of statistical models. Their descriptions are brief and intended simply as introductions, rather than instructions, to the analyses.

*Mediation analysis.* When SEM estimates simultaneous regression models, a modeler may be interested in evaluating the indirect paths in mediation analysis. Mediation analysis is a common use of SEM in communication research. Most SEM software packages can provide robust estimates of indirect effects using a Sobel test, delta method, or other appropriate estimator.

*Latent factor interactions.* Modelers can also use SEM to test for moderation effects among latent factors. Modeling interaction terms in SEM is structurally equivalent to modeling interaction terms in multiple regression, with the addition of a measurement model. As with moderation analyses in multiple regression, an SEM that includes an interaction of two latent factors should control for both main effects.

*Latent growth modeling.* When data are longitudinal, a modeler can use SEM to estimate the intercept and slope of a growth curve. The growth curve is equivalent to a regression equation, where the intercept is the predicted outcome at time 0 and the slope can be used to predict outcomes at all subsequent times. Latent growth modeling is useful for understanding attitudinal and behavioral changes over time and also for forecasting future change.

*Latent class analysis.* When modelers are interested in segmenting a population into different homogeneous subsets (e.g., when conducting audience segmentation), a modeler can use SEM to define latent classes. The latent classes are factors whose indicators include a common set of variables. Factor loadings across variables and variable scores between cases are used to assign cases to classes. Class assignment can then be used to predict relevant outcomes.

**Further Reading**

Bowen, N. K., & Guo, S. (2012). *Structural equation modeling*. Oxford: Oxford University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Kline, P. (1994). *An easy guide to factor analysis*. New York: Routledge.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Tanaka, J. S. (1987). How big is big enough? Sample-size and goodness of fit in structural equation models with latent-variables. *Child Development, 58*(1), 134-146.

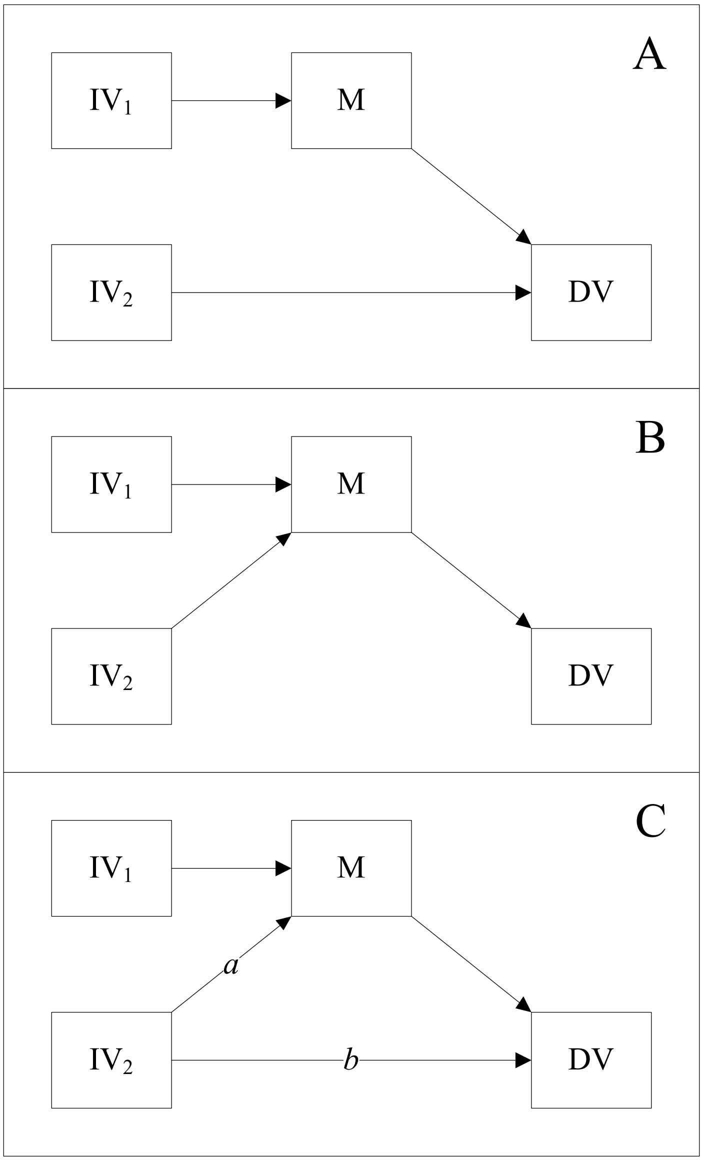


Figure 1. Examples of nesting among three structural models in which two independent variables ($IV_1$ and $IV_2$) and a mediator (M) predict a dependent variable (DV).

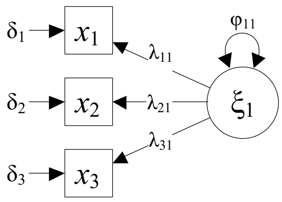


Figure 2. Confirmatory factor analysis of one factor with three indicators.

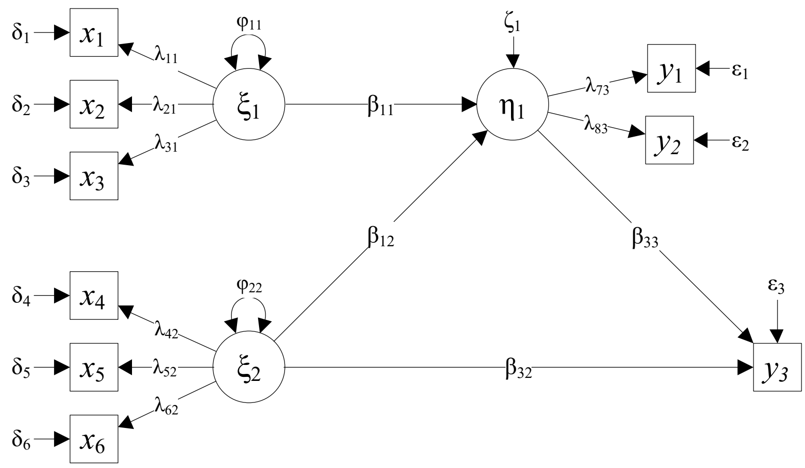


Figure 3. Structural equation model including exogenous and endogenous latent factors and an observed endogenous variable. Factor loadings (λ) indicate the measurement model and regression paths (β) indicate the structural model.