Strategy Structure and Performance – Factor Analysis, Path Analysis and Structural Equation Modelling

***Yoginder Singh Kataria**

*****Associate Professor, PIET – Panipat Institute of Engineering & Technology, Pattikalayana, Samalkha – 132102 (Panipat) Haryana.

Abstract:

Structural Equation Modelling (SEM) is gaining importance as an important multivariate technique for analysis as it provides for advantages compared to other traditional techniques. It has evolved in to a mature and popular methodology to investigate theory driven structural and causal hypothesis. Practitioners have little formal SEM background leading to misapplication. The article focuses on basics of Factor Analysis, Path Analysis and Structural Equation Modelling (SEM). Structural Equation Modelling (SEM)

Background – Multiple Regression and Factor Analysis are good statistical methods, but they don't go far in analysis due to their limitation of use in social sciences. In multiple regressions, a variable can be either a predictor (an independent variable) or an outcome (a dependent variable). But however in real life a variable may be an outcome with respect to some variables but may in turn become a predictor of other variables. It is very cumbersome to use Multiple Regression. Path analysis, an extension of multiple regression, Structural equation modelling extends path analysis by looking at latent variables. These statistical methods allow us to deal with more than one dependent variable simultaneously and allows for variables to be dependent with respect to some variables and independent with respect to others.

AIM - The primary aim of article is to provide better clarity and understanding on Factor Analysis (FA), Path Analysis (PA) and Structural Equation Modelling (SEM) theory to enhance and encourage its use in social sciences.

Design / Methodology / Approach – Factor Analysis (FA), Path Analysis (PA) and Structural Equation Modelling (SEM) theory is reviewed from national and international research stream. The paper attempts to provide better understanding to Factor Analysis, Path Analysis and Structural Equation Modelling (SEM) and aspects related to operationalizing these methods. Proposition for use of Structural Equation Modelling (SEM) method for research in Indian context based upon this enhanced understanding.

Findings – Structural Equation Modelling (SEM) besides its limitation and complexity offers tremendous scope in researches in social sciences for its ability to handle complex multivariate analysis.

Research Limitation- Path Analysis (PA) and Structural Equation Modelling (SEM) theory is relatively new in use and often most complex to use. Literature review is very broad on this concept leading to varying interpretation and use. There are very few studies available in Indian context.

Keywords – Factor Analysis (FA), Organizational Performance, Path Analysis (PA), Strategy, Structure, Structural Equation Modelling (SEM).

Introduction

Factor Analysis, Path Analysis and Structural Equation Modelling (SEM) are used when testing theoretical models and constructs along with measuring interactions between independent variables. This article explores Factor Analysis, Path Analysis and Structural Equation Modelling (SEM) multivariate statistical techniques for measuring and analysing complex theoretical models and distinct advantages it provides against multiple regression analysis methods.

The following flow chart explains the logical steps on deciding to conduct Structural Equation Modelling (SEM):

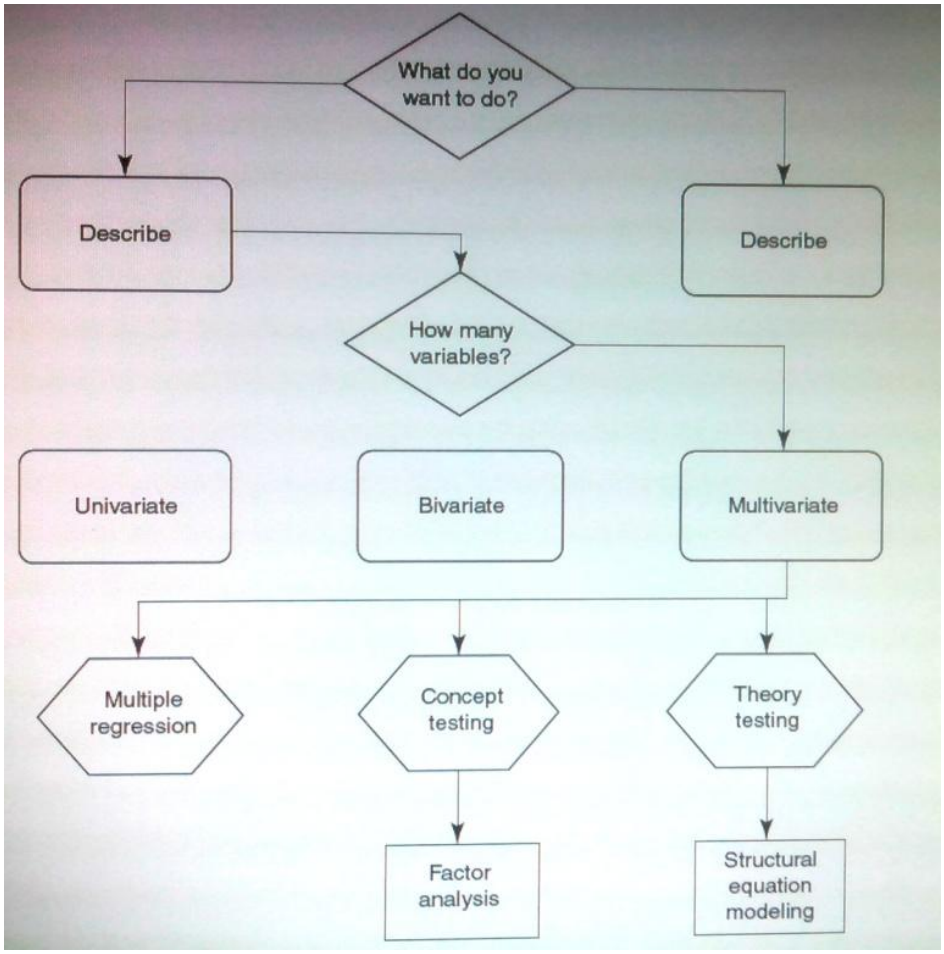

Figure: 1: Steps in selecting SEM

Factor Analysis:

It is a multivariate analysis procedure aims to identify underlying "factors" responsible for co-variation amongst group independent variables and attempts to reduce the number of variables used to explain a relationship. A factor is linear combination of group of variables (items) combined to represent a scale measure of construct. To successfully use factor analysis variables must represent indicators of some common underlying dimension so that they can be grouped together.

There are two types of factor analysis:

1. **Exploratory Factor Analysis** is used to explore the loading of variables to arrive at best model. Variables are grouped together in a model where they are expected and then we see how the factor analysis groups them.

2. **Confirmatory Factor Analysis** is more rigorous and is used to confirm the previously defined hypothesis concerning relationship between variables.

Generally the combination of two methods of factor analysis is used as the researcher has some idea about where the variables are going to load and how to use factor analysis to support the hypothesis by accepting some minor modifications in terms of

grouping. Thus factor analysis is a data reduction procedure as it places a number of variables in a model and determining which variables are to be removed from the model to make it more parsimonious. Factor analysis can be used to check multi co linearity as variables that group together and have high factor loading are multi collinear.

Factor analysis as a multivariate analysis has two concepts: variance and factoral complexity. There are three components to variance concept: communality, uniqueness and error variance.

Communality is the part of variance shared with one or more other variables, represented by sum of the squared loading for a variable (across factors). Factor analysis attempts to explain maximum variance possible with the least variables (parsimony).

Uniqueness is the variance specific to a particular variable. It measures the variance reflected in single variable alone. It is assumed to be uncorrelated with the component factors or other unique factors.

Error variance is the variance due to random or systematic error in the model.

Factoral complexity is the number of variables loading on a given factor. Ideally a variable should load only one factor, which confirms theoretically that you have accurately determined conceptually how the variables will group. This confirms that you have an accurate measure of the underlying dimension. Variables that load or cross load on more than one factor represents complex model and is more difficult to determine the true relationships between variables, factors, and the underlying dimension.

Assumptions in Factor Analysis

Basic assumption of factor analysis is that the data is interval and normally distributed (linear). The second assumption is that there is no specification error in the model, which refers to exclusion of relevant variable from the analysis or conversely inclusion of irrelevant variable in the model. This problem though can only be rectified during conceptual stages of research planning. The third assumption is that sample size is sufficient to conduct base analyses. Hatcher (1994) suggested the sample size to be at least 5 times the number of variables in a model.

While multi Co linearity is problematic in regression analysis, in factor analysis it is necessary because variables must be highly co associated with other variables to load in to factors, with caveat that all the variables should not be highly correlated with other group of variables.

Steps Involved in Factor Analysis and Interpretation:

STEP1: EXAMINE UNIVARIATE ANALYSIS: measures of skewness and kurtosis. If distribution is skewed or kurtosis it may not be normally distributed and / or non linear.

STEP 2: PRELIMINARY ANALYSES: scanning and examining to see data if it is appropriate for the factor analysis.

a) Bartlett's test of sphericity to determine if the correlation matrix in the factor analysis is Identity Matrix (where diagonals are 1 and off diagonals are 0). This would mean that none of the variables are correlated with each other. If the Bartlett's test is non significant than the factor analysis is not to be used because the variables will not load together properly.

b) Anti-Image correlation matrix shows there is low degree of correlation between the variables when the other variables are held constant. Anti-image means that low

correlation values will produce large numbers. Diagonal values are large and off diagoal values should be small. If there are large values in off diagonals then factor analysis should not be done.

c) Kaiser-Meyer-Olin measure of sampling adequacy determines if the sampling is adequate for analysis. The KMO compares the observed correlation coefficients to the partial correlation coefficients. Small values of KMO indicate problems with sampling. KMO of 0.90 is best and below 0.50 is unacceptable and you need to look in to individual measures that are located on the diagonal in the anti-image matrix to see what variables must be bringing KMO down.

STEP 3: Extract Factors:

Principal component analysis is most common methods of extracting factors, others competing methods include Maximum Likelihood. These analyses determine how well the factor explains the variation. It helps in identifying the linear combination of variables that account for greatest amount of common variance. Factors with Eigen value of more than 1 represents the number of factors needed to describe the underlying dimensions of the data. Each of the factors at this point is not correlated with each other (orthogonal). Eigen value represents the strength of a factor.

The scree plot is the graphical representation of the incremental variance accounted for by each factor in the model. Factors that are in level part of scree plot may need to be excluded from the model. Excluding variables should be supported both theoretically and methodologically.

At this point and after rotation you should examine factor matrix (component matrix) to determine what variables to be combined (those that load together) and if any variables should be dropped. This can be done through factor loading value. Any score more than 0.40 in the factor matrix indicate that a variable is closely associated with the factor.

STEP 4: Factor Rotation

The variables are related to factors seemingly at random. It is difficult to determine clearly which variables load together. Examining factor matrix values of more than 0.4 can be included in a factor. For those variables that reach the cut-off, it is now possible to determine which variables load (group) with other variables. During Exploratory factor analysis, this is where a determination can be made concerning which variable to combine with scales or factors. During the confirmatory factor analysis, this step determines how well the theoretical model faired under testing.

There are two main categories of rotation: orthogonal and oblique. In orthogonal category most commonly used rotation is varimax, as this rotation attempts to minimize the number of variables that have high loading on a factor.

Oblique rotation used in spss is oblimin. The closer the factors are clustered, especially if there are only two clusters' of factors, the better an oblique rotation will identify a model but it is not suited for model with more than three clusters of factors due to increased complexity.

STEP 5: Use of Factor in other Analysis

After completing factor analysis, the factors can be used in other analysis by saving the factors as variables; therefore the factor values become the values of that variable. Due care must be observed while using factored variables in regression due to their higher R2 than might be expected in the model, because each factor is scale measure in the underlying dimension.

Factor analysis has largely fallen to disuse due to statistical advancement in other tools. Original rival of factor analysis, path analysis has given way to now Structural Equation Modelling (SEM).

Structural Equation Modelling (SEM)

Structural equation modelling (SEM) is a multi equation technique in which there can be multiple variables, best suited for social sciences. While in all forms of multiple regressions we used single formula y = a+bx+e with one dependant variable. Multiple equation system allows for analysis of multiple indicators for the concepts; which requires matrix algebra. But with advancement of technology and software, these calculations of complex measures are handled by statistical programs like AMOS or LISREL.

In this article we will try to set up a basic SEM model and understand some key concepts of SEM. Structural equation modelling (SEM) consists of two primary models : measurements (Null) and structural models. Measurement model pertains to how the observed variables relate to unobserved variables.

In general researches there are confounding variables that are not included in model or accounted for in the analysis, but which have an effect on outcome. Major advantage of using Structural equation modelling (SEM) is that it can deal with how concepts relate to one another alongside attempting to account for these confounding variables. For example combination of Strategy and Structure may be enhancing organizational effectiveness but it may be confounding variables like strategy – structure alignment, knowledge management or culture that may be making this relationship. SEM addresses this complex relationship by creating a theoretical model of the relationship and then testing with data to see if the data supports it. So, essentially SEM is a confirmatory procedure where a model is proposed, theoretical diagram is generated and examination of data is done to see how close the data supports the model. We want the two models to be not statistically significantly different.

Variables in SEM:

In SEM we do not use the terms dependant and independent variables unlike other research procedures. There are two forms of variables: Exogenous and Endogenous. Exogenous variables are always analogous to independent variables, they have path coming from it and none leading to it. Endogenous variables are at some point in model dependant variable and at other point may be independent variables, so they have at least one path leading to it.

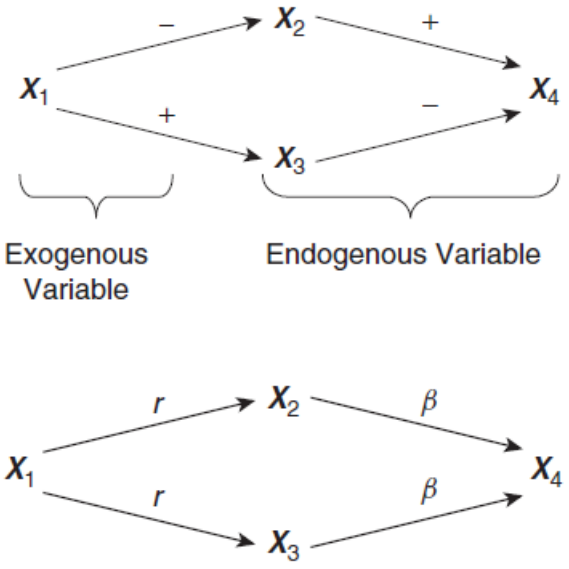

Figure 2: Showing exogenous and endogenous variables

A factor in factor analysis is called latent variable in SEM and used by an oval. Measured variables are the one that are measured directly. Arrows come from latent variable to measured variable, reflecting that measured variable are due to or the result of latent variable.

Sample Size

In PA and SEM adequacy of sample size is very important. Though there is no universal suggestion on sample size but it is recommended to have 10 subjects per parameter (not per variable). Though caution is advised here that too large sample size will result in X2 gof may be statistically significant, even if model relatively is tested to be fit.

Assumptions in SEM

Five assumptions must hold good while conducting SEM, though similar to the assumptions of multivariate analyses. First, the relationship between coefficients and error term must be linear. Second the residual must have a mean of 0 (zero), be independent, normally distributed and have uniform variances across the variable. Third variables must be continuous, interval level data. Fourth, SEM has no specification error (necessary variables are not omitted and unnecessary variables are not included in model) and finally, variables included in model must have kurtosis of an acceptable level. Therefore, an examination of univariate statistics of variables will be crucial for completing SEM. Any variables with kurtoses outside the acceptable level will result in inaccurate calculations. This often limits the data in social sciences. Generally scaled data and dichotomous data are not used in SEM.

Advantages of SEM

1. SEM allows for identifying direct and indirect effects. Principally we look for direct effects in research where in dependant variable is explained in terms of changes in independent variable by establishing cause and effect relationship.

So in direct effect an arrow goes from one variable only into another variable for example:

Strategy ⟶ Structure

An indirect effect occurs when one variable goes through another variable on the way to dependant or independent variable. For example:

Strategy / Structure → Strategy – Structure Alignment → Performance

Complicating SEM models may have both direct and indirect effects simultaneously.

A curved line indicates the covariance or correlation between pairs of error terms or the correlation between a pair of exogenous variables. A curved line indicates that a causal interpretation is not invited for the relationship.

2. It is possible to have multiple indicators of a concept. In multiple regression analysis we evaluate only the effect of individual independent variable on the dependant variable in a simple path model. SEM allows for estimation of combined effects of independent variables in to concepts / constructs as shown in figure 3 below. It can be seen that variables are no longer acting alone but in concert and combination with other variables that are conceptually adding to the prediction of dependant variable.

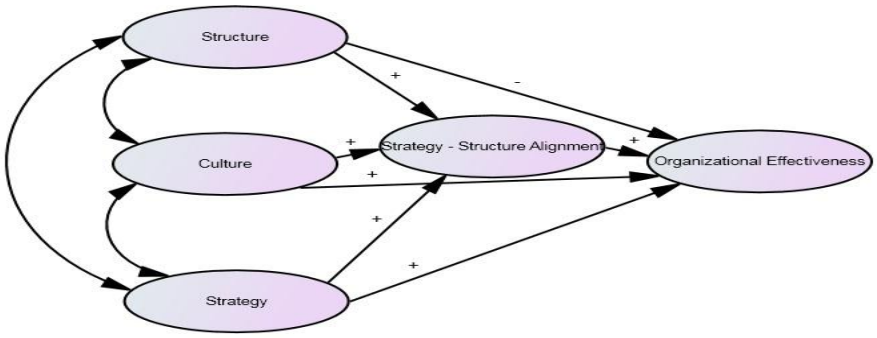

Figure 3: Model showing multiple indicators of concept

3. SEM includes / accommodates measurement error terms in to the model, which is not possible in regression analysis. While path analysis using regression include error terms for its prediction but they do not adequately control for measurement error. SEM accounts for measurement error thereby adding to our understanding on how good the theoretical model predicts the actual behaviour (see figure 4).

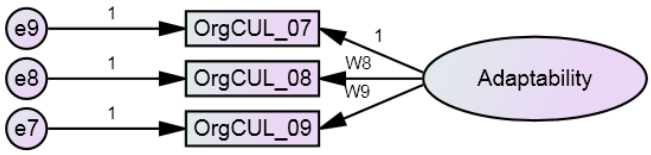

Figure 4: Model showing error terms and numbers over the path are standardised path co-efficient.

However it is to be understood with Path Analysis (PA) that if variable A has an arrow pointing towards variable B and B points to C it does not necessarily mean that A causes B and B causes C. Also it is important to understand here that PA is not to develop theoretical models. PA and SEM are model – testing methods, not model developing. We develop our models based upon our theory, knowledge or hunches.

Types of Path Models

1. **SIMPLE MODEL:** there are two exogenous variable X1 and X2 and one endogenous variable Y (see model A in figure 5)

2. **MEDIATED MODEL:** Y modifies the effect of X on Z (see model B in figure 5).

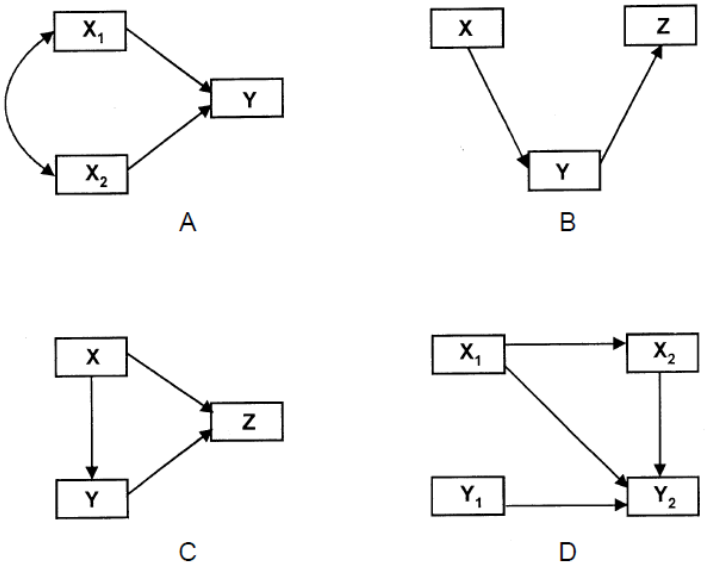

Figure 5: Some examples of Path Models

3. **COMPLEX MODEL:** It combines elements of previous two models. Variable X has direct effect on variable Z but also acts on variable Y which in turn effects Z (see model C in figure 5)

4. **MULTIPLE PATH MODEL:** There is no path between Y1 and X1 and Y1 to X2 or Y2 to X2. We could use this model to look at different variables at the same time. (see model D in figure 5)

Types of Model

1. Recursive Model: Arrows / paths goes in one direction only e.g. $A \rightarrow B$. There must not be any feedback loop in endogenous variable. $(\text{no } A \rightleftarrows B)$ and no reciprocal pattern between pairs of variables $(\text{no } A \longleftrightarrow B)$.

2. Non Recursive Model: where the path goes forward and backward. These models are more difficult to ascertain. To fit a non recursive model to the data, "ordinary least squares regression will provide good estimates of the parameters when the necessary assumptions are made about the properties of the residual terms" (Asher, 1983, p. 15).

Steps in Sem Construction

1. Model specification

2. Model identification

3. Estimation

4. Test of fit

5. Respecification

Model Specification

Model specification is also called measurement model, consists of specifying relationship between latent variables and determining how latent variables will be measured. It is most important and crucial step as everything else follows from here. It is important to specify model correctly, choose right measured variable to reflect the latent variables. The model specification takes place with your knowledge in the field or by supporting it through literature review.

Model Identification

Check for whether a unique value for each and every free parameter can be obtained from the observed data or not. Important thing to remember here is that number of parameters cannot exceed number of observations. To solve the equation one path from the latent variable to the measured variable must be fixed. Conventionally we fix t to 1. Similarly we fix the path from error (disturbance) terms to the exogenous variable to 1. We can even fix the error variance terms but very rarely we know the value in advance, hence we fix the path. We can also limit the number of parameters by constraining some variables. The most common way is to constrain the error/ disturbance term of variables that are expected to have similar variances.

At this point our model looks something like below in figure 6:

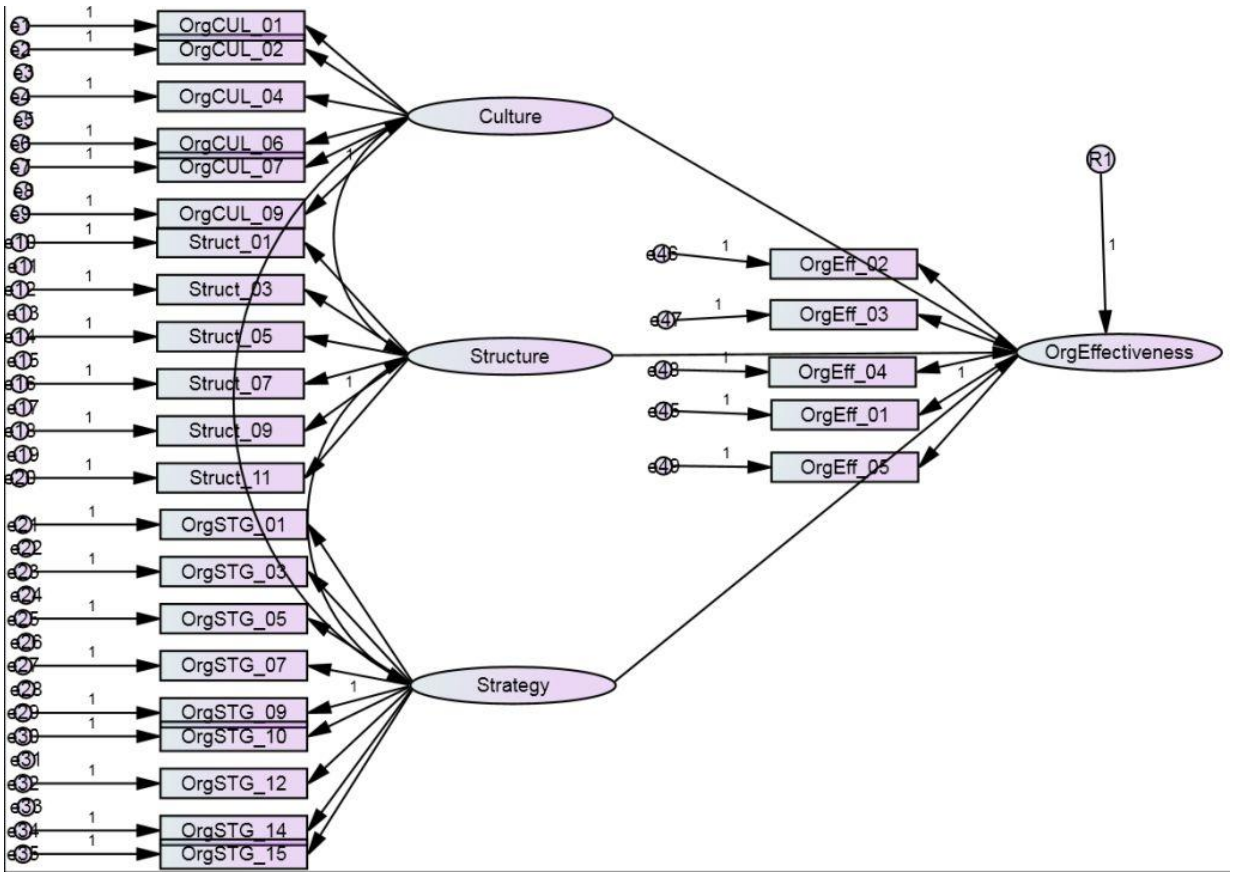

Figure 6: Showing complete model

Models need to be over-identified in order to be estimated (Step 3 in SEM construction) and in order to test hypotheses about relationships among variables. A necessary condition for over identification is that the number of data points (number of variances

and co - variances) is less than the number of observed variables in the model. Figure 7 shows, a flow-chart (extracted from Ullman 1996) summarizing the procedures of model identification:

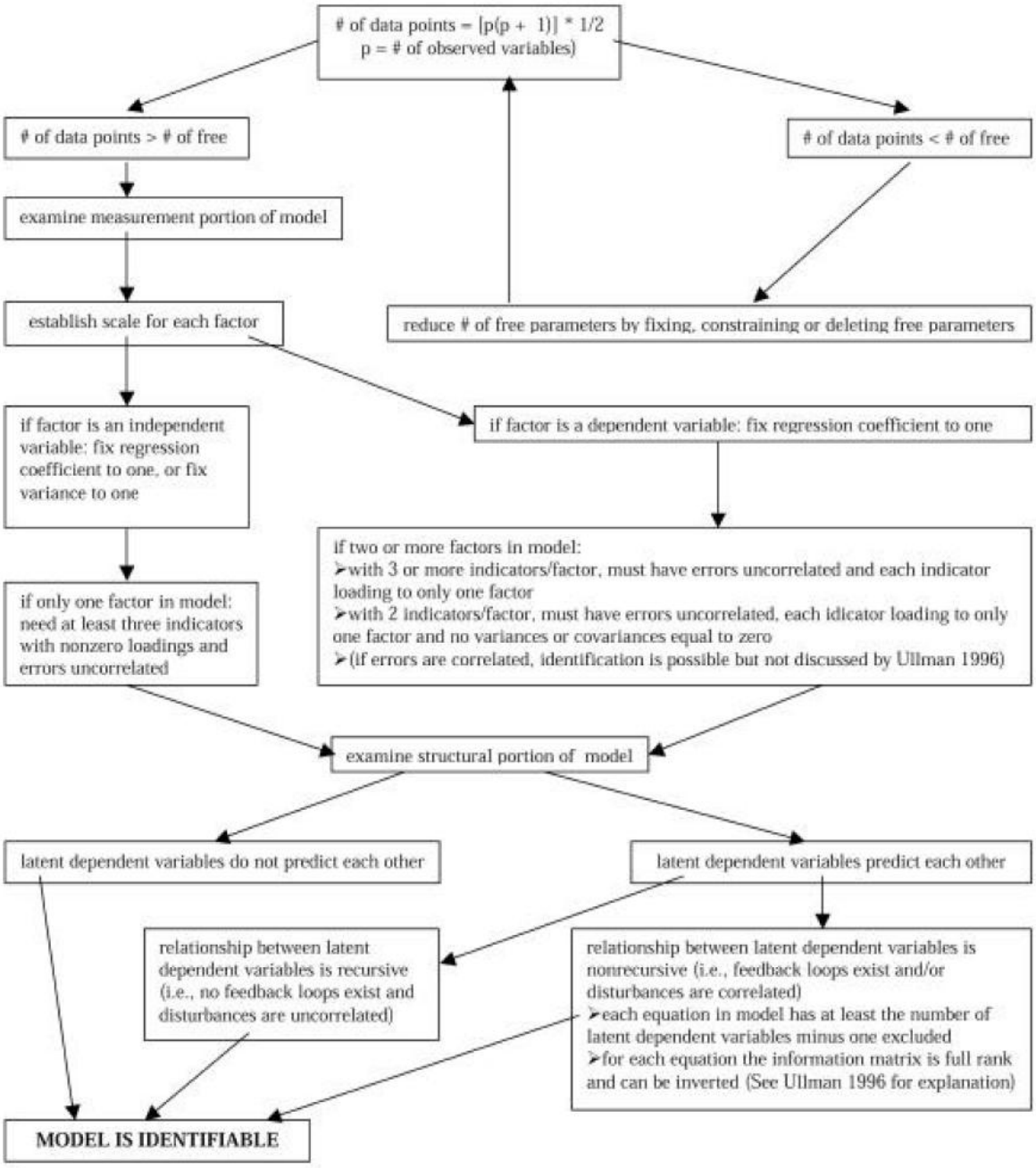

Figure 7: Flowchart to determine Model Identifiability

Estimation

Now that we have specified the model and fixed or constrained some of the path or variance we are ready to estimate the parameters. There are various methods but **unweighted least squares** approach is the most advantageous as it does not make any assumptions about underlying distribution of variables but has major drawback that its results are dependent on the scale of measurement, so this method is rarely used. Weighted least square approach and retains the desired property of remaining

distribution free, but not suited for large sample sizes which are generally the case in SEM. Therefore by default **Maximum likelihood** is the most used method for estimation. It is not dependant on scale of measurement but it requires multivariate normality (i.e. all variables are normal when looked at collectively), makes it unsuitable for ordinal measures and highly skewed data set.

Test of Fit

During Path analysis we looked at signs of coefficients, determined if they are statistically significant and used the **X2 gof** to look at the model as whole. X2 gof is sensitive to the normality of data and sample size. Too many subjects it is always significant and too less it never is. But major advantage of X2 gof against all other tests is that it the only one which has test of significance attached to it. As a thumb rule we want X2 gof to be non significant and X2 gof/df to be less than 2.

The other tests can be clubbed in to two: **Comparative fit** and **variance explained** all are scaled to be between 0 and 1 with larger numbers indicating better fit with a minimum criterion of 0.90. The most common comparative fit is **Normed fit index (NFI)** which test the model differs from null hypothesis and none of the variables are related. It faces limitation that it shows better results when you add more parameters in to model. To overcome this limitation we have **Normed fit Index 2 (NFI2).**

Most common test for variance explained is **Goodness of fit index (GFI)** and its variant **Adjusted goodness of fit index (AGFI),** which adjusts number of parameters (fewer the better) and sample size (more the better). Another test **Akaike's Information Criterion (AIC**) it's not scaled between 0 and 1. It is helpful in comparing models; the one with the lower value has a better fit with the data.

So if, NFI 2 and AGFI. If they are both over 0.90, and the x2GoF is not significant, your model pass the test of fit. If the indices are above 0.90 but the x2 GoF is significant, look to see if your sample size may be "too large". If only some of the indices are over 0.90 and the x2 GoF is significant, you've got problems and your model needs re-specification.

Respecification

Here you need to improve your model for better fit. There are some tests available that will help you in respecification by telling which path or variable are unnecessary or there is need to add path or covariances. Many a times misfit is due to omission of variable in to model in PA and SEM.The **Lagrange multiplier tests** tell you how much the model could be improved if fixed or constrained variables were allowed to be free (in other words, if parameters were added to the model). Conversely, the **Wald test** shows the effects of deleting parameters from the model, such as by removing paths or constraining variables.

It is important to bear in mind, whether or not to free a parameter or to add or drop a covariance term must be based on your theory and knowledge.

Conclusion

There are two perspectives to SEM: Measurement model and the model as a whole. Measurement model pertains to each latent variable and its associated measured variable and seeks to answer how well the measured variable reflect the latent variable, are some observed variables better and how reliable is measured variable. Each latent variable is mini factor analysis by itself. We can go back and remove variables that don't seem to be adding to model fit but instead adding to error variance. Once we arrive at good measured variable we can look at model to see how well it fits the data. Are there some latent variables that don't have significant paths to others or are having

wrong signs? PA and SEM though offer great advantage over other statistical methods but it is very difficult to administer them.

References

Benter, P. M. & Chou, C. P. (1987). Practical issues in structural modelling. *Sociological Methods and Research, 16(1)*, 78-117.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley & Sons.

Bollen, K. A., & Long, J. S. (1993). Testing structural equation models. Thousand Oaks, CA: Sage.

Byrne, B.M. (2001), *Structural equation modelling with AMOS: Basic concepts, application and programming.* New Jersey: Lawrence Erlbaum Associates, Publishers.

Caroll, J.B. (1953). Approximating simple structure in Factor Analysis. *Psychometrika,* 18:23-38.

Duncon, O.D. (1966). Path Analysis: Sociological Examples, *America Journal of Sociology* 73: 1-16.

Everitt, B.S. and G. Dunn. (1991). *Applied Multivariate Data Analysis*. Halsted Press. New York, NY. pp. 257-275.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modelling.* Cary, NC: SAS Institute.

Hayduk. L.A. (1987). *Structural Equation Modelling with LISREL: Essentials and Advances.* Baltimore: The John Hopkins University Press.

Hotelling, H. (1933). Analysis of Complex of Statistical Variables in to Principal Components. Journal of Educational Psychology, 24: 417.

Hoyle, R. (1995). *Structural equation modelling: concepts, issues and applications.* Thousand Oaks, CA: Sage Publications.

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

Johnson, R.A., and D.W. Wichern. (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall, Inc. Englewood Cliffs, NJ. pp. 326-333.

Kaiser, H.F. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 23: 187-200.

Kaiser, H.F. (1974a). An Index of Factoral Simplicity Analysis. *Psychometrika*, 39: 31-36.

Kaiser, H.F. (1974b). A Note on Equamax Criterion. *Multivariate Behavioural Research*, 9: 501-503.

Kelloway, E.K. (1998). *Using LISREL for Structural Equation Modelling.* SAGE Publications, Inc. Thousand Oaks, CA. Ch 6, Ch 7.

Kim, J.O., and Mueller, C.W. (1978a). *Introduction to Factor Analysis*. London: Sage Publications.

Kim, J.O., and Mueller, C.W. (1978b). *Factor Analysis: Statistical Methods and Practical Issues*. London: Sage Publications.

Kline, R. B. (2004). *Principles and practices of structural equation.* New York: Guilford Press.

Lawly, D.N. (1940). The Estimation of Factor Loading by the Method of Maximum Likelihood. *Proceedings of Royale Society of Edinburgh*, 60: 64-82.

Little, R. J. A. & Rubin, D. A. (1987). *Statistical analysis with missing data.* New York NY: John Wiley & Sons.

Loehlin, J. C. (1992). *Latent variable models.* Hillsdale, NJ: Lawrence Erlbaum Publishers.

Long, L.S. (1983). *Confirmatory Factor Analysis*, London: Sage Publications.

Maruyama, G. M. (1998). *Basics of structural equation modelling.* Thousand Oaks, CA: Sage.

Mitchell, R.J. (1993). Path analysis: pollination. In: *Design and Analysis of Ecological Experiments* (Scheiner, S.M. and Gurevitch, J., Eds.). Chapman and Hall, Inc. New York, NY. pp. 211-231.

Pearson, K. (1901). On Lines and Planes of Closest Fit to System of Points in Space. *Philosophical Magazine*, 6(2):559-572.

Rigdon, E. (1997). *Approaches to testing identification.* http://www.gsu.edu/~mkteer/identifi.html

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537-560.

Schumacker, R.E. and R.G. Lomax. (1996). *A Beginner's Guide to Structural Equation Modelling.* Lawrence Erlbaum Associates, Inc. Mahwah, NJ.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. American Journal of Psychology, 15: 201-293.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Lawrence Erlbaum Publishers.

Thurstone, L.L. (1931). Multiple Factor Analysis, *Psychological Review*, 38 :406-427.

Thurstone, L.L. (1947). *Multiple Factor Analysis.* Chicago: University of Chicago Press.

Ullman. J.B. (1996). Structural equation modelling (In: *Using Multivariate Statistics*, Third Edition, B.G. Tabachnick and L.S. Fidell, Eds.). HarperCollins College Publishers. New York, NY. pp. 709-819.

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D.R. Heise (Eds.): *Sociological Methodology.* San Fransisco: Jossey-Bass.

Wright, S. (1934). The Method of Path Coefficients. *Annals of Mathematical Statistics*, 5 : 161-215.