Clin Child Fam Psychol Rev (2014) 17:217–229
DOI 10.1007/s10567-014-0165-3

Why So Many Arrows? Introduction to Structural Equation Modeling for the Novitiate User

Olga V. Berkout · Alan M. Gross · John Young

Published online: 8 February 2014

Abstract Structural equation modeling (SEM) is the term for a broadly applicable set of statistical techniques that allow researchers to precisely represent constructs of interest, measure the extent to which data are consistent with a proposed conceptual model, and to adjust for the influence of measurement error. Although SEM may appear intimidating at first glance, it can be made accessible to researchers. The current manuscript provides a non-technical overview of SEM and its major constructs for a novitiate user. Concepts are illustrated using a simple example, representing a potential study performed in the field of youth and family research. The purpose of this manuscript is to offer interested scholars a conceptual overview and understanding of research questions and issues that may be addressed with this family of techniques.

Keywords Statistics · Structural equation modeling · Measurement · Confirmatory factor analysis

Introduction

Structural equation modeling (SEM) offers analytic flexibility to researchers working in the biomedical and behavioral sciences. SEM allows researchers to work with directly measured variables and latent factors to represent relationships among data.

Structural equation modeling encompasses a family of techniques. Although path analysis and latent growth curve modeling are considered members, the term SEM most commonly refers to confirmatory factor analysis (CFA) and structural regression (SR). A major advantage of this group of techniques is its ability to model underlying (latent) variables and error (MacCallum and Austin 2000). Latent variables (factors) can be thought of as constructs underpinning variable scores obtained by researchers (Brown 2006). Youth externalizing behavior, for example, is a widely studied construct within the behavioral sciences. Scholars can conceptualize a number of different behaviors as falling under the umbrella of externalizing (e.g., physical aggression, vandalism, and non-compliance). Specific expressions may vary, but these behaviors are all indicators of a common response group and could be thought of as caused by an externalizing factor. As such, youth responses to questions about aggression, theft, and vandalism would all be viewed as indicators of externalizing. Responses to these questions would be considered manifest (directly observed) variables. If these were statistically represented as caused by a youth externalizing factor, these responses would be discussed as factor indicators (as represented in Fig. 1).

In addition to asking different kinds of questions to serve as indicators of externalizing, we could combine information from various questionnaires, numerous respondents, or multiple measurement strategies (e.g., self report, parent report, and direct observation). Researchers interested in youth externalizing could gather information from different sources to get a more accurate representation of this construct. Parents, who see youth in a number of settings for an extended period of time, may have data that would not be observed by researchers in a discrete time period. Conversely, researchers could be freer from the bias parents may have toward their child, offering a unique perspective in their behavioral observations. Teachers contribute additional insight from their experience with

O. V. Berkout (✉) · A. M. Gross · J. Young
University of Mississippi, Oxford, MS, USA
e-mail: oberkout@gmail.com

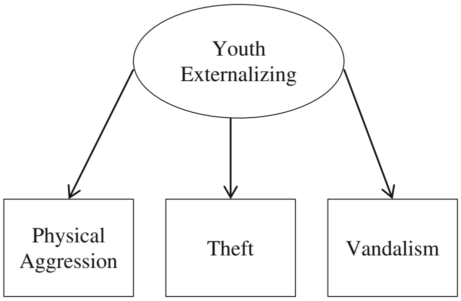

Fig. 1 Youth externalizing factor with question indicators

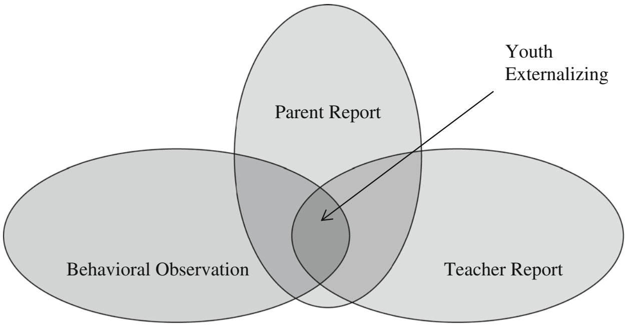

Fig. 2 Shared variability between parent report, teacher report, and behavioral observations of youth externalizing behavior

youth in the school setting. Combining information from various sources provides a more accurate assessment than if a single measure was used (MacCallum and Austin 2000).

Although researchers could create a summary variable of externalizing (e.g., adding standardized parent and teacher report questionnaires and a behavioral observation severity score), representing the latent factor causing these scores offers advantages over this more direct combinatorial approach. Latent factors represent the shared variability among their indicators (as seen in Fig. 2). This shared information is thought to represent the true construct (youth externalizing), rather than idiosyncratic aspects of each instrument. Latent factor models are conceptualized as psychological (or biological) constructs causing scores on measured indicators.

In addition to allowing for more accurate representation of constructs, latent variable modeling allows researchers to incorporate error into the model. Variables are rarely measured perfectly (e.g., due to misreading a question, inaccuracy in retrospective reporting, and inter-individual differences in responses to items). Error, inherent in all psychosocial measurement, thus likely affects obtained values and can lead to under- or overestimation of relationships between variables (DeShon 1998; Schmidt and Hunter 1999). Attenuations in estimated relationships due to lower internal consistency are adjusted for in SEM (DeShon 1998). Although this ability to correct is of great utility and can compensate for psychometric imperfections to some extent, SEM cannot counteract extensive problems (Kline 2011). Additionally, other sources of error not accounted for within the model can still have an effect similar to that seen in more traditional analytic methods (DeShon 1998).

It is easy to get excited about the benefits offered by SEM. Complex arrow and circle diagrams can sometimes take on alchemic properties, particularly when novitiate researchers initially begin utilizing this technique. It is important to note, however, that SEM loses its benefits outside conceptually sound applications, and although these approaches offer considerable advantages compared with traditional generalized linear modeling statistics, they are neither infallible nor magical in nature. For example, SEM is primarily a confirmatory technique (Byrne 2012), meaning that researchers have a solid theory and basic research supporting the proposed model. If this conceptual foundation does not exist to propel SEM analyses, then another technique would be more appropriate. There are mitigations to this general statement, in that causal relationships between latent variables in a less studied area may still be represented using exploratory SEM (described in Asparouhov and Muthen 2009), but a fairly coherent theoretical model is still needed. Although complex statistical techniques may be appealing or dazzling at first glance, it is important to recognize that choosing a simpler method may often be the better way to answer a research question, just as discretion may be the better part of valor.

Despite the above admonishments, SEM has been demonstrated to be a broadly applicable technique that will be a useful addition to the toolkits of many researchers. The purpose of this manuscript is to provide researchers with a broad understanding of SEM. Given the extensive scope of SEM, this paper is by no means all encompassing; however, this manuscript is intended to provide a conceptual and practical starting point. In order to demonstrate the principles of SEM, a computer-generated dataset will be used [created using Markov Chain Monte Carlo (MCMC) in Mplus 7.0]. Readers interested in using this dataset for practice may contact the corresponding author to request a copy. Authors specified relationships between variables in this example, and readers should know that the model will perform better than would typically be seen in applied samples. Values presented in the example will reflect those within our generated dataset, based upon roughly the strength of relationships that may be found within SEM studies. Additionally, while MCMC datasets provide a number of statistics useful in simulation studies, results will be reported in a manner consistent with an SEM analysis to facilitate demonstration.

Practical Issues in SEM

Structural equation modeling concepts will be demonstrated using a model based on Patterson et al. (1989) conceptualization of youth externalizing behavior. Patterson et al. (1989) proposed that poor parenting behaviors lead to the development of youth externalizing. Although a number of parenting behaviors are important, use of harsh and inconsistent discipline is thought to be particularly influential in shaping youth externalizing. Youth struggling with externalizing difficulties tends to behave aggressively with peers, leading to rejection by the normative peer group (Patterson et al. 1989). Use of harsh and inconsistent discipline is also likely to have some effect upon peer rejection; parents modeling ineffective social interactions may have youth who are less socially skilled and more likely to be rejected by others. A SEM model in which harsh and inconsistent discipline predicts youth externalizing and peer rejection, with youth externalizing affecting peer rejection, could be used to capture these relationships and will demonstrate concepts throughout this manuscript.

Notations

Structural equation modeling analysis usually begins with a visual diagram of the model. This allows researchers to ensure that all of the conceptual relationships have been represented and reduces otherwise complex networks of relationships to a more easily discernible form. To ensure uniformity and communication between researchers implementing SEM approaches, these diagrams use a common language. Oval shapes represent latent factors, and rectangular ones connote directly measured variables. A single-headed arrow indicates that one variable predicts another, whereas a double-headed arrow implies that the two variables covary or correlate (Hox and Bechger 1998; Kline 2011). Latent factors have single-headed arrows pointing at their indicators, representing factor influence on variable scores.

Measurement Model: CFA

In our example model, we are interested in harsh and inconsistent discipline, youth externalizing, and peer rejection. We use multiple measures of each construct in order to get a more accurate representation; parent report, teacher report, and behavioral observation data are gathered on all variables. First, we must make sure that data gathered through these different assessment strategies can be accurately thought of as caused by the appropriate underlying variables. This would be accomplished by testing whether the directly measured indicators of harsh and inconsistent discipline, youth externalizing, and peer rejection load onto their respective factors through a CFA (Fig. 3). Using SEM notation, our three factors are represented by ovals to indicate that these are latent variables. Directly measured factor indicators (e.g., parent report, teacher report, and behavioral observation assessments of each variable) are represented by rectangular shapes within Fig. 3. Factors have single-headed arrows pointing toward their indicators, because we are conceptually modeling factors as underlying constructs causing scores received on each measure. These represent regression paths within the model (i.e., indicators are regressed upon their respective factors). Variables without an arrow between them are not allowed to correlate within the model.

Structural Model: SR

While CFA is useful for confirmation of the way data should align given a theoretical conceptualization, SR extends the utility of SEM by allowing regression examinations among latent variables. This allows researchers to test predictive relationships between factors. In the context of a SR analysis, the relationships between factors and their indicators have been termed the measurement model (Byrne 2012). SR analysis combines the CFA model (creating latent variables) with the regression relationships, as represented within path analysis; the later portion has been discussed as the structural model within SEM (Kline 2011). Although it is possible to run these portions simultaneously, Kline (2011) recommends that the measurement model analysis be conducted prior to testing the structural model to allow researchers to check for potential problems. Within our example, the measurement model tested is represented within Fig. 3 and the structural regression (adding the structural model) within Fig. 4. Notice that in Fig. 4, we have single-headed arrows between our three factors indicating regression paths. We have posited that harsh and inconsistent discipline will predict externalizing behavior, which will in turn predict peer rejection. Use of harsh and inconsistent discipline is further expected to have a direct effect on peer rejection. In our measurement model in Fig. 3, the three factors are allowed to correlate, but are not regressed upon other factors.

Measurement Error

In addition to modeling relationships between factors and their indicators, SEM offers researchers the ability to represent error within the model. We add error terms to the visual representation of our model in Fig. 5 (these would have been estimated in the above models, but were not pictured for simplicity's sake). Error terms associated with factor indicators are marked with an E, whereas those influencing factors are represented with a D. This is

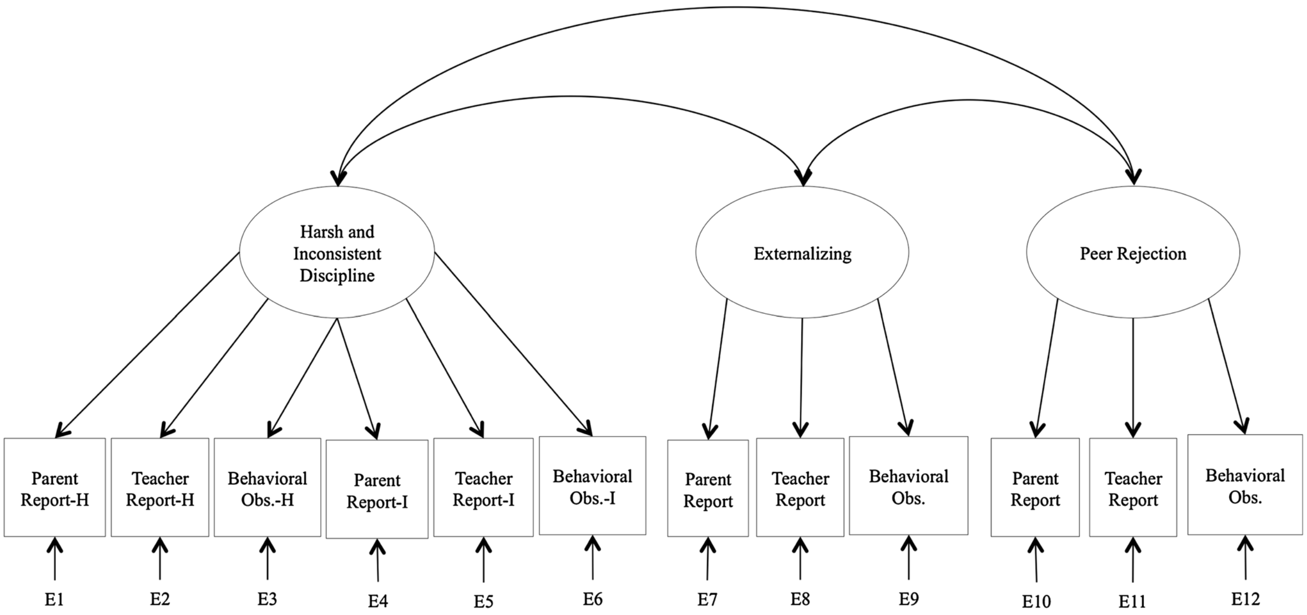

Fig. 3 Measurement model

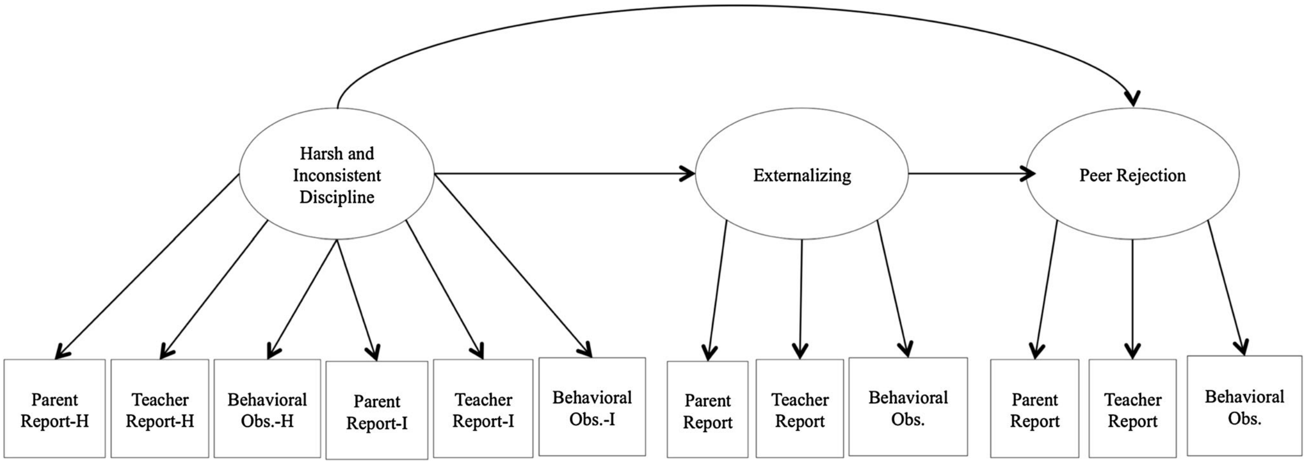

Fig. 4 Structural regression

because error terms associated with latent factors are called disturbances, whereas those of indicators (e.g., specific measures) are discussed as error (Kline 2011). Dependent (also termed endogenous) variables have associated error terms within SEM (Byrne 2012). Variables not affected by others in the model are discussed as exogenous and do not have error terms (Byrne 2012). Error terms have causal arrows pointing toward factors and their indicators in Fig. 5 to demonstrate their hypothesized influence. We expect obtained indicator scores to be predicted by both their underlying factors and error (e.g., unreliable measurement and chance variation, Kline 2011). Conceptually, this means that we would think of a score on "teacher report of youth externalizing" to be due to both actual youth externalizing behavior, error in reporting, and idiosyncratic aspects of teacher reports of externalizing that are not shared with other measurement strategies. If we believe that two variables share a source of measurement error (e.g., using a questionnaire requiring a reading level that was inappropriately high for participants in the sample), their error terms can be correlated.

Sample Size in SEM

Statistical analyses differ in the number of data points (often conceptualized as number of participants) they require to detect true existing relationships and obtain reliable estimates of these (Schreiber et al. 2006). Power is

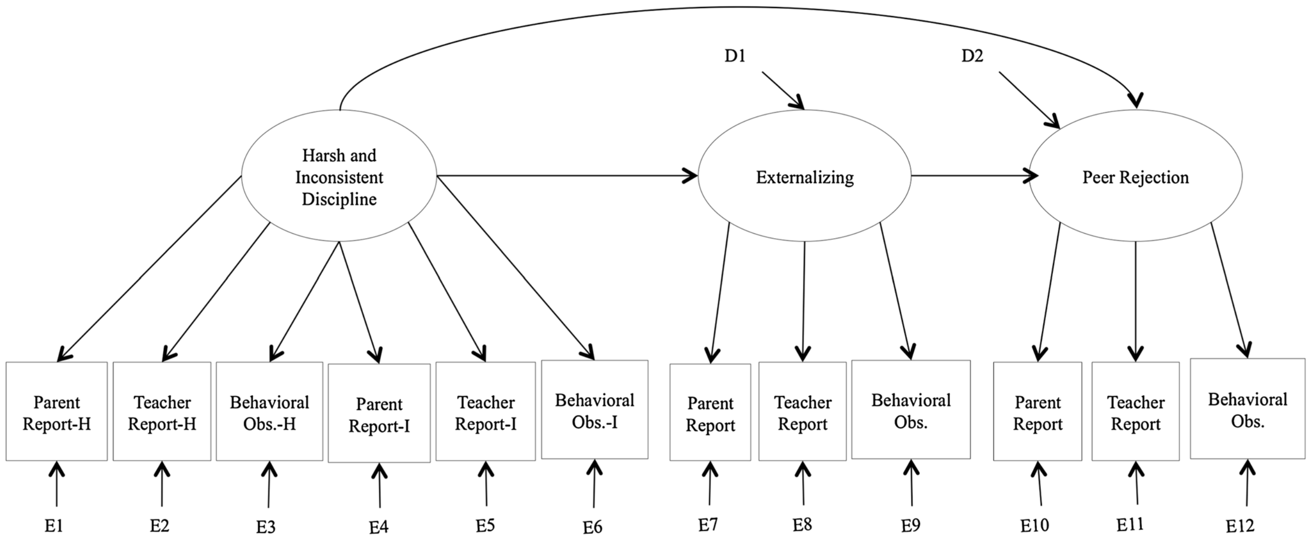

Fig. 5 Structural regression with error terms shown

the ability of a statistical test to detect an extant effect and is influenced by strength of the relationship between variables of interest (effect size), the level set for critical p value (i.e., alpha, typically ≤0.05), and sample size. In general, SEM is better suited to large samples, with at least 200 participants being common in published studies (Kline 2011; Ullman 2006). Some have put forth rules of thumb for determining the appropriate sample size. The N: q rule, as outlined in Kline (2011), is a popular convention for the most widely used maximum likelihood (ML) estimation. By this convention, a ratio of at least 10 (ideally 20) data points per parameter estimated is recommended. Rules like this can be helpful, but are not applicable or effective in all situations. Strength of the relationships between variables, the number of values that need to be estimated (e.g., causal paths between factors, variances of variables, and error variances), and the method of estimation all influence whether an extant effect will be detected (Kline 2011). This estimation is conducted in multivariate space and differs substantially from traditional power analysis with which readers may already be familiar.

Performing an a priori power analysis can ensure that researchers have a large enough sample to detect an effect. SEM tests a number of relationships between variables (parameters) and general fit of overall model (e.g., model fit indices, Hancock 2006). Power analysis in SEM can either focus on the ability to assess how well the model overall represents the data or on whether specific relationships (and other values) can be successfully subjected to tests of statistical significance (Hancock 2006). Conducting a power analysis for tests of overall model fit is a bit simpler and depends upon the number of degrees of freedom within a particular model (our example model has 51 degrees of freedom). Readers looking for a discussion of these issues are directed to Hancock (2006).

Although conducting a power analysis in SEM is more complex than doing so for other statistical techniques, the advantage is determining the approximate number of observations needed. In studies requiring extensive resources, researchers may want to know the minimum number of participants required. Alternatively, researchers who fail to find expected relationships may want to assess whether this was due to an overly small sample. Precise power analyses may be conducted, although this is not a common practice among published SEM manuscripts. It is more common to think of required sample size in terms of popular rules of thumb, such as N: q or aiming to gather at least 200 observations as common among published SEM studies (Kline 2011).

Model Identification

In SEM analysis, a number of values representing the obtained data and specified relationships are calculated (e.g., error terms, factor loadings, variances, and covariances, Bollen 1989; Hox and Bechger 1998). In order to obtain these, we must have sufficient data to get a unique solution for each parameter, such a model is said to be identified (Brown 2006). Analogously, if we were trying to determine the values of x and y from the equation $x + y = 15$, we would not have enough information to do so (Byrne 2012). There are any number of values for *x* and *y* that could be added to equal 15, and we cannot determine which would be more accurate. If such was the case in an SEM model, we would be unable to solve for each parameter and our model would be said to be under-identified (Brown 2006). The SEM equation used to build models is of course much more complex; however, the general concept is the same (Kline 2011).

In SEM, we want to have enough information to not only calculate parameters, but also to compare between different values to see which works better. If we had just enough information to get a single value for each parameter, our model would be just identified. In such a model, all possible relationships would be represented and the model would seem to fit the data perfectly (we would not have the room afforded by degrees of freedom to see whether this was indeed the case, see Kline 2011 for further discussion). In order to test whether specific relationships are a more accurate representation of the data, we need an overidentified model. This is the reason model identification is important to consider within SEM. Scholars should keep in mind that statistical software will enable computation of a model that is under-identified, but will provide invalid estimates (Kline 2011), obviating the utility of these advanced statistical approaches. Without careful attention to this somewhat intricate aspect of SEM, it is very possible to apply techniques correctly only to produce inaccurate results.

This manuscript will address identification of recursive models: those without correlated disturbances between latent factors or reciprocal causal paths between variables. Our example model falls into this category, as there are no correlated disturbance terms or reciprocal causation. Discussion of identification issues among non-recursive models is limited by the introductory scope of this manuscript; however, a highly readable discussion may be found in Kline (2011).

Identification among recursive models is relatively simple. The model must have more than 0 degrees of freedom, each factor must have at least three indicators (or two indicators and multiple factors in the model), and latent factors need to be given a scale (Kline 2011). Latent factors must be scaled either through unit loading or unit variance identification (ULI and UVI, respectively, Kline 2011). Factors are estimated on the basis of manifest variable scores and do not have an inherent metric of their own. In order for the model to be identified and factor scores to be interpreted, they must be given a scale (Hox and Bechger 1998). Generally, the factor is fixed to be on the same scale as one of the directly measured indicators [this is called the unit loading identification (ULI), Brown 2006; Kline 2011]. In this case, we are setting the path between the factor and one of its indicators to be equal to 1, with the factor taking on the indicator's scale. ULI was used in our example model, with unstandardized paths between parent report of parenting skills, parent report of youth externalizing, and parent report of youth social skills being fixed to 1 in order to give the associated latent factors their scale.

Alternatively, the variance of the latent factor can be fixed to 1; this is called unit variance identification (UVI, Brown 2006; Kline 2011). ULI is generally used more frequently and serves as the default setting in most programs (Kline 2011). ULI should be used anytime there may be changes in overall variability between independent samples or over the course of longitudinal measurement (Kline 2011).

Model Estimation

Data, Distribution, and Decisions

Structural equation modeling makes a number of assumptions about the data being analyzed, which are influenced by the method used to calculate its parameters. Estimator functions are equations used to determine numeric values representing relationships within the model.

Many estimators assume that data are continuous and multivariate normal (or that ordinal category scores represent values of a normally distributed continuous variable). Software can be used to examine univariate skew and kurtosis of individual variables. Multivariate skew and kurtosis can similarly be assessed using macros (SPSS syntax for computing values) from Lawrence DeCarlo's Columbia University home page (http://www.columbia.edu/~ld208/, as suggested by Finney and DiStefano 2006). Additionally, SEM assumes that observations are independent and that participants are a random sample of the population (Bentler and Chou 1987). Linear relationships among variables are expected (Bentler and Chou 1987), although quadratic relationships can be incorporated into more complex models (see Marsh et al. 2006 for a discussion). It is again important to note that standard computation programs will often compute values even when assumptions have not been met, providing inaccurate results. Discussions of assumptions have been unfortunately absent from many recent SEM studies (Schreiber et al. 2006), and readers are urged not to follow this trend when applying these techniques in their own work.

Missing Data Issues

Researchers should consider and report the proportion of missing data within their sample. Missing data can be handled in a number of ways. Tabachnick and Fidell (2007) suggest that data missing less than 5 % likely require no further treatment for most analyses. However, a number of more sophisticated methods exist, chief among these maximum likelihood estimation (ML, Brown 2006; Graham 2009). ML is sometimes called full information maximum likelihood, because it uses all of the information in the dataset to estimate parameters (Enders 2001; Schlomer et al. 2010).

Dealing with missing data presents a major issue within applied statistics (see Little and Rubin 2002 for an overview). Strategy for handling missing data is determined by the pattern responsible for its absence. Data are categorized as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Under MAR, missingness of values depends on available data (Graham 2009). Whether data will be missing is not related to the value of the missing data and can be predicted by other variables in the dataset (Enders 2001). Data missingness that does not depend on either values in the dataset or the values of the missing data is said to be MCAR (Graham 2009). When data are NMAR, whether or not the data are missing depends upon their values (Graham 2009). The latter case introduces complications for replacement or estimation of values that are missing.

Maximum likelihood assumes that data are at least MAR (Brown 2006). Although the MCAR assumption may be assessed using Little's MCAR test (available in SPSS), there is no way to statistically test whether data are MAR (Brown 2006). Scholars are advised to consider potential aspects of research design and data collection when deciding whether data may be NMAR. Data that are NMAR depend on these variables not in the dataset, potentially because a confound that was not assessed is influencing the presence of data missingness (Graham 2009). Analytic difficulties related to NMAR data may be addressed by modeling the process responsible for missingness (see Enders 2011 for a discussion).

Choosing an Estimator

Structural equation modeling produces parameter values, which minimize the discrepancy between observed data (i.e., what was collected and measured) and data implied by the model (i.e., what would be expected as extrapolated from observed data, Olsson et al. 2000). A number of methods for estimating parameters are available. Estimator functions vary in robustness against assumption violations (distribution, model misspecification, etc.) and data for which they are appropriate (e.g., continuous vs. ordinal). Given the scope of this manuscript, only a limited number of estimator issues will be discussed. We will focus on the broadly applicable maximum likelihood estimator and on the polychoric correlation-based weighted least squares, suitable for working with ordinal data.

Given its ability to deal with missing data (Brown 2006; Enders and Bandalos 2001), ML is commonly the estimation method first considered by researchers. Missing values for each case are estimated on the basis of available data; these are then used in parameter determination (Enders 2001). ML essentially estimates parameter values that are most probable given the obtained data and specified model. Parameters are thought of as drawn from a normally distributed set of all possible parameter values, and those that have the highest probability of occurring are selected.

Maximum likelihood assumes that endogenous variables (those predicted by others in the model) are multivariate normal, continuous, and have missing values that are at least MAR (Kline 2011). Simulation studies have demonstrated that univariate kurtosis values nearing seven, univariate skew values approaching two, and multivariate kurtosis values greater than three may be problematic for ML (Finney and DiStefano 2006). If assumptions of ML are not met, inaccurate results will be obtained. Applying ML significance tests to non-normally distributed data can lead to the occurrence of type I errors (e.g., concluding a relationship exists where one does not, Finney and DiStefano 2006).

Robust ML (MLR) has been developed to compensate for inaccurate findings resulting from use of non-normally distributed data. MLR provides adjusted statistics of overall model fit and modified standard errors (Brown 2006; Kline 2011). Brown (2006) suggests that MLR be used with non-multivariate normal continuous data. Byrne (2012) also argues that MLR may be used with ordinal data provided it has at least five categories (e.g., Likert-type scales). Samples of at least 400 participants have been suggested as a rule of thumb for MLR estimation (Schermelleh-Engel et al. 2003).

Although ML has a number of advantages, its iterative process is susceptible to local maximum values. The ML algorithm starts out with a set of best guesses for model parameters and then continues to modify to improve upon these until it cannot do so further (Kline 2011). However, if the initial values are off to a substantial degree, ML may run into problems. Rindskopf (1998) describes the ML iterative process as analogous to a person with a bucket on his head trying to find the highest point around. The person only sees a small section of the ground beneath his feet when looking for the highest point and may miss the hill off to the side if he gets stuck on a smaller elevation (Rindskopf 1998). This issue has been termed as the local maximum problem, and researchers can manually select starting values for parameters (i.e., override the computer's guesses) to try to deal with this issue, as well as general failure of the algorithm to converge (Kline 2011). For a more nuanced and practical approach to this problem, Kline (2011) offers a helpful discussion on starting value selection.

Weighted least squares (WLS) and robust weighted least squares (WLSMV) estimation techniques are aimed at handling ordinal data (Flora and Curran 2004). Both utilize polychoric correlations, which measure relationships between ordinal variables assuming categories serve as values of an underlying continuous variable (Flora and Curran 2004). WLS was initially developed for this type of data; however, it can run into calculation difficulties with a large number of observed variables and requires large samples ($n > 1{,}000$; Flora and Curran 2004; Schermelleh-Engel et al. 2003). WLSMV is a modification of WLS created to deal with these problems, providing superior performance in small to medium samples (Byrne 2012). Hence, we agree with Brown's (2006) recommendation that WLSMV be used with ordinal data if possible; however, to the best of our knowledge, this estimator is currently only available in Mplus software (Byrne 2012), potentially limiting some researchers' ability to make use of it.

Choosing between estimators depends upon data characteristics and distribution. Within our example analysis, all variables were continuous and normally distributed, allowing us to use ML estimation. If our data had been ordinal, we would have chosen between WLSMV and MLR; although WLSMV is specifically designed for ordinal data, MLR would allow for greater flexibility in model comparison (i.e., non-nested models). Use of MLR for ordinal data becomes less problematic as the number of categories increases and the distribution approaches normal.

Interpreting Results

Model Fit

Model fit refers to the extent to which the pattern of relationships implied by the researcher's model is obtained within the data. Fit indices provide estimates for overall fit of the model. Notably, these statistics provide an average measure of model fit, rather than information on specific aspects (these may still be inaccurate, Kline 2011). Model fit is evaluated by examining a number of fit indices, rather than focusing on a single one. Reporting multiple indices of fit is generally recommended as indices have different strengths and weaknesses with regard to various data characteristics and areas of model misspecification (Bentler 2007). ML estimation was used to obtain parameters and calculate fit of our model within the generated data.

A widely used measure of model fit is the exact fit chi square (χ^2). χ^2 assesses whether there is a statistically significant difference between the data pattern obtained and that implied by the model (Kline 2011). χ^2 is testing whether there is a statistically significant difference between the observed and predicted data: the null model is that there is no difference. Thus, larger significant χ^2 values indicate greater model misfit. A non-statistically significant χ^2 on the other hand would suggest that there were no differences beyond chance between the data predicted by the researcher's model and obtained (i.e., providing support for the model, or essentially perfect fit). The χ^2 statistic is influenced by sample size, with scholars arguing that very minor differences could appear statistically significant in a large enough sample due to chance alone (Brown 2006). Due to this concern, Brown (2006) suggests that other fit indices be given stronger consideration in applied studies. Kline (2011) argues that the χ^2 needs to be given serious weight and its rejection can lead to the acceptance of incorrect models. Specifically, Kline (2011) notes that cut offs for other indices vary in their ability to reject a poorly fitting model depending on the model and sample characteristics. While Kline (2011) offers a valid criticism, Brown's (2006) perspective is generally more common among published SEM studies.

The Bentler comparative fit index (CFI, Bentler 1990) is a widely used incremental index of model fit (Hooper et al. 2008). The CFI assesses the extent to which the predicted model is better than one in which variables are completely independent of each other (Kline 2011). The null hypothesis for this statistic is that there is no difference between the proposed model and a model where there are no relationships between variables. The greater the CFI value, the more the researcher's model offers an improvement over total lack of relationships: CFI ranges from 0 to 1 (Brown 2006). Monte Carlo simulations suggest that CFI values above 0.95 generally suggest close fit (Hu and Bentler 1999). Values greater than 0.90 may indicate acceptable model fit, particularly if other indices support the model (Brown 2006). CFI has the advantage of correcting for sample size and is thought to provide an accurate assessment of overall fit even in small samples (Hooper et al. 2008).

Standardized root-mean-squared residual (SRMR) gives a measure of the discrepancy between observed and predicted correlations between variables (Brown 2006). SRMR can be conceptualized as the average error in relationship prediction, and the bigger the error, the more mistakes the model is making. The null model could be conceptualized as a lack of error in model prediction, and the bigger the SRMR, the more errors (on average) the model is making. Smaller SRMR values indicate better fit. As it is standardized, SRMR can range between 0 and 1 (Brown 2006). Hu and Bentler (1999) suggest that values less than 0.08 indicate good model fit. Others argue that SRMR less than 0.05 is needed to conclude the model fits well (Hooper et al. 2008). SRMR tends to perform worse with categorical data and to be generally lower in large samples and models with many parameters (Brown 2006; Hooper et al. 2008). Kline (2011) generally suggests that inspecting the matrix of residuals is more helpful than examining the SRMR, although this remains a popular index to report.

Root-mean-square error of approximation (RMSEA; Steiger and Lind 1980) measures the misfit of the model while adjusting for model complexity and sample size (Brown 2006). The null hypothesis for the RMSEA statistic is that the proposed model does not differ from that supported by the data (MacCallum et al. 1996). The bigger the RMSEA the greater the discrepancy between the data and the researchers model; smaller RMSEA values suggest better model fit (Kline 2011). Kline (2011) suggests that values less than 0.05 indicate good fit, although Hu and Bentler (1999) propose that a cutoff of 0.06 be used. Schermelleh-Engel et al. (2003) provide some interpretive guidelines for higher values: that RMSEA less than 0.08 may be viewed as representing adequate fit and that RMSEA less than 0.10 may still be acceptable although it is far from ideal. MacCallum and Austin (2000) argue that RMSEA be given stronger consideration than other indices, because it provides confidence intervals and has received support in simulation studies. Although RMSEA has been extensively studied and has a number of advantages, it can sometimes result in type II error (rejecting true model) in small samples (Brown 2006). The indices discussed above are popular and widely used, but by no means all encompassing. There is a plethora of other measures of model fit available to the interested reader within the SEM literature (see Hooper et al. 2008; Schermelleh-Engel et al. 2003 for further overview).

Numerous methods of assessing model fit allow us to draw on strengths of varied indices in statistical evaluation. However, a well-fitting model may not accurately represent the true state of the world: the case of alternative equivalent models offers an illustrative example of this point (MacCallum et al. 1993). Models with the same implied covariance matrices, and number of restrictions may fit equally well, regardless of the direction of causal relationships (Tomarken and Waller 2003). Thus, a model in which variable A predicts variable B could fit as well as one where variable B predicts variable A, although these would have discrepant theoretical implications. Nonequivalent alternative model structures with differing covariance matrices may similarly fit as well as the model proposed by researchers, and a well-fitting model may omit key variables (Tomarken and Waller 2003). Parameter estimates drawn from a model that does not accurately represent the phenomena of interest would offer scholars an inaccurate perspective on variable relationships (Henley et al. 2006; Tomarken and Waller 2003). Scholars are urged to consider these conceptual issues during analysis and interpretation of model fit.

Continuing the applied example from previous sections, we can examine fit of our current model by examining various fit indices. Following the suggested two-step process, we first run a CFA, making sure that the variables load onto their hypothesized factors as predicted. We expected to obtain three latent factors of harsh and inconsistent discipline, youth externalizing, and peer rejection. Our generated sample had 500 participants and 100 MCMC datasets. ML was used to analyze the model, and no missing data were present. Model fit information was as follows: $\chi^2 = 50.26$, $df = 51$, $p > 0.10$, CFI = 0.998, SRMR = 0.02, RMSEA = 0.01 (95 % CI 0.00–0.02 *Note RMSEA confidence intervals are not provided with MCMC datasets with MPlus, and these values were calculated by authors). In the current example, all of the indices indicate that this model provides a close fit for the data. The χ^2 is not statistically significant, indicating that the proposed model is not discrepant from the obtained data. CFI is well over 0.95, suggesting that our model is a much better representation of relationships than one where all variables were independent. SRMR is below the smallest suggested cutoff of 0.05, and its value indicating that the average of standardized error terms in the model is about 0.02. RMSEA is similarly small and falls below the lowest suggested value of 0.05, letting us know that data does not deviate a great deal from what would be predicted by the model.

Although this model provided a close fit for that data, we may have conceptualized the parenting variables as falling onto separate harsh discipline and inconsistent discipline factors. Testing this measurement model, we also find a close fit for the data: $\chi^2 = 47.27$, $df = 48$, $p > 0.10$, CFI = 0.998, SRMR = 0.02, RMSEA = 0.01 (95 % CI 0.00–0.02), even though we know that we had specified that the harsh and inconsistent scales fall onto a single factor during data generation. This example highlights the importance of considering alternative models with equivalent fit in SEM. Despite close fit of both models, we note that the four factor model does not provide a superior fit to the three factor model (see subsequent section for a discussion of model comparison). Based on our conceptualization of harsh and inconsistent discipline as a single factor and considering the superior parsimony of such a model, we use the original three factor conceptualization for the structural regression.

We hypothesize that harsh and inconsistent discipline will predict youth externalizing behavior and peer rejection and that externalizing would also have a direct effect on peer rejection. Model fit information is as follows: $\chi^2 = 50.26$, $df = 51$, $p > 0.10$, CFI = 0.998, SRMR = 0.02, RMSEA = 0.01 (95 % CI 0.00–0.02). As discussed above, these values indicate that the model in general closely represents the data.

Local Areas of Misfit

Model fit indices provide information on the model as a whole: however, even a model that overall appears to describe the data well may have major problems in specific areas (Brown 2006). Modification indices and residual values offer information on parts of the model that may not fit well (Brown 2006; Kline 2011). The residual matrix represents the difference between values observed and predicted by the model (Kline 2011). This matrix is a collection of numbers displaying the prediction error in each relationship specified by the model (Brown 2006). Individual residual values can be examined to see which relationships are problematic. Standardized residual values can be considered analogous to z scores, which represent departure from a perfect model in standard deviations (Brown 2006). As such, standardized residual values can be subjected to statistical testing to see whether a significant error exists. Varied cutoffs for this value have been suggested, with some researchers arguing that a number of values greater than 1.96 (significant at $p < 0.05$) may indicate that a model change is in order (Anderson and Gerbing 1988). Brown (2006) notes that others suggest that values greater than 2.58 (significant at $p < 0.01$) may be considered extreme, particularly in bigger samples where standardized residual values will tend to be larger. A negative residual suggests that a relationship between two variables has been overestimated, whereas a positive residual suggests that it has been underestimated (Brown 2006; Kline 2011).

Modification indices provide information on how much better fitting a model would be if a particular change were made (specifically the estimated decrease in χ^2, Hox and Bechger 1998). As the larger χ^2 values represent greater discrepancy between model predicted and actual data, decreases in χ^2 would mark improvements in model fit. Like the χ^2, modification indices are influenced by sample size and are more likely to be larger in bigger samples (Brown 2006). Information available from standardized residuals and modification indices may make it tempting to make a number of changes on the basis of statistics alone; however, changes made to improve model fit without consideration for theory can lead to a nonsensical model, which is unlikely to be replicated in another sample (Brown 2006). Scholars are urged to consider any changes they may be making from a conceptual perspective. Unfortunately, modification index statistics and standardized residuals are not available with MCMC datasets (and thus our example); however, most SEM programs will allow researchers to request this option.

Parameters in the Model

A number of relationships between variables are estimated within an SR model. Values associated with relationships between variables are termed path coefficients. Path coefficients are similar to standardized and unstandardized coefficients in a multiple regression and may be interpreted as such (Byrne 2012; Hox and Bechger 1998). Among standardized factor loadings, paths may be squared to obtain how much variance in the indicator a factor can explain, so long as the indicator is only caused by a single factor (Kline 2011). If multiple factors cause an indicator, the standardized loadings are analogous to regression beta weights (Kline 2011). Double-headed arrows represent covariances between variables (if unstandardized) and correlations (if standardized, Hox and Bechger 1998). Along with path coefficients, researchers obtain standard errors associated with each path. Dividing a path coefficient by its standard error provides a z test statistic, which can be used to test statistical significance (Ullman 2006).

We can examine values in Fig. 6 to get a sense of relationships within our example model. We can see that indicators of the harsh and inconsistent discipline factor range in standardized loading size from 0.50 to 0.77. Because these variables load only onto harsh and inconsistent discipline, we can square these values to obtain the amount of variance parenting skill can explain in each of its indicators (around 25 to 59 %). Loadings on externalizing range from 0.59 to 0.75 and those on peer rejection between 0.65 and 0.84. These values can similarly be squared to provide further information: the latent variable of externalizing accounts for approximately 35 to 56 % and peer rejection for about 42 to 71 % of the variance in their respective indicators. In addition to factor loadings, relationships between factors are represented in the model. We expected harsh and inconsistent discipline to have a positive effect on youth externalizing and youth peer rejection. Youth externalizing was hypothesized to predict peer rejection. The pattern of relationships was as expected. Peer rejection was predicted by youth externalizing (0.45) and harsh and inconsistent discipline (0.44). Youth externalizing was predicted by harsh and inconsistent discipline (0.49). These coefficients are standardized and may be interpreted similarly to beta weights in a multiple regression.

One of the main advantages of SEM is its ability to adjust for error. Along with the values we have obtained for relationships between factors and their indicators, SEM provides estimates for residual variances of each variable. These are represented as predicting their respective variables. The residual terms let us know how much of the

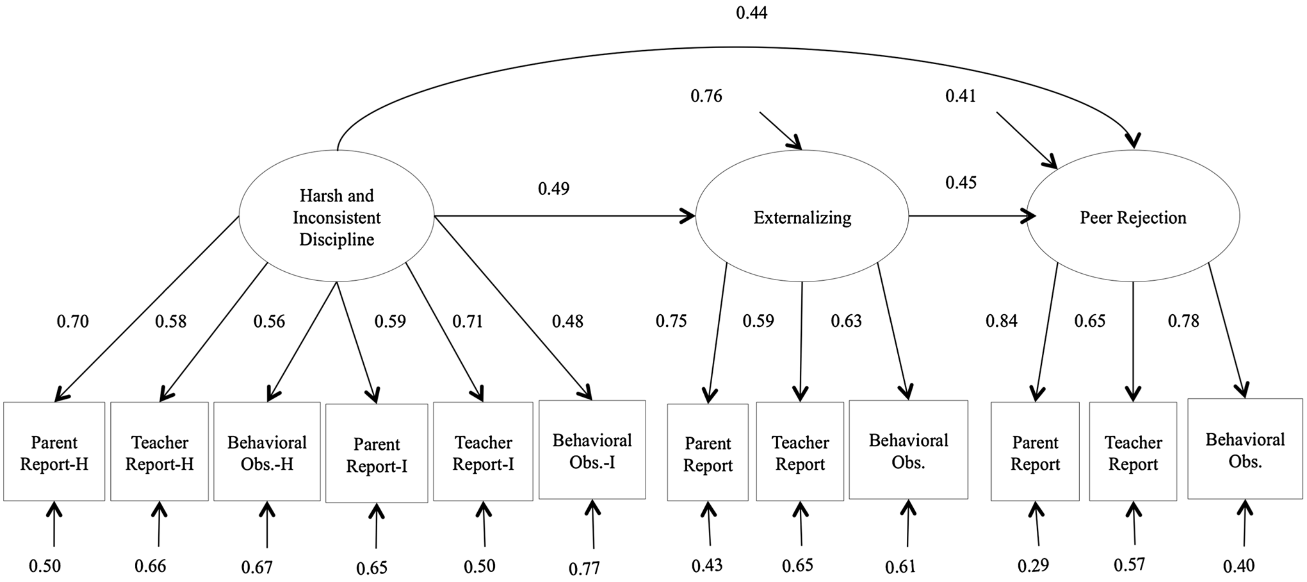

Fig. 6 Structural regression with parameter estimates

variance in each particular variable remains unexplained within the model. Residual terms for indicators range from 0.29 to 0.77. Larger residual variance terms suggest that some indicators are not accounted for as well by their latent factors (or perhaps were more prone to measurement error). Residual values (disturbances) for endogenous latent factors are also represented.

Comparing Alternative Models

Although SEM is primarily a confirmatory technique, scholars may compare several models to determine which one fits the data best. Models being compared can fall under two categories: nested and non-nested. When a model is nested under another, it is a more restricted version of its parent model (Brown 2006). Relationships that are allowed to exist in the parent model are fixed to zero in the nested model (Bentler and Chou 1987). Examples of nested models would include those with the same number of indicators but varied numbers of factors, a model that specifies an additional causal relationship between variables versus one that does not, a model that sets a constraint on factor loadings to be equal across gender versus one that allows those to vary. Nested models can be statistically compared using χ^2 difference testing (Bentler and Chou 1987).

Let us say for example that we wanted to compare the model described above to an alternative. In the initial model, harsh and inconsistent discipline predicted youth externalizing and peer rejection, with externalizing behavior having a direct effect on peer rejection. In the alternative model, harsh and inconsistent discipline predicts externalizing and externalizing predicts peer rejection. In this case, the path between harsh and inconsistent discipline and peer rejection could be thought of as constrained to zero. In the parent model, harsh and inconsistent discipline is allowed to have an influence on peer rejection, but in the nested model this relationship is not permitted. The procedure for χ^2 difference testing is fairly straightforward. To conduct a difference test, we calculate the difference between the χ^2 values and degrees of freedom for the two models (Brown 2006). The resultant χ^2 value with its degrees of freedom could be assessed for statistical significance using a χ^2 table of critical values. For our example, the parent model has a χ^2 of 50.26 with 51 degrees of freedom. The nested model, where the relationship between harsh and inconsistent discipline and peer rejection is removed, has a χ^2 of 104.35 with 52 degrees of freedom. The difference χ^2 is equal to 54.09 with 1 degree of freedom, which is statistically significant at $p < 0.001$. Remember that in the context of evaluating model fit, higher χ^2 values indicate *greater* deviation between the proposed model and data. When we take the difference of the two χ^2 values, we are looking to see whether the statistic associated with the parent model is significantly smaller than that of the nested model. A smaller χ^2 value suggests that there is less discrepancy between model and data.

Non-nested models cannot be statistically compared using χ^2 difference testing (Kline 2011), although a number of other measures exist. Notably, researchers should consider that models that allow for more relationships can appear to fit the data better regardless of whether this accurately represent the true state of the world (Bollen

1989). To correct for this issue, indices used to compare non-nested models adjust for model complexity. A number of indices of model fit are suitable for comparing non-nested models; among them, Akaike information criterion (AIC), Bayes information criterion (BIC), and expected cross-validation index (ECVI; Schreiber et al. 2006). AIC and ECVI both adjust for model complexity, with AIC adjusting for the number of model parameters and ECVI adjusting for complexity and sample size (Brown 2006; Vrieze 2012). Lower AIC, BIC, and ECVI values indicate better model fit (Brown 2006; Vrieze 2012). BIC adjusts its evaluation of the model for both the number of parameters and observations within the sample (Vrieze 2012). Vrieze (2012) advocates for the use of BIC due to its guaranteed ability to select the correct model given a large enough sample and provided that the true model is among those tested. A number of other fit indices exist, and statistical software varies in the indices provided. Interested readers are referred to Rust et al. (1995) for a more extensive discussion of model comparison. Given our use of Mplus, we present the available AIC and BIC measures of model fit.

Although AIC and BIC do not allow for statistical significance testing, the model with the smallest AIC and BIC values is more likely to be replicated in an independent sample, (Kline 2011). The AIC value of our parent model was 14,866.68, and the AIC for our nested model was 14,918.780. The BIC value of our parent model was 15,031.05, and the BIC for our nested model was 15,078.94. In both cases, the parent model was suggested to provide a closer fit, consistent with our χ^2 difference test. Allowing for a relationship between harsh and inconsistent discipline and peer rejection was found to be the better representation.

Conclusion

This manuscript offers an introduction to a flexible and useful statistical technique. We hope that this brief description of the relevant issues has given scholars an understanding of basic concepts and a desire to learn more, perhaps even applying SEMs in their own research. SEM can be particularly useful in studies concerning youth, where directly measured data from multiple reporters can be represented using latent variables. Researchers can also apply techniques to experimental and longitudinal studies and are encouraged to continue exploring these issues in their own work. In order to encourage further learning, we offer resources we have found helpful below. These are by no means all encompassing or the only good sources for SEM; however, these are tools we have found helpful in our quest to understand and apply SEM techniques.

Resources

- Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling, Third Edition.* New York, NY: Guilford Press. A well-written introductory text covering the basics in SEM in an easy to understand fashion for the novice user.
- Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press. Comprehensive and readable book on confirmatory factor analysis (with most topics being applicable to structural regression models).
- Hancock, G.R. & Mueller, R.D. (2006). *Structural Equation Modeling: A Second Course*. Greenwich, CT: Information Age Publishing. A more advanced discussion of practical issues within SEM and deviations from basic models (e.g., non-recursive models, categorical data, and non-linear relationships).
- Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ: Wiley-Interscience. An advanced discussion of mathematical procedures and issues within SEM.
- SEMNET listserv (accessible at http://www2.gsu.edu/~mkteer/semnet.html)—a popular email listserv scholars throughout the world use to discuss conceptual and practical issues associated with SEM.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423.

Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*, 825–829.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural equation modeling. *Sociological Methods Research, 16*(1), 78–117.

Bollen, K. A. (1989). *Structural equations with latent variables*. Hoboken, NJ: Wiley-Interscience.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Taylor & Francis Group.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*(4), 412–423.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128–141.

Enders, C. K. (2011). Missing not at random models for latent growth curve analysis. *Psychological Methods, 16*, 1–16.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age Publishing.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.

Henley, A. B., Shook, C. L., & Peterson, M. (2006). The presence of equivalent models in strategic management research using structural equation modeling: Assessing and addressing the problem. *Organizational Research Methods, 9*(4), 516–535.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53–60.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review, 11*, 354–373.

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, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*(1), 185–199.

Marsh, H. W., Wen, Z., & Hau, K. T. (2006). Structural equation models of latent interactions and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 225–265). Greenwich, CT: Information Age Publishing.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*(4), 557–595.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist, 44*(2), 329–335.

Rindskopf, D. (1998). *Explaining maximum likelihood estimation*. Retrieved from http://www.rasch.org/rmt/rmt1237.htm.

Rust, R. T., Lee, C., & Valente, E. (1995). Comparing covariance structure models: A general methodology. *International Journal of Research in Marketing, 12*, 279–291.

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23–74.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*(1), 1–10.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183–198.

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research, 99*(6), 323–337.

Steiger, J. H. & Lind, J.C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society: Iowa City, IA.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with 'well fitting' models. *Journal of Abnormal Psychology, 112*(4), 578–598.

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment, 87*(1), 35–50.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. doi:10.1037/a0027127.