Open Access Indonesia Journal of Social Sciences

Journal Homepage: https://journalsocialsciences.com/index.php/OAIJSS

Data Analysis Procedures with Structural Equation Modelling (SEM): Narrative Literature Review

Rachmat Hidayat[1*], Patricia Wulandari[2]

[1]Department of Biology, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia

[2]Cattleya Mental Health Center, Palembang, Indonesia

ARTICLE INFO

Keywords:
Data analysis
Structural equation modelling
Variable

***Corresponding author:**
Rachmat Hidayat

E-mail address:
rachmathidayat@fk.unsri.ac.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.37275/oaijss.v5i6.142

A B S T R A C T

The relationship between variables in structural equation modeling (SEM) forms a structural model. This structural model can be explained through structural equations, such as in regression analysis. This structural equation describes the prediction of the latent (exogenous) independent variable on the latent (endogenous) dependent variable. Researchers who use analysis with structural equation models need to know whether the model built with empirical data has a unique value or not so that the model can be estimated. If the model does not have a unique value, then the model cannot be identified (unidentified). The cause of a model is categorized as unidentified because the information contained in empirical data is not sufficient to produce a unique solution in calculating model estimation parameters. This literature review aims to describe the process of data analysis using SEM.

1. Introduction

There is a principal difference between regression analysis and path (path analysis) and SEM in terms of measuring variables. In the regression analysis, the dependent and independent variables are variables that can be measured directly (observable), whereas, in SEM, the dependent and independent variables are variables that cannot be measured directly (unobservable). Unobserved variables are also often called latent variables. The structural equation model or SEM is a model that explains the relationship between latent variables, so the SEM model is often referred to as latent variable analysis or linear structural relationship. The relationship between variables in SEM is the same as the relationship in path analysis (Astrachan, 2014). However, in explaining the relationship between latent variables, the SEM model differs from path analysis, where path analysis uses observable variables while SEM uses unobservable variables (Babin et al., 2008). The relationship between variables in SEM forms a structural model. This structural model can be explained through structural equations, such as in regression analysis. This structural equation describes the prediction of the latent (exogenous) independent variable on the latent (endogenous) dependent variable.

Model specifications

SEM begins by specifying the research model. The analysis will not begin until the researcher specifies a model that shows the relationship between the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

variables to be analyzed. Through the steps below, researchers can obtain the desired model: 1) Define the latent variables of the study, 2) Define the observed variable, and 3) Define the relationship between latent variables and observed variables (Bagozzi et al., 2012). Pay attention to aspects of the unidimensional or multidimensional construct variable. Unidimensional constructs (first-order constructs) are constructs that directly describe the relationship between latent variables and observed variables reflectively (arrows away from latent variables) or formatively (arrows towards latent variables). Multidimensional constructs (second-order constructs) are constructs formed from several unidimensional constructs.

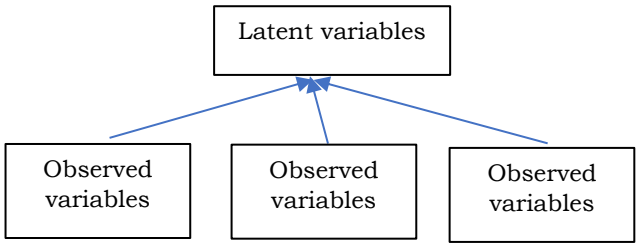

Figure 1. Unidimensional construct (First order construct).

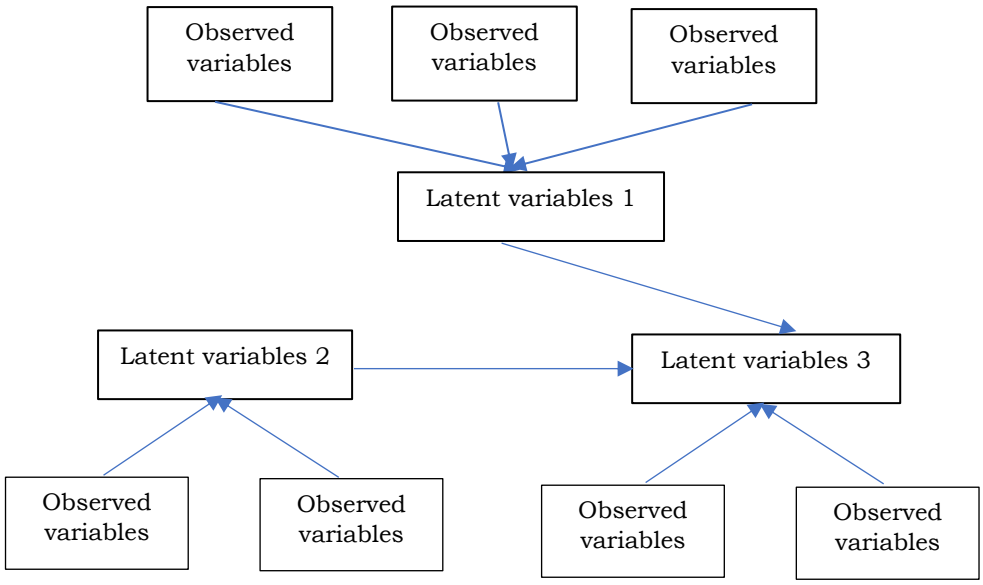

Figure 2. Multidimensional construct (Second order construct).

Model identification

Researchers who use analysis with structural equation models need to know whether the model built with empirical data has a unique value or not so that the model can be estimated. If the model does not have a unique value, then the model cannot be identified (unidentified). The cause of a model is categorized as unidentified because the information contained in empirical data is not sufficient to produce a unique solution in calculating model estimation parameters (Baker, 1999; Bido et al., 2012). An example of an under-identified case is A x B = 1000. The question is, what is the value of A or B? To determine the value of A or B, of course, the answers vary widely. It can be 100 x 10, 500 x 2, 250 x 4, 200 x 5 or 1 x 1000. To ensure an answer, we must choose the most

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

appropriate (unique) answer, called model identification. This example also occurs in SEM, where the theoretical model built and empirical data are not sufficient to produce a unique solution in calculating model parameter estimates (Byrne, 2010). However, if we determine the value of A = 100, then the value of B = 10 will be automatic. This can also be done in SEM analysis to overcome the unidentified model by constraining the model by 1) Adding indicators or manifest variables from latent constructs, 2) Determining the value of additional fixed parameters so that the calculation of the degree of freedom becomes positive (this method is most often used by researchers), 3) Assuming that the parameters with each other have the same value.

It should be noted that the use of the three methods above to change an under-identified model must be, in theory, not merely done so that the model can be identified. There are three possible identification models in SEM: 1) Under-identified model, where the value of $t \geq s/2$; namely, a model with a greater number of estimated parameters than the known amount of data (the data is the variance and covariance of the observed variables). For example, there is the equation X + Y = 10, representing 1 known piece of data and 2 parameters to be estimated, namely X and Y, so the number of df = 1-2 = -1. From the understanding of the unidentified model in SEM, it has df = the amount of data that is known - the number of parameters estimated < 0. So it can be concluded that the under-identified model has a negative df, 2) Just-Identified model, where $t = s/2$; i.e., a model with the same number of estimated parameters as the known data. For example, there are two equations X + Y = 10 and X + 2Y = 16, which are 2 known pieces of data and 2 parameters to be estimated, namely X and Y, then the number of df = 2-2 = 0. So it can be concluded that the model just - identified has a zero df, 3) Over-Identified model, where $t \leq s/2$; namely, a model with a smaller number of estimated parameters than the known amount of data (Churchill, 1979). For example, there are three equations X + Y = 10, X + 2Y = 16 and 3X + 2Y = 22, which are 3 known pieces of data and 2 parameters to be estimated, namely X and Y, so the number of df = 3-2 = 1. So it can be concluded that the model is over-identified *df* positive.

Information:

t = number of parameters estimated,

s = total variance and covariance between indicators.

The goodness of fit (Assessment of fit)

Overall model fit

The first stage of the fit test is intended to evaluate, in general, the degree of fit or goodness of fit (GOF) between the data and the model. Assessing the overall fit of the model in SEM cannot be done directly as in other multivariate techniques (multiple regression, discriminant analysis, MANOVA, etc.). SEM does not have a single best statistical test that can explain the predictive power of the model (Delaney, 1996). Instead, researchers have developed several measures of GOF or goodness of fit indices (GOFI) that can be used together or in combination. This situation causes the thorough compatibility test stage, which is a step that invites a lot of debate and controversy. The controversy and debate around GOF arise when the question of size arises in the discussion, i.e., what is the acceptable level of fit? Despite the controversy, there is finally a consensus among researchers, some of which are: 1) The best guide in assessing model fit is strong substantive theory. If the model only shows or represents a substantive theory that is not strong, and even though the model has a very good fit, it is rather difficult for us to judge the model, 2) The Chi-square (X2) statistical test should not be the only basis for determining the fit of the data to the model, 3) None of the GOF or GOF Indices (GOFI) measures can be used exclusively as a basis for evaluating the fit of the entire model. GOFI is grouped into three parts, namely absolute fit measures, incremental fit measures, and parsimonious fit measures (parsimony fit size).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Absolute fit measures

The absolute fit measure determines the degree of prediction of the overall model (structural and measurement models) on the correlation and covariance matrices Deshpande, 1989; Deshpande, 1998). This measure contains measures that represent the overall fit point of view mentioned earlier. Of the various absolute fit measures, the measures usually used to evaluate SEM are 1) Chi-square (X2). A low Chi-square value (X2) results in a significance level > 0.05 or (p > 0.05), which indicates the null hypothesis is accepted. This means that the predicted input matrix is not statistically different from the actual one. Chi-square (X2) cannot be used as the sole measure of the overall fit of the model. One of the reasons is that Chi-square (X2) is sensitive to sample size. 2) Non-Centrality Parameters (NCP). Like X2, NCP is also a measure of the badness of fit where the greater the difference between Σ and Σ(θ) the more**,** the greater the NCP value. So, we need to find an NCP whose value is small or low. 3) The goodness of fit index (GFI). The GFI value ranges from 0 (poor fit) to 1 (perfect fit), and a GFI value > 0.90 is a good fit, while 0.80 < GFI < 0.90 is often referred to as marginal fit. 4) Root Mean Square Residual (RMR). RMR represents the average value of all standardized residuals and ranges from 0 to 1. A model with a good fit will have an RMR value smaller than 0.05. 5) Root Mean Square Error of Approximation (RMSEA). An RMSEA value between 0.08 and 0.10 indicates marginal fit, and an RMSEA value > 0.10 indicates poor fit. 6) Single Sample Cross-Validation Index/Expected Cross-Validation Index (ECVI). ECVI is used for model comparison, and the smaller the ECVI of a model, the better the level of fit.

Incremental fit measures

The incremental fit measure compares the proposed model to the baseline model, which is often referred to as the null model or the independence model, and the saturated model. The null model is the model with the worst fit of the model data (worst fit). A saturated model is the best fit for the model data (best fit). The concept of incremental fit will place the model-data match level between the null model and the saturated model. The level of model-data compatibility that is between the null model and the saturated model is called a nested model. Of the various incremental fit measures, the measures usually used to evaluate SEM are 1) Adjusted Goodness of Fit Index (AGFI). AGFI values range from 0 to 1, and AGFI values > 0.90 indicate a good fit. Whereas 0.80 < GFI < 0.90 is often referred to as marginal fit. 2) tucker-lewis index/non-normal fit index (TLI/NNFI). TLI values ranged from 0 to 1.0, with TLI values > 0.90 indicating good fit and 0.80 < TLI < 0.90 indicating marginal fit. 3) normed fit index (NFI). This NFI has values ranging from 0 to 1. An NFI value > 0.90 indicates a good fit, while 0.80 < NFI < 0.90 is often referred to as a marginal fit. 4) relative fit index (RFI). The RFI value will range from 0 to 1. An RFI value > 0.90 indicates a good fit, while 0.80 < RFI < 0.90 is often referred to as a marginal fit. 5) incremental fit index (IFI). IFI values will range from 0 to 1. IFI values > 0.90 indicate a good fit, while 0.80 < IFI < 0.90 is often referred to as a marginal fit. 6) comparative fit index (CFI). CFI values will range from 0 to 1. CFI values > 0.90 indicate good fit, while 0.80 < CFI < 0.90 are often referred to as marginal fit.

Parsimony fit measures

Models with relatively few parameters (and relatively many degrees of freedom) are often known as models that have high parsimony or frugality. Meanwhile, a model with many parameters (and few degrees of freedom) can be said to be a model that is complex and lacks parsimony. Of the various parsimony fit measures, the measures that are usually used to evaluate SEM are: 1) Parsimonious Normal Fit Index (PNFI). The higher the PNFI value, the better. The use of PNFI is primarily for comparisons of two or more models that have different degrees of freedom. PNFI was used to compare alternative models, and no

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

acceptable match level is recommended. However, when comparing the 2 models, the difference in the PNFI value of 0.06 to 0.09 indicates a fairly large model difference. 2) Parsimonious goodness of fit index (PGFI). PGFI values range between 0 and 1, with higher values indicating a better parsimony model. 3) Normed Chi-square. Recommended values: lower limit: 1.0, upper limit: 2.0 or 3.0, and looser 5.0. 4) Akaike information criterion (AIC). A small AIC value close to zero indicates a better fit and higher parsimony.

Measurement model fit (measurement model analysis)

Measurement model fit is carried out by testing its validity and reliability. Validity relates to whether variable measures what it is supposed to measure. Although validity can never be proven, support for such evidence can be developed. A variable is said to have good validity against its construct or latent variable if the factor loading t value (loading factors) is greater than the critical value or >1.96 or for practice >2), and standardized loading factors >0.70. Reliability is the consistency of measurement. High reliability indicates that the indicator has high consistency in measuring its latent constructs. To measure the reliability in SEM will be used: the composite reliability measure (composite reliability measure) and variance extracted measure (variance extract size). A construct has good reliability if the Construct Reliability (CR) value is > 0.70 and the variance Extracted value is (VE)> 0.50.

Structural model fit (Structural model analysis)

Analysis of the structural model includes examining the significance of the estimated coefficients. The SEM method and its software provide not only the estimated coefficients but also the t-count values for each coefficient. By specifying a significant level (usually α= 0.50), then each coefficient representing the hypothesized causal relationship can be tested for statistical significance. In addition to this, it is also necessary to evaluate the standard solution where all the beta coefficients are in multiple regression. That is the coefficient value that is close to zero indicates a smaller effect. An increase in the value of this coefficient is associated with an increase in the importance of the variable in question in a causal relationship. As an overall measure of the structural equation, the overall coefficient of determination (R2) is calculated as in multiple regression.

Respecification/modification and modeling strategy

Re-specification is the next step after the compatibility test is carried out. The implementation of respecification is highly dependent on the modeling strategy to be used. There are 3 modeling strategies that can be chosen in SEM, namely: 1) Confirmatory modeling strategy or confirmatory modeling strategy. In this modeling strategy, a single model is formulated or specified, and then empirical data is collected to test its significance. This test will result in an acceptance or rejection of the model. This strategy does not require respecification. 2) Model competition strategy or competing for modeling strategy. In this modeling strategy, several alternative models are specified, and based on an analysis of a group of empirical data, one of the most suitable models is selected. In this strategy, respecification is only needed if alternative models are developed from several existing models. 3) Model development strategy or model development strategy. In this modeling strategy, an initial model is specified, and empirical data is collected. If the initial model does not match the existing data, then the model is modified and tested again with the same data. Several models can be tested in this process with the aim of finding a model that not only fits the data well but also has the property that each parameter can be interpreted properly. Respecification of the model can be done based on theory-driven or data-driven.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

However, respecification based on theory-driven is more recommended (Freeman, 1984).

A confirmatory modeling strategy (CS) is rarely encountered because, generally, researchers are not satisfied with simply rejecting a model without proposing an alternative model. Currently, the most widely used in research is the model development strategy. After the model estimation is done, the researcher can still modify the developed model if it turns out that the resulting estimate has a large residual. However, modifications can only be made if the researcher has sufficiently strong theoretical justification because SEM is not intended to generate theory but to test models that have a correct theoretical basis, therefore providing an interpretation of whether the theory-based model being tested can be accepted directly or not (Fornell, 1981). If modifications are needed, researchers must direct their attention to the predictive power of the model by observing the number of residuals produced. If in the standardized residual covariances matrix, there are values outside the range of -2.58 < residual < 2.58 and probability (P) if < 0.05, then the estimated model needs to be further modified based on the modified index by selecting the modification index (MI) is the largest and has a theoretical basis. The largest MI will give an indication that if the coefficient is estimated, there will be a significant reduction in the chi-square (X2) value (Grinstein, 2008). In SEM software, the modification index is included in the output so that the researcher only has to choose which coefficient to estimate. If the chi-square value (X2) is still not significant, look for the next largest MI value and so on.

2. Conclusion

The process of data analysis using SEM starts with model specification, model identification, model suitability testing, and respecification/modification and modeling strategy. modifications can only be made if the researcher has a sufficiently strong theoretical justification because SEM is not intended to generate theory but to test models that have a correct theoretical basis, therefore providing an interpretation of whether the theory-based model being tested can be directly accepted or needs modification. Then the researcher must direct his attention to the predictive power of the model by observing the amount of residual generated.

3. References

Astrachan CB, Patel VK, Wanzenried G. 2014. A comparative study of CB-SEM and PLS-SEM for theory development in family firm research. Journal of Family Business Strategy. 5; 116-28.

Babin BJ, Boles JS, Robin DP. 2000. Representing the perceived ethical work climate among marketing employees. Journal of the Academy of Marketing Science. 28(3): 345-58.

Babin BJ, Hair JF, Boles JS. 2008. Publishing research in marketing journals using structural equation modeling. Journal of Marketing Theory & Practice. 16(4): 279-285.

Bagozzi R, Yi Y. 2012. Specification, evaluation, and interpretation of structural equation models. Journal of the Academy of Marketing Science, 40(1): 8-34.

Blocker C, Flint D, Myers M, Slater S. 2011. Proactive customer orientation and its role for creating customer value in global markets. Academy of Marketing Science Journal. 39(2): 216-33.

Baker WE, Sinkula JM. 1999. Learning orientation, market orientation, and innovation: Integrating and extending models of organizational performance. Journal of Market-Focused Management. 4(4): 295-308.

Bido DDS, Souza CAD, Silva DD, Godoy AS, Torres RR. 2012. Quality of reporting methodological procedures in national publications in the area of business administration: the case of structural equation modelling. Organizações & Sociedade, 19(60): 125-44.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Brei VA, Liberali NG. 2006. O uso da técnica de modelagem em equações estruturais na área de marketing: um estudo comparativo entre publicações no Brasil e no exterior. Revista de Administração Contemporânea. 10(4): 131-51.

Byrne BM. 2010. Structural equation modeling with AMOS: Basic concepts, applications, and programming, 2nd ed. New York: Routledge.

Chin WW, Peterson RA, Brown SP. 2008. Structural equation modeling in marketing: some practical reminders. Journal of Marketing Theory & Practice, 16(4): 287-9.

Churchill GAJr. 1979. A paradigm for developing better measures of marketing constructs. JMR, Journal of Marketing Research, 16(1): 64-73.

DeConinck JB. 2010b. The influence of ethical climate on marketing employees' job attitudes and behaviors. Journal of Business Research, 63(4): 384-91.

Delaney JT, Huselid MA. 1996. The impact of human resource management practices on perceptions of organizational performance. Academy of Management Journal, 39(4): 949-69.

DeVellis RF. 2011. Scale development: Theory and applications. Sage Publications, Inc. 26.

Deshpande R, Farley J. 1998. Measuring market orientation: Generalization and synthesis. Journal of Market-Focused Management, 2(3): 213-32.

Deshpande R, Webster JrFE. 1989. Organizational culture and marketing: defining the research agenda. The Journal of Marketing. 3-15.

Dillman DA, Smyth JD, Christian LM. 2009. Internet, mail, and mixed-mode surveys: The tailored design method. 3rd ed. Hoboken: John Wiley & Sons Inc.

Fabrigar LR, Porter RD, Norris ME. 2010. Some things you should know about structural equation modeling but never thought to ask. Journal of Consumer Psychology. 20(2): 221-5.

Ferrell O, Gonzalez-Padron T, Hult G, Maignan I. 2010. From market orientation to stakeholder orientation. Journal of Public Policy & Marketing, 29(1): 93-6.

Freeman R. 1984. Strategic management: A stakeholder approach: Pitman Boston, MA.

Fornell C, Larcker DF. 1981. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research. 18(1): 39–50.

Gallagher D, Ting L, Palmer A. 2008. A journey into the unknown; taking the fear out of structural equation modeling with AMOS for the first-time user. The Marketing Review. 8(3): 255-75.

Grinstein A. 2008. The effect of market orientation and its components on innovation consequences: A meta-analysis. Journal of the Academy of Marketing Science. 36(2): 166-73.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License