CHAPTER 5

STRUCTURAL EQUATION MODELING APPROACH

5.1 INTRODUCTION

Structural Equation Modeling (SEM) has become an active area of research in various disciplines; research studies such as Lamb et al., [2011], Aburatani [2011], Beran and Violato [2010], Mayo et al., [2015], Xiong et al., [2015] extend the application of SEM in the areas of plant sciences, gene regulation and system biology, medical research, health economics and construction engineering. This indicative literature and other similar recent studies have shown the impact of SEM beyond social science, which has witnessed an extensive application of SEM for so many decades. Meyer and Smith [2000], Cheng [2001], Holbert and Stephenson [2003], Salanova et al., [2005] and references there on provide sufficient illustrations in the previous decade. Recent studies such as Guay et al., [2015] in experimental education research can be considered for understanding the importance of SEM in many branches of social science.

The extensive use of SEM mainly stems from the reason that it provides a convenient platform for the researchers to formulate, test and validate their research models and / or hypothesis. Though there are other statistical techniques available for such scientific endeavor, SEM has been considered as more useful and unified approaches in many study environments. These kind of observations are not only prevalent among the researchers / practitioners, but such messages and comparisons are evidently available in text books [Norman and Striener, 2003; Kline, 2015] that could extend in classroom practices. In particular Kline [2015] is an excellent
resource for understanding the history of SEM as well as the areas of application together with computing platform.

The advent of computing power can also be considered as one of reasons for an extensive practice of SEM in wide spread applications. LISREL, M-plus, EQS, SEM R packages are few computing platform for modern day theoretical and/or applied researchers. Narayanan [2012] provides a review of eight software for SEM including three R packages which are distributed as open source software packages.

Another aspect that helps to align researchers is to consolidate the results from variety of SEM software packages. Though software differs the way of presenting results, objective of most or all the researchers can be unified. In this view, many studies have indicated important quantities to extract from a software and to decipher the required information pertaining SEM study [Boomsma, 2000; McDonald and Ho, 2002; Schreiber, 2008]. These and similar other studies indicate the most important statistical measure to report understand and to use for interpretation of results. This includes goodness-of-fit tests, model fit indices, path coefficients (similar to regression models) with standard errors and statistical significance such as p-values. In most of SEM models these practices are uniformly followed so that understanding between the researchers and the audience or practitioners are practically visible from the reports of a study, methodical and/or practical applications.

This section first quickly reviews the important definitions, notations, and symbols followed in most of the present day SEM work. This exercise can be done easily through a hypothetical path diagram to show the number of latent variables with respective indicators, single or double sided arrows, and error terms.
Beyond such diagram for a SEM model it is highly necessary to understand the mathematical, statistical inferential and computational aspects for any modeler. This is usually referred as measurement or structural model and associated estimates and interpretation there on. Table 5.1 helps to complete these three phases with the usual standard notations adopted in SEM practices. It represents the pronunciation and
understanding of SEM abbreviations where terminology is abbreviated with the combination of Roman and Greek characters.

Table 5.1: Structural equation modeling abbreviations.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Pronunciation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \xi )</td>
<td>Xi (KZI)</td>
<td>A construct associated with measured ( X ) variables</td>
</tr>
<tr>
<td>( \lambda_X )</td>
<td>lambda &quot;X&quot;</td>
<td>A path representing the factor loading between a latent construct and a measured ( X ) variable</td>
</tr>
<tr>
<td>( \lambda_Y )</td>
<td>lambda &quot;y&quot;</td>
<td>A path representing the factor loading between a latent construct and a measured ( Y ) variable</td>
</tr>
<tr>
<td>( \Lambda )</td>
<td>capital lambda</td>
<td>A way of referring to a set of loading estimates represented in a matrix where rows represent measured variables and columns represent latent constructs</td>
</tr>
<tr>
<td>( \eta )</td>
<td>eta(&quot;eight-ta&quot;)</td>
<td>A construct associated with measured ( Y ) variables</td>
</tr>
<tr>
<td>( \phi )</td>
<td>phi(( \phi ))</td>
<td>A path represented by an arced two-headed arrow representing the co-variation between one ( \xi ) and another ( \xi )</td>
</tr>
<tr>
<td>( \Phi )</td>
<td>capital phi</td>
<td>A way of referring to the covariance or correlation matrix between a set of ( \xi ) constructs</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>gamma</td>
<td>A path representing a causal relationship (regression coefficient) from a ( \xi ) to an ( \eta )</td>
</tr>
<tr>
<td>( \Gamma )</td>
<td>capital gamma</td>
<td>A way of referring to the entire set of ( \gamma ) relationships for a given model</td>
</tr>
<tr>
<td>( \beta )</td>
<td>beta</td>
<td>A path representing a causal relationship (regression coefficient) from one ( \eta ) construct to another ( \eta ) construct</td>
</tr>
<tr>
<td>( \ Beta )</td>
<td>Capital beta</td>
<td>A way of referring to the entire set of ( \beta ) relationships for a given model</td>
</tr>
<tr>
<td>( \delta )</td>
<td>delta</td>
<td>The error term associated with an estimated, measured ( X ) variable</td>
</tr>
<tr>
<td>( \theta_0 )</td>
<td>theta</td>
<td>A way of referring to the residual variances and covariances associated with the ( X ) estimates; the error variance items are the diagonal</td>
</tr>
<tr>
<td>( \varepsilon )</td>
<td>epsilon</td>
<td>The error term associated with an estimated, measured ( Y ) variable</td>
</tr>
<tr>
<td>( \theta_\varepsilon )</td>
<td>theta-epsilon</td>
<td>A way of referring to the residual variances and covariance associated with the ( Y ) estimates; the error variance items are the diagonal</td>
</tr>
<tr>
<td>( \xi )</td>
<td>zeta</td>
<td>A way of capturing the co-variation between ( \eta ) construct errors</td>
</tr>
<tr>
<td>( \tau )</td>
<td>tau</td>
<td>The intercept terms for estimating a measured variable</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>chi(k)squared</td>
<td>The likelihood ratio</td>
</tr>
</tbody>
</table>
The main concern of the present work is to clarify the statistical aspects of mediation analysis; this notion in SEM is interestingly provide a scope to understand the present day computations and interpretation of mediated effect. Because SEM, apart from a model testing mechanism, it can extended to derive mediation analyses of hypotheses of interest. Hence, this chapter has concerned a threefold objectives; making a clear mathematical description of SEM models so as to calculation of degrees of freedom in the statistical inference can directly be evaluated by the researchers. This would give direct information for the researchers regarding the knowns and unknowns of the model. Second, the statistical aspect of mediation analysis is SEM that always needs a careful investigation. The main reason is a typical SEM model provides a direct path coefficients, specific indirect path coefficients and composite and total indirect path coefficients. Hence a need is understood to maintain a protocol for deciphering direct /indirect effects. This work has augmented this practice with useful computational platform for both numerical and graphical results. To illustrate the data analysis and comprehensive comparison is the third issue in this present chapter. This third aspect will illustrate the different ways for interpreting mediated effect such as Baron and Kenny [1986] or Zhao et al., [2010] approach that are discussed in chapter 3; a natural way to extend to a SEM model and related analyses.

5.2 MATHEMATICAL ASPECTS OF SEM

Let a theoretical or hypothesized model have following components with respective notations:

- $w$: Number of variables,
- $u$: Number of exogenous variables ($\xi$),
- $v$: Number of endogenous variables ($\eta$), (Here $u + v = w$),
- $p$: Number of indicators related to $u$, 

q: Number of indicators related to v, m: p+q

\( \xi_i \): Indicators related to u (i=1,2,3...p),

\( \eta_j \): Indicators related to v (j=1,2,3...q)

\( \delta_i \): Error terms associated with \( x_i \) (i=1,2,3...p),

\( \epsilon_j \): Error terms associated with \( y_j \) (j=1,2,3...q)

Associated matrix notation can be:

\[
X = (x_1, x_2, ..., x_p)^T \\
Y = (y_1, y_2, ..., y_q)^T \\
\theta_\delta = (\delta_1, \delta_2, ..., \delta_p)^T \\
\theta_\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_q)^T
\]

so that order of X and \( \theta_\delta \) is px1 and that and of Y and \( \theta_\epsilon \) is qx1. Consider two more matrices

\( \Lambda_x = (\lambda_{rc})_{pxu} \) and \( \Lambda_x = (\lambda_{rc})_{qvx} \)

which refer to factor loading of u and v variable respectively. Hence

\( \lambda_{rc} = 0 \) if there is no path from \( \xi_c \) to \( x_r \) or \( \eta_c \) to \( y_r \)

Further let \( \xi = (\xi_1, \xi_2, ..., \xi_u)^T \) and \( \eta = (\eta_1, \eta_2, ..., \eta_v)^T \) as two matrices of respective order ux1 and vx1. Now measurement model in SEM can be defined as

\[
X = \Lambda_x \xi + \theta_\delta \\
Y = \Lambda_y \eta + \theta_\epsilon
\]

It can easily be verified that the order of the matrices are compatible with matrix algebraic operation. For example order of \( \Lambda_x \) \textit{and} \( \xi \) is \( p \times u \) and \( u \times 1 \) and
hence their product is a matrix of order $p \times 1$ that can be added to the matrix $\theta_\delta$ of order $p \times 1$. Hence the entire matrix operations of the right hand side of the above equations (relating to a measurement model) equal the order of left hand side matrices respectively. Further, it can be observed that if no difference between $\xi$ and $\eta$ is assumed then the measurement model can be represented in a single matrix equation as

$$X = \Lambda_x \xi + \theta_\delta$$

with the appropriately modified matrices where

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = (x_1, x_2, \ldots, x_m)^T$</td>
<td>$m \times 1$</td>
</tr>
<tr>
<td>$\Lambda_x = (\lambda_{rc})$</td>
<td>$m \times w$</td>
</tr>
<tr>
<td>$\lambda_{rc} = 0$ if no path from $\xi_c$ to $x_r$</td>
<td></td>
</tr>
<tr>
<td>$\xi = (\xi_1, \xi_2, \ldots, \xi_w)^T$</td>
<td>$w \times 1$</td>
</tr>
<tr>
<td>$\theta_\delta = (\delta_1, \delta_2, \ldots, \delta_m)^T$</td>
<td>$m \times 1$</td>
</tr>
</tbody>
</table>

Hence, the data input is restricted with $m$ variable as there are as many as $m$ observed variables. The covariance or correlations for each pair of these $m$ variables constitute the inputs for the underlying measurement model.

Hence, the order of covariance – variance ($C$) matrix correspond to $X$ is $m \times m$ and has $m^2$ elements. The leading diagonal elements of $C$ has variance and off-diagonal elements are covariance ($C_{ij}$) between $x_i$ and $x_j$ ($1 \leq i < j \leq m$). That is

$$
\begin{bmatrix}
  x_1 & x_2 & x_3 & \cdots & x_m \\
  x_1 & c_{11} & c_{12} & \cdots & c_{1m} \\
  x_2 & c_{21} & c_{22} & \cdots & c_{2m} \\
  x_3 & c_{31} & c_{32} & \cdots & c_{3m} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_m & c_{m1} & c_{m2} & \cdots & c_{mm}
\end{bmatrix}
$$
It can be observed that $C_{ij} = C_{ji}$ and $C_{ij}$ is variance of $x_i \ (i = 1,2,...,m)$. Hence it is customary to consider only the diagonal and any one off-diagonal elements; without losing generality upper diagonal may be omitted from the further consideration. This makes the notion for the number of inputs as the sum of number of diagonal elements and lower off-diagonal elements of $C$. :: Number of unique inputs from the observed variables from $C$ is

$$m + \binom{m}{2} = m + \frac{m(m - 1)}{2}$$

$$= m \left[ 1 + \frac{m - 1}{2} \right]$$

$$= \frac{m(m + 1)}{2}$$

On the other hand number of unknowns (or parameters in the notion of statistical inference) is from the matrices of $\Lambda_x, \theta_\delta$ covariance between $\xi_{ij} \ (1 \leq i < j \leq w)$ and variances of $\xi_i \ (i = 1,2,...,w)$ in a matrix form $\phi$. However for each factor $x_r$ of the latent variable($\xi_c$), $\lambda_{rc}$ is fixed to some constant. These observations help to understand the number of parameters involved in a measurement model.

That is number of parameters is $(m - w) + m + w + \binom{w}{2}$. First terms corresponds to elements of the matrix $\Lambda_x$ second term to the matrix $\theta_\delta$, third and fourth terms are associated with the matrix.

$$\phi = \begin{bmatrix}
\xi_1 & \xi_2 & \cdots & \xi_w \\
\xi_1 & \phi_{11} & - & - \\
\xi_2 & \phi_{12} & \phi_{22} & \\
\vdots & \vdots & \vdots & \ddots \\
\xi_w & \phi_{w1} & \phi_{w2} & \cdots & \phi_{ww}
\end{bmatrix}$$
Here, the similarity between $\phi$ and $C$ is evident. ∴ Number of parameters = $2m + \binom{w}{2}$. From the above calculations, it is straightforward to find the degrees of freedom (df) as Input – Parameter or

$$df = \frac{m(m+1)}{2} - 2m - \frac{w(w-1)}{2}$$

where $m$ is the total number of factors and $w$ is that of latent variables. Now if such approach is extended to a structural model of the SEM approach, then it is only to include additional matrices of unknowns or parameters. It is because the input still remains with C matrix, number of input is still $\frac{m(m+1)}{2}$. However the two matrices of regression coefficients pertaining to the paths from $\xi$’s to $\eta$’s and between $\eta$’s.

That is $\Gamma = (\gamma_{ts})$ of path a $v \times u$ matrix containing the regression coefficients from $\xi_t$ to $\eta_t$ so that

$$\gamma_{ts} = 0 \text{ if no paths is identified (theoretically) from } \xi_t \text{ to } \eta_t.$$  In a similar manner, $B = (\beta_{ij})$ is a $v \times v$ matrix whose elements are the regression coefficients of path from $\eta_i$ to $\eta_i$; if no such path is perceived in the model then $\beta_{ij} = 0$. Care must be taken to understand that row of both $\beta$ and $\Gamma$ corresponds to the destination variable and column relates the originating variable. For example $\beta_{12}$ is a path coefficient for a path from $\eta_2$ to $\eta_1$ with all these assumptions the structural model can be represented as

$$\eta = B\eta + \Gamma\xi + \zeta$$

Here $\zeta$ represents the matrix of measurement error associated with the endogenous latent variables. Hence order of the above matrix equation can be revisited as

120
so to enumerate the number of parameters for a structural model is not so straightforward. The main reason is η’s are generally assumed to be uncorrelated and hence covariance is zero, yet in some models one or more pairs of η’s can be allowed to correlate depending on the underlying model specifications. Similarly there may be one or more error variance that is fixed and imposing equality constraints.

If a structural model is assumed without any such (or more) additional constraint, calculation of \( df \) may be attempted in an easier way. That is parameters are from the elements of the matrices \( B \), \( \Gamma \), \( \Phi = (\phi_{ij}) \) variance – covariance matrix of order \( u \times u \) related to \( \xi \)'s, \( \Psi = (\psi_{ij}) \) variance – covariance matrix of order \( v \times v \) related to η’s and \( \psi_{ij} = 0 \) for \( i \neq j \). It can be recalled that the upper off-diagonal elements of \( \Phi \) and \( \Psi \) are not considered. Two more diagonal matrices related to the error terms associated to \( X \) and \( Y \); that is

\[
\theta_\delta = (\theta_{ij}) \text{ of order } p \times p \text{ and } \theta_e = (\theta_{ij}) \text{ of order } q \times q
\]

Here also, it should be observed that off diagonal elements of \( \theta_\delta \) and \( \theta_e \) are zero because of the assumption that error terms are uncorrelated.

Hence, the numbers of parameters are collected from

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>((\beta_{ij}) )</td>
</tr>
<tr>
<td>( \Gamma )</td>
<td>((\gamma_{ij}) )</td>
</tr>
<tr>
<td>( \zeta )</td>
<td>((\zeta_1, \zeta_2, \ldots \zeta_v)^T )</td>
</tr>
<tr>
<td>( \eta )</td>
<td>((\eta_1, \eta_2, \ldots \eta_v)^T )</td>
</tr>
<tr>
<td>( \xi )</td>
<td>((\xi_1, \xi_2, \ldots \xi_w)^T )</td>
</tr>
</tbody>
</table>
(i) Regression coefficients

(ii) Variances of exogenous variables

(iii) Covariance of exogenous variables

(iv) Error terms of the endogenous variables (not the variance of an endogenous variable)

(v) Factor loadings

That is numbers of parameters = $B + G + u + \left(\frac{u}{2}\right) + v + m$

= $B + G + 2m + \left(\frac{u}{2}\right)$  \hspace{1cm} (\because m = u + v)

This relation assumes $u > 1$; in case, if $u = 1$ then it becomes $B + G + 2m$. $b$ and $g$ are respectively the number of nonzero elements in the matrices $B$ and $\Gamma$ respectively. This discussion yields an expression for $df$ of a structural model (with stated assumptions) as

$$
df = \begin{cases} 
\frac{m(m + 1)}{2} - [B + G + 2m] & \text{if } u = 1 \\
\frac{m(m + 1)}{2} - \left[B + G + 2m + \frac{u(u - 1)}{2}\right] & \text{if } u > 1 
\end{cases}
$$

The second term is separated because it would accommodate more inclusion of terms. It may not be possible to derive the scope to include more terms in the second part of above equation for $df$ still some indicative approaches can be listed.

(i) Number of fixed error variance(s) (such as when a construct has only one indicator)

(ii) Some measures may be allowed to load on more than one latent variable let $l$ be number of such factors.
(iii) e: number of free error term covariance as some models may allow measurement error terms to correlate.

(iv) If some exogenous construct are orthogonal to others, then corresponding off diagonal elements of $\phi$ are fixed to zero. Let $f$ be number of such fixed covariances.

(v) In some cases, certain structural error terms are allowed to covary and let $z$ be such quantities.

Hence, consolidating (may not be exhaustive) above five points the formula for $df$ becomes,

$$df = \begin{cases} 
\frac{m(m + 1)}{2} - \left[ b + g + 2m - s - f + l + e + z \right] & \text{if } u = 1 \\
\frac{m(m + 1)}{2} - \left[ b + g + 2m + \frac{u(u - 1)}{2} - s - f + l + e + z \right] & \text{if } u > 1 
\end{cases}$$

(i.e) fixed parameters are subtracted and free parameters are added to the second term of $df$ formula. This indicates that $df$ will increase with inclusion of fixed parameters and decreases when parameters are made free in the model.

To illustrate analysis from Hair et al., [2006] can be considered. A measurement model described in p827 has $m=21$ $w=5$ and hence $df = \binom{21 	imes 22}{2} - \left[2(21) + \frac{5(4)}{2}\right]$ or $df = 179$. The corresponding structural model can be found in p884 that has $b=3$; $g=4$; and $u=2$, so that $df$ is $231 - [3 + 4 + 2(21) + 1] = 181$. Further a re-specification of this model is discussed in p889 which results in adding a $\Gamma$element to the model. This makes the change in $g$ from 4 to 5 and hence $df$ is 180.
5.3 INFERENCE FOR MEDIATION USING SEM

The input data for SEM is usually a covariance matrix of observed variables or a correlation matrix of observed variables and means and standard deviations of the variables. The input provides a key mechanism for testing the quality of a CFA or general standard model. The main aspect of SEM estimation results is measured in terms of how well the SEM model being tested can reproduce the analyzed matrix. Any statistical estimation method available for SEM has to obtain estimates of model parameters that minimize the difference between the input matrix and the model implied matrix. The implied matrix is the matrix of covariance that is as closer to the input matrix as possible, given the hypothesized model. Hence the null hypothesis in SEM is that population covariance matrix ($\Sigma$) equals the matrix that is implied by the CFA or general structural model; that is $H_0: \Sigma = S$.

Like many other statistical inferential approaches this equation also invokes the inference of population values from sample statistics. However, the population matrix $\Sigma$ is not available in most of the cases the sample covariance matrix ($C$) is substituted in the equation

$$\Rightarrow H_0: C = S$$

The difference between two matrices is presented in residual matrix which is calculated as element wise differences between the elements of $C$ and $S$. Hence the main goal of statistical inference in SEM is to estimate the parameters based on $S$. Such that $C$ is as close to $S$ for this, a fitting function defined as $F(C, S) = C - S$ is minimized using fitting functions. Commonly available fitting functions include

(i) MLE, maximum likelihood estimators

$$F = \frac{1}{2} \text{tr}[(C - S)S^{-1}]^2$$
(ii) ULS, unweighted least square

\[ F = \frac{1}{2} \text{tr}[(C - S)^2] \]

(iii) GLS, Generalized least squares

\[ F = \frac{1}{2} \text{tr}[(C - S)C^{-1}]^2 \]

In almost all cases it may be possible to obtain explicit solutions for the unknown parameters and hence iterative numeric procedures are applied to get model implied matrix S. This is based on four matrix block given by

\[ S_{XX} = \Lambda_X \phi \Lambda'_X + \theta_\delta \]

\[ S_{XY} = \Lambda \phi \Gamma'[(I - B)^{-1}]\lambda' \]

\[ S_{YX} = \Lambda_Y (I - B)^{-1} \Gamma \phi \Lambda'_X \]

\[ S_{YY} = \Lambda_Y (I - B)^{-1} (\Gamma \phi \Gamma' + \Psi) [(I - B)^{-1}] \Lambda'_Y + \theta_\epsilon \]

so that \( S = \begin{bmatrix} S_{YY} & S_{YX} \\ S_{XY} & S_{XX} \end{bmatrix} \)

Once the fitting function is minimized, an appropriate statistical test has to be run to determine how similar the two matrices (C and S) are and also one can check whether the differences are statistically significant or not as well. Values obtained with fitting function are \( \chi^2 \) distributed and can be evaluated at the number of df. A non significant \( \chi^2 \) value indicates that the null hypothesis cannot be rejected in the sense that the researcher’s model is consistent with the data. However limitations of \( \chi^2 \) statistic include

(i) Highly sensitive to sample size – large sample chi-square values are inflated
(ii) Affected by the departure from multivariate normality. Hence a large number of additional tests of fit are used to support the claims of good fit even \( \chi^2 \) test is significant.

In the case of absolute fit indices root mean square error (RMSEA) is the second fit statistic next to \( \chi^2 \) statistic and associated p value, RMSEA favours the parsimony in that it will choose the model with lesser number of parameters. The Goodness-of-Fit-statistic (GFI) is considered as an alternative to the \( \chi^2 \) test and calculates the proportion of variance that is accounted for by the estimated population covariance. Another index is the adjusted goodness-of-fit (AGFI) which adjusts the GFI based upon the degrees of freedom. Additional indices in this sequence include root mean square residual (RMR) and its standardized from (SRMR). It is based on the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance model.

Incremental or comparative or relative fit indices are a group of indices which do not use the \( \chi^2 \) in their raw form but can be compared to base line model. Normed Fit Index (NFI) is a statistic that compares \( \chi^2 \) value of the model to check the null model. Another index Non-Normed fit index (NNFI) or Tucker-Lewis Index (TLI) is also used to overcome an issue pertaining to NFI for smaller samples. Yet in some cases NNFI can exceed its maximum cut-off value (1.0) so that interpretations become difficult. Another revised form of the NFI is the Comparative Fit Index (CFI) that performs well even for smaller samples.

There are another two indices in the group of parsimony used in practice. These two are recommended to use together with other measures of goodness-of-fit. The Parsimony Goodness-of-fit Index (PGFI) is based on the GFI by adjusting for
loss of df. Second one is the Parsimonious Normed Fit Index (PNFI) also adjusts for 
df but it is based on NFI. Model complexity is peanalised by these two indices but 
practically no threshold values have been recommended.

More aspect of SEM other than these statistic include, type and number of fit 
indices to be reported; improving the model fit; comparing the completing of 
alternative models. These issues are more on model theory or concept based rather 
than statistical issues. Hence they are not included in the present work; however a vast 
list of literature on this applied area of research like (may not be exhaustive) Barrett 
[2007], Boomsma [2000], McDonald and Ho [2002], Schreiber et al., [2006], 
Schreiber [2008], Preacher and Merkle [2012], Kenny et al., [2014], Kenny and 
McCoach [2003], Marsh et al., [2004], McIntosh [2007], Miles and Shevlin [2007].

The importance of latent variable models can be understood from the 
influence of measurement error on the estimation of parameters involved in any 
analysis or model building process. Measurement error is common in all field of 
scientific investigation. Measurement error can be random due to chance factors and 
can distort estimates of relations between variables. Especially, literature has 
demonstrated the effect of measurement error in mediations analysis [Hoyle, 1999, 
Mackinnon, 2008 and Kline, 2015]. Reliabilities of the underling variables (X, M, or 
Y) affect the respective relations; say unreliability of mediator with an error free 
independent variable affects the estimation of path coefficient (b) between M and Y.

It has been observed that the use of latent variable models would remove the 
effects of unreliability in the mediators compared with using a single measure of the 
mediator with less than perfect reliability. Inclusion of mediators in a latent variable 
model does not alter the mathematical aspects of underlying estimation procedures
such as SEM. Further deciphering and interpretation of indirect effects have drawn reasonable attention. Most of the available SEM softwares also do not accommodate all possible ways of defining indirect effects in a mediation analysis. As an illustration consider the following hypothetical two mediation model with usual notation of a SEM (Showing only single arrow & relations).

![Hypothetical two mediation model](image)

It is customary to define total indirect effect as a combination of all possible paths between independent and dependent variables. For example if $\xi_1$ and $\eta_3$ are considered the path $\xi_1 \rightarrow \eta_1 \rightarrow \eta_3$, $\xi_1 \rightarrow \eta_1 \rightarrow \eta_2 \rightarrow \eta_3$, and $\xi_1 \rightarrow \eta_2 \rightarrow \eta_3$ constitute a total indirect effect which is provided in most of the present day software also. However the interest may lie on understanding the specific interest of $\eta_1$ as mediator between $\xi_1$ and $\eta_3$. This leads to define a specific path $\xi_1 \rightarrow \eta_1 \rightarrow \eta_3$.

Hence, a need arises to understand indirect effect through the product of respective path coefficients ($\gamma_{11}\beta_{31}$). Mackinnon [2008] has indicated two procedures by “tweaking” the matrices involving $\beta$ and / or $\gamma$ coefficients by assuming zeros and
estimate the parameters to obtain the specific effects. This may not be possible for many of the SEM practitioners. This exercise needs a careful manipulation of path coefficients while modeling in a SEM software.

Another approach that can be found in similar literature is to have hand calculations based on the path coefficients estimated using any SEM software. This will include the standard error of respective quantities. Then any method described in chapter 3 can be used to understand the indirect effect. To complete the discussion on the above hypothetical example it provides totally four specific paths $\xi_1 \rightarrow \eta_1 \rightarrow \eta_3$; $\xi_1 \rightarrow \eta_2 \rightarrow \eta_3$; $\xi_2 \rightarrow \eta_1 \rightarrow \eta_3$ and $\xi_2 \rightarrow \eta_2 \rightarrow \eta_3$. Hence, a simpler model (inspired by an example provided in Hair et al., 2006) that has four possible specific effects which practically a cumbersome process to define an exhaustive list of specific effects in a more complex model. So, such attempt is left with the researcher and driven by the research hypotheses. Little et al., [2007] can also be referred for interpretation of different levels of mediation by including specific effects; Full, Partial, Inconsistent and No mediation. Many applied researchers have followed this type of recommendations based on Baron and Kenny [1986] approach [for eg. Salanova et al., 2005]. However, statistical inference warrants to use Sobel’s test of any other suitable method to complete mediation analysis in a latent variable model.

5.4 DATA ANALYSIS

In this study twenty four published data sets have been considered for a comprehensive data analysis. Chapter 1 provides the description of all the data sets. Additionally for the aspects of this chapter, data sets have been regrouped based on the inclusion of latent variables. This is due to the fact that SEM models are mainly used to account the measurement errors. In this sense only eleven data sets (I to XI)
are considered of which 6 data sets (IV to VII, X and XI) have one exogenous and two endogenous variables that portray typical single mediation model as shown in figure 5.3.

Data sets VIII and IX are extracted from Hair et al., [2010]; that have two exogenous and three endogenous variables. Data sets I and II have one exogenous and three endogenous, III has eight exogenous and two endogenous variables. Appropriate mediation hypotheses provide a scope to perform mediation analysis with SEM models. This classification of data sets provides more clarity on the purpose of selected data sets; for example data set III can be used to understand the treatment of more than one independent variable in regression approach. Compared to SEM model treating exogenous variables; similarly data set VIII will illustrate the way of handling specific and composite mediated effect and issues there on.

Figure 5.3: Latent variable mediation model
The causal step of Baron and Kenny [1986] explained in chapter 2 can easily be revisited as

(i) Significance of the path from an antecedent to mediator(path a)
(ii) Significance of the path from the mediator to consequent (path b)
(iii) Significance of the path from antecedent to consequent controlling the effect of mediator (path $c'$)

In SEM analysis also classification of mediation relations are following the step (i) – (iii) which is uniformly visible in almost all SEM research reports such as Salanova et al., [2005]. However, most of similar studies do not report the statistical significance of indirect effects through normal theory based or any other resampling methods. This could cause the statistical aspect incomplete or a thorough mediation analysis may not be available. Hence, based on a research hypothesis it is necessary to identify the nature of indirect effect as specific or composite paths. Further most or the software packages are providing the indirect effect due to composite paths rather than specific paths. Hence, data analysis procedure is based on following step-wise approach that includes routine SEM steps such as improving model fit and other recommendations.

Step 1: calculate the reliability for each variable to understand measurement error.
Step 2: Perform a Confirmatory Factor Analysis (CFA) with constrained (from step 1) or unconstrained approach.
Step 3: Adopt model improvement, if necessary
Step 4: Perform the hypothesized structural model and check the model fit, improve with necessary steps if necessary so as to confirm the model fit.
Step 5: if the hypothesis warrants a composite indirect path, extract from the software package such as AMOS or LISREL (used in the present work) else; if specific indirect is hypothesized, then

Step 6: Extract the path coefficients, a, b and c with respective standard errors (SE)

Step 7: using two R-functions “typmedBK” where BK means Baron and Kenny [1986] approach and “typmedZH” where ZH means Zhao et al., [2010] approach implemented in R programming environment developed during the present work conclusion can be made.

“typmed()” function provides the estimates and statistical significance of direct and indirect effects (using Sobel approach); both aspects of interpretations are provided for each set of hypothesized effects. A graphical display will help the user to understand the effects using z value ($z = \frac{Estimates}{SE}$). Following figure

![Graphical display of Zhao approach](image)

Figure 5.4: Graphical display of Zhao approach
Illustrates the display for Zhao et al., [2010] procedure assuming the level of significance $\alpha = 0.05$. NS refers to non-significance and S refers to significance, IDE refers indirect effect, arrow mark shows the NS for direct and indirect effects. Another graphical display provides the estimates of direct and indirect effect (pairwise) so that their sign pattern can be understood and the approach of Zhao et al., [2010] is readily followed.

Similar output can be obtained for Baron and Kenny [1986] also using “typmedBK” function in R. Based on these observations results are presented in tables 5.2 (data sets I,II), 5.3 (data set III) 5.4 (data sets IV to VII, IX, X and XI), and 5.5 (data set VIII). Also the following display lists the variables involved in the data sets considered for the study.
Figure 5.5: Multiple independent variables using Baron and Kenny approach

Figure 5.6: Multiple independent variables using Zhao approach

Figure 5.7: Single mediator models using Baron and Kenny approach
Figure 5.8: Single mediator models using Zhao approach

Figure 5.9: Composite indirect effect using Baron and Kenny approach

Figure 5.10: Composite indirect effect using Zhao approach
For data set I Tangibility (TAN) is exogenous variable, Reliability (REL), Affordability (AFF) and Satisfaction (SAT) are endogenous variables in which REL and AFF are two mediators. Results from table 5.2 indicate the complementary mediated effects for data set I whereas the causal step approach fails to identify any relationship because of non significant relationship between exogenous and endogenous as well as between two endogenous variables \((TAN \to AFF \ and \ AFF \to SAT)\). For data set II, Activities (ACT) is exogenous variable, Lateral thinking skills (LAT), Mindset (MIN) and Communication skills (CSK) are endogenous variables in which LAT and MIN are mediators. Similar conclusion is arrived for data set II through causal step because of non significance effect between endogenous variables \((MIN \to CSK \ and \ LAT \to CSK)\). Zhao et al. [2010] approach also yields no direct non-mediation effect that is thought in similar line with Baron and Kenny [1986] approach.

In the notion of mediation analysis, eight exogenous variables such as Value for Education (VFE), Aspiration (ASP), Parent Teacher Encouragement (PTP), Need for Cognition (NFC), Personal involvement (PIN), Perceived Cost (PCO), Perceived Benefits (PBE), College Attribute (CAT) with Motivation (MOT) and Extent of search (EOS) are endogenous variables in which MOT as mediating variable are considered in data set III (Table 5.3). There are 17 specific paths in which eight paths towards MOT and nine paths towards EOS including the direct path. Baron and Kenny [1986] approach is used to find the mediated effect. It can be observed that all independent variables have indirect effect except for PTP, NFC, and CAT due to non significant effects in the paths between exogenous and mediating variables. When these three relations are further examined with Zhao et al., [2010] approach which also indicates no mediating effect. While indirect effect \((ab)\) are considered for other
five paths, three estimates such as PTP, NFC and PCO have positive sign whereas the rest have negative estimates. When direct effect (c') are considered five estimates are positive in the case of VFE, PTP, PIN, PCO and PBE whereas rest have negative estimates. Also, PIN, PCO and PBE showed significant direct effects when compared to other independent variables. Based on the sign of ab\*c', it can be observed that PIN, PBE have competitive and PCO has complementary mediation whereas VFE and ASP have indirect effect only.

Table 5.4 provides the results from mediation analysis for seven data sets (IV to VII and IX to XI) as classified earlier; all the data sets are having only one mediating, independent and dependent variable. All data sets except VI and XI show complementary indirect effect (positive direct and indirect effects) when Zhao et al., [2010] approach is used; XI has direct-non mediation; an indirect only mediation can be observed for data set VI. This pattern is similarly followed in Baron and Kenny [1986] approach also except that dataset XI exhibits non-significant effect between exogenous and mediating variable. Data set II shows full mediation and remaining five have partial mediation.

For data set VIII, Attitude towards coworkers(AC) and Environmental perceptions(EP) are exogenous variables & Job satisfaction(JS), Staying intention(SI) and Organizational commitment(OC) are endogenous variables in which JS is mediating variable. The underlying model is quite similar to IX except the difference that the mediation hypothesis is to test whether JS mediates EP and SI. From Table 5.5 it can be observed that all three paths (EP → JS, EP → JS and JS → SI) are significant and hence Baron and Kenny (BK) approach indicates a partial mediation; but the product approach yields a non significant indirect effect (\(\hat{ab} = 0.016\); z value =
1.256) and Zhao et al. (ZH) approach shows a Direct-only non-mediation effect. Further, if the entire model is considered and indirect effect can be extracted from SEM analysis. Then composite indirect effect results with (combination of EP $\rightarrow$ JS $\rightarrow$ SI, EP $\rightarrow$ OC $\rightarrow$ SI and EP $\rightarrow$ JS $\rightarrow$ OC $\rightarrow$ SI) an estimate of 0.175 (SE=0.033, Z=5.331) and shows partial mediation. But the third approach is not intended to test based on the mediation hypothesis.
### Table 5.2: Structural equation modeling approach for multiple mediator models.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Paths</th>
<th>Estimates</th>
<th>Direct Effect</th>
<th>Total indirect effect</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From</td>
<td>To</td>
<td>Est</td>
<td>SE</td>
<td>t value</td>
</tr>
<tr>
<td>I</td>
<td>a1</td>
<td>TAN  REL</td>
<td>-0.223</td>
<td>0.065</td>
<td>-3.425</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>TAN  AFF</td>
<td>-0.087</td>
<td>0.076</td>
<td>-1.150</td>
</tr>
<tr>
<td></td>
<td>a3</td>
<td>REL  SAT</td>
<td>-0.781</td>
<td>0.400</td>
<td>-1.953</td>
</tr>
<tr>
<td></td>
<td>a4</td>
<td>AFF  SAT</td>
<td>-0.074</td>
<td>0.368</td>
<td>-0.201</td>
</tr>
<tr>
<td></td>
<td>a5</td>
<td>TAN  SAT</td>
<td>0.826</td>
<td>0.222</td>
<td>3.721</td>
</tr>
<tr>
<td>II</td>
<td>a1</td>
<td>ACT  LAT</td>
<td>0.871</td>
<td>0.256</td>
<td>3.402</td>
</tr>
<tr>
<td></td>
<td>a2</td>
<td>ACT  MIN</td>
<td>0.910</td>
<td>0.172</td>
<td>5.291</td>
</tr>
<tr>
<td></td>
<td>a3</td>
<td>LAT  CSK</td>
<td>0.732</td>
<td>0.506</td>
<td>1.447</td>
</tr>
<tr>
<td></td>
<td>a4</td>
<td>MIN  CSK</td>
<td>0.352</td>
<td>0.565</td>
<td>0.623</td>
</tr>
<tr>
<td></td>
<td>a5</td>
<td>ACT  CSK</td>
<td>-0.486</td>
<td>0.793</td>
<td>-0.613</td>
</tr>
</tbody>
</table>

### Table 5.3: Structural equation modeling approach for multiple independent variables.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Direct Effect</th>
<th>Hyp</th>
<th>Indirect effect (Specific Paths)</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c'</td>
<td>SE</td>
<td>t</td>
<td>sig</td>
</tr>
<tr>
<td>III</td>
<td>0.035</td>
<td>0.080</td>
<td>0.435</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>-0.038</td>
<td>0.078</td>
<td>-0.487</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>-0.029</td>
<td>0.048</td>
<td>-0.606</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>0.049</td>
<td>0.056</td>
<td>0.874</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>0.225</td>
<td>0.046</td>
<td>4.934</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>0.254</td>
<td>0.053</td>
<td>4.802</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>0.256</td>
<td>0.072</td>
<td>3.570</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>-0.060</td>
<td>0.049</td>
<td>-1.225</td>
<td>N</td>
</tr>
</tbody>
</table>
Table 5.4: Structural equation modeling approach for single mediator models.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Direct Effect</th>
<th>Hyp</th>
<th>Indirect effect (Specific Paths)</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$e'$</td>
<td>SE</td>
<td>t value</td>
<td>sig</td>
</tr>
<tr>
<td>IV</td>
<td>0.488</td>
<td>0.068</td>
<td>7.176</td>
<td>Y</td>
</tr>
<tr>
<td>V</td>
<td>0.515</td>
<td>0.078</td>
<td>6.603</td>
<td>Y</td>
</tr>
<tr>
<td>VI</td>
<td>0.038</td>
<td>0.076</td>
<td>0.500</td>
<td>N</td>
</tr>
<tr>
<td>VII</td>
<td>0.499</td>
<td>0.074</td>
<td>6.743</td>
<td>Y</td>
</tr>
<tr>
<td>IX</td>
<td>0.101</td>
<td>0.041</td>
<td>2.463</td>
<td>Y</td>
</tr>
<tr>
<td>X</td>
<td>0.660</td>
<td>0.302</td>
<td>2.185</td>
<td>Y</td>
</tr>
<tr>
<td>XI</td>
<td>0.573</td>
<td>0.284</td>
<td>2.018</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 5.5: Composite indirect effect.

<table>
<thead>
<tr>
<th>data set</th>
<th>Estimates</th>
<th>Direct Effect</th>
<th>Indirect effect (Specific Paths)</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>t Value</td>
<td>c'</td>
</tr>
<tr>
<td>VIII</td>
<td>0.245</td>
<td>0.062</td>
<td>3.984</td>
<td>0.371</td>
</tr>
</tbody>
</table>
5.5 CONCLUSION

The ever growth of handling latent variables and associated measurement errors in conceptual theory building or hypotheses testing provide ample scope to formulate and validate models in many applications. Recent years have witnessed such extensive studies across many disciplines in medical epidemiological, engineering, psychological, organizational sciences etc. The structural equation modeling (SEM) is one of techniques that is highly suitable to handle such scenarios through a more established mathematical and statistical principles. This is augmented with the availability of not so expensive computing facilities; yet successful utility of SEM through proper software should be initiated and driven by reasonable theory underlying the problems of interest. Norman and Streiner [2003] observed the model building as “the computer can’t help at this stage; you actually have to think on your own, you specify the model, based on your knowledge of the field, on your reading of the literature, or on theory.”

Hence, keeping aside the model specification or related aspects of SEM, this chapter has focused on mathematical. Statistical and interpretation part of mediation analysis when latent variables are involved. This not only provides the inclusion of latent variable but emphasis on understanding the inclusion of measurement errors in the underlying model. The comparative data analysis has brought out this observation that SEM technique can be considered as a unified approach for handling multiple independent, mediators and dependent variables. Though SEM literature adopts different names the basic frame work remains somewhat similar to regression approaches.
A careful evaluation of hypothesized model through the two step model building (measurement and structural) is an essential precursor to any outcome of SEM – Model testing, validating with or without mediation analysis. Hence an appropriately fitted model is alone be prepared for subsequent hypothesis testing procedures.

However mediation analysis from a properly tested model has its own computational aspects. This has been highlighted from an interpretation point of view where the statistical facts are playing important roles. This starts from ably distinguishing the indirect effects between specific and composite effects. Recent tutorial kind of studies may not reveal these details but they are essential for practitioners because most of the studies solely depend on the routine output of software packages [for example, Gunzler et al., 2013]. On the other hand causal mediated effect has drawn recent research activities such as Muthen and Asparouhov [2015].

The present work made an attempt to illustrate the main difference between two types of indirect effect. It provides through which the significance of indirect effect can be studied appropriately for stated hypotheses when SEM is in practice. This helps to extend the interpretation aspects involved in a mediation analysis detailed in earlier chapter. This will supplement the active research on SEM in a variety of scientific investigations.