Available online at https://journal.rescollacomm.com/index.php/ijeer/index

International Journal of Ethno-Sciences and Education Research

e-ISSN 2776-7590

Vol. 1, No. 3, pp. 52-60, 2021

Study on Structural Equation Modeling for Analyzing Data

M. Ihsan Khairi[1*], Dwi Susanti[2], Sukono[3]

[1,2,3] Department of Mathematics, Faculty Of Mathemathics and Natural Sciences, Universitas Padjadjaran, Indonesia

**Corresponding author email: ihsan.khairi@outlook.com*

Abstract

The Structural Equation Model (SEM) is a combination of two separate statistical methods, namely factor analysis developed in psychology and psychometry and simultaneous equation model developed in econometrics. Factor analysis was first introduced by Galton in 1869 and Pearson (Pearson and Lee, 1904). Spearman's (1904) research is the development of a general factor analysis model in research relating to the structure of mental abilities, Spearman stated that the intercorrelation test between mental abilities can determine general ability factors and special ability factors. SEM is a combination of factor analysis and path analysis into one comprehensive statistical method. Path analysis itself is the forerunner of the structural equation of Sewwl Wright's research in the field of biometrics. Wright's contribution is to be able to show that the correlation between variables is related to the parameters of a model described by a path (path diagram). In SEM there are 2 variables, namely latent variables (exogenous and endogenous) and indicator variables. SEM has 2 equation models, namely the measurement equation model and the structural equation model. SEM also has 2 errors, namely the error for the measurement equation model and the error for the structural equation model. In general, SEM is formed from the relationship between latent variables and their respective indicator variables. To test whether the existing indicator variables are valid indicators for measuring the latent construct, Confirmatory Factor Analysis (CFA) is used. Data analysis with SEM must meet the existing SEM assumptions. The model feasibility test is carried out based on the goodness of fit criteria. The stages in SEM analysis are theoretical model development, flow chart drawing, flow chart conversion into equation form, input matrix and model parameter estimation techniques, model problem identification, evacuating model parameter estimates, model interpretation, and model modification.

Keywords: Structural Equation Model, Latent Variables, Indicator Variables, Errors in SEM, Measurement Equation Model, Structural Equation Model, CFA

1. Introduction

Structural Equation Modeling (SEM) was first discovered by a scientist named Joreskog in 1970, Structural equation modeling is a second-generation multivariate analysis technique that combines factor analysis and path analysis, allowing researchers to simultaneously test and estimate latent variables, both exogenous and endogenous which also involves the indicator variables. In the structural equation model, there are several estimation methods. The method commonly used is the Maximum Likelihood (ML) method or the maximum likelihood method. ML is a method that has an unbiased estimator and a minimum variance (Ghozali, 2017).

SEM is a combination of factor analysis and path analysis into one comprehensive statistical method. Path analysis itself was the forerunner of the structural equations of Sewwl Wright's research in 1918, 1921, 1934, and 1960 in the field of biometrics. Wright's contribution is to be able to show that the correlation between variables is related to the parameters of a model described by a path (path diagram). Wright also stated that the resulting equation model can be used to estimate the direct effect, indirect effect, and total effect. The first application of path analysis developed by Spearman is statistically equivalent to factor analysis (Ghozali, 2017).

Structural Equation Modeling (SEM) is a multivariate analysis technique developed to cover the limitations of previous analytical models that have been widely used in statistical research (Sasongko et al., 2016).

2. Strutural Equation Model

Variabel-Variabel Dalam SEM.

A. Latent variables

Latent variables are abstract concepts, for example, people's behavior, attitudes, feelings, and motivations. Latent variables can only be observed imperfectly through their effects on the observed variables. SEM has two types of latent variables, namely endogenous latent variables and exogenous latent variables. Exogenous variables appear as independent variables in all equations in the model, with the mathematical notation letter ("ksi"). Endogenous variables are dependent variables in at least one equation in the model, with the mathematical notation letter ("eta") (Wijanto, 2008).

B. Indicator variables

Indicator or measurable variables are variables that can be observed or measured empirically and are often called indicators. The observed variable is the effect or measure of the latent variable. In the survey method using a questionnaire, each question on the questionnaire represents an observed variable (so if a questionnaire has 50 questions, then there will be 50 observed variables). The observed variables related to or an effect of the exogenous latent variable (ξ) are given mathematical notation with the label X, while those related to the endogenous latent variable (η) are labeled Y. The symbol of the path diagram of the observed variable is a square (Wijanto, 2008).

Models in SEM

There are 2 types of models in SEM, namely:

A. Structural model

Structural Model describes the relationship that exists between latent variables. Parameters that show the relationship between endogenous and exogenous variables are denoted by . Parameters that show the relationship between endogenous variables and other endogenous variables are denoted by . In SEM, the exogenous latent variable is independently covariant and the covariance matrix of this variable is denoted by (Wijanto, 2008). The general structural model equation can be written as follows:

Suppose the random vectors $\eta^T = (\eta_1, \eta_2, \dots \eta_m)$ and $\xi^T = (\xi_1, \xi_2, \dots \xi_n)$ are endogenous and exogenous variables, respectively, forming a simultaneous equation with a linear relationship system:

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \quad (1)$$

B and Γ are the coefficient matrix of latent variables and $\zeta = (\zeta_1, \zeta_2, \dots \zeta_m)$ are error vectors in structural equations. The element B represents the effect of variable η in the other variable η, and element Γ presents the direct effect of variable ξ in variable η. It is assumed that ξ is not correlated with ζ and $I - \mathrm{B}$ is nonsingular. (Joreskog and Sorbon, 1989).

Structural model equations can be described again and the following results are obtained:

$$\eta = (I - \mathrm{B})^{-1}(\Gamma\xi + \zeta) \quad (2)$$

with:

η = vector of endogenous variables of size m× 1
ξ = vector of exogenous variables of size n× 1
B = coefficient matrix of endogenous variables of size $m \times m$
Γ = exogenous variable coefficient matrix of size $m \times n$
ζ = vector error in structural equation

B. Measurement Model

In SEM, each latent variable usually has several measures or indicators. SEM users most often associate latent variables with their indicators through measurement models in the form of factor analysis which are widely used in psychometrics and sociometry. In the SEM model, each latent variable is modeled as a factor that underlies the related indicators. The load of factors that relate the latent variable to the indicator is denoted by (lambda). For indicator X is denoted by _x and for indicator Y is denoted by _y (Wijanto, 2008). The equation of the measurement model can be written as follows:

$$X = \Lambda_x \xi + \delta \quad (3)$$
$$Y = \Lambda_y \xi + \epsilon \quad (4)$$

with:

X : indicator variable vector of exogenous variable $q \times 1$
Λ_x : matrix for loading factor (λ) or coefficient that shows relationship between ξ and X with size $q \times n$
δ : vector of measurement model error for X with size $q \times 1$
Y : indicator variable vector of endogenous variable $p \times 1$
Λ_y : matrix for loading factor (λ) or coefficient that shows relationship between η and Y with size $p \times m$
ϵ : vector of measurement model error for Y with size $p \times 1$

Errrors in SEM

A. Structural Error

In general, SEM users do not expect that the independent variable can perfectly predict the dependent variable, so that a structural error component is usually added in a model, which is given the Greek symbol (zeta). In order to obtain a consistent parameter estimate, this structural error is assumed to be uncorrelated with the exogenous variables of the model. However, structural faults can be modeled to correlate with other structural faults (Wijanto, 2008).

B. Measurement Error

In SEM the indicators or observed variables cannot perfectly measure the related latent variables. To model this imperfection, a component representing measurement error was added to the SEM, which was labeled with the Greek letter (delta) for the measurement error associated with the observed variable X, while that associated with the variable Y was labeled with the Greek letter (epsilon). . The measurement errors of may be covariant with each other, although by default they are not covariant with each other. The covariance matrices of are denoted by the Greek letter _δ (Theta delta) and are by default a diagonal matrix, while the covariance matrices of are denoted by the Greek letter _ϵ (Theta epsilon). When a latent variable is only reflected/measured by a single observed variable, estimating the value of the associated measurement error is difficult/impossible. In this case, the measurement error must be specified before estimating the parameter or the measurement error can be considered as non-existent or zero (Wijanto, 2008).

General Form of SEM

Based on the previous explanation, a model from SEM can be drawn up. SEM is formed based on the relationship between latent variables where each latent variable is measured by its respective indicator variables. An example of the general form of SEM can be seen in Figure 1 (Wijanto, 2008).

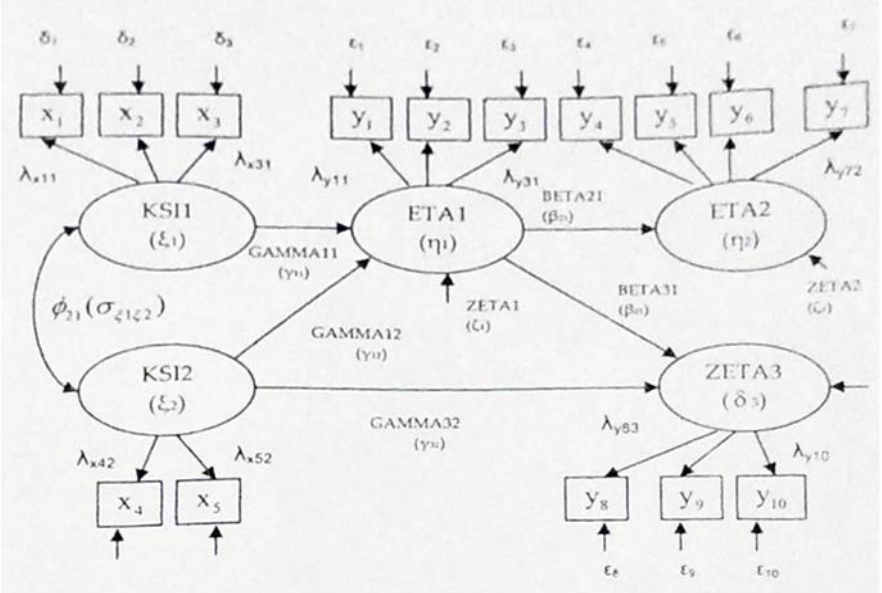

Figure 1. General form of SEM

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis (CFA) is a measurement model that shows whether a latent variable is measured by one or more indicator variables (Sasongko et al., 2016). CFA tests whether the existing indicator variables are valid indicators of measuring the latent construct (Ghozali, 2017). The general model of the CFA is as follows:

$$X = \Lambda_x \xi + \delta \quad (5)$$

with:

X : vector for the indicator variable with size q x 1

Λ_x : matrix for loading factor (λ) a or coefficient that shows the relationship between x with ξ with size q x n

ξ : vector for the latent variable of size n x 1

δ : vector for the measurement error of size q x 1

The level of significance with a loading factor value > 0,4 indicates the indicator can explain each factor variable (Hair et al., 2010).

3. Assumptions and Criteria for Goodness of Fit

3.1. SEM Assumptions

Ferdinand in (Santosa, 2006), the assumptions that must be met in the data collection and processing procedures analyzed by SEM modeling are:

a. The minimum sample size is 100 and then using a comparison of 5 observations for each estimate parameter.
b. The distribution of the data must be analyzed to see if the assumption of normality is met.
c. Outliers are observations with extreme values both univariate and multivariate that arise because of the combination of unique characteristics they have and are bound to be very different from other observations.
d. Detect multicollinearity and singularity of the determinant of the covariance matrix. The very small value of the determinant of the covariance matrix indicates the existence of a multicollinearity or singularity problem.

3.2. Criteria Goodness of Fit

There are three types of Goodness-of-Fit measures, namely Absolute Fit Indices, Incremental Fit Indices, and Parsimonious fit indices (Haryono and Wardoyo, 2012).

Absolut Fit Measures

According to Wijanto (2008) the absolute fit measure determines the degree of prediction of the overall model (structural and measurement models) to the correlation and covariance matrices. Of the various absolute fit measures, the most commonly used measures to evaluate SEM are:

a. *Chi-Square* (χ^2)

The first statistic and the only statistical test in GOF is χ^2. Chi-square is used to test how close the match is between the covariance matrix of the sample S and the covariance matrix of the model $\sum \theta$. The statistical test χ^2 is

$$\chi^2 = (n-1)F(S, \sum(\theta)) \quad (6)$$

Which is a Chi-square distribution with a *degree of freedom* (df) of c – p ; in this case, c = (nx+ny) (nx+ny+1)/2 is the number of non-redundant variance-covariance matrices of the observed variable, nx is the number of observed variables x, ny is the number of observed variables y. The p is the number of estimated parameters and n is the sample size.

The expected value of Chi-square (χ^2) is low. A low Chi square value indicates that the null hypothesis is accepted. This means that the predicted and actual input metrics are not statistically different

Joreskog dan Sobron in Wijanto (2008) say that *Chi-square* (χ^2) should be treated more as a measure of goodness of fit or badness of fit and not as a statistical test. *Chi-square* (χ^2) is referred to as *badness of fit* because the higher *Chi-square* (χ^2) value indicates a bad fit while the Chi-square value (χ^2) a small one indicates a good fit.

Chi-square (χ^2) cannot be used as the only measure of the overall fit of the model, one of the reasons is because *Chi-square* (χ^2) is sensitive to sample size. When the sample size increases, the *Chi-square* (χ^2) value will also increase and lead to model rejection even though the value of the difference between the sample covariance matrix (S) and the model covariance matrix or $\sum(\Theta)$ has been minimal and small.

b. *Goodness Of Fit Idices* (GFI)

At first GFI was proposed by Jöreskog and Sörbom in Wijanto (2008) for estimation with ML and ULS, then generalized to other estimation methods by Tanaka and Huba (1985). GFI can be classified as an absolute fit measure, because it basically compares the hypothesized model with no model at all ($\sum(\theta)$). The formula for GFI is as follows:

$$GFI = 1 - \frac{F}{F_0} \tag{7}$$

with F is the minimum value of the hypothesized model

F_0 is the minimum value of F when no model is hypothesized.

The GFI value ranges from 0 (poor fit) to 1 (perfect fit), and the GFI value 0,90 is good fit, while 0,80 < GFI < 0,90 is often referred to as marginal fit.

c. *Root Mean Square Error Of Approximation* (RMSEA).

This index was first proposed by Steinger and Lind in Wijanto (2008) and is currently one of the most informative indices in SEM. The RMSEA calculation formula is as follows:

$$RMSEA = \sqrt{\frac{F_0}{df}} \tag{8}$$

with $F_0 = Max\{F - df/(n-1), 0\}$

RMSEA value < 0,05 indicates close fit, while 0,05 RMSEA 0,08 indicates good fit (Brown and Cudeck, 1993).

Incremental Fit Indices

The measure of incremental fit compares the proposed model with the baseline model which is often referred to as the null model or the independence model and the saturated model. Null model is a model with the worst fit of the data-model ("worst fit"). The saturated model is the one with the best data-model fit (“best fit”). The basic model or null model is a model in which all variables in the model are independent of each other (or all correlations between variables are zero) and most restricted (most restricted), Byrne (1998) in Wijanto (2008).

The concept of incremental fit will place the data-model fit level between the null model and the saturated model. The level of fit of the data-model that is between the null model and the saturated model is called the nested model, Mueller in Wijanto (2008). This incremental fit measure contains a measure that represents the comparative fit to base model point of view. The closer to the saturated model, the better the fit. Of the various measures of incremental fit, the ones commonly used to evaluate SEM are:

a. *Adjusted Goodnes Of Fit Index* (AGFI)

Jöreskog and Sörbom in Wijanto (2008), AGFI is an extension of GFI adjusted for the ratio between the degree of freedom of the null / independence / baseline model and the degree of freedom of the hypothesized or estimated model. AGFI can be calculated by the following formula:

$$AGFI = 1 - \frac{df_o}{df_h}(1 - GFI) \quad (9)$$

with df_o is the degree of freedom of no model

df_h is the degree of freedom of the hypothesized model.

As with GFI, the AGFI value ranges from 0 to 1 and the AGFI value 0,90 indicates good fit. While 0,80 < GFI < 0,90 is often referred to as marginal fit.

b. *Tucker-Lewis Index / Non Normed Fit Index* (TLI/NNFI).

It was first proposed as a tool for evaluating factor analysis, but is now being developed for SEM.

$$TLI = \frac{(\chi_i^2 / df_i) - (\chi_h^2 / df_h)}{(\chi_h^2 / df_i) - 1} \quad (10)$$

With χ_i^2 is the chi square of the null/independence model.

χ_h^2 is the chi square of the hypothesized model.

df_i is *degree of freedom* from *null model*.

df_h is *degree of freedom* from hypothesized model.

TLI value ranges from 0 to 1 with TLI value 0,90 indicating good fit and 0,80 < TLI < 0,90 is marginal fit.

c. *Norm Fit Index* (NFI)

In addition to NNFI, Bentler and Bonnet in Wijanto (2008) also proposed a GOF measure known as NFI. This NFI has a value ranging from 0 to 1. The NFI value 0,90 indicates good fit, while 0,80 < NFI < 0,90 is often referred to as marginal fit. To obtain the NFI value, the following formula can be used:

$$NFI = \frac{\chi_i^2 - \chi_h^2}{\chi_i^2} \quad (11)$$

d. *Comparative Fit Index* (CFI)

Bentler in Wijanto (2008) adds to the inventory of incremental matches through CFI, whose value can be calculated by the formula:

$$CFI = 1 - \frac{l_1}{l_2} \quad (12)$$

with $l_1 = \max(l_h, 0)$ dan $l_2 = \max(l_h, l_i, 0)$

$l_h = [(n-1)F_h - df_h]$ dan $l_i = [(n-1)F_i - df_i]$

The CFI value will range from 0 to 1. The CFI value 0,90 indicates good fit, while 0,80 < CFI < 0,90 is often referred to as marginal fit.

e. *Incremental Fit Index* (IFI)

Bollen in Wijanto (2008) proposes IFI, whose value can be obtained from:

$$IFI = \frac{nF_i - nF_h}{nF_i - df_h} \quad (13)$$

The IFI value will range from 0 to 1. The IFI value 0,90 indicates good fit, while 0,80 < IFI < 0,90 is often referred to as marginal fit.

f. *Relative Fit Index* (RFI).

RFI from Bollen in Wijanto (2008) can be calculated using the formula:

$$RFI = \frac{F_h / df_h}{F_i / df_i} \quad (14)$$

with F_h is the minimum value of F from the hypothetical model

F_i is the minimum value of F from *null/independence*

Parsimonious Fit Indices

According to Wijanto (2008) models with relatively few parameters (and relatively many degrees of freedom) are often known as models that have parsimony or high efficiency. Meanwhile, a model with many parameters (and a few degrees of freedom) can be said to be a complex model and lack parsimony. The parsimony fit measure relates the model's GOF to the number of parameters estimated, i.e., required to achieve a fit at that level. In this case, parsimony can be defined as obtaining the highest degree of fit for each degree of freedom. This means that higher parsimony is better. Of the various parsimony fit measures, the measures commonly used to evaluate SEM are:

a. *Parsimonious Normed Fit Index* (PNFI).

According to James, Begink and Brett in Wijanto (2008) PNFI is a modification of NFI. PNFI takes into account the number of degrees of freedom to achieve a level of fit. PNFI is defined as follows:

$$PNFI = \frac{df_h}{df_i} NFI \tag{15}$$

with df_h is the degree of freedom of hypothesis model
df_i is the degree of freedom of *null model*

The higher the PNFI score the better. The use of PNFI is mainly for comparison of two or more models that have different degrees of freedom. The PNFI was used to compare alternative models, and no acceptable fit was recommended. However, when comparing the 2 models, the difference in PNFI values of 0,06 to 0,09 indicates a fairly large model difference (Haryono and Wardoyo, 2012).

b. *Parsimonius Goodness of Fit Index* (PGFI).

In contrast to AGFI which modifies GFI based on the degree of freedom, PGFI is based on the parsimony of the model (Wijanto, 2008). PGFI makes adjustments to GFI in the following ways:

$$PGFI = \frac{df_h}{df_0} GFI \tag{16}$$

The PGFI value ranges between 0 and 1 with a higher value indicating a better parsimony model.

In empirical research practice, a researcher does not have to meet all the goodness of fit criteria. According to Hair et al (2010) in Latan (2011), the use of 4 to 5 goodness of fit criteria is considered sufficient to assess the feasibility of a model, provided that each goodness of fit group is absolute fit indices, incremental fit indices and parsimonious fit. indices are represented.

4. SEM Stages

The steps of data analysis using SEM are as follows (Ghozali, 2017):

Theoretical Model Development

The theoretical development of the model is based on an existing and strong theory. The strength of the causal relationship between the two variables does not lie in the analytical method, but in the theoretical justification to support the model analysis (Ghozali, 2017).

Flowchart Drawing

In the second step, the relationship between exogenous and endogenous variables and their measuring variables will be described based on the previous theoretical development of the model (Ghozali, 2017).

Convert Flowchart into Equation

There are 2 equation models in SEM, namely a measurement equation model to express the relationship between latent variables and their indicator variables and a structural equation model to express the relationship between various latent variables (Ghozali, 2017).

Input Matrix and Model Parameter Estimation Technique

SEM only uses the variance/covariance matrix or correlation matrix as input data for the overall estimation it performs. Individual observations can be used, but the input data will be converted into a covariance matrix or correlation matrix before estimation. The covariance matrix is used when the purpose of the analysis is to test a model that already has a theoretical concept. The correlation matrix is used when the purpose of the analysis only wants to see the pattern of relationships between variables (Ghozali, 2017).

Parameter estimation techniques that can be used in SEM are Maximum Likelihood, Generalized Least Square Estimation, and Asymptotically Distribution-Free Estimation (Ghozali, 2017).

Model Problem Identification

According to Ghozali (2017) there are 3 possibilities that can occur in SEM:

- *Model Under–identified* if the value of $t \geq \frac{s}{2}$
- *Model Just–identified* if the value of $t = \frac{s}{2}$
- *Model Over –identified* if the value of $t \leq \frac{s}{2}$

where t is the estimated number of parameters
s is the number of variance and covariance between the manifest variables which is $(p+q)(p+q+1)$
p is the number of variable y (indicator of endogenous latent variable)
q is the number of variables x (exogenous latent variable indicator)

If the model is an under-identified model, then the model cannot be analysed

Evaluating Model Parameter Estimation

In the process of estimating the model parameters, firstly, a confirmatory analysis (CFA) is carried out on the measurement equation model for exogenous and endogenous variables. Secondly, a feasibility test of the model is carried out by taking into account the loadin factor value (expected value 0,5) and the goodness of fit criteria. If the model is not feasible, modify the model by eliminating indicator variables that do not meet the model's eligibility criteria. Third, combine the final result model from the confirmatory analysis of the exogenous and endogenous variables and estimate the parameters of the model. Fourth, testing the SEM assumptions, goodness of fit feasibility test, validity test and reliability test (Ghozali, 2017).

Some goodness of fit criteria are chi- quare, probability, Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Tucker-Lewis Index (TLI), Root Mean Square of Approximation (RMSEA) (Ghozali, 2017).

Interpreting Models and Modifying Models

Interpreting the model and modifying the model for models that do not meet the test requirements. Before modifying the model, it is necessary to observe the value of the loading factor and standardized residuals generated by the model. An indicator that has a loading factor value of more than or equal to 0,5 indicates that the indicator meets convergent validity. Meanwhile, an indicator with a loading factor value below 0,5 means that it does not meet convergent validity, so the indicator is eliminated from the model (Ghozali, 2017).

5. Conclusion

There are 2 variables, 2 equation models, and 2 errors in SEM. Variables in SEM are latent variables and indicator variables. The equation model in SEM is a measurement equation model and a structural equation model. Errors in SEM are structural errors and measurement errors. CFA on the measurement model is carried out to test whether the existing indicator variables are valid indicators of measuring the latent construct. There are several goodness of fit test criteria that can be used in SEM including GFI, AGFI RMSEA, and TLI. The SEM steps are as follows: 1) Theoretical model development, 2) Flowchart drawing, 3) Convert flow chart into equation form, 4) Input matrix and model parameter estimation techniques, 5) Identify the model problem, 6) Evacuating model parameter estimates, and 7) Interpreting the model and modifying the model

References

Ghozali, I. (2017). *Model Persamaan Struktural konsep dan Aplikasi with Program Amos 24.* Semarang: Universitas Diponegoro.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis Seventh Edition.* New York: Prentice Hall International.

Haryono, S., & Wardoyo, P. (2012). *Structural Equatuon Modeling Untuk Penelitian Manajemen Menggunakan AMOS 18.00*, Bekasi: Intermedia Personalia Utama

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 User's Reference Guide.*

Santosa, J. (2006). *Analisis Faktor-faktor Yang Mempengaruhi Strategi Integrasi untuk Meningkatkan Kinerja Pemasaran (Studi Kasus Distributor Makanan dan Minuman di Kota Semarang).* Tesis. Semarang: Universitas Diponegoro.

Sasongko, E. N., Mustafid, & Rusgiyono, A. (2016). Penerapan Metode Structural Equation Modeling untuk Analisis Kepuasan Pengguna Sistem Informasi Akademik Terhadap Kualitas Website (Studi Kasus Pada Website Sia.undip.ac.id). *GAUSSIAN*, *5*(3), 395-404.

Wijanto, S. H. (2008). *Structural Equation Modeling with LISREL 8.8: Konsep & Tutorial.* Yogyakarta: Graha Ilmu.