Methods of Psychological Research Online 2003, Vol.8, No.2, pp. 23-74
Internet: http://www.mpr-online.de

Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures

Karin Schermelleh-Engel[1] **and Helfried Moosbrugger**
Goethe University, Frankfurt

Hans Müller
University of Erfurt

For structural equation models, a huge variety of fit indices has been developed. These indices, however, can point to conflicting conclusions about the extent to which a model actually matches the observed data. The present article provides some guidelines that should help applied researchers to evaluate the adequacy of a given structural equation model. First, as goodness-of-fit measures depend on the method used for parameter estimation, maximum likelihood (ML) and weighted least squares (WLS) methods are introduced in the context of structural equation modeling. Then, the most common goodness-of-fit indices are discussed and some recommendations for practitioners given. Finally, we generated an artificial data set according to a "true" model and analyzed two misspecified and two correctly specified models as examples of poor model fit, adequate fit, and good fit.

Keywords: *Structural equation modeling, model fit, goodness-of-fit indices, standardized residuals, model parsimony*

In structural equation modeling (SEM), a model is said to fit the observed data to the extent that the model-implied covariance matrix is equivalent to the empirical covariance matrix. Once a model has been specified and the empirical covariance matrix is given, a method has to be selected for parameter estimation. Different estimation methods have different distributional assumptions and have different discrepancy functions to be minimized. When the estimation procedure has converged to a reasonable solu-

[1] Correspondence concerning this article should be addressed to Dr. Karin Schermelleh-Engel, Goethe University, Institute of Psychology, Mertonstrasse 17, 60054 Frankfurt am Main, Germany. E-mail: schermelleh-engel@psych.uni-frankfurt.de.

tion, the fit of the model should be evaluated. Model fit determines the degree to which the structural equation model fits the sample data. Although there are no well-established guidelines for what minimal conditions constitute an adequate fit, a general approach is to establish that the model is identified, that the iterative estimation procedure converges, that all parameter estimates are within the range of permissible values, and that the standard errors of the parameter estimates have reasonable size (Marsh & Grayson, 1995). Furthermore, the standardized residuals should be checked for patterns in the residual matrix as a sign of ill fit. As Hayduk (1996) states, it is the difference between the empirical and the model-implied covariance matrix "that drives all tests of overall fit, and systematic differences here, even if small, warrant caution" (p. 198).

Applied researchers often have difficulty determining the adequacy of structural equation models because various measures of model fit point to conflicting conclusions about the extent to which the model actually matches the observed data. Software programs such as LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), Mplus (Muthén & Muthén, 1998), AMOS (Arbuckle & Wothke, 1999), SEPATH (Steiger, 1995), or RAMONA (Browne & Mels, 1992), among others, provide a variety of fit indices for model evaluation. As there does not exist a consensus about what constitutes a "good fit" (Tanaka, 1993), the fit indices should be considered simultaneously.

In the following, we will first give a short overview over two methods frequently used for parameter estimation in structural equation modeling, i.e., maximum likelihood (ML) and weighted least squares (WLS)[2]. Second, we will discuss some common goodness-of-fit measures provided by the LISREL program, as most of these indices are provided by other software programs, too. Third, we will give some recommendations for evaluating the fit of structural equation models. Finally, using an artificial data set we will evaluate the fit of four alternative models as examples of poor model fit, adequate fit, and good fit.[3]

[2] Besides these, several other methods exist, but they will not be discussed here in detail. For further information on estimation methods we recommend the respective book chapters in Bollen (1989), Bollen and Long (1993), and Kaplan (2000).

[3] To facilitate orientation, a detailed table of contents is provided as a separate file next to this article at http://www.mpr-online.de

Methods for Parameter Estimation

Maximum Likelihood (ML)

Maximum Likelihood (ML) is the most widely used fitting function for structural equation models. Nearly all of the major software programs use ML as the default estimator. This method leads to estimates for the parameters θ which maximize the likelihood L that the empirical covariance matrix $\mathbf{S}$ is drawn from a population for which the model-implied covariance matrix $\mathbf{\Sigma}(\theta)$ is valid. The log-likelihood function log L to be maximized is (Bollen, 1989, p. 135)

$$\log L = -\frac{1}{2}(N-1)\{\log|\mathbf{\Sigma}(\theta)| + \mathrm{tr}[\mathbf{S}\mathbf{\Sigma}(\theta)^{-1}]\} + c \qquad (1)$$

where

log is the natural logarithm,

L is the likelihood function,

N is the sample size,

θ is the parameter vector,

$\mathbf{\Sigma}(\theta)$ is the model-implied covariance matrix and $|\mathbf{\Sigma}(\theta)|$ its determinant,

tr is the trace of a matrix, and

c is a constant that contains terms of the Wishart distribution that do not change once the sample is given (Bollen, 1989, p. 135).

Maximizing log L is equivalent to minimizing the function

$$F_{\mathrm{ML}} = \log|\mathbf{\Sigma}(\theta)| - \log|\mathbf{S}| + \mathrm{tr}[\mathbf{S}\mathbf{\Sigma}(\theta)^{-1}] - p \qquad (2)$$

where

F_{ML} is the value of the fitting function evaluated at the final estimates (cf. Hayduk, 1989, p. 137), and

p is the number of observed variables.

The ML estimator assumes that the variables in the model are multivariate normal, i.e., the joint distribution of the variables is a multivariate normal distribution. Furthermore it is assumed that $\mathbf{\Sigma}(\theta)$ and $\mathbf{S}$ are positive definite, which implies that these matrices must be nonsingular. ML estimators have several important properties (cf. Bollen, 1989). If the observed data stem from a multivariate normal distribution, if the model is specified correctly, and if the sample size is sufficiently large, ML provides parameter estimates and standard errors that are asymptotically unbiased, consistent, and efficient. Furthermore, with increasing sample size the distribution of the estimator approximates a normal distribution. Thus, the ratio of each estimated parameter to its standard error is approximately z-distributed in large samples.

An important advantage of ML is that it allows for a formal statistical test of overall model fit for overidentified models. The asymptotic distribution of $(N-1)F_{\mathrm{ML}}$ is a χ^2 distribution with $df = s - t$ degrees of freedom, where s is the number of nonredundant elements in $\mathbf{S}$ and t is the number of free parameters. Another advantage of ML is that its estimates are in general scale invariant and scale free (Bollen, 1989, p. 109). As a consequence, the values of the fit function do not depend on whether correlation or covariance matrices are analyzed, and whether original or transformed data are used.

A limitation of ML estimation is the strong assumption of multivariate normality, as violations of distributional assumptions are common and often unavoidable in practice and can potentially lead to seriously misleading results. Nevertheless, ML seems to be quite robust against the violation of the normality assumption (cf. Boomsma & Hoogland, 2001; Chou & Bentler, 1995; Curran, West & Finch, 1996; Muthén & Muthén, 2002; West, Finch, & Curran, 1995). Simulation studies suggest that under conditions of severe nonnormality, ML parameter estimates are still consistent but not necessarily efficient. Using the χ^2 as a measure of model fit (see below) will lead to an inflated Type I error rate for model rejection (Curran, West, & Finch, 1996; West, Finch, & Curran, 1995). For ML estimation with small samples, bootstrapping (Efron & Tibshirani, 1993) may be an alternative (Shipley, 2000).

Corrections have been developed to adjust ML estimators to account for nonnormality (for an overview of robust, corrective statistics, see Satorra & Bentler, 1994). The Satorra-Bentler scaled χ^2 is computed on the basis of the model, estimation method, and sample fourth-order moments and holds regardless of the distribution of the observed variables (Hu & Bentler, 1995, p. 79). As simulation studies demonstrate, robust maximum likelihood estimators based on the Satorra-Bentler scaled χ^2 statistic have relatively good statistical properties compared to least squares estimators (Boomsma & Hoogland, 2001; Hoogland, 1999).

In robustness studies, the scaled χ^2 statistic outperformed the standard ML estimator (Curran, West, & Finch, 1996; Chou, Bentler, & Satorra, 1991), and robust standard errors yielded the least biased standard errors, especially when the distributions of the observed variables were extremely nonnormal (Chou & Bentler, 1995). But as robust maximum likelihood estimation needs relatively large sample sizes of at least $N \geq 400$ (Boomsma & Hoogland, 2001) or even $N \geq 2{,}000$ (Yang-Wallentin & Jöreskog, 2001), further studies are necessary to determine the advantages and disadvantages of these methods in small to moderate sample sizes.

Weighted Least Squares (WLS)

If the data are continuous but nonnormal, the estimation method most often recommended is the asymptotically distribution free (ADF) method (Browne, 1984), although simulation studies suggest that ML estimation with or without a correction for nonnormality seems to perform better than ADF and should be preferred (Boomsma & Hoogland, 2001; Hu, Bentler, & Kano, 1992; Olsson, Foss, Troye, & Howell, 2000). The ADF method is available in LISREL under the name "weighted least squares (WLS)" and in EQS under "arbitrary distribution generalized least squares (AGLS)". In contrast to ML, raw data are needed for data analysis. This method may also be used if some of the observed variables are ordinal and others continuous, if the distributions of the continuous variables deviate considerably from normality, or if models include dichotomous variables.

WLS minimizes the fit function

$$F_{\mathrm{WLS}} = [\mathbf{s} - \boldsymbol{\sigma}(\theta)]' \mathbf{W}^{-1} [\mathbf{s} - \boldsymbol{\sigma}(\theta)] \tag{3}$$

where

s is the vector of nonredundant elements in the empirical covariance matrix,

$\boldsymbol{\sigma}(\theta)$ is the vector of nonredundant elements in the model-implied covariance matrix,

θ is the (t × 1) vector of parameters,

$\mathbf{W}^{-1}$ is a $(k \times k)$ positive definite weight matrix with $k = p(p + 1)/2$ and $p =$ number of observed variables.

WLS requires that the matrix **W** is a consistent estimate of the asymptotic covariance matrix of the sample variances and covariances (or correlations) being analyzed (Browne, 1984; see also Kaplan, 2000, p. 81f.). The elements of $\mathbf{W}^{-1}$ are usually obtained by inverting a matrix **W** which is chosen in such a way that a typical element w_{ijgh} is proportional to a consistent estimate of the asymptotic covariance of s_{ij} with s_{gh}:

$$w_{ijgh} = (N-1)\, acov(s_{ij}, s_{gh}) = \sigma_{ig}\sigma_{jh} + \sigma_{ih}\sigma_{jg} + \frac{N-1}{N}\kappa_{ijgh} \tag{4}$$

where

κ_{ijgh} is the fourth-order cumulant, a component of the distribution related to the multivariate kurtosis,

s_{ij} , s_{gh} are sample covariances, and

σ_{ig}, σ_{jh}, σ_{ih}, and σ_{jg} are population covariances of X_i with X_g, X_j with X_h, X_i with X_h, and X_j with X_g, respectively.

The WLS method has several advantages, yet also some disadvantages (Bollen, 1989, p. 432). One main advantage is that it requires only minimal assumptions about the distribution of the observed variables. Simulation research with nonnormal data shows that the WLS test statistic is relatively unaffected by distributional characteristics (Hoogland & Boomsma, 1998; West, Finch, & Curran, 1995). Another advantage is that WLS may also be used as a means of analyzing correlation matrices, if the corresponding matrix **W** contains the covariances of the correlations r_{ij} and r_{gh}. It should be noted that this matrix differs from the one containing the covariances s_{ij} and s_{gh}. In general, WLS produces an accurate χ^2 test statistic and accurate standard errors if sample size is sufficiently large.

A limitation of the WLS method can be seen in the fact that the weight matrix grows rapidly with increasing numbers of indicator variables. As the asymptotic covariance matrix is of order $(k \times k)$, where $k = p(p + 1)/2$ and p is the number of observed variables, the weight matrix of a model containing 10 variables would be of order (55×55) with 1540 nonredundant elements. Thus, the WLS method compared to ML requires large samples in order to obtain consistent and efficient estimates. If the distribution of the observed variables does not deviate from the normal distribution by a considerable amount, one may also apply ML. Consistent with previous findings (cf. Chou, Bentler, & Satorra, 1991; Muthén & Kaplan, 1985, 1992), Chou and Bentler (1995) do not recommend WLS for practical applications when models are complex and when the sample size is small.

Recent developments of WLS-based estimators under nonnormality suggest that WLS for categorical outcomes as implemented in Mplus (Muthén & Muthén, 1998) does not require the same large sample sizes as WLS for continuous non-normal data (cf. Kaplan, 2000, p. 85ff.). But further studies are needed to determine the advantages and disadvantages of this method compared to other robust estimators.

Special cases of WLS estimation are the Generalized Least Squares (GLS) and the Unweighted Least Squares (ULS) estimators.

Under the assumption of multivariate normality, the WLS fitting function F_{WLS} can be rewritten as

$$F_{\text{WLS}} = \frac{1}{2}\operatorname{tr}\{[\mathbf{S} - \mathbf{\Sigma}(\theta)]\mathbf{V}^{-1}\}^2 \tag{5}$$

(Bollen, 1989, p. 428), where

tr is the trace of the matrix,

$\mathbf{S}$ is the empirical covariance matrix,

$\mathbf{\Sigma}(\theta)$ is the model-implied covariance matrix,

θ is the $(t \times 1)$ vector of parameters, and

$\mathbf{V}^{-1}$ is a $p \times p$ weight matrix (a weight matrix of lower dimensions).

Equation 5 is the general form of the Generalized Least Squares (GLS) estimator in which the $k \times k$ weight matrix $\mathbf{W}^{-1}$ of Equation 3 is replaced by the $p \times p$ weight matrix $\mathbf{V}^{-1}$.

Inserting $\mathbf{S}$ for $\mathbf{V}$ (the most common choice which is found both in LISREL and in EQS) leads to the Generalized Least Squares (GLS) estimator, a special case of WLS, with the fitting function

$$F_{\text{GLS}} = \frac{1}{2}\operatorname{tr}\{[\mathbf{S} - \mathbf{\Sigma}(\theta)]\mathbf{S}^{-1}\}^2 \ , \tag{6}$$

where

tr is the trace of the matrix,

$\mathbf{S}$ is the empirical covariance matrix and $\mathbf{S}^{-1}$ its inverse,

$\mathbf{\Sigma}(\theta)$ is the model-implied covariance matrix, and

θ is the $(t \times 1)$ vector of parameters.

Generalized Least Squares (GLS) is a frequently used estimation method that is asymptotically equivalent to F_{ML}. As GLS is based on the same assumptions as ML, this estimation method is used under the same conditions. But as it performs less well in small samples, maximum likelihood should generally be preferred with small sample sizes.

Finally, with the identity matrix $\mathbf{I}$ chosen as the weight matrix $\mathbf{V}^{-1}$, GLS reduces to Unweighted Least Squares (ULS) estimator, another special case of WLS. For parameter estimation, the following fit function is minimized:

$$F_{\text{ULS}} = \frac{1}{2}\text{tr}\{[\mathbf{S} - \mathbf{\Sigma}(\theta)]\}^2, \tag{7}$$

where

tr is the trace of the matrix,

$\mathbf{S}$ is the empirical covariance matrix,

$\mathbf{\Sigma}(\theta)$ is the model-implied covariance matrix, and

θ is the $(t \times 1)$ vector of parameters.

The fitting function F_{ULS} minimizes the sum of squares of each element in the residual matrix $(\mathbf{S} - \mathbf{\Sigma}(\theta))$. ULS offers the advantage that it leads to a consistent estimator of θ comparable to ML and WLS, yet in contrast to ML, distributional assumptions are not necessary. Disadvantages are that ULS does not provide the most efficient estimates for θ and that these estimates are neither scale invariant nor scale free (Bollen, 1989). Furthermore, some software programs do not provide χ^2 statistics and standard errors if ULS is applied. Others estimate standard errors and χ^2 for ULS only under the assumption of multivariate normality. Therefore, ULS results should be interpreted with caution.

Evaluation of Model Fit

In structural equation modeling, evaluation of model fit is not as straightforward as it is in statistical approaches based on variables measured without error. Because there is no single statistical significance test that identifies a correct model given the sample data, it is necessary to take multiple criteria into consideration and to evaluate model fit on the basis of various measures simultaneously. For each estimation procedure, a large number of goodness-of-fit indices is provided to judge whether the model is consistent with the empirical data. The choice of the estimation procedure depends on the type of data included in the model.

Generally, the fit criteria of a structural equation model indicate to what extent the specified model fits the empirical data. Only one goodness-of-fit measure, i.e., the χ^2 test statistic, has an associated significance test, while all other measures are descriptive. Thus, following successful parameter estimation, model evaluation can be assessed inferentially by the χ^2 test or descriptively by applying other criteria.

For inferential statistical evaluation, only the χ^2 test is available, whereas for descriptive evaluation, three main classes of criteria exist, i.e., measures of overall model fit, measures based on model comparisons, and measures of model parsimony (cf. Schumacker & Lomax, 1996, p. 119 ff.). Most of the descriptive fit criteria are based on the χ^2 statistic given by the product of the sample size $(N - 1)$ and the optimized fitting function.

In the following, we will present a selection of fit indices that are provided by the LISREL program (Jöreskog & Sörbom, 1993). Most of these indices are reported by other software programs as well so that this presentation should be also beneficial for users of EQS, Mplus, and other software programs for SEM.

Part I: Tests of Significance

χ^2 Test Statistic

The χ^2 test statistic is used for hypothesis testing to evaluate the appropriateness of a structural equation model. If the distributional assumptions are fulfilled, the χ^2 test evaluates whether the population covariance matrix $\mathbf{\Sigma}$ is equal to the model-implied covariance matrix $\mathbf{\Sigma}(\theta)$, i.e., it tests the null hypothesis that the differences between the elements of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}(\theta)$ are all zero: $\mathbf{\Sigma} - \mathbf{\Sigma}(\theta) = \mathbf{0}$. Being population parameters, these matrices are unknown, so researchers examine their sample counterparts, the empirical covariance matrix **S** and the model-implied covariance matrix $\mathbf{\Sigma}(\hat{\theta})$, where $\hat{\theta}$ is the $(t \times 1)$ vector of estimated parameters. If the null hypothesis is correct, the minimum fit function value times $N - 1$ converges to a χ^2 variate

$$\chi^2(df) = (N-1)\,F[\mathbf{S}, \mathbf{\Sigma}(\hat{\theta})] \tag{8}$$

with $df = s - t$ degrees of freedom, where

s is the number of nonredundant elements in **S**,

t is the total number of parameters to be estimated,

N is the sample size,

S is the empirical covariance matrix, and

$\mathbf{\Sigma}(\hat{\theta})$ is the model-implied covariance matrix.

LISREL provides different χ^2 test statistics for ML, WLS, GLS, or ULS, so that the obtained χ^2 value depends on the estimation method. In general, high χ^2 values in relation to the number of degrees of freedom indicate that the population covariance matrix $\mathbf{\Sigma}$ and the model-implied covariance matrix $\mathbf{\Sigma}(\theta)$ differ significantly from each other. As the residuals, namely, the elements of $\mathbf{S} - \mathbf{\Sigma}(\hat{\theta})$, should be close to zero for a good model fit, the researcher is interested in obtaining a nonsignificant χ^2 value with associated degrees of freedom. If the *p*-value associated with the χ^2 value is larger than .05, the null hypothesis is accepted and the model is regarded as compatible with the population covariance matrix $\mathbf{\Sigma}$. In this case the test states that the model fits the data. But still an uncertainty exists that other models may fit the data equally well (cf. Lee & Hershberger, 1990; Stelzl, 1986).

There are several shortcomings associated with the χ^2 test statistic.

- Violation of assumptions. The χ^2 test is based on the assumptions that the observed variables are multivariate normal and that the sample size is sufficiently large. However, these assumptions are not met in many practical applications.
- Model complexity. One disadvantage of the χ^2 value is that it decreases when parameters are added to the model. Thus, the χ^2 value of a more complex, highly parameterized model tends to be smaller than for simpler models because of the reduc-

tion in degrees of freedom. For example, the saturated model with as many free parameters as there are variances and covariances in **S** yields a χ^2 of zero, whereas the independence model, a very restrictive model, usually has a very large χ^2 value (cf. Mueller, 1996, p. 89). Thus, a good model fit may result either from a correctly specified model or from a highly overparameterized model.

- Dependence on sample size. With increasing sample size and a constant number of degrees of freedom, the χ^2 value increases. This leads to the problem that plausible models might be rejected based on a significant χ^2 statistic even though the discrepancy between the sample and the model-implied covariance matrix is actually irrelevant. On the other hand, as sample size decreases, the χ^2 value decreases as well and the model test may indicate nonsignificant probability levels even though the discrepancy between the sample and the model-implied covariance matrix is considerable. Therefore not too much emphasis should be placed on the significance of the χ^2 statistic. Jöreskog and Sörbom (1993) even suggest to use χ^2 not as a formal test statistic but rather as a descriptive goodness-of-fit index. They propose to compare the magnitude of χ^2 with the expected value of the sample distribution, i.e., the number of degrees of freedom, as $E(\chi^2) = df$. For a good model fit, the ratio χ^2/df should be as small as possible. As there exist no absolute standards, a ratio between 2 and 3 is indicative of a "good" or "acceptable" data-model fit, respectively. However, the problem of sample size dependency cannot be eliminated by this procedure (Bollen, 1989, p. 278).

χ^2 Difference Test

In applications of covariance structure analysis, researchers often face the problem of choosing among two or more alternative models. The choice of which measure to use for selecting one of several competing models depends on whether or not the models are nested.

A specific model (Model A) is said to be nested within a less restricted model (Model B) with more parameters and less degrees of freedom than Model A, if Model A can be derived from Model B by fixing at least one free parameter in Model B or by introducing other restrictions, e.g., by constraining a free parameter to equal one or more other parameters. For example, in multi-sample comparisons of factorial invariance, any model with some parameters constrained to be invariant over the multiple groups is nested under the corresponding model in which the respective parameters are unconstrained, and the model in which *all* parameters are invariant is nested under both these models. Any two models are nested when the free parameters in the more restrictive model are a subset of the free parameters in the less restrictive model.

As the test statistic of each of the nested models follows a χ^2 distribution, the difference in χ^2 values between two nested models is also χ^2 distributed (Steiger, Shapiro, & Browne, 1985), and the number of degrees of freedom for the difference is equal to the difference in degrees of freedom for the two models. Under appropriate assumptions, the difference in model fit can be tested using the χ^2 difference test

$$\chi^2_{\text{diff}}(df_{\text{diff}}) = \chi^2_{\text{A}}(df_{\text{A}}) - \chi^2_{\text{B}}(df_{\text{B}}) \tag{9}$$

(Bentler, 1990; Bollen, 1989; Jöreskog, 1993), where

χ^2_{A} denotes the χ^2 value of Model A, a model that is a restricted version of Model B, i.e., Model A has less free parameters and more degrees of freedom (df_{A}) and is thus nested within Model B,

χ^2_{B} denotes the χ^2 value of Model B, a model that is less restricted and therefore has more free parameters and less degrees of freedom (df_{B}) than Model A, and

$df_{\text{diff}} = df_{\text{A}} - df_{\text{B}}$.

If the χ^2 difference is significant, the null hypothesis of equal fit for both models is rejected and Model B should be retained. But if the χ^2 difference is nonsignificant, which means that the fit of the restricted model (Model A) is not significantly worse than the fit of the unrestricted model (Model B), the null hypothesis of equal fit for both models cannot be rejected and the restricted model (Model A) should be favored.

The χ^2 difference test applied to nested models has essentially the same strengths and weaknesses as the χ^2 test applied to any single model, namely, the test is directly affected by sample size, and for large samples trivial differences may become significant. For the χ^2 difference test to be valid, at least the least restrictive model of a sequence of models (in our example Model B) should fit the data.

It should be noted that the Satorra-Bentler scaled χ^2 values resulting from robust estimation methods cannot be used for χ^2 difference testing because the difference between two scaled χ^2 values for nested models is not distributed as a χ^2 (Satorra, 2000). Recently, Satorra and Bentler (2001) developed a scaled difference χ^2 test statistic for moment structure analysis. They could show that simple hand calculations based on output from nested runs can give the desired χ^2 difference test of nested models using the scaled χ^2. These calculations may be obtained from Mplus by asking for the MLM estimator.

If models are not nested, they may be compared on the basis of descriptive goodness-of-fit measures that take parsimony as well as fit into account, e.g., the Akaike Information Criterion (Akaike, 1974, 1987), which can be used regardless of whether models for the same data can be ordered in a nested sequence or not (see below). A more detailed discussion of alternative methods for comparing competing models is given in Kumar and Sharma (1999), Raykov and Penev (1998), and Rigdon (1999).

Part II: Descriptive Goodness-of-Fit Measures

Because of the drawbacks of the χ^2 goodness-of-fit tests, numerous descriptive fit indices have been developed that are often assessed intuitively. These indices are derived from ML, WLS, GLS, or ULS, but in the following we will not differentiate between these methods (for further information on ML-, WLS-, and GLS-based descriptive fit indices, cf. Hu & Bentler, 1998). Many of these measures are intended to range between zero (no fit) and one (perfect fit), but as Hu and Bentler (1995) note, the sampling distributions of goodness-of-fit indices are unknown with the exception of χ^2 so that critical values for fit indices are not defined. As a reasonable minimum for model acceptance, Bentler and Bonett (1980) proposed a value of .90 for normed indices that are not parsimony adjusted (cf. Hoyle & Panter, 1995), while .95 should be indicative of a good fit relative to the baseline model (Kaplan, 2000). But recently, Hu and Bentler (1995, 1998, 1999) gave evidence that .90 might not be a reasonable cutoff for all fit indices under all circumstances: "The rule of thumb to consider models acceptable if a fit index exceeds .90 is clearly an inadequate rule" (Hu & Bentler, 1995, p. 95). They suggested to raise the rule of thumb minimum standard for the *CFI* and the *NNFI* (see below) from .90 to .95 to reduce the number of severely misspecified models that are considered acceptable based on the .90 criterion (Hu & Bentler, 1998, 1999).

Descriptive Measures of Overall Model Fit

Due to the sensitivity of the χ^2 statistic to sample size, alternative goodness-of-fit measures have been developed. Measures of overall model fit indicate to which extent a structural equation model corresponds to the empirical data. These criteria are based on the difference between the sample covariance matrix **S** and the model-implied covariance matrix $\mathbf{\Sigma}(\hat{\theta})$. The following indices are descriptive measures of overall model fit: Root Mean Square Error of Approximation (*RMSEA*), Root Mean Square Residual (*RMR*), and Standardized Root Mean Square Residual (*SRMR*).

Root Mean Square Error of Approximation (RMSEA)

The usual test of the null hypothesis of exact fit is invariably false in practical situations and will almost certainly be rejected if sample size is sufficiently large. Therefore a more sensible approach seems to be to assess whether the model fits approximately well in the population (cf. Kaplan, 2000, p. 111). The null hypothesis of exact fit is replaced by the null hypothesis of "close fit" (Browne & Cudeck, 1993, p. 146). Thus, the Root Mean Square Error of Approximation (*RMSEA*; Steiger, 1990) is a measure of approximate fit in the population and is therefore concerned with the discrepancy due to approximation.

RMSEA is estimated by $\hat{\varepsilon}_a$, the square root of the estimated discrepancy due to approximation per degree of freedom:

$$\hat{\varepsilon}_a = \sqrt{\max\left\{\left(\frac{F(\mathbf{S},\mathbf{\Sigma}(\hat{\theta}))}{df} - \frac{1}{N-1}\right), 0\right\}} \qquad (10)$$

where

$F(\mathbf{S},\mathbf{\Sigma}(\hat{\theta}))$ is the minimum of the fit function,

$df = s - t$ is the number of degrees of freedom, and

N is the sample size.

The *RMSEA* is bounded below by zero. Steiger (1990) as well as Browne and Cudeck (1993) define a "close fit" as a *RMSEA* value less than or equal to .05. According to Browne and Cudeck (1993), *RMSEA* values ≤ .05 can be considered as a good fit, values between .05 and .08 as an adequate fit, and values between .08 and .10 as a mediocre fit, whereas values > .10 are not acceptable. Although there is general agreement that the value of *RMSEA* for a good model should be less than .05, Hu and Bentler (1999) suggested an *RMSEA* of less than .06 as a cutoff criterion. In addition, a 90% confidence interval (CI) around the point estimate enables an assessment of the precision of the *RMSEA* estimate. On the basis of the CI, it is possible to say with a certain level of confidence that the given interval contains the true value of the fit index for that model in the population (MacCallum, Browne, & Sugawara, 1996). The lower boundary (left side) of the confidence interval should contain zero for exact fit and be < .05 for close fit. Note that when the model fits well in the population, the lower end of the confidence interval is truncated at zero, which leads to an asymmetry of the confidence interval. *RMSEA* is regarded as relatively independent of sample size, and additionally favors parsimonious models (Browne & Cudeck, 1993; Kaplan, 2000).

For an understanding of *RMSEA* it is important to distinguish between two different kinds of error. The *error of approximation*, which is of primary interest here, represents the lack of fit of the model to the population covariance matrix $\mathbf{\Sigma}$. The minimum fit function value one would obtain if the model could be fitted to the population covariance matrix is a possible measure of this error. In contrast, the *error of estimation* reflects the differences between the model fitted to the population covariance matrix $\mathbf{\Sigma}$ (if this could be done) and the model fitted to the sample covariance matrix **S** (Browne & Cudeck, 1993, p. 141). From the viewpoint of model fit in the population, the error of estimation is of secondary interest only. Because the fit function value $F(\mathbf{S},\mathbf{\Sigma}(\hat{\theta}))$, which refers to the sample covariance matrix, would be a biased estimator of the population error of approximation, *RMSEA* includes a term inversely proportional to $N - 1$ which serves as a correction for bias (Browne & Cudeck, 1993, p. 143).

Root Mean Square Residual (RMR) and Standardized RMR (SRMR)

It was already mentioned that the residuals are given by the elements of the matrix $\mathbf{S}-\mathbf{\Sigma}(\hat{\theta})$. These are sometimes called "fitted residuals" because they express the remaining discrepancies between the covariance matrices **S** and $\mathbf{\Sigma}(\hat{\theta})$ once the parameters of the model are estimated.

The Root Mean Square Residual index (*RMR*) of Jöreskog and Sörbom (1981, p. 41; 1989) is an overall badness-of-fit measure that is based on the fitted residuals. Concretely, *RMR* is defined as the square root of the mean of the squared fitted residuals,

$$RMR=\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}(s_{ij}-\hat{\sigma}_{ij})^2}{p(p+1)/2}} \tag{11}$$

where

s_{ij} is an element of the empirical covariance matrix **S**,

$\hat{\sigma}_{ij}$ is an element of the model-implied matrix covariance $\mathbf{\Sigma}(\hat{\theta})$, and

p is the number of observed variables.

In principle, *RMR* values close to zero suggest a good fit. But as the elements of **S** and $\mathbf{\Sigma}(\hat{\theta})$ are scale dependent, the fitted residuals are scale dependent, too, which implies that *RMR* depends on the sizes of the variances and covariances of the observed variables. In other words, without taking the scales of the variables into account it is virtually impossible to say whether a given *RMR* value indicates good or bad fit.

To overcome this problem, the Standardized Root Mean Square Residual (*SRMR*) has been introduced (Bentler, 1995, p. 271). Here, the residuals $s_{ij} - \hat{\sigma}_{ij}$ are first *divided by the standard deviations* $s_i = \sqrt{s_{ii}}$ and $s_j = \sqrt{s_{jj}}$ *of the respective manifest variables*, which leads to a standardized residual matrix with elements $(s_{ij} - \hat{\sigma}_{ij})/(s_i s_j) = r_{ij} - \hat{\sigma}_{ij}/(s_i s_j)$ where r_{ij} is the observed correlation between the respective variables (Bentler, 1995, p. 90). In contrast to LISREL (see below), EQS provides this matrix explicitly. Calculating the root mean square of the such defined standardized residuals in analogy to Equation 11 leads to the *SRMR* which is available in both LISREL and EQS. Again, a value of zero indicates perfect fit, but it is still difficult to designate cut-off values for good and for acceptable fit because of sample size dependency and sensitivity to misspecified models (Hu & Bentler, 1998). A rule of thumb is that the *SRMR* should be less than .05 for a good fit (Hu & Bentler, 1995), whereas values smaller than .10 may be interpreted as acceptable.

The standardized residuals given above are similar – but in general not identical – to the correlation residuals suggested by Bollen (1989, p. 258). Both share the main idea that in an overall fit measure based on residuals, the observed variables should enter in standardized form such that all matrix elements contributing to the fit measure are on comparable scales. Thus the *SRMR*, same as the *RMR*, remains a purely descriptive fit index.

Unfortunately the term "standardized residuals" also appears in a second meaning that is independent of the *SRMR*. In this meaning, the residuals themselves get standardized: Besides the fitted residuals, the LISREL program provides a matrix of standardized residuals which are obtained by *dividing each fitted residual by its large-sample standard error* (Jöreskog & Sörbom, 1989, p. 28). Being independent of the units of measurements of the variables as the standardized residuals discussed before, they also allow an easier interpretation than the fitted residuals. In the present case, however, the standardized residuals can be interpreted approximately in an inferential sense, namely, in a way similar to z scores. Provided that the *SRMR* or other fit indices signalize bad fit, single standardized residuals whose absolute values are greater than 1.96 or 2.58 can be useful for detecting the source of misfit. The largest absolute value indicates the element that is most poorly fitted by the model. Because the kind of standardized residuals considered here refers to a standard error, the absolute values tend to increase with increasing sample size if the magnitudes of the fitted residuals remain essentially constant.

As the *RMR* and the *SRMR* are overall measures based on squared residuals, they can give no information about the directions of discrepancies between $\mathbf{S}$ and $\mathbf{\Sigma}(\hat{\theta})$. In a residual analysis, regardless of whether unstandardized or standardized residuals are used and which kind of standardization is preferred, it is important to take the sign of a residual into account when looking for the cause of model misfit. Given that an empirical covariance is positive, a positive residual indicates that the model underestimates the sample covariance. In this case, the empirical covariance is larger than the model-implied covariance. A negative residual indicates that the model overestimates the sample covariance, that is, the empirical covariance is smaller than the model-implied covariance.

Descriptive Measures Based on Model Comparisons

The basic idea of comparison indices is that the fit of a model of interest is compared to the fit of some baseline model. Even though any model nested hierarchically under the target model (the model of interest) may serve as a comparison model, the independence model is used most often. The independence model assumes that the observed variables are measured without error, i.e., all error variances are fixed to zero and all factor loadings are fixed to one, and that all variables are uncorrelated. This baseline model is a very restrictive model in which only p parameters, namely the variances of the variables, have to be estimated. An even more restrictive baseline model than the independence model is the null model, a model in which all parameters are fixed to zero (Jöreskog & Sörbom, 1993, p. 122) and hence, no parameters have to be estimated. The fit index for a baseline model will usually indicate a bad model fit and serves as a comparison value. The issue is whether the target model is an improvement relative to the baseline model.

Often used measures based on model comparisons are the Normed Fit Index (*NFI*), the Nonnormed Fit Index (*NNFI*), the Comparative Fit Index (*CFI*), the Goodness-of-Fit Index (*GFI*), and the Adjusted Goodness-of-Fit Index (*AGFI*), which will be explained below in more detail.

Normed Fit Index (NFI) and Nonnormed Fit Index (NNFI)

The Normed Fit Index (*NFI*) proposed by Bentler and Bonnett (1980) is defined as

$$NFI = \frac{\chi_i^2 - \chi_t^2}{\chi_i^2} = 1 - \frac{\chi_t^2}{\chi_i^2} = 1 - \frac{F_t}{F_i}, \tag{12}$$

where

χ_i^2 is the chi-square of the independence model (baseline model),

χ_t^2 is the chi-square of the target model, and

F is the corresponding minimum fit function value.

NFI values range from 0 to 1, with higher values indicating better fit. When $F_t = F_i$, *NFI* equals zero; when $F_t = 0$, *NFI* equals one, which suggests that the target model is the best possible improvement over the independence model. Although the theoretical boundary of *NFI* is one, *NFI* may not reach this upper limit even if the specified model is correct, especially in small samples (Bentler, 1990, p. 239). This can occur because the expected value of χ_t^2 is greater than zero: $E(\chi_t^2) = df$. The usual rule of thumb for this index is that .95 is indicative of good fit relative to the baseline model (Kaplan, 2000, p. 107), whereas values greater than .90 are typically interpreted as indicating an acceptable fit (Marsh & Grayson, 1995; Schumacker & Lomax, 1996).

A disadvantage of the *NFI* is that it is affected by sample size (Bearden, Sharma, & Teel, 1982). In order to take care of this problem, Bentler and Bonnett (1980) extended the work by Tucker and Lewis (1973) and developed the Nonnormed Fit Index (*NNFI*), also known as the Tucker-Lewis Index (*TLI*). The *NNFI* measures relative fit and is defined as

$$NNFI = \frac{(\chi_i^2/df_i) - (\chi_t^2/df_t)}{(\chi_i^2/df_i) - 1} = \frac{(F_i/df_i) - (F_t/df_t)}{(F_i/df_i) - 1/(N-1)}, \tag{13}$$

where

χ_i^2 is the chi-square of the independence model (baseline model),

χ_t^2 is the chi-square of the target model,

F is the corresponding minimum fit function value, and

df is the number of degrees of freedom.

The *NNFI* ranges in general from zero to one, but as this index is not normed, values can sometimes leave this range, with higher *NNFI* values indicating better fit. A rule of thumb for this index is that .97 is indicative of good fit relative to the independence model, whereas values greater than .95 may be interpreted as an acceptable fit. As the independence model almost always has a large χ^2, *NNFI* values are often very close to one (Jöreskog & Sörbom, 1993, p. 125), so that a value of .97 seems to be more reasonable as an indication of a good model fit than the often stated cutoff value of .95.

NNFI takes the degrees of freedom of the specified model as well as the degrees of freedom of the independence model into consideration. More complex, i.e., less restrictive models are penalized by a downward adjustment, while more parsimonious, i.e., more restrictive models are rewarded by an increase in the fit index. An advantage of the *NNFI* is that it is one of the fit indices less affected by sample size (Bentler, 1990; Bollen, 1990; Hu & Bentler, 1995, 1998).

Comparative Fit Index (CFI)

The Comparative Fit Index (*CFI*; Bentler, 1990), an adjusted version of the Relative Noncentrality Index (*RNI*) developed by McDonald and Marsh (1990), avoids the underestimation of fit often noted in small samples for Bentler and Bonett's (1980) normed fit index (*NFI*). The *CFI* is defined as

$$CFI = 1 - \frac{\max[(\chi_t^2 - df_t), 0]}{\max[(\chi_t^2 - df_t), (\chi_i^2 - df_i), 0]} \tag{14}$$

where

max denotes the maximum of the values given in brackets,

χ_i^2 is the chi-square of the independence model (baseline model),

χ_t^2 is the chi-square of the target model, and

df is the number of degrees of freedom.

The *CFI* ranges from zero to one with higher values indicating better fit. A rule of thumb for this index is that .97 is indicative of good fit relative to the independence model, while values greater than .95 may be interpreted as an acceptable fit. Again a value of .97 seems to be more reasonable as an indication of a good model fit than the often stated cutoff value of .95. Comparable to the *NNFI*, the *CFI* is one of the fit indices less affected by sample size (Bentler, 1990; Bollen, 1990; Hu & Bentler, 1995, 1998, 1999).

Goodness-of-Fit-Index (GFI) and Adjusted Goodness-of-Fit-Index (AGFI)

The Goodness-of-Fit-Index (*GFI*; Jöreskog & Sörbom, 1989; Tanaka & Huba, 1984) measures the relative amount of the variances and covariances in the empirical covariance matrix **S** that is predicted by the model-implied covariance matrix $\mathbf{\Sigma}(\hat{\theta})$. According to Jöreskog and Sörbom (1993, p. 123), this implies testing how much better the model fits as compared to "no model at all" (null model), i.e., when all parameters are fixed to zero. The *GFI* seems to be inspired by analogy with the concept of a coefficient of determination (Mulaik et al., 1989, p. 435) and is defined as

$$GFI = 1 - \frac{F_t}{F_n} = 1 - \frac{\chi_t^2}{\chi_n^2}\,, \qquad (15)$$

where

χ_n^2 is the chi-square of the null model (baseline model),

χ_t^2 is the chi-square of the target model, and

F is the corresponding minimum fit function value.

The *GFI* typically ranges between zero and one with higher values indicating better fit, but in some cases a negative *GFI* may occur. The usual rule of thumb for this index is that .95 is indicative of good fit relative to the baseline model, while values greater than .90 are usually interpreted as indicating an acceptable fit (Marsh & Grayson, 1995; Schumacker & Lomax, 1996).

Joreskog and Sorbom (1989) also developed the Adjusted Goodness-of-Fit Index *AGFI* to adjust for a bias resulting from model complexity. The *AGFI* adjusts for the model's degrees of freedom relative to the number of observed variables and therefore rewards less complex models with fewer parameters. The *AGFI* is given by

$$AGFI = 1 - \frac{df_n}{df_t}(1 - GFI) = 1 - \frac{\chi_t^2 / df_t}{\chi_n^2 / df_n} \tag{16}$$

where

χ_n^2 is the chi-square of the null model (baseline model),

χ_t^2 is the chi-square of the target model,

$df_n = s = p(p+1)/2$ is the number of degrees of freedom for the null model, and

$df_t = s - t$ is the number of degrees of freedom for the target model.

AGFI values typically range between zero and one with larger values indicating a better fit, but it is also possible that a large N in combination with small df_t can result in a negative *AGFI*. If the number of degrees of freedom for the target model approaches the number of degrees of freedom for the null model, the *AGFI* approaches the *GFI*. A rule of thumb for this index is that .90 is indicative of good fit relative to the baseline model, while values greater than .85 may be considered as an acceptable fit.

Simulation studies suggest that *GFI* and *AGFI* are not independent of sample size (Hu & Bentler, 1995, 1998, 1999). Furthermore, both indices decrease with increasing model complexity, especially for smaller sample sizes (Anderson & Gerbing, 1984).

Descriptive Measures of Model Parsimony

Parsimony is considered to be important in assessing model fit (Hu & Bentler, 1995; Mulaik et al., 1989) and serves as a criterion for choosing between alternative models.

Several fit indices, among others the Parsimony Goodness-of-Fit Index (*PGFI*), the Parsimony Normed Fit Index (*PNFI*), the Akaike Information Criterion (*AIC*), the Consistent *AIC* (*CAIC*), and the Expected Cross-Validation Index (*ECVI*) adjust for model parsimony when assessing the fit of structural equation models.

Parsimony Goodness-of-Fit Index (PGFI) and Parsimony Normed Fit Index (PNFI)

The Parsimony Goodness-of-Fit Index (*PGFI*; Mulaik et al., 1989) and the Parsimony Normed Fit Index (*PNFI*; James, Mulaik, & Brett, 1982) are modifications of *GFI* and *NFI*:

$$PGFI = \frac{df_t}{df_n} GFI \tag{17}$$

where

df_t is the number of degrees of freedom of the target model,

df_n is the number of degrees of freedom of the null model, and

GFI is the Goodness-of-Fit Index,

and

$$PNFI = \frac{df_t}{df_i} NFI \tag{18}$$

where

df_t is the number of degrees of freedom of the target model,

df_i is the number of degrees of freedom of the independence model, and

NFI is the Normed Fit Index.

PGFI and *PNFI* both range between zero and one, with higher values indicating a more parsimonious fit. Both indices may be used for choosing between alternative models.

The effect of multiplying *GFI* and *NFI* by the respective parsimony ratio of the degrees of freedom is to reduce the original indices to a value closer to zero. Comparing *AGFI* and *PGFI* , both indices adjust the *GFI* downward for more complex and therefore less parsimonious models, but the *PGFI* exerts a stronger penalty on complex models with less degrees of freedom than *AGFI* (Tanaka, 1993). Similarly, the *PNFI* adjusts *NFI* downward.

Akaike Information Criterion (AIC)

The Akaike Information Criterion (*AIC*; Akaike, 1974, 1987) adjusts χ^2 for the number of estimated parameters and can be used to compare competing models that need not be nested. Various versions of the *AIC* exist (Hayduk, 1996, pp. 198-200) which are essentially equivalent as long as the version is not changed during the comparisons. Furthermore, all calculations must be based on the same covariance matrix.

The LISREL program provides

$$AIC = \chi^2 + 2t\,, \tag{19}$$

where t is the number of estimated parameters.

Other software programs, e.g., EQS (Bentler, 1995), adopt *AIC* as

$$AIC = \chi^2 - 2df\,, \tag{20}$$

where df is the number of degrees of freedom.

Originally, *AIC* had been introduced by Akaike as "an information criterion" (Akaike, 1985, p. 12) in the form

$$AIC = -2\log L + 2t\,, \tag{21}$$

where log$L$ is the maximized value of the log likelihood function for the respective model.

All versions share the feature that within a set of models for the same data, the model with the minimum *AIC* value is regarded as the best fitting model. In other words, *AIC* is actually a "badness of fit" index (Kaplan, 2000, p. 116).

The derivation of *AIC* unifies the information-theoretical concept of entropy and the method of maximum likelihood (Akaike, 1985) and is rather demanding. On the other hand, the above formulas show that the concrete calculation of *AIC* is quite simple once χ^2 or the log-likelihood is known. However, because the criterion (in whatever version) is not normed, it is not possible to interpret an isolated *AIC* value. Therefore, *AIC* should first be obtained for all models of interest, including the values for the independence model and the saturated model (a model with zero degrees of freedom) which automatically appear in the LISREL output. Researchers should then consider to select the model with the lowest *AIC* value. To avoid misunderstandings, it should be recognized that *AIC* is only a descriptive measure and not a test of significance although it may be used to decide between competing models.

The main implications of *AIC* can be understood as follows. Provided a set of competing models for given data, the task is to select the model which serves best as an approximation to "reality". In approximating reality, two kinds of errors can occur that have already been addressed in the context of *RMSEA*, namely, systematic discrepancies (the error of approximation) introduced by the population model, and random discrepancies (the error of estimation) introduced by the use of sample data. In contrast to *RMSEA* which implies a focus on the population model while disregarding the estimation of its parameters, *AIC* takes the view that only the *estimated* model can be used for the prediction of further observations.

From this predictive point of view, extremely imprecise parameter estimates would jeopardize the practical applicability of an otherwise appropriate model. Considering the overall error, it can be better to tolerate a model that is slightly oversimplified if this is compensated by a reduction of sampling fluctuations. Therefore, *AIC* reflects the search for a compromise between the approximation and estimation errors that minimizes the overall error. For example, note that the LISREL version of *AIC* increases both with χ^2 (which is connected to the approximation error) and the number of free parameters (which is connected to the estimation error). Because *AIC* penalizes a high number of estimated parameters (or, equivalently, rewards a high number of degrees of freedom), it rewards parsimony.

To examine more closely how *AIC* works, it is instructive to consider the special case that a single target model is to be contrasted with the saturated model. For the sake of simplicity, we use the EQS version of *AIC*. According to Equation 20, *AIC* for the target model is given by $AIC_t = \chi^2 - 2df$, whereas for the saturated model, $\chi^2 = 0$, $df = 0$, and thus $AIC_s = 0$. The target model should be selected if $AIC_t < AIC_s$, which is equivalent to the condition that for the target model $\chi^2/df < 2$. This sheds new light on the recommendation that $\chi^2/df < 2$ is indicative of a good fit (see above, subsection "χ^2 Test Statistic"). Note that the result does not depend on the particular *AIC* version as two models were compared.

Consistent AIC (CAIC)

After the introduction of *AIC*, similar information indices have been proposed for various reasons. We restrict ourselves to a single example: The LISREL program offers the so-called Consistent *AIC* (*CAIC*; Bozdogan, 1987) in the version

$$CAIC = \chi^2 + (1 + \log N)t\ , \tag{22}$$

where

$\log N$ is the natural logarithm of the sample size N, and

t is again the number of estimated parameters.

Given a set of competing models, *CAIC* is used in the same way as *AIC*. In the present context, consistency means that the correct model is selected as the sample size tends to infinity ($N \rightarrow \infty$).

At first sight, the only practically important difference between *AIC* and *CAIC* is that the factor 2 in the penalty term of Equation 19 is replaced by the factor ($1 + \log N$), which implies that the weight of the number of estimates now depends on the sample size and that parsimonious models are rewarded more generously. These obvious features may however distract from a fundamental problem inherent in *CAIC*. The promise that the "true" model will be selected as $N \rightarrow \infty$ rests on the assumption that the true model is contained in the set of competing models. As Burnham and Anderson (1998) pointed out, such an assumption would be rather unrealistic in the biological and social sciences. First, if "an investigator knew that a true model existed and that it was in the set of candidate models, would not he know which one it was?" (p. 69). Second, even if the true model could be identified for a certain finite number of cases, this model would hardly remain the true model as $N \rightarrow \infty$, because in the biological and social sciences, increasing the number of cases usually increases the number of relevant variables, too.

Bozdogan himself (1987, p. 357) conceded that the concept of a true model becomes "suspect" in view of real data, and that the virtue of consistency (which is a large-sample property) should not be exaggerated. Therefore we generally recommend to prefer the original *AIC*, which is intended to identify the "best fitting" model (Takane & Bozdogan, 1987). In any case, the "magic number 2" (Akaike, 1985, p. 1) in the penalty term of *AIC* is not a factor that may be arbitrarily changed.

Expected Cross Validation Index (ECVI)

Another index to be mentioned in the neighborhood of *AIC* is the Expected Cross Validation Index (*ECVI*) of Browne and Cudeck (1989, 1993). The *ECVI* is actually a population parameter which is estimated by the statistic

$$c = F(\mathbf{S}, \mathbf{\Sigma}(\hat{\theta})) + \frac{2t}{N-1}, \tag{23}$$

where

$F(\mathbf{S}, \mathbf{\Sigma}(\hat{\theta}))$ is the minimum value of the fit function,

t is the number of estimated parameters, and

N is the sample size.

Whereas *AIC* was derived from statistical information theory, *ECVI* is a measure of the discrepancy between the model-implied covariance matrix in the analyzed sample and the covariance matrix that would be expected in another sample of the same size (Jöreskog & Sörbom, 1993, p. 120). Thus, *ECVI* evaluates how well a model fitted to the calibration sample would perform in comparable validation samples (Kaplan, 2000). When choosing between several models, the smallest *ECVI* estimate indicates the model with the best fit. In addition, a 90% confidence interval allows to assess the precision of the estimate.

Although derived independently, the *ECVI* estimate leads to the same rank order of competing models as *AIC*, provided that the ML fitting function is used (Browne & Cudeck, 1993, p. 148). In fact, multiplication of *ECVI* by $N - 1$ then leads to the LISREL version of *AIC* given above. Therefore it suffices to take either *AIC* or *ECVI* into account when looking for the model that minimizes the overall error, and it is not necessary to report both indices.

Conclusions and Recommendations

Sample Size and Estimation Method

When the structural equation model is correctly specified and the observed variables are multivariate normal, it can be analytically derived that different estimation procedures, e.g., ML, WLS, and GLS, will produce estimates that converge to the same optimum and have similar asymptotic properties (Browne, 1984). Under ideal conditions the choice of the estimation method is therefore quite arbitrary. But under the more realistic assumption of models that are more or less misspecified and data that are not multivariate normally distributed, the different procedures may not converge to the same optimum (Olsson et al., 2000).

If all variables are measured on an interval scale, if they are normally distributed, and if the sample size is sufficiently large, ML should be applied. Because this method is relatively robust to violations of the normality assumption, it may also be used for models with variables that are not normally distributed, given that the deviation from normality is not too extreme. Robust maximum likelihood estimation may be an alternative but it needs relatively large sample sizes of at least $N \geq 400$.

If the sample size is large and data are non-normally distributed, WLS (ADF) is often recommended. This method may also be considered if some of the observed variables are ordinal and others continuous, or if models including dichotomous variables are being analyzed. But several simulation studies suggest that WLS may not perform too well.

Boomsma and Hoogland (2001, p. 148) found that for a sample size of $N \leq 200$, "ADF is a disaster with more than one-third of the solutions being improper". For nonnormally distributed variables, recommendations are quite divergent. Minimum sample size for WLS estimation should be at least 1,000 (Hoogland & Boomsma, 1998), and depending on the model and data analyzed, in some cases even more than 4,000 or 5,000 cases are required (Boomsma & Hoogland, 2001; Hoogland, 1999; Hu, Bentler & Kano, 1992). In general, results based on small to medium sample sizes should be interpreted with caution (Bollen, 1989, p. 432).

In a simulation study, Olsson et al. (2000) generated data for eleven conditions of kurtosis, three conditions of misspecification, and five different sample sizes. Three estimation methods, ML, GLS, and WLS were compared in terms of overall fit and the discrepancy between estimated parameter values and the true parameter values. Their results show that WLS under no condition was preferable to ML and GLS. In fact, only for large sample sizes of at least $N = 1,000$ and mildly misspecified models WLS provided estimates and fit indices close to the ones obtained for ML and GLS. In general, ML estimation with or without a correction for non-normality (by using either robust χ^2 or bootstrapping) seems to perform better than WLS and should be preferred. WLS for categorical outcomes as implemented in Mplus may not require the same large sample sizes as WLS for continuous, non-normal data, but further simulation studies are needed.

For choosing the adequate estimation method, sample size is an important criterion. Sample size requirements do not only depend on the distribution of the variables (as data become more nonnormal, larger sample sizes are needed), but also on the size of the model, the number of indicator variables, the amount of missing data, the reliability of the variables, and the strength of the relationships among the variables. Therefore one should not simply trust the rules of thumb given in the literature, not even the one stated by Bentler (1995, p. 6) which recommends at least five times the number of free parameters in the model (Bentler & Chou, 1987; Bentler, 1995). In reality there does not exist a rule of thumb that applies to all situations (Muthén & Muthén, 2002).

Most sample size recommendations refer to ML estimation. As simulation research has shown, a reasonable sample size for a correctly specified model and multivariate normally distributed data is about $N = 150$ (Muthén & Muthén, 2002) or $N = 200$ (cf. Hoogland & Boomsma, 1998; Boomsma & Hoogland, 2001).

For a confirmatory factor model with non-normally distributed variables and missing data (missing completely at random), larger samples of about $N = 300$ may be needed (Muthén & Muthén, 2002). With strongly kurtotic data, the minimum sample size should be ten times the number of free parameters (Hoogland & Boomsma, 1998).

The number of indicator variables should also be considered for chosing a sufficient large sample size. Marsh, Hau, Balla, and Grayson (1998) as well as Marsh and Hau (1999) support Boomsma's (1985) recommendations by stating that for confirmatory factor analyses with 6 to 12 indicator variables per factor a sample size of $N = 50$ is sufficient, whereas for 3 to 4 indicators per factor a sample size of $N = 100$ is necessary. With two indicators per factor one should at least have a sample size of $N \geq 400$ (cf. Marsh & Hau, 1999; Boomsma & Hoogland, 2001). There seems to be a mutual compensatory effect of sample size and number of indicators per factor: More indicators may compensate for small sample size, and a larger sample size may compensate for few indicators.

Some evidence exists that simple models could be meaningfully tested even if sample size is quite small (cf. Hoyle, 1999; Hoyle & Kenny, 1999; Marsh & Hau, 1999). Nevertheless, models with moderate to small sample sizes should only be analyzed if a greater sample is not available and if convergence problems or improper solutions, such as negative variance estimates or Heywood cases (cf. Chen, Bollen, Paxton, Curran, & Kirby, 2001), do not occur. As a minimum requirement for parameter estimation using the ML method, sample size should be larger than or at least equal to the number of observed variables ($N \geq p$) (MacCallum, Browne, & Sugawara, 1996, p. 144).

Evaluation of Model Fit

There is a consensus that one should avoid to report all fit indices that have been developed since the first days of SEM, but there is a certain disagreement on just which fit indices to consider for model evaluation.

As the χ^2 test is not only sensitive to sample size but also sensitive to the violation of the multivariate normality assumption (Curran, West, & Finch, 1996; Hu, Bentler, & Kano, 1992; West, Finch, & Curran, 1995), it should not serve as the sole basis for judging model fit. Bollen and Long (1993) as well as Mueller (1996) recommend to evaluate several indices simultaneously which represent different classes of goodness-of-fit criteria. The following criteria form an adequate selection of indices which are frequently presented in current publications: χ^2 and its associated p value, χ^2/df, *RMSEA* and its associated confidence interval, *SRMR*, *NNFI*, and *CFI*. The fit indices *RMSEA*, *NNFI* and *CFI* are sensitive to model misspecifications and do not depend on sample size as strongly as χ^2 (Fan, Thompson, & Wang, 1999; Hu & Bentler, 1998; Rigdon, 1996), therefore they should always be considered. Hu and Bentler (1998) recommend to use *SRMR*, supplemented by *NNFI*, *CFI*, or *RMSEA* derived from ML and GLS estimation (*NNFI* and *RMSEA* are less preferable at small sample sizes), and *SRMR*, *NNFI*, and *CFI* derived from WLS estimation.

For model comparisons (nested models) it is necessary to report additionally the value of the χ^2 difference test and the *AIC* values of all models investigated. In most publications, *GFI* and *AGFI* are also reported. However, these indices depend on sample size and tend to underestimate the fit of complex models (Steiger, 1989).

Other things being equal, a model with fewer indicators per factor may have a higher fit than a model with more indicators per factor, because more indicators per factor provide a more powerful, precise test than a comparable model with fewer indicators (MacCallum et al., 1996). Fit coefficients which reward parsimony, e.g., *RMSEA*, *AIC*, *PGFI*, and *PNFI*, are one way to adjust for this tendency. For choosing between alternative models, parsimony indices may provide important information and should be reported.

As we have demonstrated by discussing different goodness-of-fit indices, it is quite difficult to decide on data-model fit or misfit, especially if various measures of model fit point to conflicting conclusions about the extent to which the model actually matches the observed data. Although there are no well-established guidelines for what minimal conditions constitute an adequate fit, some rules of thumb exist. Table 1 provides an overview over some rule of thumb criteria for goodness-of-fit indices.

Table 1

Recommendations for Model Evaluation: Some Rules of Thumb

Fit Measure	Good Fit	Acceptable Fit
χ^2	$0 \le \chi^2 \le 2df$	$2df < \chi^2 \le 3df$
p value	$.05 < p \le 1.00$	$.01 \le p \le .05$
χ^2/df	$0 \le \chi^2/df \le 2$	$2 < \chi^2/df \le 3$
RMSEA	$0 \le RMSEA \le .05$	$.05 < RMSEA \le .08$
p value for test of close fit ($RMSEA < .05$)	$.10 < p \le 1.00$	$.05 \le p \le .10$
Confidence interval (CI)	close to *RMSEA*, left boundary of CI = .00	close to *RMSEA*
SRMR	$0 \le SRMR \le .05$	$.05 < SRMR \le .10$
NFI	$.95 \le NFI \le 1.00$[a]	$.90 \le NFI < .95$
NNFI	$.97 \le NNFI \le 1.00$[b]	$.95 \le NNFI < .97$[c]
CFI	$.97 \le CFI \le 1.00$	$.95 \le CFI < .97$[c]
GFI	$.95 \le GFI \le 1.00$	$.90 \le GFI < .95$
AGFI	$.90 \le AGFI \le 1.00$, close to *GFI*	$.85 \le AGFI < .90$, close to *GFI*
AIC	smaller than *AIC* for comparison model	
CAIC	smaller than *CAIC* for comparison model	
ECVI	smaller than *ECVI* for comparison model	

Note. *AGFI* = Adjusted Goodness-of-Fit-Index, *AIC* = Akaike Information Criterion, *CAIC* = Consistent *AIC*, *CFI* = Comparative Fit Index, *ECVI* = Expected Cross Validation Index, *GFI* = Goodness-of-Fit Index, *NFI* = Normed Fit Index, *NNFI* = Nonnormed Fit Index, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = Standardized Root Mean Square Residual.

[a] *NFI* may not reach 1.0 even if the specified model is correct, especially in smaller samples (Bentler, 1990). [b] As *NNFI* is not normed, values can sometimes be outside the 0-1 range. [c] *NNFI* and *CFI* values of .97 seem to be more realistic than the often reported cutoff criterion of .95 for a good model fit.

It should be clear that these rule of thumb cutoff criteria are quite arbitrary and should not be taken too seriously. Fit indices may be affected by model misspecification, small-sample bias, effects of violation of normality and independence, and estimation-method effects (Hu & Bentler, 1998). Therefore it is always possible that a model may fit the data although one or more fit measures may suggest bad fit.

Examples

Data Generation

In the following, we will evaluate the fit of four alternative models as examples of poor model fit, adequate fit, and good fit. Using the EQS program (Bentler, 1995), we generated data for $N = 200$ cases according to the "true" model depicted in Figure 1. In this population model, two latent predictor variables and two latent criterion variables are each measured by two indicator variables, and each predictor variable influences both criterion variables. We generated data for the eight X- and Y-variables of the model and computed the sample covariance matrix, which is given in Table 2.

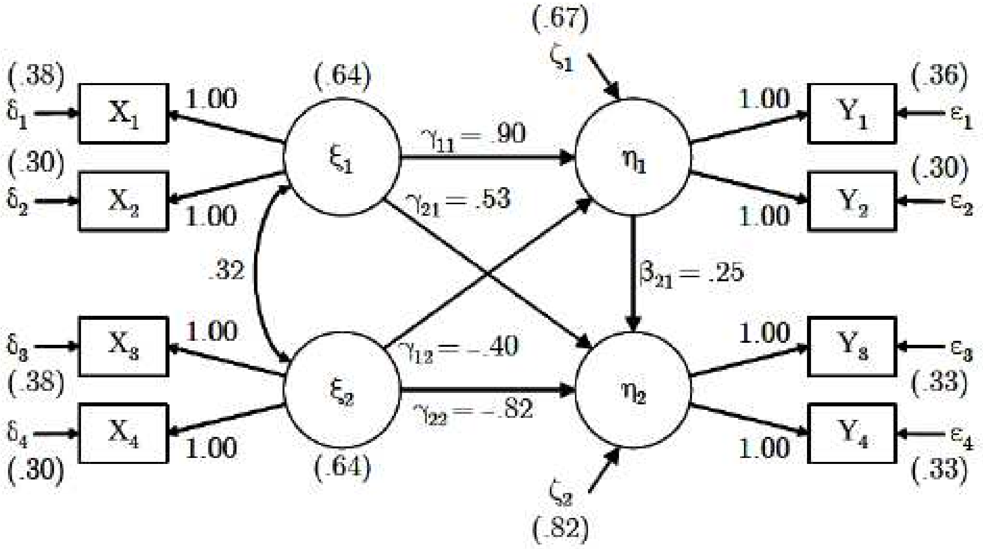

Figure 1. Population model with known parameter values used for data generation.

Specifications of Models A, B, C, and D

According to Hu & Bentler (1998), there are four major problems involved in using fit indices for evaluating goodness of fit: sensitivity of a fit index to model misspecification, small-sample bias, estimation-method effect, and effects of violation of normality and independence. In our present examples, we will only demonstrate the problem of model misspecification.

Table 2

Empirical Covariance Matrix (N = 200)

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4
Y_1	1.429							
Y_2	1.069	1.369						
Y_3	0.516	0.536	1.681					
Y_4	0.436	0.425	1.321	1.621				
X_1	0.384	0.485	0.192	0.183	1.021			
X_2	0.494	0.424	0.181	0.191	0.640	0.940		
X_3	0.021	0.045	-0.350	-0.352	0.325	0.319	1.032	
X_4	0.035	0.013	-0.347	-0.348	0.324	0.320	0.642	0.941

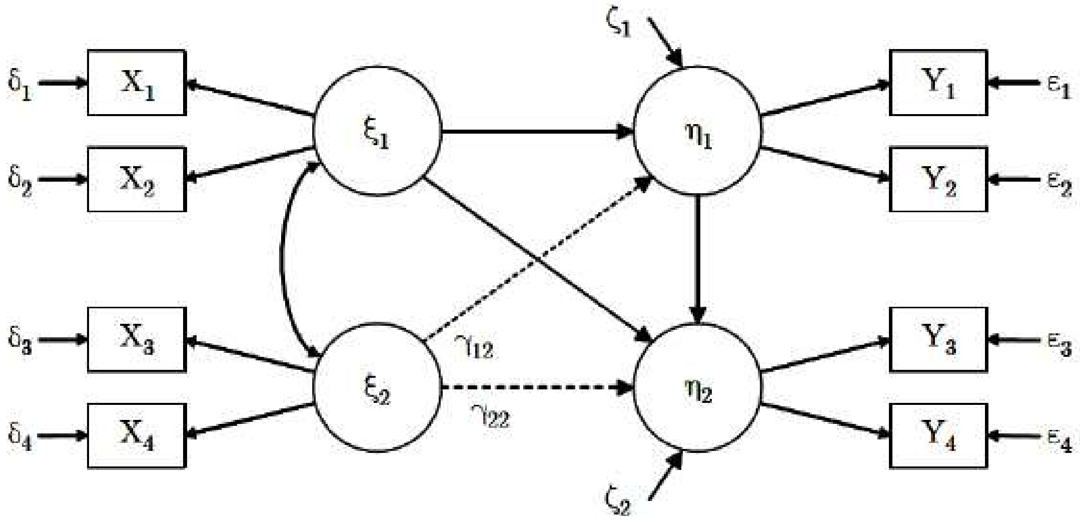

Figure 2. Path diagram for hypothesized models. Four models were analyzed, two misspecified models and two correctly specified models. In the misspecified models, path coefficients γ_{12} and γ_{22} (Model A) or γ_{12} (Model B) were fixed to zero (indicated by dashed lines). In the correctly specified models, all parameters were estimated freely (Model C) or additionally, all factor loadings pertaining to a latent construct were constrained to be equal (Model D).

Four models were specified and analyzed, two misspecified models and two correctly specified models (cf. Figure 2). The first misspecified model (Model A) is severely misspecified with path coefficients γ_{12} and γ_{22} fixed to zero as indicated by dashed lines in Figure 2. In the second misspecified model (Model B) only γ_{12} was fixed to zero. In the first correctly specified model (Model C), all parameters were estimated freely. With the second correctly specified model (Model D), we will demonstrate the effects of parsimony. In this model, all factor loadings pertaining to a latent construct were constrained to be equal, i.e., all factor loadings were fixed to one.

Model Evaluation Procedure

Generally, the first gauge of model-data fit should be the inspection of the fit indices. If there is some indication of misfit, the second gauge should be the inspection of the residual matrix. As the fitted residuals, i.e., discrepancies between the covariance matrices $\mathbf{S}$ and $\mathbf{\Sigma}(\hat{\theta})$, are difficult to interpret if the manifest variables have different variances (or scale units), a more general recommendation is to inspect standardized residuals. Standardized residuals provided by the LISREL program (see above, section on *RMR* and *SRMR*) that are greater than an absolute value of 1.96 ($p < .05$) or 2.58 ($p < .01$) can be useful for detecting the cause of model misfit. The largest absolute value indicates the element that is most poorly fit by the model. A good model should have a high number of standardized residuals close to zero, implying high correspondence between elements of the empirical and the model-implied covariance matrix.

The third gauge should be the inspection of the modification indices provided by LISREL (or the Lagrange multiplier test provided by EQS). These indices provide an estimate of the change in the χ^2 value that results from relaxing model restrictions by freeing parameters that were fixed in the initial specification. Each modification index possesses a χ^2 distribution with $df = 1$ and measures the expected decrease in the χ^2 value when the parameter in question is freed and the model reestimated. Thus, the modification index is approximately equal to the χ^2 difference of two nested models in which the respective parameter is fixed or constrained in one model and set free in the other. The largest modification index belongs to the parameter that improves the fit most when set free. A good model should have modification indices close to one, because $E(\chi^2) = df$.

If the fit indices suggest a bad model-data fit, if one or more standardized residuals are more extreme than ±1.96 ($p < .05$) or ±2.58 ($p < .01$), and if at least one modification index is larger than 3.84 ($p < .05$) or 6.63 ($p < .01$), then one may consider to

Table 3

Goodness-of-Fit Indices for Misspecified and Correctly Specified Models Based on an Artificial Data Set of $N = 200$

	Misspecified models		Correctly specified models	
Fit Index	Model A *Two* paths fixed to zero	Model B *One* path fixed to zero	Model C All parameters estimated freely	Model D All factor loadings fixed to 1
$\chi^2(df)$	54.340 (16)	27.480 (15)	17.109 (14)	17.715 (18)
p value	.000	.025	.250	.475
χ^2/df	3.396	1.832	1.222	0.984
RMSEA	.110	.065	.033	.000
p value for test of close fit (*RMSEA* < .05)	.001	.238	.668	.871
90% CI	.079 ; .142	.023 ; .102	.000 ; .080	.000 ; .062
SRMR	.120	.058	.016	.018
NFI	.922	.963	.977	.977
NNFI	.896	.966	.991	1.000
CFI	.941	.982	.995	1.000
GFI	.936	.967	.979	.978
AGFI	.856	.920	.946	.956
PGFI	.416	.403	.381	.489
PNFI	.527	.516	.489	.628
Model *AIC*	94.340	69.480	61.109	53.715
Saturated *AIC*[a]	72.000	72.000	72.000	72.000
Model *CAIC*	180.306	159.744	155.672	131.085
Saturated *CAIC*[a]	226.739	226.739	226.739	226.739
Model *ECVI*	.474	.349	.307	.271
90% CI	.380 ; .606	.294 ; .443	.291 ; .381	.271 ; .341
Saturated *ECVI*[a]	.362	.362	.362	.362

Note. *AGFI* = Adjusted Goodness-of-Fit-Index, *AIC* = Akaike Information Criterion, *CAIC* = Consistent *AIC*, *CFI* = Comparative Fit Index, *ECVI* = Expected Cross Validation Index, *GFI* = Goodness-of-Fit Index, *NFI* = Normed Fit Index, NNFI = Nonnormed Fit Index, *PGFI* = Parsimony Goodness-of-Fit Index, *PNFI* = Parsimony Normed Fit Index, *RMSEA* = Root Mean Square Error of Approximation, *SRMR* = Standardized Root Mean Square Residual.

[a] Saturated *AIC*, *CAIC*, and *ECVI* serve as possible comparative values for the model *AIC*, *CAIC*, and *ECVI*, respectively.

modify the model by freeing a fixed parameter. But this practice is controversial, as it implies changing the model only in order to improve fit. It would be better to have a substantive theory on which to base modifications, as modifications devoid of a theoretical basis are ill advised (Field, 2000).

In our examples, we will assume to have a theoretical basis for model modifications according to the model depicted in Figure 1. We will modify our initial model (Model A) three times and analyze the covariance matrix given in Table 2 repeatedly. The goodness-of-fit indices taken from the LISREL outputs pertaining to the four models are listed in Table 3.

Although Table 3 lists all fit measures discussed in this article, not all of these measures are necessary for evaluating the fit of the models. It is quite sufficient to assess χ^2 and its associated p-value, χ^2/df, *RMSEA* and its associated confidence interval, *SRMR*, *NNFI*, and *CFI*. For model comparisons it is recommended to assess additionally the χ^2 difference tests (nested models only) and the *AIC* values of all models investigated.

Model A (Severely Misspecified Model)

An inspection of the fit indices for Model A (Table 3) suggests to reject the model, as $\chi^2/df > 3$, $RMSEA > .08$, p-value for test of close fit ($RMSEA < .05$) is almost zero, lower boundary of the confidence interval does not include zero, $SRMR > .10$, $NNFI <$ .95, and $CFI < .95$. Only $NFI > .90$ and $GFI > .90$ are indicative of an acceptable model fit.

An inspection of the standardized residuals in Table 4 reveals 11 significant residuals. Because of the misspecifications in Model A, relations of Y_1 and Y_2 with X_3 and X_4 are overestimated by the model (residuals are negative while the respective sample covariances are positive, cf. Table 2), whereas relations of Y_3 and Y_4 with X_3 and X_4 are underestimated (residuals are negative while the respective empirical covariances are negative).

An inspection of the modification indices (Appendix B1) reveals that the largest modification index of 24.873 pertains to parameter γ_{22}. This suggests that the most substantive change in the model in terms of improvement of fit would arise from relaxing the constraint on the respective structural coefficient, which would result in an unstandardized γ_{22} estimate of about $-.715$ according to the LISREL output. We followed this suggestion and set γ_{22} free.

Table 4

Standardized Residuals of Model A (Severely Misspecified Model)

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4
Y_1	--							
Y_2	--	--						
Y_3	-0.256	0.237	--					
Y_4	0.379	-0.111	--	--				
X_1	-1.056	**2.079**	1.216	1.168	--			
X_2	**2.408**	-0.587	1.268	1.425	**-2.727**	--		
X_3	**-2.620**	**-2.390**	**-4.953**	**-4.752**	0.871	0.481	--	
X_4	**-2.540**	**-3.026**	**-5.194**	**-4.955**	0.944	0.642	--	--

Note. Significant values are in boldface.

An inspection of the standardized residuals in Table 4 reveals 11 significant residuals. Because of the misspecifications in Model A, relations of Y_1 and Y_2 with X_3 and X_4 are overestimated by the model (residuals are negative while the respective sample covariances are positive, cf. Table 2), whereas relations of Y_3 and Y_4 with X_3 and X_4 are underestimated (residuals are negative while the respective empirical covariances are negative).

An inspection of the modification indices (Appendix B1) reveals that the largest modification index of 24.873 pertains to parameter γ_{22}. This suggests that the most substantive change in the model in terms of improvement of fit would arise from relaxing the constraint on the respective structural coefficient, which would result in an unstandardized γ_{22} estimate of about $-.715$ according to the LISREL output. We followed this suggestion and set γ_{22} free.

Model B (Misspecified Model)

The fit indices for Model B point to conflicting conclusions about the extent to which this model actually matches the observed data. All fit indices – with the exception of the parsimony indices *PNFI* and *PGFI* – are improved compared to Model A, and the model modification resulted in a significant χ^2 difference test $(54.34 - 27.48 = 26.86,$

$df = 1$, $p < .01$). Furthermore, model *AIC*, model *CAIC*, and model *ECVI* are smaller for Model B than for Model A. Indications of a good model fit are $\chi^2/df < 2$, $NFI > .95$, $GFI > .95$, and $CFI > .97$, whereas other descriptive goodness-of-fit indices suggest only an acceptable fit ($RMSEA > .05$, $SRMR > .05$, $NNFI < .97$). Moreover, the lower boundary of the *RMSEA* confidence interval still does not include zero.

An inspection of the standardized residuals for Model B (Table 5) reveals that relations of Y_1 and Y_2 with X_3 and X_4 are still overestimated by the model (note that the empirical covariances are positive), but that the largest standardized residuals show an underestimation of the relation between Y_3 and Y_4 and an overestimation of the relation between X_1 and X_2.

An inspection of the modification indices (Appendix B2) shows that two modification indices are equally large (10.533) suggesting to either free parameter γ_{12} (with an expected unstandardized estimate of about −.389), or to free β_{12} (with an expected unstandardized estimate of about .488). As both modification indices are of the same size, the decision which constraint to relax can only be made on a theoretical basis. In our example, we relaxed the constraint on parameter γ_{21}.

Table 5

Standardized Residuals of Model B (Misspecified Model)

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4
Y_1	--							
Y_2	--	--						
Y_3	**2.329**	**2.923**	**3.245**					
Y_4	1.181	1.005	**3.245**	**3.245**				
X_1	-1.204	**1.997**	-0.069	0.033	–			
X_2	**2.465**	-0.478	-0.533	0.139	**-3.246**	--		
X_3	**-2.569**	**-2.315**	-0.920	-1.344	0.708	0.423	--	
X_4	**-2.480**	**-2.927**	-1.075	-1.493	0.859	0.652	--	--

Note. Significant values are in boldface.

Model C (Correctly Specified Model)

After the second model modification with parameter γ_{21} now set free, the fit of Model C (correctly specified model) improved substantially with a significant drop in the χ^2 value ($27.480 - 17.109 = 10.371$, $df = 1$, $p < .01$) and smaller *AIC*, *CAIC*, and *ECVI* values for Model C compared to Model B. The descriptive goodness-of-fit indices point to a good model fit with $\chi^2/df < 2$, $RMSEA < .05$ (lower boundary of the *RMSEA* confidence interval contains zero), *NNFI* and $CFI > .97$, *NFI* and $GFI > .95$, $AGFI > .90$, and $SRMR < .05$. The only exceptions are again the parsimony fit indices *PNFI* and *PGFI* with smaller values for Model C than for Model B. The standardized residuals (Table 6) for Model C are now all close to zero and not significant.

Table 6

Standardized Residuals of Model C (Correctly Specified Model)

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4
Y_1	--							
Y_2	--	--						
Y_3	0.432	0.982	--					
Y_4	-0.667	-1.216	--	--				
X_1	-1.823	1.161	0.018	0.107	--			
X_2	1.736	-1.139	-0.297	0.291	--	--		
X_3	-0.134	0.381	0.224	-0.424	0.110	-0.117	--	
X_4	0.175	-0.400	0.356	-0.403	0.107	-0.081	--	--

An inspection of the modification indices (Appendix B3) reveals that the parameters of four covariances between error variables should be freed to improve the model fit. But as our theory (Figure 1) does not support this modification, such suggestions can be ignored.

Model D (Correctly Specified Parsimonious Model)

In order to demonstrate the effects of parsimony on model fit, we fixed all factor loadings to one. The resulting Model D (correctly specified parsimonious model) is still in accordance with the population model depicted in Figure 1. The model fit again improved with $\chi^2/df < 1$, $RMSEA = .00$, $NNFI = 1.00$, $CFI = 1.00$, and $AGFI > .95$. Model D is more parsimonious than Model C with smaller values of *AIC*, *CAIC*, *ECVI*, and *RMSEA*, and with larger parsimony indices *PNFI*, *PGFI*, and larger *AGFI*. Standardized residuals are all small (Table 7), most of them close to zero. But still modification indices (Appendix B4) suggest to lower the constraints on four error covariances, a suggestion that we will not follow because our theory (Figure 1) does not support such modification.

Table 7

Standardized Residuals of Model D (Correctly Specified Parsimonious Model)

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4
Y_1	-0.108							
Y_2	-0.108	0.108						
Y_3	0.729	1.251	0.764					
Y_4	-0.815	-1.157	0.763	-0.764				
X_1	-1.326	0.811	0.107	-0.074	-0.075			
X_2	1.087	-0.648	-0.135	0.104	-0.075	0.075		
X_3	-0.139	0.379	-0.019	-0.060	0.074	-0.068	0.014	
X_4	0.172	-0.391	0.048	0.025	0.059	-0.050	0.014	-0.014

Discussion

Our examples demonstrate how to evaluate model fit and model misfit and how to modify a model. If goodness-of-fit indices suggest that a model does not fit, an inspection of significant standardized residuals and the largest modification indices may be extremely helpful in deciding at which part of the model a modification should be considered. But one should never modify a model solely on the basis of modification indices, although the program might suggest to do so. Model modification based on modification indices may result in models that lack validity (MacCallum, 1986) and are highly susceptible to capitalization on chance (MacCallum, Roznowski, & Necowitz, 1992). Therefore the modifications should be defensible from a theoretical point of view.

To sum up, model fit improved substantially from Model A to Model D as expected from the construction of our simulated example. One may ask why there is no "perfect fit" of Models C and D although both models are correctly specified. The first reason is that even Model D is not yet the most parsimonious model. The most parsimonious model requires that *all* parameters are fixed at their true population values, but these values are of course unknown in practice. Usually the best one can do is to fix *some* parameters or to impose certain equality constraints, which also leads to more parsimonious models. The second reason is that only sample covariance matrices are analyzed instead of the population covariance matrix. Whereas the population covariance matrix should fit a correctly specified model "perfectly", sample covariance matrices will usually do not, as they are subject to sampling fluctuations. Of course, the χ^2 test takes these fluctuations into account and most fit indices should signalize "good fit", as we could demonstrate for Models C and D.

As has been pointed out by Jöreskog (1993), it is important to distinguish between three types of analyses with SEM: strictly confirmatory approach (CM), alternative models approach (AM), and model generating approach (MG). If the analysis is not strictly confirmatory, that is, if the first model for a given data set must be modified several times until a good or acceptable fit is reached – either because of testing alternative models (AM) or because of developing a new model (MG) – it is possible or even likely that the resulting model to a considerable degree only reflects capitalization on chance. In other words, subsequent modifications based on the same data always imply the danger that a model is only fitted to some peculiarities of the given sample. If our example would rest on real (instead of simulated) empirical data, we would strongly suggest to cross-validate Model D.

It should be kept in mind that even if cross-validation is successful, it is not allowed to infer that the respective model has been "proved". As is known from the work of Stelzl (1986), Lee and Hershberger (1990), and MacCallum, Wegener, Uchino, and Fabrigar (1993), several different models may fit the data equally well. Different models can imply the same covariance matrix and thus be empirically indistinguishable from the SEM viewpoint. This even holds if the competing models are contradictory from a causal perspective. In order to decide if the model of interest is the "best" model one must be able to exclude all equivalent models on logical or substantive grounds (Jöreskog, 1993).

Let us therefore close with the friendly reminder that it is not possible to confirm any tested model but that it is only possible to reject improper models, as we have done with Models A and B. At the same time, significance tests and descriptive goodness-of-fit indices are a valuable tool in finding a structural equation model that can be accepted in a preliminary sense. If a model is accepted, it may be consistent with the data, but still it may not be consistent with reality. "If a model is consistent with reality, then the data should be consistent with the model. But, if the data are consistent with a model, this does not imply that the model corresponds to reality" (Bollen, 1989, p. 68). Therefore researchers should pay attention "to the important but difficult question regarding the discrepancy between the simplified, formalized theoretical model and the actual operating processes that govern the phenomena under study in the real empirical world" (Boomsma, 2000, p. 477).

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE transactions on Automatic Control, 19*, 716-723.

Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson & S. E. Fienberg (Eds.), *A celebration of statistics* (pp. 1-24). New York: Springer.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155-173.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 users' guide*. Chicago: SmallWaters.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425–430.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16,* 78–117.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47,* 563–592.

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

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Bollen, K., & Long, J. S. (Eds.). (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 52,* 345–370.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling, 7,* 461–483.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Browne, M. W., & Mels, G. (1992). *RAMONA user's guide.* The Ohio State University: Department of Psychology.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach.* New York: Springer.

Chen, F., Bollen, K., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research, 29,* 468–508.

Chou, C.-P., & Bentler, P. M. (1995). Estimation and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Thousand Oaks, CA: Sage.

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347–357.

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56–83.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.

Hayduk, L. A. (1989). *Structural equation modeling with LISREL*. Baltimore: Johns Hopkins University Press.

Hayduk, L. A. (1996). *LISREL issues, debates, and strategies*. Baltimore: Johns Hopkins University Press.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*, 329–367.

Hoyle, R. H. (1999). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage.

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195-222). Thousand Oaks, CA: Sage.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 76–99). London: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.

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

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351–362.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods* (Research Report 81–8). Uppsala, Sweden: University of Uppsala, Department of Statistics.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's reference guide*. Chicago: SPSS Publications.

Jöreskog, K. G., & Sörbom, D. (1993). *Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Chicago: Scientific Software.

Kaplan, D. (2000). *Structural equation modeling: Foundation and extensions*. Thousand Oaks, CA: Sage Publications.

Kumar, A., & Sharma, S. (1999). A metric measure for direct comparison of competing models in covariance structure analysis. *Structural Equation Modeling, 6*, 169–197.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25*, 313–334.

MacCallum R. C. (1986). Specification searches in covariance structure modelling. *Psychologival Bulletin*, 100, 107–20.

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods* 1, 130–149.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185–199.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 177–198). Thousand Oaks, CA: Sage.

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample size* (pp. 251–306). Thousand Oaks, CA: Sage.

Marsh, H. W., Hau, K-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 103*, 391–411.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York: Springer.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430–445.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19–30.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus: The comprehensive modeling program for applied researchers*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599–620.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557–595.

Raykov, T., & Penev, S. (1998). Nested structural equation models: Noncentrality and power of restriction test. *Structural Equation Modeling, 5,* 229–246.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling, 3,* 369–369.

Rigdon, E. E. (1999). Using the Friedman method of ranks for model comparison in structural equation modeling. *Structural Equation Modeling, 6,* 219–232.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra, (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233–247). London: Kluwer Academic Publishers.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shipley, B., 2000. *Cause and correlation in biology.* University Press, Cambridge.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

Steiger J. (1995). SEPATH structural equation modeling, *Statistica, 3,* 3539–3689.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–264.

Stelzl, I. (1986). Changing causal relationships without changing the fit: Some rules for generating equivalent LISREL models. *Multivariate Behavioral Research, 21,* 309–331.

Takane, Y., & Bozdogan, H. (1987). Introduction to special section [on AIC]. *Psychometrika, 52,* 315.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–40). Newbury Park, CA: Sage.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology, 46,* 621–635.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

West, S. G, Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Yang-Wallentin, F., & Joreskog, K. G. (2001). Robust standard errors and chi-squares for interaction models. In G. A. Marcoulides and R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 159-171). Mahwah, NJ: Lawrence Erlbaum.

APPENDIX A[4]

A1 LISREL Input File for Model A (Severely Misspecified Model)

```
!Model A: GA(1,2) and GA(2,2) fixed to zero
DA NI=8 NO=200 MA=CM
CM SY
1.429
1.069  1.369
0.516  0.536   1.681
0.436  0.425   1.321   1.621
0.384  0.485   0.192   0.183  1.021
0.494  0.424   0.181   0.191  0.640  0.940
0.021  0.045  -0.350  -0.352  0.325  0.319  1.032
0.035  0.013  -0.347  -0.348  0.324  0.320  0.642  0.941
LA
Y1 Y2 Y3 Y4 X1 X2 X3 X4
MO NX=4 NK=2 NY=4 NE=2 LX=FU,FI LY=FU,FI TD=DI,FR TE=DI,FR GA=FU,FR C
        BE=FU,FI PH=FU,FR
LK
KSI1 KSI2
LE
ETA1 ETA2
VA 1 LX(1,1) LX(3,2)
VA 1 LY(1,1) LY(3,2)
FR LX(2,1) LX(4,2)
FR LY(2,1) LY(4,2)
FI GA(1,2) GA(2,2)
FR BE(2,1)
PATH DIAGRAM
OU ND=3 ME=ML RS SC MI
```

A2 LISREL Input File for Model B (Misspecified Model)

```
!Model B: GA(1,2) fixed to zero
DA NI=8 NO=200 MA=CM
CM SY
1.429
1.069  1.369
0.516  0.536   1.681
0.436  0.425   1.321   1.621
0.384  0.485   0.192   0.183  1.021
0.494  0.424   0.181   0.191  0.640  0.940
0.021  0.045  -0.350  -0.352  0.325  0.319  1.032
0.035  0.013  -0.347  -0.348  0.324  0.320  0.642  0.941
LA
Y1 Y2 Y3 Y4 X1 X2 X3 X4
MO NX=4 NK=2 NY=4 NE=2 LX=FU,FI LY=FU,FI TD=DI,FR TE=DI,FR GA=FU,FR BE=FU,FI
PH=FU,FR
LK
KSI1 KSI2
LE
ETA1 ETA2
VA 1 LX(1,1) LX(3,2)
VA 1 LY(1,1) LY(3,2)
FR LX(2,1) LX(4,2)
```

[4] Appendix A is stored as an ASCII file next to this article at http://www.mpr-online.de

```
FR LY(2,1) LY(4,2)
FI GA(1,2)
FR BE(2,1)
PATH DIAGRAM
OU ND=3 ME=ML RS SC MI
```

A3 LISREL Input File for Model C (Correctly Specified Model)

```
!Model C: correctly specified model
DA NI=8 NO=200 MA=CM
CM SY
1.429
1.069  1.369
0.516  0.536   1.681
0.436  0.425   1.321   1.621
0.384  0.485   0.192   0.183  1.021
0.494  0.424   0.181   0.191  0.640  0.940
0.021  0.045  -0.350  -0.352  0.325  0.319  1.032
0.035  0.013  -0.347  -0.348  0.324  0.320  0.642  0.941
LA
Y1 Y2 Y3 Y4 X1 X2 X3 X4
MO NX=4 NK=2 NY=4 NE=2 LX=FU,FI LY=FU,FI TD=DI,FR TE=DI,FR GA=FU,FR C
   BE=FU,FI PH=FU,FR
LK
KSI1 KSI2
LE
ETA1 ETA2
VA 1 LX(1,1) LX(3,2)
VA 1 LY(1,1) LY(3,2)
FR LX(2,1) LX(4,2)
FR LY(2,1) LY(4,2)
FR BE(2,1)
PATH DIAGRAM
OU ND=3 ME=ML RS SC MI
```

A4 LISREL Input File for Model D (Correctly Specified Parsimonious Model)

```
!Model D: equality constraints on LX and LY
DA NI=8 NO=200 MA=CM
CM SY
1.429
1.069  1.369
0.516  0.536   1.681
0.436  0.425   1.321   1.621
0.384  0.485   0.192   0.183  1.021
0.494  0.424   0.181   0.191  0.640  0.940
0.021  0.045  -0.350  -0.352  0.325  0.319  1.032
0.035  0.013  -0.347  -0.348  0.324  0.320  0.642  0.941
LA
Y1 Y2 Y3 Y4 X1 X2 X3 X4
MO NX=4 NK=2 NY=4 NE=2 LX=FU,FI LY=FU,FI TD=DI,FR TE=DI,FR GA=FU,FR C
        BE=FU,FI PH=FU,FR
LK
KSI1 KSI2
LE
ETA1 ETA2
VA 1 LX(1,1) LX(3,2)
VA 1 LY(1,1) LY(3,2)
VA 1 LX(2,1) LX(4,2)
VA 1 LY(2,1) LY(4,2)
FR BE(2,1)
PATH DIAGRAM
OU ND=3 ME=ML RS SC MI
```

APPENDIX B

B1 Modification Indices for Model A (Severely Misspecified Model)

```
No Non-Zero Modification Indices for BETA

        Modification Indices for GAMMA

               KSI1        KSI2
           --------    --------
    ETA1       - -       13.505
    ETA2       - -       24.873

        Expected Change for GAMMA

               KSI1        KSI2
           --------    --------
    ETA1       - -       -0.440
    ETA2       - -       -0.715

        Standardized Expected Change for GAMMA

               KSI1        KSI2
           --------    --------
    ETA1       - -       -0.344
    ETA2       - -       -0.448
```

B2 Modification Indices for Model B (Misspecified Model)

```
        Modification Indices for BETA

               ETA1        ETA2
           --------    --------
    ETA1       - -       10.533
    ETA2       - -          - -

        Expected Change for BETA

               ETA1        ETA2
           --------    --------
    ETA1       - -        0.488
    ETA2       - -          - -

        Standardized Expected Change for BETA

               ETA1        ETA2
           --------    --------
    ETA1       - -        0.406
    ETA2       - -          - -

        Modification Indices for GAMMA

               KSI1        KSI2
           --------    --------
    ETA1       - -       10.533
    ETA2       - -          - -

        Expected Change for GAMMA
```

```
                KSI1       KSI2
            --------   --------
    ETA1          - -     -0.389
    ETA2          - -        - -

        Standardized Expected Change for GAMMA

                KSI1       KSI2
            --------   --------
    ETA1          - -     -0.304
    ETA2          - -        - -
```

B3 Modification Indices for Model C (Correctly Specified Model)

```
 No Non-Zero Modification Indices for BETA

 No Non-Zero Modification Indices for GAMMA

        Modification Indices for THETA-DELTA-EPS

                  Y1         Y2         Y3         Y4
            --------   --------   --------   --------
      X1      11.327      9.719      0.005      0.040
      X2      10.705      9.216      0.942      0.812
      X3       0.537      0.768      0.126      0.198
      X4       0.517      0.734      0.251      0.197

        Expected Change for THETA-DELTA-EPS

                  Y1         Y2         Y3         Y4
            --------   --------   --------   --------
      X1      -0.145      0.133     -0.003      0.008
      X2       0.137     -0.126     -0.038      0.035
      X3      -0.030      0.035      0.016     -0.019
      X4       0.028     -0.032      0.022     -0.019

        Completely Standardized Expected Change for THETA-DELTA-EPS

                  Y1         Y2         Y3         Y4
            --------   --------   --------   --------
      X1      -0.120      0.112     -0.002      0.006
      X2       0.119     -0.111     -0.030      0.028
      X3      -0.025      0.029      0.012     -0.015
      X4       0.024     -0.028      0.017     -0.015
```

B4 Modification Indices for Model D

(Correctly Specified Parsimonious Model)

```
No Non-Zero Modification Indices for BETA

No Non-Zero Modification Indices for GAMMA

        Modification Indices for THETA-DELTA-EPS

                   Y1          Y2          Y3          Y4
              --------    --------    --------    --------
      X1        11.327       9.719       0.005       0.040
      X2        10.705       9.216       0.942       0.812
      X3         0.537       0.768       0.126       0.198
      X4         0.517       0.734       0.251       0.197

        Expected Change for THETA-DELTA-EPS

                   Y1          Y2          Y3          Y4
              --------    --------    --------    --------
      X1        -0.145       0.133      -0.003       0.008
      X2         0.137      -0.126      -0.038       0.035
      X3        -0.030       0.035       0.016      -0.019
      X4         0.028      -0.032       0.022      -0.019

        Completely Standardized Expected Change for THETA-DELTA-EPS

                   Y1          Y2          Y3          Y4
              --------    --------    --------    --------
      X1        -0.120       0.112      -0.002       0.006
      X2         0.119      -0.111      -0.030       0.028
      X3        -0.025       0.029       0.012      -0.015
      X4         0.024      -0.028       0.017      -0.015
```