(or at least one plausible) theory might be. In particular, there i information in the data you have collected that you may not be using tı its fullest advantage.

A second approach to an ill-fitting model is to use the availablı information to try to generate a more appropriate model. This is thı "art" of model modification—changing the original model to fit thı data. Although model modification is fraught with perils, I do no believe that anyone has ever "gotten it right" on the first attempt a model fitting. Thus, the art of model fitting is to understand the danger and try to account for them when you alter your model based oı empirical observations.

The principal danger in post hoc model modification is that thi procedure is exploratory and involves considerable capitalization oı chance. Thus, you might add a path to a model to make it fit the dat: only to find that you have capitalized on chance variation within you sample and the results will never be replicated in another sample. Therı are at least two strategies for minimizing this problem.

First, try to make model modifications that have some semblance o theoretical consistency (bearing in mind Steiger's comments about ou ability to rationalize). If there are 20 studies suggesting that job satisfac tion and job performance are unrelated, do not hypothesize a patl between satisfaction and performance just to make your model fit Second, as with any scientific endeavor, models are worthwhile onl; when they can be replicated in another sample. Post hoc modification to a model should always be (a) identified as such and (b) replicated iı another sample.[2]

Notes

[1] It also helps to remember that in path diagrams, the hypothesized causal "flow is traditionally from left to right (or top to bottom); that is, the independent (exogenous variables or predictors are on the left (top), and the dependent (endogenous) variable or criteria are on the right (bottom).

[2] Note that the use of a holdout sample is often recommended for this purpose. Se aside 25% of the original sample, then test and modify the model on the remaining 75% When you have a model that fits the data on the original 75%, test the model on th remaining 25%. Although this procedure does not always result in replicated finding: it can help identify which paths are robust and which are not.

CHAPTER 3 Assessing Model Fit

Perhaps more has been written about the assessment of model fit than any other aspect of structural equation modeling. Indeed, many researchers are attracted to structural equation modeling techniques because of the availability of global measures of model fit (Brannick, 1995). In practice, such measures often are used as an omnibus test of the model whereby one first assesses global fit before proceeding to a consideration of the individual parameters composing the model (Jöreskog, 1993). A variety of fit indices are currently available to researchers wishing to assess the fit of their models, and it is instructive to consider exactly what we mean when we claim that a model "fits" the data.

At least two traditions in the assessment of model fit are apparent (Tanaka, 1993): the assessment of the *absolute* fit of the model and the assessment of the *comparative* fit of the model. The assessment of the comparative fit of the model may be further subdivided into the assessment of comparative fit and *parsimonious* fit. The assessment of absolute fit is concerned with the ability of the model to reproduce the actual covariance matrix. The assessment of comparative fit is concerned with comparing two or more competing models to assess which provides the better fit to the data.

The assessment of parsimonious fit is based on the recognition that one can always obtain a better fitting model by estimating more parameters. (At the extreme, one can always obtain a perfect fit to the data by estimating the just-identified model containing all possible parameters.) Thus, the assessment of parsimonious fit is based on the idea of a

"cost-benefit" trade-off and asks: Is the cost (loss of a degree of freedo)
worth the additional benefit (increased fit) of estimating more paran -
ters? Although measures of comparative and absolute fit will alwa s
favor more complex models, measures of parsimonious fit provide a
"fairer" basis for comparison by adjusting for the known effects f
estimating more parameters.

In the remainder of this chapter, I present the most commonly us d
indices for assessing absolute, comparative, and parsimonious fit. f
necessity, the presentation is based on the formulae for calculating the e
indices; however, it should be remembered that structural equati n
modeling programs such as LISREL do the actual calculations for yc .
The researcher's task, therefore, is to understand what the fit indices a e
measuring and how they should be interpreted. The chapter conclud s
with some recommendations on assessing the fit of models.

Absolute Fit

Tests of absolute fit are concerned with the ability to reproduce t e
correlation/covariance matrix. As shown in the previous chapter, pe -
haps the most straightforward test of this ability is to work backwarc ,
that is, from the derived parameter estimates, calculate the impli l
covariance matrix and compare it, item by item, with the observ l
matrix. There are at least two major stumbling blocks to this procedu .

First, the computations are laborious when models are even mode -
ately complex. Second, there are no hard and fast standards of ho ,
"close" the implied and observed covariance matrices must be to clai n
that the model fits the data. For example, if the actual correlatic n
between two variables is 0.45 and the correlation implied by the mod l
is 0.43, does the model fit the data or not?

Early in the history of structural equation modeling, researche s
recognized that for some methods of estimation a single test statist :
(distributed as χ^2) was available to test the null hypothesis that

$$\Sigma = \Sigma(\Theta)$$

where Σ is the population covariance matrix and $\Sigma(\Theta)$ is the covarian :
matrix implied by the model (Bollen & Long, 1993). The developme t
of the χ^2 test statistic for structural equation models proceeds direct
from early accounts of path analysis in which the attempt was to speci

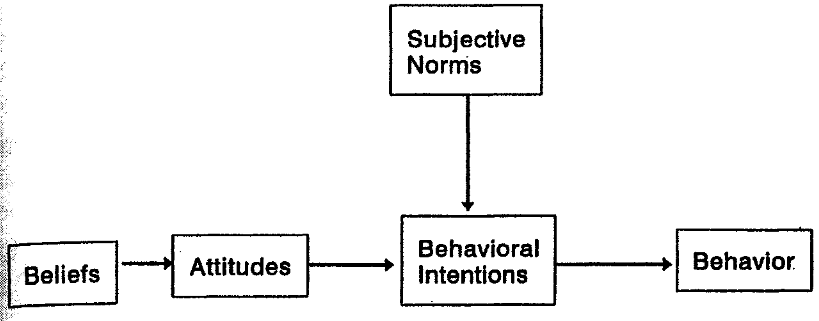

Figure 3.1.

a model that reproduced the original covariance matrix (e.g., Blalock, 1964). In the obverse of traditional hypothesis testing, a nonsignificant χ^2 implies that there is no significant discrepancy between the covariance matrix implied by the model and the population covariance matrix. Hence, a nonsignificant χ^2 indicates that the model "fits" the data in that the model can reproduce the population covariance matrix.

The test is distributed with degrees of freedom equal to

$$1/2(q)(q+1)-k$$

where q = the number of variables in the model and
k = the number of estimated parameters.

For example, the Fishbein and Ajzen (1975) model introduced in Chapter 2 and repeated in Figure 3.1 is based on five variables and incorporates four paths:

1. Behavioral intentions predict behavior.
2. Attitudes predict behavioral intentions.
3. Subjective norms predict behavioral intentions.
4. Beliefs predict attitudes.

The model therefore has

$$df = 1/2(5)(6)-4$$
$$df = 1/2(30)-4$$

$$df = 15 - 4$$
$$df = 11.$$

Although the test is quite simple (indeed, LISREL calculates it fo you), there are some problems with the χ^2 test in addition to the logica problem of being required to accept the null hypothesis. First, th approximation to the χ^2 distribution occurs only for large samples (e.g, $N \geq 200$). Second, just at the point where the χ^2 distribution becomes tenable assumption, the test has a great deal of power. Recall that th test is calculated as $N - 1 \times$ (the minimum of the fitting function) therefore, as N increases, the value of χ^2 must also increase. Thus, for minimum fitting function of .5, the resulting χ^2 value would be 99.5 fo $N = 200$, 149.5 for $N = 300$, and so on. This makes it highly unlikel that you will be able to obtain a nonsignificant test statistic with larg sample sizes.

Although not typically presented as fit indices, the LISREL outpu also includes some indications of model fit, including the following:

1. the noncentrality parameter (estimated as $\chi^2 - df$ and used in the calcu lation of some fit indices),
2. the 90% confidence interval for the noncentrality parameter,
3. the minimum of the fitting function,
4. the discrepancy function (used in calculating other fit indices), and
5. the 90% confidence interval for the discrepancy function.

This output is presented largely for the information of the researcher and the values presented typically have no straightforward interpreta tion. As noted above, however, much of this information is used in th calculation of other fit indices, as explained below.

Given the known problems of the χ^2 test as an assessment of mode fit, numerous alternate fit indices have been proposed. Gerbing an Anderson (1992, p. 134) describe the ideal properties of such indices to

1. indicate degree of fit along a continuum bounded by values such as 0 an 1, where 0 represents a lack of fit and 1 reflects perfect fit.
2. be independent of sample size. . . . and
3. have known distributional characteristics to assist interpretation an allow the construction of a confidence interval.

With the possible exception of the root mean squared error of approximation (Steiger, 1990; see below), thus far none of the fit indices commonly reported in the literature satisfy all three of these criteria; the requirement for known distributional characteristics is particularly lacking. The current version of LISREL (LISREL VIII) reports 18 such indices of model fit, only four of which address the question of absolute fit.

The simplest fit index provided by LISREL is root mean squared residual (RMR). This is the square root of the mean of the squared discrepancies between the implied and observed covariance matrices. The lower bound of the index is 0, and low values are taken to indicate good fit. The index, however, is sensitive to the scale of measurement of the model variables. As a result, it is difficult to determine what a "low" value actually is. LISREL therefore also now provides the standardized RMR, which has a lower bound of 0 and an upper bound of 1. Generally for this index, values less than 0.05 are interpreted as indicating a good fit to the data.

LISREL also reports the root mean squared error of approximation (RMSEA) developed by Steiger (1990). Similar to the RMR, the RMSEA is based on the analysis of residuals, with smaller values indicating a better fit to the data. Steiger (1990) suggests that values below 0.10 indicate a good fit to the data, and values below 0.05 a very good fit to the data. Values below 0.01 indicate an outstanding fit to the data, although Steiger (1990) notes that these values rarely are obtained.

Unlike all other fit indices discussed in this chapter, the RMSEA has the important advantage of going beyond point estimates to the provision of 90% confidence intervals for the point estimate. Moreover, LISREL also provides a test of the significance of the RMSEA by testing whether the value obtained is significantly different from 0.05 (the value that Steiger suggests indicates a very good fit to the data). Perhaps because of its recent inclusion in the LISREL program, the RMSEA is not frequently reported in the literature; however, the advantages of confidence intervals and formal hypothesis testing available with this index will likely increase its use as a measure of model fit.

The goodness-of-fit index (GFI) is based on a ratio of the sum of the squared discrepancies to the observed variances (for generalized least squares, the maximum likelihood version is somewhat more complicated). The GFI ranges from 0 to 1, with values exceeding 0.9 indicating a good fit to the data. It should be noted that this guideline is based on experience. Like many of the fit indices that will be presented, the GFI has no known sampling distribution. As a result, "rules" about when an index indicates a good fit to the data are highly arbitrary and should b treated with caution.

Finally, the adjusted goodness-of-fit index (AGFI) adjusts the GFI fo degrees of freedom in the model. The AGFI also ranges from 0 to 1 with values above 0.9 indicating a good fit to the data. A discrepanc between the GFI and AGFI typically indicates the inclusion of trivi (i.e., small) and often nonsignificant parameters.

Early in the discussion of fit indices, researchers proposed assessin the fit of the model by taking the ratio of the χ^2 and its degrees c freedom. Unfortunately, conflicting standards of interpretation for thi index abound (Medsker et al., 1994). For example, χ^2/df ratios of les than 5 have been interpreted as indicating a good fit to the data as hav ratios between 2 and 5, with ratios less than 2 indicating overfittin Interpretative standards for the χ^2/df ratio have very little justificatio other than modelers' experience, and as a result, use of the index appea to be unwise and in decline (Kelloway, 1996).

Researchers also have reported the coefficient of determination fo the model as an index of overall fit. Structural equation modelin programs typically report R^2 values for each endogenous variable as we as an overall coefficient of determination for the model. Althoug researchers have interpreted these indices as measures of model fit, the clearly do not address the question of whether the model can reproduc the covariance matrix. Rather, the model coefficient of determinatio and the R^2 values for individual endogenous variables are measures c variance accounted for, rather than measures of model fit (Medske et al., 1994). It is quite possible to have a well-fitting model that explai only a modest amount of variance in the endogenous variables.

With the exception of R^2 values, the indices discussed thus far asse whether or not the model as a whole provides an adequate fit to th data. More detailed information can be acquired from tests of specifi parameters composing the model. James and colleagues (1982) describ two types of statistical tests used in structural equation modelin Condition 9 and Condition 10 tests. A Condition 10 test assesses th overidentifying restrictions placed on the model. The most commo example of a Condition 10 test is the χ^2 likelihood test for goodness c fit. Using the term "test" loosely to include fit indices with unknow distributions, the fit indices discussed above would also qualify a Condition 10 tests.

In contrast, Condition 9 tests are tests of the specific paramete composing the model. Programs such as LISREL commonly report bot the parameter and the standard error of estimate for that parameter. Th ratio of the parameter to its standard error is also reported as a *t* test. In practice, however, and given the large sample sizes involved in structural equation modeling, these *t* values are interpreted using the critical values for the *Z* test. That is, values above 1.96 are significant at the $p < .05$ level. A Condition 9 test, therefore, assesses whether parameters predicted to be nonzero in the structural equation model are in fact significantly different from zero.

Again, it is important to note that consideration of the individual parameters composing the model is important for assessing the accuracy of the model. The parameter tests are not, in and of themselves, tests of model fit. Two likely results in testing structural equation models are that (a) a proposed model fits the data even though some parameters are nonsignificant and/or (b) a proposed model fits the data but some of the specified parameters are significant and opposite in direction to that predicted. In either case, the researcher's theory is disconfirmed even though the model may provide a good absolute fit to the data. The fit of the model has nothing to say about the validity of the individual predictions composing the model. One must move beyond the assessment of global fit to truly evaluate the results of structural equation modeling (Jöreskog, 1993).

Comparative Fit

Perhaps because of the problems inherent in assessing the absolute fit of a model to the data, researchers increasingly have turned to the assessment of comparative fit. The question of comparative fit deals with whether the model under consideration is better than some competing model. For example, many of the indices discussed below are based on choosing a model as a "baseline" and comparing the fit of theoretically derived models to the baseline model.

In some sense, all tests of model fit are based on a comparison of models. The tests discussed previously implicitly compare the theoretical model against the just-identified model. Recall that the just-identified model consists of all possible recursive paths between the variables. As a result, the model has 0 degrees of freedom (because the number of estimated paths is the same as the number of elements in the covariance matrix) and always provides a perfect fit to the data.

Indices of comparative fit are based on the opposite strategy. Rather than comparing against a model that provides a perfect fit to the data,

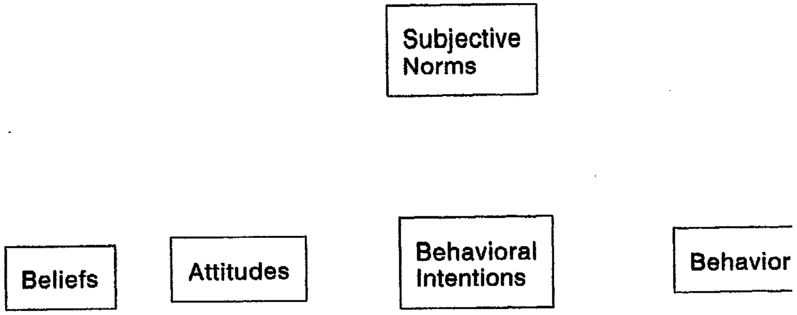

Figure 3.2.

indices of comparative fit typically choose as the baseline a model th t is known a priori to provide a poor fit to the data. The most comm n baseline model is the "null" or "independence" model (the two ter s are used interchangeably; LISREL printouts refer to the independen e model, but much of the literature makes reference to the null mode . The null model is a model that specifies no relationships between t e variables composing the model. That is, if one were to draw the pa n model for the null model, it would have no paths connecting t e variables (see Figure 3.2).

For example, Bentler and Bonett (1980) have suggested a normed t index (NFI), defined as

$$(\chi^2_{\text{indep}} - \chi^2_{\text{model}})/\chi^2_{\text{indep}}$$

The NFI ranges from 0 to 1, with values exceeding 0.9 indicating a go d fit.[1] As Bentler and Bonett (1980) point out, the NFI indicates t e percentage improvement in fit over the baseline independence mod l. Thus, an NFI of 0.90 means that the model is 90% better fitting th n the null model. Although the NFI is widely used, it may underestim e the fit of the model in small samples.

The nonnormed fit index (NNFI) uses a similar logic but adjusts t e normed fit index for the number of degrees of freedom in the mod l. The NNFI is given by

$$(\chi^2_{\text{indep}} - df_{\text{indep}}/df_{\text{model}}\,\chi^2_{\text{model}})/(\chi^2_{\text{indep}} - df_{\text{model}})$$

Although this correction reduces the problem of underestimating fit it introduces a new complication in that it may result in numbers outs le of the 0 to 1 range. That is, the NNFI results in numbers with a lower bound of 0 but an upper bound greater than 1. Higher values of the NNFI indicate a better fitting model, and it is common to apply the 0.90 rule as indicating a good fit to the data.

Bollen's (1989) incremental fit index (IFI) reintroduces the scaling factor, so that IFI values range between 0 and 1, with higher values indicating a better fit to the data. The IFI is given by

$$(\chi^2_{\text{indep}} - \chi^2_{\text{model}})/(\chi^2_{\text{indep}} - df_{\text{model}})$$

Bentler (1990) proposed a comparative fit index (CFI) based on the noncentral χ^2 distribution. The CFI also ranges between 0 and 1, with values exceeding 0.90 indicating a good fit to the data. The CFI is based on the noncentrality parameter, which can be estimated as $\chi^2 - df$. Thus, the CFI is given by

$$1 - [(\chi^2_{\text{model}} - df_{\text{model}})/(\chi^2_{\text{indep}} - df_{\text{indep}})]$$

Marsh and colleagues (1988) proposed a relative fit index (RFI) defined as

$$\frac{(\chi^2_{\text{indep}} - \chi^2_{\text{model}}) - [\text{df}_{\text{indep}} - (\text{df}_{\text{model}}/\text{n})]}{\chi^2_{\text{indep}} - (\text{df}_{\text{indep}}/\text{n})}$$

Again, the RFI ranges between 0 and 1, with values approaching unity indicating a good fit to the data. The use of 0.90 as an indicator of a well-fitting model is also appropriate with this index.

Finally, Cudeck and Browne (1983) suggested the use of the cross-validation index as a measure of comparative fit. Cross-validation of models is well established in other areas of statistics (e.g., regression analyses; Browne & Cudeck, 1993; Cudeck & Browne, 1983). Traditionally, cross-validation required two samples: a calibration sample and a validation sample. The procedure relied on fitting a model to the calibration sample and then evaluating the discrepancy between the covariance matrix implied by the model to the covariance matrix of the validation sample. If the discrepancy was small, then the model was judged to fit the data in that it cross-validated to other samples.

The obvious practical problem with this strategy is the requirement for two samples. Browne and Cudeck (1989) suggested a solution to the problem by estimating the expected value of the cross-validation index using only data from a single sample. Although the mathematics of the expected value of the cross-validation index (ECVI) will not be presented here (the reader is referred to the source material cited above), the ECVI is thought to estimate the expected discrepancy (i.e., difference between the implied and actual covariance matrices) over all possible calibration samples. The ECVI has a lower bound of zero but no upper bound. Smaller values indicate better-fitting models. In addition to the point estimate of the ECVI, LISREL provides both the confidence intervals for the estimate and the ECVI values for the independence (null) and saturated (just-identified) models.

Parsimonious Fit

Parsimonious fit indices are concerned primarily with the cost-benefit trade-off of fit and degrees of freedom. It is not surprising that several of the indices can be calculated by adjusting other indices of fit for model complexity. For example, James and colleagues (1982) have proposed the parsimonious normed fit index (PNFI), which adjusts the NFI for model parsimony. The PNFI is calculated as

$$(df_{\text{model}} / df_{\text{indep}}) \times \text{NFI}$$

Similarly, the parsimonious goodness-of-fit index (PGFI) adjusts the GFI for the degrees of freedom in the model and is calculated as

$$1 - (P/N) \times \text{GFI}$$

where P = the number of estimated parameters in the model and
N = the number of data points.

Both the PNFI and the PGFI range from 0 to 1, with higher values indicating a more parsimonious fit. Unlike the other fit indices we have discussed, there is no standard for how "high" either index should be to indicate parsimonious fit. Indeed, neither the PNFI nor the PGFI will likely reach the 0.90 cutoff used for other fit indices. Rather, these indices are best used to compare two competing theoretical models; that is, they would calculate an index of parsimonious fit for each model and choose the model with the highest level of parsimonious fit.

The Akaike Information Criterion (AIC) and Consistent Akaike Information Criterion (CAIC) (Akaike, 1987; Bozdogan, 1987) are also measures of parsimonious fit that consider both the fit of the model and the number of estimated parameters. AIC is defined as

$$\chi^2_{\text{model}} - 2df_{\text{model}}$$

and CAIC is defined as

$$\chi^2_{\text{model}} - (\ln N + 1)df_{\text{model}}$$

where N is the number of observations. For both indices, smaller values indicate a more parsimonious model. Neither index, however, is scaled to range between 0 and 1, and there are no conventions or guidelines to indicate what "small" means. Like the PNFI and PGFI, interpretation of the AIC and CAIC is based on comparing competing models and choosing the model that shows the most parsimony. McDonald and Marsh (1990) note that following this strategy using the AIC might result in overly complex models (i.e., although the AIC adjusts for parsimony, the adjustment is not sufficient to overcome a bias in favor of more complex models).

Nested Model Comparisons

As should be apparent at this point, the assessment of model fit is not a straightforward task. Indeed, from the discussion thus far, it should be clear that there are at least three views of what "model fit" means:

1. the absolute fit of the model to the data,
2. the fit of a model to the data relative to other models, or
3. the degree of parsimonious fit of the model relative to other models.

Given the problems inherent in assessing model fit, it is commonly suggested that models of interest be tested against reasonable alternative models. If we cannot show that our model fits the data perfectly, we can at least demonstrate that our model fits better than some other reasonable model. Although this may sound suspiciously like the question of comparative fit, recall that indices of comparative fit are based on comparison with the independence model, which is purposely defined as a model that provides a poor fit to the data. In contrast, the procedures we are about to discuss are based on comparing two plausible models of the data.

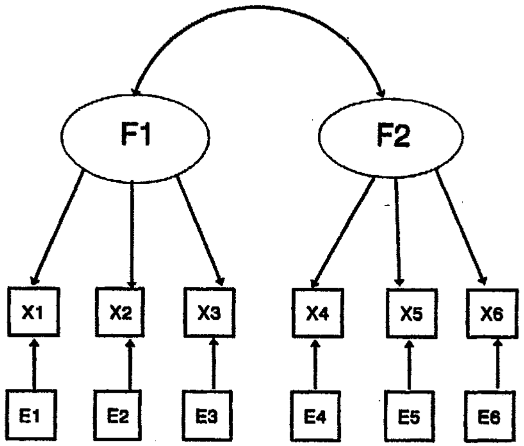

Figure 3.3.

Many such plausible and rival specifications exist. For example consider the case of a confirmatory factor analysis model shown i Figure 3.3. The model suggests that there are two common factors which are correlated, causing six indicators. Plausible rival hypothese might include a model suggesting two orthogonal common factor (Figure 3.4) or a unidimensional model (Figure 3.5).

In the case of path analyses, an alternative specification of th Fishbein and Ajzen (1975) model might include the hypothesis tha subjective norms about a behavior influence both attitudes and behavioral intentions (see Figure 3.6). Support for this modification has beefound in some tests of Fishbein and Ajzen-based models (e.g., Fullagar, McCoy, & Shull, 1992; Kelloway & Barling, 1993). Although we are always interested in whether the model(s) fit the data absolutely, we also may be interested in which of these competing specifications provides the best fit to the data.

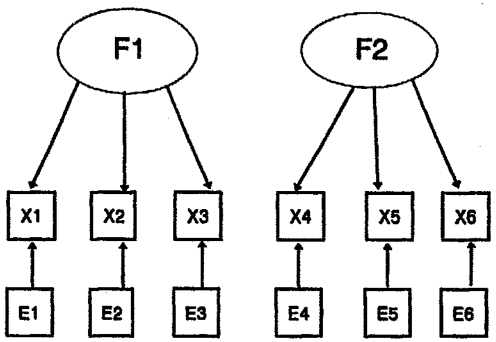

Figure 3.4.

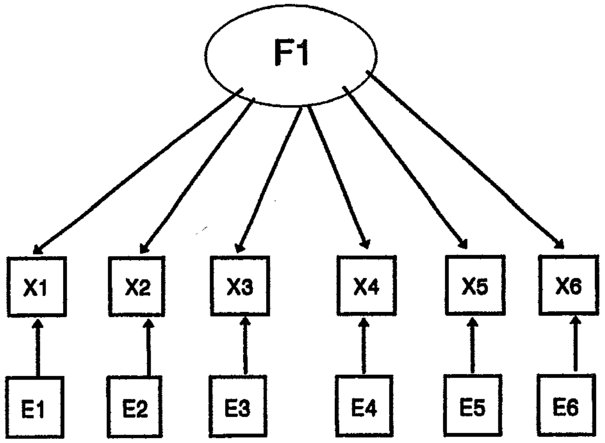

Figure 3.5.

If the alternative models are in hierarchical or nested relationships, then these model comparisons may be made directly. A nested relationship exists between two models if one can obtain the model with the fewest number of free parameters by constraining some or all of the parameters in the model with the largest number of free parameters.

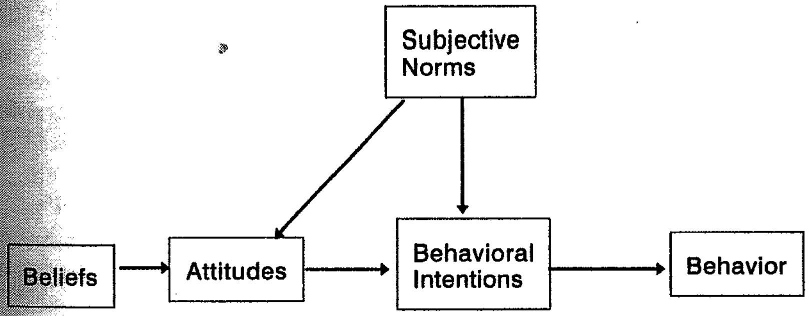

Figure 3.6.

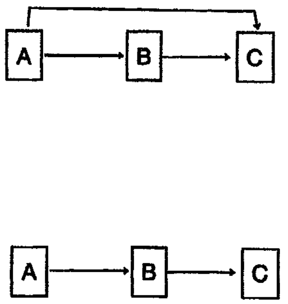

Figure 3.7.

That is, the model with the fewest parameters is a subset of the model with more parameters.

For example, consider two factor models of an eight-item test. Model A suggests that four items load on each factor and that the two factors are correlated (oblique). Model B suggests that the same four items load on each factor but that the two factors are orthogonal. In this case, Model B is nested in Model A. By taking Model A and constraining the interfactor correlation to equal 0, one obtains Model B. The nesting sequence results from the observation that Model B is composed of all the same parameters as Model A, with the exception of the interfactor correlation (which is not estimated in Model B).

Similarly, for a path model example, consider the model shown in Figure 3.7. The model at the top suggests that *A* predicts both *B* and C directly. In contrast, the model at the bottom suggests that *A* predicts *B*, which in turn predicts C. Again, these models stand in a nested sequence. By deleting the direct prediction of C from *A* from the first model, we obtain the second model (ergo, the second model is nested within the first).

When two models stand in a nested sequence, the difference between the two may be directly tested with the $\chi^2_{\text{difference}}$ test. The difference between the χ^2 values associated with each model is itself distributed as χ^2 with degrees of freedom equal to the difference in degrees of freedom for each model. For example, assume that the two-factor model with correlated factors generated $\chi^2(19) = 345.97$. Constraining the interfactor correlation between the two models to equal 0 results in $\chi^2(20) = 347.58$. The $\chi^2_{\text{difference}}$ is

$$347.58 - 345.97 = 1.61$$

which is distributed with

$$20 - 19 = 1 \text{ degree of freedom}$$

Because the critical value for χ^2 with 1 degree of freedom is 3.84 and the obtained value is less than the critical value, we conclude that there is no significant difference between the two models. By inference, then, the model hypothesizing two oblique factors is overly complex. That is, the additional parameter (the interfactor correlation) did not result in a significant increase in fit.

In this simple example, the models differ in only one parameter, and the results of the $\chi^2_{\text{difference}}$ test probably do not provide any information beyond the tests of individual parameters discussed at the beginning of this chapter. When models differ in more than one parameter, however, the $\chi^2_{\text{difference}}$ test is a useful omnibus test of the additional parameters that can be followed up by the Condition 9 tests of specific parameters.

Before leaving the subject of nested model comparisons, it is important to note that the test is valid only when the models stand in nested sequence. If the nesting sequence is not present, use of the $\chi^2_{\text{difference}}$ test is inappropriate. The key test of whether Model A is nested in Model B is whether all the relationships constituting Model A exist in Model B. That is, if Model B simply adds relationships to Model A, then the two models are nested. If there are other differences (e.g., Model B is obtained by deleting some parameters from Model A and adding some others), then the models are not nested.

Although the test is not conclusive, it should be apparent that given two models in nested sequence, the model with the fewest parameters will always provide the worst fit to the data (i.e., be associated with the highest χ^2 value and the larger degrees of freedom). Moreover, the degrees of freedom for the $\chi^2_{\text{difference}}$ test should always equal the number of additional paths contained in the more complex model.

Model Respecification

As noted earlier, perhaps no aspect of structural equation modeling techniques is more controversial than the role of model respecification. Despite the controversy (see, for example, Brannick, 1995; Kelloway, 1995; Williams, 1995), structural equation programs such as LISREL commonly provide the researcher with some guidelines for finding sources of model misspecification. That is, given that the proposed model does not fit the data, is there anything we can do to improve the fit of the model?

Two sources of information are particularly valuable. First, the test of model parameters discussed at the beginning of this chapter provide some information about which parameters are contributing to the fit o the model and which parameters are not making such a contribution Theory trimming (Pedhazur, 1982) is a common approach to mode improvement. It essentially consists of deleting nonsignificant paths from the model to improve model fit.

Although these tests provide information about the estimated mode parameters, LISREL also provides information about the nonestimated parameters. That is, the use of a theory-trimming approach asks, "What parameters can be deleted from the model?"; one can also adopt a theory-building approach that asks, "What parameters should be added to the model?"

These tests are technically known as Lagrange multiplier tests but are referred to in LISREL as the modification indices. For each parameter in the model that is set to zero, LISREL calculates the decrease in the model χ^2 that would be obtained from estimating that parameter. The amount of change in the model χ^2 is referred to as the modification index for that parameter.

Obviously, there is a trade-off between estimating more parameters (with the corresponding loss of degrees of freedom) and improving the fit of the model. Commonly, we would estimate any parameter that is associated with a modification index greater than 5.0; however, this rough guideline should be used with caution for several reasons.

First, recall that such specification searches are purely exploratory in nature. In contrast with the other parameters in the model, which are based on theory or previous research, parameters added on the basis of the modification indices (or, indeed, deleted on the basis of significance tests) may be reflecting sample-specific variance. The modifications made to the model following these procedures may not generalize to other samples.

Second, the process of theory trimming or theory building is analogous to the procedures of stepwise regression through either backward elimination (theory trimming) or forward entry (theory building). As such, both procedures are based on univariate procedures in which each parameter is considered in isolation. As a result, both theory trimming and theory building are based on a large number of statistical tests, with a corresponding inflation of Type I error rates. Moreover, the tests may

be misleading in that adding a parameter to a model based on the modification indices may change the value of parameters already in the model (i.e., making theoretically based parameters nonsignificant).

Third, even when the modification indices are greater than 5.0, the improvement in model fit obtained from freeing parameters may be trivial to the model as a whole. For example, if the overall χ^2 for the model is 349.23, then it is questionable whether the improvement in fit (reduction of the χ^2 by 5.0) is worth the dangers of adding a parameter based on the modification indices.

Toward a Strategy for Assessing Model Fit

As has been evident throughout this discussion, the assessment of model fit is a complex question. Numerous fit indices have been proposed, each with a slightly different conception of what it means to say a model "fits the data." The literature is in agreement, however, on several fundamental points that provide the basis for a strategy of model testing.

First, the focus of assessing model fit almost invariably should be on comparing the fit of competing and theoretically plausible models. Simply stated, the available techniques for assessing model fit do a better job of contrasting models than they do of assessing one model in isolation. The researcher's task, then, is to generate plausible rival specifications and test them. Ideally, such rival specifications will consist of nested models allowing the use of direct methods of comparison such as the $\chi^2_{\text{difference}}$ test.

Second, and in a similar vein, rather than relying on modification indices and parameter tests to guide the development of models, researchers should be prepared a priori to identify and test the sources of ambiguity in their models. I elaborate on this theme in subsequent chapters, where examples of the major types of structural equation models are presented.

Third, given the varying definitions of model fit presented above, it is incumbent on researchers to use multiple measures of fit. As Loehlin (1987) notes, the use of multiple fit indices may place the researcher in the position of an individual with seven watches: If they all agree, then you know what time it is, but if they don't, you are no better off than if you had no watch. I suggest that the situation with regard to assessing model fit is not that serious. Understanding what each of the various fit indices means would suggest that, at a minimum, researchers would want to consider the issues of absolute fit, comparative fit, and parsimonious fit for each model tested. Fit indices therefore should be chosen so as to reflect each of these concerns (i.e., choosing one or two indices of each type of fit).

Finally, it is important for researchers to recognize that "model fit" does not equate to "truth" or "validity." The fit of a model is, at best, a necessary but not sufficient condition for the validity of the theory that generated the model predictions. Although the question of model fit is important, it is by no means the most important or only question we should ask about our data.

Note

[1] Recall the previous discussion about the arbitrariness of such guidelines and the resultant need for cautious interpretation.

CHAPTER 4 *Using LISREL*

Having considered the general approach to be used in structural equation modeling, it is now time to consider the specifics of using LISREL to estimate and evaluate such models. It should be noted that other programs (e.g., EQS and EZPATH) are available to evaluate structural equation models. LISREL remains, however, a popular and widely available software package for structural equation modeling.

LISREL works by defining eight matrices; within each matrix are free and fixed parameters. Free parameters are unknowns to be estimated by the program, whereas fixed parameters are set to some predetermined value (usually zero).

For example, a typical LISREL matrix might contain three rows and two columns, as follows.

	K1	K2
X1	Free	Fixed
X2	Free	Fixed
X3	Fixed	Free

In this case, matrix elements (1, 1), (2, 1) and (3, 2) are going to be freely estimated by the program. Matrix elements (1, 2), (2, 2), and (3, 1) are fixed (set to zero). Researchers would specify the model to be tested by manipulating these matrices and whether their elements were fixed or free. Using the matrix formulation of LISREL, the researcher's task is to translate the model (i.e., the path diagram) into the LISREL matrices.