Testing Structural Equation Models or Detection of Misspecifications?[1]

Abstract

Assessing the correctness of a structural equation model is essential to avoid drawing wrong conclusions from empirical research. In the past, the CHI2 test was recommended for assessing the correctness of the model but this test has been criticized because of its sensitivity to the sample size. As a reaction, an abundance of fit indices has been developed. The result of these developments is that SEM packages are now producing a large list of fit measures. One would think that this progression has led to a clear understanding of evaluating models with respect to model misspecifications.

However, in this paper we question the validity of approaches for model evaluation based on overall goodness of fit indices. The argument against the use of fit indices is that they do not provide an adequate indication of the "size" of the model's misspecification. That is, they vary dramatically with the values of incidental parameters that are unrelated with the misspecification in the model. This is illustrated using simple but fundamental models. As an alternative method of model evaluation, we suggest using the Expected Parameter Change (EPC) in combination with the Modification Index (MI) and the power of the MI test.

Keywords: Structural Equation Models (SEM), Likelihood Ratio Test (LRT), Chi-square Goodness-of-fit test, Power, Sensitivity Analysis, Goodness-of-fit Indices, Expected Parameter Changes (EPC)

[1] Acknowledgements: We appreciate the very useful comments on an earlier version of two anonymous reviewers.

.

1. Introduction

In an influential paper, MacCallum, Browne and Sugawara (1996: 131) write: *"if the model is truly a good model in terms of its fit in the population, we wish to avoid concluding that the model is a bad one. Alternatively, if the model is truly a bad one, we wish to avoid concluding that it is a good one."* The mentioned two types of wrong conclusions correspond to what in statistics are known as Type I and Type II errors, whose probabilities of occurrence are called α and β respectively. Although everybody would agree that α and β should be as small as possible, in the practice of SEM these probabilities are seldom controlled. In this paper we will show the consequences of not controlling the probabilities α and β.

To discuss this issue we first have to define what is a good and a bad model in terms of misspecifications. MacCallum *et al*. (1996) do not give a *definition of good and bad models*. We suggest that good and bad is defined in this context by the absence (*good*) or presence (*bad*) of misspecifications in the model, as done by Hu and Bentler (1998: 427), who state that: "*a model is said to be misspecified when (a) one or more parameters are estimated whose population values are zeros (i.e. an over-parameterized misspecified model) (b) one or more parameters are fixed to zeros whose population values are non-zeros (i.e. an under-parameterized misspecified model) or both.*" In line with Hu & Bentler (1998), we believe that (b) is the type of misspecification that has more serious consequences, so in this paper we merely discuss that type of misspecification. In the case of just one parameter of a model being misspecified, *the size of the misspecification is the absolute difference between the true value of the parameter and the value specified in the analysis*. If there is more than one parameter misspecified, the size of the misspecification of the model is also determined by the differences between the restricted values in the specified model and the true population values of the parameters under the correct model. This definition of the size of the misspecifications deviates from the definition of other scholars such as Fan and Sivo (2006), who define the size of the misspecification on the basis of the non-centrality parameter or the power of the test.

Some authors, e.g. Browne & Cudeck (1993) and MacCallum *et al.* (1996), have argued that models are always simplifications of reality and are therefore always misspecified. Although there is truth in this argument, this is not a good reason to completely change the approach to model testing. What is needed is for (a) models with substantially

relevant misspecifications to be rejected and (b) models with substantially irrelevant misspecifications to be accepted.

To make our discussion more concrete, we now provide examples using population data on what we mean by substantially relevant misspecifications and substantially irrelevant misspecifications.

A substantively relevant misspecification

As an example of a substantively relevant misspecification, consider the fundamental causal model example M_1:

$$y_1 = \gamma_{11}x_1 + \zeta_1 \quad (1)$$

$$y_2 = \beta_{21}y_1 + \gamma_{22}x_1 + \zeta_2 \quad (2)$$

where: $E(x_i)=E(y_i)=E(\zeta_i)=0$, $E(x_i,\zeta_j) = 0$ and $E(\zeta_1,\zeta_2) = \psi_{21}$and

where all variables are standardized except for the disturbance term.

The purpose of many studies is to determine whether there is an effect of one variable, i.e. y_1 on another one, i.e. y_2. To test this hypothesis, it is essential to ensure that all variables causing spurious relationships between the two variables have been introduced. If that is not the case, the covariance between the disturbance term (ψ_{21}) will not be zero.

If ψ_{21} is other than zero and the researcher specifies a model M_0 where $\psi_{21} = 0$, the effect β_{21} will be over- or under-estimated[2] and wrong conclusions about this parameter may be drawn. Depending on the size of the misspecification of the model M_0 (absolute value of the deviation of ψ_{21} form zero) a substantial misspecification can be attained, in which case the model should be rejected.

To make the example more complete, consider the following (true) population parameter values for model M_1 (see figure 1); $\gamma_{11} = .4$, $\beta_{21} = .0$, $\gamma_{22} = .1$ and $\psi_{21} = .2$. According to Hu-Bentler's definition of misspecification, the model M_0 (see figure 2) is misspecified since it imposes the incorrect restriction of $\psi_{21} = 0$.

Figure 1 and 2 about here

[2] The effect (β_{21}) could also be underestimated if the correlation between the disturbance term is negative.

The size of the misspecification is .2, i.e. the difference between the value of ψ_{21} under M_0 and its value under the correct model M_1. Note that the size of the misspecification would always be .2 regardless of the size of the other parameters in the model.

The consequence of the misspecification is that the effect β_{21} will be overestimated when fitting M_0 instead of the true model M_1. The expected value would be .2 and not .0 and so the wrong conclusion will be drawn that the variable y_1 has an effect on y_2. This example illustrates a case where a misspecification yields wrong conclusions, so this is a case of a "bad model" that should be rejected.

A substantively irrelevant misspecification

As an example of a model with an irrelevant misspecification we use a simple but important example from factor analysis. Consider the following 2-factor model M1:

$$x_1 = b_{11} f_1 + e_1$$
$$x_2 = b_{21} f_1 + e_2$$
$$x_3 = b_{31} f_2 + e_3 \qquad (3)$$
$$x_4 = b_{41} f_2 + e_4$$

where $E(f_i)= 0$ and $E(f_i) = 1$ $E(f_i e_j) = 0$ $E(e_i e_j) = 0$ while $E(f_1 f_2) = \rho$

and suppose that our interest lies in assessing whether this is a one-factor model, i.e. whether the correlation ρ is equal to 1. Let's note as M_0 the model that imposes ρ to be equal to 1. Suppose that population values for the parameters of loadings equal .8 and the correlation coefficient (ρ) equals .95. In that case, substantive researchers would agree that the two factors are the same, that is, that there is only one factor and not two. According to the definition stated previously, in this case the size of the misspecification of M_0 is .05, regardless of the size of the other parameters in the model. The size of this misspecification is substantively irrelevant, and therefore one would not like to reject the model M_0 since the model is adequate for all practical purposes even though it is not exactly correct. This illustrates the situation of a model with substantively irrelevant misspecifications which should be accepted. Figures 3 and 4 show the corresponding path diagram of the true and the approximate models.

Figure 3 and 4 about here

The above two examples should not imply that relevant misspecifications occur only for path analysis models, while irrelevant misspecifications occur only in factor analysis models. The mentioned problems can occur in both types of models and, of course, in the combination of both models.

In this paper we want to show, first of all, that the standard procedures for the evaluation of models do not work as required. After that we want to suggest an alternative approach for the evaluation of structural equation models.

The structure of the paper is as follows. Section 2 reviews the standard procedure of using goodness-of-fit testing and goodness-of-fit indexes for evaluation of models in the SEM tradition. Section 3 illustrates how not controlling for Type I and Type II errors does not work well, in that, even in the case of very simple but fundamental models, a bad model is typically accepted whilst a model that is good for all practical purposes is typically rejected. Section 4 describes the reason for these problems. Section 5 suggests an alternative approach based on detection of model misspecification. Section 6 describes an illustration with empirical data of the proposed procedures. Section 7 concludes with a discussion.

2. Traditional goodness-of-fit testing in SEM

In SEM, the goodness-of-fit of a specified model M is typically tested using a chi-square goodness-of-fit test statistic T (the so called CHI2 test), defined as n (the sample size) times the value of a discrepancy function that evaluates the differences between the observed covariance matrix in the sample and the fitted covariance matrices based on the parameter estimates and the specified model. Different discrepancy functions that take care of different distributional assumptions can be used (see Bollen, 1989). Under standard assumptions and when the model holds, T is asymptotically χ^2 distributed with degrees of freedom (df) equal to the number of over-identifying restrictions implied by the specified model. So, in the standard approach, M is rejected when

$$T > c_\alpha \qquad (4)$$

where c_α is the critical value of the test, that is, the value for which $pr(\chi^2(df) > c_\alpha) = \alpha$, and α being the chosen significance level of the test. Typically, researchers choose $\alpha =$

0.05, so the probability α of rejecting the model when the model is exactly correct (Type I error) is 0.05. The power of this test, that is, the probability of rejecting M_0 when the model does not hold, can be computed conditional to specific values of parameters of an alternative model (M_a) that deviates from M_0 (c.f., Satorra and Saris, 1985), with M_0 nested in M_a. Power values are seldom computed in applications, which means that no proper control of the probability β of a Type II error (the wrong decision of accepting the model when it is a wrong model) is exerted[3].

To evaluate models, most researchers and editors of journals prefer to rely on the information provided by one or more (goodness/badness) fit indices (FI) that measure deviation of the analysed model from baseline models instead of the CHI2 test. One of the reasons for this shift is that the use of the CHI2 test for model evaluation is not problem-free. A classical criticism of the CHI2 test is its severe dependence on sample size, in the sense that any small misspecification in the model will be detected by the CHI2 test (leading to rejection of the model) provided the sample size is large enough. Hu and Bentler (1998: 429) say about this: "*the decision for accepting or rejecting a particular model may vary as a function of sample size, which is certainly not desirable.*" This problem with the CHI2 test has led to the development of a plethora of Fit indices. Marsh, Hau, and Grayson (2005) provide a detailed overview of these fit indices, from which it is clear that many of them are functions of the CHI2 test statistic (see Appendix 1).

Nowadays, the traditional model evaluation method has been replaced by a similar procedure using FI statistics, with the model being rejected if:

$$FI < C_{fi} \qquad (5a)$$

where C_{fi} is a fix cut-off value developed specifically for each FI.

This procedure is used for FIs such as AGFI and GFI (Jöreskog & Sörbom, 1981), that have a theoretical upper value of 1 for good fitting models. There are however, also FIs for which a theoretical lower value of 0 indicates a good fit; e.g. RMSEA (Steiger & Lind, 1980), for which the model is rejected if:

$$FI > C_{fi} \qquad (5b)$$

[3] Note that power = 1 – β.

In Equation 5a and 5b, C_{fi} is the cut-off value for the specific FI. Such values have been derived from analyses of simulated data (see for example Hu in Bentler (1999)). In the Appendix we report the cut-off values (C_{fi}) of the FIs discussed in this paper. Marsh (1995) emphasized however that no rationale has been given for using fixed cut-off values. In fact, by using a fixed cut-off value for FI in the way described, the FI acts as a statistic for hypothesis testing: that is, if the critical value is exceeded, the model is rejected; if not, the model is accepted. However, the choice for a specific cut-off value is not based on controlling either Type I or Type II errors, i.e. the probabilities α and β mentioned above. Some researchers, e.g. Marsh, Hau, & Wen (2004), Fan & Sivo (2005), and Barrett (2006) have criticized using FI with fixed cut-off values as if they were test statistics.

Now we will argue that regardless of whether we use the traditional CHI2 test or the FIs, without attending to the probabilities α and β models with serious misspecifications (i.e. "*bad"* models) may have a high chance of being accepted, whilst models with irrelevant misspecification (i.e. "good" models) may have a high chance of being rejected. Illustrations of both instances are given below using the two models discussed above and depicted in Figures 2 and 4.

3. Illustration of the problem using population data

In this section, we use population data to illustrate the consequences of the standard procedures for model evaluation. We begin with a model that is definitely wrong but which has a high chance of being accepted when using the standard procedures. This is the case of the path analysis model M_1 whose path diagram and population values of parameters are shown in Figure 1. Table 1 shows the covariance matrix implied by the model in Figure 1

Table 1 about here

These data have been analyzed with the *bad* model M_0 depicted in Figure 2. Just as a reminder, we say it is a *bad* model since the size of the misspecification of assuming no correlation between the disturbance term ($\psi_{21}=0$) is big (.2) and this misspecification leads to severely wrong conclusions regarding the effect of y1 on y2. The results of the analysis of M_0 using the covariance matrix presented in Table 1 are shown in the first row of Table 2.

Attending to all goodness-of-fit measures, it follows – without exception – that the model should be accepted.

We can generate more implied covariance matrices by specifying other parameter values. These covariance matrices can also be analyzed with the same model in order to find out whether the size of the population parameter values influences the ability of the goodness-of-fit measures to detect the misspecification in model M_0. This has been carried out for different values of the parameter γ_{22}, while keeping the other parameters of the model fixed to the population values specified earlier. Note that the size of the misspecification does not change. The results of those analyses are summarized in Table 2.

Table 2 about here

Table 2 shows the values of CHI2 statistic, the power of the CHI2 test of this model[4] based on the population covariance matrices implied by the model in Figure 1, with varying values for γ_{22} (first column Table 2). We also provide the population values for the fit indices SRMR, RMSEA, CFI and AGFI to illustrate their performance in model evaluation. In this study the sample size is fixed at 400, which is not an unusual sample size in research. If the CHI2 test statistic would only be affected by the size of the misspecification of the model, then the CHI2 test statistic would have the same value across all analyses because the size of the misspecification is the same for all of them. However, this table shows that the CHI2 test statistic is not only affected by the size of the misspecification of the model but is also affected by the value of the parameter γ_{22}. This parameter has nothing to do with the misspecification, thus the CHI2 test statistic is also (next to sample size) affected by the value of incidental parameters in the model.

The CHI2 values show that the misspecification of the model is likely to be detected only for very large values of the parameter γ_{22}. Therefore, in practice, with the 5% level CHI2 test, the misspecification of the model will not be detected, unless the γ_{22} is very large or unless the sample size is very large. As a consequence, a biased estimate for the parameter of major interest (β_{21}) will be reported. This problem with the CHI2 test statistic, has already been discussed by Saris, Satorra and Sörbom (1987) and Saris, den Ronden and Satorra

[4] The power of the test is estimated on the basis of the non-centrality parameters (NCP) obtained by analysing population data (Satorra and Saris 1985). The non-centrality parameter is equal to the CHI2 statistic in this case.

(1987) who suggested considering the power of the test. The third column shows that the 5% level CHI2 test statistic has reasonable power to reject a wrong model only when γ_{22} exceeds the value of .8.

This problem with the CHI2 test statistic is inherited by most of the FIs (such as RMSEA, CFI, and AGFI). This may not be that surprising, because these fit indices are functions of the CHI2 test; for example, in Table 2 we see that in the same way as with the CHI2 test, the RMSEA increases with the value of γ_{22} and only when γ_{22} is above 0.7 does RMSEA exceed the suggested cut-off value of 0.05. The other indices react in similar ways, except for the SRMR which remains constant.[5] Furthermore, besides similar behavior to the CHI2 test statistic, the other FIs – except the RMSEA – do not reject the model in any circumstance (the FIs do not exceed the threshold values). Finally, we see that the use of the Modification Index (MI) for ψ_{21} is not the solution (Saris, Satorra and Sörbom 1987) because the MI behaves exactly in the same way as the CHI2 test statistic. In this particular study, with only one misspecification, the MI equals the CHI2 test.

Under typical circumstances - a sample size of 400 and commonly found parameter values - the seriously misspecified model depicted in Figure 2 will not be rejected. This holds regardless of whether we use the CHI2 test or the commonly used FIs.

Let us now look at the second example where the model is adequate for all practical purposes and should therefore be accepted in a substantive research. Using the same approach, we calculate the correlation matrices for different values of the loadings of model M_1, keeping the correlation between the factors equal to .95. Thereafter, the different computed population correlation matrices were used as input to estimate the parameters for a factor model under the restriction that $\rho_{21} = 1$, i.e. assuming that the correlations between the observed variables can be explained by a "single-factor" (model M_0). Table 3 summarizes the results of this analysis.

Table 3 about here

This table shows that the model, which could be seen as a good model for all practical purposes would very likely be rejected when the loadings are larger than .8, using the standard procedures for model evaluation. An exception is found for the fit index CFI that accepts the model. This table shows that the model is rejected for most of the statistics if the

[5] In general, this is not the case. It is due to a specific character of this model.

size of the loadings increases. This is a very inconvenient result, because the better the measurement model – high loadings – the higher probability of getting rejected.

The above two examples illustrate the fact that the standard methods for model evaluation can lead to precisely the decisions that MacCallum *et al.* (1996) stated should be avoided. These problems associated with the CHI2 test have been documented in several papers that appeared 20 years ago (see Saris & Satorra (1986); Saris, den Ronden, and Satorra (1987); Saris, Satorra and Sörbom (1987)). Given the relationship between the standard test statistic T and the fit indices, the same problems occur for the fit indices as have been documented in detail by Fan and Sivo (2007); Corten, Saris and Satorra (unpublished).

4. What is the problem of the model test and fit indices?

As the above examples illustrate, there is a fundamental problem with the use of the standard 5% level CHI2 test as well as with the fit indices to assess model misspecification. The problem is that the FIs as well as the CHI2 test are not only affected by the size of the misspecification of the model but also by other characteristics of the model. We have shown the effect on the FIs of the size of incidental parameters unrelated with the misspecification. The phenomenon shown is similar to the classical problem of dependency of the 5% level CHI2 test on sample size. While the FIs have been developed mainly to cope with the effect of sample size on the CHI2 test, they offer no protection from parameter values unrelated with the misspecification of the model (Saris & Satorra, 1987); therefore, whether a misspecification is detected or not will depend heavily on characteristics unrelated with the misspecification, e.g. sample size, values of the parameters, number of indicators, etc.

The situation is even more complex when it comes to multiple hypothesis testing. In this case, the CHI2 test and other fit indices will have different sensitivity for different misspecifications of the model – this is discussed in Saris, Satorra and Sörbom (1987) – and one may therefore doubt whether a critical value can be specified at all for the model as a whole. A rejection of the model may be due to the test's high sensitivity to a specific misspecification; acceptance of the model may be due to low test sensitivity to important misspecifications.

From all of the above, we can conclude that the standard model evaluation procedures do not satisfy the above-mentioned requirements of MacCallum *et al.* (1996: 131): "if the *model is truly a good model in term of its fit in the population, we wish to avoid concluding*

that the model is a bad one. Alternatively, if the model is truly a bad one, we wish to avoid concluding that it is a good one."

Saris, Satorra and Sörbom (1987) argued that there is no simple procedure to test the model as a whole. The standard CHI2 test and the FIs, as they are nowadays used, do not give a proper answer to the issue of validity of the model, as they are affected by characteristics other than just the size of the misspecification.

To tackle the issue of the dependency of the CHI2 test on other model characteristics, Satorra & Saris (1985) suggested *taking the power of the test into account*. This approach is rather tedious for routine practice and can only be applied to limited sets of parameter restrictions if the rest of the model does not contain misspecifications. It is clear that this is a rather unlikely situation in most research.

5. An alternative approach

An alternative to the goodness-of-fit test is to turn attention to investigating whether specific misspecifications are present in the model. According to our definition, a model that contains one or more *relevant* misspecifications is not a good model. Starting from that principle, Saris, Satorra, and Sörbom (1987) suggested evaluating the quality of a model using the combination of Expected Change Parameters (EPC) and the modification index (MI). They noted that the EPC gives a direct estimate of the size of the misspecification for all fixed parameters, while the MI provides a significance test (with one degree of freedom) for the estimated misspecification (for more details we refer to the paper by Saris, Satorra and Sörbom (1987)). However, one should realize that the modification index has the same problem as the CHI2 test which is that the size of the estimate depends on other characteristics of the model. In addition, the direct EPC misspecification estimates are problematic because sampling fluctuations can be rather large as will be shown below. To tackle this issue we will introduce the standard error of the EPC and the power of the MI test.

6.1 More information about misspecifications

Fortunately, the following simple but fundamental relationship exists between the three statistics mentioned above (Saris, Satorra, and Sörbom, 1987, p. 121):

$$MI = (EPC / \sigma^2) \qquad (6)$$

where σ is the standard error of the EPC.

From this relationship it follows that

$$\sigma = EPC / \sqrt{MI}; \quad (7)$$

Thus, σ can be estimated from EPC and MI, statistics which are nowadays provided by most of the SEM software.

This formula allows one to estimate the EPC standard error for alternative restrictions. Information on the EPC standard error is helpful because it can be used to construct a confidence interval for the EPC. Since EPC is normally distributed, the 95% confidence interval is defined for any parameter (θ) as:

$$EPC-1.96\,\sigma < \theta < EPC+1.96\,\sigma \quad (8)$$

Knowing the size of the EPC and the MI also provides a simple way to estimate the power of the test for the size of each misspecification. Consider a specific deviation δ for which one would like to know the power. Hence, δ would be the minimum size of the misspecification that one would like to be detected by the test with a high likelihood (power). By standard theory, under deviation from the null hypothesis, the asymptotic distribution of the MI is non-central χ^2-with the non-centrality parameter (ncp) given by

$$ncp = (\delta / \sigma)^2 \quad (9)$$

By combining (7) and (9) we obtain:

$$ncp = (MI/EPC^2)\,\delta^2 \quad (10)$$

an expression of the ncp that is a function of statistics provided by the standard software and the user-specified value δ of maximally acceptable misspecification. This ncp can be used to determine the power of the test of a misspecification of δ for any value of the significance level α of the test, and for all restricted parameters. The power of the test can be

obtained from the tables of the non-central χ^2-distribution (or using any computer-based routine[6]) as:

$$\text{Prob}(\chi^2(1, \text{ncp}) > c_\alpha) \qquad (11)$$

where c_α is the critical value of an α-level test based on a χ^2-distribution with df=1 and $\chi^2(1, \text{ncp})$is the non-central chi-square distribution with non-centrality parameter ncp.

Note that this approach requires the specification of the deviation δ. We suggest that for a standardized structural parameter and for a correlated error term a misspecification of 0.1 is significant enough to be detected by a test. For factor loadings, one may follow the standard approach where loadings smaller that .4 are ignored. These values are merely suggestions and one could use other values for δ that are more appropriate within a specific theory of interest.

In the examples of the two misspecified models discussed above, for all the misspecified parameters we have computed: the standard error (σ) of the expected parameter change (EPC); the confidence interval for the EPC; the non-centrality parameter for a deviation δ; and the power of the test to detect a misspecification of δ or larger. These statistics obtained using formulae (7), (8) and (10) are presented in Table 4 and 5. We will discuss the results from the population study presented in Table 4 first.

Table 4 about here

The results show that the EPC values vary, in this case, around the proper value of the parameter and that the size of the interval depends on the value of γ_{22}. For small values of γ_{22}, the EPC values for ψ_{21} can be negative but values that are close to 0.4 are also possible. If γ_{22} is 0.5 or larger, the probability of obtaining a value for ψ_{21} close to zero becomes smaller, which is an indication that there is a misspecification in the model. Under this condition the MI also shows that the EPC is significantly different from zero which was not true for smaller values of γ_{22}. Note that the power of this test for a misspecification of 0.1 or larger for the parameter ψ_{21} is fairly low for all different values of γ_{22}.

[6] A computer program, JRule, that produces statistics for all the restricted parameters, based on the output of LISREL, has been developed by us. The program can be requested by sending an e-mail to vdveld@telfort.nl putting JRule in the subject line.

Table 5 about here

Let us now look at the second example. The results in Table 5 indicate that the MI increases rapidly with the increase of the size of the loadings. For values of the loadings larger than .8, the MI is significant even though the misspecification is minimal for all practical purposes. Accordingly, the power of the test also increases rapidly. This explains why this model, although it is good for all practical purposes, is likely to be rejected if the loadings get larger than 0.8.

6.2. What should be done?

As mentioned above, we suggest switching from goodness-of-fit testing, based on the CHI2 test and/or FIs, to searching for possible (one parameter) misspecifications in the model, using the MI, the EPC and the power of the MI test.

The approach we propose distinguishes the following four possible situations shown in Table 6, which result from combining the significance or not of the MI test and the high/low power of the MI test:

Table 6 about here

When MI is significant and the power of the MI test is low, we conclude that there is a misspecification because the test is not very sensitive (low power) and nevertheless a significant value of the MI has been obtained. This situation appears in Table 4 for values of $\gamma_{22} > .4$. This is the cell in Table 6 labeled “m” for misspecification.

Using a reversed argument, the decision is also simple if the MI is not significant and the power of the MI is high. In that case, the conclusion is that there is no misspecification, so the corresponding cell in Table 6 is labeled “nm”. This situation does not occur in the tables presented.

The situation is more complex if the MI is significant but the power of the MI test is high. In that case it may be a serious misspecification, but it may also be that the MI is significant due to a high sensitivity of the test for this misspecification. Therefore, in that situation, we suggest looking at the substantive relevance of the EPC: If the EPC is rather small, one concludes that there is no serious misspecification. This makes sense because,

generally, we do not want to adjust our model for a standardized coefficient of .001 even though this coefficient is significant. However, when the EPC is large, for example larger than .2, it is concluded that there is a relevant misspecification in the model. The first situation, with small EPC, occurs in Table 5 for values of the loadings larger .8 and the decision would again be correct for all practical purposes. This cell in Table 6 is labeled "EPC", for EPC use. If the decision is that there is a misspecification we will denote it as "EPC:m". If it is decided that there is no misspecification, this will be denoted as "EPC:nm".

The fourth and last situation is that in which MI is low, and the power of the MI test is also low. In that case it should be concluded that one lacks sufficient information to make a decision. This is the most frequently occurring situation in our examples. It occurs in Table 4 for values of $\gamma_{22}<.4$, and also in Table 5 for loadings of .8 or smaller. Concluding that not enough information is available to reach a decision for the validity or not of a specific restriction should in itself be informative. This case is labeled as inconclusive "I".

Thus, Table 6 helps to classify the different options we may be confronted with in conducting model evaluation.

6.3. Some complications

Unfortunately, the situation is more complex than that presented above because in empirical research we do not know which parameter is misspecified and SEM analysis software provides EPCs for all restricted parameters. This point has also been discussed extensively in several papers in the journal *Multivariate Behavioral Research* (number 1989; number 1990). We can compute the 95% confidence interval for the EPC and the power for every restricted parameter. As an illustration, we did so for two restricted parameters, β_{12} and γ_{21}, in model M_0 (Figure 2). The results are presented in Table 7.

Table 7 about here

For this specific model (M_0), the introduction of either ψ_{21} or γ_{21} will lead to a perfect fitting model. Hence, model M_0 extended with either ψ_{21} or γ_{21} are equivalent. The EPC estimates for the parameter γ_{21} are consistent estimates of the possible parameter values in this population study. In addition, for γ_{21} – just as for ψ_{21} – the standard errors of the EPC differ considerably for different values of γ_{22} (an incidental parameter) and so does the power

of the test. So the choice to include γ_{21} or ψ_{21} cannot be made solely on statistical grounds but should also be based on substantive arguments. This was also the conclusion of Kaplan (1990) at the end of the discussion about this issue in *Multivariate Behavioral Research.*

The situation for the parameter β_{12} is rather different. The introduction of the parameter β_{12} does not lead to a perfect-fitting model. The EPC gives an impression of what would be the most likely value for this parameter if estimated given the specified model. This value is decreasing with increasing values of γ_{22}. The confidence interval behaves in a similar way, so that it becomes more and more likely that the population (true) value is zero for this parameter. In addition, the MI never indicates that the EPC is significantly different from zero but on the other hand the power of this test is rather low for all data sets.

The situation becomes even more complicated if there is more than one important misspecification in the model. The EPCs are consistent estimates of the true value of the parameter provided that the other restrictions in the model are (approximately) correct. If this is not the case, multivariate EPC (Satorra, 1989) should be used to obtain consistent estimates of the change in a restricted parameter vector. This however would complicate the matter considerably and is not pursued here.

As indicated above, one could construct confidence intervals for the EPCs. To illustrate this, consider the population data corresponding to M_1 with γ_{22}=.1 and ψ_{21}=.2 and consider the fit of M_0. Suppose the EPC for ψ_{21} is found to be smaller than .20, say .10 with a standard error (σ) of .112. In this case, the 95% level confidence interval would run from -.124 to .324, and would thus contain zero, and the MI would not be significant. In such a case, one would typically conclude that *there is no misspecification*. However, we know that there is a misspecification in this model given the way in which the data have been generated (see model M_1), hence *this conclusion is incorrect*. The cause of this wrong conclusion is that the power is too low to detect a misspecification of 0.1 in this situation. This illustration shows that a non-significant MI does not necessarily mean that the EPC is zero in the population. A non-significant MI can also mean that there is not enough information (low power) to detect whether the value of the parameter deviates from zero. This is a rather different conclusion to just reporting non-significance of the parameter.

7. An illustration: Modelling school career

For the sake of illustration, the above methodology will be applied to a study in which the principal author of this paper was consulted some years ago. It corresponds to the analysis of school career data in The Netherlands (see Blok and Saris 1980 for details of this project). At the end of primary school, the type of secondary education that children should go on to study at has to be decided. This choice is very important because only the highest types of secondary school allow pupils to continue on to higher educational studies. The causal model, presented in Figure 5, was initially formulated and on the basis of prior substantive information on this issue.

Figure 5 about here

The correlation matrix, means and standard deviations of the variables involved, based on a sample of 383 pupils, are presented in Appendix 2. Using this data, the model in Figure 5 was fitted obtaining the following values for CHI2 and fit indices: CHI2= 161 with df =9, SRMR =.073, RMSEA = .21, CFI = .95 and AGFI =.67. According to the suggested cut-off values, all fit indices, with the only exception of CFI, would reject the model. How can we be sure of this conclusion? It is also possible that there are only very small misspecification(s) for which all the test statistics and fit indices, with the exception of CFI, are very sensitive.

Using the methods developed in this paper we will now sketch the typical steps to assess whether or not the model is substantially misspecified. Table 8 lists the MI test for each of the restricted parameters of the model. The table reports the MI, EPC and the power for each restricted parameter and the decisions based on Table 6. In the calculations, δ=.1 was chosen and the power was classified as high when it was above .75. The JRule software (Van der Veld, Saris and Satorra, 2008) was used to obtain the information reported in this table.

Table 8 about here

From this we can see that the model is misspecified, since for several parameters the decision is “m”, which corresponds to the case of low power but substantially large MI. In addition to that, we have the EPC decision for several parameters which means that one has to inspect the reported EPC for substantive significance (this is the case of high power). For several of these restricted parameters, BE 1 4, BE 2 4 and BE 3 4, the EPC is smaller than .01

(those labeled EPC:nm); but there are other restrictions with EPC large enough to indicate a serious misspecification (those labeled EPC:m).

We also see that there are a number of parameters which are most likely to be not misspecified (nm) because the power is high but nevertheless the MI is not significant. And finally there is only one parameter for which the status is unclear because the power is too low (the one labeled I for "inconclusive").

Given the number of restrictions that are found to be severely misspecified, the model can be adjusted in many different directions. The number of possibilities can be reduced by theoretical information on the described process: for example time ordering and other theoretical reasons exclude certain effects. Substantive considerations not to be discussed in the present paper (but detailed in Saris and Stronkhorst 1986) lead to the alternative model specification depicted in Figure 6.

Figure 6 about here

Analysis of this alternative model leads to the following results regarding CHI2 test and FIs: CHI2 = 3.88, with df=5 (p-value =.57), SRMR =.0076, RMSEA = .0, CFI= 1.0 AGFI = .98. Now all indices suggest that the model fits the data. However, this decision is also doubtful as it is possible that the power of the tests is so low for this model that the misspecifications are not detected. If we apply the method discussed in this paper we get the MI test results reported in Table 9.

Table 9 about here

In this table we can see that there are several parameters for which the power of the test is too low to decide if there is a misspecification or not (parameters labeled I). The table also shows that for many parameters we can conclude that there is no misspecification (nm) because the power is high but the MI is not significant.

In contrast with the standard evaluation of the model, we do not conclude that the model is acceptable. On the basis of these tests we should conclude that the model is not misspecified for those parameters for which the power is high enough to test for misspecification but that there are some parameters for which this study cannot determine whether they are misspecified due to of lack of power. This conclusion is quite different form the conclusion derived using the standard model evaluation procedures.

8. Conclusions

We have argued that the commonly used evaluation procedures for structural equation models can not be trusted. The reason is that the test statistics and fit indices used are not only affected by the size of the misspecifications but also by other characteristics of the model unrelated to the size of the misspecifications. For a more elaborate study of this phenomenon, providing data for more different types of misspecifications and more fit indices, we refer to Miles and Shevlin (2007) and for the effect of the model specification to Fan and Sivo (2007).

By power analysis for different possible misspecifications in the model, one can see that the CHI2 test statistic and FIs are unequally sensitive for different misspecifications. So a standard test for the complete model with a fixed critical value could lead to rejection because of a small misspecification for which the test is very sensitive. On the other hand, it could just as well lead to accepting a model with a large misspecification because the test is not sensitive enough for that misspecification. The conclusion is that, based on a general model test, it is hard to draw conclusions as to possible misspecification of a model.

An alternative to the model test is to look for possible misspecified restrictions in the model. Estimates of the misspecifications can be obtained from EPCs and the significance of that misspecification can be evaluated using the MI test. We have argued, however, that a decision as to whether a restriction is misspecified should include information on the power of the MI test. In many situations there is not enough information (i.e. power too low) to say whether or not the restriction is misspecified. The standard practice of concluding that a model is a good model if the fit is acceptable, or no significant MIs are found, is unjustified because non-significance may just be due to lack of power. Non-significance should not imply that the parameter is zero, except when there is reasonable power.

We propose using the MI for detection of misspecifications in combination with the power of the MI test. This allows one to specify four different situations (Table 6) for which the decision concerning the presence/absence of misspecification can be made. In some situations, where the power is low and the MI is not significant, one will come to the conclusion that not enough information is available regarding the validity of that restriction.

Besides the power of the test, one has also to take into account the substantive relevance of the misspecifications. Very small misspecifications can lead to significant MIs if

the power is high. However, if these deviations are very small, one should consider whether it makes sense to reduce the parsimony of the model by introducing parameters that do not deviate substantively from zero or any other fixed value.

We have also shown that many different corrections will be suggested in some cases, with the MI statistics giving insufficient evidence regarding the best correction. The decision as to the specific "direction" in which a model needs to be augmented should be based on theoretical grounds.

In our approach, we did not specify what constitutes high and low power. It should be made clear that these specifications are rather arbitrary. The choice of the critical deviations and the threshold value for power should be dictated by the standards of the specific discipline. We suggested, as a critical deviation, .1 for standardized structural parameters and .4 for factor loading (because these are often used as critical values in social science research); also .75 for threshold of high power value. These values, however, are rather arbitrary and can change in different areas, and over time as the research advances. Different disciplines should choose their standards. We do not wish to claim any higher precision than this. Further research should show if greater precision is possible and thus whether there is scope for making these cutting points more precise. Important is that once a choice is made regarding the power and the unacceptable misspecification (d) it is clear what the result of the test means. This is in sharp contrast to the standard model test and the use of the FI where the power is neglected.

References

Barret, P. (2006). Structural equation modeling: Adjudging model fit. In *Personality and Individual Differences, 42, pp. 815-824.*

Beauducel, A. and Wittmann, W. W. (2005). Simulation study on fit indices in CFA based on data with slightly distorted simple structure. In *Structural equation modeling, 12, pp. 41-75*

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

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

Blok, H. and Saris, W. E. (1980). Relevante variabelen bij het doorverwijzen na de lagere school: een structureel model. *Tijdschrift voor onderwijs,* Research 5, pp. 63-80.

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

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

Corten, I. W., Saris, W. E. and Satorra A. (forthcoming) Can Fit Indices be used to evaluate Structural Equation Models?

Fan, X. and Sivo, S. A. (2005). Sensitivity of fit indices to misspecified structural or measurement model components: Rational of two. Index strategy revisited. In *Structural equation modeling,* 12, pp. 343-367.

Fan, X and Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. In Multivariate Behavioral Research, 42, 509-529

Hu, L. and Bentler, P.M. (1995). Evaluating model fit. In R. Hoyle (Ed.), *Structural equation modeling: Issues, concepts and applications* (pp. 76-99). Newbury Park, CA: Sage.

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

Marsh, H. W., Hau, K. T. and Wen, Z. (2004). In search of golden rules: Comment on hypothesis. Testing approaches to setting cut-off values for fit indices and dangers in overgeneralising Hu and Bentlers´ (1999) findings. In *Structural equation modeling, 11,* pp. 320-341.

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, pp. 130-149.

McDonald, R. P. and Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7. pp. 64- 82.

Miles, J. and Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and individual differences*, 42, pp. 869-874.

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. and Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. In *Sociological Methodology, 17,* pp. 105-129.

Saris, W. E. and 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. and Stronkhorst, H. (1984). *Causal Modelling in Nonexperimental Research*. Sociometric Research Foundation.

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

Satorra, A. and Saris, W. E. (1983). Power evaluations in structural equation models. In K. A. Bollen and J. S. Long (Eds). *Testing structural equation models. London*, Sage, pp. 163-181.

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

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

Van der Veld, W.M., Saris W.E. and Satorra, A. (2008). *JRule 2.0: User manual*.

Appendix 1: The definition of the Fit Indices included in this study

Fit index	Formula[1)]	Reference	Cut-off value
AGFI	$1-\left[\frac{p(p+1)}{2df_h}\right]\left[1-GFI\right]$	Jöreskog & Sörbom (1989)	.9
GFI	$1-\left(\frac{\chi_h^2}{\chi_u^2}\right)$	Jöreskog & Sörbom (1989)	.95
SRMR	$\sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\left(s_{ij}-\hat{\sigma}_{ij}\right)/s_{ii}s_{jj}\right]^2}{p(p+1)/2}}$	Jöreskog & Sörbom (1989)	.05
NFI or BBI	$\frac{\left(\chi_b^2-\chi_h^2\right)}{\chi_b^2}$	Bentler & Bonett (1980)	.95
CFI	$\frac{\hat{\lambda}_b-\hat{\lambda}_h}{\hat{\lambda}_b}$	Bentler (1990)	.95
RMSEA	$\sqrt{\frac{\hat{F}_0}{df_h}}=\sqrt{Max\left\{\left(\frac{\hat{F}_h}{df_h}-\frac{1}{n}\right),0\right\}}$	Steiger (1990), Steiger & Lind (1980)	.05

[1)] *Meaning of subscripts and symbols:*

- *F is the fitting function, $\chi^2 = n*F$.*
- *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.*
- *λ is the non-centrality parameter.*

Appendix 2: The correlations, means and standard deviations of the variables of the school career model.

Variables	Correlations							SD	Mean
School achievement	1.000							22.5	53.7
Advise teacher	.8113	1.000						1.8	3.0
Preference parents	.7858	.8534	1.000					1.8	3.3
School test score	.8109	.7641	.7611	1.000				27.8	50.5
School choice	.7921	.8605	.9879	.7747	1.000			1.8	3.3
Quality school	.2763	.1905	.2799	.4664	.2847	1.000		28.0	50.9
Background parents	.1963	.2821	.2969	.2435	.2966	.1399	1.000	1.5	3.2

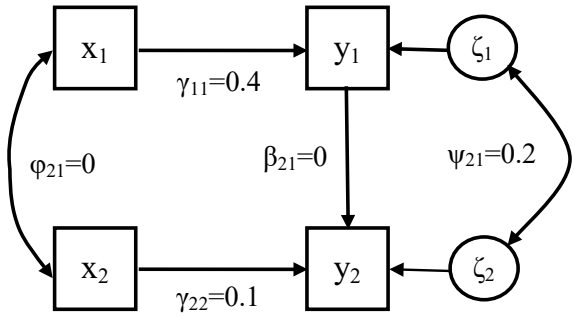

Figure 1: The causal population model M_1, with correlated disturbance term.

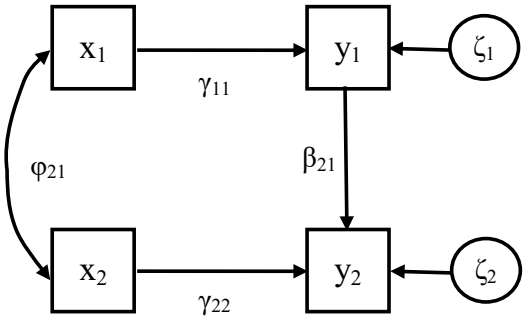

Figure 2: The hypothesized causal model M_0, without correlated disturbance term ($\psi_{21}=0$).

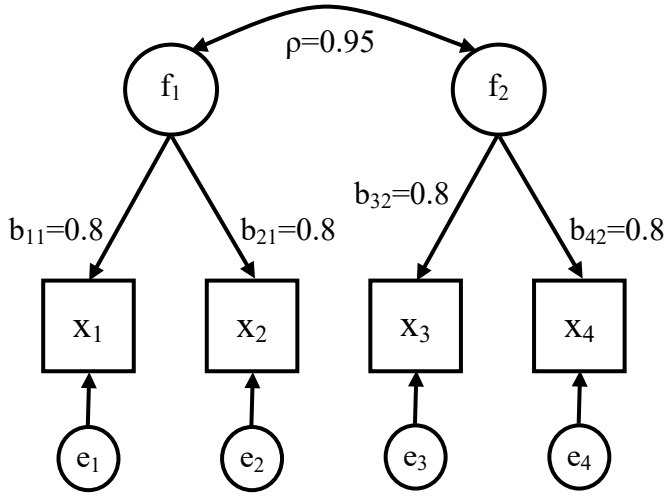

Figure 3: The population factor model (M_1), with a correlation of .95 between the two factors.

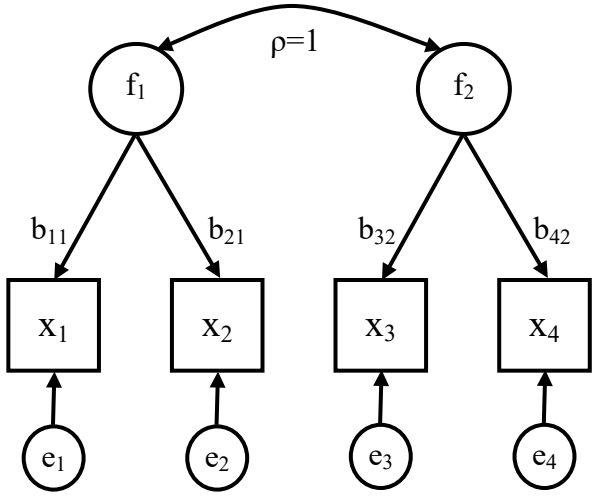

Figure 4: The hypothesized factor model (M_0), with perfect correlation between the two factors.

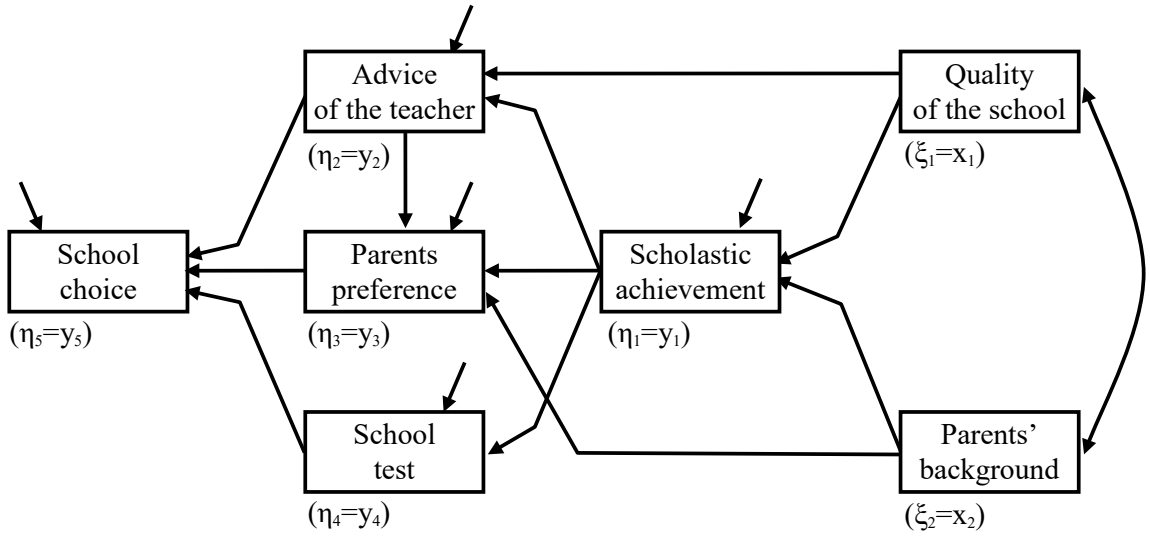

Figure 5: The school career model of Blok and Saris (1980) where it is assumed that all variables were measured without errors, therefore $\eta_i=y_i$ and $\xi_j=x_j$.

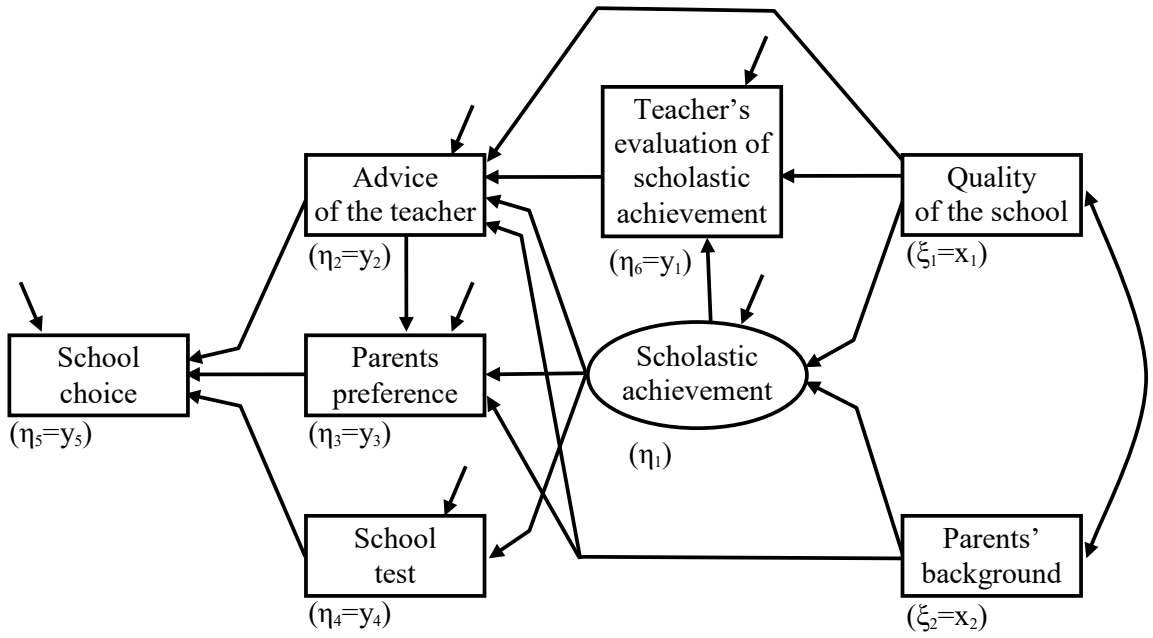

Figure 6: The adjusted model for the School career model of Blok and Saris (1980) where it is assumed that all variables were measured without errors, therefore $\eta_i=y_i$ and $\xi_j=x_j$, except $\eta_1$which is not equal to y_1 while $\eta_6= y_1$.

Table 1: Implied correlation matrix of the model depicted in Figure 1.

	y_1	y_2	x_1	x_2
y_1	1.00			
y_2	0.20	1.00		
x_1	0.40	0.00	1.00	
x_2	0.00	0.10	0.00	1.00

Table 2: Goodness-of-fit measures* for model M_0 with a constant misspecification ($\psi_{21}=0$) and with an increasing size of the incidental parameter γ_{22}.

γ_{22}	CHI2**	power***	RMSEA	CFI	AGFI	SRMR	MI of ψ_{21}
0.1	3.20	0.34	0.00	1.00	0.99	0.025	3.20
0.2	3.30	0.35	0.00	1.00	0.99	0.025	3.30
0.3	3.49	0.37	0.00	1.00	0.99	0.025	3.49
0.4	3.80	0.38	0.00	1.00	0.99	0.025	3.80
0.5	4.20	0.43	0.01	1.00	0.99	0.025	4.20
0.6	5.07	0.50	0.03	1.00	0.98	0.025	5.07
0.7	6.47	0.62	0.04	0.99	0.98	0.025	6.47
0.8	9.50	0.79	0.06	0.98	0.97	0.025	9.50
0.9	20.27	0.99	0.10	0.96	0.94	0.025	20.27

* *Estimates obtained with the quasi-maximum likelihood procedure in LISREL8.80. The population size is 400.*

** *The CHI2 is equal to the non-centrality parameter in this case, because population data are analyzed (Saris and Satorra, 1985).*

*** *The power-values were computed with $\alpha=0.05$ and 4 degrees of freedom (because the model has 4 df). To obtain the correct power value, we used tables, such as can be found in Saris and Stronkhorst (1984), that relate the power, the degrees of freedom of the test, and the non-centrality parameter.*

Table 3: Goodness-of-fit measures* for model M_0 with a constant misspecification ($\rho_{21}=1$) and with increasing size of the incidental parameters b_{ij}.

b_{ij}	ncp**	power***	RMSEA	CFI	AGFI	SRMR	MI of ρ_{21}
0.70	1.09	0.18	0.000	1.00	0.99	1.00	1.09
0.75	2.15	0.24	0.014	1.00	0.99	1.00	2.15
0.80	3.75	0.39	0.047	1.00	0.98	1.00	3.75
0.85	8.81	0.76	0.092	0.99	0.95	0.99	8.81
0.90	18.13	0.96	0.140	0.99	0.89	0.99	18.13

* *Estimates obtained with the quasi-maximum likelihood procedure in LISREL8.80. The population size is 400.*

** *The non-centrality parameter (ncp) is equal to the CHI2 in this case because population data are analyzed (Saris and Satorra, 1985).*

*** *The power-values were computed with $\alpha=0.05$ and 2 degrees of freedom (because the model has 2 df). To obtain the correct power value, we used tables, such as can be found in Saris and Stronkhorst (1984), that relate the power, the degrees of freedom of the test, and the non-centrality parameter.*

Table 4: Statistical information about the misspecification (ψ_{21}=0) in model M_0; including the power related to the size of the incidental parameter γ_{22}*.

γ_{22}	$EPC_{\psi 21}$	MI	95% confidence interval		σ	ncp	power
			Low	*high*			
0.1	0.2	3.20	-0.019	0.419	0.112	0.80	0.146
0.2	0.2	3.30	-0.016	0.416	0.110	0.83	0.149
0.3	0.2	3.49	-0.010	0.410	0.107	0.87	0.155
0.4	0.2	3.80	-0.001	0.401	0.103	0.95	0.164
0.5	0.2	4.20	0.009	0.391	0.098	1.05	0.176
0.6	0.2	5.07	0.026	0.374	0.089	1.27	0.202
0.7	0.2	6.47	0.046	0.354	0.079	1.62	0.245
0.8	0.2	9.50	0.073	0.327	0.065	2.38	0.336
0.9	0.2	20.27	0.113	0.287	0.044	5.07	0.615

* *The expected parameter change (**EPC**) and modification index (**MI**) are taken from the LISREL output (carried out for Table 2). The 95% confidence interval for the EPC is estimated with expression (8), the standard error of the EPC (**σ**) is estimated with (7), the non-centrality parameter (**ncp**) is estimated with (10) for a misspecification of δ=0.1 or larger, and the power can be found in power tables in the literature, e.g. Saris and Stronkhorst (1984). The power was estimated with α=0.05 and df=1, because the test is on a single parameter.*

Table 5: Statistical information about the misspecification ($\rho_{21}=1$) in model M_0; including the power related to the size of the incidental parameters b_{ij} (the factor loadings)*.

b_{ij}	$EPC_{\rho 21}$	MI	95% confidence interval		σ	ncp	Power
			Low	*high*			
0.70	-0.05	1.09	-0.15	0.05	0.050	4.36	0.550
0.75	-0.05	2.15	-0.12	0.02	0.034	8.65	0.840
0.80	-0.05	3.75	-0.11	0.01	0.026	15.0	0.972
0.85	-0.05	8.81	-0.08	-0.16	0.017	34.6	1.000
0.90	-0.05	18.13	-0.07	-0.03	0.012	72.5	1.000

* *The expected parameter change (**EPC**) and modification index (**MI**) are taken from the LISREL output (carried out for Table 3). The 95% confidence interval for the EPC is estimated with expression (8), the standard error of the EPC (**σ**) is estimated with (7), the non-centrality parameter (**ncp**) is estimated with (10) for a misspecification of $\delta=0.1$ or larger, and the power can be found in power tables in the literature, e.g. Saris and Stronkhorst (1984). The power was estimated with $\alpha=0.05$ and $df=1$, because the test is on a single parameter.*

Table 6: The decisions to be made in the different situations defined on size of the modification index (MI) and the power of the test.

	High power	**Low power**
Significant MI	Inspect EPC (EPC)	Misspecification present (m)
Nonsignificant MI	No misspecification (nm)	Inconclusive (I)

Table 7: Statistical information about the misspecification (β_{12} and γ_{21}) in model M_0; including the power related to the size of the incidental parameter γ_{22}*.

	Parameter β_{12}					**Parameter γ_{21}**				
γ_{22}	**EPC$_{\beta 12}$**	**95% interval**		**MI**	**power**	**EPC$_{\gamma 21}$**	**95% interval**		**MI**	**power**
0.1	0.200	-0.026	0.426	3.00	0.140	-0.095	-0.199	0.009	3.20	0.478
0.2	0.170	-0.037	0.377	2.60	0.158	-0.095	-0.197	0.007	3.30	0.491
0.3	0.140	-0.049	0.329	2.10	0.179	-0.095	-0.195	0.005	3.49	0.517
0.4	0.110	-0.055	0.275	1.70	0.219	-0.095	-0.191	0.001	3.80	0.536
0.5	0.090	-0.065	0.245	1.30	0.243	-0.095	-0.186	-0.004	4.20	0.578
0.6	0.070	-0.061	0.201	1.10	0.321	-0.095	-0.178	-0.012	5.07	0.659
0.7	0.060	-0.064	0.184	0.90	0.352	-0.095	-0.168	-0.022	6.47	0.764
0.8	0.050	-0.067	0.167	0.70	0.388	-0.095	-0.155	-0.035	9.50	0.900
0.9	0.040	-0.061	0.141	0.60	0.503	-0.095	-0.136	-0.054	20.27	0.997

* *The expected parameter change (**EPC**) and modification index (**MI**) are taken from the LISREL output (carried out for Table 3). The 95% confidence interval for the EPC is estimated with expression (8), the standard error of the EPC (**σ**) is estimated with (7), the non-centrality parameter (**ncp**) is estimated with (10) for a misspecification of δ=0.1 or larger, and the power can be found in power tables in the literature, e.g. Saris and Stronkhorst (1984). The power was estimated with α=0.05 and df=1, because the test is on a single parameter.*

Table 8: The test on misspecifications in the school career model (see figure 5).

Parameter			From	to	MI	EPC	Power	Decision
BE	5	1	η5	η1	0.02	0.00	0.999	nm
BE	1	2	η1	η2	18.66	-0.68	0.098	m
BE	4	2	η4	η2	36.82	0.10	0.999	EPC:m
BE	1	3	η1	η3	22.41	-0.52	0.149	m
BE	2	3	η2	η3	27.35	0.66	0.125	m
BE	4	3	η4	η3	46.07	0.11	0.999	EPC:m
BE	1	4	η1	η4	68.81	0.00	0.999	EPC:nm
BE	2	4	η2	η4	43.36	0.00	0.999	EPC:nm
BE	3	4	η3	η4	13.24	0.00	0.999	EPC:nm
BE	1	5	η1	η5	20.35	-0.46	0.165	m
BE	2	5	η2	η5	18.97	0.35	0.236	m
BE	3	5	η3	η5	2.57	0.26	0.095	I
BE	4	5	η4	η5	45.97	0.11	0.999	EPC:m
GA	3	1	η3	ξ1	10.30	0.08	0.980	EPC:nm
GA	4	1	η4	ξ1	70.93	0.26	0.899	EPC:m
GA	5	1	η5	ξ1	0.00	0.00	0.999	nm
GA	2	2	η2	ξ2	18.66	0.13	0.914	EPC:m
GA	4	2	η4	ξ2	8.25	0.09	0.890	EPC:nm
GA	5	2	η5	ξ2	0.03	0.00	0.999	nm
PS	2	1	η2	η1	18.66	-0.74	0.090	m
PS	3	1	η3	η1	10.30	-0.30	0.187	m
PS	5	1	η5	η1	0.00	0.00	0.999	nm
PS	3	2	η3	η2	10.30	0.79	0.060	m
PS	4	2	η4	η2	43.36	0.12	0.999	EPC:m
PS	5	2	η5	η2	0.02	0.00	0.999	nm
PS	4	3	η4	η3	13.24	0.05	0.999	EPC:nm
PS	5	3	η5	η3	0.03	0.00	0.999	nm
PS	5	4	η5	η4	0.02	0.00	0.999	nm
TE	1	1	y1	y1	95.53	0.23	0.989	EPC:m
TE	2	1	y2	y1	9.86	-0.05	0.999	EPC:nm
TE	3	1	y3	y1	0.97	0.00	0.999	nm
TE	4	1	y4	y1	78.50	-0.17	0.999	EPC:m
TE	5	1	y5	y1	8.01	0.00	0.999	nm
TE	2	2	y2	y2	2.08	-0.28	0.081	I
TE	3	2	y3	y2	0.05	0.00	0.999	nm
TE	4	2	y4	y2	10.25	0.04	0.999	EPC:nm
TE	5	2	y5	y2	0.01	0.00	0.999	nm
TE	3	3	y3	y3	0.03	0.00	0.999	nm
TE	4	3	y4	y3	3.32	0.01	0.999	nm
TE	5	3	y5	y3	0.03	0.00	0.999	nm
TE	5	4	y5	y4	0.02	0.00	0.999	nm
TD	1	1	x1	x1	77.17	-0.29	0.857	EPC:m
TD	2	1	x2	x1	16.10	-0.10	0.980	EPC:m
TD	2	2	x2	x2	19.29	0.13	0.925	EPC:m

* M= misspecification
Nm= no misspecification
EPC:m= inspection of the EPC leads to conclusion: misspecification
EPC:nm= inspection of the EPC leads to conclusion: no misspecification
I= inconclusive

Table 9: The test on misspecifications in the adjusted school career model (see figure 6).

Parameter			from	to	MI	EPC	Power	Decision
BE	5	1	η5	η1	0.01	0.00	0.999	nm
BE	1	2	η1	η2	0.86	-0.10	0.153	I
BE	4	2	η4	η2	0.65	-0.04	0.522	I
BE	6	2	η6	η2	0.33	0.03	0.482	I
BE	1	3	η1	η3	0.02	-0.01	0.294	I
BE	2	3	η2	η3	1.84	-0.05	0.774	nm
BE	4	3	η4	η3	0.04	0.00	0.999	nm
BE	6	3	η6	η3	1.08	0.02	0.999	nm
BE	1	4	η1	η4	3.76	0.21	0.152	I
BE	2	4	η2	η4	0.17	0.01	0.985	nm
BE	3	4	η3	η4	0.21	0.00	0.999	nm
BE	6	4	η6	η4	0.84	-0.03	0.863	nm
BE	1	5	η1	η5	0.04	-0.01	0.516	I
BE	2	5	η2	η5	1.73	-0.05	0.749	I
BE	3	5	η3	η5	0.00	0.01	0.000	I
BE	4	5	η4	η5	0.02	0.00	0.999	nm
BE	6	5	η6	η5	1.03	0.02	0.999	nm
BE	2	6	η2	η6	0.50	-0.02	0.947	nm
BE	3	6	η3	η6	1.38	0.02	0.999	nm
BE	4	6	η4	η6	0.16	-0.01	0.979	nm
BE	5	6	η5	η6	0.01	0.00	0.999	nm
GA	3	1	η3	ξ1	2.30	-0.06	0.714	I
GA	4	1	η4	ξ1	3.59	0.14	0.271	I
GA	5	1	η5	ξ1	0.00	0.00	0.999	nm
GA	6	1	η6	ξ1	1.07	-0.08	0.251	I
GA	2	2	η2	ξ2	1.91	0.05	0.789	nm
GA	4	2	η4	ξ2	3.85	-0.06	0.904	EPC:nm
GA	5	2	η5	ξ2	0.03	0.00	0.999	nm
GA	6	2	η6	ξ2	1.07	0.04	0.734	I
PS	2	1	η2	η1	0.86	-0.07	0.261	I
PS	3	1	η3	η1	0.01	0.00	0.999	nm
PS	4	1	η4	η1	3.76	0.30	0.100	I
PS	5	1	η5	η1	0.03	0.00	0.999	nm
PS	3	2	η3	η2	2.30	-0.04	1.000	nm
PS	4	2	η4	η2	0.17	0.01	0.985	nm
PS	5	2	η5	η2	0.02	0.00	0.999	nm
PS	6	2	η6	η2	0.00	0.00	0.999	nm
PS	4	3	η4	η3	0.21	0.01	0.996	nm
PS	5	3	η5	η3	0.02	0.00	0.999	nm
PS	6	3	η6	η3	0.55	0.01	0.999	nm
PS	5	4	η5	η4	0.01	0.00	0.999	nm
PS	6	4	η6	η4	0.84	-0.04	0.630	I
PS	6	5	η6	η5	0.02	0.00	0.999	nm
TE	2	1	y2	y1	0.22	-0.01	0.997	nm
TE	3	1	y3	y1	0.01	0.00	0.999	nm
TE	4	1	y4	y1	0.80	-0.03	0.846	nm
TE	5	1	y5	y1	0.02	0.00	0.999	nm
TE	2	2	y2	y2	2.20	0.07	0.563	I
TE	3	2	y3	y2	0.14	0.00	0.999	nm
TE	4	2	y4	y2	0.00	0.00	0.999	nm
TE	5	2	y5	y2	0.01	0.00	0.999	nm
TE	3	3	y3	y3	0.02	0.00	0.999	nm
TE	4	3	y4	y3	0.09	0.00	0.999	nm
TE	5	3	y5	y3	0.02	0.00	0.999	nm
TE	4	4	y4	y4	0.01	0.02	0.079	I

Parameter			from	to	MI	EPC	Power	Decision
TE	5	4	y5	y4	0.01	0.00	0.999	nm
TD	1	1	x1	x1	0.89	-0.05	0.471	I
TD	2	1	x2	x1	0.95	0.03	0.901	nm
TD	2	2	x2	x2	1.72	0.04	0.906	nm

* m= misspecification
nm= no misspecification
EPC:m= inspection of the EPC leads to conclusion: misspecification
EPC:nm= inspection of the EPC leads to conclusion: no misspecification
I= inconclusive