

Goodness-of-fit indices in confirmatory factor-analysis: an unreliable guide

by

Peter Prudon

Independent investigator and author in (clinical) psychology at Amsterdam
Primary care psychologist for SPEL Haarlemmermeer, Hoofddorp, the Netherlands

Main fields of interest:

Theoretical psychopathology
Obsessive-compulsive disorders
Motivation and emotion

Drs. P.C.H. Prudon
Sajetplein 31
1091 DB Amsterdam
The Netherlands
Tel. 0031 20 6723087
Tel. 0031 6 30498710 (mobile)
pprudon@hotmail.com

Advised referees:

Prof. Dr. J.F.M. ten Berge
Universiteit van Groningen: Faculteit Gedrags- & Maatschappijwetenschappen
Afdeling: Psychometrische & Statistische Technieken — Basiseenheid Psychometrie & Statistiek

Prof. Dr. H.A.L. Kiers
University of Groningen: Faculteit Gedrags- & Maatschappijwetenschappen
Methoden & Technieken van Gegevensverwerking - Basiseenheid Psychometrie & Statistiek

Keywords:

goodness of fit
fit-indices
confirmatory factor analysis
structural equation modelling,
unique variance and goodness of fit

**Goodness-of-fit indices in confirmatory factor-analysis:
an unreliable guide**

Abstract

A much cited part of the output of confirmatory factor analysis consists of goodness-of-fit (GOF) indices: measures for the correctness of the prediction of the factor structure. A small number of recent studies are reviewed to show that GOF-indices cannot be relied on: Many are strongly and paradoxically affected (overestimation of fit) by the degree of unique variance, especially in combination with high factor correlations. Comparative fit-indices do not survive that combination either. This makes it impossible to specify benchmark values per GOF-index to indicate degree of fit, let alone stringent cut-off values to split models in falsified and confirmed ones. An explanation of these spurious effects is offered.

An important contribution to construct-validation of multidimensional measuring instruments in psychology and related disciplines is investigating its factor structure. An empirically found factor structure which corresponds with the a priori division in subtests is a strong support for the ideas that guided construction of the instrument. Conversely, a psychological theory about, for instance, a psychopathological category may be translated into a number of items, and thereupon expectations about the way these items cluster may be formulated.

A direct and controlled way to test these clusters is confirmatory factor analysis (CFA). For the last few decades the method of choice in CFA has often been *confirmatory common¹ factor analysis* (CCFA) as a part of LISREL (Jöreskog, 1969) and similar programs for *structural equation modelling* (SEM). It can be used for testing both the *measurement* part and the *structural* aspect of the model. (The structural aspect relates to a more complicated theoretical model of which the measurement model is a part).

The output of LISREL and similar programs provides several clues as to the correctly and incorrectly specified parameters of the model, but an important part of it are χ^2 and a number of measures of *global* fit of the model, the so-called goodness-of-fit indices (GOF-indices). These should give a quick impression of the accuracy of prediction, if not a test whether the model holds for the population.

There has been much critical discussion in the specialized literature about the validity and reliability of these GOF-indices, since Hu & Bentler (1998, 1999) proposed their stringent cut-off values. A number of studies that played or could play an important part in this discussion will be discussed. They will raise serious doubts about the current GOF-indices.

The impetus to this paper is the observation that for the last decade or so construct validation of questionnaires is often done with LISREL by which authors contend themselves with reporting GOF-indices for a few rival models and primary factor loadings for the “winning” model, and nothing more, even if the values of the fit-indices are mediocre and some of the factor loadings are clearly too low (for an example, see Hoekstra et al., 2008). The sophistication of LISREL in combination with closeness to advised criterion values of ill-understood fit-indices seems to be an excuse to refrain from critically investigating sources and meaning of remaining misfit. This is problematic the more since GOF-indices can absolutely not be relied upon as I hope to show with this paper.

1. Testing predicted factor structure by means of LISREL

The prediction of the factor structure includes at least the number of factors and the assignment of items to the factors (indicators: variables with high *primary* factor loadings. It may also include the prediction of a number of *secondary* factor loadings. It may include a prediction of the factor correlations. The measurement error may be included in the model. If a structural model is tested, then there may be predictions about *extraneous* causal variables.

All these correlations, covariances and error variances (parameters) of the factor structure are translated back into a covariance (or correlation) matrix over *all* measured variables: the *implied covariance matrix*.

- This implied covariance matrix is adjusted to the empirically found *sample* covariance matrix by means of some iterative method, mostly maximum likelihood estimation (MLE), in such a way that the difference between the two is minimized.
- This *estimated implied covariance matrix* is subsequently compared with the *empirical sample matrix*. This comparison results in a *residual matrix*.
- These residuals are a function of the: 1) *approximation discrepancy* between model and population values, 2) *estimation discrepancy*, which is the difference between sample and population values due to sample fluctuations. (See Cudeck & Henly, 1991.)
- The hypothesis is that on the population level the approximation discrepancy is negligible, so the empirically found difference on the sample level must be only due to the estimation discrepancy.

¹ Common here refers to the common part of the variance in the measuring variables, to be distinguished from the unique part: error and item-specific variance.

- The estimation difference between the two distributions (matrices) is expressed in χ^2 , with degrees of freedom (*df*) equalling the number of covariances in the matrix minus the number of free parameters (parameters to be freely estimated by the program).
- This χ^2 should be small enough - in relation to *df* - to be only the result of chance fluctuations in the sample.

2. Problems with χ^2 : goodness-of-fit indices

The latter means that, in contrast to what is customary in statistics, χ^2 should be *non-significant* to prove a significant fit. That may be asking for trouble. Indeed, with big samples even trivial differences may be deemed significant, suggesting a poor fit, in spite of the greater representativeness of a big sample. For this reason, a number of indices of “goodness of fit” or “approximate fit” have been devised. Many of them are based on χ^2 and *df*; one is based on the residuals directly.

Over the years these indices have been investigated in numerous studies with empirical data and - more often - simulated data. Time and again they turned out to be unsatisfactory in some respect, so adapted and new ones were devised. Now many of them are available. Often reported nowadays are:

- **RMSEA** (root mean square error of approximation): based on χ^2 , *df* and *N*. It was devised by Steiger (1990). Its formula (derived from Kenny, 2010) is:

$$\sqrt{(\frac{\chi^2}{df} - 1)/(N - 1)}$$

By dividing by *df*, RMSEA penalizes for free parameters. It also rewards for a big sample size because of *N* in the denominator; in spite of this RMSEA is not notorious for sample bias.

A value of 0 indicates perfect fit. Hu & Bentler (1999) suggested $\leq .06$ as a cut-off value for good fit.

A GOF-index resembling RMSEA is Gamma-hat (Steiger, 1989). It contains a correction for the number of variables in its formula.

- **TLI** (Tucker-Lewis index, 1973), also known as **NNFI** (non-normed fit index): based on the comparison of χ^2 for the implied matrix with χ^2 for the matrix of a null-model (in practice: all measuring variables are uncorrelated). Its formula (derived from Kenny, 2010) is:

$$\frac{\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{implied}}{df_{implied}}}{(\frac{\chi^2}{df} - 1)_{null}}$$

A value of 1 indicates perfect fit. TLI is called non-normed because it may assume values < 0 and > 1 .

Hu & Bentler (1999) proposed $\geq .95$ as a cut-off value for a good fit.

TLI belongs to the class of *comparative fit-indices*: these are all based on a comparison of the χ^2 's of two models. Those indices that are not, like RMSEA, Gamma-hat and SRMR, are called *absolute fit-indices*.

- **CFI** (comparative fit index: Bentler, 1990) is based on the comparison of χ^2 for the implied matrix with χ^2 for the matrix of a null-model (all variables are uncorrelated). Its formula (derived from Kenny, 2010) is:

$$1 - \frac{(\chi^2 - df)_{implied}}{(\chi^2 - df)_{null}}$$

Values > 1 are truncated to 1, values < 0 are raised to 0. Without this “normalization”, this fit-index is the one devised by McDonald & Marsh, (1990), the **RNI** (relative non-centrality index). By subtracting *df*, it penalizes for free parameters.

Hu & Bentler (1999) suggested $CFI \geq .95$ as a cut-off value for good fit. Marsh et al. (2005, p. 295) warned that CFI has a slight downward bias, due to the truncation of values greater than 1.0.

- **SRMR** (standardized root mean square residual: Jöreskog & Sörbom, 1988). To calculate this index the residuals ($S_{ij} - I_{ij}$) in the residual correlation matrix (including the variances) are squared and then summed; this sum is divided by the number of residuals q [equals $p \cdot (p+1)/2$, where p is the number of variables], and next the square root of this mean is drawn. (S stand for correlation sample matrix, I for implied correlation matrix.)

$$SRMR = \sqrt{\frac{1}{q} \sum (S_{ij} - I_{ij})^2}$$

A value of 0 indicates perfect fit. Hu & Bentler (1999) advise a cut-off value of $\leq .08$ for good fit. Notice that χ^2 is not used to calculate SRMR.

3. Hypothesis testing approach

Testing with χ^2 is a statistical test, meant to justify generalization of some test value from sample to population. Typical for such test is an all-or-none decision: the model is confirmed or falsified. This hypothesis-testing approach has been transferred to the application of GOF-indexes in CCFA in the studies of Hu & Bentler (1998, 1999).

In order to accept correct models within this hypothesis-testing approach, the cut-off values for the indices should be lenient, but in order to reject incorrect models they should be strict. How to determine them? A controlled way is to work with simulated data: generate data in agreement with a predefined factor structure; formulate correct and a few incorrect factor models; draw a great many samples of different sizes; and see what the values of a number of the fit-indices of interest will do.

One of the convenient things about simulation studies is that the correct model is known beforehand. That provides a basis, independently of the fit-index values, to judge whether a predicted model should be rejected or accepted. To express the suitability of the selected cut-off value of the GOF-index, the percentage of rejected samples is reproduced, that is, the samples for which the fit-index value is on the "rejection side" of the cut-off value. Of course, within the hypothesis testing approach the rejection rate should be very small for correct models, and very large for incorrect models. Which percentage is to be demanded as a basis for advising a certain cut-off value is often not stated explicitly by the researchers who have conducted the simulation study. Within the hypothesis testing approach it seems reasonable (in my eyes) to demand a rate of $\leq 10\%$ for correct models and $\geq 90\%$ for incorrect models, considering that mere tossing would already lead to a rate of 50%.

Although the population is large (e.g. 500.000 cases), it is known as well. It would be wise to calculate the fit-index values per GOF-index for both correct and all incorrect models with respect to this population. It places one in a position to judge whether a selected cut-off value is too liberal anyway even before having drawn any sample.

How valid are the cut-off-values proposed by Hu & Bentler (1998, 1999)? To investigate this Marsh et al. (2004) replicated their study:

Hu & Bentler had set up a population which corresponded to the following factor structures: three correlated factors with five indicators each, with either 1) no cross-loadings (the "simple model" → 33 nonzero parameter estimates), or 2) three cross-loadings (the "complex model" → 36 nonzero parameter estimates). The misspecification in the simple model involved one or two factor correlations misspecified (set to zero), and in the complex model one or two cross-loadings misspecified (set to zero).

The population of Marsh et al. (2004) involved 500.000 cases. A number of samples of, respectively, 150, 250, 500, 1000 and 5000 cases were drawn. (The number of replications was not mentioned). MLE was applied. The dependent variable was the rejection rate per GOF-index with the cut-off value advised by Hu & Bentler (1999).

Unlike Hu & Bentler (1999) Marsh et al. (2004) calculated the population values of χ^2 and the indices.

The main findings:

- χ^2 ($p > .05$) was doing better than the indices in almost all cases. It nevertheless rejected between 8 and 14,5% of the correct models in both conditions (type I-error).
- The population values of the GOF-indices for many of the misspecified models were often on the acceptance side of the advised cut-off value. So in spite of being more stringent than had been the rule before, these cut-off values were often too liberal to detect the misspecifications. Marsh et al. (2004) characterized these models as “acceptable misspecified models”.
- If the population values are on the acceptance side of the cut-off value but very close to it, then - because of larger sampling fluctuations (Marsh et al., 2004) - the smaller samples are doing paradoxically “better” than the bigger ones, that is, lead to higher rejection rates of the misspecified models. (However, these rates stay far below the 90% rate demanded above.)
- If the population value for an incorrect model is on the rejection side of the cut-off value, as it should, but again very close to it, then the smaller samples may remain far below 90% rejection, again because of these bigger chance fluctuations. Only the larger samples will approach or reach 100%.
- SRMR shows a sample bias (higher values in smaller samples such as $N=150$ in table 1. Meade (2008), too, concluded from an analysis with ANOVA that SRMR - unlike the other fit-indices - is biased by sample size. Fan and Sivo (2007) found a bias of 13% variance for SRMR over 5 models. This is an additional reason to use only bigger samples in CCFA (the other reason being that MLE needs it).
- Cut-off values should depend on the *type(s)* of misspecification one suspects or one is interested in, because they may influence the various GOF-indices differently.
- Next, they should depend on the *degree* of misspecification one is willing to tolerate or not. For instance, should one misspecified cross-loading be considered a trivial misspecification or an unacceptable one?

Difficulty reconciling avoidance of both type 1-error and 2-error

Within the hypothesis testing approach one should avoid two types of error at the same time. One should not reject a model while in fact it is correct; this represents type-I error. Alternatively, one should not accept a model that is incorrect; that represents type-II error. To prevent the first error the cut-off value of the fit-index should be sufficiently lenient. To avoid the second error, it should be strict enough. Can these two opposing interests always be reconciled, for any sample size, in CCFA?

A study by Chen et al. (2008) offered an excellent opportunity of determining this for, at least, RMSEA.

Chen et al. (2008) investigated three models:

- 1) Three factors with five indicators each plus one cross-loading indicator each,
- 2) three factors with five indicators each plus one cross-loading indicator each,
- 3) like 1, but with an additional four inter-correlating exogenous variables that correlated with one factor, and of which two additionally correlated with the two other factors.

In model 1 and 2 the misspecification involved omitting one, two or three of the cross-loading indicators from the prediction. In model 3 the *smallest* misspecification was omitting all three cross-loadings from the prediction, the *moderate* misspecification was omitting the four additional correlations of exogenous variables with the factors, and the *large* misspecification combined the small and moderate misspecification. Parameter values were chosen so that omission of one or more of them would result in a meaningful impact on overall model fit.

Sample sizes were 50, 75, 100, 200, 400, 800 and 1000. They generated 800 samples for each of the 84 experimental conditions.

The authors investigated the rejection rates for RMSEA cut-off points ranging from 0 to 0.15 with increments of .005. From their results it was clear that avoiding type-I and type-II error at the same time was often not possible for $N \leq 200$. To give an impression of their results, I only summarize those for $N=200$ and $N=1000$ in table 1:

If $N=200$, only the moderate and severest misspecification in model 3 allow for cut-off values which are suited for both accepting 90% or more correct models and rejecting 90% or more of incorrect models. In the seven other cases no such cut-off values can be found. In three cases, even $RMSEA = 0$ is not strict enough to arrive at 90% rejection! For $N=1000$, the results are much better.

By the way, in this study, too, the population values of the misspecified models were often below the cut-off -value of .06 or even .05 (six out of nine were smaller)!

Table 1: Cut-off values of RMSEA in order to reach good rejection rates (Chen et al., 2008, figures 4 to 15)

Misspecification	N=200			N=1000		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
	To effect \geq 90% acceptance of correct models, computed RMSEA must be:					
None	$\geq .045$	$\geq .035$	$\geq .040$	$\geq .020$	$\geq .015$	$\geq .020$
	To effect \geq 90% rejection of incorrect models, computed RMSEA is restricted to be:					
Smallest	fails, $\pm 64\%$	fails, $\pm 77\%$	$\leq .030$, no	$\leq .010$, no	$\leq .015$, \pm yes	$\leq .040$, yes
Moderate	fails, $\pm 80\%$	$\leq .01$, no	$\leq .070$, yes	$\leq .025$, yes	$\leq .025$, yes	$\leq .075$, yes
Severest	$\leq .035$, no	$\leq .025$, no	$\leq .080$, yes	$\leq .055$, yes	$\leq .035$, yes	$\leq .090$, yes

The word "Fails" in a cell means that even a cut-off value of 0 is not suited to effect 90% rejection of incorrect models. Behind it the actually found rejection rate for RMSEA = 0 is printed.

"No" in a cell means that the cut-off value for 90% rejection of incorrect models is irreconcilable with that for 90% acceptance of correct models.

"Yes" in a cell means that the cut-off value for 90% rejection of incorrect models is reconcilable with that for 90% acceptance of correct models.

4. Cut-off values or benchmarks?

The preceding paragraph indicates that the hypothesis testing approach cannot be very successful:

- 1) Cut-off values are too much dependent on the *type* of misspecification one is expecting, as well as the *degree* of misspecification one is willing to accept.
- 2) Avoidance of type I and type II error at the same time is often not feasible, certainly not for the smaller sample sizes.

According to Marsh et al. (2004) a hypothesis testing approach with GOF-indices does not make sense in the first place. They state: "All GOF indices considered by Hu & Bentler were specifically designed to reflect approximation discrepancy at the population level and to be independent of sample size. However, Hu & Bentler's (1999) approach to evaluating these indices used a hypothesis-testing approach that is based substantially on estimation discrepancy and is primarily a function of sampling fluctuation. Hence there is a logical inconsistency between the intended purpose of the GOF indices and the approach used to establish cut-off values." (p. 322-323)

Besides, in practice researcher are not willing to reject the fruit of their efforts altogether as a consequence of a result falling short from significance. The advised cut-off values are therefore rarely used as sharp decision rules, but rather as rules of thumb whether or not one's model would need more or less re-specification in one or more respects. For the latter purpose other parts of the SEM output have to be investigated.

Even with the pretention so tempered, there should be some consensus about what constitutes an excellent fit, a good fit, a fair fit and a poor fit. This asks for *benchmarks* of the fit-indices on the basis of which the values can be rated on a scale from excellent to poor. These benchmarks do not ask for an unconditional rejection or acceptance of the model, but serve as the basis for a more informal evaluation of the model.

If simulation studies are used to determine such benchmark values, it does not make sense anymore to demand rejection rates of 90% or 10% (as the case may be). Instead one can use the values that effectuate a 50% rejection rate for big samples (N=1000 or more), because these are close to the population values. A confidence interval per value per sample size, however, would be convenient then.

A table with benchmarks as fancied above, however, will only be possible and useful if the GOF-indices can be relied on. They *may* vary with type of misspecification and *should* vary with the degree of misspecification, but it is essential that no other factors would complicate the picture any further. Unfortunately, there is evidence of the opposite.

5. Undesirable influence of degree of unique variance: Browne et al. (2002)

Browne et al. (2002) wrote a theoretical article which seems devastating to a great number of current fit-indices. The authors had been consulted by colleagues who were puzzled by an unwelcome phenomenon: when working with very reliable and homogeneous measurements, as in biological or physical research, χ^2 and χ^2 -based fit-indices often indicated a poor fit of the model while the fit was in fact excellent, as could be judged from inspecting the residual matrix. On the other hand, this very residual matrix could result in favourable values of these fit-indices if it came from the comparison of a model with data containing more unique variance.

For an example the authors discussed data that came from a clinical trial of the efficacy of a psychological intervention to reduce stress and improve health behaviours for women with breast cancer (Andersen et al., 1998). One focus was on two types of biological responses of the immune system to the intervention: 1) natural killer cell lysis, 2) response of these cells to recombinant interferon gamma. For each of them there were four replicates, differing in the ratios of effector to target cells, which were measured by means of peripheral blood leukocytes in blood samples. These 2 x 4 replicates were treated as indicators of two corresponding and related but distinct factors.

Because the measures were biological and replicates, one may expect a high reliability and homogeneity of them, which means a small unique variance. Indeed, on the face of it the correlation matrix showed two distinct clusters of highly inter-correlating variables: 1 to 4 (mean inter-correlation .853) and 5 to 8 (mean inter-correlation .960). A MLE CFA was carried out hypothesizing these two factors.

In spite of the clear picture (the residual matrix only contained values very close to zero) and the small sample (N=72), χ^2 was highly significant, urging rejection of the model, and so did RMSEA and three other absolute GOF-indices. The comparative fit-indices RNI and NFI did better, but were below the advised cut-off values. Only SRMR, being 0.2, indicated strongly acceptance of the model, in line with the fact that this index is based on the residuals alone.

The investigators went on to frame a correlation matrix consisting again of two clusters, but with much lower inter-correlations between the variables, especially within the first "cluster" (mean inter-correlation only .136). The second cluster inter-correlated on the average .608: in itself quite high but much below the original cluster. MLE CFA with a model of two related factors should yield the same residual matrix as above, and it did. Now χ^2 was highly non-significant, indicating perfect fit, like all GOF-indices, all this in spite of the first cluster being unconvincing. SRMR - of course - was again .02.

Browne et al. (2002) reasoned that this is caused by the fact that χ^2 as a result of CCFA conducted with MLE or generalized least squares (GLS), is affected not only by the residual matrix but equally by the sample matrix. This is because χ^2 is based on an *estimate* of the approximation discrepancy between implied matrix and population matrix. So it can be influenced by the degree of unique variance in the measuring variables. If the unique variance is very small, this will give rise to very small eigenvalues of factors beyond the first two that had been predicted. And these small eigenvalues, in turn, will lead to a very high χ^2 , because of peculiarities of the formula's used to calculate it. Since all but one GOF-index are based on χ^2 , they will indicate poor fit too. Only SRMR will escape this unfortunate fate.

Browne et al. concluded that the other fit-indices are measuring *detectability of misfit* rather than misfit directly. In other words, high statistical power of a test may easily leads to rejection of a fit, whereas mediocre statistical power leads to acceptance of it (see also Steiger, 2000). In social science research one is rarely confronted with this undesirable phenomenon, because most of the measures are of mediocre quality. Nevertheless they may vary from rather reliable and homogenous to rather poor. It would be ironic if only the latter would scoop with the honour of seemingly good fit.

Hayduk et al. (2005)

The study of Browne et al. (2002) had a remarkable sequel in a reanalysis of the data by Hayduk et al. (2005). These investigators had reasons to believe that the two factor model of Browne et al. had not been correct and that two "progressively interfering" factors should have been included in the model, which were thought to affect the respective two clusters of measurements. This alternative model appeared to fit very well the data, so that now χ^2 was far from significant in spite of the minute unique variance. Eliminating one of the progressively interfering

factors from the model spoiled the results. The same held for a few other variations on the model.

Mulaik (2010) proposed a further refinement of the model, which involved a correlation of .40 between each of the progressively interfering factors and the clusters of measures they were supposed to affect. This refinement did not improve the already good fit, but was theoretically more plausible.

Conclusion

The lesson to be learned from this is that one should not be *too* eager to resist indications of ill fit just because the unique variance is so small. A serious search for an alternative model should always be undertaken (Hayduk et al., 2007; McIntosh, 2007). That does not mean, however, that the indications of ill fit should *never* be considered trivial. But the latter cannot mechanically be concluded from a very small unique variance; it has to be defended with substantive arguments.

A much more important lesson to be learned, however, is that a non-significant χ^2 and favourable GOF-index values can absolutely not be trusted, because these may merely be promoted by a large unique variance! This was the main message of Browne et al. (2002), more than the ill fit in case of low unique variance.

Mistrust is warranted the more because the findings of Browne et al. do not stand alone, as witnessed by §6 and 8.

6. Undesirable influence of degree of unique variance: Miles & Shevlin (2007)

Miles & Shevlin (2007), by way of comment on Barrett (2007), did a study on the effect of reliability of measurement on χ^2 , RMSEA, SRMR and a number of *comparative* fit-indices, in order to demonstrate the latter's greater robustness against the paradoxical influence of reliability. They set up a population, corresponding to two related factors (correlation .3) with four high-loading indicators each, and two minor factors, the one loading very low on two indicators of one major factor, the other loading very low on two indicators of the other major factor. They "drew" an imaginary sample of N=500, perfectly mirroring the population values. The tested model was a 2-factor structure leaving out the two minor factors.

With a perfect reliability of 1, the disturbance by the minor factors was enough to have the model rejected by χ^2 ; RMSEA indicated doubtful fit, but the other fit-indices indicated good fit, as they were hoped to do. However, with a reliability of 0,8, both χ^2 and RMSEA now suggested good fit (the comparative indices and SRMR continued to indicate good fit). So, as in the study of Browne et al., (2002), reliability appeared to have a diminishing effect on χ^2 and RMSEA, but not on SRMR. In the present study, however, the comparative fit-indices did well, more convincingly so than in the study of Browne et al.

Miles and Shevlin (2007) did a second study with this model, but now left out the two minor factors while the correlation between the two factors was increased to .5. This time the model tested was a severely misspecified one: a 1-factor model. If the reliability was .8, χ^2 indicated misfit, as it should, and so did all the fit-indices. However, if the reliability was decreased to a meagre 0,5, χ^2 suddenly indicated good fit, and so did RMSEA, and - in spite of not being χ^2 -based - SRMR. The comparative fit-indices were hoped to be robust against the spurious influence of reliability on χ^2 , but only two of them were: NFI (normed fit-index), and RFI (relative fit-index: Bollen, 1986). Three of them did not indicate misfit: CFI, TLI, and IFI. (An explanation will be given in §10.)

Saris, Satorra & van der Veld (2009) reported a complementary study: They devised population data for a one factor model. Strictly speaking, the data could be more precisely explained by a two factor model in which the two factors correlated 0,95. This difference they considered trivial, so the one factor model should be deemed acceptable. They drew an imaginary perfectly representative sample (N=400) from this population and calculated χ^2 , RMSEA, CFI and SRMR. If factor loadings were 0,85 or 0,9 - which presupposes high reliability and low specific variance -, then the values of χ^2 and RMSEA led to rejection of the one factor model. Only CFI (in line with the expectations of Miles & Shevlin, 2007) and SRMR (not being χ^2 -based) had values pleading for acceptance of the model under all loadings (0,7 till 0,9). If the loadings were decreased to $\leq 0,80$, χ^2 and RMSEA now indicated acceptability of the model.

7. Minor and major model error

The preceding two paragraphs raise the question: what kind of misspecification regarding the number of factors is to be considered trivial? The one in the Browne et al. (2002) study was less trivial than it seemed on first glance (Hayduk et al., 2005), and with respect to the Saris et al. (2009) study it can be argued that a factor correlation of .95 is not trivial if (but only if) the factor loadings are as high as .85 or .90. On the other hand, as McCallum (2003) contended, models can never be perfect since they are always simplifications of reality. So models always contain minor model error. If one executes an exploratory factor analysis, there will always be a number of factors which still contribute to the total explained variance, but so little that it mostly does not make sense to take them into account; they often cannot be well interpreted. Such minor factors should be excluded from the model in CFA as well.

It could, therefore, be argued that χ^2 and the fit-indices should be robust against this type of minor model error. Stuiwe (2007), considering this, included such error in her simulation study. This was done by introducing 9 unmodeled, minor factors in a part (2/3) of the data, explaining either 10% or 20% of the variance (remember: 12 items, 3 factors).

Stuiwe's (2007; see also Stuiwe et al., 2008) study was done with continuous data, simulating a 12-item questionnaire with three subtests of four items each. The misspecification here was an incorrect assignment of one or more items to subtests (three levels of gravity). Note this is a more serious misspecification than one or two unpredicted cross-correlations.

Independent variables were: 1) the unique variance (25%, 49% or 81%), 2) the amount of minor model error, 3) the correlations between the three factors (0.0, 0.3 or 0.7). These independent variables were *not* specified parameters, that is, were not part of the tested model. The sample size was another independent variable: 50, 100, 200, 400 and 1000 cases. For each combination of conditions 50 samples were drawn.

The dependent variables were: the percentage of a) accepted correct assignments, and b) rejected incorrect assignments. Acceptance of the assignments depended on the p-value ($> .05$) of ML χ^2 , plus on different cut-off values for three fit-indices, and for three combinations thereof².

Stuiwe (2007, her figure 5.2) found that introducing 10% model error had a very strong effect on χ^2 ($p > .05$), resulting in a fall of the rejection rate of correct models from 90% to 60%. Ten percent model error had also a rather strong effect on RMSEA, leading to much more rejection of correct assignments with the two strictest cut-off values, $\leq .03$ and $\leq .06$ (her figure 5.2). CFI was robust against 10% model error but not against 20%. SRMR was robust even against 20% model error. However, Meade (2008) found that $SRMR \leq .08$ was completely unsuited to detect even a major unmodeled factor (his figure 8).

So SRMR may be *insensitive* to any unmodeled factors, major or minor. χ^2 and RMSEA, on the other hand, could be considered too sensitive. CFI's sensitivity to minor unmodeled factors may be just good. So if one agrees with MacCallum (2003), then χ^2 ($p > .05$) and RMSEA should not be used.

Hayduk does not agree. He notices that " χ^2 locates more problems when N is larger [and the unique variance smaller, PP], so that some people blame chi-square (the messenger) rather than the culprit (probably the model)" (Hayduk on SEMNET, 3 June 2005). He argues that we should profit from χ^2 's sensitivity to model error and take the rejection of our model as an invitation for further investigation and improvement of the model (see Hayduk et al., 2007.)

However, if we accept the invitation, the model error found may well appear to be trivial indeed.

8. Undesirable influence of degree of unique variance: Stuiwe (2007)

Stuiwe (2007) investigated unique variance as an explicitly independent variable. She varied the percentage of unique variance: 25, 49 and 81% respectively. Her results are presented in her

² Here she followed a suggestion by Hu & Bentler, 1999). The combinations are ignored in the present discussion.

figures 5.1 (correct models, including those with 10% and 20% minor model error), factor correlation 0, .3 and .7 combined) and 5.4 (incorrect models; factor correlation .3³). However, these are not easy to interpret, certainly not when one tries to judge them in combination, so I summarize the results for RMSEA SRMR and CFI in separate tables, with inferred population values of correct and incorrect models. These population values were inferred from the rejection rates at different cut-off values in combination with sample size. These are, of course, not very exact but will facilitate comprehension of the data. It may be helpful to download Stuive's book (open access) to have a look these figures at the same time (chapter 5).

Chi square:

- With a small unique variance of 25%, many correct models were rejected on the basis of ML χ^2 ($p > .05$). The acceptance rate was even below chance level. Remember, however, that correct models with 0% minor model error *did* lead to an acceptance rate of 90%; see §7. The “correct” models tested here included one third with 0% model error, one third with 10% minor model error, and one third with 20% minor model error.
- If the unique variance was 49% or 81%, ML χ^2 ($p > .05$) suggested a much better fit. This paradoxical because a high unique means lower correlations in general, rendering the matrix more diffuse.
- The rejection of incorrect models was all-right.

RMSEA:

Table 2: Estimated population values, inferred from Stuive (2007), figures 5.1 and 5.4.

RMSEA	correct model				incorrect models				
	unique variance	estimated population value	cut-off value	sample size	accept. rate	estimated population value	cut-off value	sample size	reject. rate
25%		≈ .08	≤ .10	N ≥ 50	85-90%	≈ .12	> .10	N ≥ 50	95-100%
49%		≈ .05	≤ .06	N ≥ 100	90-100%	≈ .10	> .08	N ≥ 50	90-95%
81%		≈ .02	≤ .03	N ≥ 400	90-100%	≈ .03	> .03	N ≥ 400	90-95%

Column estimated population value: inferred from the different rejection rates.

Column cut-off value: if the calculated fit-index obeys the condition mentioned, the rate of correct judgments will be 85-100% for N = 400 and N=1000 at least.

Column sample size: 85-90% correct judgment rate is realized for the sample sizes mentioned.

Remember, the calculated RMSEA values will be higher as a consequence of minor model error, deliberately introduced in two thirds of the data.

In table 2 the results for RSMEA are summarized. With regard to *correct* models, the (estimated) population value of the *correct* model decreased dramatically as a function of an increasing unique variance:

- If the unique variance was only 25%, it needed a lenient cut-off value RMSEA ≤ .10 to realize an 85-90% acceptance rate (granted, for all sample sizes).
- With a unique variance of 49%, however, this acceptance rate could already be achieved with the prescribed cut-off value RMSEA ≤ .06 (except for N=50).
- With a unique variance of 81% even the strict cut-off value RMSEA ≤ .03 led to 90% acceptance of *correct* models only for N=1000 and N=400.

The (estimated) population of *incorrect* models also decreased dramatically with an increasing unique variance, especially from 49% to 81%:

- With a small unique variance of 25%, full rejection of *incorrect* models was guaranteed already with the lenient cut-off value RMSEA > .10.
- With a unique variance of 49%, RMSEA became smaller. Now the RMSEA cut-off value had to be >.08 for 90% rejection.
- With a unique variance of 81%, however, RMSEA became so small, that only the strict cut-off value RMSEA > .03 realized a rejection rate of 90%, and such only for only N=400 and N=1000.

³ The results for a factor correlation .3 were very close to those for factor correlation 0.

SRMR:

See table 3: The rejection rates of SRMR in case of *correct* models indicated that the degree of unique variance did not matter: the population values stayed constant and small.

Table 3: Estimated population values, inferred from Stuive (2007), figures 5.1 and 5.4.

SRMR	correct model				incorrect models			
	unique variance	estimated population value	cut-off value	sample size	accept. rate	estimated population value	cut-off value	sample size
25%	≈ .04	≤ .06	N ≥ 200	90-100%	≈ .12	> .10	N ≥ 200	100%
49%	≈ .04	≤ .06	N ≥ 200	95-100%	≈ .09	> .08	N ≥ 200	90-100%
81%	≈ .04	≤ .06	N ≥ 200	90-100%	≈ .04	fails	-	0-15%

See table 2.

Fails: even SRMR ≤ .06 is too lenient for rejection above chance.

Column sample size: 85-100% acceptance rate is for N ≥ 200 (N = 50 and n = 100 are ignored because of the sample size bias, see §4).

Stuive also investigated the cut-off value .09, not mentioned in the table.

Remember, the calculated SRMR-values will *not* be affected by the minor model error.

The rejection rate for *incorrect* models showed a different picture. Now, as was the case with RMSEA, the larger the unique variance, the smaller SRMR, and vice versa:

- 25% unique variance produced a SRMR larger than .10, so that even the most lenient cut-off value (> .10) led to 100% rejection of *incorrect* models, a welcome result.
- However, with 81% unique variance the opposite was true: now SRMR became clearly smaller than .06, so that all cut-off values led to an acceptance of almost all *incorrect* models.
- 49% unique variance produced a picture in between these two extremes. Obviously the population value of SRMR must have been about .09 in this case. As a result, if the cut-off value was > .08, then the rejection rate for *incorrect* models was about 85% for N=1000 and about 90% for N=200 and N=400 (because of the larger sample fluctuations of smaller samples).

So in spite of not being χ^2 -based, for *incorrect* models SRMR is not guarded against the impact of unique variance.

CFI:

The results for CFI are summarized in table 4.

Table 4: Estimated population values, inferred from Stuive (2007), figures 5.1 and 5.4.

CFI	correct model				incorrect models			
	unique variance	estimated population value	cut-off value	sample size	accept. rate	estimated population value	cut-off value	sample size
25%	≈ .965	≥ .94	N ≥ 100	85-90%	≈ .90	< .94	N ≥ 50	100%
49%	≈ .975	≥ .95	N ≥ 100	85-95%	≈ .91	< .94	N ≥ 50	95-100%
81%	≈ .975	≥ .95	N ≥ 400	85-95%	≈ .93	< .94	N ≥ 400	90-100%

See table 4.

Column sample size: 85-100% correct judgment rate is realized for these sample sizes.

Remember, the calculated CFI-values will be slightly lowered as a consequence of minor model error.

For *correct* models holds:

- If the unique variance is 25%, then CFI is a little lower (suggesting lesser fit) than when it is 49% or 81%, so the lenient cut-off value ≥ .94 is needed to realize an 85-90% acceptance rate.
- If the unique variance is 49%, the calculated CFI raises enough to score an 85-95% acceptance rate at the cut-off value ≥ .95.
- If the unique variance is 81%, the population value is about the same, but the acceptance rate is more dependent on sample size: for N=1000 the strict CFI ≥ .96 scores 95%, for N=400 it needs the cut-off value CFI ≥ .95 to score 85% acceptances. For N ≤ 200, even the lenient CFI ≥ .94 results in only 80% acceptance.

As to *incorrect* models, there was a very good rejection rate for 25% and 49% unique variance, while 81% unique variance asked for N=400 and N=1000 to achieve such a high rejection rate. (This is *not* a paradoxical picture.)

Conclusion

In line with the findings and reasoning of Browne et al. (2002), Miles and Shevlin (2007), and Saris et al. (2009), ML χ^2 ($p > .05$) lost much of its statistical power with increasing unique variance, meaning acceptance of both correct models (partly containing, however, "minor" model error) and incorrect models.

Being χ^2 -based in a very direct way, the same held for RMSEA, albeit in a much more attenuated degree.

CFI was rather robust against the undesirable influence of unique variance, in line with the expectations of Miles and Shevlin (2007) who argue that the effect of unique variance on χ^2 of the predicted model is more or less compensated by its effect of χ^2 of the null-model.

SRMR, however, appeared to lose its power to detect incorrect models as a consequence of high unique variance (81%, not 49%), even more so than RMSEA, in spite of not being χ^2 -based. This finding is in line with that of Miles and Shevlin (2007) with respect to their test of a false one-factor model when the reliability was only 0,5.

9. The influence of level of inter-correlations in the matrix

SRMR

Why should SRMR assume lower values as a consequence of high unique variance in case incorrect models are tested? thereby promoting type II-error? The following explanation is offered:

- A large average unique variance depresses the common variance between the variables and consequently implies a correlation matrix in which the clusters of variables have a relatively low mean inter-correlation. A small unique variance implies the opposite.
- If the empirical correlation matrix has small correlations on the average, then the implied correlation matrix will also have small ones; in case of high correlations the opposite is true.
- Subtracting two matrices with generally low values will often result in smaller residuals than subtracting two matrices with generally high correlations.
- As a consequence SRMR will be smaller (implying better fit) if the correlations in the empirical matrix are depressed by a high unique variance.

This can also be shown mathematically: Let the correlation between variable i and j in the sample be called S_{ij} , and the corresponding correlation in the implied matrix I_{ij} . Then the residuals are $S_{ij} - I_{ij}$. If both correlations decrease with a factor 0,67 because of a large unique variance, then the new residuals will be $0,67 \cdot S_{ij} - 0,67 \cdot I_{ij} = 0,67 \cdot (S_{ij} - I_{ij})$. Squaring such a residual will give $0,67^2 \cdot (S_{ij} - I_{ij})^2$. Summing these squared residuals will yield:

$$\Sigma[0,67^2 \cdot (S_{ij} - I_{ij})^2] = 0,67^2 \cdot \Sigma(S_{ij} - I_{ij})^2$$

There are q residuals. $q = p \cdot (p + 1) / 2$, where p is the number of variables. Averaging these q summed squared residuals will yield:

$$\frac{0,67^2 \cdot \Sigma(S_{ij} - I_{ij})^2}{q}$$

Drawing the square root, which is SRMR, produces:

$$0,67 \cdot \sqrt{\frac{1}{q} \Sigma(S_{ij} - I_{ij})^2}$$

So SRMR is proportionally lower, suggesting better fit. For instance, 0.11 will become .074.

The effect is, of course, much greater for incorrect models than for correct models, because in the former case there are bigger residuals to reduce.

This proportionately lower residual matrix - in case of incorrect models - will also affect χ^2 and the GOF-indices based on it, apart from the effect via the eigenvalues, demonstrated by Browne et al. (2002).

Undesirable effect of high factor correlation on comparative indices

Stuive (2007) varied the factor correlation (0, .3 and .7) in her simulation study, not as a predicted parameter which could be misspecified, but as an independent, free parameter which might affect the values of the fit-indices.

From her figure 5.3 it can be learned that the acceptance rate for *correct* assignments of RMSEA and χ^2 remained the same, irrespective of factor correlation being 0, .3 or .7. The rate of CFI, however, improved for factor correlation moving from 0 to .3, and even more so from .3 to .7. This implies that CFI is artificially augmented by a high factor correlation, leading to acceptance of the model. A similar but much smaller (probably insignificant) effect was visible for SRMR with cut-off values .06 and .08.

Why has a high factor correlation this effect on CFI? Because a high factor correlation implies higher correlations between all variables on the average. For instance, in Stuive's factor structure there were 3 clusters of 4 variables each, implying there are 66 correlations of which 3 X 6 within-cluster correlations and 48 cross-correlations. Say the mean inter-correlation per cluster is .50. If the factors correlate 0, then the cross-correlations may on the average amount to 0. If the factors correlate .3, the cross-correlations may average .10. If the factors correlate .7 the cross-correlations may be on the average .30. In the first case the average correlation of the total observed matrix is $9/66 = 0,136$. In the second case this average correlation is $13,8/66 = 0,209$. In the third case it is $23,4/66 = 0,355$.

Why does an empirical matrix with a high average inter-correlation promote high CFI's? As Kenny (2010) argues it can be deduced from the formula for CFI:

$$1 - \frac{(\chi^2 - df)_{\text{implied}}}{(\chi^2 - df)_{\text{null}}}$$

Usually the null model in comparative fit-indices involves the (unlikely) situation that all variables correlate zero. So if the correlations between the variables in the empirical matrix are on the average high, then χ^2 for the null-model will be very large, because it rests on the difference between the empirical matrix and the null-model matrix. As a consequence the fraction above will assume a lower value, and this means that CFI will be higher, suggesting better fit. If the correlations between the variables in the empirical matrix are on the average low, the opposite holds. This argument was already raised by Rigdon (1996).

The same holds for TLI and probably all other comparative indices, as long as they start from the null-model mentioned above.

10. Interaction between unique variance and factor correlation

The results of Stuive (2007) mentioned in §8 held when the factors correlated .3 or 0. However, if the factors correlated .7, a large unique variance had an even more devastating effect on the power of RMSEA and SRMR to reject *incorrect* models (Stuive's figure 5.5), while CFI, too, was spared no more, on the contrary:

- Chi square: With 25% and 49% unique variance χ^2 ($p > .05$) urged rejection of all incorrect models ($N \geq 200$), but with 81% unique variance the rejection rate decreased to 85% ($N = 1000$), whereas for lower sample sizes the rejection rate was much below chance level. (Besides, remember that with 25% unique variance too many *correct* models were rejected.)

Table 5: Estimated population values for three GOF-indices in case of incorrect models when factors correlate .7, inferred from Stuive (2007), figures 5.5.

unique variance	RMSEA			SRMR			CFI		
	estim. popul. value	cut-off value	sample size	estim. popul. value	cut-off value	sample size	estim. popul. value	cut-off value	sample size
25%	≈ .11	> .10	$N \geq 50$	≈ .08	> .06	$N \geq 200$	≈ .950	< .96	-
49%	≈ .08	> .06	$N \geq 200$	≈ .05	fails	$N \geq 200$	≈ .975	fails	-
81%	≈ .02	fails	-	≈ .04	fails	$N \geq 200$	≈ .985	fails	-

Explanation, see table 4 to 6.

Fails: even the strictest cut-off values are too lenient for rejection above chance, or any rejection.

- **RMSEA:** The results for *incorrect* models are summarized in table 5: If the unique variance is 81%, even the strict $RMSEA \leq .03$ has a rejection rate below chance level. This means the population value of RMSEA must have become very small (.02 or so), suggesting excellent fit. For a unique variance of 25% and 49% the rejection rates are normal.
- **SRMR:** See table 5: If the unique variance is 25%, then the population value of SRMR must be about .08, because $SRMR > .08$ results in 50% rejection for $N = 1000$ (no sensitivity for minor model error), and $SRMR > .06$ in 100% rejection. However, if the unique variance is 49%, then the population value of it must be .05 at best, because even the strict $SRMR \leq .06$ is too lenient to reject above chance level. If the unique variance is 81% the population value of SRMR must be .04 or less, because the rejection rate for incorrect models is zero, even with $SRMR < .06$.
- **CFI:** See table 5: With 25% unique variance the population value must be somewhat below .96 because the rejection rate for incorrect models approaches 90% only for $CFI > .96$. With 49% and 81% unique variance the population value must be much above .96, because the cut-off value $CFI > .96$ leads to a rejection rate much below chance level, the more so for the bigger sample sizes (because these have smaller sample fluctuations).

Explanation

Why does the combined influence of high factor correlation and high unique variance decrease RMSEA and SRMR, thus promoting type II-error? As we saw a high unique variance lowers the inter-correlations within the clusters of variables, whereas a high factor correlation raises cross-correlations, thereby reducing the distance between misspecified correlations in the implied matrix and their counterpart in the empirical matrix. So the combination of the two reduces this difference even more, even in case of a severely misspecified model. If this is hard to follow, the reader may profit from a look at the simplistic tables in:

<http://home.kpn.nl/p.prudon/Tablesuvfc.pdf>.

These smaller residuals will also affect χ^2 and the GOF-indices based on it, apart from the effect via the eigenvalues, demonstrated by Browne et al. (2002).

Why CFI is not spared this time, as Shevlin and Milner (2007) expect the comparative indices to do? We have seen that low variable correlations (implying a high unique variance) lower χ^2 for the null model as well as for the implied model. If the factors correlate moderately (.3), the first effect is compensated for by the second effect, resulting in a CFI sufficiently low to reject incorrect models. However, if the cross-correlations in the matrix are well above zero because of high factor correlations, then the average correlation in the matrix will be relatively high, which now spoils this compensation (see table 8 in the link above). So CFI is raised again, the compensation is spoiled, and now it is no longer suited to detect incorrect models.

That is the reason why CFI, TLI and IFI in Miles & Shevlin's study (2007) no longer indicate misfit when a factor correlation of .5 was combined with a reliability of .5. That NFI and RFI continued to do so cannot be relied upon without further investigation.

In sum

From Stuive's simulation study it can be concluded that the more messy the factor structure the better fit is suggested by χ^2 , CFI, SRMR, and to a lesser degree by RMSEA, for all rather severely misspecified models which were typical for her study.

11. One more spurious influence on some GOF-indices

Brevik & Olsson (2001) reasoned and found that RMSEA tends to favour models that include more variables and constructs over models that are simpler, due to the parsimony adjustment built-in. The same problem seems to hold for SRMS (Anderson & Gerbing, 1984; Brevik & Olsson, 2001; Kenny & McCoach, 2003).

These conclusions are in line with those of Meade (2008). Meade investigated the suitability of various fit-indices to detect an unmodeled factor present in the data. He found that $RMSEA \leq .06$ was strict enough to detect the unmodeled factor, only if the factors consisted of four items. However, if the factors consisted of eight items each, $\leq .06$ appeared too liberal (figure 9 in his article). $SRMR \leq .08$ was not strict enough to detect the unmodeled factor. CFI did all-right, and so did TLI, RNI and IFI. $\chi^2 (p > .05)$ showed itself useless.

Such a direct and spurious effect is to be distinguished from the fact that a prediction error should be seen in proportion to the total number of variables. A misspecified cross-correlation in a 3-item subscale is a more serious error than one in a 5-item subscale, as is demonstrated in table 2 based on the data of Chen et al. (2008): model 1 vs. model 2 respectively.

The bias of the number of variables was also substantiated by Fan & Sivo (2007), who were primarily interested in a possible influence of *type of model* on GOF-indices. They found that - with 15 and 12 observed variables (first three models) - RMSEA discriminated well between the correct specification and the two levels of misspecification, no matter the type of model. The same held for "Gamma-hat" and "McDonalds Centrality index" (not discussed in this paper). However, with 4 or 6 observed variables (model 4 and 5 respectively), RMSEA-values were much higher (especially for the misspecified models) and to a greater degree affected by the type of model. Now only Gamma-hat survived the test. Remember (see §2), this index is similar to RMSEA, but contains a correction for the number of variables in its formula. Next best were CFI and RNI, as well as Bollen's Delta-2 (not discussed in this paper).

12. Conclusion

It seems there is a fundamental problem with χ^2 as a kind of similarity coefficient for comparing complete correlation matrices, apart from the significance complication with big samples. This problem is: if the unique variance is small (homogeneous and reliable measurements) correct models may be rejected on the basis of merely trivial model error⁴, whereas with increasing unique variance goodness of fit is overestimated, which is a much worse error. This bias is transferred to the fit-indices based upon it, which is all but one. Especially RMSEA mirrors the effect unique variance on χ^2 . The comparative fit-indices are less affected because the imprecision in χ^2 for the implied matrix is more or less compensated for by the imprecision in the χ^2 for the null-model.

More or less implies that the compensation is not very stable. Indeed, if the factors correlate high, then some, if not all, of the comparative GOF-indices equally overestimate fit, because high factor correlation implies a higher inter-correlation of the measuring variables, and as a consequence a bigger distance from the null-model, resulting in higher values of CFI, TLI and so on.

The only GOF-index that is not χ^2 -based, SRMR, appeared no less immune for the impact of unique variance. This was explained by the average level of correlations in the implied and empirical matrix. If both are rather low as a consequence of a high unique variance, then the residual matrix will consist of smaller residuals than will be the case if the average correlation would have been rather high. Besides, very small residuals do not always imply that the model fit well, as the study of Hayduk et al. (2005) demonstrated. This effect of unique variance on the residual matrix will affect the χ^2 -based indices along a second route, undermining them even more.

The small discriminating power of the fit-indices becomes even more obvious if a moderate or high unique variance is combined with a high factor correlation (.7 was enough for the effect). Then, their performance with respect to incorrect models tumbles down.

The utmost consequence of the paradoxical relation between unique variance and goodness of fit was drawn by Marsh et al. (2004), repeated in Marsh et al. (2005, p.318), where they stated: "... assume a population-generating model in which all measured variables were nearly uncorrelated. Almost any hypothesized model would be able to fit these data because most of the variance is in the measured variable uniqueness terms, and there is almost no co-variation to explain. In a nonsensical sense, a priori models positing one, two, three, or more factors would all be able to "explain the data" (as, indeed, would a "null" model with no factors). The problem with ... this apparently good fit would be obvious in an inspection of ... all factor loadings and factor correlations ... close to zero." On p. 329 they add: "... scientific evaluation of unexplained variance can be as important a criterion as GOF in some applications."

An additional troubling influence is the number of variables in a model: more variables increase the chance that χ^2 , RMSEA and SRMR will indicate a good fit of misspecified models. Gamma-hat and the comparative indices, however, are much less affected (see §11).

⁴ Whether the error is really trivial has to be carefully investigated and defended with substantive arguments. The arguments of Hayduk et al. (2007) should be kept in mind.

All this makes it unwarranted and useless to formulate guidelines as to what values for what GOF-indices indicate a good fit. The difficulties in applying and interpreting the GOF-indices made Barrett (2007) call for their abandonment: "I would recommend banning *ALL* such indices from ever appearing in any paper as indicative of 'model acceptability' or 'degree of misfit' (p. 821). Others felt there could be still a place for them (e.g. Goffin, 2007; Markland, 2007; Mulaik, 2007). On the basis of this review such optimism seems unwarranted.

Even if one day a few real good indices would have been devised within the present tradition, recommendations about which values constitute good fit should depend on the type of misspecification one is interested in (primary loadings, cross-loadings, factor correlations, disturbance terms, causal relation with exogenous variables, number of factors, etc.), because each will probably have a different impact. It should be made clear beforehand *what* constitutes "trivial model error" and how much of it one is willing to accept. In specifying this multitude of advised fit-index values it would also be necessary to find a balance between avoiding both underestimating correct models and overestimating incorrect models.

From the above it will be also clear that even perfect GOF-indices have no diagnostic value whatsoever. They may only be useful as tests, telling whether a model needs further investigation and improvement or can be unconditionally accepted, but will never tell where possible mistakes lie or what kind of improvements may be needed. For this purpose other parts of the output have to be inspected. This made McIntosh (2007, p. 859), likewise commenting on Barret (2007), conclude that GOF-indices "offer little value-added in SEM practice, given the wide variety of available methods for performing detailed model assessments". However, McIntosh left "the issue of whether AFIs should be completely abandoned to future research".

The question how well the job of detecting and correcting the specific errors in CCFA of the measurement model is done by LISREL and other SEM-programs deserves a second review.

The only use of the current fit-indices that is not challenged by the foregoing is comparative testing of alternative models on the same data. But this job could also be done by any GOF-index, yet to be devised, which is less vulnerable to spurious influences. What kind of GOF-index that could be, that question would demand a third article.

References

- Anderson, B.L., Farrar, W.B., Golden-Kreutz, D., Katz, L.A., MacCallum, R.C., Courtney, M.E. & Glaser, R. (1998). Stress and immune response after surgical treatment for regional breast cancer. *Journal of the National Cancer Institute*, 90, 30-36.
- 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.
- Barrett, P. (2007). Structural equation modeling: adjudging model fit, *Personality and Individual Differences*, 42 (5), 815–824. Available online 7 Nov. 2006.
- 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.
- Bollen, K.A. (1986). Sample size and Bentler & Bonett's nonormed fit index. *Psychometrika*, 51, 375-377.
- Bollen, K.A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research*, 17, 303-316.
- Breivik, E. & Olsson, U.H. (2001). Adding variables to improve fit: the effect of model size on fit assessment in LISREL. In: Cudeck, R., Jöreskog, K. G., du Toit, S. H. C., and Sörbom, D. (2001). *Structural equation modeling: present and future : a festschrift in honor of Karl Jöreskog*. (pp. 169-194).
- Browne, M.W., MacCallum, R.C., Kim, C-T., Andersen, B.L., and Glaser, R. (2002). When fit indices and residuals are incompatible, *Psychological Methods*, 7, 403-421.
- Chen, F., Curran, P.J., Bollen, K.A., Kirby, J. & Paxton, P. (2008). An empirical evaluation of the use of fixed cut-off points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36 (4), 462-494.
- 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.
- Fan, X. & Sivo, S.A. (2007). Sensitivity of fit Indices to model misspecification and model types. *Multivariate Behavioral Research*, 42 (3), 509-529.
- Goffin, R.D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences*, 42 (5), 831-839. Available online 7 Nov. 2006.
- Hayduk, L.A., Cummings, G.C., Boadu, K., Pazderka-Robinson, H. & Boulianne, S. (2007). Testing! testing! one, two, three – Testing the theory in structural equation models! *Personality and Individual Differences*, 42 (5), 841-850.
- Hayduk, L.A., Pazderka-Robinson, H., Cummings, G.C., Levers, M-J.,D. & Beres, M.A. (2005). Structural equation model testing and the quality of natural killer cell activity measurements. *BMC Medical Research Methodology*, 5:1. Open access via <http://www.biomedcentral.com/1471-2288/5/1>
- Hoekstra, R.A., Bartels, M., Cath, D.C. & Boomsma, D.I. (2008). Factor structure, reliability and criterion validity of the autism-spectrum quotient (AQ): a study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders*, 38 (8), 1555-1566. Published online 2008 February 27.
- Hu, L.T. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Hu, L.T. & 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.
- Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 31, 183-202.
- Jöreskog, K.G. & Sörbom, D. (1988). *LISREL 7 - A guide to the program and applications* (2nd ed.). Chicago: SPSS.
- Kenny, D.A. (2010). <http://www.davidakenny.net/cm/fit.htm>
- Kenny, D.A. & McCoach, D. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10, 333-351.
- MacCallum, R.C. (2003). Working with imperfect models, *Multivariate Behavioral Research*, 38, 113–139.
- Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, 42 (5), 851-858. Available online 7 Nov. 2006.
- Marsh, H.W., Hau, K.-T. & Wen, Z. (2004). In search of golden rules. Comment on hypothesis-testing approaches to setting cut-off values for fit indices and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320-341.
- Marsh, H.W., Hau, K-T. & Grayson, D. (2005): Goodness of fit in structural equation models. Ch. 10 in: Maydeu-Olivares, A. & McArdle, J.J. (eds.): *Contemporary Psychometrics. A Festschrift to Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum.
- McDonald R.P. & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin*, 107, 275-255.
- McIntosh, C.N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, 42 (5), 859-867.

- Meade, A.W. (2008). *Power of AFI's to Detect CFA Model Misfit*. Paper presented at the 23th Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA. Open access via: [http://www4.ncsu.edu/~awmeade/Links/Papers/AFI\(SIOP08\).pdf](http://www4.ncsu.edu/~awmeade/Links/Papers/AFI(SIOP08).pdf)
- Millsap, R.E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, 42 (5), 875-881. Available online 26 Dec. 2006.
- Miles, J. & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, 42 (5), 869-874. Available online 20 Dec. 2006.
- Mulaik, S. (2007). There is a place for approximate fit in structural equation modeling. *Personality and Individual Differences*, 42 (5), 883-891. Available online 19 Jan. 2007.
- Mulaik, S. (2010). *Another look at a cancer research model. Theory and indeterminacy in the BMC model by Hayduk et al.* Paper presented at the Annual meeting of the Society for Multivariate Experimental Psychology. Atlanta, GA, October 2010.
- Rigdon, E.E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, 3 (4), 369-379
- Saris W.E. & Satorra, A. (1988). Characteristics of structural equation models which affect the power of the Likelihood Ratio Test. In W.E. Saris & I.N. Gallhofer (Eds.), *Sociometric research* (Vol. 2, pp. 222-236). London: MacMillan.
- Saris, W.E., Satorra, A. & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16: 561-582.
- Steiger, J. H. (1989). EzPATH: Causal modeling. Evanston, IL: SYSTAT.
- Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.
- Steiger, J.H. (2000). Point estimation, hypothesis testing and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling*, 7, 149-162.
- Stuive, I. (2007). *A comparison of confirmatory factor analysis methods. Oblique multiple group method versus confirmatory common factor method*. Thesis, University of Groningen, The Netherlands.
- Stuive, I. Kiers, H.A.L., Timmermans, M. & Berge, J.M.F. (2008). The empirical verification of an assignment of items to subtests. the Oblique Multiple Group method versus the Confirmatory Common Factor method. *Educational and Psychological Measurement*, 68, 6, 923-939.
- Tucker, L.R. & Lewis, C. (1973), The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.