Structural Equation Modeling (SEM) is a statistical approach to testing hypothesis about the relationships among observed and latent variables. The use of SEM in research has increased in psychology, sociology, and economics in recent years. In particular, educational researchers try to obtain the complete image of the process of education through the measurement of personality differences, learning environment, motivation levels and host of other variables that affect the teaching and learning process. With the use of survey instruments and interviews with students, teachers and other stakeholders as a lens, educators can assess and gain valuable information about the social ecology of the classrooms that could help in improving the instructional approach, classroom management and the learning organizations. A considerable number of research have been conducted to identify the factors and interactions between students' characteristics, personal preferences, affective traits, study skills, and various other factors that could help in better educational performance. In recent years, educational researchers use Structural Equation Modeling (SEM) as a statistical technique to explore the complex and dynamic nature of interactions in educational research and practice. SEM is becoming a powerful analytical tool and making methodological advances in multivariate analysis. This book presents the collective works on concepts, methodologies and applications of SEM in educational research and practice. The anthology of current research described in this book will be a valuable resource for the next generation educational practitioners.
APPLICATION OF STRUCTURAL EQUATION MODELING
IN EDUCATIONAL RESEARCH AND PRACTICE
CONTEMPORARY APPROACHES TO RESEARCH
IN LEARNING INNOVATIONS

Volume 7

Series Editors:
Myint Swe Khine – Curtin University, Australia
Lim Cher Ping – Hong Kong Institute of Education, China
Donald Cunningham – Indiana University, USA

International Advisory Board
Jerry Andriessen – University of Utrecht, the Netherlands
Kanji Akahori – Tokyo Institute of Technology, Japan
Tom Boys – London Metropolitan University, United Kingdom
Thanasis Daradoumis – University of Catalonia, Spain
Arnold Depickere – Murdoch University, Australia
Roger Hartley – University of Leeds, United Kingdom
Victor Kaptelinin – Umea University, Sweden
Paul Kirschner – Open University of the Netherlands, the Netherlands
Konrad Morgan – University of Bergen, Norway
Richard Oppermann – University of Koblenz-Landau, Germany
Joerg Zumbach – University of Salzburg, Austria

Rationale:
Learning today is no longer confined to schools and classrooms. Modern information and communication technologies make the learning possible anywhere, any time. The emerging and evolving technologies are creating a knowledge era, changing the educational landscape, and facilitating the learning innovations. In recent years educators find ways to cultivate curiosity, nurture creativity and engage the mind of the learners by using innovative approaches.

Contemporary Approaches to Research in Learning Innovations explores approaches to research in learning innovations from the learning sciences view. Learning sciences is an interdisciplinary field that draws on multiple theoretical perspectives and research with the goal of advancing knowledge about how people learn. The field includes cognitive science, educational psychology, anthropology, computer and information science and explore pedagogical, technological, sociological and psychological aspects of human learning. Research in this approaches examine the social, organizational and cultural dynamics of learning environments, construct scientific models of cognitive development, and conduct design-based experiments.

Contemporary Approaches to Research in Learning Innovations covers research in developed and developing countries and scalable projects which will benefit everyday learning and universal education. Recent research includes improving social presence and interaction in collaborative learning, using epistemic games to foster new learning, and pedagogy and praxis of ICT integration in school curricula.
Application of Structural Equation Modeling in Educational Research and Practice

Edited by

Myint Swe Khine
Science and Mathematics Education Centre
Curtin University, Perth, Australia
# TABLE OF CONTENTS

## Part I – Theoretical Foundations

Chapter 1  
Applying Structural Equation Modeling (SEM) in Educational Research: An Introduction  
*Timothy Teo, Liang Ting Tsai and Chih-Chien Yang*  

Chapter 2  
Structural Equation Modeling in Educational Research: A Primer  
*Yo In‘nami and Rie Koizumi*  

## Part II – Structural Equation Modeling in Learning Environment Research

Chapter 3  
Teachers’ Perceptions of the School as a Learning Environment for Practice-based Research: Testing a Model That Describes Relations between Input, Process and Outcome Variables  
*Marjan Vrijnsen-de Corte, Perry den Brok, Theo Bergen and Marcel Kamp*  

Chapter 4  
Development of an English Classroom Environment Inventory and Its Application in China  
*Liyan Liu and Barry J. Fraser*  

Chapter 5  
The Effects of Psychosocial Learning Environment on Students’ Attitudes towards Mathematics  
*Ernest Afari*  

Chapter 6  
Investigating Relationships between the Psychosocial Learning Environment, Student Motivation and Self-Regulation  
*Sunitadevi Velayutham, Jill Aldridge and Ernest Afari*  

Chapter 7  
In/Out-of-School Learning Environment and SEM Analyses on Attitude towards School  
*Hasan Şeker*
# TABLE OF CONTENTS

Chapter 8  
Development of Generic Capabilities in Teaching and Learning Environments at the Associate Degree Level  
*Wincy W.S. Lee, Doris Y. P. Leung and Kenneth C.H. Lo*  
169

## Part III – Structural Equation Modeling in Educational Practice

Chapter 9  
Latent Variable Modeling in Educational Psychology: Insights from a Motivation and Engagement Research Program  
*Gregory Arief D. Liem and Andrew J. Martin*  
187

Chapter 10  
Linking Teaching and Learning Environment Variables to Higher Order Thinking Skills: A Structural Equation Modeling Approach  
*John K. Rugutt*  
217

Chapter 11  
Influencing Group Decisions by Gaining Respect of Group Members in E-Learning and Blended Learning Environments: A Path Model Analysis  
*Binod Sundararajan, Lorn Sheehan, Malavika Sundararajan and Jill Manderson*  
241

Chapter 12  
Investigating Factorial Invariance of Teacher Climate Factors across School Organizational Levels  
*Christine DiStefano, Diana Mîndrilă and Diane M. Monrad*  
257

## Part IV – Conclusion

Chapter 13  
Structural Equation Modeling Approaches in Educational Research and Practice  
*Myint Swe Khine*  
279

Author Biographies  
285
PART I

THEORETICAL FOUNDATIONS
1. APPLYING STRUCTURAL EQUATION MODELING (SEM) IN EDUCATIONAL RESEARCH: AN INTRODUCTION

INTRODUCTION

The use of Structural Equation Modeling (SEM) in research has increased in psychology, sociology, education, and economics since it was first conceived by Wright (1918), a biometrician who was credited with the development of path analysis to analyze genetic theory in biology (Teo & Khine, 2009). In the 1970s, SEM enjoyed a renaissance, particularly in sociology and econometrics (Goldberger & Duncan, 1973). It later spread to other disciplines, such as psychology, political science, and education (Kenny, 1979). The growth and popularity of SEM was generally attributed to the advancement of software development (e.g., LISREL, AMOS, Mplus, Mx) that have increased the accessibility of SEM to substantive researchers who have found this method to be appropriate in addressing a variety of research questions (MacCallum & Austin, 2000). Some examples of these software include LISREL (LInear Structural RELations) by Joreskog and Sorbom (2003), EQS (Equations) (Bentler, 2003), AMOS (Analysis of Moment Structures) by Arbuckle (2006), and Mplus by Muthén and Muthén (1998-2010).

Over the years, the combination of methodological advances and improved interfaces in various SEM software have contributed to the diverse usage of SEM. Hershberger (2003) examined major journals in psychology from 1994 to 2001 and found that over 60% of these journals contained articles using SEM, more than doubled the number of articles published from 1985 to 1994. Although SEM continues to undergo refinement and extension, it is popular among applied researchers. The purpose of this chapter is to provide a non-mathematical introduction to the various facets of structural equation modeling to researchers in education.

What Is Structural Equation Modeling?

Structural Equation Modeling is a statistical approach to testing hypotheses about the relationships among observed and latent variables (Hoyle, 1995). Observed variables also called indicator variables or manifest variables. Latent variables also denoted unobserved variables or factors. Examples of latent variables in education are math ability and intelligence and in psychology are depression and self-
confidence. The latent variables cannot be measured directly. Researchers must
define the latent variable in terms of observed variables to represent it. SEM is also
a methodology that takes a confirmatory (i.e. hypothesis-testing) approach to the
analysis of a theory relating to some phenomenon. Byrne (2001) compared SEM
against other multivariate techniques and listed four unique features of SEM:

(1) SEM takes a confirmatory approach to data analysis by specifying the
relationships among variables \textit{a priori}. By comparison, other multivariate
techniques are descriptive by nature (e.g. exploratory factor analysis) so that
hypothesis testing is rather difficult to do.

(2) SEM provides explicit estimates of error variance parameters. Other
multivariate techniques are not capable of either assessing or correcting for
measurement error. For example, a regression analysis ignores the potential error
in all the independent (explanatory) variables included in a model and this raises
the possibility of incorrect conclusions due to misleading regression estimates.

(3) SEM procedures incorporate both unobserved (i.e. latent) and observed
variables. Other multivariate techniques are based on observed measurements only.

(4) SEM is capable of modeling multivariate relations, and estimating direct and
indirect effects of variables under study.

\textit{Types of Models in SEM}

Various types of structural equation models are used in research. Raykov and
Marcoulides (2006) listed four that are commonly found in the literature.

(1) Path analytic models (PA)
(2) Confirmatory factor analysis models (CFA)
(3) Structural regression models (SR)
(4) Latent change model (LC)

Path analytic (PA) models are conceived in terms of observed variables. Although they focus only on observed variables, they form an important part of the
historical development of SEM and employ the same underlying process of model
testing and fitting as other SEM models. Confirmatory factor analysis (CFA)
models are commonly used to examine patterns of interrelationships among
various constructs. Each construct in a model is measured by a set of observed
variables. A key feature of CFA models is that no specific directional relationships
are assumed between the constructs as they are correlated with each other only.
Structural regression (SR) models build on the CFA models by postulating specific
explanatory relationship (i.e. latent regressions) among constructs. SR models are
often used to test or disconfirm proposed theories involving explanatory
relationships among various latent variables. Latent change (LC) models are used
to study change over time. For example, LC models are used to focus on patterns of growth, decline, or both in longitudinal data and enable researchers to examine both intra- and inter-individual differences in patterns of change. Figure 1 shows an example of each type of model. In the path diagram, the observed variables are represented as rectangles (or squares) and latent variables are represented as circles (or ellipses).

![Path Diagram of SEM Models]

Figure 1. Types of SEM models.

**Example Data**

Generally, SEM undergoes five steps of model specification, identification, estimation, evaluation, and modifications (possibly). These five steps will be illustrated in the following sections with data obtained as part of a study to examine the attitude towards computer use by pre-service teachers (Teo, 2008, 2010). In this example, we provide a step-by-step overview and non-mathematical using with AMOS of the SEM when the latent and observed variables are
continuous. The sample size is 239 and, using the Technology Acceptance Model (Davis, 1989) as the framework data were collected from participants who completed an instrument measuring three constructs: perceived usefulness (PU), perceived ease of use (PEU), and attitude towards computer use (ATCU).

**Measurement and Structural Models**

Structural equation models comprise both a measurement model and a structural model. The measurement model relates observed responses or ‘indicators’ to latent variables and sometimes to observed covariates (i.e., the CFA model). The structural model then specifies relations among latent variables and regressions of latent variables on observed variables. The relationship between the measurement and structural models is further defined by the two-step approach to SEM proposed by James, Mulaik and Brett (1982). The two-step approach emphasizes the analysis of the measurement and structural models as two conceptually distinct models. This approach expanded the idea of assessing the fit of the structural equation model among latent variables (structural model) independently of assessing the fit of the observed variables to the latent variables (measurement model). The rationale for the two-step approach is given by Jöreskog and Sörbom (2003) who argued that testing the initially specified theory (structural model) may not be meaningful unless the measurement model holds. This is because if the chosen indicators for a construct do not measure that construct, the specified theory should be modified before the structural relationships are tested. As such, researchers often test the measurement model before the structural model.

A measurement model is a part of a SEM model which specifies the relations between observed variables and latent variables. Confirmatory factor analysis is often used to test the measurement model. In the measurement model, the researcher must operationally decide on the observed indicators to define the latent factors. The extent to which a latent variable is accurately defined depends on how strongly related the observed indicators are. It is apparent that if one indicator is weakly related to other indicators, this will result in a poor definition of the latent variable. In SEM terms, model misspecification in the hypothesized relationships among variables has occurred.

Figure 2 shows a measurement model. In this model, the three latent factors (circles) are each estimated by three observed variables (rectangles). The straight line with an arrow at the end represents a hypothesized effect one variable has on another. The ovals on the left of each rectangle represent the measurement errors (residuals) and these are estimated in SEM.

A practical consideration to note includes avoiding testing models with constructs that contains a single indicator (Bollen, 1989). This is to ensure that the observed indicators are reliable and contain little error so that the latent variables can be better represented. The internal consistency reliability estimates for this example ranged from .84 to .87.
Structural models differ from measurement models in that the emphasis moves from the relationship between latent constructs and their measured variables to the nature and magnitude of the relationship between constructs (Hair et al., 2006). In other words, it defines relations among the latent variables. In Figure 3, it was hypothesized that a user’s attitude towards computer use (ATCU) is a function of perceived usefulness (PU) and perceived ease of use (PEU). Perceived usefulness (PU) is, in turn, influenced by the user’s perceived ease of use (PEU). Put differently, perceived usefulness mediates the effects of perceived ease of use on attitude towards computer use.

**Effects in SEM**

In SEM two types of effects are estimates: direct and indirect effects. Direct effects, indicated by a straight arrow, represent the relationship between one latent variable to another and this is indicated using single-directional arrows (e.g. between PU and ATCU in Figure 2). The arrows are used in SEM to indicate directionality and do not imply causality. Indirect effects, on the other hand, reflect the relationship between an independent latent variable (exogenous variable) (e.g. PEU) and a dependent latent variable (endogenous variable) (e.g. ATCU) that is mediated by one or more latent variable (e.g. PU).
STAGES IN SEM

From the SEM literature, there appears an agreement among practitioners and theorists that five steps are involved in testing SEM models. These five steps are model specification, identification, estimation, evaluation, and modification (e.g., Hair et al., 2006; Kline, 2005; Schumacker & Lomax, 2004).

Model Specification

At this stage, the model is formally stated. A researcher specifies the hypothesized relationships among the observed and latent variables that exist or do not exist in the model. Actually, it is the process by the analyst declares which relationships are null, which are fixed to a constant, and which are vary. Any relationships among variables that are unspecified are assumed to be zero. In Figure 3, the effect of PEU on ATCU is mediated by PU. If this relationship is not supported, then misspecification may occur.

Relationships among variables are represented by parameters or paths. These relationships can be set to fixed, free or constrained. Fixed parameters are not estimated from the data and are typically fixed at zero (indicating no relationship...
between variables) or one. In this case where a parameter is fixed at zero, no path (straight arrows) is drawn in a SEM diagram. *Free parameters* are estimated from the observed data and are assumed by the researcher to be non-zero (these are shown in Figure 3 by asterisks). *Constrained parameters* are those whose value is specified to be equal to a certain value (e.g. 1.0) or equal to another parameter in the model that needs to be estimated. It is important to decide which parameters are fixed and which are free in a SEM 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. The choice of which parameters are free and which are fixed in a model should be guided by the literature.

There are three types of parameters to be specified: directional effects, variances, and covariances. Directional effects represent the relationships between the observed indicators (called factor loadings) and latent variables, and relationships between latent variables and other latent variables (called path coefficients). In Figure 3, the directional arrows from the latent variable, PU to PU2 and PU3 are examples of factor loading to be estimated while the factor loading of PU1 has been set at 1.0. The arrow from PU to ATCU is an example of path coefficient showing the relationship between one latent variable (exogenous variable) to another (endogenous variable). The directional effects in Figure 3 are six factor loadings between latent variables and observed indicators and three path coefficients between latent variables, making a total of nine parameters.

Variances are estimated for independent latent variables whose path loading has been set to 1.0. In Figure 3, variances are estimated for indicator error (er1~er9) associated with the nine observed variables, error associated with the two endogenous variables (PU and ATCU), and the single exogenous variable (PEU). Covariances are nondirectional associations among independent latent variables (curved double-headed arrows) and these exist when a researcher hypothesizes that two factors are correlated. Based on the theoretical background of the model in Figure 3, no covariances were included. In all, 21 parameters (3 path coefficients, 6 factor loadings, and 12 variances) in Figure 3 were specified for estimation.

**Model Identification**

At this stage, the concern is whether a unique value for each free parameter can be obtained from the observed data. This is dependent on the choice of the model and the specification of fixed, constrained and free parameters. Schumacker and Lomax (2004) indicated that three identification types are possible. If all the parameters are determined with just enough information, then the model is ‘just-identified’. If there is more than enough information, with more than one way of estimating a parameter, then the model is ‘overidentified’. If one or more parameters may not be determined due to lack of information, the model is ‘under-identified’. This situation causes the positive degree of freedom. Models need to be overidentified in order to be estimated and in order to test hypotheses about the relationships among variables. A researcher has to ensure that the elements in the
correlation matrix (i.e. the off-diagonal values) that is derived from the observed variables are more than the number of parameters to be estimated. If the difference between the number of elements in the correlation matrix and the number of parameters to be estimated is a positive figure (called the degree of freedom), the model is over-identified. The following formula is used to compute the number of elements in a correlation matrix:

$$[p \ (p + 1)]/2$$

where $p$ represents the number of observed(measured) variables. Applying this formula to the model in Figure 3 with nine observed variables, $[9(9+1)]/2 = 45$. With 21 parameters specified for estimation, the degree of freedom is 45-21 = 24, rendering the model in Figure 3 over-identified. When the degree of freedom is zero, the model is just-identified. On the other hand, if there are negative degrees of freedom, the model is under-identified and parameter estimation is not possible.

Of the goals in using SEM, an important one is to find the most parsimonious model to represent the interrelationships among variables that accurately reflects the associations observed in the data. Therefore, a large degree of freedom implies a more parsimonious model. Usually, model specification and identification precede data collection. Before proceeding to model estimation, the researcher has to deal with issues relating to sample size and data screening.

Sample size. This is an important issue in SEM but no consensus has been reached among researchers at present, although some suggestions are found in the literature (e.g., Kline, 2005; Ding, Velicer, & Harlow, 1995; Raykov & Widaman, 1995). Raykov and Widaman (1995) listed four requirements in deciding on the sample size: model misspecification, model size, departure from normality, and estimation procedure. Model misspecification refers to the extent to which the hypothesized model suffers from specification error (e.g. omission of relevant variables in the model). Sample size impacts on the ability of the model to be estimated correctly and specification error to be identified. Hence, if there are concerns about specification error, the sample size should be increased over what would otherwise be required. In terms of model size, Raykov and Widaman (1995) recommended that the minimum sample size should be greater than the elements in the correlation matrix, with preferably ten participants per parameter estimated. Generally, as the model complexity increases, so does the larger sample size requirements. If the data exhibit nonnormal characteristics, the ratio of participants to parameters should be increased to 15 in to ensure that the sample size is large enough to minimize the impact of sampling error on the estimation procedure. Because Maximum Likelihood Estimation (MLE) is a common estimation procedure used in SEM software, Ding, Velicer, and Harlow (1995) recommends that the minimum sample size to use MLE appropriately is between 100 to 150 participants. As the sample size increases, the MLE method increases its sensitivity to detect differences among the data.
Kline (2005) suggested that 10 to 20 participants per estimated parameter would result in a sufficient sample. Based on this, a minimum of $10 \times 21 = 210$ participants is needed to test the model in Figure 3. The data set associated with Figure 3 contains 239 cases so this is well within the guidelines by Kline. Additionally, Hoelter’s critical $N$ is often used as the standard sample size that would make the obtained fit (measured by $\chi^2$) significant at the stated level of significance (Hoelter, 1983). Hoelter’s critical $N$ is a useful reference because it is found in most SEM software (e.g., AMOS).

**Multicollinearity.** This refers to situations where measured variables (indicators) are too highly related. This is a problem in SEM because researchers use related measures as indicators of a construct and, if these measures are too highly related, the results of certain statistical tests may be biased. The usual practice to check for multicollinearity is to compute the bivariate correlations for all measured variables. Any pair of variables with a correlations higher than $r = .85$ signifies potential problems (Kline, 2005). In such cases, one of the two variables should be excluded from further analysis.

**Multivariate normality.** The widely used methods in SEM assume that the multivariate distribution is normally distributed. Kline (2005) indicated that all the univariate distributions are normal and the joint distribution of any pair of the variables is bivariate normal. The violation of these assumptions may affect the accuracy of statistical tests in SEM. For example, testing a model with nonnormally distributed data may incorrectly suggest that the model is a good fit to the data or that the model is a poor fit to the data. However, this assumption is hardly met in practice. In applied research, multivariate normality is examined using Mardia’s normalized multivariate kurtosis value. This is done by comparing the Mardia’s coefficient for the data under study to a value computed based on the formula $p(p+2)$ where $p$ equals the number of observed variables in the model (Raykov & Marcoulides, 2008). If the Mardia’s coefficient is lower than the value obtained from the above formula, then the data is deemed as multivariate normal. As with the Hoelter’s critical $N$, the Mardia’s coefficient is found most SEM software (e.g., AMOS).

**Missing data.** The presence of missing is often due to factors beyond the researcher’s control. Depending on the extent and pattern, missing data must be addressed if the missing data occur in a non-random pattern and are more than ten percent of the overall data (Hair et al., 2006). Two categories of missing data are described by Kline (2005): missing at random (MAR) and missing completely at random (MCAR). These two categories are ignorable, which means that the pattern of missing data is not systematic. For example, if the absence of the data occurs in $X$ variable and this absence occur by chance and are unrelated to other variables; the data loss is considered to be at random.

A problematic category of missing data is known as not missing at random (NMAR), which implies a systematic loss of data. An example of NMAR is a
situation where participants did not provide data on the interest construct because they have few interests and chose to skip those items. Another NMAR case is where data is missing due to attrition in longitudinal research (e.g., attrition due to death in a health study). To deal with MAR and MCAR, users of SEM employ methods such as listwise deletion, pairwise deletion, and multiple imputations. As to which method is most suitable, researchers often note the extent of the missing data and the randomness of its missing. Various comprehensive reviews on missing data such as Allison (2003), Tsai and Yang (2012), and Vriens and Melton (2002) contain details on the categories of missing data and the methods for dealing with missing data should be consulted by researchers who wish to gain a fuller understanding in this area.

Model Estimation

In estimation, the goal is to produce a $\Sigma(\theta)$ (estimated model-implied covariance matrix) that resembles $S$ (estimated sample covariance matrix) of the observed indicators, with the residual matrix ($S - \Sigma(\theta)$) being as little as possible. When $S - \Sigma(\theta) = 0$, then $\chi^2$ becomes zero, and a perfect model is obtained for the data. Model estimation involves determining the value of the unknown parameters and the error associated with the estimated value. As in regression, both unstandardized and standardized parameter values and coefficients are estimated. The unstandardized coefficient is analogous to a Beta weight in regression and dividing the unstandardized coefficient by the standard error produces a $z$ value, analogous to the $t$ value associated with each Beta weight in regression. The standardized coefficient is analogous to $\beta$ in regression.

Many software programs are used for SEM estimation, including LISREL (Linear Structural Relationships; Jöreskog & Sörbom, 1996), AMOS (Analysis of Moment Structures; Arbuckle, 2003), SAS (SAS Institute, 2000), EQS (Equations; Bentler, 2003), and Mplus (Muthén & Muthén, 1998-2010). These software programs differ in their ability to compare multiple groups and estimate parameters for continuous, binary, ordinal, or categorical indicators and in the specific fit indices provided as output. In this chapter, AMOS 7.0 was used to estimate the parameters in Figure 3. In the estimation process, a fitting function or estimation procedure is used to obtain estimates of the parameters in $\theta$ to minimize the difference between $S$ and $\Sigma(\theta)$. Apart from the Maximum Likelihood Estimation (MLE), other estimation procedures are reported in the literature, including unweighted least squares (ULS), weighted least squares (WLS), generalized least squares (GLS), and asymptotic distribution free (ADF) methods.

In choosing the estimation method to use, one decides whether the data are normally distributed or not. For example, the ULS estimates have no distributional assumptions and are scale dependent. In other words, the scale of all the observed variables should be the same in order for the estimates to be consistent. On the other hand, the ML and GLS methods assume multivariate normality although they are not scale dependent.
When the normality assumption is violated, Yuan and Bentler (1998) recommend the use of an ADF method such as the WLS estimator that does not assume normality. However, the ADF estimator requires very large samples (i.e., \( n = 500 \) or more) to generate accurate estimates (Yuan & Bentler, 1998). In contrast, simple models estimated with MLE require a sample size as small as 200 for accurate estimates.

**Estimation example.** Figure 3 is estimated using the Maximum Likelihood estimator in AMOS 7.0 (Arbuckle, 2006). Figure 4 shows the standardized results for the structural portion of the full model. The structural portion also call structural regression models (Raykov & Marcoulides, 2000). AMOS provides the standardized and unstandardized output, which are similar to the standardized betas and unstandardized B weights in regression analysis. Typically, standardized estimates are shown but the unstandardized portions of the output are examined for significance. For example, Figure 4 shows the significant relationships (\( p < .001 \) level) among the three latent variables. The significance of the path coefficient from perceived ease of use (PEU) to perceived usefulness (PU) was determined by examining the unstandardized output, which is 0.540 and had a standard error of 0.069.

Although the critical ratio (i.e., z score) is automatically calculated and provided with the output in AMOS and other programs, it is easily determined whether the coefficient is significant (i.e., \( z \geq 1.96 \) for \( p \leq .05 \)) at a given alpha level by dividing the unstandardized coefficient by the standard error. This statistical test is an approximately normally distributed quantity (z-score) in large samples (Muthén & Muthén, 1998-2010). In this case, 0.540 divided by 0.069 is 7.826, which is greater than the critical z value (at \( p = .05 \)) of 1.96, indicating that the parameter is significant.

![Figure 4. Structural model with path coefficients](image-url)
Model Fit

The main goal of model fitting is to determine how well the data fit the model. Specifically, the researcher wishes to compare the predicted model covariance (from the specified model) with the sample covariance matrix (from the obtained data). On how to determine the statistical significance of a theoretical model, Schumacker and Lomax (2004) suggested three criteria. The first is a non-statistical significance of the chi-square test and. A non-statistically significant chi-square value indicates that sample covariance matrix and the model-implied covariance matrix are similar. Secondly, the statistical significance of each parameter estimates for the paths in the model. These are known as critical values and computed by dividing the unstandardized parameter estimates by their respective standard errors. If the critical values or $t$ values are more than 1.96, they are significant at the .05 level. Thirdly, one should consider the magnitude and direction of the parameter estimates to ensure that they are consistent with the substantive theory. For example, it would be illogical to have a negative parameter between the numbers of hours spent studying and test scores. Although addressing the second and third criteria is straightforward, there are disagreements over what constitutes acceptable values for global fit indices. For this reason, researchers are recommended to report various fit indices in their research (Hoyle, 1995, Martens, 2005). Overall, researchers agree that fit indices fall into three categories: absolute fit (or model fit), model comparison (or comparative fit), and parsimonious fit (Kelloway, 1998; Mueller & Hancock, 2004; Schumacker & Lomax, 2004).

Absolute fit indices measure how well the specified model reproduces the data. They provide an assessment of how well a researcher’s theory fits the sample data (Hair et al., 2006). The main absolute fit index is the $\chi^2$ (chi-square) which tests for the extent of misspecification. As such, a significant $\chi^2$ suggests that the model does not fit the sample data. In contrast, a non-significant $\chi^2$ is indicative of a model that fits the data well. In other word, we want the $p$-value attached to the $\chi^2$ to be non-significant in order to accept the null hypothesis that there is no significant difference between the model-implied and observed variances and covariances. However, the $\chi^2$ has been found to be too sensitive to sample size increases such that the probability level tends to be significant. The $\chi^2$ also tends to be greater when the number of observed variables increases. Consequently, a non-significant $p$-level is uncommon, although the model may be a close fit to the observed data. For this reason, the $\chi^2$ cannot be used as a sole indicator of model fit in SEM. Three other commonly used absolute fit indices are described below.

The Goodness-of-Fit index (GFI) assesses the relative amount of the observed variances and covariances explained by the model. It is analogous to the $R^2$ in regression analysis. For a good fit, the recommended value should be GFI > 0.95 (1 being a perfect fit). An adjusted goodness-of-fit index (AGFI) takes into account differing degree of model complexity and adjusts the GFI by a ratio of the degrees of freedom used in a model to the total degrees of freedom. The standardized root mean square residual (SRMR) is an indication of the extent of error resulting from the estimation of the specified model. On the other hand, the amount of error or
residual illustrates how accurate the model is hence lower SRMR values (<.05) represents a better model fit. The root mean square error of approximation (RMSEA) corrects the tendency of the \(\chi^2\) to reject models with large same size or number of variables. Like SRMR, a lower RMSEA (<.05) value indicates a good fit and it is often reported with a confidence level at 95% level to account for sampling errors associated with the estimated RMSEA.

In comparative fitting, the hypothesized model is assessed on whether it is better than a competing model and the latter is often a baseline model (also known as a null model), one that assumes that all observed variables is uncorrelated. A widely-used index example is the Comparative Fit Index (CFI) which indicates the relative lack of fit of a specified model versus the baseline model. It is normed and varies from 0 to 1, with higher values representing better fit. The CFI is widely used because of its strengths, including its relative insensitivity to model complexity. A value of > .95 for CFI is associated with a good model. Another comparative fit index is the Tucker-Lewis Index (TLI), also called the Bentler-Bonnet NNFI (nonnormed fit index) by Bentler and Bonnet (1980) is used to compare a proposed model to the null model. Since the TLI is not normed, its values can fall below 0 or above 1. Typically, models with a good fit have values that approach 1.0.

Parsimonious indices assess the discrepancy between the observed and implied covariance matrix while taking into account a model’s complexity. A simple model with fewer estimated parameters will always get a parsimony fit. This is because although adding additional parameters (thus increasing the complexity of a model) will always improve the fit of a model but it may not improve the fit enough to justify the added complexity. The parsimonious indices are computed using the parsimony ratio (PR), which is calculated as the ratio of degrees of freedom used by the model to the total degrees of freedom available (Marsh, Balla, & McDonald, 1988). An example of parsimony fit indices is the parsimony comparative-of-fit index (PCFI), which adjust the CFI using the PR. The PCFI values of a model range from 0 to 1 and is often used in conjunction with the PCFI of another model (e.g. null model). Because the AGFI and RMSEA adjust for model complexity, they may be also used as indicators of model parsimony.

Test of Model Fit Using Example Model

Most of the above fit indices are used to test the model in Figure 3 and their results shown in Table 1. These model fit indices represent the three fit indices categories absolute fit, comparative fit, and parsimonious fit. It can be seen the fit indices contradict each other. Although the GFI, SRMR, CFI, and the TLI, the significant \(\chi^2\), high RMSEA and AGFI suggest that the model may be a poor fit to the data. The fit indices suggests that some misspecification may exist that suggests that the model may not fit well.
Table 1. Model fit for Figure 3

<table>
<thead>
<tr>
<th>Fit Index</th>
<th>Model in Figure 3</th>
<th>Recommended level</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>61.135, significant</td>
<td>Non-significant</td>
<td>Hair et al. (2006)</td>
</tr>
<tr>
<td>GFI</td>
<td>.94</td>
<td>&lt; .95</td>
<td>Schumacker &amp; Lomax (2004)</td>
</tr>
<tr>
<td>AGFI</td>
<td>.89</td>
<td>&lt; .95</td>
<td>Schumacker &amp; Lomax (2004)</td>
</tr>
<tr>
<td>SRMR</td>
<td>.04</td>
<td>&lt; .08</td>
<td>Hu &amp; Bentler (1998)</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.08</td>
<td>&lt; .07</td>
<td>Hair et al. (2006)</td>
</tr>
<tr>
<td>CFI</td>
<td>.97</td>
<td>&gt; .95</td>
<td>Schumacker &amp; Lomax (2004)</td>
</tr>
<tr>
<td>TLI</td>
<td>.95</td>
<td>&gt; .95</td>
<td>Schumacker &amp; Lomax (2004)</td>
</tr>
</tbody>
</table>

Note: GFI= Goodness-of-Fit; AGFI=Adjusted Goodness-of-Fit; SRMR=Standardized Root Mean Residual; RMSEA= Root Mean Square Error of Approximation; CFI=Comparative Fit Index; TLI=Tucker-Lewis Index

Parameter estimates. Having considered the structural model, it is important to consider the significance of estimated parameters. As with regression, a model that fits the data well but has few significant parameters is not desirable. From the standardized estimates in Figure 4 (the path coefficients for the observed indicators are not shown here because they would have been examined for significance during the confirmatory factor analysis in the measurement model testing stage), it appears that there is a stronger relationship between perceived ease of use (PEU) and perceived usefulness (PU) ($\beta = .60$) than between perceived ease of use (PEU) and attitude towards computer use (ATCU) ($\beta = .43$). However, the relationship between PEU and ATCU is also mediated by PU, so two paths from PEU and ATCU can be traced in the model (PEU $\rightarrow$ PU $\rightarrow$ ATCU). Altogether, PU and PEU explain 60.8% of the variance in ATCU. This is also known as squared multiple correlations and provided in the AMOS output.

Model Modification

If the fit of the model is not good, hypotheses can be adjusted and the model retested. This step is often called re-specification (Schumacker & Lomax, 2004). In modifying the model, a researcher either adds or removes parameters to improve the fit. Additionally, parameters could be changed from fixed to free or from free to fixed. However, these must be done carefully since adjusting a model after initial testing increases the chance of making a Type I error. At all times, any changes made should be supported by theory. To assist researchers in the process of model modification, most SEM software such as AMOS compute the modification indices (MI) for each parameter. Also called the Lagrange Multiplier (LM) Index or the Wald Test, these MI report the change in the $\chi^2$ value when parameters are adjusted. The LM indicates the extent to which addition of free
parameters increases model fitness while the Wald Test asks whether deletion of free parameters increases model fitness. The LM and Wald Test follow the logic of forward and backward stepwise regression respectively.

The steps to modify the model include the following:

- Examine the estimates for the regression coefficients and the specified covariances. The ratio of the coefficient to the standard error is equivalent to a z test for the significance of the relationship, with a $p < .05$ cutoff of about 1.96. In examining the regression weights and covariances in the model you originally specified, it is likely that one will find several regression weights or covariances that are not statistically significant.

- Adjust the covariances or path coefficients to make the model fit better. This is the usual first step in model fit improvement.

- Re-run the model to see if the fit is adequate. Having made the adjustment, it should be noted that the new model is a subset of the previous one. In SEM terminology, the new model is a nested model. In this case, the difference in the $\chi^2$ is a test for whether some important information has been lost, with the degrees of freedom of this $\chi^2$ equal to the number of the adjusted paths. For example, if the original model had a $\chi^2$ of 187.3, and you remove two paths that were not significant. If the new $\chi^2$ has a value of 185.2, with 2 degrees of freedom (not statistically significant difference), then important information has not been lost with this adjustment.

- Refer to the modification indices (MI) provided by most SEM programs if the model fit is still not adequate after steps 1 to 3. The value of a given modification index is the amount that the $\chi^2$ value is expected to decrease if the corresponding parameter is freed. At each step, a parameter is freed that produces the largest improvement in fit and this process continues until an adequate fit is achieved (see Figure 5). Because the SEM software will suggest all changes that will improve model fit, some of these changes may be nonsensical. The researcher must always be guided by theory and avoid making adjustments, no matter how well they may improve model fit. Figure 5 shows an example of a set of modification indices from AMOS 7.0.

Martens (2005) noted that model modifications generally result in a better-fitting model. Hence researchers are cautioned that extensive modifications may result in data-driven models that may not be generalizable across samples (e.g., Chou & Bentler, 1990; Green, Thompson, & Babyak, 1998). This problem is likely to occur when researchers (a) use small samples, (b) do not limit modifications to those that are theoretically acceptable, and (c) severely misspecify the initial model (Green et al., 1998). Great care must be taken to ensure that models are modified within the limitations of the relevant theory. Using Figure 3 as an example, if a Wald test indicated that the researcher should remove the freely estimated parameter from perceived ease of use (PEU) to perceived usefulness (PU), the researcher should not apply that modification, because the suggested relationship between PEU and PU has been empirically tested and well documented. Ideally, model modifications suggested by the Wald or Lagrange Multiplier tests should be tested on a separate sample (i.e. cross-validation). However, given the large
samples required and the cost of collecting data for cross-validation, it is common to split an original sample into two halves, one for the original model and the other for validation purposes. If the use of another sample is not possible, extreme caution should be exercised when modifying and interpreting modified models.

<table>
<thead>
<tr>
<th>Covariance: (Group number 1 – Default model)</th>
<th>M.I.</th>
<th>Par Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>er7 &lt;--&gt; er10</td>
<td>17.060</td>
<td>.064</td>
</tr>
<tr>
<td>er9 &lt;--&gt; er10</td>
<td>4.198</td>
<td>-.033</td>
</tr>
<tr>
<td>er6 &lt;--&gt; er9</td>
<td>4.784</td>
<td>-.038</td>
</tr>
<tr>
<td>er5 &lt;--&gt; er11</td>
<td>5.932</td>
<td>-.032</td>
</tr>
<tr>
<td>er5 &lt;--&gt; er7</td>
<td>5.081</td>
<td>.032</td>
</tr>
<tr>
<td>er4 &lt;--&gt; er11</td>
<td>8.212</td>
<td>.039</td>
</tr>
<tr>
<td>er4 &lt;--&gt; er8</td>
<td>4.532</td>
<td>-.032</td>
</tr>
<tr>
<td>er3 &lt;--&gt; er7</td>
<td>4.154</td>
<td>-.042</td>
</tr>
<tr>
<td>er2 &lt;--&gt; er10</td>
<td>4.056</td>
<td>-.032</td>
</tr>
<tr>
<td>er2 &lt;--&gt; er9</td>
<td>8.821</td>
<td>.049</td>
</tr>
<tr>
<td>er1 &lt;--&gt; er10</td>
<td>5.361</td>
<td>.038</td>
</tr>
</tbody>
</table>

Figure 5. An example of modification indices from AMOS 7.0

CONCLUSION

This chapter attempts to describe what SEM is and illustrate the various steps of SEM by analysing an educational data set. It clearly shows that educational research can take advantage of SEM by considering more complex research questions and to test multivariate models in a single study. Despite the advancement of many new, easy-to-use software programs (e.g., AMOS, Lisrel, Mplus) that have increased the accessibility of this quantitative method, SEM is a complex family of statistical procedures that requires the researcher to make some decisions in order to avoid misuse and misinterpretation. Some of these decisions include answering how many participants to use, how to normalize data, what estimation methods and fit indices to use, and how to evaluate the meaning of those fit indices. The approach to answering these questions is presented sequentially in this chapter. However, using SEM is more than an attempt to apply any set of decision rules. To use SEM well involves the interplay of statistical procedures and theoretical understanding in the chosen discipline. Rather, those interested in using the techniques competently should constantly seek out information on the appropriate application of this technique. Over time, as consensus emerges, best practices are likely to change, thus affecting the way researchers make decisions.

This chapter contributes to the literature by presenting a non-technical, non-mathematical, and step-by-step introduction to SEM with a focus for educational researchers who possess little or no advanced Mathematical skills and knowledge. Because of the use of the variance-covariance matrix algebra in solving the simultaneous equations in SEM, many textbooks and ‘introductory’ SEM articles
contained formulas and equations that appear daunting to many educational researchers, many of whom consume SEM-based research reports and review journal articles as part of their professional lives. In addition, this chapter embedded an empirical study using a real educational data set to illustrate aspects of SEM at various junctures aimed to enhance the readers' understanding through practical applications of the technique. In view of the need for continuous learning, several suggestions and resources are listed in this chapter to aid readers in further reading and reference. In summary, while this author acknowledge that similar information may be obtained from textbooks and other sources, the strength of this chapter lies in its brevity and conciseness in introducing readers on the background, features, applications, and potentials of SEM in educational research.

APPENDIX

As with many statistical techniques, present and intending SEM users must engage in continuous learning. For this purpose, many printed and online materials are available. Tapping on the affordances of the internet, researchers have posted useful resources and materials for ready and free access to anyone interested in learning to use SEM. It is impossible to list all the resources that are available on the internet. The following are some websites that this author has found to be useful for reference and educational purposes.

Software (http://core.ecu.edu/psyc/wuenschk/StructuralSoftware.htm)
The site Information on various widely-used computer programs by SEM users. Demo and trails of some of these programs are available at the links to this site.

Books (http://www2.gsu.edu/~mkteer/bookfaq.html)
This is a list of introductory and advanced books on SEM and SEM-related topics.

General information on SEM (http://www.hawaii.edu/sem/sem.html)
This is one example of a person-specific website that contains useful information on SEM. There are hyperlinks in this page to other similar sites.

Journal articles (http://www.upa.pdx.edu/IOA/newsom/semrefs.htm)
A massive list of journal articles, book chapters, and whitepapers for anyone wishing to learn about SEM.

SEMNET (http://www2.gsu.edu/~mkteer/semnet.html)
This is an electronic mail network for researchers who study or apply structural equation modeling methods. SEMNET was founded in February 1993. As of November 1998, SEMNET had more than 1,500 subscribers around the world. The archives and FAQs sections of the SEMNET contain useful information for teaching and learning SEM.
REFERENCES


TEO, T., & KHINE, M. S. (2009). Modeling educational research: The way forward. In T. Teo & M. S. Khine (Eds.), Structural equation modeling in educational research: Concepts and applications (pp. 3-10). Rotterdam, the Netherlands: Sense Publishers.

TIMOTHY TEO
University of Auckland
New Zealand

LIANG TING TSAI
National Taichung University
Taiwan

CHIH-CHIEN YANG
National Taichung University
Taiwan
2. STRUCTURAL EQUATION MODELING IN EDUCATIONAL RESEARCH: A PRIMER

INTRODUCTION

Structural equation modeling (SEM) is a collection of statistical methods for modeling the multivariate relationship between variables. It is also called covariance structure analysis or simultaneous equation modeling and is often considered an integration of regression and factor analysis. As SEM is a flexible and powerful technique for examining various hypothesized relationships, it has been used in numerous fields, including marketing (e.g., Jarvis, MacKenzie, & Podsakoff, 2003; Williams, Edwards, & Vandenberg, 2003), psychology (e.g., Cudeck & du Toit, 2009; Martens, 2005), and education (e.g., Kieffer, 2011; Teo & Khine, 2009; Wang & Holcombe, 2010). For example, educational research has benefited from the use of SEM to examine (a) the factor structure of the learner traits assessed by tests or questionnaires (e.g., Silverman, 2010; Schoonen et al., 2003), (b) the equivalency of models across populations (e.g., Byrne, Baron, & Balev, 1998; In’nami & Koizumi, 2012; Shin, 2005), and (c) the effects of learner variables on proficiency or academic achievement at a single point in time (e.g., Ockey, 2011; Wang & Holcombe, 2010) or across time (e.g., Kieffer, 2011; Marsh & Yeung, 1998; Tong, Lara-Alecio, Irby, Mathes, & Kwok, 2008; Yeo, Fearrington, & Christ, 2011). This chapter provides the basics and the key concepts of SEM, with illustrative examples in educational research. We begin with the advantages of SEM, and follow this with a description of Bollen and Long’s (1993) five steps for SEM application. Then, we discuss some of the key issues with regard to SEM. This is followed by a demonstration of various SEM analyses and a description of software programs for conducting SEM. We conclude with a discussion on learning more about SEM. Readers who are unfamiliar with regression and factor analysis are referred to Cohen, Cohen, West, and Aiken (2003), Gorsuch (1983), and Tabachnick and Fidell (2007). SEM is an extension of these techniques, and having a solid understanding of them will aid comprehension of this chapter.

ADVANTAGES OF SEM

SEM is a complex, multivariate technique that is well suited for testing various hypothesized or proposed relationships between variables. Compared with a number of statistical methods used in educational research, SEM excels in four aspects (e.g., Bollen, 1989; Byrne, 2012b). First, SEM adopts a confirmatory,
hypothesis-testing approach to the data. This requires researchers to build a hypothesis based on previous studies. Although SEM can be used in a model-exploring, data-driven manner, which could often be the case with regression or factor analysis, it is largely a confirmatory method. Second, SEM enables an explicit modeling of measurement error in order to obtain unbiased estimates of the relationships between variables. This allows researchers to remove the measurement error from the correlation/regression estimates. This is conceptually the same as correcting for measurement error (or correcting for attenuation), where measurement error is taken into account for two variables by dividing the correlation by the square root of the product of the reliability estimates of the two instruments \( \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}} \). Third, SEM can include both unobserved (i.e., latent) and observed variables. This is in contrast with regression analysis, which can only model observed variables, and with factor analysis, which can only model unobserved variables. Fourth, SEM enables the modeling of complex multivariate relations or indirect effects that are not easily implemented elsewhere. Complex multivariate relations include a model where relationships among only a certain set of variables can be estimated. For example, in a model with variables 1 to 10, it could be that only variables 1 and 2 can be modeled for correlation. Indirect effects refer to the situation in which one variable affects another through a mediating variable.

**FIVE STEPS IN AN SEM APPLICATION**

The SEM application comprises five steps (Bollen & Long, 1993), although they vary slightly from researcher to researcher. They are (a) model specification, (b) model identification, (c) parameter estimation, (d) model fit, and (e) model respecification. We discuss these steps in order to provide an outline of SEM analysis; further discussion on key issues will be included in the next section.

**Model Specification**

First, model specification is concerned with formulating a model based on a theory and/or previous studies in the field. Relationships between variables – both latent and observed – need to be made explicit, so that it becomes clear which variables are related to each other, and whether they are independent or dependent variables. Such relationships can often be conceptualized and communicated well through diagrams.

For example, Figure 1 shows a hypothesized model of the relationship between a learner’s self-assessment, teacher assessment, and academic achievement in a second language. The figure was drawn using the SEM program Amos (Arbuckle, 1994-2012), and all the results reported in this chapter are analyzed using Amos, unless otherwise stated. Although the data analyzed below are hypothetical, let us suppose that the model was developed on the basis of previous studies. Rectangles represent observed variables (e.g., item/test scores, responses to questionnaire items), and ovals indicate unobserved variables. Unobserved variables are also
called factors, latent variables, constructs, or traits. The terms *factor* and *latent variable* are used when the focus is on the underlying mathematics (Royce, 1963), while the terms *construct* and *trait* are used when the concept is of substantive interest. Nevertheless, these four terms are often used interchangeably, and, as such, are used synonymously throughout this chapter. Circles indicate measurement errors or residuals. Measurement errors are hypothesized when a latent variable affects observed variables, or one latent variable affects another latent variable. Observed and latent variables that receive one-way arrows are usually modeled with a measurement error. A one-headed arrow indicates a hypothesized one-way direction, whereas a two-headed arrow indicates a correlation between two variables. The variables that release one-way arrows are independent variables (also called exogenous variables), and those that receive arrows are dependent variables (also called endogenous variables). In Figure 1, self-assessment is hypothesized to comprise three observed variables of questionnaire items measuring self-assessment in English, mathematics, and science. These observed variables are said to *load on* the latent variable of self-assessment. Teacher assessment is measured in a similar manner using the three questionnaire items, but this time presented to a teacher. The measurement of academic achievement includes written assignments in English, mathematics, and science. All observed variables are measured using a 9-point scale, and the data were collected from 450 participants. The nine observed variables and one latent variable contained measurement errors. Self-assessment and teacher assessment were modeled to affect academic achievement, as indicated by a one-way arrow. They were also modeled to be correlated with each other, as indicated by a two-way arrow.

Additionally, SEM models often comprise two subsets of models: a measurement model and a structured model. A measurement model relates observed variables to latent variables, or, defined more broadly, it specifies how the theory in question is operationalized as latent variables along with observed variables. A structured model relates constructs to one another and represents the theory specifying how these constructs are related to one another. In Figure 1, the three latent factors – self-assessment, teacher assessment, and academic achievement – are measurement models; the hypothesized relationship between them is a structural model. In other words, structural models can be considered to comprise several measurement models. Since we can appropriately interpret relationships among latent variables only when each latent variable is well measured by observed variables, an examination of the model fit (see below for details) is often conducted on a measurement model before one constructs a structural model.
Model Identification

The second step in an SEM application, namely model identification, is concerned with whether one can derive a unique value for each parameter (in the model) whose value is unknown (e.g., factor loadings, factor correlations, measurement errors) using the variance/covariance matrix (or the correlation matrix and standard deviations) of the measured variables that are known. Models are not identified when there are more parameters than can be estimated from the information available in the variance/covariance matrix. Models that are complex, even if theoretically sound, are likely to have identification problems, particularly when there are a large number of parameters to be estimated relative to the number of variances and covariances in the matrix. Two important principles are applicable to the identification of SEM models. First, latent variables must be assigned a scale (metric) because they are unobserved and do not have predetermined scales. This can be achieved by fixing either a factor variance, or one of the factor loadings, to be a specific value, usually 1. Second, the number of data points in the variance/covariance matrix – known information – must be at least equal to the number of parameters to be estimated in the model (i.e., free parameters) – unknown information. For example, for the academic achievement model, there are 21 estimated parameters: 8 factor loadings, 10 measurement error variances, 1 covariance, and 2 factor variances. Three of the factor loadings are each fixed to be 1 and do not have to be estimated. The number of data points is \( p(p + 1)/2 \), where \( p \) refers to the number of observed variables. For the academic achievement factor in Figure 1, there are nine observed variables, and therefore \( 9(9 + 1)/2 = 45 \) data points. This is larger than the number of parameters to be estimated in the model, which is 21. Thus, this model is identifiable. The degrees of freedom (df) are the difference between the number of data points and the number of parameters to be estimated. In the current example, the \( df \) are 24. When \( df \) are positive (one or
above), models can be identified. When \( df \) are negative, models cannot be identified, and are called unidentified. When \( df \) are zero, models can be identified but cannot be evaluated using fit indices (for fit indices, see below).

**Parameter Estimation**

Third, once the model has been identified, the next step is to estimate parameters in the model. The goal of parameter estimation is to estimate population parameters by minimizing the difference between the observed (sample) variance/covariance matrix and the model-implied (model-predicted) variance/covariance matrix. Several estimation methods are available, including maximum likelihood, robust maximum likelihood, generalized least squares, unweighted least squares, elliptical distribution theory, and asymptotically distribution-free methods. Although the choice of method depends on many factors, such as data normality, sample size, and the number of categories in an observed variable, the most widely used method is maximum likelihood. This is the default in many SEM programs because it is robust under a variety of conditions and is likely to produce parameter estimates that are unbiased, consistent, and efficient (e.g., Bollen, 1989).

Maximum likelihood estimation is an iterative technique, which means that an initially posited value is subsequently updated through calculation. The iteration continues until the best values are attained. When this occurs, the model is said to have converged. For the current example in Figure 1, the data were analyzed using maximum likelihood. The subsequent section entitled Data Normality provides more discussion on some recommendations for choice of estimation method.

**Model Fit**

Fourth, when parameters in a model are estimated, the degree to which the model fits the data must be examined. As noted in the preceding paragraph, the primary goal of SEM analysis is to estimate population parameters by minimizing the difference between the observed and the model-implied variance/covariance matrices. The smaller the difference is, the better the model. This is evaluated using various types of fit indices. A statistically nonsignificant chi-square (\( \chi^2 \)) value is used to indicate a good fit. Statistical nonsignificance is desirable because it indicates that the difference between the observed and the model-implied variance/covariance matrices is statistically nonsignificant, which implies that the two matrices cannot be said to be statistically different. Stated otherwise, a nonsignificant difference suggests that the proposed model cannot be rejected and can be considered correct. Note that this logic is opposite to testing statistical significance for analysis of variance, for example, where statistical significance is usually favorable.

Nevertheless, chi-square tests are limited in that, with large samples, they are likely to detect practically meaningless, trivial differences as statistically significant (e.g., Kline, 2011; Ullman, 2007). In order to overcome this
problem, many other fit indices have been created, and researchers seldom depend entirely on chi-square tests to determine whether to accept or reject the model. Fit indices are divided into four types based on Byrne (2006) and Kline (2011), although this classification varies slightly between researchers. First, incremental or comparative fit indices compare the improvement of the model to the null model. The null model assumes no covariances among the observed variables. Fit indices in this category include the comparative fit index (CFI), the normal fit index (NFI), and the Tucker-Lewis index (TLI), also known as the non-normed fit index (NNFI). Second, unlike incremental fit indices, absolute fit indices evaluate the fit of the proposed model without comparing it against the null model. Instead, they evaluate model fit by calculating the proportion of variance explained by the model in the sample variance/covariance matrix. Absolute fit indices include the goodness-of-fit index (GFI) and the adjusted GFI (AGFI). Third, residual fit indices concern the average difference between the observed and the model-implied variance/covariance matrices. Examples are the standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA). Fourth, predictive fit indices examine the likelihood of the model to fit in similarly sized samples from the same population. Examples include the Akaike information criterion (AIC), the consistent Akaike information criterion (CAIC), and the expected cross-validation index (ECVI).

The question of which fit indices should be reported has been discussed extensively in SEM literature. We recommend Kline (2011, pp. 209-210) and studies such as Hu and Bentler (1998, 1999) and Bandalos and Finney (2010), as they all summarize the literature remarkably well and clearly present how to evaluate model fit. Kline recommends reporting (a) the chi-square statistic with its degrees of freedom and $p$ value, (b) the matrix of correlation residuals, and (c) approximate fit indices (i.e., RMSEA, GFI, CFI) with the $p$ value for the close-fit hypothesis for RMSEA. The close-fit hypothesis for RMSEA tests the hypothesis that the obtained RMSEA value is equal to or less than .05. This hypothesis is similar to the use of the chi-square statistic as an indicator of model fit and failure to reject it is favorable and supports the proposed model. Additionally, Hu and Bentler (1998, 1999), Bandalos and Finney (2010), and numerous others recommend reporting SRMR, since it shows the average difference between the observed and the model-implied variance/covariance matrices. There are at least three reasons for this. First, this average difference is easy to understand by readers who are familiar with correlations but less familiar with fit indices. Hu and Bentler (1995) emphasize this, stating that the minimum difference between the observed and the model-implied variance/covariance matrices clearly signals that the proposed model accounts for the variances/covariances very well. Second, a reason for valuing the SRMR that is probably more fundamental is that it is a precise representation of the objective of SEM, which is to reproduce, as closely as possible, the model-implied variance/covariance matrix using the observed variance/covariance

...
matrix. Third, calculation of the SRMR does not require chi-squares. Since chi-
squares are dependent on sample size, this indicates that the SRMR, which
is not based on chi-squares, is not affected by sample size. This is in contrast with
other fit indices (e.g., CFI, GFI, RMSEA), which use chi-squares as part of the
calculation. For the assessment and academic achievement data, the chi-square
is 323.957 with 24 degrees of freedom at the probability level of .464 ($p > .05$).
The matrix of correlation residuals is presented in Table 1. If the model is
correct, the differences between sample covariances and implied covariances
should be small. Specifically, Kline argues that differences exceeding |0.10|
indicate that the model fails to explain the correlation between variables.
However, no such cases are found in the current data. Each residual correlation
con be divided by its standard error, as presented in Table 2. This is the same
as a statistical significance test for each correlation. The well-fitting model
should have values of less than [2]. All cases are statistically nonsignificant. The
RMSEA, GFI, and CFI are 0.000 (90% confidence interval: 0.000, 0.038), .989,
and 1.000, respectively. The $p$ value for the close-fit hypothesis for RMSEA is
.995, and the close-fit hypothesis is not rejected. The SRMR is .025. Taken
together, it may be reasonable to state that the proposed model of the relationship
between self-assessment, teacher assessment, and academic achievement is
supported.

The estimated model is presented in Figure 2. The parameter estimates
presented here are all standardized as this facilitates the interpretation of
parameters. Unstandardized parameter estimates also appear in an SEM output and
these should be reported as in Table 3 because they are used to judge statistical
significance of parameters along with standard errors. Factor loadings from the
factors to the observed variables are high overall ($\beta = .505$ to .815), thereby
suggesting that the three measurement models of self-assessment, teacher
assessment, and academic achievement were each measured well in the current
data. A squared factor loading shows the proportion of variance in the observed
variable that is explained by the factor. For example, the squared factor loading of
English for self-assessment indicates that self-assessment explains 53% of the
variance in English for self-assessment ($.731 \times .731$). The remaining 47% of the
variance is explained by the measurement error ($.682 \times .682$). In other words, the
variance in the observed variable is explained by the underlying factor and the
measurement error. Finally, the paths from the self-assessment and teacher
assessment factors to the academic achievement factor indicate that they
moderately affect academic achievement ($\beta = .454$ and .358). The correlation
between self-assessment and teacher assessment is rather small ($-.101$), thereby
indicating almost no relationship between them.
<table>
<thead>
<tr>
<th></th>
<th>Self-assessment</th>
<th></th>
<th>Teacher assessment</th>
<th></th>
<th>Academic achievement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Mathematics</td>
<td>Science</td>
<td>English</td>
<td>Mathematics</td>
<td>Science</td>
</tr>
<tr>
<td>Self-assessment</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Math</td>
<td>0.008</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Science</td>
<td>–0.026</td>
<td>0.005</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Teacher assessment</td>
<td>English</td>
<td>0.032</td>
<td>–0.003</td>
<td>–0.023</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Math</td>
<td>0.002</td>
<td>0.015</td>
<td>0.036</td>
<td>0.003</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Science</td>
<td>–0.013</td>
<td>–0.065</td>
<td>–0.014</td>
<td>–0.006</td>
<td>–0.001</td>
<td>–</td>
</tr>
<tr>
<td>Academic achievement</td>
<td>English</td>
<td>0.002</td>
<td>–0.064</td>
<td>0.046</td>
<td>0.048</td>
<td>0.008</td>
</tr>
<tr>
<td>Math</td>
<td>0.012</td>
<td>–0.016</td>
<td>0.047</td>
<td>–0.011</td>
<td>0.003</td>
<td>–0.032</td>
</tr>
<tr>
<td>Science</td>
<td>0.081</td>
<td>0.010</td>
<td>0.029</td>
<td>–0.083</td>
<td>–0.009</td>
<td>0.058</td>
</tr>
</tbody>
</table>
Table 2. Standardized correlation residuals

<table>
<thead>
<tr>
<th></th>
<th>Self-assessment</th>
<th>Teacher assessment</th>
<th>Academic achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Mathematics</td>
<td>Science</td>
</tr>
<tr>
<td>Self-assessment</td>
<td>English</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>0.102</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>–0.380</td>
<td>0.092</td>
</tr>
<tr>
<td>Teacher</td>
<td>English</td>
<td>0.389</td>
<td>–0.045</td>
</tr>
<tr>
<td>assessment</td>
<td>Mathematics</td>
<td>0.030</td>
<td>0.314</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>–0.219</td>
<td>–1.297</td>
</tr>
<tr>
<td>Academic</td>
<td>English</td>
<td>0.029</td>
<td>–0.959</td>
</tr>
<tr>
<td>achievement</td>
<td>Mathematics</td>
<td>0.211</td>
<td>–0.349</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>1.340</td>
<td>0.195</td>
</tr>
</tbody>
</table>

SEM IN EDUCATIONAL RESEARCH: A PRIMER
Model Respecification

Fifth, model re-specification is concerned with improving the model-data fit, for example, by deleting statistically nonsignificant paths or adding paths to the model. Any decision must be theoretically defensible and should not be statistically driven. The results are no longer confirmatory and should be viewed as explanatory. For the assessment and academic achievement data, we could, for example, delete the correlation between self-assessment and teacher assessment as it is very small in size (r = -.101) and statistically nonsignificant. This could be done only if it were supported by previous studies. Since this is not the case, no change is made in the model.

Table 3. Unstandardized and standardized estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Standard error</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-assessment ( \rightarrow ) English</td>
<td>1.000*</td>
<td>–</td>
<td>.731</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>.910*</td>
<td>.073</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>.703*</td>
<td>.060</td>
</tr>
<tr>
<td>Teacher assessment ( \rightarrow ) English</td>
<td>1.000*</td>
<td>–</td>
<td>.712</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>.736*</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>.528*</td>
<td>.066</td>
</tr>
<tr>
<td>Academic achievement ( \rightarrow ) English</td>
<td>1.000*</td>
<td>–</td>
<td>.784</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>.483*</td>
<td>.060</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>.534*</td>
<td>.065</td>
</tr>
<tr>
<td>Self-assessment ( \rightarrow ) Academic achievement</td>
<td>.498*</td>
<td>.072</td>
<td>.454</td>
</tr>
<tr>
<td>Teacher assessment ( \rightarrow ) Academic achievement</td>
<td>.380*</td>
<td>.073</td>
<td>.358</td>
</tr>
<tr>
<td>Self-assessment ( \leftarrow ) Teacher assessment</td>
<td>–0.092</td>
<td>.058</td>
<td>–.101</td>
</tr>
</tbody>
</table>

Note. *Fixed to 1.000 for scale identification. *p < .05. \( B \) refers to unstandardized estimates, \( \beta \) refers to standardized estimates.
SOME KEY ISSUES

Thus far, we have discussed an SEM analysis with minimal details. In practice, there are several other issues that must be considered in order to use SEM appropriately. We will discuss these issues surrounding data screening, model fit indices, and sample size because of their prevalence in SEM.

Data Screening

Before being put to appropriate use, SEM must undergo data screening. Such preliminary analysis may initially seem tedious; however, if it is done properly, it often saves time and leads to a more precise understanding of the results. Data screening is often discussed in terms of linearity, data normality, outliers, and missing data. Researchers examine these issues in slightly different ways. Readers are referred to Byrne (2006, 2010), Kline (2011), and Tabachnick and Fidell (2007) for further details.

Linearity. SEM models are estimated by examining the relationship—usually a linear one—among measured variables that are represented in the variance/covariance matrix (or the correlation matrix and standard deviations). Such a linear relationship between variables is called linearity: One variable increases/decreases in proportion to a change in another variable. Figure 3A shows an example of this relationship. As with regression and factor analysis, excessive linearity is problematic. This can be examined through inspection of scatterplots or correlation matrices. For example, high correlations among variables (e.g., +/- .90; Tabachnick & Fidell, 2007) are troublesome. Table 4 shows that the correlations between the observed variables range from -.103 to .601. They are not high enough to cause a problem. Statistical tests for multicollinearity are also available, which include squared multiple correlations, tolerance, and the variance inflation factor. These tests are also used in statistical analysis in general and are not limited to SEM. High linearity can be adjusted for by deleting or aggregating redundant variables.

Nonlinear relationships can also be examined in quadratic or cubic models. A quadratic relationship is one in which one variable affects another up to some point, after which the effect levels off or decreases. Figure 3B shows a data distribution that looks like an inverse U-shape, where as one variable increases (1, 2, 3, 4, 5, 6, 7, 8) the other increases and then decreases (2, 3, 4, 5, 4, 3, 2, 1). A cubic relationship is similar to a quadratic relationship—one variable affects another up to some point, the effect levels off or decreases beyond that point, but this time comes back to influence once again after a certain point. Figure 3C shows a cubic relationship. Quadratic and cubic relationships are also called curvilinear relationships. Figure 3D shows an interactive relationship, in which scores in one group increase while those in the other group decrease. It is possible that a moderator variable is at play. It should be noted that there are a variety of nonlinear relationships in addition to those presented in Figures 3B, 3C, and 3D.
(e.g., U-shaped relationship for a quadratic one). As a standard SEM assumes linear relations, modeling a nonlinear effect requires advanced techniques (see Kline, 2005, 2011; Marsh, Wen, Nagengast, & Hau, 2012).

Data normality. Data normality is divided into univariate normality and multivariate normality. Univariate normality refers to the situation in which one variable is normally distributed. Multivariate normality refers to the situation in which, in addition to the normality of each variable, each variable is also normally distributed for each other variable (Tabachnick & Fidell, 2007). Numerous SEM application studies use the maximum likelihood estimation method. This method assumes multivariate normal distribution of the data for the dependent (i.e., endogenous) variable. Although maximum likelihood methods are robust against non-normality, it is still important to assess whether the data satisfy the assumption of normality. Since multivariate normality is related to univariate normality, both types of normality need to be examined.

Univariate normality can be examined by inspecting absolute skewness and kurtosis values or the statistical significance of those values. First, with regard to the inspection of skewness and kurtosis, data normality is ensured when both values are zero. Unfortunately, there are no clear-cut guidelines on the degree of non-normality. Kline (2011) reviewed relevant studies (e.g., Curran, West, & Finch, 1996) and suggested viewing skewness and kurtosis exceeding 3 and 20 respectively as extremely non-normal. Note that this is a rule-of-thumb and is not an agreed-upon definition. For example, Curran et al. (1996) consider a skewness of 2 and a kurtosis of 7 as moderately non-normal, and a skewness of 3 and a kurtosis of 21 as severely non-normal. Chou and Bentler (1995) and Muthén and Kaplan (1985) argue that skewness and kurtosis values approaching 2 and 7, respectively, indicate problems. Table 4 shows that skewness and kurtosis values for all the observed variables are well below these cut-offs and are in fact very near to zero, thereby indicating that the data are univariately normal.
<table>
<thead>
<tr>
<th></th>
<th>Self-assessment</th>
<th>Teacher assessment</th>
<th>Academic achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Mathematics</td>
<td>Science</td>
</tr>
<tr>
<td>Self-assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>.601</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.531</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Science</td>
<td>.452</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Teacher assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>–.034</td>
<td>–.061</td>
<td>–.063</td>
</tr>
<tr>
<td>Mathematics</td>
<td>–.052</td>
<td>–.044</td>
<td>–.011</td>
</tr>
<tr>
<td>Science</td>
<td>–.048</td>
<td>–.103</td>
<td>–.046</td>
</tr>
<tr>
<td>Academic achievement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>.241</td>
<td>.220</td>
<td>.246</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.172</td>
<td>.164</td>
<td>.193</td>
</tr>
<tr>
<td>Science</td>
<td>.235</td>
<td>.200</td>
<td>.180</td>
</tr>
<tr>
<td>Mean</td>
<td>4.089</td>
<td>4.102</td>
<td>4.088</td>
</tr>
<tr>
<td>SD</td>
<td>1.284</td>
<td>1.048</td>
<td>1.021</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.412</td>
<td>1.099</td>
<td>1.281</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.095</td>
<td>0.035</td>
<td>0.016</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.819</td>
<td>0.299</td>
<td>0.137</td>
</tr>
<tr>
<td>z</td>
<td>0.520</td>
<td>–0.031</td>
<td>–1.453</td>
</tr>
</tbody>
</table>

Table 4. Correlations between variables and their descriptive statistics
Second, the statistical significance of skewness and kurtosis also serves as an indicator of data normality. In particular, the critical ratio or $z$ value is computed by dividing either skewness or kurtosis by its standard error. Data normality is ensured when the absolute value is within $+/- 2.58 \ (p < .01)$ or $3.29 \ (p < .001)$. However, as emphasized by Kline (2011) and Tabachnick and Fidell (2007), the standard errors of skewness and kurtosis shrink in large sample sizes, which can produce statistically significant skewness and kurtosis values even when the data distribution appears normal. Thus, with large samples, making substantive decisions on the basis of the visual inspection of the data – for example, using histograms or box plots – is preferred. However, it is difficult to define what is meant by a large sample. For example, Byrne (2006, 2010) only uses absolute skewness and kurtosis values for her dataset with a sample size of 372. Ullman (2007) uses both absolute values and statistical significance of skewness and kurtosis for her two datasets with sample sizes 175 and 459. In actuality, it appears that researchers are more likely to use estimation methods that are robust against non-normality, such as Satorra-Bentler correction or weighted least square parameter estimate methods. In any case, Table 4 shows that $z$ values for skewness and kurtosis are all within $+/-2.58 \ (p < .01)$ or $3.29 \ (p < .001)$, thereby suggesting data normality.

Additionally, multivariate normality can be measured using Mardia’s coefficient of multivariate kurtosis. The statistical significance of Mardia’s coefficient is examined using a $z$ value, but this time using the $z$ values of 5 or 6, not $+/-2.58 \ (p < .01)$ or $3.29 \ (p < .001)$, since Bentler (2005) argues that multivariate non-normality would not affect the model in practice unless its values were 5, 6, or above. Univariate normality can be estimated using general-purpose software programs (e.g., SAS or SPSS) or SEM programs, whereas multivariate normality can only be estimated using SEM programs (for SEM programs, see the Software section). Mardia’s coefficient for the current data is $-0.157$ with a $z$ value of $-0.119$. This indicates the multivariate normality of the data.

As seen above, numerous issues surrounding the treatment of non-normal data complicate decision making during data analysis. We reviewed previous studies and found Finney and DiStefano (2006) the most accessible, synthetic, and up to date. They summarize relevant studies and recommend that, for continuous data, if the variables are approximately normally distributed, the maximum likelihood estimation is recommended; if the variables are moderately non-normal (skewness $< 2$ and kurtosis $< 7$) the maximum likelihood estimation or Satorra-Bentler correction method are recommended; if the variables are severely non-normal (skewness $> 2$ and kurtosis $> 7$), the Satorra-Bentler correction or bootstrapping methods is recommended. For categorical data, regardless of the number of categories, they recommend using weighted least square parameter estimates (WLSMV), available in the SEM program Mplus. If Mplus is not available, they recommend that if the variables are approximately normally distributed, the maximum likelihood estimation should be used for scales with five or more categories and the Satorra-Bentler correction method for scales with four or more categories. This also applies to moderately non-normal data (skewness $< 2$ and
kurtosis < 7). If the variables are severely non-normal (skewness > 2 and kurtosis > 7), the Satorra-Bentler correction method is recommended.

**Outliers.** An outlier is an extremely large or small value of one variable (a univariate outlier) or a combination of such values of two or more variables (a multivariate outlier). Univariate outliers can be detected by drawing a histogram or inspecting the z values of variables using, for example, the SPSS EXPLORE or DESCRIPTIVES functions. Multivariate outliers can be detected using the Mahalanobis distance (i.e., Mahalanobis d-squared) statistic. It shows how one observation in the data is distantly located from the others. It is distributed as a chi-square statistic with degrees of freedom equal to the number of observed variables. Observations are arranged according to the size of the statistics, and those exceeding the critical value of the chi-square given degrees of freedom (e.g., $p < .001$) can be judged as outliers. For the current data, the histograms appear normal. There are five responses out of 4050 (450 × 9 items) exceeding the z value 3.29 ($p < .001$). As this is just 0.001% of the total responses, it is considered negligible. With regard to multivariate outliers, the critical value of chi-square for 24 degrees of freedom is 51.179. The most deviated case was participant 4, whose responses produced a Mahalanobis distance of 27.192—still below 51.179. Taken together, it is reasonable to say that the current dataset does not include univariate or multivariate outliers.

**Missing data.** The ideal situation is to be able to analyze a complete dataset that contains all examinees' responses to all items. In reality, this rarely occurs and one often has to analyze a dataset with missing values. Therefore, how to treat missing data is a widely discussed issue in the application of statistics, including SEM. Missing data treatment is classified into three types: (a) the deletion of those data, (b) the estimation of those data, and (c) the use of parameter estimation methods that take missingness into consideration. Deletion of missing data is a traditional approach, and includes listwise deletion (elimination of all cases with missing values from subsequent analysis) or pairwise deletion (removal of paired cases in which at least one case has a missing value). Although both methods are easy to implement, they may result in substantial loss of data observations. More importantly, Muthén, Kaplan, and Hollis (1987) argue that the two methods work only when data are missing completely at random, a case that is often violated in practice. Thus, both listwise and pairwise deletion methods may bias results if data missingness is not randomly distributed through the data (Tabachnick & Fidell, 2007).

A preferred approach is to estimate and impute missing data. Methods abound, such as mean substitution, regression, and expectation maximization methods; however, according to Tabachnick and Fidell (2007), the most recommended method is multiple imputation (Rubin, 1987). It replaces missing values with plausible values that take into account random variation.

Another way to address missing data is to use parameter estimation methods that take missingness into consideration. This is implemented in (full information)
maximum likelihood estimation, which uses all available data regardless of completeness (e.g., Enders, 2001). Both expectation maximization and maximum likelihood estimation methods are available in SEM programs. As the current data do not include missing responses, it is not necessary to eliminate, estimate, or impute such responses.

**Model Fit Indices**

Although no agreed-upon guidelines exist regarding which fit indices should be reported, some recommendations can be found in the literature. In an often-cited article, Hu and Bentler (1998, 1999) recommend reporting (a) the SRMR, and (b) the CFI, TLI, RMSEA, or other indices (e.g., Gamma Hat, Incremental Fit Index [IFI], McDonald’s Centrality Index [MCI], Relative Noncentrality Index [RNI]). Similarly, Kashy, Donnellan, Ackerman, and Russell (2009) recommend reporting the CFI or TLI along with the chi-square and RMSEA. Bandalos and Finney (2010) recommend the chi-square, CFI, TLI, RMSEA, and SRMR, whereas Mueller and Hancock (2010) recommend RMSEA and its confidence interval, the SRMR, and at least one of CFI, NFI, and TLI. Widaman (2010) encourages reporting the chi-square, CFI, TLI, and RMSEA. For testing measurement invariance across groups (e.g., whether the factor loadings are the same across groups), Cheung and Rensvold (2002) recommend reporting the CFI, Gamma Hat, and McDonald’s Noncentrality Index and interpreting a reduction in each index as evidence of measurement invariance. Summarizing studies that provide guidelines for reporting fit indices, In’nami and Koizumi (2011) report that the indices recommended most often are the chi-square (with degrees of freedom and *p* values), CFI, TLI, RMSEA (and its confidence interval), and the SRMR.

**Sample Size**

One rule-of-thumb is that a sample size below 100, between 100 and 200, and over 200 is often considered small, medium, and large, respectively (Kline, 2005). Similarly, Ding, Velicer, and Harlow (1995) argue that the minimum sample size adequate for analysis is generally 100 to 150 participants. Another approach is to consider model complexity in terms of the ratio of the sample size to the number of free parameters to be estimated in a model. A minimum sample size is at least 10 times the number of free model parameters (Raykov & Marcoulides, 2006). For example, a model with 30 free parameters would require at least 300 observations (30 × 10). Nevertheless, as the authors of the aforementioned articles emphasize, these are only rough guidelines. This is particularly because the requisite sample size depends on numerous factors, including the number and patterns of missing data, strength of the relationships among the indicators, types of indicators (e.g., categorical or continuous), estimation methods (e.g., [robust] maximum likelihood, robust weighted least squares), and reliability of the indicators. Complex issues surrounding sample size determination seem to hamper creating definitive rules –
or even rules of thumb – concerning necessary sample size (e.g., Mundfrom, Shaw, & Ke, 2005).

Instead of elaborating on general guidelines for sample size, more empirically grounded, individual-model-focused approaches to determining sample size in relation to parameter precision and power have been proposed. These approaches include Satorra and Saris (1985), MacCallum, Browne, and Sugawara (1996), and Muthén and Muthén (2002). The methods of both Satorra and Saris (1985) and MacCallum et al. (1996) estimate sample size in terms of the precision and power of an entire model using the chi-square statistic and RMSEA, respectively. In contrast, Muthén and Muthén (2002) evaluate sample size in terms of the precision and power of individual parameters in a model, while allowing the modeling of various conditions that researchers frequently encounter in their research, such as non-normality or type of indicator. Such modeling flexibility is certainly useful for estimating sample size, given that sample size and many variables affect each other in intricate ways.

In order to evaluate sample size, Muthén and Muthén (2002) use four criteria. First, parameter bias and standard error bias should not exceed |10%| for any parameter in the model. Second, the standard error bias for the parameter for which power is of particular interest should not exceed |5%|. Third, 95% coverage – the proportion of replications for which the 95% confidence interval covers the population parameter value – should fall between 0.91 and 0.98. One minus the coverage value equals the alpha level of 0.05. Coverage values should be close to the correct value of 0.95. Finally, power is evaluated in terms of whether it exceeds 0.80 – a commonly accepted value for sufficient power.

An analysis of the sample size of the current data based on Muthén and Muthén (2002) is presented in Table 5. Columns 2 and 3 show population and sample parameters. Population parameters are unstandardized parameters in Table 3. They are viewed as correct, true parameters from which numerous samples (replications) are generated in each run, and results over the replications are summarized. For example, using these values, the parameter bias for self-assessment measured by mathematics is calculated in the following manner: |0.9130 – 0.910|/|0.910| = 0.00330, or in other words, 0.330%. This is far below the criterion of 10%, thereby suggesting a good estimation of the parameter. The result is presented in Column 4. Column 5 shows the standard deviation of the parameters across replications. Column 6 shows the average of the standard errors across replications. The standard error bias for self-assessment measured by mathematics is |0.0743 – 0.0754|/|0.0754| = 0.01459, or in other words, 0.330%. This is far below the criterion of 10%, thereby suggesting a good estimation of the parameter. The result is presented in Column 7. In particular, we are interested in the effect of self-assessment and teacher assessment on academic achievement. The standard error biases for these parameters of interest are 0.413% and 0.545%, respectively. Neither exceeds 5%, thereby suggesting a good estimation of the parameter. Column 8 provides the mean square error of parameter estimates, which equals the variance of the estimates across replications plus the squared bias (Muthén & Muthén, 2007). Column 9 shows coverage, or the proportion of
replications where the 95% confidence interval covers the true parameter value. The value of 0.947 for self-assessment measured by mathematics is very close to 0.95, thereby suggesting a good estimation of the parameter. The last column shows the percentage of replications for which the parameter is significantly different from zero (i.e., the power estimate of a parameter). Column 10 shows that the power for self-assessment measured by mathematics is 1.000, which exceeds 0.80 and suggests sufficient power for the parameter. Together, these results provide good evidence for parameter precision and power for self-assessment measured by mathematics and suggest that the sample size for self-assessment measured by mathematics is sufficient. The same process is repeated for the remaining parameters. It should be noted that the power for the correlation between self-assessment and teacher assessment is low (0.339; see the last row). This suggests that the current sample size of 450 is not enough to distinguish the correlation from zero. Thus, although the sample size for the current model is adequate overall, the underpowered correlation indicates that caution should be exercised when interpreting it. The Appendix shows the Mplus syntax used for the current analysis.

Table 5. Mplus output for the Monte Carlo analysis to determine the precision and power of parameters

<table>
<thead>
<tr>
<th>Population parameter</th>
<th>Sample parameters averaged</th>
<th>Parameter bias</th>
<th>SD of sample parameters</th>
<th>Standard error of sample parameters</th>
<th>Standard error bias</th>
<th>Mean square error of parameters</th>
<th>95% coverage</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-assessment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>by English</td>
<td>0.910</td>
<td>0.9130</td>
<td>0.330</td>
<td>0.0754</td>
<td>0.0743</td>
<td>1.419</td>
<td>0.0057</td>
<td>0.947</td>
</tr>
<tr>
<td>Science</td>
<td>0.703</td>
<td>0.7024</td>
<td>0.885</td>
<td>0.0613</td>
<td>0.0609</td>
<td>0.613</td>
<td>0.0038</td>
<td>0.950</td>
</tr>
<tr>
<td>Teacher assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>by English</td>
<td>0.796</td>
<td>0.7421</td>
<td>0.226</td>
<td>0.0890</td>
<td>0.0876</td>
<td>1.573</td>
<td>0.0079</td>
<td>0.951</td>
</tr>
<tr>
<td>Science</td>
<td>0.528</td>
<td>0.5318</td>
<td>0.8720</td>
<td>0.0662</td>
<td>0.0660</td>
<td>0.684</td>
<td>0.0044</td>
<td>0.953</td>
</tr>
<tr>
<td>Academic achievement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>by English</td>
<td>0.483</td>
<td>0.4811</td>
<td>0.393</td>
<td>0.0613</td>
<td>0.0604</td>
<td>1.448</td>
<td>0.0038</td>
<td>0.945</td>
</tr>
<tr>
<td>Mathematics</td>
<td>0.534</td>
<td>0.5310</td>
<td>0.562</td>
<td>0.0638</td>
<td>0.0635</td>
<td>1.216</td>
<td>0.0034</td>
<td>0.947</td>
</tr>
<tr>
<td><strong>Self-assessment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and Teacher assessment</td>
<td>-0.092</td>
<td>-0.0894</td>
<td>2.826</td>
<td>0.0386</td>
<td>0.0373</td>
<td>1.345</td>
<td>0.0034</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Note. The column labels were slightly changed from original Mplus outputs to enhance clarity. Self-assessment by English refers to a path from the self-assessment factor to the English variable. Self-assessment with Teacher assessment refers to the correlation between these two factors.
Various types of models can be analyzed within the SEM framework. In addition to the models presented in Figures 1 and 2, we describe models often used in educational studies: confirmatory factor analysis, multiple-group analysis, and latent growth modeling. First, confirmatory factor analysis is used to examine whether the factor structure of a set of observed variables is consistent with previous theory or empirical findings (e.g., Brown, 2006). The researcher constructs a model using knowledge of the theory and/or empirical research, postulates the relationship pattern, and tests the hypothesis statistically. This reinforces the importance of theory in the process of model building. The models of self-assessment, teacher assessment, and academic achievement in Figures 1 and 2 represent different measurement models and must be verified through confirmatory factor analysis in terms of whether each of the three constructs are well represented by the three measurements of English, mathematics, and science. Unfortunately, each measurement model has only three observed variables, and this results in zero degrees of freedom (6 parameters to estimate – two factor loadings, three measurement errors, and one factor variance – and 3(3 + 1)/2 = 6 data points). The measurement models cannot be evaluated in the current model specification (see model identification in the Five Steps in an SEM Application above).

Various models can be analyzed using confirmatory factor analysis. For example, the often-cited study Holzinger and Swineford (1939) administered a battery of tests to measure seventh- and eighth-grade students in two Chicago schools. The tests were designed to measure mental ability, hypothesized to comprise spatial, verbal, speed, memory, and mathematics abilities. Although Holzinger and Swineford (1939) did not use SEM, the model closest to the one they hypothesized is shown in Figure 4A, and competing models that we postulated are shown in Figures 4B, 5A, and 5B. Figure 4A shows that mental ability comprises a general ability and five sub-abilities. Figure 4B is similar to Figure 4A but assumes a hierarchical relationship between a general ability and sub-abilities. Figure 5A assumes only a single general ability. Figure 5B hypothesizes no general ability and instead assumes correlated sub-abilities. A series of models can be tested on a single dataset using SEM by comparing model fit indices or using a chi-square difference test (see, for example, Brown, 2006; Shin, 2005).

Second, multiple-group or multiple-sample analysis aims to fit a model to two or more sets of data simultaneously. It allows us to test whether and to what extent measurement instruments (tests and questionnaires) function equally across groups, or, put another way, whether and to what extent the factor structure of a measurement instrument or theoretical construct of interest holds true across groups (e.g., Bollen, 1989). Multiple-group analysis involves testing across the samples whether factor loadings, measurement error variances, factor variances, and factor covariances are the same. Equivalence across groups suggests the cross-
IN’NAMI AND KOIZUMI

Figure 4. Confirmatory factor analysis of a model of mental ability: Bi-factor model (left) and higher-order model (right). The spatial test battery comprises (1) visual perception, (2) cubes, (3) paper form board, and (4) flags. The verbal test battery comprises (5) general information, (6) paragraph comprehension, (7) sentence completion, (8) word classification, and (9) word meaning. The speed test battery comprises (10) addition, (11) coding, (12) counting groups of dots, and (13) straight and curved capitals. The memory test battery comprises (14) word recognition, (15) number recognition, (16) figure recognition, (17) object-number, (18) number-figure, and (19) figure-word. The math test battery comprises (20) deduction, (21) numerical puzzles, (22) problem reasoning, (23) series completion, and (24) Woody-AcCall Mixed Fundamentals.
Figure 5. Confirmatory factor analysis of a model of mental ability: Single-factor model (left) and correlated-factor model (right)
validation or generalizability of findings. It is possible that factor loadings are similar in size across groups, while factor covariances are different. For example, Holzinger and Swineford’s (1939) data include seventh- and eighth-grade students of both genders from two different schools. It would be of interest to examine whether the bi-factor model of mental ability in Figure 4A is stable across grades, gender, and/or schools. For examples of applications of multiple-group analysis, see Byrne, Baron and Balev (1998), In’nami and Koizumi (2012), and Shin (2005).

Third, latent growth modeling is useful for evaluating longitudinal changes in aspects of individuals over time. It provides a great deal of information, including change at the individual and the group levels, pattern of change (e.g., linear, quadratic), and variables related to change, such as age, gender, motivation, and socioeconomic status (i.e., income, education, and occupation). For example, Tong et al. (2008) hypothesized that second language oral proficiency develops linearly when measured by vocabulary and listening tests at three time points over two years. Their model is presented in Figure 6A. Initial status, also called intercept,
that all factor loadings are fixed, unlike confirmatory factor analysis and multiple-group analysis.

More complex models can also be analyzed using latent growth modeling. Yeo, Fearrington, and Christ (2011) investigated how demographic variables – gender, income, and special education status – affect reading growth at school. Their model, shown in Figure 6B, differs primarily from the model in Figure 6A in two ways. First, the loadings for the growth rate factor are fixed to be 0, 5, and 9 – three time points of data collection (August = 0, January = 5, and May = 9) – because we assume that the authors were interested in nine-month growth and rescaled the slope factor loadings accordingly. It should be noted that growth rate factors, whether fixed to be 1, 2, and 3, or 0, 5, and 9, do not change the data-model fit (e.g., Hancock & Lawrence, 2006). Second, the three demographic variables are incorporated into the model as predictors of initial status and growth rate. The results indicate the relative impact of the external variables on the initial level of reading proficiency and on the growth rate of reading proficiency over nine months. For further examples of latent growth modeling, see Kieffler (2011) and Marsh and Yeung (1998).

SOFTWARE

Since Byrne (2012a) provides a detailed, comparative review of SEM software, we will present only a brief treatment of SEM software programs (also see Narayanan, 2012). There are several major commercial programs for performing SEM, including Amos (Analysis of Moment Structures; Arbuckle, 1994-2012), CALIS (SAS Institute, 1990-2012), EQS (Equations; Bentler, 1994-2011), LISREL (Linear Structural Relationships; Jöreskog & Sörbom, 1974-2012), and Mplus (Muthén & Muthén, 1998-2012). Free programs are also available, including Mx (Neale, Boker, Xie, & Maes, 2003) and three R-language packages: the OpenMx package (Boker, Neale, Maes, Wilde, Spiegel, Brick, et al., 2007-2012), the “sem” package (Fox, Nie, Byrnes, Culbertson, Friendly, Kramer, & Monette, 2012), and the “lavaan” package (Rosseel, 2012). The choice of software depends on the purpose of the SEM analysis and the proficiency of the user’s computing skills. Byrne (2012a) indicates three aspects related to deciding on the best software: (a) familiarity with SEM concepts and application, (b) the types of SEM model to be tested, and (c) preference concerning manual or graphic interface. She argues that beginners may find Amos or EQS the easiest to use, and that more advanced learners may prefer to use EQS, LISREL, or Mplus. Unlike Amos, EQS, and LISREL, Mplus requires command-based inputs, and learners who are used to graphic interfaces may need some time to become comfortable with the program. In order to familiarize themselves with software, novice learners are referred to Byrne (1998, 2006, 2010, 2012b), whereas advanced learners wishing to use R-based packages are referred to Fox, Byrnes, Boker, and Neale (2012).
SOME DIRECTIONS FOR LEARNING MORE ABOUT SEM

Since SEM is a versatile technique, a single book chapter would not be able to cover a wide range of analyses that can be modeled using SEM. In order to deepen learning regarding SEM, we recommend reading through Byrne (1998, 2006, 2010, 2012b) for LISREL, EQS, Amos, and Mplus, trying to analyze the accompanying datasets, and ensuring that one can replicate findings. Based on our own experience with Byrne (2010) for Amos, and Byrne (2006) for EQS datasets, as well as on discussion with skilled SEM users, we believe that this is probably the best approach to familiarize oneself with SEM and apply the techniques to one’s own data.

For providing answers to questions that may arise with regard to particular issues related to SEM, the following recent references may be useful: Bandalos and Finney (2010), Brown (2006), Cudeck and du Toit (2009), Hancock and Mueller (2006), Hoyle (2012), Kaplan (2009), Kline (2011), Lomax (2010), Mueller and Hancock (2008, 2010), Mulaik (2009), Raykov and Marcoulides (2006), Schumacker and Lomax (2010), Teo and Khine (2009), and Ullman (2007). For more on how researchers should report SEM results, see Boomsma, Hoyle, and Panter (2012); Gefen, Rigdon, and Straub (2011); Jackson, Gillaspy Jr., and Purc-Stephenson (2009); Kahn (2006); Kashy, Donnellan, Ackerman, and Russell (2009); Martens (2005); McDonald and Ho (2002); Schreiber, Nora, Stage, Barlow, and King (2006), and Worthington and Whittaker (2006). Reporting a correlation matrix with means and standard deviations is strongly recommended as this allows one to replicate a model, although replication of non-normal and/or missing data requires raw data (for example, see In'nami & Koizumi, 2010). Of particular interest is the journal *Structural Equation Modeling: An Interdisciplinary Journal* published by Taylor & Francis, which is aimed at those interested in theoretical and innovative applied aspects of SEM. Although comprising highly technical articles, it also includes the Teacher’s Corner, which features instructional modules on certain aspects of SEM, and book and software reviews providing objective evaluation of current texts and products in the field.

For questions pertaining to particular features of SEM programs, user guides are probably the best resource. In particular, we find the EQS user guide (Bentler & Wu, 2005) and manual (Bentler, 2005) outstanding, as they describe underlying statistical theory in a readable manner as well as stepwise guidance on how to use the program. A close look at manuals and user guides may provide answers to most questions. LISREL and Mplus users should take full advantage of technical appendices, notes, example datasets, and commands, which are all available online free of charge (Mplus, 2012; Scientific Software International, 2012). The Mplus website also provides recorded seminars and workshops on SEM and a schedule listing of upcoming courses.

For problems not addressed by the abovementioned resources, we suggest consulting the Structural Equation Modeling Discussion Network (SEMNET). It was founded in February 1993 (Rigdon, 1998) and archives messages by month. Because of the large number of archived messages collected over the past two
decades (thanks to the mushrooming popularity of SEM across many disciplines), SEMNET is a treasure trove of questions and answers on virtually every aspect of SEM. Questions should only be posted if answers to them cannot be found in the archive. As with any other academic online discussion forum, contributors to SEMNET take questions seriously and spend precious time responding to them. We recommend that anyone wishing to receive a good reply should mention that answers were not found in the archive and articulate problems in enough detail for others to respond. Posting a command/script/syntax file is a good idea.

SEM is constantly evolving and expanding. The development and application of new techniques are causing numerous academic disciplines to move increasingly toward a better understanding of various issues that require tools that are more precise. SEM analysis offers powerful options for analyzing data from educational settings, and techniques discussed in this chapter will enable educational researchers to be in a better position to address a wide range of research questions. By employing SEM analysis appropriately, we will be able to contribute much in years to come.

APPENDIX

Mplus Input for the Monte Carlo Analysis for Determining the Precision and Power of Parameters

TITLE: THREE-FACTOR, NORMAL DATA, NO MISSING

MONTECARLO:
NAMES ARE X1-X9;
NOBSERVATIONS = 450; ! SAMPLE SIZE OF INTEREST
NREPS = 10000;
SEED = 53567;

MODEL POPULATION:
f1 BY X1@1 X2*.91 X3*.70;
f2 BY X4@1 X5*.74 X6*.53;
f3 BY X7@1 X8*.48 X9*.53;
X1*.77; X2*.37; X3*.61; X4*.91; X5*.48; X6*.76; X7*.66; X8*.63;
X9*.66;
f1*.88; f2*.94; f3*.74;
f3 ON f1*.50; f3 ON f2*.38;
f1 WITH f2*.09;

MODEL:
f1 BY X1@1 X2*.91 X3*.70;
f2 BY X4@1 X5*.74 X6*.53;
f3 BY X7@1 X8*.48 X9*.53;
X1*.77; X2*.37; X3*.61; X4*.91; X5*.48; X6*.76; X7*.66; X8*.63;
X9*.66;
f1*.88; f2*.94; f3*.74;
f3 ON f1*.50; f3 ON f2*.38;
f1 WITH f2*.09;

ANALYSIS: ESTIMATOR = ML;
OUTPUT: TECH9;
REFERENCES


Teo, T., & Khine, M. S. (Eds.). (2009). Structural equation modeling in educational research: Concepts and applications. Rotterdam, the Netherlands: Sense Publishers.


Yo In’nami
Shibaura Institute of Technology
Japan

Rie Koizumi
Juntendo University
Japan