7

Statistical Fundamentals: Structural Equation Modeling

Numbers never lie; they simply tell different stories

- LUIS ALBERTO URREA

INTRODUCTION

In this chapter, I introduce the basics of structural equation modeling (SEM). There are many different versions of SEM, including full information SEM and limited information SEM and many variants within each of these types. In Chapter 8, I delve into these variants but in the current chapter, I focus on traditional full information estimation SEM. Some researchers make a distinction between **path analysis** and **SEM**, defining the former as models that have no latent variables; each construct is represented by a single measure. I personally find the distinction arbitrary. The phrase "structural equation modeling" implies we "model" a set of "equations" that specify "structural" or causal relationships between variables. This applies to both path analysis and models with latent variables. As well, path analytic models have latent variables and measurement models implied within them. A path analytic model with an observed measure consisting of a self-report of income assumes that the self-report is impacted by a latent construct of true income and that the measure has no measurement error (see Chapter 3). Path analytic models also include unmeasured latent variables in the form of disturbance terms. In this book, I will not make a distinction between path analysis and SEM. To me, they both represent SEM.

SEM is a complex method that can't be adequately summarized in a single chapter. The current chapter is long and not well suited to reading in a single sitting. My emphasis is on describing core concepts including discussion of the concepts of covariance/correlation decompositions, model identification status, maximum likelihood estimation, global indices of model fit, localized fit indices, the inclusion of latent variables in causal models, making model revisions when faced with poor model fit, and comparing competing models in SEM. I recommend you roughly divide the chapter into thirds and read each third in a separate sitting. Useful references for learning more about SEM include the books by Bollen (1989), Brown (2015), Kaplan (2008), Kline (2023), Little (2013) and Schumacker and Lomax (2016). I introduce SEM using a highly simplified example outside the context of an RET. I do so for pedagogical purposes. Future chapters beginning with Chapter 11 apply SEM to RETs directly.

THE BASICS OF SEM

The general logic of SEM is as follows:

1. We posit a causal model that we think may be operating in a given context, such as an RET

2. We specify the predictions the model makes about how the data that we collect to evaluate the model should pattern themselves

3. We analyze the data to determine if they do indeed pattern themselves in the predicted fashion

4. We reject the model if the data do not pattern themselves in the predicted ways. If the data pattern themselves as predicted by the model, our confidence in the viability of the model usually increases, depending on the quality of the research design we used.

5. If the model proves to be consistent with the data, then we interpret the parameter estimates of the model to gain further insights into the phenomena we are studying.

Consider the model in Figure 7.1. The substantive focus of this model is on the relationship between the size of the family that a child grows up in and the child's intelligence. According to the model, family size influences the amount of time a parent can devote to each child; parents with a large number of children devote less time to each child than parents with few children. This decreased attention from the parent, in turn, is thought to negatively impact the intellectual development of the child. For the sake of pedagogy, set aside for now the problem of confounding and omitted variables. It turns out that the model in Figure 7.1 makes certain predictions about how the correlations or covariances between the three variables should pattern themselves. The question is, do the predicted correlations, in fact, pattern themselves in the way the model predicts. SEM addresses this question and the correspondence between data and model represents a test of model viability.

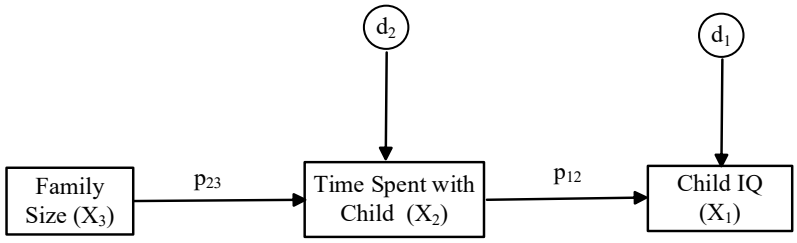


FIGURE 7.1. Example model of child IQ

In traditional SEM, a common practice is to label a causal path with a *p* and two subscripts. The first subscript indicates the variable number the path is directed at and the second number is the variable number the path emanates from. I adopt this notation when

explaining the logic of SEM because it is convenient for the proofs I present later. However, I deviate from it in future chapters for the sake of simplicity.

The model in Figure 7.1 can be translated into a set of linear equations. We need to perform such a translation in order to be specific about the model's predictions about how the data should pattern themselves. The first step in the translation process is to identify the endogenous and exogenous variables in the influence diagram. An **endogenous variable** is a variable with at least one cause that is explicitly represented in the system, i.e., any variable with a straight causal arrow pointing directly to it. An **exogenous variable** is a variable with no represented causes in the system or no arrows pointing directly to it. In Figure 7.1, family size is an exogenous variable and time spent with the child and child intelligence are endogenous variables. The number of equations to be specified equals the number of endogenous variables. For Figure 7.1, there are two equations.[1]

To specify each equation, pick one of the endogenous variables and treat it as an outcome. Express it as a linear function of all variables that have an arrow pointing directly to it and then add a disturbance term. For the variable of child IQ, using traditional regression notation, I have

$$X_1 = a_1 + b_1 X_2 + d_1$$

where a_1 is the intercept, b_1 is the regression coefficient and d_1 is the disturbance term (I use sample notation here). For time spent with the child, the equation is:

$$X_2 = a_2 + b_2 X_3 + d_2$$

If I use the notation scheme mentioned earlier, the equations substitute *p*s for *b*s, but they represent the same thing, namely regression coefficients, like this:

$$X_1 = a_1 + p_{12} X_2 + d_1$$

$$X_2 = a_2 + p_{23} X_3 + d_2$$

To introduce the logic of SEM in ways that are somewhat easier to understand, I am going to work with the standardized equations rather than unstandardized equations. Thus, I standardize X_1, X_2, and X_3 and turn them into Z scores. This allows me to develop the core logic of SEM using correlations rather than covariances, which will be simpler for many of you. Technically, the underlying statistical theory of SEM uses raw

[1] Technically, the equations that describe a full SEM model are more complex than my characterizations here and are best represented using matrix algebra (see Bollen, 1989). I simplify matters now for the sake of pedagogy.

scores and covariances but such data and statistics makes it more challenging to understand the general logic of SEM that I am going to present. Using standardized equations makes it easier to gain a general appreciation of the method. The two model equations expressed in standardized form are

$Z_1 = p_{12} Z_2 + d_1$

$Z_2 = p_{23} Z_3 + d_2$

where Z_1 is the standard score of intelligence, Z_2 is the standard score of time spent with each child, Z_3 is the standard score for family size, and d_1 and d_2 are the disturbance terms for these standardized equations.[2] There is no intercept in these equations because intercepts are always equal to 0 in linear equations with fully standardized scores, a concept that is taught in introductory statistics.

Suppose I collect data on the three variables in the causal model in Figure 7.1 and I calculate a correlation matrix for the measures of them. Here is the correlation matrix that might result:

	Family Size	Time Spent	IQ
Family Size	1.00	-.30	-.60
Time Spent	-.30	1.00	0.30
IQ	-.60	0.30	1.00

The model in Figure 7.1 represents a theory about this correlational structure, i.e., it is thought to explain why the observed correlations pattern themselves as they do. According to the theory, intelligence and the amount of time that a parent spends with each individual child are correlated *because* the latter variable influences the former variable. Similarly, family size and the amount of time spent with a child are correlated *because* family size influences time spent with the child. As I stated, there are many confounds for these relationships, but I set those aside for now in the interest of exposition.

So, what are the predictions that the model makes about the values and patterning of the observed correlations? To specify this in a precise way, I need to **decompose** each correlation in the correlation matrix to express that correlation to be a function of the path

[2] Do not confuse these Z scores with my use of Z scores in Chapter 5 for probit analysis. In Chapter 5, the Z scores referred to scores in a cumulative standard normal distribution. Here they are traditional Z scores in which a mean is subtracted from person's raw score and this result is divided by the standard deviation

coefficients in the model. This decomposition process is the heart of SEM and is what separates it from many other forms of statistical analysis. It is important that you understand this decomposition process conceptually. The actual decomposition mechanics are not difficult but the process involves numerous algebraic manipulations. Bear with me as I illustrate the procedure to you. In order to simplify my explanation, I am make some simplifying assumptions, almost all of which can be relaxed in real world applications:

1. First, I assume there is no measurement error in the variables.

2. Second, I assume all the relationships in the model are linear.

3. Third, I assume the metrics of the variables have interval level properties.

4. Fourth, I assume that a disturbance term is uncorrelated with the presumed determinants of the endogenous variable in the equation that generates the disturbance term. In the present case, d_1 is assumed to be uncorrelated with X_2 and d_2 is assumed to be uncorrelated with X_3.

5. Finally, I assume the disturbances (d_1 and d_2) are uncorrelated.

To perform the decomposition of each correlation, I use a special formula for the correlation coefficient. You have undoubtedly encountered formulae for calculating a correlation in your introductory statistics texts but I use one that is not widely known. The formula illustrated for r_{23} from our example is:

$$r_{23} = (1/N) \Sigma Z_2 Z_3$$

That is, I calculate a correlation between two variables by (1) standardizing each variable, (2) multiplying the two standardized scores for each individual, and then (3) averaging these products. Executing this formula will produce the exact same result as other formulae for correlations that you have encountered.

The decomposition process for r_{23} is summarized in Table 7.1. I explain each row of that table shortly. Here is a video link that talks you through the decomposition process if you prefer that form of learning (I recommend you watch the video as well as reading the text): SEM decomposition. [*If you are not reading this pdf in a browser, then left click on the link. If you are reading this pdf from within Chrome, right click the link and choose to open the link in a new window; if reading it from within Safari, hold down the command key while clicking the link. If reading a printed copy, see the video link on the Resources tab of my webpage for Chapter 7.*]

Table 7.1: Decomposition of r23

Row	Expression	Principle Used
1	$r_{23} = (1/N) \sum Z_2 Z_3$	*correlation formula*
2	$= (1/N) \Sigma (p_{23} Z_3 + d_2) Z_3$	*Substitution*
3	$= (1/N) \Sigma (p_{23} Z_3 Z_3 + Z_3 d_2)$	$a(b+c) = ab + ac$
4	$= (1/N) [\Sigma p_{23} Z_3 Z_3 + \Sigma Z_3 d_2]$	$\sum (X + Z) = \sum X + \sum Z$
5	$= (1/N) \Sigma p_{23} Z_3 Z_3 + (1/N) \Sigma Z_3 d_2$	$c (\sum X + \sum Z) = c \sum X + c \sum Z$
6	$= p_{23} (1/N) \sum Z_3 Z_3 + (1/N) \sum Z_3 d_2$	$\sum c X = c \sum X$
7	$= (p_{23})(1.0) + 0$	-
8	$= p_{23}$	-

Row 1 of Table 7.1 presents the prior formula for the correlation coefficient. Recall that according to the model in Figure 7.1, $Z_2 = p_{23} Z_3 + d_2$. I can substitute the right-hand side of this equation for Z_2 in the correlation formula, as follows (per row 2 of Table 7.1):

$$r_{23} = (1/N) \Sigma (p_{23} Z_3 + d_2) Z_3$$

In row 3, I expand the product of $(p_{23}Z_3 + d_2)Z_3$ by multiplying each term within the parentheses by Z_3, i.e., Z_3 times $p_{23}Z_3$ and Z_3 times d2. This yields

$$r_{23} = (1/N) \Sigma (p_{23} Z_3 Z_3 + Z_3 d_2)$$

There is a rule for summations in algebra that states that

$$\Sigma (X + Y) = \Sigma X + \Sigma Y$$

I invoke this rule for the term $\Sigma (p_{23} Z_3 Z_3 + Z_3 d_2)$ to obtain the expression in row 4 that splits up the summation into two summations:

$$r_{23} = (1/N) [\Sigma p_{23} Z_3 Z_3 + \Sigma Z_3 d_2]$$

Another summation rule is that for a constant, c,

$$c (\Sigma X + \Sigma Y) = c \Sigma X + c \Sigma Y$$

I use this rule in row 5 to multiply each of the summation terms in the expression [Σp_{23}

$Z_3 Z_3 + \Sigma Z_3 d_2$] by 1/N, yielding

$(1/N) \Sigma p_{23} Z_3 Z_3 + (1/N) \Sigma Z_3 d_2$

Yet another summation rule is that if *c* is a constant then

$\Sigma c X = c \Sigma X$

I invoke this rule in row 6 to bring p_{23} outside of the summation because p_{23} is a constant:

$r_{23} = p_{23} (1/N) \Sigma Z_3 Z_3 + (1/N) \Sigma Z_3 d_2$

The above equation is noteworthy in several respects. First, note that the expression $(1/N) \Sigma Z_3 Z_3$ in it is the formula for a correlation coefficient and, in this case, it is the correlation of Z_3 with itself. It must equal 1.0. Also, the expression $(1/N) \Sigma Z_3 d_2$ is a correlation between Z_3 and d_2, which must be zero given the assumptions I made at the outset. This yields the following (see rows 7 and 8 of Table 7.1):

$r_{23} = (p_{23})(1.0) + 0$

$r_{23} = p_{23}$

According to the decomposition exercise, the correlation between Z_2 and Z_3 for this model is simply the path coefficient, p_{23}. This is not particularly earth shaking and underscores a well-known statistical fact; in bivariate regression, the standardized regression coefficient for the regression of one variable on another equals the value of the correlation coefficient. If I conduct a comparable decomposition for r_{12}, the result would be similar, namely I would find that $r_{12} = p_{21}$ because this path also is focused on a bivariate relationship. I can now re-draw the influence diagram but substitute into it the values of the (standardized) path coefficients taken from the observed correlation matrix because the path coefficients equal the correlations. I do so in Figure 7.2.

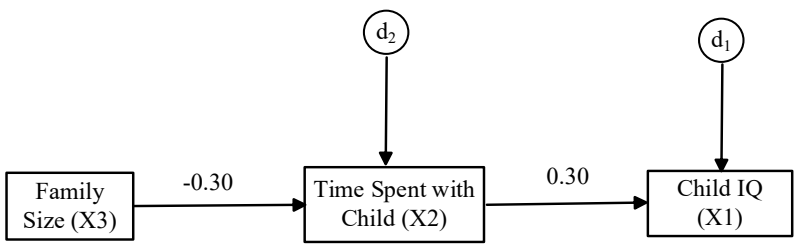


FIGURE 7.2. Path model with standardized values

The significance tests for the path coefficients in this case are simply the significance tests for the correlation coefficients. As is well known, one minus the squared correlation is the proportion of unexplained variance, so I can further fill out the influence diagram with values for the variances for the disturbance terms, which I do in Figure 7.3. Family size accounts for 9% of the variation in the time parents spend with a child (with 91% of the variance being due to other factors) and the amount of time a parent spends with a child accounts for 9% of the variation in the child IQ.

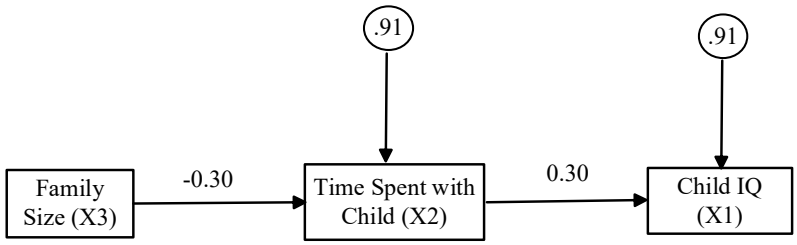


FIGURE 7.3. Model with disturbance values

None of the above is very insightful beyond what we do in traditional statistics. The unique logic of SEM becomes more apparent when I perform the same decompositional analysis but now focused on the correlation between Z_1 and Z_3. Everything I am about to do I have already illustrated in the prior decomposition; there is nothing new here. Let's go through the decomposition. The correlation is

$$r_{13} = (1/N) \Sigma Z_1 Z_3$$

According to Figure 7.1, $Z_1 = p_{12} Z_2 + d_1$. I therefore substitute the right-hand side of this equation for Z_1 into the formula:

$$r_{13} = (1/N) \Sigma (p_{12} Z_2 + d_1) Z_3$$

I next expand the product of Z_3 times $(p_{12}Z_2 + d_1)$, so that

$$r_{13} = (1/N) \Sigma (p_{12} Z_2 Z_3 + Z_3 d_1)$$

Next, I break the above summation into two summations and apply (1/N) to each one:

$$r_{13} = (1/N) \Sigma p_{12} Z_2 Z_3 + (1/N) \Sigma Z_3 d_1$$

and because p_{12} is a constant, I bring it to the left of the summation

$r_{13} = p_{12} (1/N) \Sigma Z_2 Z_3 + (1/N) \Sigma Z_3 d_1$

Note that the expression $(1/N) \Sigma Z_2 Z_3$ in the above equation is the formula for a correlation coefficient and in this case it is the correlation of Z_2 and Z_3. As shown above, the correlation between Z_2 and Z_3 equals p_{23}. The expression $(1/N) \Sigma Z_3 d_1$ also is a correlation in this case between Z_3 and d_1, which is zero, by assumption. This yields

$r_{13} = (p_{12})(r_{23}) + 0 = (p_{12})(p_{23})$

Based on this equation, if the underlying causal model in Figure 7.1 is correct, r_{13} should equal the product of p_{12} and p_{23}. Because I know the estimated values of these path coefficients (see Figure 7.3), I can multiply the two path coefficients to obtain a prediction of what the correlation r_{13} should be assuming the model in Figure 7.1 is correct. The values of p_{12} and p_{23} are 0.30 and -.30, respectively, and the product of these two coefficients is -.09. According to the model in Figure 7.1, the correlation between family size and IQ should be -.09. However, when I examine the observed correlation between these two variables that occurred in the data, the correlation is -0.60. The discrepancy between the predicted and observed correlation is large and this calls the model into question. The data are not patterning themselves as the model predicts.

This is a simplified example, but it conveys the basic logic of SEM. The researcher specifies a conceptual model that s/he believes can account for the variances and covariances among a set of observed variables. This model is translated into a set of linear equations. A decomposition analysis is performed on each variance/covariance, similar to what I did above. SEM software derives values for the path coefficients in the model (a process I comment more on shortly) and then calculates from these values and the decomposition analyses the predicted variances and covariances by the causal model under consideration. We then compare the predicted variances and covariances with the observed variances and covariances in the data. If the predicted values are close to the observed values, then the model is said to be consistent with the data (note: the model is not proven; it is merely consistent with the data) and our confidence in the model's viability increases. If the predicted values deviate substantially from the observed values, the model is rejected as viable.

Table 7.2 presents the observed correlation matrix for the current example (repeated from earlier), the predicted correlation matrix based on the model (via the decomposition analyses), and the difference between the two matrices, where each cell of the predicted matrix has been subtracted from the corresponding cell of the observed matrix. This latter matrix is called a **residual matrix**. A perfect fitting model will yield a residual matrix that has all zeros in it, i.e., it will be a **zero matrix**.

Table 7.2: Predicted, Observed and Residual Correlations

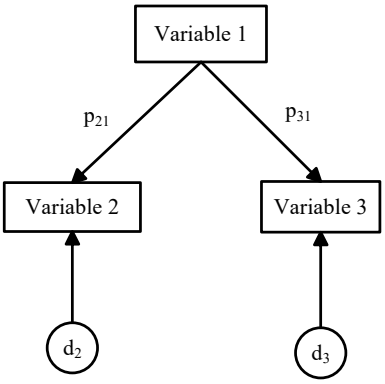
Observed Matrix			Predicted Matrix			Residual Matrix		
1.00	-.30	-.60	1.00	-.30	-.09	0.00	0.00	-0.51
-.30	1.00	0.30	-.30	1.00	0.30	0.00	0.00	0.00
-.60	0.30	1.00	-.09	0.30	1.00	-0.51	0.00	0.00

If a model fits the data well and is judged to be consistent with the data, then researchers often interpret the estimated path coefficients in the model. The path coefficients are interpreted like regression coefficients in a linear model because, after all, they *are* regression coefficients. In the present case, I would not interpret the path coefficients because the model was a poor fitting model given the non-zero character of the residual matrix. It makes no sense to conclude that a model is inconsistent with the data and then to interpret the parameter estimates for that bad fitting model.

In sum, one property that separates SEM from traditional regression modeling is that it translates a causal model into a set of equations and then evaluates the viability of the causal model by comparing the predictions the model makes about the patterning of data with the actual patterns of data that occur. In traditional regression, one simply assumes that the regression model one tests is correct and calculates and interprets the regression coefficients for the model without testing model viability. SEM takes the analysis a step further by allowing us to test the viability of a casual model. To be sure, we can never prove causality from correlational data nor prove that the model is correct. But we *can* use correlational data to evaluate if a causal model accurately predicts data patterns and in that sense, gain perspectives on the viability of the causal model.

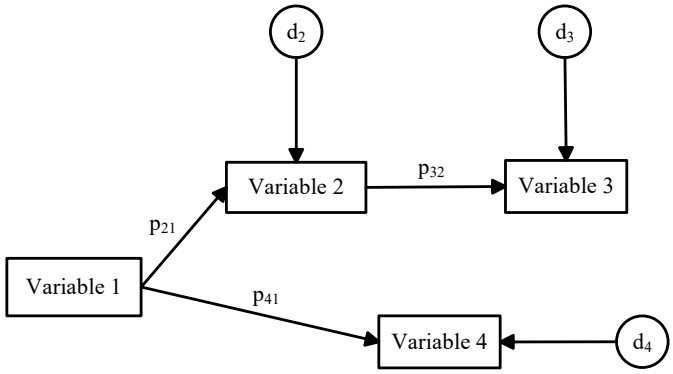
You will see in future chapters many other advantages of SEM over traditional regression methods, which is why I like to use SEM for the analysis of RETs. To be sure, there are challenges that one encounters in using SEM, but in the final analysis, it is a powerful method for analyzing RETs and conducting program evaluations. The remainder of this book illustrates this fact. Although the decomposition analyses I outlined above seem cumbersome, SEM computer software does all decompositional analyses for you automatically and it also produces predicted versus observed correlations and covariances for most models. It also generates summary indices of the degree of fit between model predictions and the observed data, which I consider below.

For completeness, let's examine a few additional examples of predictions that different models make based on SEM decomposition analyses. Consider the model



Decomposition analysis of this model shows that the model predicts that r_{23} = p_{21} times p_{31} when the path coefficients are standardized. If the standardized p_{21} = 0.50 and p_{31} = .20, what should r_{23} equal for this model to be viable? Suppose you found in your data that p_{21} and p_{31} had these values and that r_{23} equals 0.50. What would you conclude?

For the model



decomposition analysis shows the correlation between variables 3 and 4 should equal p_{21} times p_{32} times p_{41} for the standardized path coefficients. If the standardized values are p_{21} = 0.20 and p_{32} = .20, and p_{41} = .50, what should r_{34} equal? Suppose the correlation r_{34} equals 0.02. What would you conclude?

I now discuss numerous qualifications and elaborations of the core SEM logic.

TAUTOLOGICAL PREDICTED AND OBSERVED COVARIANCES

Inspection of Table 7.2 reveals some subtleties in the logic of comparing predicted and observed covariance matrices in SEM. On the one hand, we test the viability of a model by examining how well it reproduces the covariance/correlation matrix between the observed variables from the path coefficients in the model. However, we also use the

covariance/correlation matrix to derive the values of the path coefficients. This can sometimes leads to cases where perfect prediction of a correlation or covariance is guaranteed because it is a mathematical tautology with the path coefficient(s). For example, I found in our decomposition of correlations in the family size and IQ example that r_{23} equals p_{23}. In this case, the "prediction" of r_{23} from p_{23} is tautological; perfect "prediction" of the correlation from the path coefficient is guaranteed because, after all, I *defined* the values of the path coefficient and the correlation coefficient as being the same Stated another way, I am telling you the value of the path coefficient is the value of the correlation coefficient and then asking you to think I have done something meaningful by predicting the value of the correlation coefficient from the path coefficient.

Such a tautology, however, was not true for r_{13}. This correlation, according to the model, should equal $p_{12}\ p_{23}$ and I found that the product of the two estimated path coefficients did not predict the observed value of r_{13} well. In short, some of the variances and covariances will be perfectly reproducible because of mathematical regularities and tautologies within the theoretical system but this may not be true of other variances and covariances in the model. Only in the latter case are comparisons between predicted and observed values meaningful as formal tests of model viability.

Sometimes a theorist may specify a causal model to evaluate in which *every* observed variance and covariance is perfectly predictable because of rampant mathematical dependencies and tautologies in the model. Such models are said to be **just-identified** and these models cannot be evaluated by comparing predicted and observed correlations/covariances because it is mathematically guaranteed that the predicted and observed correlations will be identical. Another term sometimes used for a just-identified model is a **saturated model**. These models are not testable by virtue of traditional SEM logic. We need to use specialized methods for testing just identified models and I discuss these strategies in future chapters (as well as below).

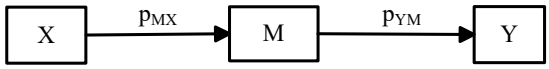
Other models that might be specified by a theorist will have at least one variance or covariance/correlation that is not a mathematical tautology vis-a-vis the underlying path coefficients. These models are said to be **over-identified**. Over-identified models can be meaningfully tested by comparing predicted and observed covariances because there is at least one variance or covariance that is not guaranteed to be perfectly reproduced by the underlying mathematics. The model in Figure 7.1 is an example of an over-identified because the correlation r_{13} is not guaranteed to be perfectly reproduced relative to the path coefficients in the model; the model can be disconfirmed if the observed value of r_{13} is discrepant from the value of $p_{12}\,p_{23}$. And in this case, the model was disconfirmed.

There also are situations where a model is **under-identified**. This occurs when a theorist specifies a model for which there is no unique solution for the path coefficients

and an infinite number of possible values exist for the path coefficients, all of which result in perfect reproduction of the observed variances and covariances. Such models cannot be uniquely tested in SEM and are problematic more generally because we do not know which of the many solutions that perfectly reproduce the correlation matrix to use. It is analogous to the problem in algebra where you are given the equation 6 = a + b and asked to specify the single, correct values of a and b that satisfy the equation. There are an infinite set of "correct" values for *a* and *b* and we do not know which ones to choose. If your model is under-identified, it turns out there are strategies you can use to still gain perspectives on the viability of the model, but they are tricky. I discuss these strategies in Chapter X.

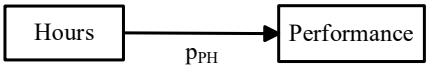
An important step in applying SEM is to know if the model you are working with is just identified, over-identified, or under-identified. Most SEM software, including Mplus (the featured software in this book), perform identification checks and will inform you of the identification status of your model. You then adjust your model evaluation strategy accordingly.

As some examples of identification status, here is an influence diagram for a model that is over-identified (I omit disturbance terms to avoid clutter):



The variable X is assumed to influence the outcome Y through the mediator M. Decomposition analysis finds that the correlation between X and Y should equal p_{YM} times p_{MX}, which is not tautological. Given there is at least one non-tautological correlation in this model, the model is said to be over-identified. Note that within this model, the correlation between r_{MY} is tautological because the p_{YM} path is set equal to this correlation. The same is true for the relationship between r_{TM} and p_{MT}. By contrast, a meaningful comparison in this model is between the predicted and observed covariances/correlations of r_{XY}.

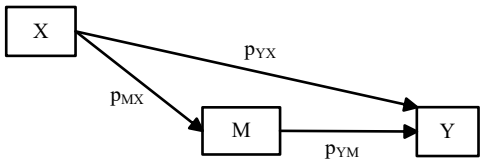
To provide an obvious example of a just identified model, consider an unrealistically simple two variable model that states that the number of hours studied for a test impacts one's test performance:



Suppose I obtain a measure of both of these variables and find a correlation of 0.30 between them. This correlation also is my estimate of the path coefficient p_{PH}. If you ask

me what evidence I have that this model is valid, I might answer that it is valid because the path coefficient perfectly reproduces the correlation between two variables. Of course, this is meaningless because I *defined* the path coefficient as the correlation. The model is just identified model and such a model "test" is meaningless. To be sure, the fact that the number of hours studied and test performance are correlated 0.30 is indeed consistent with my predictions and this is model affirmative. However, the fact that the path coefficient perfectly reproduces the observed correlation is in this case irrelevant.

Here is an influence diagram for another model that is just-identified:



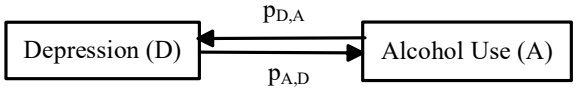
This model has two endogenous variables and, hence, two linear equations:

$Y = a_1 + p_{YM} M + p_{YX} X + d_1$

$M = a_2 + p_{YMT} X + d_2$

To test this model, I need to conduct two regressions, one for each of the above equations. Note that for this model every variable is connected to every other variable by a direct causal arrow. This usually is a recipe for a just-identified model. If I applied SEM to this model, I would find a zero residual matrix. However, such a model test is not meaningful because the mathematics of this model are such that it is guaranteed I will perfectly reproduce every observed correlation in the data.

Here is an influence diagram for an under-identified model applied to cross sectional data, i.e., measures of depression and alcohol use taken at the same time:



This model has only two variables in it, depression and alcohol use. The model asserts that depression and alcohol use measured at a given point in time are correlated because of two causal dynamics that have played themselves out prior to the assessment of the variables: (1) depression has increased the tendency for people to drink excessively and (2) drinking excessively has caused people to become depressed. There are two endogenous variables in this model, hence two equations that capture its essence:

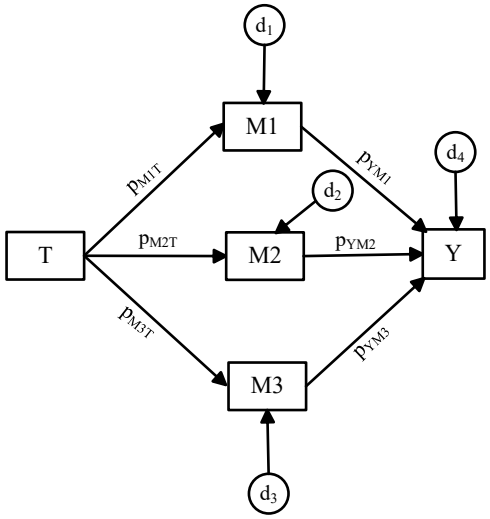
Alcohol Use = a_1 + $p_{A,D}$ Depression + d_1

Depression = a_2 + $p_{D,A}$ Alcohol Use + d_2

It turns out that a decomposition of $r_{A,D}$ would show that it equals $p_{A,D}$ times $p_{D,A}$. The problem is that there are an infinite number of pairs of values of the two path coefficients that will perfectly reproduce $r_{A,D}$. If $r_{A,D}$ (and hence $r_{D,A}$) equals 0.24, then one possible set of path values that will perfectly reproduce this correlation is $p_{A,D}$ = 0.8 and $p_{A,D}$ = 0.30. Another possible set of coefficient values that will perfectly reproduce 0.24 is $p_{A,D}$ = 0.60 and $p_{D,A}$ = 0.40. Yet another possible set of coefficient values that will perfectly reproduce 0.24 is $p_{A,D}$ = 0.70 and $p_{D,A}$ = 0.34. The model is under-identified in that there are many possible solutions and we have no idea which solution to choose. By contrast, for the just-identified models presented above, there is only one set of path coefficient values that will perfectly reproduce the observed covariance/correlation matrix despite the fact that the model estimation is guaranteed to produce a zero residual matrix. This is what makes those models just identified. It turns out I can augment the above under-identified model with additional variables to make it just identified or over-identified but how I do so is a topic for a different chapter.

MAXIMUM LIKELIHOOD ESTIMATION

Another way that SEM differs from traditional regression is the way in which it estimates the values of the intercepts and the path coefficients for the equations comprising the model. Consider the following model for an RET design where T is a dummy variable for the treatment condition, M1, M2 and M3 are mediators that the program targets, and Y is the outcome that is thought to be influenced by the three mediators:



This model has four endogenous variables and can be captured in four equations:

$M1 = a_1 + p_{M1T}\, T + d_1$

$M2 = a_2 + p_{M2T}\, T + d_2$

$M3 = a_3 + p_{M3T}\, T + d_3$

$Y = a_4 + p_{YM1}\, M1 + p_{YM2}\, M2 + p_{YM3}\, M3 + d_4$

In traditional regression frameworks, we use OLS to estimate values of the path/regression coefficients by conducting, in this case, four separate regression analyses, one for each equation. The path/regression coefficients are obtained one regression equation at a time. In contrast to this approach, full information SEM estimates the values of the path coefficients for all four equations *simultaneously* in a single analysis. This multivariate rather than piecewise approach to coefficient estimation can yield more efficient estimates of the path coefficients and it also lends itself well to global tests of model fit, as discussed below.

Traditional regression analyses select values for coefficients in a given equation by minimizing the sum of the squared differences between the predicted and observed scores for individuals in the sample (called the **ordinary least squares criterion**). Full information SEM, by contrast, uses algorithms to select path coefficient values that yield the closest possible correspondence between the predicted and the observed covariance matrices thereby taking into account all the model equations simultaneously. This is a different statistical orientation to coefficient estimation than traditional OLS regression although the results of the two frameworks often will be quite similar.

The full information SEM algorithms for selecting coefficient values use a "trial and error" process, also called an **iterative** process. An initial set of values for all the path coefficients is specified and then used to derive a predicted covariance matrix from the decomposition analyses. If the predicted and observed covariances are highly discrepant, then a different set of values for the path coefficients is tried to see if the prior correspondence can be improved upon. Then another set of values is tried, each time with the idea of improving model fit as one iterates through different sets of values for the path coefficients. The choice of values to use on a given iteration is not random. Rather, it is a very intelligent search process that quickly zeros in on the values that will produce the best possible correspondence between the predicted and observed covariances given the model. Once it is determined that the best possible correspondence has been achieved, the iterative process stops. The solution is said to have **converged**.

Minimizing the disparity between the predicted and observed covariances requires that we have a quantitative index of the degree of correspondence or "fit" between the predicted and observed covariance matrices that we seek to minimize during the iteration process. There are many candidates for such indices. For example, we could find path values that minimize the sum of the absolute values of the cells of the residual matrix. Alternatively, we could derive path values that minimize the sum of the square of the elements of the residual matrix. Statisticians have found it useful to minimize what is known as a maximum likelihood criterion that is defined in SEM using matrix algebra as:

$$F_{ML} = \ln|\Sigma'| - \ln|\Sigma| + \text{trace}[(\Sigma)(\Sigma'^{-1})] - k$$

where Σ is the observed covariance matrix, Σ' is the predicted covariance matrix by the model, *k* is the number of variables in the covariance matrix, ln is the natural log function and | | signifies the determinant of the matrix between the two bars.[3] This criterion is sometimes called the **maximum likelihood fit function** or the **maximum likelihood discrepancy function**. Although it appears formidable, it actually is straightforward if one knows matrix algebra (if you do not, you can skip the remainder of this paragraph and just accept the fact that it is a viable fit criterion). Consider the case where there is perfect model fit and Σ' equals Σ. In this case, the determinant of Σ will equal the determinant of Σ' and the difference between the logs of these determinants in the first part of the right-hand side of the equation will equal 0. Similarly, if $\Sigma = \Sigma'$, then $(\Sigma)(\Sigma'^{-1})$ in the equation is equivalent to $(\Sigma)(\Sigma^{-1})$. In matrix algebra, any matrix multiplied by its inverse yields an identity matrix, which is a matrix that has the same number of rows and columns of the matrix being operated on but with all 1s in the diagonal and all zeros in the off-diagonal. The trace function says to sum the diagonal elements, the result of which, for a perfect fitting model, will be the sum of the diagonal elements of an identity matrix, the value of which must be k. Subtracting k from this value yields zero. Thus, when there is perfect model fit, F_{ML} equals zero. As model fit to the data becomes worse, values of F_{ML} become larger.

The reason statisticians prefer to minimize F_{ML} as compared with other fit functions is that F_{ML} has many useful statistical properties. Specifically, it is possible to use the final value of F_{ML} that results after the solution converges during the iteration process in the sample to calculate estimated standard errors for each path coefficient in the model and to perform significance tests for the path coefficients. F_{ML} also can be used to define a variety of intuitive indices to evaluate overall model fit. I discuss these indices shortly.

Although F_{ML} has many desirable statistical properties, there are scenarios where it

[3] The determinant of a covariance matrix results from a complex set of operations that yield an index of multivariate variability. See Namboodiri (1984).

is ill-behaved (I elaborate some of these scenarios below). When this occurs, different minimization criteria typically are used by statisticians. Alternative fit criteria include unweighted least squares and weighted least squares, among others. Unweighted least squares minimizes the sum of the squared residuals in the residual matrix. Weighted least squares also minimizes the sum of the squared residuals in the residual matrix, but it gives an empirically determined weight to each squared residual before summing the residuals into an overall index. Thus, in weighted least squares, some squared residuals are given more weight than others. I generally do not use F_{ML}. Instead, I use a robust variation of it based on Huber-White estimation principles, as detailed later. See the resources page on my website for different fit functions options offered by Mplus, the SEM software I use in this book.

GLOBAL INDICES OF MODEL FIT

Because interpretation of F_{ML} is not intuitive, statisticians have developed additional indices of the correspondence between the predicted and observed covariance matrices to make it easier to evaluate a hypothesized model. These global fit indices give us a sense of how well a model fares in reproducing the observed correlations and covariances separate from F_{ML}. Over 30 such indices have been proposed and there is little agreement as to which index is best. In this section, I discuss the more popular indices.

When characterizing the fit between model predictions and data, there are three classes of fit indices that are often used. One class measures absolute fit by comparing in various ways the predicted versus observed variances and covariances. A second class uses absolute fit but includes a penalty function for lack of model parsimony. One can always improve fit by including additional paths in the model to the point where perfect fit is virtually guaranteed because the model becomes just-identified. The second class of indices penalizes this practice. The third class of fit indices compares the absolute fit of the model to a competing or alternative model that is either *a priori* specified or imposed arbitrarily on the data.

I now discuss exemplars of fit from each category. The current wisdom is that evaluation of data-model correspondence should use at least one fit index from each class. If good correspondence is suggested across diverse indices, then one has increased confidence in model viability. If the different fit indices provide different conclusions about model viability, then caution is warranted. This classification of indices may be different than other categorizations you encounter in the literature. One can get into endless debates about how the indices should be parceled, but in the final analysis, you want to work with multiple indices that view model-data correspondence from different perspectives.

The Chi Square Test of Fit

In the first category, one index of model-data fit proposed early in the evolution of SEM is what is called the **chi square fit index** or the **chi square statistic**. This is simply the sample value of F_{ML} at the final step of the iterative process (which I symbolize as $\hat{F}_{ML}$) times N-1, where N is the sample size. The chi square statistic is symbolized by χ^2 (some researchers use the letter T). When the $\hat{F}_{ML}$ is zero, the chi square index will be zero and perfect fit exists in the sample; there is no disparity between the predicted and observed covariance matrices. As $\hat{F}_{ML}$ increases in value, so does the chi square index, everything else being equal.

The magnitude of the chi square index *per se* lacks intuitive interpretation for most researchers, so, instead, they rely on a significance test associated with it. This test evaluates the null hypothesis that there is perfect fit between the predicted and observed covariances in the population, i.e., that the population residual matrix is all zeros versus an alternative hypothesis that the population residual matrix is not all zeros. Statisticians have shown that under the assumption of multivariate normality among observed variables in the model and with reasonably large sample sizes, the product of $\hat{F}_{ML}$ times N-1 has a chi square sampling distribution. The degrees of freedom for the sampling distribution varies depending on the number of observed variances, covariances, and means available for estimating parameters and the number of estimated parameters, but the value is routinely reported by SEM software and need not concern us now.

As an example, for a study that has k = 5 observed variables, there are a total of (0.5)(k)(k+1) = 15 variances and covariances in the data. These statistics might be used to estimate a model that has 7 free parameters. The model degrees of freedom is 15-7 = 8. If the model also estimates means and intercepts, then the sample means are brought into the mix. Once the degrees of freedom are known, one can compute a p value to evaluate the null and alternative hypotheses for the chi square test. If the chi square test is statistically significant ($p < 0.05$), then the null hypothesis of perfect model fit in the population is rejected and the model is called into question. If the chi square is statistically non-significant, then there may or may not be perfect model fit in the population. Given this, a model is often said to be viable if it yields a statistically non-significant chi square test, but even if this is the case, it does not mean the model is the correct population model; the test may lack statistical power to detect misspecification or an alternative causal structure may account for the data equally well.

Based on the above, the chi square index serves two functions. First, it is an index of fit between model predictions and data, albeit one that many researchers find difficult to interpret. The larger the value, the worse the fit, everything else being equal. Second, it is an omnibus model test that informs you, with assumptions, if the null hypothesis of a

population zero residual matrix can be rejected. For the family size-child IQ example, the chi square statistic for the model was 37.70, $p < 0.05$. This suggests an unsatisfactory model in terms of global model fit.

The chi square index and its associated test of significance have been criticized on several grounds. First, the statistic is not always chi square distributed for purposes of testing statistical significance, such as when sample sizes are small or for some forms of non-normal data. In such cases, the p values for it may not be accurate. I discuss in Chapter 28 attempts to re-scale the chi square statistic to be better behaved with small samples but there still remains the problem of low power for small sample sizes. Second, the p value associated with the chi square statistic tends to decrease with large sample sizes for models whose deviations from zero in the population residual matrix might be trivial. This can lead researchers to giving too much weight to trivial model-data disparities when evaluating a model. Relatedly, some scientists are uncomfortable with the idea that a model is considered viable only if it *perfectly* fits the population data. The argument is that reasonable models may not fit the population data perfectly but such models might still be useful approximations to the true underlying dynamics and worthy of consideration i.e., they are good enough. Bollen (1989) frames the argument this way: "*In virtually all cases we do not expect to have a completely accurate description of reality….The assumption that we have identified the exact process generating the data would not be accepted. Yet the chi-square test derives from a comparison of the hypothesized model to a model of perfect fit. A perfect fit may be an inappropriate standard, and a high chi-square estimate may indicate what we already know – that the null hypothesis holds approximately, not perfectly*."

My own view is that the chi square test has shortcomings but that a statistically significant chi square test (under reasonable conditions that conform to a valid test) constitutes a red flag that informs you that you need to dig deeper into model-data correspondence because something might be amiss. Perhaps after doing so, you will conclude that the magnitude and nature of the disparity from a zero residual matrix is substantively non-consequential and declare the data are reasonably model consistent. Or perhaps not. In the final analysis, however, we need to be cautious about dismissing red flags of ill fit without further probing and, given ill-fit, making a good case for why it is ignorable if you decide to move forward with the model anyway.

The Standardized Root Mean Square Residual

Another global fit index in the first category is called the standardized root mean squared residual, or more simply, the standardized RMR. The **standardized RMR** is an index that roughly reflects the average absolute discrepancy between predicted and observed

correlations. A standardized RMR value of 0.10 suggests, roughly, that the predicted and observed correlations deviate from each other by 0.10 correlation units, on average. The smaller the value of the standardized RMR, the better the model-data fit, with the smallest possible value being zero. Technically, the SRMR is not the same as the average absolute disparity between the predicted and observed correlations, which is known as the **correlation root mean squared residual** or CRMR (Bollen, 1989). I explain the difference between the two indices in the Appendix, but they often are close in value.

The standardized RMR in sample data is a biased estimator of the population standardized RMR; it tends to underestimate it over the long run (Kenny, 2018; Maydeu-Olivares, 2017a; Maydeu-Olivares, Shi & Rosseel, 2017). The bias is more pronounced for small N and smaller model degrees of freedom. If your N is large and your degrees of freedom are large, the bias is likely minimal. Maydeu-Olivares (2017a) suggested a bias correction for the SRMR and a method for calculating confidence intervals and significance tests for it (see also Maydeu-Olivares et al., 2017; Pavlov, Shi & Maydeu-Olivares, 2021). The approach is implemented in lavaan (see option lavResiduals) but not Mplus. For a discussion of the use of the SRMR statistic in Mplus, see Asparouhov and Muthén (2018). Asparouhov and Muthén (2018) recommend caution when using SRMR for $N < 200$. Some researchers suggest the SRMR should be at less than 0.05 as a fit standard while others suggest 0.08. These are rough rules of thumb that sometimes work well and other times not.

One problem with the standardized RMR and the CRMR is that they can, at times, paint what I think is a deceptively positive picture of the average disparity between predicted and observed correlations. The indices include in their calculation disparities that are true tests of model fit because they are not tautologically determined as well as disparities that are tautological and guaranteed to be zero. Inclusion of the latter drive the indices downward. I think a more accurate model test is to focus on the average absolute disparity between the predicted and observed values for non-tautological disparities only. My website has a program that allows you to do this, called *generalized RMR*. For the family size-IQ example, the standardized RMR as reported on Mplus output was 0.19, which includes both the tautological and non-tautological parameters. Compare this with the absolute disparity between the predicted and observed correlations for the one non-tautological comparison in the model. It was 0.51, which is quite a bit larger than 0.19.

The Root Mean Square of Approximation

In the second class of fit indices (absolute fit with penalty functions), the most commonly reported index is the **root mean square error of approximation** (RMSEA) and the test of close fit associated with it. The RMSEA (Steiger & Lind, 1980; Browne & Cudek,

1993) is formally defined in the population as

$$\varepsilon = (F_0 / df)^{1/2} \quad [7.1]$$

where F_0 is a generic discrepancy function reflecting disparities between the predicted and observed covariances and df is the degrees of freedom associated with the discrepancy function.[4] The population F_{ML} is typically used as F_0, so that ε is defined as

$$\varepsilon = (F_{ML} / df)^{1/2}$$

where F_{ML} is the population value of the maximum likelihood fit function. Conceptually, ε reflects the lack of fit per degree of freedom because F_{ML} is divided by the model df. The larger the value of ε, the more ill fit in the model per degree of freedom.

The above formula is the population representation of the RMSEA. In sample data, the computational formula for estimating the population RMSEA is

$$\text{RMSEA} = [((\chi^2 / df) - 1) / (N-1)]^{1/2}$$

where χ^2 is the sample value of F_{ML} multiplied by (N-1). This formula includes a correction factor to adjust for the fact that the sample F_{ML} is a biased estimator of the population F_{ML}. The smallest value a sample RMSEA can take is 0 (if it is less than 0, it is set to zero). It rarely exceeds 1.00 but it can be greater than 1.00. The presence of F_{ML} in the defining formula for the RMSEA makes it difficult to interpret on an intuitive level. In general, the smaller the value of the RMSEA, the better the global fit between the model and data everything else being equal. More parsimonious models have larger degrees of freedom, so the presence of *df* in the equation acts indirectly as a penalty function for lack of parsimony. James Steiger, who proposed the first variant of the RMSEA), argues that the RMSEA can roughly be thought of as a type of (standardized) root mean square residual somewhat analogous to the SRMR but now with a penalty function attached to it (see Steiger's SEPATH manual in TIBCO, 2020; see also Kline 2023). This is a reasonable way of thinking about the RMSEA index in some contexts (e.g., for simple models in which the correlations between variables are small to moderate in size) but there are scenarios (e.g., large correlations between variables in complex models) where such an interpretation becomes dubious (see Steiger, 2000; Saris et al., 2009; Marsh et al., 2004; Savalei, 2012).

Browne and Cudek (1993) suggest that, as a rough rule of thumb, RMSEA population values less than 0.08 reflect adequate model fit, values less than 0.05 imply good model fit, and values less than 0.01 represent excellent fit. These values have

[4] Raising a quantity to the ½ is the same as taking the square root of it.

become “rules of thumb” that often are invoked when evaluating model fit. However, based on extensive simulation work, Chen et al. (2008) concluded that “any effort to identify universal cutoff points for the RMSEA is not supported and should not be pursued as a single way of assessing model fit” (see also the simulations by Savalei, 2012). I revisit this point below when I suggest a weight of the evidence perspective for model evaluation.

Browne and Cudek (1993) argue that the p value for the chi square test of model fit is too stringent because it tests a null hypothesis of perfect model fit. Framed in terms of the RMSEA index that uses F_{ML}, the traditional chi square test essentially evaluates the following null and alternative null hypotheses:

H_0: $\varepsilon = 0$

H_1: $\varepsilon > 0$

Browne and Cudek (1993) devised an inferential test for a "close" fitting population model rather than a perfectly fitting model, where “close” is defined as a population RMSEA value of 0.05 or less. The underlying idea is that sometimes a model applied to population data may not perfectly reproduce the observed population covariance matrix, but it might come very close to doing so, close enough that the model can still provide useful insights into the phenomena being studied. It is analogous to me telling you the distance between the cities of Los Angeles and New York using a new measuring device and being off by 1 inch. Technically, the device is inaccurate. But the degree of error is so small that it functionally does not matter. In the Browne and Cudek approach using RMSEAs, the null and alternative hypotheses for the test of close fit are

H_0: $\varepsilon \leq 0.05$

H_1: $\varepsilon > 0.05$

where ε is the population value of the RMSEA (do not confuse the 0.05 in the above expressions with p values; they are RMSEA population values). Just like the chi square test yields a p value for the null hypothesis test of perfect fit, the **test of close fit** also yields a p value but now for a null hypothesis of a close fit. If the p value associated with the test of close fit is non-significant ($p > 0.05$), then this is consistent with (but does not prove the presence of) a close fitting model in the population as defined by an RMSEA of 0.05 or less. A statistically significant p value for the test of close fit ($p < 0.05$) leads one to reject the null hypothesis that $\varepsilon \leq 0.05$ and one concludes the population model does not yield a close fit to the population data.

As an example, suppose when evaluating an RET model, I apply the traditional chi square test and find a statistically significant p value for it, $p < 0.05$. This might lead me to reject the viability of the model because it does not *perfectly* reproduce the population covariance matrix. I might also evaluate the test of close fit and find that the p value for it is statistically non-significant ($p > 0.05$). This means that the model *could* be close fitting, so perhaps we should not reject it even though it does not perfectly reproduce the population covariance matrix.

By contrast, suppose that both the chi square test and the test of close fit yield p values less than 0.05 leading both tests to reject their respective null hypotheses. This would suggest the RET model neither perfectly accounts for the population covariance matrix nor does it even come close (as defined by a population RMSEA < 0.05) to accurately predicting the population covariance matrix. I might therefore reject the model.

A problem with the above logic is that it assumes that an RMSEA of 0.05 is a reasonable standard for defining a close fitting model. This is controversial which makes the test controversial. The RMSEA has other shortcomings as well. A property of the sample RMSEA is that it is sample size dependent, i.e., it generally decreases as sample size becomes larger, a property that critics object to (Kenny et al., 2015). Also, the impact of the model degrees of freedom on the value of the RMSEA sometimes is counter-intuitive (Kenny et al., 2015). Kenny et al., (2015) suggest that if a model has more than 4 degrees of freedom and N > 175, then the RMSEA likely is reasonable to consider as a model-data fit index. However, they caution against use of the index in conjunction with traditional rules of thumb when model df are less than 5 and with smaller N; the strategy tends to over-reject reasonably fitting models in such cases. For example, a nonsignificant chi square of 2.10 for a model with df=1 and N = 70 yields an RMSEA of 0.13 despite a statistically non-significant chi square value.

A useful feature of the RMSEA is that methods have been developed for calculating confidence intervals for it. A 90% confidence interval often is evaluated given the one-tailed nature of significance tests for the RMSEA (e.g., we evaluate the null hypothesis that ε is zero relative to the alternative hypothesis that $\varepsilon > 0$). MacCallum et al. (1996) proposed a third type of test of model-data fit that makes use of the RMSEA confidence interval called the **test of not close fit**. In this test, the model is rejected if the upper limit of the 90% confidence interval is less than a cutoff value. For example, if one defines an acceptable model as one that has a population RMSEA < 0.08, then one would evaluate if the upper limit of the sample RMSEA 90% confidence interval is less than 0.08. If it is, one is confident that the population RMSEA is indeed less than 0.08 and concludes the fit of the population model is acceptable. Maydeu-Olivares, Shi & Rosseel (2017), however,

raise concerns about the accuracy of RMSEA confidence intervals when models have a large number of variables (>30) and when the data are non-normal. They argue for the use of an unbiased SRMR statistic and its associated confidence intervals as a basis for testing model-data fit (but see Pavlov et al. 2021 for qualifications to their arguments).

For the family size-IQ example, the RMSEA was 0.60 and the test of close fit yielded a p value <0.05. These results raise questions about model adequacy. The 90% confidence interval for the RMSEA was 0.45 to 0.78. Because the upper limit of 0.78 is larger than a cutoff of 0.08, the test of not close fit also indicates a suspect model.

In sum, although the RMSEA is quite popular, its interpretation is not straightforward and can be impacted by incidental parameters and model complexity. In some cases, the RMSEA is roughly analogous to a standardized root mean residual with a penalty function but not always. Rules of thumb for evaluating model fit using RMSEA standards have been suggested, but they are controversial. A strength of the RMSEA is that it can be adapted to perform a test of close fit and one can calculate confidence intervals. However, I personally find it to be a somewhat difficult index to work with.

The Comparative Fit Index

In the third class of model-data fit indices, we compare two models. A popular index is called the **comparative fit index** (CFI). It compares model-data fit of the target model with the fit of a competing model known as the **independence model** or the **null model** (also sometimes called the **baseline model**). The independence model is defined somewhat differently depending on the model analyzed but it typically specifies a zero correlation between the exogenous-endogenous variables and between the various endogenous variables. Such a model is not realistic in most research situations, so one expects that a "good" model will be far more in accord with the data than the null model. The CFI ranges from 0 to 1.0, with larger values implying the target model fits the data better than the independence model. A CFI of 0.90 means, roughly, that the target model improved fit by 90% relative to the independence model, after adjusting for model complexity. A CFI of 0.70 means, roughly, the target model improved fit by 70% relative to the independence model. A rule of thumb often suggested is that models with a CFI less than 0.95 are suspect.

The formal definition of the CFI is as follows: Let d_M = the χ^2 value for the target model minus its degrees of freedom and d_I = the χ^2 value for the independence model minus its degrees of freedom. Then

$$CFI = (d_I - d_M) / d_I \qquad [7.2]$$

If the index is greater than one, it is set to one. There are two noteworthy features of the

CFI. First, it incorporates a penalty function for model complexity by virtue of the inclusion of the model degrees of freedom into the formula. In this sense, it is a hybrid between indices that include penalty functions and those that focus on comparative fit. Second, the CFI reflects the improvement in fit of the model to data compared to the independence model (the numerator in the above equation) scaled against the lack of fit of the independence model (the denominator). This is a common way in the statistical literature of indexing improvement rates. For example, if a medication reduces the incidence of an adverse medical condition from 20% to 15%, then it is often said to have improved the incidence by (20-15)/20 = 0.25 or 25%.

It turns out that the CFI is affected by the size of the correlations between variables in the population. If all of the population correlations between variables are, in fact, low, then the independence model actually will fit the data well and the target model cannot improve much on it. Based on simulation work, Kenny et al. (2015) recommend not using the CFI if the RMSEA for the independence model is less than 0.16, because the CFI will be artificially low (but see van Laar and Braeken, 2021, for qualifications). I provide on my website a program for calculating the RMSEA value for the independence model (see the program called *CI for RMSEA*) so that Kenny's recommendation can be applied, if desired. The CFI also can be impacted by the type of independence model used in its calculation; for details. See van Laar and Braeken (2021).

Parenthetically, there are two other indices of relative fit that use the same structure of Equation 7.2. One index is the **Tucker-Lewis Index** that defines the d in Equation 7.2 as the chi square divided by its degrees of freedom rather than the chi square minus its degrees of freedom, while also subtracting 1 from the d_I in the denominator. This formulation imposes a harsher penalty for model complexity than the CFI. Another index is the **Bentler-Bonnet Index** (also known as the **Normed Fit Index**) that omits the model degrees of freedom from Equation 7.2 and defines the respective d as each model's chi square. As such, the BBI imposes no penalty function for model complexity.

For the family size-IQ example, the CFI was 0.32, which suggests ill model fit.

Fit Based on Information Indices

Another group of fit indices in the relative fit category is based in statistical information theory (Burnham & Anderson, 2004). The most well-known indices are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). You will not encounter these indices as frequently in the SEM literature because they are not well suited to characterizing fit of a single model. However, they are viable when formally comparing one model to another model. I consider them later in the chapter.

Summary of Global Fit Indices

In sum, a model that yields a statistically non-significant chi square statistic, a small standardized RMR value (say, less than 0.05), a small RMSEA value (say, less than 0.08), a non-significant test of close fit, and a large CFI value (say, greater than 0.95) is likely - but not guaranteed - to be a reasonable fitting model to the data. Meeting these standards does not make a model "correct" because (a) the rules of thumb associated with most of the fit indices are context dependent, and (b) other models might account for the data equally well or better (Hayduk, 2014). Rather, at a global level, a pattern of results along the above lines is suggestive that the model is reasonably consistent with the data. All of the fit indices have weaknesses and one needs to take these into account when evaluating a model. For example, Browne et al. (2002) describe cases where models provide relatively good fit to data but yield large chi squares and unfavorable global fit indices. McNeish, An, and Hancock (2017) describe scenarios where global fit evaluations of different types of measurement models can be misleading. The work of Brown et al., McNeish et al., and others underscores that there is more to evaluating model viability than just examining global fit indices and global tests of fit. For a useful exposition of the technical aspects of the fit indices, see Bollen (2026).

Some methodologists argue that model evaluation should prioritize the chi square test of fit because otherwise one can declare a misspecified model as reasonable. The other global fit indices, the argument goes, start with the premise that some degree of misfit between model and data is "permissible," but how much misfit is permissible is subject to debate and context dependent. Hayduk et al. (2007) object to the use of rules of thumb for the RMSEA, CFI, and SRMR for declaring a model as viable arguing that researchers need to be sensitive to *any* model-data disparities that arise in empirical evaluations. Statistics other than the chi square test, the argument goes, provide wiggle room for ignoring potentially meaningful misfit. Of course, one might ultimately conclude that one or more model-data disparities is chance induced or substantively trivial, but the point is that one usually needs to dig deeper beyond the global fit indices to determine this. I return to this issue below.

LOCALIZED FIT INDICES

In addition to global fit indices, SEM programs routinely provide feedback about model-data correspondence for specific portions of the model. I call these **focused indices of model fit** or **localized fit indices**. These indices are important because they help pinpoint where ill fit is occurring within the broader model. One such statistic is called a **modification index**. A modification index is provided for path coefficients or other

model parameters that have been fixed at some *a priori* value, usually zero, i.e., for paths or parameters omitted from the model. Consider our original model on family size and IQ that omitted the path with the dashed arrow shown in Figure 7.4. A modification index for the dashed path tells us how much the overall chi square value for model fit will likely decrease if the path is not omitted (i.e., not fixed at zero) but rather included in the model. When I calculated the modification index for the dashed path coefficient using Mplus software (which I describe in more detail in Chapter 11), it equaled 31.40. In other words, adding this path to the model will decrease its "badness of fit" by approximately 31.40 chi square units. A modification index of approximately 4.0 or larger signifies that if the path is added to the model, its coefficient likely will be statistically significant ($p < 0.05$) were the model to be re-estimated with the path. In the family size and IQ model, it seems that the original model is deficient because it left out this crucial path. A good fitting model will not only yield satisfactory overall fit indices, but will also have modification indices that all are small. Stated another way, a good fitting model will not omit paths or parameters that should have been in the model. Most SEM software reports modification indices for a model to help evaluate that model at a localized level.

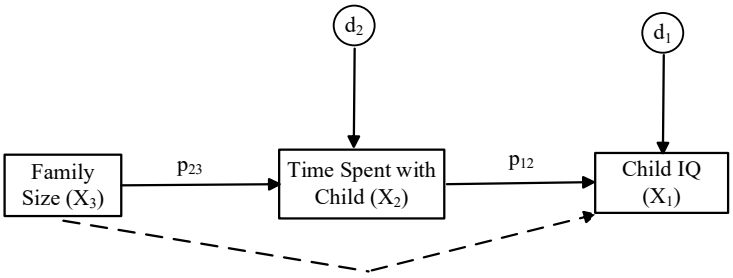


FIGURE 7.4. Illustration of model modification

Just as chi square values can be impacted by sample size, such is the case for non-zero modification indices: The larger the sample size, the larger the (non-zero) modification index will be, everything else being equal. With large sample sizes, it is not uncommon to observe one or more modification indices larger than 4. Like the chi square statistic, I use modification indices greater than 4 as warnings that something in the model might be amiss. In these cases, I carefully think about the omitted parameter called to my attention by the large modification index and whether adding it to the model makes conceptual sense. I also keep in mind that models that have a large number of zero or fixed parameters can yield some values larger than 4 just by chance. I show you in

Chapter 11 how to take these considerations into account when applying SEM to RET models. Zengh and Bentler (2023) have recently proposed a formal test for omitted paths to complement traditional modification indices and that is based on the integration of bootstrapping and Wald tests. However, the method still requires further evaluation.

A second localized index of model-data fit is the **standardized residual**. A standardized residual is the difference between a given covariance in the input data matrix and its predicted covariance by the model with this difference divided by the estimated standard error of the difference. It is analogous to a statistical test of the difference between the predicted and observed variances/covariances for each cell of the residual matrix, with the value of the standardized residual representing a z test of this difference. If an absolute value of z is larger than 1.96, this means the null hypothesis of a zero difference between them is rejected at $p < 0.05$. This ill fit can usually be traced back to the part of the causal model where the two variables reside, often to an omitted path or an omitted parameter in the model. As such, standardized residuals can be diagnostic of the same model error as modification indices, but not always.

For the family size-IQ example, I repeat from Table 7.2 the observed correlations, the predicted correlations and the residual matrix that is the difference between the predicted and observed correlation matrices:

Observed Matrix			Predicted Matrix			Residual Matrix		
1.00	-.30	-.60	1.00	-.30	-.09	0.00	0.00	-0.51
-.30	1.00	0.30	-.30	1.00	0.30	0.00	0.00	0.00
-.60	0.30	1.00	-.09	0.30	1.00	-0.51	0.00	0.00

The standardized residual for Y and X1 (corresponding to a correlation disparity of -0.51) was -4.70. This value is a z ratio for a z test that compares the predicted and observed covariance for Y with X1. (Note that if the covariance difference between Y and X1 is statistically significant (and hence non-zero), then by definition the *correlation* difference also must be statistically different). Because the absolute value of 4.70 is greater than 1.96, the p value for the null hypothesis of no difference between the predicted and observed covariance/correlation for Y with X1 is less than 0.05. There is a statistically significant difference between the predicted and observed covariance for this particular cell of the covariance matrix. In general, a good fitting model will yield small absolute standardized residuals for every variance and covariance in the model, less than 1.96 or, roughly, 2.0. Most SEM software reports standardized residuals for model evaluation.

Non-zero standardized residuals also can be impacted by sample size and test

multiplicity, so these properties must be considered as well. There are technical issues associated with the calculation of standardized residuals that I do not want to get side tracked on here. I discuss the issues in Chapter 11. For now, the above discussion captures the general spirit of these localized fit tests.

A third approach to identifying ill fit at the level of specific parameters is to examine the values of the parameters yielded by the SEM software. If the parameters take on values that do not make statistical sense (e.g., correlations greater than 1.0, negative variances) or do not make substantive sense, then the model is questioned.

Mplus recently (version 8.12 and higher) added two additional localized indices of fit. The first is straightforward and is simply the difference between the predicted and observed correlations on a cell-by-cell basis for all possible pairs of observed variables. These are called **correlation residuals** and should all be near zero. If when evaluating a model I observed a correlation residual that is large relative to a value of zero, then this is a flag that the model is having difficulty reproducing that correlation. I would then revisit the model and try to figure out why this is the case. The second index is analogous to the correlation residual but is based on partial correlation differences. Specifically, Mplus calculates the partial correlation between a each pair of variables partialling out all other variables in the model. It does so twice, first using the observed sample correlation matrix and then again using the predicted correlation matrix. Mplus then differences these two partial correlations with the idea that a good fitting model will produce a difference near zero. Mplus performs this localized test for all possible pairs of observed variables in the model and refers to the statistic as a **partial correlation residual** (Asparouhov & Muthén, 2025). As a general practice, you should scan both of these residual matrices and have more confidence in a model where disparities are small. Note that the units of the residuals are correlation units in both types of matrices. The definition of a "large" discrepancy depends on the substantive area and the overall absolute value of the correlations themselves; a discrepancy of 0.05 for correlations of 0.90 and 0.85 is qualitatively different than one for correlations of 0.06 and 0.01. Also, some disparities are large by chance, so take this into account as well.

EVALUATION OF PREDICTED PATHS

Modification indices evaluate a model by exploring whether omitted paths/parameters are, in fact, empirically reasonable to exclude. The obverse of this localized fit criterion is that the paths that are included in the model are empirically supported by the data, i.e., the paths that one predicts should be statistically significant, in fact, are statistically significant. This criterion also is used for evaluating models that are just-identified because such models are assured perfect fit to the data via global and localized fit indices.

If your model predicts a path should be statistically significant but it is not, then this questions the viability of your model as originally formulated.

A WEIGHT-OF-THE-EVIDENCE PERSPECTIVE

In sum, a reasonable model not only is well behaved in terms of global tests and global descriptive fit indices but also in terms of modification indices, standardized residuals, correlation residuals, partial correlation residuals, making both substantive and statistical sense, and whether the hypothesized paths in the model are empirically affirmed. I like to think of model evaluation using the concept of "weight of evidence" (Grace, 2020). I think about (1) the prior evidence for the model, (2) the fit of the model to the data as reflected by an appropriate chi square test (if available) and by examination of diverse global fit indices, (3) the fit of the model to the data as reflected by localized fit indices, (4) whether the *a priori* predicted paths of the model are statistically significant and non-trivial in magnitude, (5) whether the model makes substantive sense, and (6) given viable competing models, whether the model performs better than or as well as those competing models. I consider the weight of the evidence across these criteria and then form a judgment about model viability.

Model evaluation is much more than examining if model-data correspondence indexed by global fit indices satisfies a set of rules of thumb. It is often subjective and requires evaluation of model fit from multiple perspectives. Researchers disagree about how much weight to give to different criteria. The global fit indices reflect differences between an observed covariance matrix and the model-implied covariance matrix taking into account the multifaceted nature of these differences. Each global index comes at the matter of describing disparities from different perspectives, no one of which is "correct."

There are some researchers who prioritize the chi square statistic and who discount the other fit indices as arbitrary, but I personally believe the chi square test also has a certain degree of arbitrariness. Its sampling distribution can be distorted based on small sample size and certain forms of non-normality; it uses a somewhat arbitrary criterion p value to declare model misfit (usually $p < 0.05$); it may lack statistical power; and, even if everything falls into place for a valid chi square test, it only tells you if the fit of the model is not exactly perfect. It provides limited perspectives on the magnitude of model-data disparities. Also, the chi square test evaluates the overidentifying restrictions of the model which some would argue is too narrow a focus. I personally am not content to rely primarily only on the chi square test given such shortcomings. To be sure, a statistically significant chi square statistic is a red flag that I should look at model fit more closely. But I typically use a more comprehensive set of criteria to make judgments about model viability. The closeness of the sample covariance matrix to the predicted covariance

matrix needs to be examined from different vantage points. The various global fit indices provide ways of doing so. Examination of local fit indices provide yet additional perspectives. And, I always keep in mind that sometimes inaccurate models can produce perfect model-data fit (e.g., for the bivariate case, the true function might be X→Y, but the model Y→X will perfectly reproduce the data).

As noted, rules of thumb have been suggested for most of the global fit indices but the various rules of thumb are controversial because of their context dependent nature (Nye & Drasgow, 2011). West et al. (2023) argue that the rules of thumb can be viewed as *rough guidelines* for overall correspondence between the observed versus model implied covariances but that they should not be reified. McNeish (2021, 2022, 2023) builds on the work of Milsap (2007, 2013) to propose context sensitive strategies for defining model rejection cutoffs of fit indices using localized Monte Carlo simulations. McNeish's method is intriguing but it also has shortcomings, including its reliance on multivariate normality assumptions and its atheoretical incorporation of specification error. More research on his approach is needed. In the final analysis, model evaluation is a complex enterprise that requires multiple considerations from multiple perspectives.

I conclude this section by summarizing Kline (2023) on evaluating model fit:

1. If you use a simultaneous estimation method, examine the model chi-square with its degrees of freedom and p value. If the model fails the chi square (exact-fit) test by yielding a p value less than 0.05, then tentatively reject the model. Then, diagnose both the magnitude and possible sources of model misfit by inspecting local fit and other diagnostics. If there are slight discrepancies between the predicted and observed data that appear "inconsequential" or that are theoretically vacuous, rescind rejection but be sure you can explain and justify your decision.

2. If the model passes the exact-fit test, you still must inspect local fit. If local fit indicates non-trivial discrepancies, then reject the model. Among the local fit indices, examine tests of specific predicted versus obtained covariances and modification indices.

3. If you report values of approximate fit indexes, then include at a minimum the RMSEA, CFI, and SRMR. Do not try to justify retaining the model by depending solely on rule of thumb thresholds for these indices. Take a more comprehensive approach.

4. If you respecify your model, explain why and the diagnostics that led you to do so.

5. Statistical evidence of model fit is not the sole factor in deciding whether to retain a model. The model estimates must make sense and the model should not be "overfit."

6. If a model is retained, then you should explain why that model is preferred over viable, alternative models that explain the same data as well or nearly as well. If you cannot, then

view these alternative models as viable.

7. If no model is retained, then try to figure out why and address it in future research.

LATENT VARIABLES IN SEM

A strength of SEM is that it can accommodate latent variables and measurement theories. I introduced this idea in Chapter 3 and elaborate it here. Suppose I want to evaluate the causal influence of maternal depression on adolescent depression. I obtain three measures of maternal depression, (1) the CES-D measure, (2) the Beck measure, and (3) the PHQ-9 measure, all well-known depression scales. I use the same three measures to assess depression in the adolescent child of the mother. Let MD1, MD2, and MD3 represent the three depression measures for the mother and AD1, AD2, and AD3 represent the three depression measures for the adolescent. Let LMD represent latent maternal depression and LAD represent latent adolescent depression. The causal model that I think underlies the correlations between the six measures is in Figure 7.5. I signify paths from the latent variable to the indicators with Ls to correspond to (unstandardized) factor loadings. Each measure has an error term associated with it. As noted in Chapter 3, this is an explicit recognition that the measures are not perfect and likely contain measurement error.

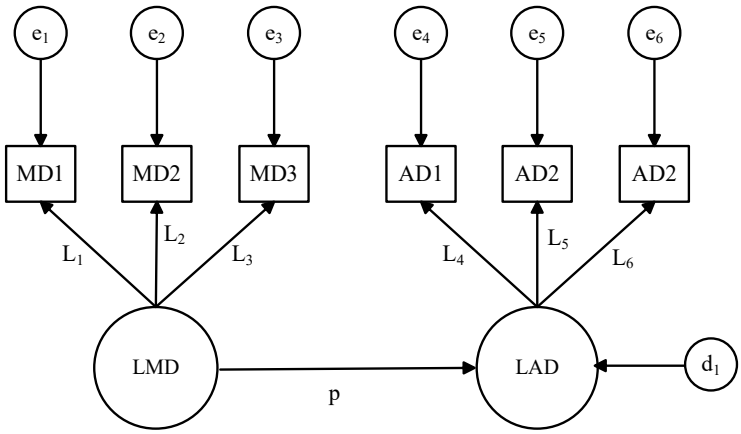


FIGURE 7.5. Latent variable model

The causal path from the latent maternal depression to latent adolescent depression is what I am most interested in substantively. The path coefficient that the SEM software estimates for this path will take into consideration all of the information available in the model and also will make adjustments for the modeled measurement error. Thus, in SEM

we combine both a measurement theory and a structural theory into one encompassing model of simultaneous equations. This is a strength of SEM, but it also raises additional complexities. For example, if a measurement theory is misspecified, then the misspecification can lead to inaccurate inferences about the properties of the measures, such as their reliability and validity. Misspecification also can reverberate to other model parts and lead to faulty inferences about structural coefficients in the model.

In total, there are six observed measures in the above model, which yield a 6X6 covariance matrix. My guiding hypothesis is that the patterning of the covariances in this matrix is due to the causal dynamics in Figure 7.5. I want to test this hypothesis and, if the model is supported by the data, examine the parameter estimates for the model. I test the model by comparing the observed covariance matrix with the predicted covariance matrix that is generated by a decompositional analysis of each of the covariances. The approach to the analysis is much the same as the simplified model in Figure 7.1. First, I translate the diagram into a set of linear equations using the same heuristics as before: (a) I identify all the endogenous variables and (b) I treat each endogenous variable as an outcome with predictors represented by all variables that have an arrow pointing directly to it. There are 7 endogenous variables, MD1, MD2, MD3, AD1, AD2, AD3 and LAD. LMD is an exogenous variable. Here are the equations for each of endogenous variable, using sample notation:

$MD1 = a_1 + L_1\ LMD + e_1$

$MD2 = a_2 + L_2\ LMD + e_2$

$MD3 = a_3 + L_3\ LMD + e_3$

$AD1 = a_4 + L_4\ LAD + e_4$

$AD2 = a_5 + L_5\ LAD + e_5$

$AD3 = a_6 + L_6\ LAD + e_6$

$LAD = a_7 + L_7\ LMD + e_7$

An unusual feature of each of these equations is that the predictor variables are unmeasured latent variables. Statisticians have creatively developed methods for estimating the path coefficients and error variances in these equations as well as a predicted covariance matrix for the observed measures to compare against the observed measure covariance matrix. SEM analysis of this type also yields the various indices of model fit describing the correspondence between the predicted and observed covariance

matrices, namely the chi square test, the RMSEA, the CFI, and so on. As before, we compare the correspondence between the predicted and observed covariances using both global fit indices and focused fit indices. I illustrate all of this for an RET in Chapter 11.

One matter that must be addressed before latent variable modeling can be pursued is the assignment a metric to the latent variables. On what metric is the latent mother depression scaled. Should scores on it range from 0 to 10, 0 to 100, 1 to 5, or what? Statisticians have made several suggestions for assigning a metric to a latent variable in SEM. The most common approach is to pass the metric of one of the indicators of the latent variable to the latent variable. Suppose the C-ESD scale, one of the indicators of maternal depression (MD1 in this case), is a measure that I am quite familiar with and whose metric is meaningful both to me and other researchers in my field. The CES-D ranges from 0 to 60 with most people having low scores on it, near 5 to 10. Higher scores indicate greater levels of depression. My desire is to score the maternal latent variable on this same metric, but to adjust it for measurement error. The strategy that statisticians use to accomplish this is to work with L1 in the above equations, namely the factor loading linking the maternal latent depression construct to the observed C-ESD measure, MD1. Instead of estimating L_1 when evaluating the measurement model, I instead tell my SEM software to force L_1 to equal 1.0 (i.e., I "fix" it at the value 1.0), so the equation becomes

$$MD1 = a_1 + 1.00\ LMD + e_1$$

Note that, based on this assigned coefficient value, for every one unit that LMD changes, MD1 changes, on average, by 1.0 unit, i.e., the slope in the linear equation equals 1.0. This fixing of the loading for the observed C-ESD to 1.0 puts the two variables on a comparable metric. Think of it like currency exchanges. If one Euro equals one US dollar and if a company charges a $5 service fee to exchange currencies, then when I change dollars to Euros, the equation for how many Euros I get is

Euros = -5.00 + 1.00 Dollars

and if I change Euros to dollars, it is

Dollars = -5.00 + 1.00 Euros

Dollars and Euros are on the same metric by virtue of the slope of 1.0. A 10 unit increase in dollars given to the agency leads to a 10 unit increase in the Euros I receive and vice versa. When I force SEM software to set the slope for MD1 on LMD to 1.0, the metric of MD1 essentially is "passed" to LMD. This means I can interpret LMD as if it was scored like the C-ESD scale but with an adjustment for measurement error.

The measure used to define the metric of the latent variable by fixing its loading to 1.0 is called the **reference indicator** or **marker variable**. We usually choose it based on how intuitive the metric is and its psychometric properties (reliability, validity). Although we lose information about the true value of L_1 by passing the metric of the reference indicator to the latent variable, we gain that information back, somewhat, in other facets of the analysis, as I explain in Chapter 11. With the use of the reference indicator strategy, the metric of the maternal depression latent variable is essentially the same as that of the reference indicator. However, the variance of LMD will usually be smaller than the variance of MD1 because it is corrected for measurement error and absent the random noise that influences MD1 (see Kline, 2023, and Chapter 11).

Suppose I use this method to set the metric of LMD to that of the CES-D and I do the same for the metric of AD1 to set the LAD metric to that of the CES-D for adolescents. The regression coefficient for LAD onto LPD will then be as if I regressed the CES-D scale for adolescents, adjusted for measurement error, onto the CES-D scale for parents, adjusted for measurement error.[5]

There are two other methods that are sometimes used to define the metric of a latent variable. I do not delve into them here, but want to at least mention them. One strategy does not fix the value of one of the factor loadings, L, but instead fixes the variance of the latent variable in question to 1.0, much like a standardized score. This is called the **fixed factor method**. It has advantages in some modeling scenarios but it is not viable when the latent variable is endogenous and, hence, it is somewhat limited. This is because the variance of a latent endogenous variables depends, in part, on the variances of the variables in the model that are specified as determinants of it and this must be taken into account accordingly. A third method for defining the metric of a latent variable is called **effects coding**. Effects coding requires that all indicators of a latent variable be on the same metric, which often is unrealistic. For details about this method, see Little, Siegers, and Card (2006). Most researchers use the reference indictor approach, ideally with metrics that are meaningful, and that is what I will do in this book. For additional discussions of criteria to consider for choosing reference metrics, see Bollen et al. (2022).

In sum, a strength of SEM is that it can model latent variables to allow us to take into account measurement error when estimating causal relationships. This is not true of traditional regression methods. I illustrate how to model latent variables for an RET in Chapter 11. For now, I merely want to establish that latent variables can be modeled and to introduce you to the need to assign metrics to latent variables when doing so.

[5] Some psychometricians conceptualize the variance of a measure as its metric. In this framework, the latent variable and the reference indicator do not strictly share a common metric because the latent variable variance is smaller by virtue of the elimination of measurement error variance from it. Nevertheless, it still is reasonable to think of the latent variable metric as roughly mapping onto the scaling of the reference indicator and interpreting it accordingly.

THEORY REVISIONS BASED ON DATA

If I fit an RET model to data I have collected for purposes of program evaluation but I obtain poor model fit, what do I do next? There is controversy about how to deal with such scenarios. SEM is traditionally viewed as a confirmatory rather than exploratory enterprise: A theory is specified *a priori*, data are collected to test that theory, the theory makes predictions about how the data should pattern themselves, and then based on the results, the theory is either rejected or declared to be consistent with the data. Faced with a poor fitting model, analysts often examine model diagnostics within the data to identify the sources of ill fit. The idea is then to modify the model so that the revised theory is now consistent with the data. Inferences are then made accordingly.

Modifying a model based on the same data used to test that model is seen by many as turning the SEM enterprise into an exploratory rather than a confirmatory approach. Critics argue that this is inappropriate because of its reliance on chance, among other things. They argue against any form of post hoc model revision and criticize those who engage in it. In the context of program evaluation, if I formulate an RET model that I think operates and I find a poor fitting model after an initial SEM analysis, I assure you that after spending a substantial sum of the funder's money and expending considerable effort to conduct the program evaluation, I am not going to walk away from the data and say "sorry, I can't conclude anything because I obtained a poor model-data fit for my *a priori* model." In program evaluation, a rigid conceptualization of SEM as purely confirmatory is unrealistic.

I discuss in this section issues surrounding (a) adding paths or parameters post hoc to poor fitting models in order to improve model-data fit and (b) trimming non-significant paths of good fitting models to make them more parsimonious. In the SEM literature, **forward searching** for model revision starts with a poor-fitting model and attempts to find a "correct," well-fitting model, usually by adding paths or parameters to it. **Backward searching** refers to the process of trying to obtain a more parsimonious model from an initially well-fitting model, usually by eliminating statistically non-significant paths, a process also known as **theory trimming**. MacCallum (1986) has suggested four ideals to increase the likelihood of model re-specifications that better capture the true population dynamics while retaining parsimony: (1) carefully think through one's initial model, (2) invoke substantive considerations when making modifications, (3) use large sample sizes, and (4) continue forward searching even if a good fitting model is suggested by global fit indices.

Forward Searching

One of the most common approaches to forward searching is that of altering an ill-fitting model by adding parameters based on the inspection of modification indices. When faced with a modification index that is larger than 3.84 (or, more roughly, 4.00), a decision must be made about whether to add the flagged path/parameter. Obviously, if the parameter that is associated with a large modification index makes no conceptual sense, I would not include it in my model (but for exceptions, see Chapter XX). SEM software has no idea about the substantive content of my variables nor of the theory surrounding those variables. It focuses only on the mathematics of the numbers it is provided with. It is up to me to introduce common sense when evaluating a modification index and what to do about it if it is large. I like to think of a large modification index as identifying a potential "point of stress" in the model where the model is having difficulty accounting for an association between two variables. The model is asking me to add a path or a correlated disturbance between the variables to allow the model to better account for the association between them. However, I ultimately must decide the most appropriate way to address that point of stress. If the path or correlated disturbance makes conceptual sense, I might add it. If it does not, I usually would not.

Often there are dependencies in modification indices. Stated another way, there can be more than one way to resolve a point of stress in a model. Consider the RET model in Figure 7.6. Suppose this model does not fit the data well. Upon inspection of modification indices, I might observe a large modification index for each of three parameters, (1) adding a causal path from Outcome 1 to Outcome 2, (2) adding a causal path from Outcome 2 to Outcome 1, and (3) adding a correlated disturbance between d_2 and d_3. The overall message of these modification indices is that the model cannot account well for the association between Outcome 1 and Outcome 2; there is a point of stress surrounding the correlation between these two variables and this is evident by the modification indices suggesting different ways of reducing that stress. In essence, the modification indices tell me that if I add a parameter to more directly link these two variables in some way, the model will better account for the association between them. However, I must decide as a theorist which of the three modes of stress resolution is most appropriate.

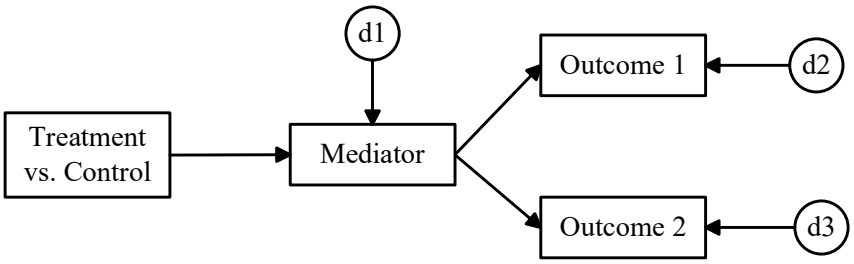


FIGURE 7.6. Model to illustrate modification index dependencies

Suppose the mediator in the above model is the motivation to lose weight, outcome 1 is dieting, and outcome 2 is exercising. In its current form, the model implies that the correlation between dieting and exercise occurs for one and only one reason, namely they share the mediator of the motivation to lose weight as a common cause. To the extent there are other unmeasured common causes of dieting and exercise, and to the extent I can build a compelling case for their existence, then correlating the two disturbance terms is a viable modification to the model. For example, it might be the case that a person's age impacts both dieting and exercise with younger adults eating more healthy and exercising more vigorously than older adults. However, if I did not measure age in my study and therefore do not formally include it in my model, then age still will be operating as part of d1 and part of d2, causing d1 and d2 to be correlated. If I add a correlation parameter between the two disturbances and re-estimate the model, the revised model should fit the data better and the other two modification indices likely will reduce to zero or near-zero given modification index dependencies; I resolved the stress in fit by introducing correlated disturbances.

When making my model modification decision, I also need to explicitly consider the viability of the other ways of reducing the targeted "point of stress." For example, I might ask if it is possible that dieting impacts exercise and, hence, whether a path from dieting to exercise should be added to the model. Without getting sidetracked into the relevant substantive literatures, there is indeed reason to believe that diet has both a positive influence on people exercising and a negative influence on people exercising, the net result of which is an overall zero effect. I might therefore rule out the possibility of adding a causal path from dieting to exercise, although I, of course, need to convince skeptics that this is the case. The same can be said for a causal path from exercise to dieting. As such, I might decide that the correlated disturbance re-specification is most viable as opposed to the other options available to me. Critics of model re-specification would argue that if a correlated disturbance between the two outcomes is so compelling,

why did I not include it in the model in the first place. My response is either (a) that I should have included it and that I just failed to think matters through (my bad), or (b) I thought it was a possibility but was not sure so I decided to let the data inform me one way or the other through the analysis of modification indices post hoc.

Scenarios sometimes occur where two fully dependent modification indices both are theoretically defensible and that my model re-specification should include both of them. This would be the case in Figure 7.6 if I could theoretically justify both the existence of unmeasured common causes for the two outcomes (producing correlated disturbances) and the existence of a dominant causal path from dieting to exercise. In such cases, I must accept the fact that I cannot choose between the two re-specifications and must accept result ambiguity. This is because if I introduced both re-specifications, this portion of the model would be under-identified.

Given the existence of modification index dependencies, many analysts adopt the practice of adding one path/parameter at a time based on a large modification index and then re-running the SEM analysis after each addition to see if the other large modification indices diminish. Analysts often begin the respecification process using the modification index that is largest and that makes the most conceptual sense. This strategy has shortcomings. One shortcoming is analogous to those that arise with stepwise regression where the order in which variables are entered into an equation can affect the search process for adding later predictors. Analogously, the order in which parameters with large modification indices are addressed can impact evaluations of later omitted paths to include or parameters to unconstrain if such constraints have been imposed. MacCallum (1986) and others (e.g., Silvia & MacCallum, 1988; Steiger, 1989) have shown that automated model corrections based on adding paths with the largest modification indices larger than 4.0 on successive steps but without regard to substantive meaning can be problematic. The bottom line is that there is no simple, rule-based way of re-specifying a model. You must bring common sense to bear and ultimately, it is subjective. In the RET examples in Chapters 11 and onward, we sometimes will encounter large modification indices and I will develop in more depth strategies for dealing with and contextualizing them.

Modification Indices and the Expected Parameter Change Index

In addition to modification indices, most SEM software reports a statistic associated with each modification index called an **expected parameter change** (EPC). The EPC estimates what the value of the parameter would be in the model if we were indeed to add it. Software reports both an unstandardized and a fully standardized version of the EPC with most researchers focusing on the latter. In the family size and IQ model, I noted that

the modification index for the path going directly from family size to IQ was 31.40. In my analysis of the data, the fully standardized EPC for this path was reported by Mplus to equal -0.56. This means that if I were to add a path going directly from family size to IQ, the standardized path coefficient for that path would be approximately -0.56 and its p value would be < 0.05 (by virtue of the modification index being > 4.0). The rationale for providing EPC in addition to the modification index is that the standardized EPC gives us a sense of the effect size of the path if you were to add it. If the EPC is small, then even if the modification index is > 4.0, we may choose not to add the path because it is trivial in magnitude. In the current instance, a standardized path coefficient of -.56 is fairly large, which makes for a stronger case for including the path.

Saris et al. (2009) describe four situations for forward model re-specification when using MIs and fully standardized expected parameter changes (S-EPCs) together:

1. When a large MI occurs with a large S-EPC value, one should consider estimating the parameter given it makes conceptual sense to do so.

2. When a large MI occurs with a small S-EPC value, one should keep in mind the effects of sample size on MIs (larger sample sizes typically lead to larger MIs, everything else being equal) and consider not estimating the parameter if ignoring it is not critical to study conclusions.

3. When a small MI occurs with a large S-EPC value, the decision is ambiguous and likely reflects a low power scenario due to a small sample size

4. When a small MI occurs with a small S-EPC value, one would not free the parameter unless theory demands

The key question becomes what is a “large” MI and a “large” S-EPC to use as criteria. There is no simple answer to these questions as the standards can vary depending on the substantive domain and context. Some researchers suggest an absolute correlation of 0.10 (for correlated disturbances) and an absolute standardized regression coefficient of 0.10 (for path coefficients) as a standard for a large S-EPC. Whitaker (2012) explored the viability of a S-EPC cut-off value of 0.20 and found it performed better than the exclusive use of modification indices. However, she also found that using both the MI and S-EPC did not improve on the simpler S-EPC only cut-off approach using values of 0.10. We simply do not have good guidance on this matter.

Parenthetically, for correlated disturbances, the fully standardized EPC value refers to a correlation between the disturbances. If the fully standardized EPC for a correlated disturbance is 0.20, this means that if I estimate the correlation, it will be close to 0.20.

To Make or Not to Make Model Modifications

As noted, some theorists caution against making post-hoc modifications of any form based on forward selection. In addition to the post hoc character of the enterprise (which, as noted in Chapter 6, Bayesians do not necessarily find objectionable), the main argument against the practice goes something like this: The statistical theory that drives the calculation of confidence intervals and p values in SEM models is applicable to the case where a model is *a priori* specified. The statistical theory used to generate p values and confidence intervals does not take into account prior exploratory analyses on the same data. For example, the theory does not presume that an initial model "screening" step is first performed to identify and apply model modifications. By including such a "screening" step, one essentially alters the sampling distribution of the model parameters in unknown ways, which, in turn, can undermine the p values and confidence intervals that rely on knowledge about the shape of the sampling distribution. Thus, the argument goes, one can no longer interpret the p values and confidence interval because the screening step has altered the entire modeling enterprise driven by statistical theory. Critics of post hoc modifications also worry about overfitting one's data in which we interpret a chance effect as meaningful.

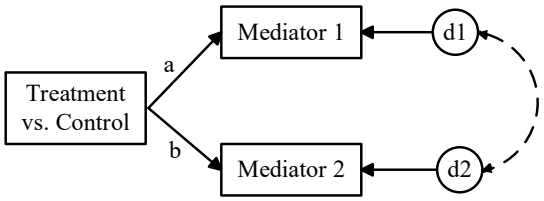
The main argument in favor of making post hoc model modifications is that without them, the model may be misspecified which can undermine the accuracy of p values and confidence intervals anyway. A large modification index is suggestive of specification error which, if ignored, can produce bias in parameter estimates throughout the model and undermine our inferential tests. Adding (meaningful) paths post hoc helps protect against such specification error. A related argument is that it makes no sense to interpret the parameters of a model that is clearly wrong. Finally, if I spent a large amount of time, money, and effort collecting data on a substantive issue, why should I close my eyes to exploring reasons why my *a priori* model did not adequately account for the data? Indeed, I should be intrigued by such a result, much like a detective trying to solve a crime who looks for new "leads" about the true state of affairs.

I personally am sympathetic to the view that we should minimize model misspecification when analyzing data and that we should let data speak to us rather than narrowly focus only on what our initial theories dictate. At the same time, I also am sympathetic to the statistical complications that result from introducing "screening" tests and the havoc that doing so can play on statistical inference. As such, when I introduce model modifications because of glaring misspecification, I interpret p values and confidence intervals with caution and tentativeness. I also express in the Discussion section of my report the limitations of post hoc model modifications and, for scientific publications, stress the need for replicating the results in future studies.

Avoiding Specification Error in RETs

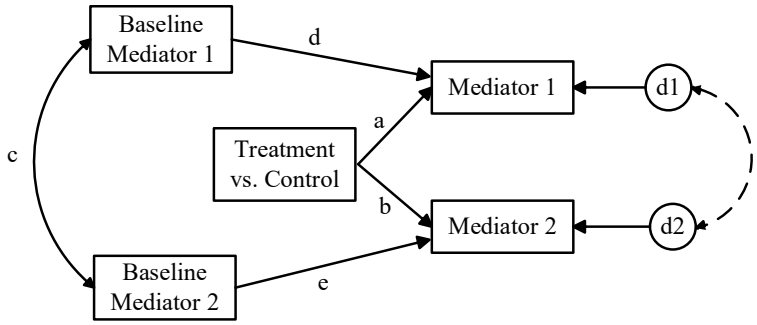
One way to avoid controversies surrounding post hoc model modifications is to make sure your model is correctly specified to begin with. When such is the case, the model should fit the data well and no modifications should be needed except to deal with chance specification error. I have found there are several types of specification error that researchers who use RETs often make. In this section, I identify these common specification errors with the idea that you should be extra careful to avoid them.

Consider the following portion of an RET model that maps the effect of the treatment condition onto mediators. Here is a simplified representation using two mediators:



Specification error in this portion of an RET model often occurs with the omission of correlated disturbances between the mediators (the dashed curved arrow). It turns out that the model without correlated disturbances implies that the only reason Mediator 1 and Mediator 2 are correlated is because the two mediators share a common cause, namely the treatment versus control condition (paths *a* and *b*). Such a specification often is unrealistic – there typically are a host of unmeasured variables that influence both mediators beyond the treatment condition a person is assigned to, such as SES, biological sex, age, and/or perhaps ethnicity. These outside disturbances, in essence, reside in both d1 and d2 causing the disturbances to be correlated. By not correlating d1 and d2, you ignore this fact and will likely underpredict the observed correlation between Mediator 1 and Mediator 2, which can reverberate through to other parts of the model as well.

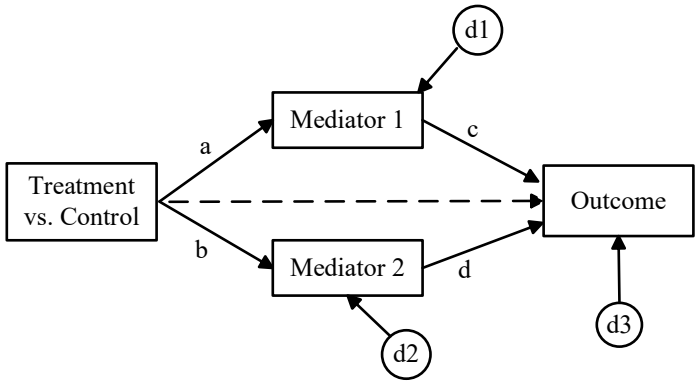
This form of specification error is mitigated to some extent if your model includes correlated baseline covariates, as follows:



In this case, the correlation between Mediator 1 and Mediator 2 is modeled to be a function of (a) the common cause of the treatment condition on each mediator *plus* (b) the degree of correlation between the two baseline mediators (parameter *c*) weighted by the strength of the baseline mediator causal effects on the posttest mediators (paths *d* and *e*). Often the unmeasured "other" variables (SES, biological sex, age, ethnicity) that influence posttest Mediator 1 and posttest Mediator 2 also influence the baseline measures of the mediators. As such, the impact of these other variables on the correlation between the mediators at the posttest are indirectly taken into account via paths *c*, *d* and *e* and might render the need for modeling a correlation between d1 and d2 moot.

If you do not include baseline measures of the mediators as covariates in your model and the only modeled cause of the posttest mediators is the treatment condition, then your model is ripe for this form of specification error. A safer practice is to a priori include correlated disturbances between mediators in your model to adequately capture the correlations between them. Of course, if you do so, you should be able to articulate one or more of the unmeasured variables that reside in both d1 and d2 or that are the source of the correlated disturbances in some other way because we generally want our model specifications to have theoretical justifications.

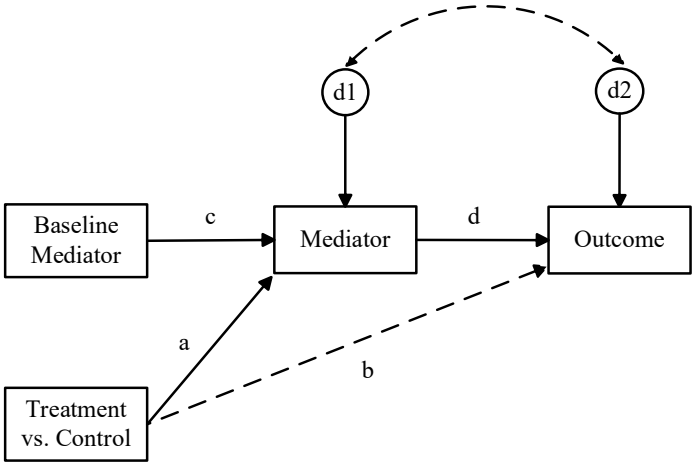
A second form of specification error is the inclusion/exclusion of the direct effect of the treatment on the outcome independent of the mediators, per the dashed path below:



One would include this path if one believes the treatment condition affects the outcome over and above the mediators that the program targets, in this case Mediator 1 and Mediator 2. One excludes the path if one believes the included mediators fully account for the effects of the program on the outcome, rendering the path coefficient for it equal to zero or a trivial value. Baron and Kenny (1987) argue that the path should *always* be included in the model because it then controls for any unmeasured mediators. However, there are scenarios where doing so can be counterproductive so we want to be circumspect about the decision to include the path. One potential disadvantage of including the path is when the strength of path *a* or path *b* is strong, i.e., the program has a strong effect on the mediator(s). Note that this is a result that we hope to achieve when we design programs. In such scenarios, when I regress the outcome onto the mediators and the treatment condition, the correlations between the predictors (i.e., between the mediators and the treatment condition) will be large because, after all, the program has a substantial impact on the mediators and is thus a primary determinant of them. The result will be high multi-collinearity between the predictors that can inflate coefficient standard errors, reduce statistical power, and inflate margins of error for the regression coefficients. Indeed, if the path for the direct effect path is zero or near zero, then coupled with high collinearity, the result can be an artifactual suppressor effect that disrupts coefficient interpretation. The bottom line is that in RETs, you should include the direct effect path if you believe the path is non-zero and meaningful. In my experience, in many RETs the path is not needed because it is reasonable to assume the treatment only affects the outcome through the mediators the treatment seeks to change.

If you exclude the path in your initial analyses but it actually belongs in the model, you usually will obtain bad model fit diagnostics. You can then revise the model by adding the path. Alternatively, if you include the path in your initial analysis and it is truly trivial, its path coefficient will be statistically non-significant and near zero. You then consider trimming it.

A third form of specification error that often is ignored in RET models are correlated disturbances between the mediator and the outcome per the following influence diagram:



A correlation between d1 and d2 is usually required if there are unmeasured confounds that are omitted from your model and that impact both the mediator and the outcome, thereby inflating (or deflating) the correlation between them. In Chapter 2, I stressed the importance of identifying such confounds when planning your study, measuring the most important ones, and then including them as covariates in your modeling efforts. If you fail to do so adequately, then you will need to consider adding correlated disturbances order to obtain unbiased estimates of the true causal coefficient for path *d* in the model.

This specification error is insidious because it does not manifest itself in any of the standard model fit diagnostics. It simply inflates or deflates your estimate of path *d* and you are none the wiser for it. Given this, it is that much more important that you use the strategies I discuss in Chapter 2 for identifying and dealing with unmeasured confounds.

Another downside of this type of specification error is that if you add the correlation between the disturbances as a parameter in your model, the model often will be under-identified and not estimable. To make the model identified, you need to introduce an instrumental variable, which I discussed in depth in Chapter 6.

In sum, be aware of the possible need for correlated disturbances between your mediators, the need to include or exclude a direct effect of the treatment condition on the outcome independent of the mediators, and the need for correlated disturbances between your mediators and outcomes. If your model does not fit the data well, these are reasonable points of stress to explore. If you think about these typical forms of specification error when first formulating your model, you likely will be able to avoid post hoc modifications.

Backward Searching

Another form of post hoc modification is to trim away paths in the model that are statistically non-significant and then to re-estimate the more parsimonious model in order to simplify matters. The arguments in favor of such trimming usually surround parsimony and the avoidance of overfitting. The arguments against trimming are that non-significance can result from low statistical power and that omitting the path or parameter can possibly introduce consequential bias in the retained parameter estimates. Also, if there are multiple paths that are non-significant, excluding any one of them may not introduce consequential bias in the included paths, but excluding all of them simultaneously can produce considerable bias. Many statisticians argue against trimming if the variables involved have a good track record in prior research and their inclusion makes substantive sense.

Concluding Observations on Model Re-specification

I tend to favor being open to making SEM model modifications after my initial analyses in order to avoid interpreting models with specification error in them. It is not uncommon when specifying an initial model to have in mind some paths that one is quite confident exist and other paths that one is confident do not exist. These beliefs dictate the essence of the influence diagram. However, there often are paths or parameters one is less confident about and that "may or not" exist in the population. Some theorists error on the side of including these paths in the model rather than leaving them out, with the idea that one can always ignore non-significant paths when interpreting results. Other theorists favor parsimony and the avoidance of overfitting by leaving these paths out of the model with the idea that if the paths are meaningful, they will generate large modification indices and can always be added when one looks at model diagnostics. Neither approach is correct. Each has pros and cons.

A common refrain is that when you make model modifications, it is good to conduct a replication study to ensure the post hoc modifications manifest themselves in a second study. Unfortunately, when you are hired to evaluate a program, this advice is somewhat vacuous because clients usually are not going to want to spend resources on an immediate repeat of the program evaluation. Some scientists recommend defining a "set aside" sample in the primary study that one can use for replication analyses. For example, after conducting the RET, one might randomly divide the sample in half and then fit the RET model separately in both samples to determine which results replicate. An argument against this strategy is the consequent reduced statistical power and inflated parameter instability that occurs by using only half of one's sample size when modeling data. For a discussion of this topic, see Koul, Becchio & Cavallo (2018) and Browne (2000).

COMPARING MODELS USING SEM

A final facet of model evaluation I consider is that of comparing models. Sometimes we want to compare models to determine if the target model is more consistent with data than viable competing models. Model comparisons also are core to selected tests of moderation. I consider in this section general SEM methods for comparing models. I first address the case of comparing nested models and then I consider comparing what are known as equivalent and non-nested models.

Comparing Nested Models

Nested models are when the models being compared focus on the same measured variables but one of the models is derivable from the other by placing restrictions on the other model. The restrictions might be fixing the value of a coefficient in one model but estimating that same coefficient in the alternative model; or it might be imposing an equality constraint between two coefficients in one model but allowing both coefficients to be freely estimated without the constraint in the other model. Stated a bit more formally, two models have a nested relationship if the parameter space for one model is a subset of the parameter space for the other model.

Consider the model in Figure 7.7 in which a treatment to reduce physical disability due to chronic back pain addressed two mediators, (1) pain catastrophizing, which is the tendency for people to describe pain experiences in exaggerated terms and to ruminate on them more than the average, and (2) fear-avoidance, which is an exaggerated tendency to avoid activities that might cause back pain which, in turn, leads to disability through "nonuse" of the back. The degree of physical disability was measured 3 months after treatment. I do not show them in the figure but the model included baseline covariates for each endogenous variable. Note that this model includes correlated disturbances between the mediators to reflect that they share unmeasured common causes. The total sample size is N = 400, 200 males and 200 females.

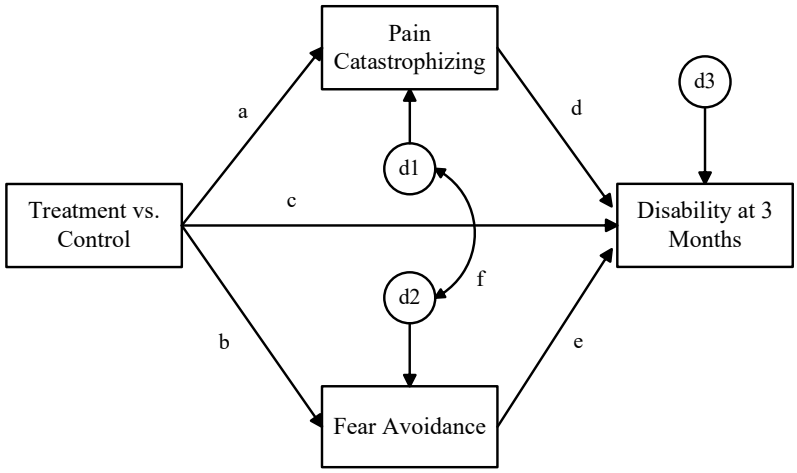


FIGURE 7.7. Example for nested models

I might want to test if path coefficients *a* through *f* are different for males and females in order to test model generalizability across biological sex. In essence, I want to compare two models, one in which the path coefficients *a* though *f* are the same in the two groups and another model in which those coefficients are allowed to differ in the two groups. In SEM, there is a strategy called **multi-group SEM** that, in a single analysis, computes the fit of the model for the different groups (males and females) and then reports the traditional overall global fit indices taking into account how well the model fits in both groups simultaneously. To test generalizability, I first fit a multigroup model that allows paths *a* through *f* to differ in value for males relative to females. I then fit a second model where I force the path coefficients for males to equal the path coefficients for females. This is known as imposing **across-group equality constraints**. There is a nested relationship between these two models because the second model is derived from the first model by imposing constraints on the first model. For the latter model, SEM algorithms find the values of the coefficients that best reproduce the covariances for males and the covariances for females but under the constraint that the coefficients for the two groups must be equal.

Suppose that the population coefficients for males and females are, in fact, equal. The equality constraint should then pose no problems for finding a good model fit, because, in reality, the same coefficients apply to both groups. However, if one or more of the population coefficients is much larger or smaller in males as opposed to females, a model that forces all of the coefficients to be equal will not fit the data well. Thus, to evaluate generalizability, I might compare the fit of a model without the across group

equality constraints with a model that imposes across group equality constraints. If the fit for the two models is comparable, then this suggests that the path coefficients for males are not much different than the path coefficients for females. However, if the imposition of the equality constraint adversely affects model fit, then this suggests that one or more of the path coefficients must differ non-trivially for males versus females.

Chi Square Difference Test

One way of testing the relative fit of two models is to (a) estimate the model with the equality constraints and obtain a chi square fit index for it, then (b) estimate the model without the equality constraints and obtain a chi square fit index for it. The former model is called a **constrained model** because I have "constrained" the paths to be equal across groups. The latter model is the **unconstrained model** because the values of the paths are free to vary across groups. The constrained model is **nested** within the unconstrained model (Bollen, 1989) or, using different terminology, the two models are **hierarchically related**. It turns out that for nested models, one can compute a **chi square difference test** to evaluate the following null and alternative hypotheses:

H_0: The difference in population model-data fit for the two models is zero

H_1: The unconstrained model population fit is better than the constrained model fit

I subtract the sample chi square for the unconstrained model from the sample chi square for the constrained model and I also difference their degrees of freedom. I can determine the p value for the chi square difference test because it turns out that for a nested model, the difference between the chi squares, χ^2_{DIFF}, is distributed as a chi square with df_{DIFF}, i.e., the difference in the models' degrees of freedom. I provide a program for applying this difference test on my website (the program is called *chi square difference test*).

There is an important qualification to the above. If you use the robust ML estimator based on Huber-White estimation (option MLR in Mplus) instead of traditional maximum likelihood estimation (which I generally recommend you do), it turns out that the χ^2_{DIFF} statistic is *not* chi square distributed. However, there is a correction you can apply so that it is. Specifically, there are two correction factors, one for each model being compared. Mplus provides them on its output, labeled as "Scaling Correction Factor for MLR." A similar label is used for the R package lavaan. The formula for the correction factors is given in Bryant and Satorra (2012) as well as how to execute the chi square difference test with them. On my website, I provide a program for the approach, called *Scaled chi square difference test*. Because I decided to use MLR for the physical disability example, I need to use the scaled difference test here.

When I fit the model in the two groups with the equality constraint, the chi square

for the constrained model was 15.32, df = 13 and the scaling correction factor was 1.01. For the unconstrained model that permitted path coefficients to vary across biological sex, the chi square was 9.01, df = 8 and the scaling correction factor was 0.99. The scaled chi square difference using the program on my website was 6.29, df = 5, $p < 0.28$. The difference in fit between the two models was statistically non-significant, which is consistent with the proposition that the results are generalizable across males and females. Had the null hypothesis been rejected, I would conclude that the difference in fit between the two models is not zero, which suggests that at least one of the path coefficients is not generalizable across males and females. I discuss in Chapter XX how to program Mplus to do these runs and how to isolate which path coefficients are responsible for the lack of generalizability. A weakness of this test is that generalizability assertions are based on obtaining a statistically non-significant result. If the difference test is underpowered, then this biases conclusions towards that of declaring generalizability.

Comparing Models using a Close Difference Approach

MacCallum, Browne and Cai (2006; see also MacCallum, Lee & Browne, 2010) describe a method for nested model comparisons using RMSEA statistics rather than chi square statistics. MacCallum et al. argue that the premise of competing models having exactly the same fit as specified in the null hypothesis of the above chi square difference test is unrealistic and will almost always be false; no two nested models in practice will ever provide the exact same fit to the data. It is more realistic, they argue, to work from a null hypothesis of a small difference between population models rather than exact model fit equivalence. Stated another way, even though two population models may not provide exactly the same fit to data, the degree of discrepancy might be so small that the models can be viewed as being *functionally equivalent*. As such, according to MacCallum et al., the null hypothesis should be stated so that it takes a functional equivalence perspective rather than a strict equivalence perspective.

Given nested models A and B, MacCallum et al. (2006) propose a null hypothesis of the form

$$H_0: \varepsilon_A - \varepsilon_B \leq d$$

where *d* is a non-negative value that represents the difference in two population RMSEAs. The value of d represents an effect size index that defines a standard for declaring the population difference in fit between the two models as being trivial to the point that we can treat the models as being functionally equivalent. In this case, ε_A is a population RMSEA for the constrained model and ε_B is a population RMSEA of the

unconstrained model. The alternative hypothesis is

$H_1: \varepsilon_A - \varepsilon_B > d$

For example, if d equals 0.01, this means that if the population difference between the population RMSEAs specified in H_0 is less than 0.01, the constrained and unconstrained models will be deemed as being functionally equivalent. MacCallum et al. (2006) and Preacher, Cai and MacCallum (2007) describe methods for testing the above null and alternative hypotheses in a hypothesis testing framework. I provide a program (called *close fit difference test*) on my website for performing the test.

A challenge for applying the close difference approach is specifying *a priori* a meaningful close fit criterion in RMSEA units. In general, it turns out that an RMSEA difference of 0.01 when the values of the RMSEA are, say, 0.04 and 0.03 implies a different amount of fit disparity at a localized level than a difference of 0.06 and 0.05 which, in turn, reflects a different disparity than when the RMSEA values are 0.08 and 0.07 (see Liu & Bentler, 2011, for elaboration). Given this, it is not sufficient to merely specify a criterion value for d, such as 0.01. Rather, one must specify specific values for ε_A and ε_B in the null and alternative hypotheses to define the effect size that is the cutoff for a small difference. This is a weakness of the approach because the RMSEA index for a given model is not very intuitive, making it difficult to a priori specify values for ε_A and ε_B (see Maydeu-Olivares, 2017; Shi, Maydeu-Olivares, DiStefano, 2018). As well, the RMSEA has been found to be influenced by incidental parameters that should have little bearing on lack of model fit, such as the magnitude of factor loadings in a measurement model and model size (Saris, Satorra, & van der Veld, 2009; Chen et al., 2008; Savalei, 2012; Shi, Lee, & Maydeu-Olivares, 2018). These considerations detract from the the MacCallum et al. (2006) strategy.

When invoking the RMSEA close fit difference test, researchers typically resort to the use of the rule of thumbs for population RMSEA differences. For example, it is not uncommon to use an RMSEA difference of 0.01 or 0.05 as a basis for applying the close difference test; but the question remains as to what values to use for the separate RMSEAs of ε_A and ε_B for this difference. It turns out that, holding the hypothesized population RMSEA difference constant, the larger the values of ε_A and ε_B (e.g., 0.08 and 0.07 versus 0.05 and 0.04), the larger will be the critical value associated with the RMSEA difference test for the same data. So, a larger effect size criterion is associated with lower values of ε_A and ε_B. Note that your choice of values for ε_A and ε_B is *not* what you think the actual values of the population RMSEAs are for the models you are testing. Rather, they are chosen to help you define what you think is an appropriate effect size standard to use to declare to models as functionally equivalent or functionally different.

The close fit difference test is typically applied to traditional maximum likelihood estimation. For the physical disability example, I used MLR for the chi square difference test between the constrained and unconstrained models that forced the path coefficients to be equal across males and females versus not imposing such an equality constraint (see the previous section for the chi square difference test). I found the scaled chi square difference to be 6.29, with the two model degrees of freedom of 13 and 8. If I (somewhat arbitrarily) define ε_A as 0.05 and ε_B as 0.04 and apply the program on my website, the resulting p value for the test of close fit difference was $p < 0.86$. This means that I cannot reject the null hypothesis and is consistent with functional equivalence of the two models. If the null hypothesis had been rejected, I would conclude that the model coefficients, as a collective, differ for males and females using a close difference standard of 0.01 RMSEA units.

As with the original chi square difference test, the results turned out to be consistent with coefficient generalizability. Note that this test also biases conclusions towards generalizability across groups if statistical power is low. Indeed, it is even more sample size demanding than the chi square test of fit difference in terms of statistical power.

Other forms of the RMSEA difference test have been proposed (Savalei, Brace & Fouladi, 2022), but in my judgment these tests require further evaluation as to their practical and substantive utility in the face of typical analytic scenarios that have non-normal and missing data that use robust estimation. MacCallum et al. (2006) provide a thoughtful discussion of factors to consider in choosing values of define ε_A and ε_B.

Comparing Models using the Comparative Fit Index

The CFI traditionally compares a model of interest with the independence model. Given nested models, you can alter the formula for the CFI to compare any two models of interest, not just the target model against the independence model. Let $d_{UNCONSTRAINED}$ = the χ^2 value for the unconstrained model minus its degrees of freedom and $d_{CONSTRAINED}$ = the χ^2 value for the constrained model minus its degrees of freedom. Then

$$CFI = (d_{CONSTRAINED} - d_{UNCONSTRAINED}) / d_{CONSTRAINED}$$

If the CFI index is greater than one, it is set to one. The CFI reflects the improvement in fit of the unconstrained model relative to that of the constrained model (the numerator in the above equation) indexed against the lack of fit of the constrained model (the denominator). For the physical disability example, the adapted CFI is

$$CFI = ((15.32-13)-(9.01-8))/(15.32-13) = 1.31/2.32 = 0.56$$

The unconstrained model improves fit by about 56% relative to the constrained model after adjusting for model complexity as reflected by the models' degrees of freedom. If I apply an adapted Tucker-Lewis Index to the data, the TLI was 0.29 and the Bentler-Bonnet Index was 0.41. Obviously, when comparing two viable models, the rule of thumb of a CFI > 0.95 for declaring a meaningful difference is not applicable. Also, there are no formal standard errors for the CFI, TLI, or BBI from which to compute p values. The results are purely descriptive.

In my opinion, the CFI, TLI and BBI indices can be misleading when used in this way. If an unconstrained model reduces the chi square from a constrained chi square of 16 to an unconstrained chi square of 8, the change of 8 units seems more to me than if the reduction is from 2 to 1. Yet, both cases reflect 0.50 improvement in the chi squares when calculated using the traditional formula for improvement rates, as discussed earlier. For the physical disability example, the chi square minus the degrees of freedom for the constrained model was 2.32 and for the constrained model it was 1.01. I personally would judge the differences in these indices to be small despite obtaining the 56% improvement reported earlier.

Comparing Models using Information Theory Indices

A final strategy I consider for comparing models in SEM is to use indices that have their basis in information theory (Burnham & Anderson, 2004; Raftery, 1995). The two most popular indices are the **Akaike Information Criterion** (AIC) and the **Bayesian Information Criterion** (BIC). Both indices make use of a construct known as a **log likelihood**, which reflects the likelihood that the observed sample data could result if the model form and the population parameters (e.g., the path coefficients) took on the values derived from the SEM maximum likelihood fit function defined earlier. Log likelihoods are not probabilities per se. Rather, they are complex functions of probability densities and yield indices that are difficult to interpret in and of themselves. To give you an idea of their complexity and non-intuitive nature, if I assume multivariate normality for the variables in a model (represented by a vector I call X), the log likelihood of the maximum likelihood estimator for a given model equals (Bollen, Harden, Ray & Zavisca, 2014):

$$LL = K - \frac{1}{2}\ln(|\Sigma'|) - \frac{1}{2}(\bar{X} - \mu')^{T}\,\Sigma'^{-1}(\bar{X} - \mu')$$

where LL is the log likelihood for the model, Σ' is the predicted covariance matrix from the model, μ' are the predicted means from the model, T is the matrix operation of a transpose, and K is constant that has no effect on the values of the parameters that maximize the log likelihood. Log likelihoods typically are negative in value and the

larger they are (i.e., the closer the negative number is to zero), the higher is the likelihood that the data are compatible with the tested model.[6] Both the AIC and BIC work with log likelihoods, as I now illustrate.

The Akaike Information Criterion. The AIC takes different forms in SEM software but the differences between them usually are not consequential for comparing models. In Mplus, the AIC is reported for most SEM models and is defined by the formula

$$AIC = (-2)\ (LL) + 2r \qquad [7.3]$$

where LL is the log likelihood associated with the model in question and *r* is the number of estimable parameters in the model. By multiplying the log likelihood by -2, the AIC becomes a positive number, with larger numbers indicating worse fit to the data. The AIC also includes what is often referred to as a penalty term for lack of parsimony, namely 2r. If the model has many parameters that must be estimated, then the AIC will be larger, everything else being equal. With the AIC, model parsimony is rewarded.[7]

There are many variations of the AIC. For example, some researchers use the above formula but with a small sample bias correction incorporated into it. This is sometimes referred to as AIC_c. The nuances of the different versions of the AIC are described in Burnham and Anderson (2004). Do not be surprised if for some software you observe AIC indices that are different in magnitude from those calculated in Mplus, the software I feature in this book. The important idea for all of them is that we can compare different models using the respective AICs and then choose models that have "better" (i.e., lower) AICs when compared to other models.

Sometimes we compare more than two models, i.e., we might compare three, four or five models. I provide examples of this in future chapters. When comparing more than two models, it is common to first identify the model with the lowest AIC value (which is the best fitting model of all the models being considered). One then subtracts this value from each AIC for the other models. For the best fitting model, the difference will be zero and for all other models, it will be positive in value, with the larger the disparity, the worse the fit of the target model relative to the best fitting model. Here is a tabled example of the AIC results for five models where the AIC column presents the computed AIC for the models and the Diff column subtracts the lowest observed AIC (202 for Model 2) from each model AIC:

[6] The log-likelihood for H_0 reported on the Mplus output is used as the value of -2 LL.

[7] Technically, the 2r term is not envisioned as a penalty for lack of parsimony but instead is part of the mathematical theory underlying the derivation of AIC. Also, choosing the value of -2 to multiply the LL by is not arbitrary. This value has a clear rationale, which is described in Burnham and Anderson (2004).

Model	AIC	Diff
1	204	2
2	202	0
3	206	4
4	206	4
5	214	12

The larger the value in the D column, the worse the model fit, with model 2 being the favored model based on the above table.

General rules of thumb have been proposed to interpret the magnitude of the AIC differences between models (see Burnham & Anderson, 2004). The most common rules of thumb are as follows:

1. If the disparity in AICs is < 2, then the two models have about the same support

2. If the disparity in AICs is > 2 and < 4, then the better fitting model has positive support relative to the model it is compared with

3. If the disparity in AICs is > 4 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the disparity in AICs is > 10, then the better fitting model has very strong support relative to the model it is compared with

One must be careful when applying rules of thumb like this because they may not apply in all contexts. Indeed, some analysts object to their invocation arguing that they can result in the same rigid and counterproductive use of a criterion like "$p < 0.05$" that plagues null hypothesis testing frameworks. Using the above criteria, model 2 is favored relative to all the other models.

Another standard for comparing any two models using the AIC is to examine what is called the **evidence ratio**. Let D = the AIC for the worse fitting model of the two models being compared minus the AIC for the better fitting model of the two models (and let *e* be the traditional Neperian constant). The evidence ratio is defined as

$$ER = 1 / e^{(-D / 2)}$$

where ER stands for "evidence ratio" and *e* is the traditional Naperian constant. It indicates how much more likely the better fitting model is (given the data) than the worse

fitting model (given the data). For example, if the AIC for the better fitting model is 200 and for the worse fitting model it is 202, then the evidence ratio is

$1 / e^{-(202-200) / 2)} = 2.72$

The better fitting model has 2.72 times more support than the model it is being compared with.

Finally, some researchers normalize AIC differences relative to all models being compared so that they sum to 1. These are called **Akaike weights** and indicate the "weight of evidence" in favor of a model relative to *all* models in the comparison set. Akaike weights are distinct from evidence ratios because Akaike weights are impacted by the particular set of models being compared when the number of models is greater than two. Let me first describe how Akaike weights are calculated and then I will make them more concrete with an example.

To calculate the Akaike weight, each model being considered is assigned an index of its likelihood relative to that of the best fitting model using the value from the denominator of the evidence ratio, $e^{(-D/2)}$ as the index. Let T = the sum of the $e^{(-D/2)}$ values across all the models being considered. Then the Akaike weight for a given model is defined as

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model in question. Suppose I fit five different models to a set of data. Here is a table with the AICs, the differences between the model AIC versus the model with the lowest AIC, and the Akaike weights (w):

Model	AIC	D	$e^{(-D/2)}$	$w = e^{(-D/2)}/T$
1	204	2	0.3678	0.2242
2	202	0	1.0000	0.6094
3	206	4	0.1353	0.0824
4	206	4	0.1353	0.0824
5	214	12	0.0024	0.0015
Sum			T = 1.6408	1.0000

The sum of the weights across all five models is 1.00. The weights represent a continuous measure of relative strength of evidence for each model. Each weight can be crudely

interpreted as the probability that the model is the best model among the set. In the present case, the data support Model 2.

The basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in AICs, the evidence ratios, the Akaike weights, and the substantive meaning/logical coherence of the models in order to choose the best one.

The Bayesian Information Criterion. In this section, I describe a second index used for comparing models that also relies heavily on log likelihoods. I focus on the BIC used in Mplus known as the **Schwartz BIC**, which is formally defined as

$$BIC = (-2)\,(LL) + \ln(N)\,r \qquad [7.4]$$

where r = the number of estimable parameters in the model, N = the sample size, and LL = the model log likelihood. Like the AIC, the smaller the BIC, the better the model fit, everything else being equal. Also like the AIC, there is a penalty function for lack of parsimony, but the penalty is different than that for the AIC. The penalty is somewhat harsher than that for the AIC.

Like the AIC, it is not uncommon for the model with the smallest BIC to be used as a reference point for comparing models, The usual practice is to calculate the difference between each model in the model set minus the model with the best BIC. For the best fitting model, this difference will, of course, be zero.

To evaluate models in terms of BIC differences, general rules of thumb are (see Raftery, 1995):

1. If the BIC disparity is < 2.2, then the better fitting model and the model it is compared with have about the same support

2. If the BIC disparity is > 2.2 and < 6, then the better fitting model has positive support relative to the model it is compared with

3. If the BIC disparity is > 6 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the BIC disparity is > 10 then the better fitting model has very strong support relative to the model it is compared with

For similar but slightly different standards, see Wasserman (1997).

One also can calculate what is called a **Bayes Factor** (BF) for each model relative to the best fitting model. It is defined as

$$BF = e^{(D'/2)}$$

where D' is the BIC difference between the target model and the best fitting model and *e* is the traditional Naperian constant. The Bayes factor is the probability that the model with the lowest BIC holds (given the data) divided by the probability the model in question holds (given the data). For example, a BF = 10 means it is 10 times more likely the model with the minimum BIC is true given the data than the model in question is true given the data. The BIC rules of thumb for model comparisons can be mapped onto Bayes Factors values which some researchers think gives BIC differences when comparing two models somewhat more intuitive meaning. Here is a table that provides such mapping as provided by Bollen (2026):

Descriptive Term	BIC Difference	Bayes Factor
Very Strong	>10	>150
Strong	6 to 10	20 to 150
Moderate	2 to 6	3 to 20
Weak	0 to 2	1 to 3

Finally, a relative model weight, analogous to the Akaike weight, can be computed by normalizing model BFs relative to *all* models in the comparison set so that they sum to 1. Let D = the difference in the BIC for the model in question minus the value of the BIC for the best fitting model, T = the sum of the index $e^{(-D/2)}$ across each model. The relative weight for a model is

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model. Again, the sum of the weights across models is 1.00.

As with the AIC, the idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in BICs, the Bayes factors, the relative weights, and the logical coherence of the models, in order to choose the best one.

You will encounter variants of the BIC, but the fundamental logic in applying them is the same. For example, like the AIC_c, there is a sample size adjusted BIC that is similar to Schwartz' BIC, but it applies a somewhat milder penalty function (Sclove, 1987).

Which Method is Better, AIC or BIC? A hotly debated topic in statistics is which approach to model comparison is better, one based on AICs or one based on BICs. There are advocates on both sides of the matter and I dare not venture into this controversy here. The BIC tends to favor simpler models more so than the AIC. This can be both a

strength and a weakness. Interested readers are referred to Burnham and Anderson (2004), Yang (2005), and Kuha (2004). Kuha argues for the use of both indices. An issue with both approaches is that researchers can be lulled into thinking that the best fitting model within a set of models is the true model. This is not necessarily the case. Researchers can choose the best of a set of wrong models, which is not our goal.

The BIC in Structural Equation Models. In structural equation modeling, the BIC has been conceptualized and implemented in multiple ways, one of which is the Schwartz BIC reported in Mplus. One difference between different formulations of the BIC in SEM is whether the BIC uses only the log-likelihood of the hypothesized model or whether it takes into account the log-likelihood of a **saturated model**. A saturated model adds parameters to the tested model to the point that the covariance and mean structure of the data are perfectly reproduced – at the cost of increasing the number of estimated parameters. When the BIC is focused on just the hypothesized model, -2LL is used in the calculation of BIC. When the saturated model also is included, the model chi square statistic is used in place of -2LL. Preacher and Merkel (2012) and Bollen et al. (2014) provide the relevant statistical details for this assertion. Thus, a common instantiation of a BIC that takes into account the fitted model and a saturated model is

$$BIC = T_{ml} + \ln(N)\, r \qquad [7.5]$$

where T_{ml} is the model chi square and r is the number of estimated parameters in the model. As before, the smaller the value of the BIC, the better the model fit. Bollen et al. (2014) use a slightly different formulation defined as

$$BIC_s = T_{ml} - (df)\,(\ln(N)) \qquad [7.6]$$

where df = the degrees of freedom associated with the chi square.[8] In this formulation, a value of BIC_s that is greater than zero supports the saturated model over the hypothesized model, with larger positive values indicating increasing support for the saturated model. A negative BIC_s supports the hypothesized model over the saturated model, with more negative values reflecting stronger support for the hypothesized model. Positive values of BIC_s are common in practice because the only reason that a hypothesized model will outperform a saturated model is the lack of parsimony associated with the saturated model. As with the traditional BIC, we still prefer the model with the lowest BIC_s value among the models being compared and we still can invoke the Raftery guidelines when comparing models, as appropriate. The BIC_s just has a different statistical basis.

[8] The degrees of freedom in Equation 7.5 is inversely related to the number of parameters estimated in Equation 7.6. The change in sign for the penalty term may appear confusing, but the essential dynamics of the BIC are preserved. See Bollen et al. (2014) for details.

An alternative to the BIC_s is **Haughton's BIC** or HBIC. It is defined as

$$HBIC = T_{ml} - (df)\,(\ln(N/(2\pi)) \qquad [7.7]$$

The HBIC slightly modifies the penalty function from Equation 7.6, the logic for which is described by Haughton (1988; Haughton et al., 1997; see also Bollen et al., 2014). Two other alternatives are the IBIC (called the **information matrix based BIC**) and the SPBIC (called the **scaled prior BIC**), both of which incorporate features of the estimated information matrix into the penalty function. Simulation studies tend to favor the use of HBIC and SPBIC, with the former having the advantage that it is easy to compute from standard SEM output (Bollen et al., 2014). On my website, I provide a program that allows you to calculate HBIC from standard SEM output.

In sum, there is considerable variability in the type of BICs reported by SEM software. In addition to a Schwartz BIC, most SEM software also reports a variant of the BIC called the sample size adjusted BIC (Enders & Tofighi, 2008; Tofighi & Enders, 2007). Aside from these indices, there are a host of additional possible indices, the major ones of which I have discussed above. It is sometimes confusing as to which BIC to use when comparing models. The Schwartz BIC and the HBIC work well in many contexts.

Sampling Error and Model Selection Uncertainty. Suppose two models are compared using BICs for data based on a random sample from a population. One almost certainly would obtain different values of the BIC disparity between the two models if a different random sample of the same size was selected from the population. Some methodologists have argued that such sample to sample fluctuations in BICs (and AICs) should be taken into account when choosing among models (Preacher & Merkel, 2012). The fluctuations typically will be greater for smaller as opposed to larger sample sizes. Preacher and Merkel (2012) develop a method for estimating confidence intervals for a BIC in structural equation modeling contexts. They use a slightly different formulation of the BIC than those described above, but their approach retains the core properties of all BIC indices, namely lower values imply more support for a model and models with more estimated parameters are penalized to a greater extent than models with fewer estimated parameters. I implement a bootstrap version of the difference between two BICs for nested models in a program called *BIC difference CIs* on my website. The confidence intervals yielded by this approach can then be used to make model choices. If the CI contains the value of 0 or a small absolute BIC value of 2 or less, then neither model has preference over the other. Merkle, You and Preacher (2016) suggest methods for comparing BIC differences for non-nested models which I describe in Appendix B. Appendix B also describes syntax in Mplus that can be used to evaluate if two models are nested.

Concluding Comments on the Different Methods for Nested Model Comparisons

In sum, there are multiple strategies for comparing two or more nested models, each with strengths and weaknesses. Both the chi square difference test and the close fit difference test emphasize statistical significance testing. As such, it is important that you have sufficient statistical power to detect meaningful model differences. I discuss relevant power analysis perspectives in Chapter 28. The close fit test uses a close fit standard in the form of RMSEA differences but RMSEAs are difficult to interpret and it can be challenging to specify substantively meaningful standards in RMSEA terms. The CFI approach has the weakness of being purely descriptive. It also can be misleading as a rate index. The BIC and AIC operate outside null hypothesis testing frameworks which some view as a strength.

The practice of comparing multiple models and then focusing on the best fitting model of the set brings with it some noteworthy cautions. Once data have been used to select a model, the estimated standard errors for the parameters of that model may no longer be strictly valid without additional corrections (Berk et al. 2013). This is because model selection is stochastic; under repeated sampling, different models might be selected based on random variation in the data. This variability should then be incorporated into the calculation of standard errors (see Berk et al. 2013).

As I discuss in later chapters on moderation, my own preference is not to work with general omnibus tests of model differences as described in this chapter, but instead I prefer to focus on model differences on a per path or per parameter basis. In the physical disability example, I would examine each separate path coefficient for males and females, document the magnitude of the difference between males and females for each path coefficient (including reporting a margin of error for the difference), and evaluate those differences on a pairwise basis with controls for multiplicity. I illustrate this approach in future chapters.

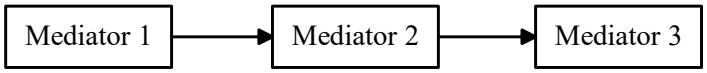
Equivalent Models and Non-Nested Models

When comparing models, statisticians distinguish between equivalent models, nested models, and non-nested models. **Equivalent models** are when two different models yield exactly the same predicted covariances and fit statistics for a data set and also share the same number of degrees of freedom. As a simplistic bivariate example, the model that X causes Y is an "equivalent model" to the model that Y causes X because each yields the same predicted covariance matrix, each has an equivalent number of degrees of freedom, and the fit statistics are the same for each model.

Sometimes models might not be equivalent but sub-portions of them are. In RETs,

the problem of equivalent sub-models usually presents itself when describing causal relations among mediators or between mediators and outcomes. Consider two possible sub-models of an RET in Figure 7.8. that focus on the subportion of a model that maps the causal relationships between three mediators. Figure 7.8a and 7.8b represent equivalent submodels because if analyzed in isolation, both have the same number of degrees of freedom and both produce the same predicted covariances among the three variables. Bentler and Satorra (2010) outline a strategy for determining empirically if two models/submodels are equivalent, nested, or non-nested. I describe this method in Appendix B. It is important in RETs that researchers be aware of the possibility of substantively viable equivalent sub-models and to acknowledge their presence when discussing RET results (Hayduk, 2014). Sometimes we can rule out an equivalent model based on logic. Bad weather causes car accidents and predicts a correlation between these two variables. An equivalent model is that car accidents cause bad weather, a model that also predicts a correlation between the two variables. However, the latter model is clearly illogical. Important to keep in mind here is that just because a theory "fits" the data, that does not make the model right. In cases where equivalent models are both logically plausible, it sometimes is possible to design studies to empirically discriminate them, but this topic is beyond the scope of this book.

(a)



(b)

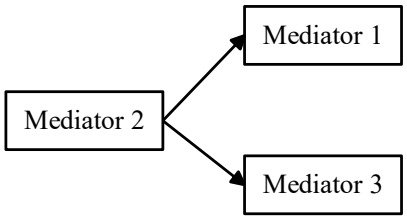


FIGURE 7.8. Example of equivalent models

If two models are not equivalent, they can be either nested or non-nested. As noted,

comparing non-nested models is more complicated than comparing nested models. A test that often can be used when comparing two models (whether they are nested or non-nested) is known as the **Vuong test** (Voung, 1989) or the **Vuong closeness test** as implemented by Merkle et al. (2016). It tests the null hypothesis that the two models are equally close to the true data generating process against the alternative that one model is closer. I describe this test in Appendix B.

CONCLUDING COMMENTS

SEM is a powerful method for testing a wide range of models of interest in the social sciences. It is particularly well suited to the analysis of RETs, as I show in future chapters. The approach translates a theory, often represented by an influence diagram, into a system of equations and then parameter estimates for each equation are derived. The overall fit of the model to the data, considered multivariately, is evaluated using global fit indices, which are supplemented by localized fit indices. The approach has been adapted to work with latent variables so that SEM has the ability to address measurement error during parameter estimation as it integrates structural models with measurement models. As with most complex, integrative frameworks, there are many issues that need to be taken into account and we will encounter them as we apply the approach to the analysis of RETs.

To preview the use of SEM with RETs, consider an RET where I compare two variants of programs to increase adherence to medical regimens for people living with chronic cancer. Each program addresses the same three facets, namely (1) teaching people effective strategies to cope with the side effects of the cancer medications they take, (2) teaching people strategies to strengthen social support for protocol adherence, and (3) increasing perceptions of the importance of adhering to the protocol. One program uses passive learning to address each mediator such that participants read engaging and informative articles on each topic. The second program uses the same methods but augmented with active learning strategies, such as role playing, writing essays, and participants explaining materials to other participants. With the addition of a control group, the RET has three conditions, (a) the education group, (b) the education plus skills group, and (c) the control group.

Figure 7.9 presents the conceptual logic model for the RET. The figure includes a path directly from the treatment condition to the adherence outcome to reflect the fact the programs may impact adherence through mediators not specifically targeted, such as program satisfaction (see Chapter 2). The adherence outcome was measured as the percent of protocol compliance over a one-month period. It ranged from 0 to 100 and was measured 3 months after program completion. The mediators were measured 1 month

post-completion. A staff rating of the quality of the patient's social support (SS) was based on a structured interview by the staff member. Scores ranged from 0 to 10 (0 = no support, 3 = slight support, 6 = moderate support, 9 = strong support). The staff member could make finer gradations across the 0 to 10 metric. Staff received training about how to use the scale, which had concrete behavioral anchors for each major scale category. A comparable 0 to 10 staff rating was used for the quality of side effect coping strategies (CS), and a 0 to 10 self-report of adherence importance (AI) was obtained from each individual as well (0 = not at all important, 3 = slightly important, 6 = moderately important, 9 = very important). The treatment conditions are represented by dummy variables, one for the education group, D_E, and one for the education plus skills group, D_{ES}. The control condition is the reference group. Normally, there would be covariates to control for confounds, but I omit them to keep exposition more manageable.

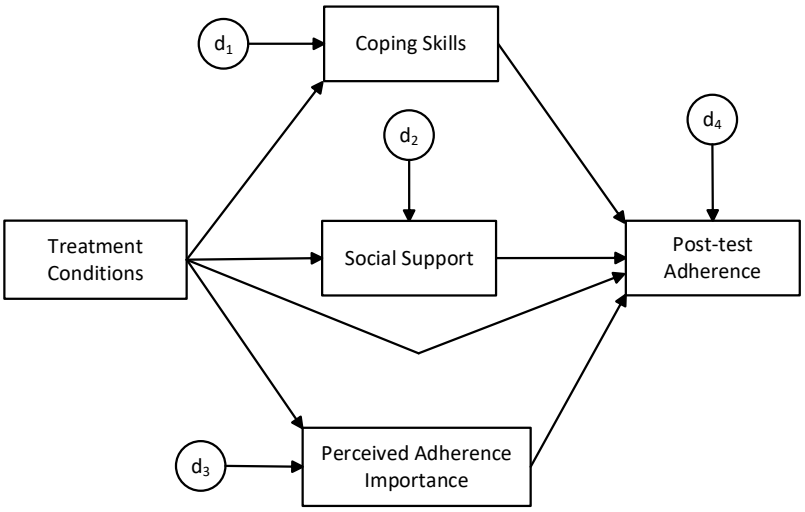


FIGURE 7.9. Logic model for adherence intervention

In the model, there are four endogenous variables and one exogenous variable. Given four endogenous variables, the core of the model is defined by four equations, one per endogenous variable. Each equation uses the endogenous variable as the outcome and each variable that has an arrow pointing directly to it as a predictor. The equations are:

$$\text{Adherence} = a_1 + b_1\,CS + b_2\,SS + b_3\,AI + b_4\,D_E + b_5\,D_{ES} + d_4 \qquad [7.8]$$

$$CS = a_2 + b_6\,D_E + b_7\,D_{ES} + d_1 \qquad [7.9]$$

$$SS = a_3 + b_8 D_E + b_9 D_{ES} + d_2 \quad [7.10]$$

$$AI = a_4 + b_{10} D_E + b_{11} D_{ES} + d_3 \quad [7.11]$$

There are no latent variables in this model, nor are there correlated disturbances. There are no causal relationships among the mediators. The logic model presumes that the three mediators, CS, SS, and AI, each influence adherence. The RET explicitly evaluates these assumptions and provides feedback to program designers about where the assumptions may fail. The programs are designed to impact CS, SS, and AI. The RET will inform us if the programs do so and, if not, which mediator(s) the programs failed to change. The RET also will inform us if the addition of active learning to the more traditional passive learning strategies creates added value to the above. This is far superior to conducting outcome only evaluations of program effects on adherence.

Once the RET data are collected, we calculate the correlations and covariances among the five variables. We hypothesize that the patterning of the observed 5X5 covariance and correlation matrices occurs because of the causal dynamics depicted in Figure 7.9. Using decompositional analysis, the model predicts the observed covariances should pattern themselves in specific ways. We calculate the actual values of the predicted covariances by the model using a robust maximum likelihood criterion via an iterative estimation process. We then evaluate if the observed covariances values do, in fact, pattern themselves in the ways predicted by the model. Specifically, we first evaluate global fit indices that reflect the correspondence between the predicted and observed covariances. These include the chi square index of fit, the standardized RMR, the RMSEA, the p value for close fit, and the CFI. We also examine localized indices of fit, including modification indices and standardized residuals.

Given reasonable support for the model using the weight-of-the-evidence approach, we proceed to examine and interpret the different path coefficients in the model to provide perspectives on the questions and issues inherent to RETs. For example, I might want to know if the program meaningfully affected each of the mediators and, if not, which mediators it failed to affect. This provides me with feedback about program facets that need to be strengthened. The statistical significance and magnitude of the path coefficients from the treatment conditions to the mediators are informative in this respect. I also want to know if the mediators I have targeted in my program are, in fact, relevant to the outcome. If a mediator is not, then I can consider streamlining the program by removing activities targeting that mediator. The statistical significance and magnitude of the path coefficients from the mediators to the outcomes are informative in this respect. There are many other facets of the SEM analysis that I can bring to bear to evaluate the program and I will outline these in future chapters.

APPENDIX A: SRMR DETAILS

Before elaborating on the SRMR, I note two points relevant to the underlying math. First, a correlation coefficient is often conceptualized as a covariance that has been standardized. The formula for a covariance between X and Y, using sample notation is

$cov_{XY} = r_{XY}\, s_X\, s_Y$

where r_{XY} is the correlation and s_X and s_Y are the standard deviations of X and Y, respectively. If I divide both sides of the equation by the product of $s_X\, s_Y$, I obtain

$$r_{XY} = \frac{cov_{XY}}{s_X s_Y}$$

which is the covariance "standardized" by the two standard deviations. The standardized RMR uses standardized covariances of this form, hence its focus is, in part, on correlations. However, it does not just use covariances in its calculations. It also uses the diagonal of the covariance matrix, which contains the variable variances. These also are standardized per the above, so they equal 1.0, and will always be perfectly reproduced by the predicted standardized variances because the latter also equal 1.0. The main message is that the SRMR focuses on both standardized covariances *and* standardized variances.

The second point I want to make is that there are many ways of calculating the mean disparity between predicted and observed covariances or correlations. Suppose I have the following two predicted and observed correlations:

Observed	Predicted	Difference
0.30	0.15	0.15
0.15	0.30	-0.15

One strategy is to sum the differences and divide by the number of differences, which in this case would be (0.15 + -0.15)/2 = 0. This mean is misleading because based on it, it appears there is perfect prediction when, in fact, there is not. The positive disparity canceled the negative disparity, yielding a mean of zero. We need to rid ourselves of the sign of the disparity. One way to do so is to use the absolute value of the disparity in the calculations. Another strategy is to square the disparities, average them, and then unsquare the average by taking the square root of it in to return to the original metric. The SRMR (and the CRMR) use this latter strategy. The result is a mean disparity, but it is a special type of mean; it is the square **root** of the **mean** of the **squared residuals**, hence

the term “standardized root mean squared residual.” When a mean disparity is calculated in this way, it tends to give more weight to the larger disparities than if we were to calculate the arithmetic mean of the absolute disparities.

With this as background, the traditional definition of the SRMR in a population is

$$\text{SRMR} = \sqrt{\frac{1}{t}\sum_{i\leq j}^{1}\frac{(\sigma_{ij}-\sigma'_{ij})^2}{\sqrt{(\sigma_{ii})(\sigma_{jj})}}}$$

where σ_{ij} is the population covariance between variables *i* and *j* (or variance if *i* = *j*), σ'_{ij} is the predicted population covariance (or variance) under the fitted model, and t = k(k+1)/2 or the number of nonredundant population variances and covariances, where k is the number of observed variables in the model. Note the term beneath the summation signifies the inclusion of both variances and the covariances. The formula basically executes a square root of the mean of the squared disparities, standardized.

For the correlation root mean squared residual, the population formula is

$$\text{CRMR} = \sqrt{\frac{1}{t-k}\sum_{i<j}^{1}(\rho_{ij}-\rho'_{ij})^2}$$

where ρ is the population correlation and ρ is the predicted population correlation under the fitted model. Note that the diagonal elements are excluded in this case.

The above is the traditional definition of the SRMR, but it becomes more complicated when a model has predicted and observed means in it, not just predicted and observed covariances. The equation needs to be modified to accommodate this case and the modification is described in Asparouhov and Muthén (2018). The mean residuals in the modified equation are standardized but they can take on any value and, consequently, can produce large SRMR values. They are important to take into account when fitting growth curve models or multi-group SEM models with mean structures. The traditional rules of thumb for what constitutes a reasonable value of the SRMR may not apply because most of those rules have been identified without mean structures as part of the model. As well, any SEM analysis in Mplus with missing data makes use of structural means, so the mean residuals are incorporated into SRMR in such cases. See Asparouhov and Muthén (2018) for discussion of how SRMR is modified by Mplus in different analytic scenarios. The method of Maydeu-Olivares (2017a) for obtaining unbiased estimates of SRMR and confidence intervals for them does not include models with mean structures. The definition of SRMR varies across SEM software, which also complicates specification of rules of thumb for asserting reasonable model fit based on it.

APPENDIX B: COMPARING NON-NESTED MODELS

There are many forms of non-nested models. In SEM, non-nested model comparisons typically compare two models that have the same variables and the same observations in them. In this Appendix, I first describe how to evaluate if two models are nested, non-nested, or equivalent. I then consider the Vuong test and a BIC test for comparing non-nested models.

Testing for Nesting and Model Equivalence

Bentler and Satorra (2010) propose a 4 step method to evaluate nesting and equivalence structures for two models, M1 and M2. If the two models have different degrees of freedom (df), let M1 be the model with the larger degrees of freedom. For nested models, this means M1 is the more restrictive model and M2 is the less restrictive model. Here are the steps:

Step 1. Conduct a standard SEM analysis of M1 using SEM software. Note the degrees of freedom for the model (df_{M1}), the predicted covariance matrix for the model ($Sigma_{M1}$) for the model, and the predicted means in a mean structure model.

Step 2. Conduct an SEM analysis of M2 using the same estimation method as in Step 1 (e.g., maximum likelihood) but instead of using the raw data as the data to be analyzed, input $Sigma_{M1}$ from Step 1 as the covariance matrix to be analyzed. Note the degrees of freedom for the model (df_{M2}) and the minimum of the fit function, $Fmin_{M2}$. Because $Fmin_{M2}$ is the minimum fit function using the predicted covariance matrix of M1 as data input, if the models are equivalent, the value of $Fmin_{M2}$ should be zero or deviate from it by a small amount due to rounding error and numerical approximations. Thus, we set an error criterion, called ε, to be a small number like 0.00001.

Step 3. Compute the difference in the degrees of freedom for the two models, $df_{DIFF} = df_{M1} - df_{M2}$.

Step 4. (a) If $df_{DIFF} > 0$ and $Fmin_{M2} < \varepsilon$, the models are nested
(b) If $df_{DIFF} = 0$ and $Fmin_{M2} < \varepsilon$, the models are equivalent
(c) If $Fmin_{M2} \geq \varepsilon$, M1 is not nested in nor equivalent to M2

Mplus has automated the Bentler and Satorra method and extended it to a wide range of variable types, including continuous variables, count variables and categorical

variables (see Asparouhov & Muthén, 2019b). For those already familiar with the Mplus programming language, I illustrate here the relevant Mplus syntax and output with two examples. If you do not know Mplus syntax, you will when you finish this book so you can then revisit this Appendix to evaluate model nesting as needed (Mplus programming is introduced in Chapter 11 and I also have resources for it on my website). For those who do not know Mplus syntax, you can skip to the section below called *The Vuong Test and Tests of BIC Differences*.

As a first step to using the Mplus strategy to determine if the models are nested, non-nested, or equivalent, you will want to conduct separate Mplus analyses of the two models in question so that you can identify the model with the larger degrees of freedom. Treat the model with the larger degrees of freedom as M1. If the degrees of freedom are the same, choose one of the models arbitrarily as M1.

I illustrate first a case where the models are nested. They are shown in Figure B.1. Model M1 posits that the treatment condition (T) impacts mediator 1 which, in turn, impacts a second mediator, which, in turn influences Y. Model M2 has these same dynamics but also includes a direct effect of mediator 1 on Y over and above its effect on Y through mediator 2. M1 is the more restricted model because, relative to M2, it omits a direct path from mediator 1 to Y. The mediators and Y are continuous whereas T is binary.

(M1) (M2)

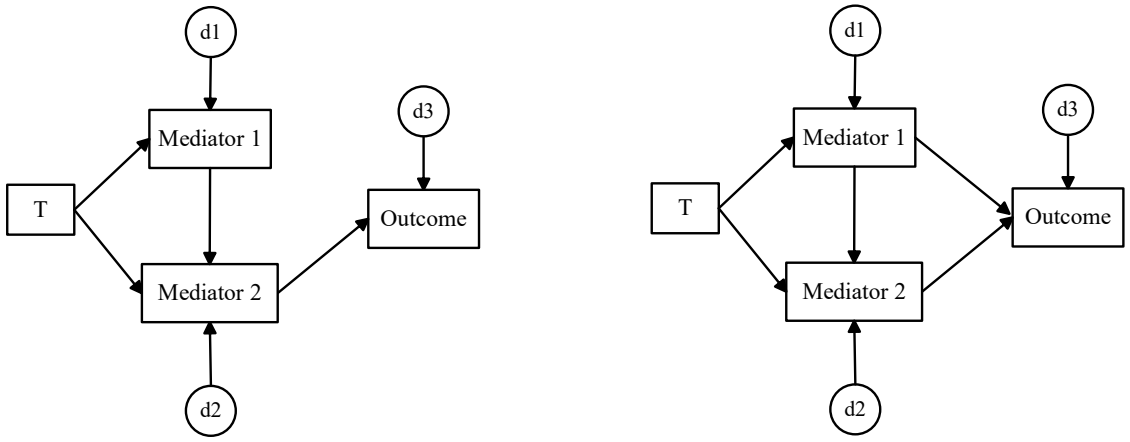


FIGURE B.1. Example 1 for Model Nesting Test

Here is the Step 1 Mplus code for evaluating if the models are nested:

```
1. TITLE: Step 1 model (M1)
2. DATA: file = maindata.dat;
3. VARIABLE:
4.   NAMES ARE t med1 med2 y ;
5. ANALYSIS: CONV=0.000001;
6. MODEL:
7.     med1 on t ;
8.     med2 on t med1 ;
9.     y on med2 ;
10. OUTPUT: ;
11. SAVEDATA: NESTED=step2.dat;
```

I number the lines for purposes of referencing here but the numbers do not appear in the Mplus code per se. Line 1 provides an arbitrary title for the analysis. Line 2 tells Mplus where to find the data file (it looks in the same folder that the input code is stored in), Lines 3 and 4 give names to the variables in the order in which they are input from the mainddata.dat file. Line 5 tells Mplus the convergence criterion to use in the analysis and Lines 6 to 9 define the M1 model using the ON statement from Mplus programming (see Chapter 11). Line 10 tells Mplus to give the default output. Line 11 tells Mplus to save the data for purposes of an evaluation of nesting and names the saved data file step2.dat. It will be saved in the same folder that the syntax input file is in.

After executing this program, you next execute the following Step 2 program:

```
1. TITLE: Step 2 model (M2)
2. DATA: file = maindata.dat;
3. VARIABLE:
4.   NAMES ARE t med1 med2 y ;
5. ANALYSIS: NESTED=step2.dat;
6. MODEL:
7.   med1 on t ;
8.   med2 on t med1 ;
9.   y on med2 med1;
10. OUTPUT: ;
```

This program has the same basic structure as the Step 1 program except the M2 model is specified on the MODEL lines (Lines 6 to 10), the SAVEDATA line is omitted, and the analysis line (Line 5) requests a nesting analysis and specifies the data file saved in the Step 1 program as input for the M1 information.

Here is the result published by Mplus in the output:

```
Nested Model Check

          Result                          Nested
          Fit Function Value         0.00000000
```

Mplus indicates the models are nested and provides the value of $Fmin_{M2}$ from the Bentler and Satorra method. Given this result, I can use the methods described in the main text to compare the two models, such as the chi square difference test.

As another example but with non-nested models, consider the two models in Figure B.2.

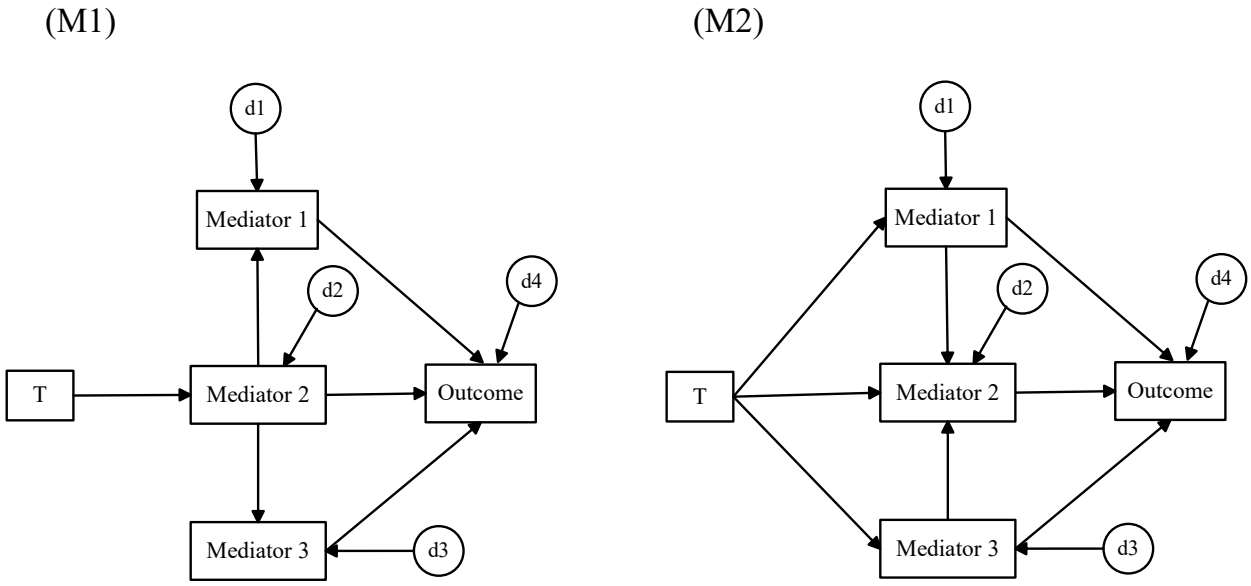


FIGURE B.2. Example 2 for Model Nesting Test

Here is the Step 1 Mplus code for evaluating if the models are nested:

```
TITLE: Step 1 model (M1)
DATA: file = maindata.dat;
VARIABLE:
 NAMES ARE med1 med2 med3 y t ;
ANALYSIS: CONV=0.000001;
MODEL:
  med1 ON med2 ;
  med2 ON t ;
  med3 ON med2 ;
  y ON med1 med2 med3 ;
  med1 ;  med2 ;  med3 ;  y ;
OUTPUT: ;
SAVEDATA: NESTED=step2.dat;
```

Here is the Step 2 code:

```
TITLE: Step 2 model (M2)
DATA: file = maindata.dat;
VARIABLE:
 NAMES ARE med1 med2 y ;
ANALYSIS: NESTED=step2.dat;
MODEL:
  med1 ON t ;
  med2 ON med1 med3 t ;
  med3 ON t ;
  y ON med1 med2 med3 ;
  med1 ; med2 ; med3 ; y ;OUTPUT: ;
```

Here is the Mplus result:

```
Nested Model Check

          Result                          Not Nested
          Fit Function Value              0.00484059
```

The models are not nested. If I want to formally compare them using statistical inference, I cannot use the traditional chi square difference test discussed in the main text. To accomplish such tests, I use the Vuong procedure described below.

Sometimes it is straightforward to identify the nesting status of two models and one can do so via a quick study of the model influence diagrams. The more experienced you get with SEM, the easier it becomes. However, for complex models even experienced SEMers may have difficulty determining the nesting status of two models. I recommend you use the Mplus `NESTED` method to evaluate the nesting and equivalence status of your models if you are going to make omnibus comparisons of them. Asparouhov and Muthén (2019b) discuss how to extend the approach to models with categorical/binary outcomes and mediators.

The Vuong Test and Confidence Intervals for BIC Differences

In some cases, it is possible to conduct inferential tests to compare non-nested models, but not always. Statistical methods for doing so have been outlined by Merkle, You and Preacher (2016) and implemented in an R library by Merkle and You called *nonest2*. One of the tests in this library is based on Vuong (1989) and the other is based on confidence intervals for BIC differences between the two models. For SEM models, the models being compared must focus on the same observed variables and the same individuals. Multivariate normality is assumed as well as maximum likelihood estimation. The underlying statistical theory is complex and I do not elaborate it here (see Merkel et al.

2016 for details and a discussion of how to extend the method beyond maximum likelihood modeling). I provide access to the tests on my website under the Programs tab in the program called *Vuong test and BIC CIs*.

In order two apply the tests to the two non-nested models, the models must have the property of being **distinguishable**; two models are indistinguishable if they provide the same fit to a population of interest, but not necessarily to an observed sample from that population (Merkle et al., 2016).[9] The Vuong and BIC difference tests apply only to cases where the non-nested models are distinguishable. If you are unsure of whether two models are distinguishable, the *nonnest2* package provides a test of distinguishability which can inform whether you move forward with the tests. The basic sequence of testing, adapted from Merkel et al. (2016), is as follows

1. Use the Mplus nested program to determine if the models are equivalent, nested, or non-nested.

2. If the models are nested, you can use a strategy of your choice from those in the main text to evaluate model differences (but see my additional discussion below).

3. If the models are not nested, test for model distinguishability using the test provided by *nonnest2*.

4. If the models are not distinguishable, then you cannot currently choose between them on empirical grounds. If they are distinguishable, then apply either the Vuong test or the BIC difference test to determine if the data favor one model over the other.

I applied the *nonnest2* program to the data I collected for the two non-nested models in Figure B.2. Here is the output for the Vuong test:

```
Variance test
  H0: Model 1 and Model 2 are indistinguishable
  H1: Model 1 and Model 2 are distinguishable
    w2 = 0.041,   p = 6.34e-05

Non-nested likelihood ratio test
  H0: Model fits are equal for the focal population
  H1A: Model 1 fits better than Model 2
    z = -2.858,   p = 0.998
  H1B: Model 2 fits better than Model 1
    z = -2.858,   p = 0.00213
```

[9] This is not the same as model equivalence because model equivalence is when the models yield the same fit in the population of interest as well as samples selected from the population.

The test for distinguishability (labeled as the `Variance test`) was statistically significant ($p < 0.001$) indicating that the null hypothesis that the two models are indistinguishable can be rejected. The Vuong test (labeled `Non-nested likelihood ratio test`) favored Model 2 over Model 1 ($p = 0.002$).

Here are the results for the difference in the AICs and the BICs for the two models:

```
Model 1
 AIC: 5601.326
 BIC: 5660.330

Model 2
 AIC: 5579.416
 BIC: 5646.850

95% Confidence Interval of AIC difference (AICdiff = AIC1 - AIC2)
  4.143 < AICdiff < 39.676

95% Confidence Interval of BIC difference (BICdiff = BIC1 - BIC2)
  -4.286 < BICdiff < 31.246
```

The confidence interval for the difference in BIC values contains the value of zero so the conclusion is that there is not sufficient evidence that Model 2 is better than Model 1. Using the rules of thumb for BIC model selection from the main text, in order to favor one model over the other, the 95% interval should not contain the value of 2, which also was the case. However, the difference in AIC values was statistically significant and the confidence interval did not contain the 2.2 standard from the main text. The BIC result also contradicts the Vuong test. The penalty function for model complexity is more harsh for the BIC index as compared to the AIC index and this likely explains the disparity in the two sets of results. In this case, I would probably conclude there is a difference between the models but the non-significant BIC difference would give me some pause.

The Vuong test (but not the BIC difference test) can be used to test fit differences between nested models but adjustments to the calculation of the test statistics are required. The program on my website gives you the option for such applications. The reason some researchers prefer it to the more traditional chi square difference test is because model fit is defined using the Kullback-Leibler distance between each model thereby not requiring that either model have good fit by and of itself. The traditional chi square difference test, by contrast, technically requires that the chi square for the unconstrained/unrestricted model be good fitting which sometimes is not the case. In other words, the traditional test can be misleading and yield incorrect p values when it subtracts a chi square for a bad fitting model from a chi square for another bad fitting

model when the models are nested. This dynamic does not apply to the Vuong test.

I provide a program on my website called *BIC difference CIs* that calculates confidence intervals for the difference between BICs for two nested models using bootstrapping. This program helps you take into account sample-to-sample fluctuations in BIC indices when comparing BICs in two nested models. See Preacher and Merkel (2012) for evaluations of how such bootstrapping approaches fares statistically. In general, they work reasonably well but for smaller N, the confidence interval coverage is slightly conservative (too wide), but not appreciably so. If the BIC confidence interval contains a value between 0 and 2 (inclusive), then one concludes there is not support that favors one of the models over the other.