UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 7

2009

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Can Fit Indices be used to evaluate Structural Equation Models?

Irmgard W. Corten

Economic Institute for Construction, Amsterdam

Willem E. Saris

ESADE / Universitat Ramon Llull , Barcelona

Albert Satorra

Universitat Pompeu Fabra. Barcelona

Correspondence should be addressed to:

Prof.Dr. W.E. Saris

C. Josep Pla 27 / 9 / 4

08019 Barcelona . Spain

Email : w.saris@telefonica.net

Can Fit Indices be used to evaluate Structural Equation Models?

Abstract

In recent years, many different fit indices have been formulated as an alternative for the standard likelihood ratio test (LRT) of Structural Equation Models (SEM). These fit indices were developed to solve specific problems associated with the LRT namely, sensitivity to sample size; the problem that no SEM will ever be an exact representation of reality; and the problem of deviation from the assumptions of the standard test. There is, however, one problem related to using the LRT that has been largely ignored by the developers of fit indices and this is the LRT's varying sensitivity to the different characteristics of a model and different types of error. Because of this problem, it is impossible to use the LRT to test the fit of models with a fixed critical value. Since most new fit indices are functions of the fitting function or the test statistic itself, it was expected that the fit indices would have the same sensitivity problems as the LRT statistic. In the present paper, we confirm this by means of a Monte Carlo experiment. We also show that fit indices do not provide a simple instrument to test the fit of models. We conclude that the current practice of evaluating the fit of a model on the basis of the value of fit index and a general specified threshold is not justified.

Keywords: Structural Equation Models (SEM), Likelihood Ratio Test (LRT), chi-square goodness-of-fit test, power, sensitivity analysis, goodness-of-fit indices

Introduction

An essential aspect of Structural Equation Modeling is assessing how well the model fits. For some time, the standard Likelihood Ratio Test (LRT) has been the most commonly used statistic to test the fit of a model. The LRT test, also referred to as the X^2 test, tests the null hypothesis that the analyzed model is true against the alternative general model that the variables are just freely correlated.

Many authors have argued that there are problems associated with the use of the LRT statistic[1] (among them Bentler and Bonett, 1980; Anderson and Gerbing, 1984; Long, 1983; Marsh and Hocevar, 1985; Saris and Satorra, 1988; Saris, Satorra and Sörbom, 1988; Saris, den Ronden and Satorra, 1984; Cudeck and Henly, 1991; Browne and Cudeck, 1992, Hu, Bentler & Kano, 1992).

First of all, the assumptions underlying the test are seldom fully met in practice. For example, the LRT tests whether the hypothesized model fits the data exactly within the bounds of sampling error. Yet, a hypothetical model is by definition an approximation of reality. Many researchers have, therefore, asked what is the point of a test for exact fit when we already know that a model can never be an exact reflection of reality (Cudeck and Henly, 1991; Browne and Cudeck, 1992).

Another problem is that of the violation of distributional assumptions, such as multivariate normality of the data. Asymptotic distribution-free X^2 goodness-of-fit tests (Browne, 1984) and other robust X^2 goodness-of-fit test statistics, such as the scaled test statistic by Satorra and Bentler (1994), have been developed to deal with this problem.

[1] and the X^2 difference test

In addition, many researchers have emphasized how great the influence of sample size is on the X^2 goodness-of-fit test; that is, the probability of rejecting a model with an irrelevant specification error tends to 1 as sample size increases. However, this is quite a desirable characteristic for a statistical test. In fact, a test statistic without this characteristic would not satisfy what is recognized to be the basic requirement of a test, the so-called property of consistency, which implies that the power should tend to 1 when sample size tends to infinity.

The problems encountered in applying the LRT to SEM models are well known. However, the fact that in the standard use of the LRT the power of the test is not taken into account is less often recognized as a problem, although there have been some exceptions (Saris and Satorra, 1988; Saris, Satorra and Sörbom, 1987; Saris, den Ronden and Satorra, 1987). Saris and Satorra (1988) and Saris, Satorra and Van der Veld (2009) showed that the power of the test varies with the characteristics of the model such as the number of indicators per factor, the size of the loadings and the size of the correlation between the factors. They showed that the power of the 5% level test for the same kind and size of error can vary from .05 to 1.00. As a consequence, under certain conditions misspecification can lead to rejection of the model whereas under other conditions the same error will pass undetected.

Saris, Satorra and Sörbom (1987) showed that the LRT is generally applied as a test for the whole model even though the test has a different sensitivity to different kinds of error. That is to say, the LRT does not weight specification errors associated with various restrictions of the model equally. There are some restrictions for which the test is more sensitive. It is therefore debatable whether it is possible to test the fit of a whole model only by means of the LRT.

The conclusion that can be drawn from the properties of LRT mentioned above is that standard procedures to test goodness-of-fit which ignore the power of the test, are too simple. If a SEM model is tested and the test statistic is significant at the 5% level, the model is

usually rejected. However, if the power of the test is high, i.e. the test statistic is very sensitive to minor specification errors, it is not clear whether the significant value of the test statistic is due to large or small specification errors in the model and, therefore, the simple rejection of the model is no longer justified. On the other hand, when the test statistic is not significant at the 5% level, one tends not to reject the model. However, when the power of the test is also low, even large specification errors have a low probability of being detected. For this reason, Saris, Satorra and Sörbom (1987) concluded that making decisions on the basis of the LRT test statistic is not justified unless the power of the test is taken into account.

To cope with the problem of the sample size and with the test of exact fit, alternative measures of fit have been developed and put forward. However, since most of these indices are based on the fitting function or on the LRT statistic, we expect these indices to display the same problems as the LRT i.e. vary according to the characteristics of the model. The problems will be analogous to those of the LRT if one uses a fixed critical value for the fit index, as is normally done: some misspecifications will be detected while others will not, depending on the characteristics of the model and the type of misspecification. This would imply that these fit indices can not be used in a simple way, that is, evaluating the fit of a model by comparing the value of the fit measure with a fixed threshold level. This means that the routine use of fit indices in this simplistic way is unjustified; The sensitivity of these indices to the characteristics of the model and the types of errors should always be taken into account when using fit indices. Saris, Satorra and Van der Veld (2009) have shown this for some simple but fundamental models. Here we would like to illustrate that this phenomenon is very general i.e. it holds for many more models and fit indices.

We begin with a brief description of well known and routinely used goodness-of-fit indices. Next, we will address the problems in the application of these fit indices and pay some attention to other studies on this subject. Then, we describe our own study design. After

that, we present the results of the Monte Carlo experiments which show that the same specification error under different circumstances can lead to considerably different values for the fit indices. Finally, the consequences of our results are discussed.

Goodness-of-fit indices

To overcome the effects of sample size on the LRT, several researchers proposed alternative goodness-of-fit indices which they claimed did not vary with the size of the sample. Most of these indices depend indirectly on the minimum value of the fitting function. They are usually classified either as stand-alone indices or incremental fit indices[2] (Bollen, 1989; Gerbing and Anderson, 1993; Marsh et al., 1988; Tananaka, 1993). Stand-alone indices are based on the results of one hypothesized model and assess the degree to which the model accounts for sample covariances. Examples include X^2 divided by its degrees of freedom, Hoelter's CN, the information criterion of Akaike (AIC) and the modification of Akaike's criterion by Schwartz (SK). [3]

Incremental fit indices[4] compare the hypothesized model under study with a baseline model. Usually the baseline model is a null model in which all indicators are specified as being uncorrelated, although other baseline models have also been suggested (Sobel & Borhnstedt, 1985). The Bentler-Bonett Index, the Relative Fit Index, the Tucker-Lewis index and the Bollen's Δ_2 indicator are all examples of this type of indicator.

[2] Also other classifications are used. For example, Fan, Thompson and Wang (1999) distinguish three types of fit indices; Tanaka (1993) suggested categorising fit indices along 6 dimensions.

[3] For recent discussion on fit indices, see the March 2000 Special issue on Model Selection in the Journal of Mathematical Psychology.

[4] It is common use to subdivide these incremental fit indices into two (see for example Marsh, Balla and McDonald (1988) or three subtypes (Hu and Bentler, 1999).

There are also two fit measures that do not use the independence model as the baseline model but use a model that assumes that all elements of the covariance matrix are zero, not only the off-diagonal elements. These two measures are the Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit-Index (AGF).

To deal with issues of deviation from an exact fit (since all the LRT tests are concerned with the null hypothesis that a model exactly fits the population covariance matrix), Bentler (1990) put forward three fit indices - Δ, FI and CFI -based on non-centrality. These three fit indices are also incremental fit measures. Browne and Cudeck (1992) introduced the RMSEA test of close fit which is a stand-alone measure[5] based on a so-called non-centrality parameter.

All fit indices mentioned so far are based on value(s)-of-fit functions, the LRT statistic(s) or non-centrality parameter(s). There is, however, also a fit measure that is not based on any of these statistics but on the residuals between observed and reproduced covariances. It is called the Standardized Root Mean Square Residual (SRMR).

Table 1 gives an overview of the names, references, cut-off values and formal definitions of the most commonly used fit indices that will be evaluated in this study.

Table 1 about here

Problems in the application of fit indices

Hu and Bentler (1998) mention four important problems in assessing the fit of a model by means of goodness-of-fit indices: 1) sensitivity to model misspecification, 2) small-sample bias, 3) effects of the estimation method and 4) effects of violation of normality and

[5] There have been many more fit indices developed along this line. For example, MCDonald and Marsh (1990) suggested the RNI index; McDonald (1989) introduced the index μ_h; Cudeck and Henly (1991) introduced F_0 ;

independence. Another important aspect is the interpretation of the value of the fit indices, in other words, determination of the cut-off criteria, above or below which a model is considered as being a correct model.

Several studies have shown that a lot of the goodness-of-fit indices are still substantially affected by sample size (Anderson and Gerbing, 1984; Bollen, 1986, 1989, 1990; Bentler, 1990; Bollen and Liang, 1988; Browne and Cudeck, 1992; Cudeck and Browne,1983, Hu and Bentler, 1998; Marsh, Balla and McDonald, 1988; Marsh, Balla and Hau, 1996; McDonald and Marsh, 1990). As a consequence, researchers with different sample sizes could come to different model choices. The effect of sample size varies according to the data set and the fit index (Marsh, Balla and McDonald, 1988).

Sobel and Bohrnstedt (1985) have argued that the baseline model recommended by Bentler and Bonett is not appropriate, except in purely exploratory cases. They claim that the choice of a baseline model should depend on theoretical considerations and prior knowledge about the concepts being studied.

Weng and Cheng (1997) emphasise that relative fit indices that use the null model as a baseline model, differ with respect to their estimation methods because the function values and the X^2 test statistic vary across the methods of estimation. Therefore, they claim that a fixed cut-off criterion for such indices, independent of the estimation method, is not advisable.

Finally, James et al. (1982) suggested adjusting (incremental) fit indices for loss of degrees of freedom, by multiplying them by the ratio d_h/d_0, where d_h represents the degree of freedom of the hypothesised model and d_0 the degree of freedom of the baseline model. The importance of parsimony is also addressed for example by Mulaik et al. (1989) and Bentler and Mooijaart (1989).

McDonald and Marsh (1990), although acknowledging the importance of parsimony, claimed that there 'does not appear to be any mathematical basis for deciding what function of the parsimony ratio and the badness-of-fit ratio would weight them appropriately'. Williams and Holahan (1994) showed that the efficiency of the parsimony correction index (the one considered by them) is influenced by the type of the model and the number of indicators per latent variable.

Fixed critical values of fit indices

As we have shown, a lot has been said about fit indices, which are currently so popular, but the sensitivity of these indices with respect to other aspects of the model is not mentioned.

As can be concluded from the goodness-of-fit indices shown in Table 1, most fit indices are somehow based on the minimum of the fitting function or on the LRT. Therefore, to a certain extent they should display the same application problems as the LRT. If this is correct, one can also not use fixed critical values or cut-off points for the different fit measures as is currently common practice.

There are, nevertheless, fixed cut-off criteria specified for most of the fit indices. The commonly used cut-off criteria for the different indices are shown in Table 1. Marsh (1995) emphasized that no rationale has been given for these values. Because many severely misspecified models are considered acceptable on basis of a cut-off value of .90 for CFI, Mulaik et al (1989) suggested raising this value to 0.95. Marsh and Hau (1996) found that the use of conventional cut-off criteria may be appropriate in some situations but not in others. Hu and Bentler (1999) investigated the number of rejections of true and misspecified models at several cut-off values for fit indices. Their results led them to propose higher cut-off values than are commonly required for model selection. They suggest a cut-off value of .95 for the ML-based TLI, Δ_2 (BL89), RNI and CFI; a value close to .08 for SRMR and a value close to

.06 for RMSEA. These values seemed to result in lower Type-II error rates with acceptable Type-I error rates. Recently, several papers have appeared that suggest that fixed critical values can not be used in order to evaluate the fit of models (Beaducel and Wittmann 2005, Fan and Sivo 2005, Marsh and Hau 1996 and Marsh, Hau and Wen 2004, Yuan 2005). This does not mean that the procedure for evaluation of models has been changed in practice.
In this paper we want to show that the problem of the fit measures in the same as the problem of the X^2 test with respect to the sensitivity of the indices to other characteristics of the model than the size of the misspecification. This is also the reason why evaluation with fixed critical values can not work.

The design of the study

To assess the influence of a model's characteristics on the fit indices, we start with a population study. This enabled us to see how, in our study-design, the non-centrality parameter and power of the X^2 test statistic vary in relation to the characteristics of the model. Based on our results and the outcome of a previous study by Saris and Satorra (1988), we formulated hypotheses on the influence of the model's characteristics on the value of the fit indices. Next, we used Monte Carlo simulations to investigate the performance of fit indices under a variety of conditions related to the model's characteristics in order to evaluate whether our predictions were correct.

Model misspecifications

For our study on population data and our Monte Carlo experiment we used the same confirmatory-factor analysis model used by Saris and Satorra (1988) for all generated data

sets. The path diagram of this model is presented in Figure 1. In this model there are two latent factors (η_1 and η_2) and eight observed variables (y_1 to y_8). It is assumed that each latent variable (or construct) affects only four observed variables: η_1 affects y_1 to y_4, whereas η_2 affects y_5 to y_8. It is also assumed that the error terms are independent of each other and of the latent variables.

Figure 1 about here

Figure 1 The model tested for each data set

The model in Figure 1 is a special example of the more general confirmatory factor analysis model for which the following model covariance structure specification holds:

$$\Sigma = \Lambda\Psi\Lambda' + \Theta, \qquad (1)$$

where Σ is the variance-covariance matrix for the observed variables, Ψ is the variance-covariance matrix for the common factors, Θ is variance-covariance matrix for the unique factors or measurement errors, and Λ is the matrix of the loadings from the common factors on the observed variables.

The model in Figure 1 specifies several restrictions in the three matrices that feature in equation (1), as shown by the matrix expressions in (2):

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \\ 0 & \lambda_{72} \\ 0 & \lambda_{82} \end{bmatrix}$$

$$\Psi = \begin{bmatrix} 1 & \psi_{21} \\ \psi_{21} & 1 \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{44} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{77} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{88} \end{bmatrix} \quad (2)$$

The parameters presented by a symbol represent the free parameters that have to be estimated. All the restrictions (zeros) introduced in a model are potential miss-specifications. This means that in the model presented in Figure 1, three different kinds of misspecifications can be distinguished:

- One or more omitted effects of latent variables on observed variables. For example, it is possible that $\lambda_{51} \neq 0$. In this case we speak of a model with an indicator flaw (Figure 2).
- An omitted correlation between the error terms of observed variables of one construct: omitted within-correlated error. For example, $\theta_{21} \neq 0$ (Figure 3).
- An omitted correlation between the error terms of observed variables of two different latent variables: an omitted across-correlated error. For example, $\theta_{51} \neq 0$ (Figure 4).

Naturally, if one makes other restrictions with regard to the model in Figure 1, other misspecifications are also possible. In fact, there is one case where we fitted a model restricting the correlation among the two factors to be 1 (i.e., we fitted a one factor model with eight indicators) when in fact the true population value of this correlation was 0.9 (see Figure 5).

Figure 2, 3, 4 and 5 about here

Figure 2 The model used to generate the data with an indicator flaw

Figure 3 The model used to generate the data with a correlated error within a construct

Figure 4 The model used to generate the data with a correlated error across constructs

Figure 5 The model used to generate the data with a factor correlation of 0.9 which is analyzed assuming that the correlation is 1

The four kinds of misspecifications described above imply that four models have been used to generate the data. The model 2 - 5 are presented in figures 2 to 5. The purpose of the experiment was to show that the value of fit indices in the misspecified model varied depending on the different circumstances. So, we examined the behavior of fit indices for each misspecification, separately, to reduce the influence of the kind and size of the misspecification. After all, it is hard to determine when different misspecifications are of the same size, for example, is an indicator flaw of 0.1 the same size as an omitted error term of 0.1?

We did not vary the size of the sample because many researchers have already addressed the influence of sample size on the value of fit indices. Nor did we vary other aspects, such as the distribution of the variables, the kind of fitting function and so on, which

have been studied by other researchers (see for example Hu and Bentler, 1998). This was not the goal of our study.

In our Monte Carlo experiment, all data were generated from a multivariate normal distribution; a fixed sample size of n=1000 was used, while the Maximum Likelihood approach was also used in all the cases.

To demonstrate the effect of the size of parameters within a specific model context, we varied the size of the loadings and the size of the correlation between the constructs. The values of the parameters were chosen as typical values occurring in practice. Because we wanted to highlight the problem rather than to go into detail, we did not include a complete range of values in our study but selected the extremes. When there was an indicator flaw or an omitted correlated error (models 2 to 4), the sizes of the loadings were 0.9 or 0.5 and the correlation between the factors was either 0.3 or 0.7 respectively. Thus, four different combinations can be made for each model. In the case of a misspecified correlation between the constructs (model 5), only the loadings between the latent and the observed variables were varied. They could be 0.9, 0.7 or 0.5.

Misspecification size was 0.1 in the case of an indicator flaw and 0.15 in the case of an omitted correlated error. With regard to the correlation between the factors, this correlation was specified as 0.9 in the data generation. In the model estimation, however, this correlation was fixed as 1, assuming a one factor model.

Study of population data

On the basis of parameter values, one can calculate the population covariance matrices associated with the Models 2 to 5 (see figures 2 to 5). For each of the Models 2 to 4 (see figure 2 to 4), four population covariance matrices were calculated. For Model 5 (see Figure 5), we calculated three population covariance matrices. Next, we analyzed all these matrices

using the incorrect (misspecified) model shown in Figure 1. In case of model 2, we assumed in our analysis that λ_{51} was equal to zero, while in the data generation λ_{51} was set at 0.1. For model 3, θ_{21} was set as 0.15 for the computation of the population covariance matrices , while in the analysis this parameter was restricted to 0. With regard to model 4, the same applies to θ_{51}. In case of model 5, in the generation of the data the correlation among factors was assumed to be 0.9, while in the analysis it was restricted at the value 1. So the analyzed model differed from the true (population) models that had generated the data in just one dimensional misspecification.

These analyses provided values for the test statistic. The deviation from zero for these test statistics can only be explained by misspecification of the model. As Satorra and Saris (1985) have shown, the value of the LRT statistic in this population study is equal to the non-centrality parameter associated to the non-central chi-square distribution of the statistic for the incorrect model. Using the non-centrality parameter values obtained in this way, the degrees of freedom of the model and a chosen significance level the power of the test can be obtained from the tables for the non-central X^2 distribution.

Table 2 about here

Table 2 shows the non-centrality parameters obtained for the specific hypothesized model H_h and the baseline independence model H_b. It can clearly be seen that the non-centrality parameters, and as a consequence also the power of the LRT, vary according to the characteristics of the model. The derived power estimates for these models will be presented in tables 3, 4, 5 and 6 together with the results form the simulation study.

In general, the results found are in agreement with the study of Saris and Satorra (1988). The size of the loadings has a strong effect on the power of the test. With high

loadings the probability that the error will be detected is substantially larger than with low loadings. This holds for all kinds of errors in this study.

The effect of the size of the correlation between the factors varies according to the type of error. With an indicator flaw, the power becomes smaller as correlation between the factors increases. On the other hand, where there is an omitted correlated error term within constructs, the power increases when the correlation increases, at least when the loadings are high. With low loadings there is hardly any effect. For an omitted correlated error across constructs, the power is hardly affected by the size of the correlations.

On the basis of the results of this study of population data and the fact that the stand-alone fit index RMSEA is a function of the non-centrality parameter, we expect the same effects of model characteristics that we found for the X^2 test.

With respect to the Incremental Fit Indices (INFI), the power of rejecting the model will depend on how severely misspecified the "base-line" model underlying the specific INFI used is. The general formulae of the incremental fit indices is:

$$INFI = (X^2_b - X^2_h)/ X^2_b = 1- X^2_h / X^2_b \qquad (3)$$

where X^2_h and X^2_b stands for the chi-square goodness-of-fit of the analyzed and baseline model respectively. Rejection of the analyzed model will occur when the INFI of (3) is smaller than a typical threshold value, usually .95, i.e. when:

$$(X^2_b - X^2_h)/ X^2_b < .95;$$

that is, when:

$$X^2_h > .05\ X^2_b \qquad (4)$$

Clearly, when the base-line model is highly misspecified, the value $.05\ X^2_b$ will be a very large number and the inequality (4) will be hard to hold, even if the analyzed model happens to suffer from relevant substantial misspecification, i.e. the power of the test will be very low

in this case. However, when the baseline model is almost correct, the right hand side of the equation (4) will be small, and therefore even extremely small misspecifications of the analyzed model will lead to its rejection by the INFI approach, that is, the power will be high. It follows that with very "noisy" data (i.e., data that has a lot of measurement error and lack of substantial structure), the baseline model specifying the independence of all the variables will be close to the true model, so small deviations from the true model may lead to rejection of the analyzed model.

Furthermore, in cases of highly structured data, the baseline model of independence among variables will be greatly misspecified and will thus have an associated high value for .05 $X^2_{b.}$ This means that in such cases there will be a lack of power to detect misspecification of the analyzed model. This explains the differential behavior of the INFI when compared with the standard chi-square goodness-of-fit model approach. Due to the influence of the baseline model, the power of the incremental fit indices will often be opposite to the power of the X^2 and stand alone tests. This can be seen in the tables below.

It is clear that incremental fit measures can attract the same kind of criticism that is directed at the standard chi-square test, which says that certain misspecifications always lead to rejection of the model (eg. large samples, high power) but not in other cases (small samples, low power). For unstructured data with high background noise, the power of a test based on incremental fit indices will be high whereas for extremely structured data, the power of a test will be low. This means that these fit measures can not be used in a simple way with a fixed critical value if the power of the fit statistic is not taken into account in each specific case.

Note that the predictions mentioned above don't apply to the SRMR because this fit measure is not a function of LRT statistic. We still included this statistic because we wanted to see how it behaves in the different models.

Note also that we can view the GFI and AGFI as incremental fit indices, where the baseline model is the model that sets variances and covariances equal to zero (and using a GLS estimator). Thus, we expect the behavior of an INFI (described above) for GFI and AGFI even though the baseline model is now different than the independence model and the fit function is not ML but GLS.

Finally, it is simple to predict that the PNFI will always lead to rejection of the models in our models. This is so because for the Model H_h, df=19 whereas df= 28 for the Model $H_{b,}$ while PNFI is equal to (df_h/df_b) NFI. Therefore, even if NFI was maximal, that is to say 1, the value of the PNFI would be 19/28 = .67 which is always lower than the critical value (.80) required for the PNFI. Although we included the PNFI in our Monte Carlo experiment, we have omitted this index because using this index will always lead to rejection of these models under all circumstances. Rejections arise due to the specific value of the ratio of the degrees of freedom and not due to incorrectness of the model.

The Monte Carlo experiment

In our study, we included two "stand-alone fit indices": RMSEA, SRMR, and several "incremental fit indices": GFI and AGFI, NFI , NNFI, RFI, IFI, PFNI and CFI . For each of these fit indices, a cut-off value or threshold was specified, above or below which a model is declared to be acceptable (see Table 1).

For each of the population covariance matrices (15 in total) that were computed on the basis of the values of the parameters (see the paragraph on population data), we generated 300 samples of size n=1000 cases from the associated multivariate normal distribution. For each sample, models H_h and H_b were fitted and the above mentioned fit measures were

calculated. This means that for each population covariance matrix and for each fit index, an empirical Monte Carlo distribution (300 values) of the index was obtained.

Next, on the basis of the conventional cut-off criteria for the fit indices, we examined the model rejection rates associated with each of the different fit measures: that is, the power of the test under the different circumstances (i.e. how often the wrong model was rejected). These rejection rates, the mean of the Monte Carlo distribution, and the population value of the index, are presented in tables 3 to 6. The population value of the fit index is the value of the fit index when the population covariance matrix is used as input matrix in the estimation procedure.

The tables also show the non-centrality parameters for the Models H_h. The power of the X^2 test was calculated on the basis of these values using the non-central X^2 distribution for an α level of 5%. Furthermore, the non-centrality parameters for the Models H_b are given. The power of the INFI statistic in (3) was approximated on the basis of the non-centrality parameter associated to the models H_h and H_b. We consider the approximation conditional to the value of X^2_b obtained when fitting H_b, and using a non-central chi-square approximation for X^2_h. We added the power of the X^2 test and the incremental fit index INFI to these empirical results to facilitate comparisons.

Results of Monte Carlo experiment

The results are summarized in Tables 3 to 6. We will show that for most fit indices the results are in line with what is observed for the power of either the X^2 test or the incremental fit index INFI. This means that these indices also depend on the characteristics of the model. We say that these results will be 'in line' because in most cases the fit indices are nonlinear functions of the value of the fitting function, the non-centrality parameter or the X^2 test statistic. This means that it is not feasible to expect a perfect relationship.

The results for data with an indicator flaw

Table 3 gives the results of estimating the model of equation (2) (Figure 1) on data generated with the model with an indicator flaw across the constructs presented in Figure 2.

Table 3 about here

The most striking aspect shown in Table 3 is that for this type of error and the population values of the parameters, most fit indices are incapable of detecting the misspecification in the model. Yet with three fit measures, the results clearly show the influence of the size of the loadings and of the size of the correlations between factors on the percentage of cases in which the model is rejected. When the RMSEA is used as an index of fit, the percentage of model rejections is higher with high loadings than with low loadings. In addition, when the loadings are .9, the model is rejected more often when the correlation between factors is lower. This result is in line with our expectations on the basis of the power of the X^2 test.

In the case of the RFI, the opposite occurs for the different values of the loadings. In the case of the NFI, the model characteristics affect the value of the index in the same way as with the RFI, although the effect is much smaller. This result is in line with the predictions of the power of the INFI index presented at the bottom of the table.

Although the percentage of model rejections on the basis of the other fit indices seems not to be influenced by the model characteristics, this does not mean that the value of the fit indices does not also vary. However, the variation is small and the mean values are always greater than the conventional cut-off criteria. For example, the mean value of the AGFI varies between .97 and .99. So the power can not vary from one model context to the next.

The results for data with a correlated error within a construct

Table 4 presents the results of estimating the model of equation (2) (Figure 1) on data generated with a correlated error term within a construct (Figure 3).

Table 4 about here

The influence of the size of the loadings and of the factor correlation on the value of the fit indices is clear if one looks at the results in Table 4, although the effect does not operate in the same direction and is not equally strong for all the fit indices.

In the case of the RMSEA, the percentage of model rejections decreases as the loadings decrease. These findings are in line with our expectations on the basis of the power of the LRT. In addition, with the RMSEA, the size of the correlation between the constructs seems to be of no real influence.

With most of the other indices, the effect operates in the opposite direction. In the case of the RFI, the percentage of model rejections increases when the loadings decrease and the percentage rejections seems to increase with the size of the factor-correlation. This also applies to the NFI and the NNFI although the effect is much smaller. These results are again in line with the power predictions for the INFI index presented at the bottom of the table.

The CFI, IFI and the SRMR seem unable to detect misspecification whatever the circumstances, since the percentage of model rejections is nearly always 0%. However, as already shown in Table 3, this does not imply that the value of this index is constant. For example, the mean value of the IFI and CFI range from .93 to .98; the mean value of the SRMR varies between .02 and .04. However, because the values of the IFI and CFI are nearly

always greater and the value of the SRMR smaller than the conventional cut-off criteria, the percentage of model rejections is always around 0%.

The results for the GFI and AGFI are particularly striking. Contrary to what we expected, here we also see high power for the highest loadings, with the influence of factor correlation being considerable in this case. Such discrepancy, however, between GFI and AGFI and the other INFI's could just be another instance of the severe effect of the estimation method on the INFI fit indexes as described in Sugawara and MacCallum (1993) (see also Fan, Thompson and Wang, 1999). As mentioned before, GFI and AGFI can be viewed as an INFI with respect to a baseline model that sets the covariance matrix equal to zero (a "zero model"), which is a highly misspecified model when the loadings are high. However, in computing GFI and AGFI, the X^2 for such a "zero-model" is obtained using a GLS fitting function, in contrast to the ML fitting function used to evaluate the independence model of the other INFI's. Again, in this case we see a value of the goodness of fit indexes being affected by other factors than the correctness or not of the model.

The results for data with a correlated error across constructs

Table 5 presents the results of estimating the model of equation (2), presented in Figure 1, on data generated with a correlated error term across constructs (Figure 4).

Figure 4 and Table 5 about here

Figure 4. The model used to generate the data with a correlated error across constructs

In the case of a correlated error across constructs, the influence of model characteristics is also obvious (Table 5). Here, the effect of the size of the loadings on the percentage of model

rejections works in the same direction for all indices: the smaller the loadings, the smaller the percentage of model rejections, which is in line with the expected power for the stand-alone and incremental fit indices. Furthermore, when using the NFI, NNFI, CFI or IFI, we see that with loadings of .5 the model is more frequently rejected with a low factor correlation than when the correlation between factors is high. This is also in line with the predictions for the incremental fit indices.

This result is not obtained for the GFI and AGFI although the same pattern can be seen in the mean values of these indices but in absolute value the values of the indices are too high to lead to differences in power for the different models. Again, in the case of the SRMR the model is accepted in all circumstances, yet the mean values of the fit indices differ again with the characteristics of the model. Although the variation in the mean values is not large in absolute terms in relative terms it is. Furthermore, when one uses the RFI to determine the fit of the model in the case of a correlated error across constructs, the model is nearly always rejected under various characteristics. Even though we see some variation in the mean values of this fit index, in this case this variation is not large in contrast to the outcome shown in the previous table.

Results for data with a factor correlation of .9

Table 6 shows the outcomes resulting from estimating the model of equation (2), presented in Figure 1, on data generated with a misspecified correlation between the latent variables shown in Figure 5.

Figure 5 and Table 6 about here

Figure 5. The model used to generate the data with a factor correlation of .9 which is analyzed assuming that the correlation is 1.

If one assumes a one-factor model when this is not the case (i.e., the data come from a two factor model with correlation .9), the model characteristics (i.e., values of the parameters) influence the decisions taken on the basis of the fit indices. With loadings of .9, the model is nearly always rejected. When the loadings are .7 or .5, the model is practically always accepted. Exceptions are the RMSEA and the RFI. In the case of the RMSEA the model-rejections is 45 percent when the loadings are 0.7 and decreases to 0%with loadings of 0.5. This is in line with our predictions but the effect is much stronger than expected. With the RFI the opposite occurs: with loadings of 0.7 the model is almost never rejected, when the loadings are 0.5 the model is rejected in 20% of the cases. This result is also in line with the expectations but much stronger than expected. When the model is evaluated on the basis of the SRMR, again it is never rejected. Still, the mean value of the SRMR varies with the model characteristics.

Conclusions

We come to the following conclusions on the effect of model characteristics. The percentage of rejections on the basis of fit indices changes substantially when the loadings and the correlations between the constructs are varied even though the type and the size of the error are kept constant. The model characteristics effect does not work in the same direction and is not of equal magnitude for all fit indices. In addition, the effect varies with the kind of misspecification. The RMSEA, used so frequently nowadays, seems to be particularly sensitive to model characteristics.

In the case of an indicator flaw, many of the fit indices in this study seem to be unaffected by the characteristics of this model context, at least in terms of the percentage of model rejections. However, when we look at the mean value of the fit indices, we see that they vary with the model characteristics, even though this variation is not very large. It would not be difficult to create more variation, for example by increasing the size of the indicator flaw to .2.

The percentage of model-rejections does not vary at all when the SRMR is used to evaluate the fit of the model. However, the mean value of this indicator varies with the model characteristics. This means that the SRMR is also influenced by the characteristics of the model, just as the other fit indices. The fact that, in almost all cases, the value of the SRMR is lower than .05 (the standard cut-off value) is probably due to the way this index is calculated. The errors in our experiments are not very large and the SRMR is the sum of the squared residuals divided by the number of distinct variances and covariances. The latter part of the formula can reduce the value of the SRMR considerably.

As we have mentioned above the results for the PNFI were omitted in the tables. In the case of the PNFI, the percentage of cases in which a wrong model was accepted seemed not to be influenced by the characteristics of the model. In all cases, the wrong model was rejected. One could assume that this means that the PNFI is a better fit-index than the others. Yet when one looks at the formula of the PNFI and the value of the conventional cut-off criterion, it becomes clear that the 100% rejection is due to the ratio of the degrees of freedom and has nothing to due with the lack of fit of the model.

One might say that the fluctuating percentage of model rejections is due to the conventional cut-off criteria value, but this is not the case. As we have seen in this paper, the population value itself of the indices is substantially affected by the model characteristics.

Fixing the cut-off criteria at other values is no solution for the dependency of the value of the fit indices on the model characteristics.

Discussion

Saris and Satorra (1988) have shown that the X^2 statistic used to test structural equation models is not only affected by the misspecifications in the model but also by other characteristics of the model. Since most of the current fit indices are based in some way on the fitting function or on the LRT, we expected these indices to show the same kind of problems as the LRT. For two fundamental but simple models this phenomenon was shown in Saris, Satorra and Van der Veld (2009). In this paper their study is extended to more models and fit indices. Keeping the kind and the size of the error constant, we showed that the percentage of model rejections varied with the characteristics of the model. It was also obvious that the kind of error influences the decision one would take on the basis of the fit indices.

One could assume that it would be justified to use these indices in the standard way to compare the fit of similar models, but this is also not the case. After all, as is obvious from our Monte Carlo experiment, the fit indices have a different sensitivity to different kinds of errors. This means that we cannot use fit indices in the standard way to evaluate the fit of the model. Of course one could look at the size of the fit index or the X^2. If the index is very large this might be an indication for misspecifications in the model. However it is possible that these large values are due to very small misspecification for which the test statistic or fit index is very sensitive. If the fit index of X^2 is very small one can assume that the model does not contain misspecifications. However, as we have shown above it can also be that there are large misspecifications for which the fit statistic is not sensitive.

In 1987 Saris, Satorra and Sorbom suggested to use the Expected Parameter Change (EPC) to detect misspecifications in SEM. This solution is re-iterated and illustrated again in Saris and Satorra (forthcoming). For the moment we think that the use of the EPC is the best option for the detection of misspecification because it gives a direct estimate of the possible misspecifications in the model. In the paper of Saris, Satorra and Van der Veld (2009) the authors adjust this approach by suggesting taking into account the test statistic (MI) and the power of the test. For further details we refer to this paper.

References

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.

Beauducel A. and W.W.Wittmann (2006) Simulation study on fit indices in CFA based on data with slightly distorted simple structure. *In Structural equation modelling, 12, 41-75*

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

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. & Mooijaart, A. (1989). Choice of structural model via parsimony: a rationale based on precision. *Psychological Bulletin, 106,* 315-317.

Bollen K.A. (1986) Sample size and Bentler and Bonnet´s nonnormed fit index, Psychimetrika, 51, 357-377.

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

Browne, M.W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21,* 230-258.

Cudeck, R. & Henly, S.J. (1991). Model selection in covariance structure analysis and the “problem” of sample size: a clarification. *Psychological Bulletin, 109,* 512-519.

Cudeck, R. & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research,* 18, 147-167.

Fan X. and S.A. Sivo (2006) Sensitivity of fit indices to misspecified structural or measurement model components: Rational of two.Index strategy revisited. . *In Structural equation modelling, 12, 343-367*

Fan, X., Thompson, B. and Wang, L.(1999). Effects of sample size, estimation methods and model specification on structural equation modeling fit indexes. *Structural equation modeling, 6*.

Gerbing, D.W & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen &

Hu, L. and Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model specification. *Psychological Methods, 3,* 424-453.

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

Hu, L., Bentler, P.M. and Kano, Y (1992). Can test statistics in covariance structure 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: Sage.

Jöreskog K.G. & Sörbom D (1989) Lisrel 7 A guide to the program and applications. Chicago, SPSS publications.

Long, J.S. (1983). Covariance Structure Models: An introduction to LISREL. Newbury Park, CA: Sage.

Marsh, H.W. & Hau, K.T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education*, 64, 364-390.

Marsh, H.W. & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. Psychological Bulletin, 97, 562-582.

Marsh, H.W. (1995). $\Delta 2$ and X^2I2 Fit indices for structural equation models: A brief note of clarification. *Structural equation modeling, 2,* 246-254.

Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Marsh, H.W., Balla, J.R. and Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In: George A. Marcoulides and Randall E. Schumacker (Eds.), *Advanced Structural Equation Modeling: Issues and techniques* (pp. 315-353). New Jersey; Lawrence Erlbaum Associates, Inc.

Marsh H.W., K.T.Hau and Z.Wen (2004) In search of golden tules: Comment on hypothesis.testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and Bentles´s (1999) findings. . *In Structural equation modelling, 11, 320-341*

McDonald, R.P. & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247-255.

McDonald, R.P. (1989) An index of goodness-of-fit based on noncentrality. Journal of classification , 6, 97-103.

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

Saris, W.E., den Ronden, J. and Satorra, A. (1984) Testing structural equation models. In P. Cuttance and R. Ecob (Eds.), *Structural modeling by example* (pp. 202-220). New York: Cambridge University Press.

Saris, W.E., Satorra, A. & Sörbom, D. (1987) The detection and correction of specification errors in structural equation models. In: *Sociological Methodology, 17,* 105-129.

Saris, W.E. & Satorra, A. (1988) Characteristics of structural equation models which affect the power of the Likelihood Ratio Test. In: Sociometric Research, vol. 2. Eds. W.E. Saris and I.N. Gallhofer. London, Macmillan.

Saris W.E., Satorra A. and W.van der Veld (2009) Testing Structural Equation Models or Detection of Misspecifications?', Structural Equation Modeling:, 16: 4, 561 – 582.

Satorra, A. & Bentler, P.M. (1994) Corrections to test statistics and standard errors in covariance structure analysis. In A. v. Eye and C.C. Clogg (eds.), *Latent variable analysis. Applications for developmental research.* Thousand Oaks, Sage

Satorra, A. & Saris, W.E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83-90.

Sobel, M.E. & Bohrnstedt, G.W. (1985). Use of null models in covariance structure models. In: N.B. Tuma (ed.), *Sociological methodology, 15,*152-178.

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

Steiger J. & Lind J.M. (1980) Statistically based tests for a number of common factpors. Paper presented at the Psychometrika Society Meeting, Iowa City.

Sugawara, H.M. , and MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. Applied Psychological Measurement, 17, 365-377

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. 136-162). Newbury Park, CA: Sage.

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

Weng, L.J. & Cheng, C.P. (1997). Why might relative fit indices differ between estimators. *Structural equation modeling,*

Williams, L.J. & Holahan, P.J. (1994). Parsimony-based fit indices for multiple-indicator models: Do they work? *Structural equation Modeling, 1*, 161-189.

TABLE 1

Fit Indices included in the study

Fit index	Formula[1)]	Reference	Cut-off value
GFI	$GFI = 1 - \left(\frac{\chi^2_h}{\chi^2_u} \right)$	Jöreskog and Sörbom (1989)	.95
AGFI	$AGFI = 1 - \left[\frac{p(p+1)}{2df_h} \right] [1 - GFI]$	Jöreskog and Sörbom (1989)	.9
SRMR	$SRMR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=1}^{i} [(s_{ij} - \hat{\sigma}_{ij})/s_{ii}s_{jj}]^2}{p(p+1)/2}}$	Jöreskog and Sörbom (1989)	.05
NFI or BBI	$BBI = \frac{(\chi^2_b - \chi^2_h)}{\chi^2_b}$	Bentler and Bonett (1980)	.95
RFI or BL86	$RFI = \frac{\left[\left(\frac{\chi^2_b}{df_b} \right) - \left(\frac{\chi^2_h}{df_h} \right) \right]}{\left(\frac{\chi^2_b}{df_b} \right)}$	Bollen (1986)	.95
NNFI or TLI	$TLI = \frac{\left(\chi^2_b / df_b - \chi^2_h / df_h \right)}{\chi^2_b / df_b - 1.0}$	Tucker and Lewis (1973)	.95
IFI (Δ_2)	$\Delta_2 = \frac{(\chi^2_b - \chi^2_h)}{(\chi^2_b - d_h)}$	Bollen (1989)/ Marsh et al. (1988)	.95
PNFI	$PNFI = NFI\left(\frac{df_h}{df_b} \right)$	James, Mulaik and Brett (1982)	.8

Table 1 *Fit Indices included in the study - continued*

Fit index	Formula	Reference	Cut-off value
CFI	$\hat{\Delta} = CFI = \frac{\hat{\lambda}_b - \hat{\lambda}_h}{\hat{\lambda}_b}$	Bentler (1990)	.95
RMSEA	$RM\hat{S}EA = \sqrt{\frac{\hat{F}_0}{df_h}} = \sqrt{Max\left\{\left(\frac{\hat{F}_h}{df_h} - \frac{1}{n}\right), 0\right\}}$	Steiger (1990), Steiger and Lind (1980)	.05

[1] Meaning of subscripts and symbols:

- F is the fitting function, X^2 = n*F
- n=N-1, N is the sample size
- 'h' refers to the hypothesized model
- 'u' refers to the ultimate null model in which all estimations are fixed at zero
- 'b' refers to the baseline model, which is usually the null model in which no common factors for the input measures and no covariances among these measures are specified. This is usually done by setting all of the covariances among the measures at zero while allowing their variances to be estimated as free parameters.
- p: number of observed variables
- t: number of free parameters
- λ is the non-centrality parameter
- $\hat{\lambda}_h = \max(\tilde{\lambda}_h, 0)$; $\tilde{\lambda}_h = nF_h - df_h$
- $\hat{\lambda}_0 = \max(\tilde{\lambda}_0, 0)$; $\tilde{\lambda}_0 = nF_0 - df_0$

TABLE 2

Non-centrality parameter (ncp) and power of the 5% level LRT for the four types of misspecified models.

kind of misspecification	loadings	correlation	ncp model H_h	ncp Base model H_b
indicator flaw	.9	.7	32.07	8824
	.9	.3	43.33	7879
	.5	.7	1.7	835
	.5	.3	6.5	676
correlated error within a construct	.9	.7	221.76	9271
	.9	.3	169.84	8781
	.5	.7	8.55	853
	.5	.3	10.88	720
correlated error across constructs	.9	.7	481.93	8585
	.9	.3	457.03	8037
	.5	.7	28.51	821
	.5	.3	29.26	669
misspecified factor correlations	.9	.9	521.0	8780
	.7	.9	50.7	2937
	.5	.9	4.7	929

TABLE 3

For indicator flaw data and misspecification error of .10, the table shows the percentage of model rejection (dark stripe), and the mean and population values of the fit index.

	loadings .9		loadings .5 *)	
	correlation .7	correlation .3	correlation .7	correlation .3
RMSEA				
% rejection	22	45	0	0
mean value	.0402	.0479	.0090	.0122
population value	.026	.036	0*)	0*)
SRMR				
% acceptance	0	0	0	0
mean value	.0159	.0259	.0186	.0224
population value	.014	.024	0*)	0*)
GFI				
% rejection	0	0	0	0
mean value	.9870	.9844	.9949	.9943
population value	.99	.99	1*)	1*)
AGFI				
% rejection	0	0	0	0
mean value	.9753	.9704	.9903	.9893
population value	.98	.98	1*)	1*)
NFI				
% rejection	0	0	1	7
mean value	.9942	.9919	.9765	.9666
population value	1.00	.99	1*)	1*)
NNFI				
% rejection	0	0	0	0
mean value	.9917	.9916	.9974	.9914
population value	1.00	1.00	1*)	1*)
CFI				
% rejection	0	0	0	0
mean value	.9964	.9943	.9961	.9926
population value	1.00	1.00	1*)	1*)
IFI				
% rejection	0	0	0	0
mean value	.9964	.9943	.9982	.9942
population value	1.00	1.00	1*)	1*)
RFI				
% rejection	0	0	9	43
mean value	.9915	.9831	.9654	.9508
Population value	.99	.99	1*)	1*)
Power analysis				
ncp H_h (df=19)	32	43	1.7	6.4
Power of the X^2 test	.96	.99	.09	.25
ncp H_b	8824	7896	835	676
Power INFI statistic	0	0	0.004	.12

*) The program LISREL does not give the values of the fit indices in case of perfect fit. The values reported are just presumed.

TABLE 4
For correlated error within a construct and misspecification error of .15, the table shows the percentage of model rejection, and the mean and population values of the fit index.

	loadings .9		loadings .5 *)	
	correlation .7	correlation .3	correlation .7	correlation .3
RMSEA				
% rejection	100	100	1	0
mean value	.1077	.0953	.0293	.0198
population value	.10	.089	.0	.0
SRMR				
% rejection	2	0	0	0
mean value	.0405	.0270	.0262	.0228
population value	.04	.025	.019	.015
GFI				
% rejection	93	18	0	6
mean value	.9395	.9551	.9905	.9330
population value	.94	.96	1.00	1.00
AGFI				
% rejection	84	9	0	6
mean value	.8853	.9150	.9820	.9268
population value	.89	.92	.99	.99
NFI				
% rejection	0	0	26	25
mean value	.9742	.9781	.9577	.9029
population value	.98	.98	.98	.98
NNFI				
% rejection	1	0	15	12
mean value	.9648	.9709	.9684	.9198
population value	.97	.97	1.00	1.02
CFI				
% rejection	0	0	3	7
mean value	.9761	.9803	.9785	.9258
population value	.98	.98	1.00	1.00
IFI				
% rejection	0	0	3	7
mean value	.9761	.9803	.9788	.9265
population value	.98	.98	1.00	1.01
RFI				
% rejection	2	0	73	65
mean value	.9619	.9678	.9376	.8853
population value	.96	.97	.97	.98
Power analysis				
ncp H_h	223	170	18.1	10.6
Power of X^2 statistic	1	1	.73	.44
ncp. H_b	9271	8781	853	720
Power of INFI statistic	0.00	0.00	.24	.18

TABLE 5
For a correlated error across construct and misspecification error of .15, the table shows the percentage of model rejection, and the mean and population values of the fit index.

	loadings .9		loadings .5 *)	
	correlation .7	correlation .3	correlation .7	correlation .3
RMSEA				
% rejection	100	100	9	6
mean value	.1590	.1551	.0378	.0387
population value	.16	.15	.022	.023
SRMR				
% rejection	0	0	0	0
mean value	.0244	.0262	.0278	.0306
population value	.023	.024	.021	.023
GFI				
% rejection	100	100	0	0
mean value	.9069	.9118	.9882	.9880
population value	.91	.92	.99	.99
AGFI				
% rejection	100	100	0	0
mean value	.8236	.8329	.9776	.9774
population value	.83	.84	.99	.99
NFI				
% rejection	96	97	62	88
mean value	.9419	.9409	.9441	.9305
population value	.94	.94	.97	.96
NNFI				
% rejection	100	100	49	72
mean value	.9174	.9161	.9489	.9351
population value	.92	.92	.98	.98
CFI				
% rejection	91	93	15	37
mean value	.9440	.9431	.9653	.9560
population value	.95	.95	.99	.98
IFI				
% rejection	90	93	14	24
mean value	.9440	.9432	.9657	.9565
population value	.95	.95	.99	.98
RFI				
% rejection	100	100	94	100
mean value	.9144	.9130	.9177	.8977
population value	.92	.92	.95	.94
Power analysis				
ncp H_h	482	457	28.6	29.3
Power of the X^2 test	1	1	.93	.94
ncp H_b	8585	8037	821	669
Power of INFI	.95	.96	.64	.86

TABLE 6
For factor correlation data, misspecification error of .1, the table shows the percentage of model rejection, and the mean and population values of the fit index.

	loadings		
	.9	.7	.5
RMSEA			
% rejection	100	45	0
mean value	.1624	.0489	.0166
population value	,16	,038	,0
SRMR			
% rejection	0	0	0
mean value	,0358	,0247	,0209
population value	,035	,021	,011
GFI			
% rejection	100	0	0
mean value	,8303	,9812	,9931
population value	,83	,99	1,00
AGFI			
% rejection	100	0	0
mean value	,6945	,9661	,9877
population value	,70	,97	1,00
NFI			
% rejection	97	0	2
mean value	,9375	,9767	,9709
population value	,94	,98	,99
NNFI			
% rejection	100	0	0
mean value	,9154	,9766	,9889
population value	,92	,99	1,02
CFI			
% rejection	93	0	0
mean value	,9396	,9833	,9914
population value	,94	,99	1,00
IFI			
% rejection	93	0	0
mean value	,9396	,9833	,9922
population value	,94	,99	1,01
RFI			
% rejection	100	1	20
mean value	,9125	,9674	,9593
population value	,92	,98	,99
Power analysis			
ncp H_h	521	50.7	4.7
Power of the X^2 test	1	.99	.19
ncp H_b	8780	2937	929
Power of INFI	.99	0.000	.004

Figure Captions

FIGURE 1. The model tested for each data set

FIGURE 2. The model used to generate the data with an indicator flaw.

FIGURE 3. The model used to generate the data with a correlated error within a construct

FIGURE 4. The model used to generate the data with a correlated error across constructs

FIGURE 5. The model used to generate the data with a factor correlation of .9 which is analyzed assuming that the correlation is 1

Figure 1

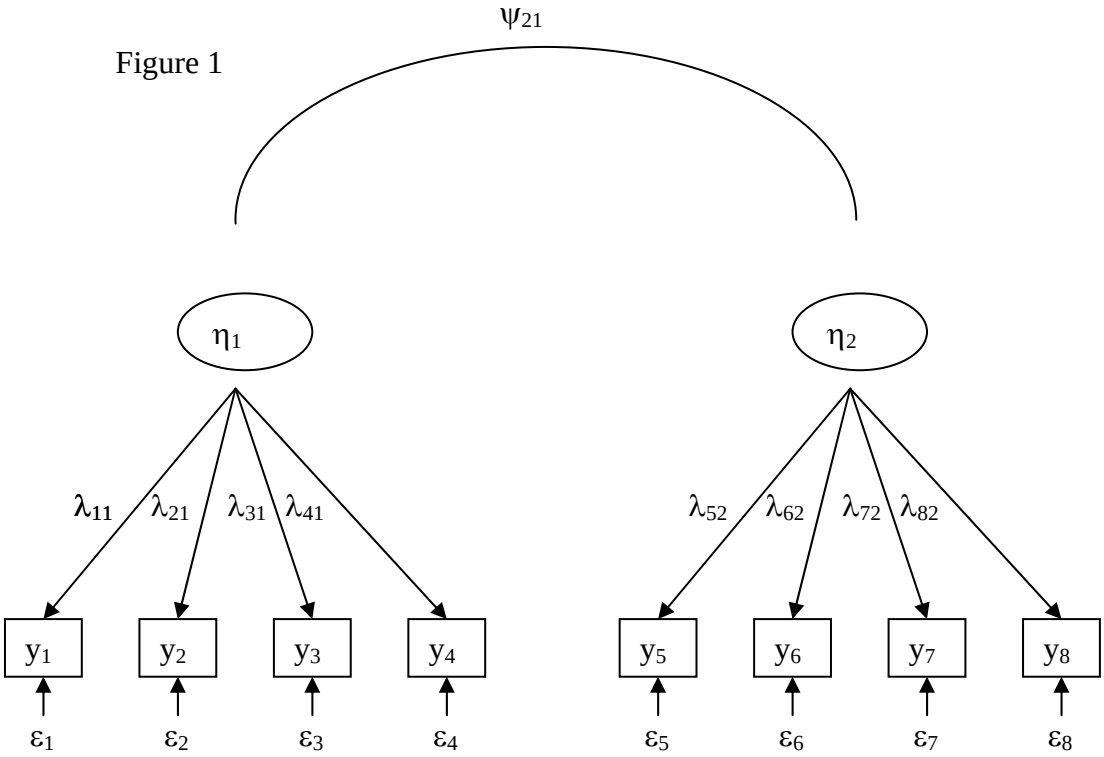

ψ21
η1
η2
λ11
λ21
λ31
λ41
λ52
λ62
λ72
λ82
y1
y2
y3
y4
y5
y6
y7
y8
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8

Figure 2

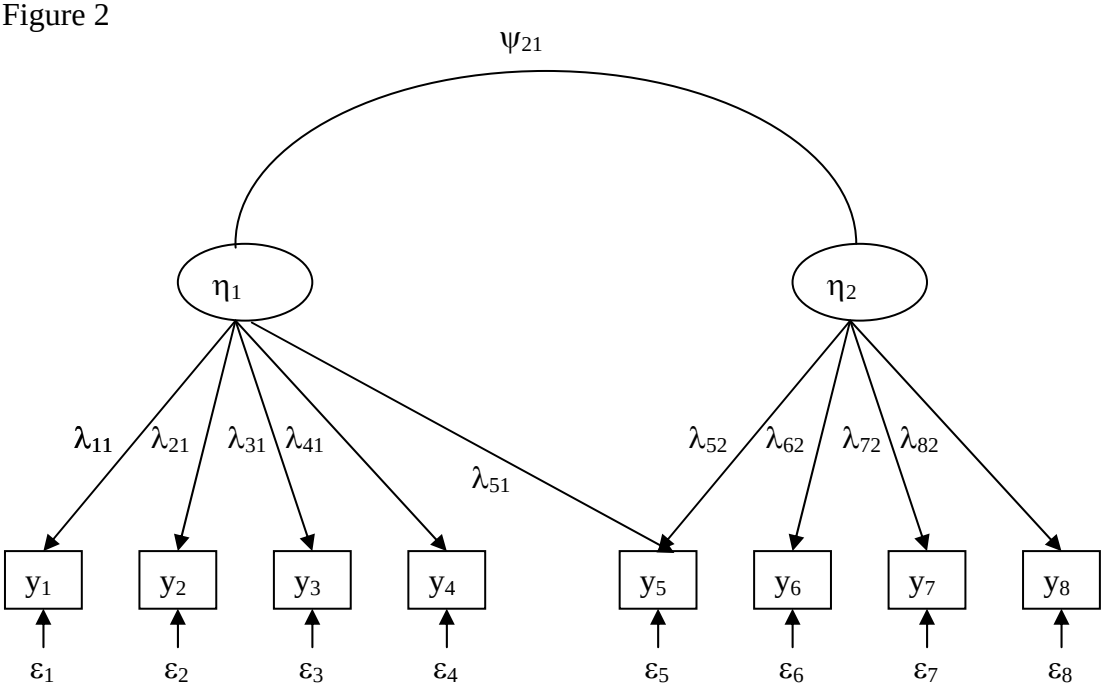

ψ21
η1
η2
λ11
λ21
λ31
λ41
λ51
λ52
λ62
λ72
λ82
y1
y2
y3
y4
y5
y6
y7
y8
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8

Figure 3

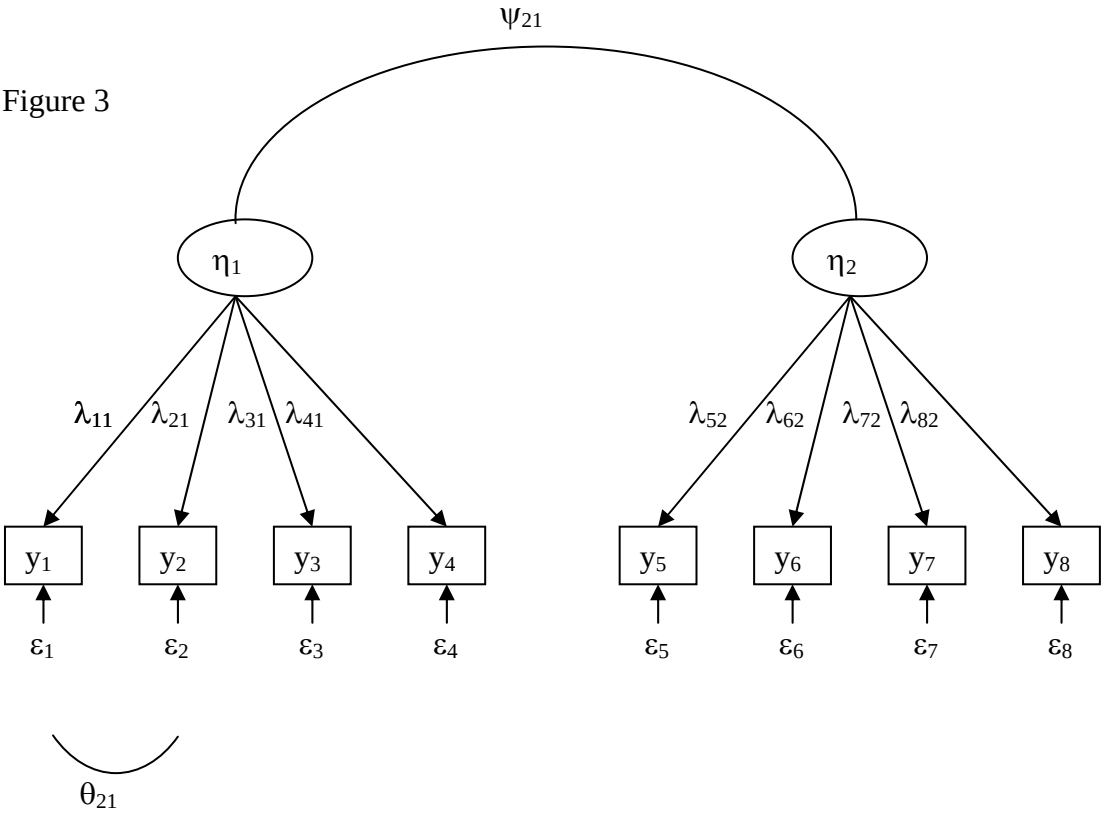

ψ21
η1
η2
λ11
λ21
λ31
λ41
λ52
λ62
λ72
λ82
y1
y2
y3
y4
y5
y6
y7
y8
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8
θ21

Figure 4

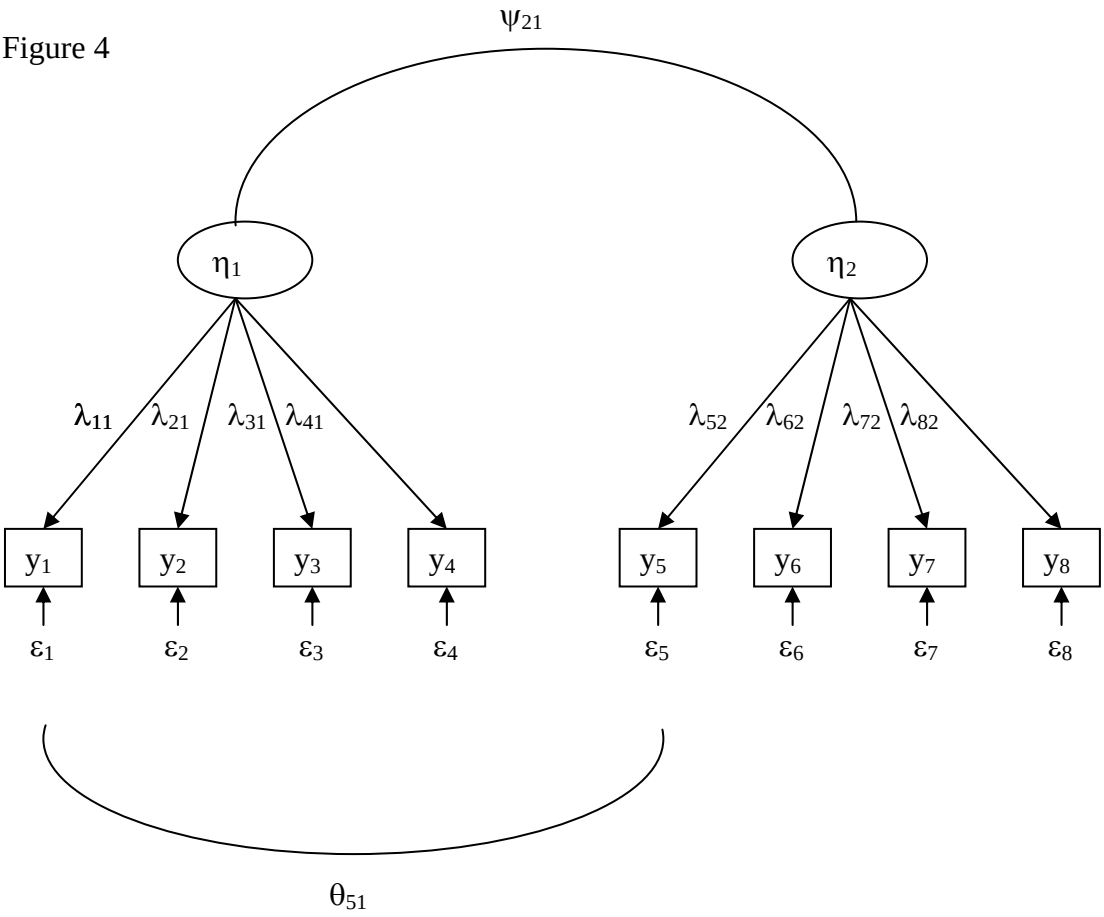
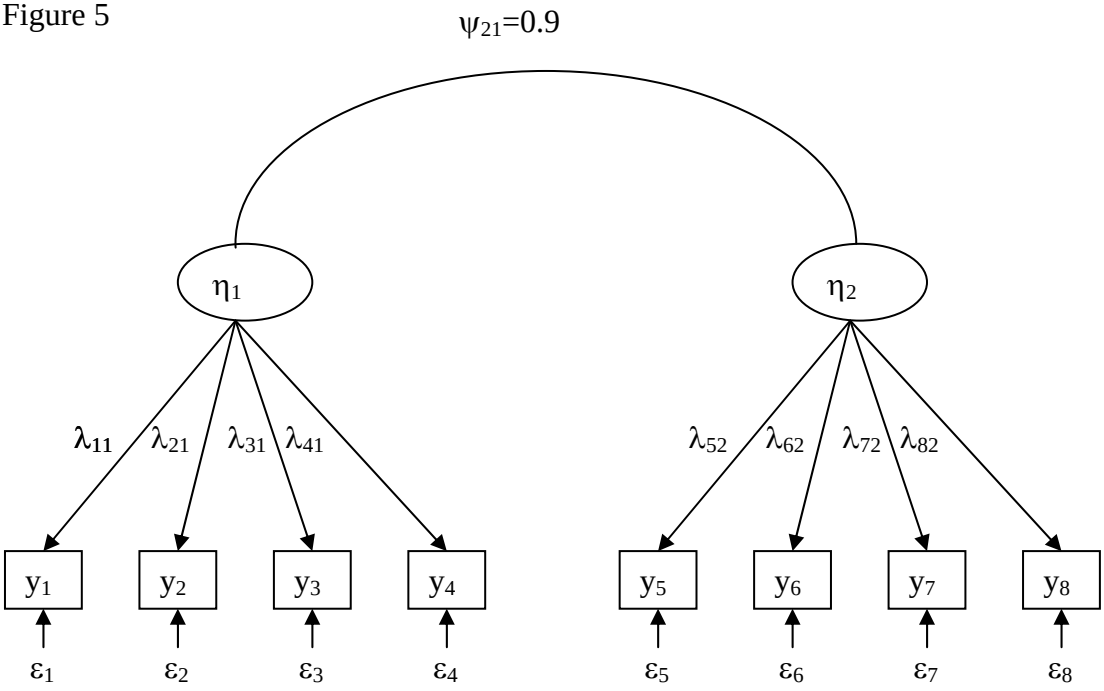

ψ21
η1
η2
λ11
λ21
λ31
λ41
λ52
λ62
λ72
λ82
y1
y2
y3
y4
y5
y6
y7
y8
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8
θ51

Figure 5


ψ21=0.9
η1
η2
λ11
λ21
λ31
λ41
λ52
λ62
λ72
λ82
y1
y2
y3
y4
y5
y6
y7
y8
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8