Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 26, December 2021 ISSN 1531-7714

Understanding the Comparative Fit Index: It's all about the base![1]

Saskia van Laar, *Centre for Educational Measurement at the University of Oslo (CEMO)*
Johan Braeken, *Centre for Educational Measurement at the University of Oslo (CEMO)*

Despite the sensitivity of fit indices to various model and data characteristics in structural equation modeling, these fit indices are used in a rigid binary fashion as a mere rule of thumb threshold value in a search for model adequacy. Here, we address the behavior and interpretation of the popular Comparative Fit Index (CFI) by stressing that its metric for model assessment is the amount of misspecification in a baseline model and by further decomposition into its fundamental components: sample size, number of variables and the degree of multivariate dependence in the data. Simulation results show how these components influence the performance of CFI and its rule of thumb in practice. We discuss the usefulness of additional qualifications when applying the CFI rule of thumb and potential adjustments to its threshold value as a function of data characteristics. In conclusion, we at a minimum recommend a dual reporting strategy to provide the necessary context and base for meaningful interpretation and even more optimal, a move to using CFI as a real incremental fit index intended to evaluate the relative effect size of cumulative theoretically motivated model restrictions in terms of % reduction in misspecification as measured by the baseline model.

Introduction

The evaluation of model fit remains a crucial yet controversial topic in the application of structural equation models. In line with concerns that a focus on mere statistical significance testing would lead to disregarding or changing relevant and theoretical sound models without proper justification for it (Bentler & Bonett, 1980), a whole range of alternative goodness-of-fit indices is currently available for model evaluation beyond the traditional chisquare significance test of exact fit. As part of the general trend to report multiple fit indices (e.g., Jackson et al., 2009; Ropovik, 2015), McDonald and Ho (2002) point out that "it is sometimes suggested that we should report a large number of these indices, apparently because we do not know how to use any of them" (p. 72). This statement highlights a common concern about current model evaluation practices that are characterized as thoughtless routine applications of binary (good/bad) rules of thumb for fit indices.

[1] This study was supported by a research grant [FRIPRO-HUMSAM261769] for young research talents of the Norwegian Research Council.

Different cut-off criteria or rules of thumb have been proposed over time (e.g., Bentler & Bonett, 1980; Hu & Bentler, 1999; Schermelleh-Engel et al., 2003). In particular, Hu and Bentler's (1999) suggested criteria gained huge popularity. Yet, Hu and Bentler (1999) themselves stressed that "it is difficult to designate a specific cutoff value for each fit index because it does not equally well with various conditions" (p. 27). Their underlying simulation study was based on only a few conditions with either a simple or a complex structure with fixed values for a three-factor confirmatory factor analysis model with 15 manifest variables. Their note of caution resonates well with more recent findings in the literature where simulation studies have illustrated the sensitivity of fit indices and their rules of thumb to various data and model features such as sample size, model size and type, strength of relations within the measurement model, and violations of distributional assumptions (for a review, see e.g., Niemand & Mai, 2018). Nevertheless, people have been universally applying the rules of thumb regardless of their own specific context, study design, data, or model. The main point of concern is exactly this thoughtless default way of applying rules of thumb (Marsh et al., 2004). One reason given for abiding by such a thoughtless rule-based approach is that "researchers need them because it is unclear how one can reach qualitative judgements in their absence" (Lai & Green, 2016, p. 221).

Overall, one major point of concern with respect to the application of SEM in practice is the lack of deliberate decision making in all parts of the process (McDonald & Ho, 2002). In order to make more informed decisions with respect to the use of fit indices it is important to know how these fit indices work. Yet what 'good' fit means and how fit indices map onto this meaning is not well understood (Lai & Green, 2016). Hence, if we would desire not mere mindless rule-following but more deliberate practice when assessing model fit, we need to better clarify what type of fit each of the different indices stand for and to provide a better insight in their inner workings to understand why fit indices behave like they do.

Here, we will try to make one step into that direction by focusing on the Comparative Fit Index (CFI) (Bentler, 1990), the most-used statistic among the class of comparative goodness-of-fit indices (for reviews covering time periods in the interval 1995-2013, see e.g., Jackson et al., 2009; McDonald & Ho, 2002; Ropovik, 2015). A decomposition in the main components that play a role in the CFI's baseline comparison allows to clarify CFI's meaning and behavior, explain some of the mixed results in the SEM simulation literature regarding its sensitivity to model and data characteristics, and highlight the (limited) generalizability of common rules of thumb for CFI and factor analysis. We hope that this exposition can help guide the decision-making process in practice and lead to smarter, more deliberate inferences when interpreting the CFI for model fit evaluation.

A Decomposition of the Comparative Fit Index

In contrast to absolute fit or parsimony fit indices (e.g., Brown, 2015), the class of comparative fit indices promotes comparison in fit between a model of interest and a more restricted baseline model. This fit assessment strategy has its foundation with Bentler and Bonett (1980) and involves a continuum of models from the worst fitting null model to the perfect fitting or saturated model. The role of the comparative fit indices is to assess where the model of interest is located within this continuum.

Within this class, Bentler's (1990) Comparative Fit Index (CFI) is an "index to summarize the relative reduction in noncentrality parameter of two nested models" (p. 238). The noncentrality parameter λ_m of a model m can be seen as an indicator of model misspecification as it quantifies the amount of deviation between the estimated χ^2 value and the expected χ^2 value (i.e., df_m, the model's degree of freedom) for the sample under the assumption that the model is correct: $\lambda_m = \chi^2_m - \mathrm{df}_m$. The value of CFI is then based on the ratio of misspecification of both models:

$$\mathrm{CFI}_{(m,b)} = 1 - \frac{\lambda_m}{\lambda_b} = 1 - \frac{\chi^2_m - \mathrm{df}_m}{\chi^2_b - \mathrm{df}_b} \quad (1)$$

where the subscript indicates whether the statistics are of the model of interest m or the baseline model b. The one-minus-noncentrality-ratio is there to turn it from a relative misspecification measure into a relative goodness-of-fit measure. Note that the CFI is usually truncated to the [0, 1] interval, although technically values higher than one can arise if the model of interest

fits better in a noncentrality sense than the saturated model (e.g., perfect fit with less than full parameters) and values below zero can arise if the model of interest fits worse than the baseline model.

Null baseline. A so-called null model in which all observed variables are uncorrelated has taken off as the default baseline model for popular applications of CFI. Following the idea of Bentler and Bonett (1980), the $\mathrm{CFI}_{(m,0)}$ can be referred to as an 'index of information gained' by the model of interest over the more restrictive null model. Hence, conceptually it is similar to an R-square, a relative reduction in 'unexplained' variance, whereas a $\mathrm{CFI}_{(m,0)}$ could then be seen as a relative reduction in 'unexplained' variance-covariance. From here on we will drop the subscripts referring to the models being compared, if we talk about the CFI with the null model as default baseline.

Rules of thumb. For determining whether a model shows adequate fit according to the CFI, different rules of thumb have been proposed. Early on up to the late 90's, values of at least .90 for comparative fit indices were assumed to indicate decent model fit (for a review, see McDonald & Ho, 2002). This rule of thumb has been mostly motivated based on experience by expert users: At CFI origins, "In our experience, models with overall fit indices of less than .90 can usually be improved substantially" (Bentler & Bonett, 1980, p. 600) or more recently, "In my experience, models with .90+ values for the CFI . . . can be quite acceptable models" (Little, 2013, p. 116). The currently most common CFI standard is based on the influential simulation study by Hu and Bentler (1999): "the results suggest that, for the ML method, a cutoff value close to .95 for . . . CFI . . . are needed before we can conclude that there is a relatively good fit between the hypothesized model and the observed data" (p. 1). As indicated earlier in the introduction, even about the core rule of thumb, stating CFI $\geq$ .95 for good model fit, there have been many cautionary notes and simulation studies have illustrated that its applicability varies depending on data and model characteristics.

If we would desire more deliberate practice when assessing model fit using CFI values, then knowing the inner workings of this measure is an essential requirement. So how does this CFI really work? Additionally, can knowledge of its inner workings indeed shed some light on the performance of the CFI rules of thumb under various data characteristics?

CFI as a relative measure with a variable metric space

Equation 1 clarifies that the CFI is a relative measure with its denominator set by the noncentrality of the baseline model. Now suppose there is a line that represents the CFI metric. The metric space endpoints are set by the null and saturated model. The length of the line is determined by the noncentrality of the null model, as the noncentrality for the saturated model is zero. Given the formulation of CFI, this metric space serves as standard for comparison. Conceptually, the length of the line, the CFI metric space, has an influence on the behavior of CFI. Having more space, will allow for a finer grained differentiation. Having less space, makes the CFI to become less useful. The rationale is that in general it is harder to differentiate between models as they are becoming more similar. When placing a model of interest in the metric space, it will always be closer related to both the null and the saturated model as the line becomes shorter. As a consequence, a comparison in terms of CFI values is no longer based on the same standard when the denominator, the baseline noncentrality, is different among the cases being compared.

As an example to drive this idea home, consider the following two cases for which the size of the CFI metric space is different. The baseline noncentrality in the first case is $\lambda_0 = 25$. Within this space two models with slightly different noncentrality values can be placed. Overall their values only differ by 2 units, with $\lambda_1 = 1$ and $\lambda_2 = 3$ being the noncentrality value of the first and second model, respectively. Translating this to CFI values, this results in values of $\mathrm{CFI}_{(1,0)} = .96$ and $\mathrm{CFI}_{(2,0)} = .88$. Now consider the second case in which there is a shorter metric space with baseline noncentrality $\lambda_0 = 5$. Here as well, we have two models that only differ by 2 noncentrality units, now with $\lambda_1 = .2$ and $\lambda_2 = 2.2$. However, translating this to CFI interval, values of $\mathrm{CFI}_{(1,0)} = .96$ and $\mathrm{CFI}_{(2,0)} = .56$ are obtained. This example demonstrates the impact of widely differing metric spaces as defined by the baseline noncentrality. The difference in CFI-fit between the two models is huge between the two cases whereas the difference in terms of absolute misspecification as expressed by the noncentrality index is exactly the same. Sampling variability can also be expected to have a huge impact in the second case, a small difference in noncentrality value

can lead to widely differing CFI values when baseline noncentrality is small. Thus, the main conclusion is that we cannot interpret a CFI-value of a model or differences in CFI between models without considering the fit of the CFI baseline model for the same sample data. This is similar advice as with any ratio or risk measure, you cannot ignore the numerator and denominator when interpreting a percent; Or more colloquially speaking, whereas a small percent of everything is a lot, a large percent of nothing, is still nothing.

Null model baseline noncentrality as key factor

For the default CFI with a null model as baseline, the null model noncentrality λ_0 is the key to CFI behavior and interpretation as it sets the metric space that serves as standard for comparison. With F being the ML discrepancy fit function (e.g., Bollen, 1989) between the observed and null-model-implied covariance matrices $\mathbf{S}$ and $\widehat{\boldsymbol{\Sigma}}_0$, the null model noncentrality can be rewritten and simplified as follows to identify its key components:

$$\begin{aligned}\lambda_0 &= \max(\chi_0^2 - \mathrm{df}_0, 0) \\ &= \max(F\left(\mathbf{S}, \widehat{\boldsymbol{\Sigma}}_0\right)(n-1) - \mathrm{df}_0, 0) \\ &= \max(-\log|\mathbf{R}|(n-1) - p(p-1)/2,\ 0)\end{aligned} \quad (2)$$

where $\mathbf{R}$ is the observed correlation matrix, n the sample size, and p the number of manifest variables (for the derivation, see Appendix A).

Equation 2 clarifies that the CFI metric space is a function of correlation (i.e., generalized variance as expressed by the determinant of the data correlation matrix), sample size, and number of variables. Notice that all three core components of the null model baseline noncentrality are completely data dependent. In an ideal situation with a lot of correlation in your data, large sample sizes and not too many variables, CFI would allow you to make a fine-grained differentiation between models in terms of relative noncentrality. These ideal conditions are quite in line with common sense guidelines for the application of SEM. There are some more general intuitions that can be derived a priori from this decomposition that can be linked to findings in the SEM model fit literature.

Sample size n. Originally, comparative fit indices were conceptualized as 'indices of information gained' and should be independent of sample size (Bentler & Bonett, 1980). However, previous studies (e.g., Heene et al., 2011; Hu & Bentler, 1999; Marsh et al., 2004; Shi et al., 2019) as well as the decomposition show that CFI is clearly dependent on sample size. In this case, with higher sample sizes resulting in higher baseline noncentrality values and better expected performance.

Number of variables p. In the literature (e.g., Shi et al., 2019) a general trend has been reported that more variables complicate the use of CFI and its default rule of thumb. At first sight the decomposition supports this notion as more variables leads to lower baseline noncentrality making model differentiation more difficult. However there is a confounding factor that is easily forgotten, the determinant $|\mathbf{R}|$ is also a function of the number of variables p, and with more variables more non-zero correlations can in principle occur in the correlation matrix $\mathbf{R}$. Hence, the number of variables only has a clear negative effect on CFI if $p(p-1)/2$ the degrees of freedom of the null model outweighs the contribution by $-\log|\mathbf{R}|(n-1)$.

In the extreme theoretical situation in which only additional uncorrelated variables are added this will be always the case, as this has no impact on the latter factor. Yet the more correlation the added variables contribute the faster the negative effect of the number of variables disappears (i.e., the logdeterminant factor increases nonlinearly). Hence, it should thus not be surprising that Shi et al. (2019) found that, for correctly specified models, the effect of p on performance of CFI's rule of thumb was dependent on the size of the factor loadings they used. Hence, CFI also follows the general principle that having more signal in the data facilitates matters, whereas adding more noise further confounds matters.

Data correlation $\mathbf{R}$. As already indicated in the previous paragraph, the more the data is unlike the null model, the higher the baseline noncentrality and the easier CFI can differentiate between models. The study by Heene et al. (2011) also showed that performance of CFI's rule of thumb is dependent on used factor loadings. It should also not be surprising that performance issues became more severe as the sample size decreased (Heene et al., 2011), as there is a synergistic interaction between n and $-\log|\mathbf{R}|$ as reflected by the prominent role of their product in the

decomposition. Given the formulation, a decrease in both components will provide the smallest metric space, providing worse conditions for model differentiation.

Now that we have identified the core components that play an integral part in the baseline comparison for CFI we will first zoom in further on CFI in relation to different data characteristics, by assessing the impact of sampling variability on the proposed metric space principle and the extent to which this relates to the general applicability of the common rule of thumb for CFI. Secondly, we will follow up on an additional qualification on when the general CFI rule of thumb can be used. We end the paper with a more general discussion on implications of these results and with recommendations for the use of CFI and its common rule of thumb in practice.

Sampling variability & CFI

At population level, CFI is determined by the population model noncentrality $\lambda_m^{(\Sigma)}$ and the population null baseline noncentrality $\lambda_0^{(\Sigma)}$. When the estimated model is the true population model, $\lambda_m^{(\Sigma)}$ shows perfect fit $(\lambda_m^{(\Sigma)} = 0)$ and consequently the population CFI will always equal one. This means there is only systematic variation in $\lambda_0^{(\Sigma)}$, caused by variation in the components that make up the CFI metric space. Even though this does not have a direct influence on the CFI value at population level, it will set the basis for sample performance of CFI: a larger null baseline noncentrality $\lambda_0^{(\Sigma)}$ provides a more solid basis for model differentiation. In practice, the two noncentralities at sample level $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ will be prone to sampling variability and potentially also sample bias. Depending on the extent that both noncentralities are somewhat differently affected, this could lead to differences in results compared to our expectations.

Monte Carlo Simulation Design

We considered a simple one-factor data-generating population model with equal factor loadings implying equal correlations between all items. The focus was on the use of correctly specified models, as it seems that the goal of most people is not to falsify their model, but to find an adequate model as starting point for further analysis (e.g., Ropovik, 2015). Given this focus on adequate model fit, it would be good to know whether CFI's rule of thumb can meet its purpose in the ideal case of a correctly specified model.

Experimental Factors. The conditions studied are related to the three components of the baseline noncentrality provided by the decomposition of CFI: sample size n, number of variables p, and data correlation $\mathbf{R}$.

First, sample size is varied ($n \in \{100, 200, 500, 1000\}$). More information is present with increasing sample size, such that there is less uncertainty in making inferences about model fit. Minimum sample size requirements around 150-200 have been proposed for SEM (e.g., Barrett, 2007; Boomsma, 1985; Kenny, 2015; Muthén & Muthén, 2002), yet in practice about 1 in 5 studies uses sample sizes below 200 (MacCallum & Austin, 2000) and around 8-18% uses sample sizes below 100 (Jackson et al., 2009).

Second, the number of variables is varied ($p \in \{4, 8, 12, 24\}$), as previous research has shown that the number of variables does have an influence on model evaluation (e.g., Moshagen, 2012; Shi et al., 2019; Shi et al., 2018).

Third, the degree of data correlation as expressed by $|\mathbf{R}|$ is varied through the chosen data-generating population model. The use of the one factor homogeneous factor loading model as population model allows to make this determinant a direct function of one correlation number r, where $|\mathbf{R}| = [1 + (p - 1)r][1 - r]^{(p-1)}$ (e.g., Graybill, 1983) with $r \in \{.1, .2, .3, .5, .7, .9\}$. According to Brown (2015), in practice standardized factor loadings of at least .3 or .4 are considered the norm for a meaningful interpretation, which corresponds in our simulation setup to values of $r = .09$ and $r = .16$, respectively. Hair et al. (2006) are stricter and require factor loadings to be above .5 or even .7 in the context of validation studies, which corresponds to values of $r = .25$ and $r = .49$.

Experimental Design. These three experimental factors are combined into a full factorial simulation design leading to $n(4) \times p(4) \times r(6) = 96$ experimental conditions. Within each condition, 1000 sample covariance matrices $\mathbf{S}$ were drawn from a Wishart distribution, $\mathbf{S} \sim W(\mathbf{\Sigma}, df)$, where $\mathbf{\Sigma}$ is the model's population covariance matrix and df the model's degrees

of freedom. The model was then refitted to each of the generated samples. The simulation and analyses were conducted in R (R Core Team, 2020) through custom scripts in combination with the lavaan package for R (Rosseel, 2012).

Outcome measures. For each sample, the sample noncentrality of the baseline model and of the fitted model – being the numerator and denominator of the CFI, respectively – are computed. The CFI of the fitted model is assessed and used to decide whether or not the fitted model is judged to be of good fit according to the .95 rule of thumb (i.e., CFI < .95 leads to rejection of the model).

Monte Carlo Simulation Results

Full results of the 96 experimental conditions of the Monte Carlo simulation study are reported in table-format in Appendix B. In what follows, we will report on general trends for the respective outcome measures and zoom into specific conditions when relevant.

Null baseline noncentrality $\lambda_0^{(S)}$. Given that noncentrality parameters are shifted-versions of the chisquare statistic (i.e., $\lambda_0 = \chi_0^2 - \mathrm{df}_0$), the same sampling distributions would apply under asymptotical theory given regularity conditions (e.g., Steiger et al., 1985), implying a central or noncentral chisquare distribution depending on whether or not the model is correctly specified. Yet note that for the null baseline model it has been found that a noncentral chisquare distribution does not properly describe its sampling distribution beyond its central tendency (Curran et al., 2002). However, the sample null baseline noncentrality does follow nicely the population trends (see Table 1) that are function of the earlier identified three

Table 1. Eta square (η^2) effect size patterns for the main components of the CFI metric space across different outcome measures in the main simulation study.

	η^2				
term	$\lambda_0^{(\Sigma)}$	$\lambda_0^{(S)}$	$\lambda_m^{(S)}$	CFI	$< .95$
p	.124	.134	.276	.010	.025
r	.268	.265	.000	.129	.364
n	.146	.144	.033	.076	.212
$p \times r$	.135	.134	.000	.011	.028
$p \times n$	.081	.080	.075	.022	.068
$r \times n$	.163	.160	.000	.081	.227
$p \times r \times n$	.082	.081	.000	.029	.076
total	1	.999	.384	.358	1

Note. $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality; $\lambda_m^{(S)}$= sample noncentrality for the estimated true model; CFI = sample CFI value for the estimated true model (i.e., $\mathrm{CFI} = \lambda_m^{(S)}/\lambda_0^{(S)}$; <.95 = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95. η^2's are based on the type-III sum of squares in a full factorial ANOVA.

components of the metric space. Where an increase in either of the components has a positive effect on the baseline noncentrality. Notice that the sampling variation unaccounted for by the design factors is almost non-existing (i.e., $1 - \eta^2_{total} = .001$).

Comparing the theoretically expected $\lambda_0^{(\Sigma)}$ with the sample average $\bar{\lambda}_0^{(S)}$ (see Table B1) indicates that a small upward sampling bias for $\bar{\lambda}_0^{(S)}$ is present. This bias tends to become more severe with additional variables p. The relative effect of this upwards bias is worse for the lower sample size conditions, but has less of an impact with increased correlation r as the corresponding increase in the absolute value of $\bar{\lambda}_0^{(S)}$ dwarfs the bias. One consequence of the upward bias is that all small-sample-with-limited-correlation conditions that had a similarly restricted non-optimal baseline at population level, now at sample level are ordered as a function of the number of variables p.

Model noncentrality $\lambda_m^{(S)}$. Under asymptotical theory given regularity conditions (e.g., Steiger et al., 1985), the χ^2_m fit statistic when the true model is estimated, is expected to follow a central chisquare sampling distribution with mean *df*. Hence, the sample noncentrality of the model $\bar{\lambda}_m^{(S)}$ should tend to its expected value 0.

However, some upward sampling bias in $\bar{\lambda}_m^{(S)}$ is present for almost all simulation conditions, although in absolute terms this is smaller than for $\bar{\lambda}_0^{(S)}$. The true model's noncentrality (and hence its sampling bias) is most affected by the number of variables p (see Table 1), and in contrast to its prominent role in the null model unaffected by the amount of correlation r. The most severe bias is observed in the low-sample-size-many-variables conditions ($p = 24$, $n = 100$). Overall, increasing sample size seemed to reduce the biasing effect of the additional variables. The finding of large sampling bias as a function of increasing number of manifest variables and moderated by sample size corresponds to earlier findings in the literature (e.g., Moshagen, 2012). Notice that the sampling variation unaccounted for by the design factors (i.e., $1 - \eta^2_{total} = .671$) is also much higher for the model noncentrality than for the null baseline noncentrality (i.e., $1 - \eta^2_{total} = .001$).

Comparative Fit Index (CFI). The asymptotically-derived sampling distribution of the CFI has not yet been established in the literature although logically it would conform to the sampling distribution of a ratio of two dependent shifted (non)central chisquare distributions, with the caveat that even a shifted noncentral chisquare is not fully applicable for the null baseline model. What we identified so far in the simulation study is that sampling affects the numerator $\lambda_m^{(S)}$ and denominator $\lambda_0^{(S)}$ of the CFI in a slightly different fashion. The resulting effect patterns on CFI in our simulation design (see Table 1) reflect this duality and lead to a mix of both λ-patterns, with the most central role for correlation r followed by sample size n, whereas the effect of the number of variables p has become negligible.

As we looked at CFI values for estimated true models, all observed CFI values should be indicative of the kind of sample values that can be expected to express good model fit. The 5% CFI quantile shows that the expected range of realistic CFI values actually varies greatly and covers a broad range across conditions (see Table B1). This difference becomes most prominent in those conditions where low sample size co-occurs with low correlation. In the most extreme situation (i.e., $n = 100$, $p = 24$, $r = .1$), 5% of the replications even have CFI values below or equal to .57. As reference to get the picture of the whole range, 16% of replications in this condition still have CFI values above or equal to .95. At the same time, for some conditions (e.g., but not exclusively, the conditions where correlation $r = .9$) the range of realistic CFI values is much more limited as the 5% quantile was already as high as .99 or even 1.

Rule of thumb CFI ≥ .95. The common rule of thumb for CFI states that CFI should be at least .95 to speak of acceptable goodness of fit, and otherwise if CFI < .95 one would reject the model. Given that the true model is fitted each time, the ideal outcome is of course a rejection rate of 0%. The results in Table B1 however, show that this is not accurate for all conditions. The median rejection rate is 0% but the average is 8% with a maximum of 84%. Of our 96 conditions, 43 had a non-zero rejection rate and 27 a rejection rate larger than 5%.

These results follow automatically from the observed ranges of CFI values for a true model not being consistent with the range implied by the rule of thumb [.95, 1]. The much wider or at times more narrower range of observed CFI for the estimated true model

would imply that the rule of thumb should/could in fact be made more lenient or strict depending on the situation. A point to which we will return in the discussion.

Metric space principle $CFI|\lambda_0^{(S)}$. In line with our starting 'metric space' principle that the baseline determines differentiation power of CFI, the effect size patterns (see Table 1) for the model rejection rates given the rule of thumb follow the trends for the (sample and population) null baseline noncentrality yet with a diminished role of the number of variables p. Hence, increasing the metric space by increasing CFI's denominator through increasing either of the three design components has a positive effect on the size of $\lambda_0^{(S)}$, the size and range of CFI values, and the resulting model rejection rates according to the common rule of thumb (see also Table B1 for a detailed overview of results).

The observed diminished role of p is due to the set of conditions where low sample sizes are combined with low correlation in the data (i.e., $n = 100$ & $r \leq .5$ or $n = 200$ & $r \leq .2$) where a larger number of variables p leads to higher (see the excerpted conditions in Table 2) instead of the generally expected lower rejection rates. Sampling variability and bias in those conditions destroy the regularity of the metric space principle. Focusing on one of the low-sample-size-low-correlation conditions, Figure 1 shows an example of how sampling variation in $\lambda_m^{(S)}$ relates to sampling variation in $\lambda_0^{(S)}$ as a function of the number of variables p. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. Given the definition of CFI (see Equation 1), the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that result in CFI = .95. When replications are positioned in the area above this line, the corresponding CFI value will always be below .95, leading to rejection of the model. In other words, the values of $\lambda_m^{(S)}$ in these situations are becoming too large compared to their $\lambda_0^{(S)}$ counterpart to acquire good model fit according to CFI. While replications positioned on or below the diagonal line correspond to good model fit according to the .95 rule of thumb for CFI.

For both $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$, their mean values increase with additional variables p as seen in their respective marginal distributions. However, the trend in $\lambda_m^{(S)}$ seems to be dominant over the trend in $\lambda_0^{(S)}$, as with additional variables p, $\lambda_m^{(S)}$ results in more extreme values relative to the $\lambda_0^{(S)}$ counterparts as seen in the heavier right tail in the distribution of the former. As a consequence, more replications are wrongly classified as showing inadequate model fit. In these specific conditions, problems in CFI performance are due to the strong sampling variation and bias in the numerator $\lambda_m^{(S)}$ that counteracts the positive effect of increased average size of the metric space reflected by the denominator $\lambda_0^{(S)}$.

In the majority of the cases, this bias-interference is not applicable and the general metric-space principle works out despite sampling variation and bias in CFI's numerator and denominator. Figure 2 serves as an illustration of this principle. Whereas the distribution of $\lambda_m^{(S)}$ remains relatively constant across increasing correlation, the distribution of $\lambda_0^{(S)}$ takes big steps upwards, dwarfing any sampling bias in $\lambda_m^{(S)}$. The increase in correlation leads to a big increase in null baseline noncentrality which goes together with a decrease in the rejection rates of the CFI for the correctly specified model. The same results hold with increasing sample size n, whereas for increasing number of variables p it is less demarcated due to the opposing bias in $\lambda_m^{(S)}$.

Don't interpret CFI depending on RMSEA of null model?

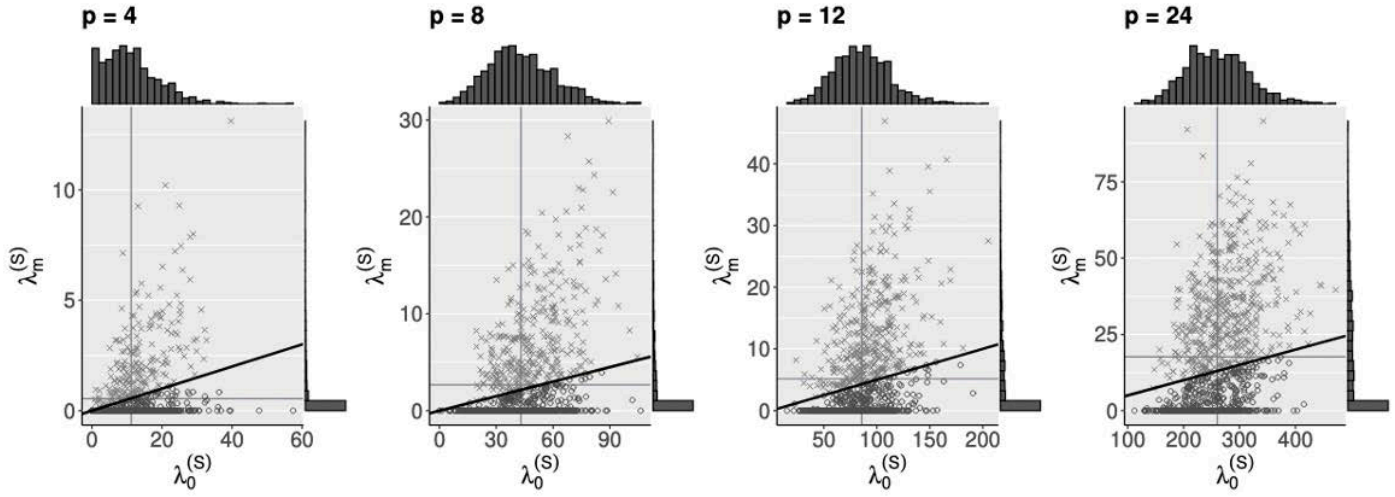
As indicated before, additional specifications on the use of the general rule of thumb for CFI have been around. For example, one lesser known qualification advocated for on a popular web resources on SEM fit indices recommends that "CFI should not be computed if the RMSEA of the null model is less than .158 or otherwise one will obtain too small a value of the CFI" (Kenny, 2015). However, formal support for this recommendation was not given. Hence, we used the results from the main simulation study to follow up on the usefulness of this specific qualification in practice.

Table 2. Contradicting the metric space principle: Negative effect of the number of variables p on the performance of CFI.

n	r	$p = 4$			$p = 8$			$p = 12$			$p = 24$		
		$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	$< .95$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	$< .95$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	$< .95$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	$< .95$
100	0.1	6.3	0.3	18.8%	22.7	2.7	41.6%	46.7	6.0	51.7%	152.6	32.2	84.2%
	0.2	19.7	0.6	20.1%	69.0	3.1	34.3%	133.0	5.7	34.4%	371.9	31.4	68.2%
	0.3	43.7	0.7	11.3%	137.0	3.4	19.2%	250.6	6.8	21.6%	653.9	32.3	45.8%
	0.5	116.2	0.8	1.6%	334.6	3.0	1.8%	583.2	6.6	2.6%	1372.4	32.3	7.6%
200	0.1	11.2	0.5	20.6%	43.0	2.7	31.8%	85.8	5.2	37.6%	260.1	17.6	47.4%
	0.2	40.5	0.7	12.9%	139.4	2.7	14.8%	258.9	4.9	12.4%	696.4	18.1	17.6%

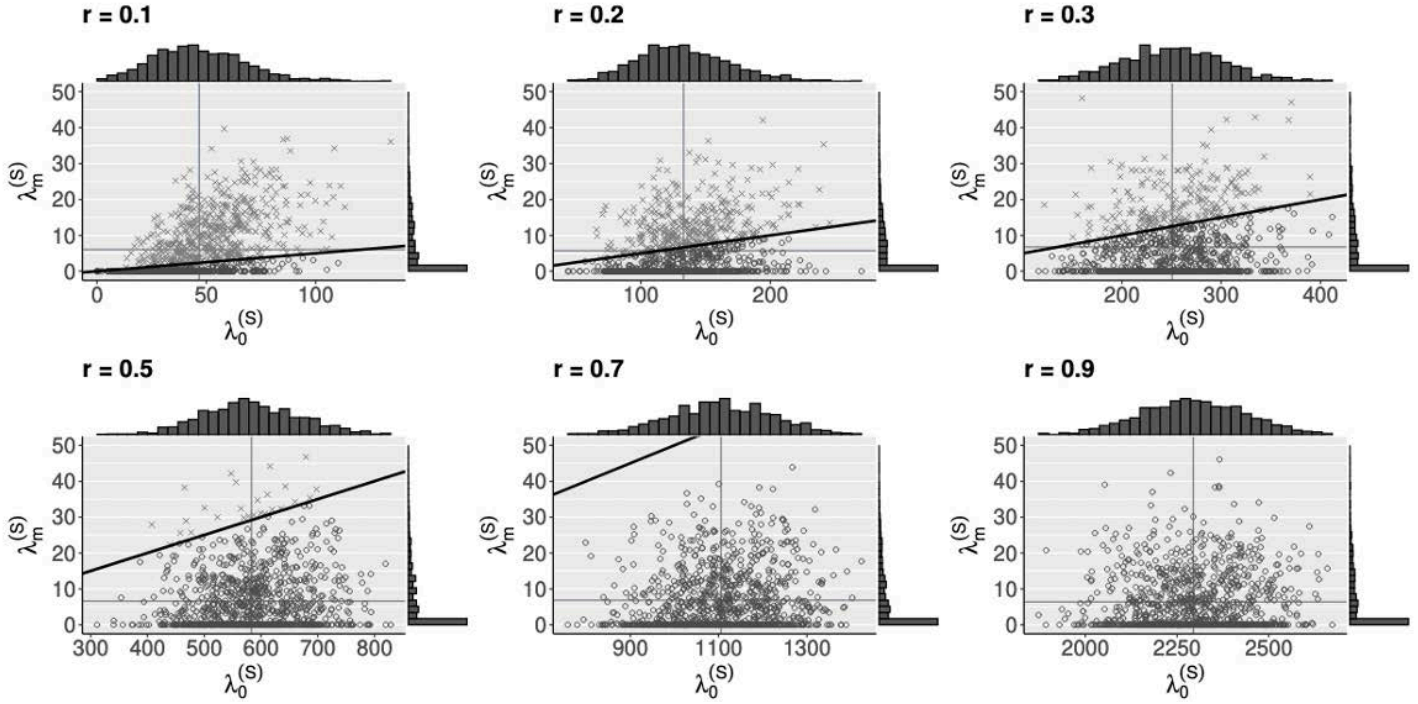
Note. In general an increase in the size of the metric space is expected to have a positive effect on the CFI model rejection rates. However, the results, excerpted from Table B1, show those conditions where additional variables p result in increased rejection rates for CFI, even though $\bar{\lambda}_0^{(S)}$ increases as expected. It should however be noted that in some conditions the rejection rates are still close to zero (e.g., when n = 100 and r = .5). With $\bar{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\bar{\lambda}_m^{(S)}$ = average sample noncentrality for the estimated true model; < .95 = model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95.

Figure 1. Scatterplot with marginal distributions of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ as a function of the number of variables p for the conditions where $n = 200$ and $r = 0.1$.



Note. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. With $\lambda_m^{(S)}$ = sample noncentrality for the estimated true model; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality. Given that CFI = $1 - \frac{\lambda_m}{\lambda_0}$, the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that results in CFI = .95. Replications that are positioned in the area above this line will always result in CFI values below .95, leading to rejection of the model. While replications positioned on or below the diagonal line will result in good model fit according to the .95 rule of thumb for CFI. The pattern observed is for the low-sample-size-low-correlation conditions not conforming to the metric space principle, for which theoretically unexpected higher rejection rates occur with increasing number of variables (see also Table 2).

Figure 2. Scatterplot with marginal distributions of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ as a function of the correlation r for the conditions where n = 100 and p = 12.



Note. The horizontal and vertical line in the figure respectively show the average value of $\lambda_m^{(S)}$ and $\lambda_0^{(S)}$ within a specific condition. With $\lambda_m^{(S)}$ = sample noncentrality for the estimated true model; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality. Given that CFI = $1 - \frac{\lambda_m}{\lambda_0}$, the diagonal line is the critical line representing the combination of $\lambda_m^{(S)}$ values and $\lambda_0^{(S)}$ values that results in CFI = .95. Replications that are positioned in the area above this line will always result in CFI values below .95, leading to rejection of the model. While replications positioned on or below the diagonal line will result in good model fit according to the .95 rule of thumb for CFI. In contrast to Figure 1, the pattern seen here is the dominant pattern conforming to the metric space principle, instead of the exception to the rule.

We expected that if this rule of thumb works, cases where $RMSEA_0 < .158$ co-occur with a CFI value below the commonly adopted .95 threshold more often than not for models that fit.

As an initial rough effectiveness indicator of this rule of thumb we cross-classified all replications for each condition from the main simulation study based on whether the sample $RMSEA_0$ and CFI values were below or above their respective thresholds (see Table 3). On average the incidence of $RMSEA_0 < .158$ amounted to 31% of the cases. Given $RMSEA_0 < .158$, the probability for also obtaining a CFI value below .95 was on average 17.5% with a range across conditions between 0 and 84.2%. The reason for this wide range can be clearly illustrated by translating the $RMSEA_0 < .158$ into a corresponding required value for the null baseline noncentrality $\lambda_0^{158} = \text{RMSEA}_0^2 \times (n-1) \times \text{df}_0$. This threshold null baseline noncentrality λ_0^{158} value indeed only depends on two design factors – the number of variables p ($\eta^2 = .482$), sample size n ($\eta^2 = .225$) –, and their interaction $n \times p$ ($\eta^2 = .293$), but not on the third factor data correlation r (i.e., $\eta^2 = .000$ for r, $p \times r$, $r \times n$, & $p \times r \times n$). As one example, Table 4 clearly illustrates the ignorance of this $RMSEA_0 < .158$ threshold for the conditions where sample size $n = 200$ and $\text{df}_0 = 28$ (i.e., number of variables $p = 8$). Note that these results generalize across the other conditions. The $RMSEA_0 < .158$ specification wrongly assumes a null baseline noncentrality λ_0^{158} that remains constant regardless of the correlation r in the data, whereas CFI and its denominator the null baseline noncentrality λ_0 are highly sensitive to exactly this correlation.

In the end, the overall negative predictive value of the .158 rule of thumb appears to be not too reliable (i.e., $Pr(CFI < .95 | RMSEA_0 < .158)$). Hence, it varies highly whether we can indeed expect too low CFI values given a correctly specified model when $RMSEA_0 < .158$. On the other hand, the correct decision of acceptable fit (i.e., $CFI \geq .95$) is taken in on average 95.8% (range across conditions = 52.9-100%) of the cases that $RMSEA_0 \geq .158$. Hence, the overall positive predictive value (i.e., $Pr(CFI \geq .95 | RMSEA_0 \geq .158)$) of the .158 rule of thumb is more promising. The reason for this difference is that for specific settings the null baseline noncentrality corresponding to the $RMSEA_0 = .158$ threshold is unreachable. This is illustrated in the latter columns of Table 4, where for this particular case of $n = 200$ and $p = 8$, $RMSEA_0$ values below .158 can only occur in conditions with correlations r below .3 (i.e., $\lambda_0^{(s)} < \lambda_0^{158}$). Note that the specific breakdown point does vary depending on sample size n and number of variables p. In the end, this leads exactly to flagging down some of the conditions in which the CFI baseline for comparison is rather too small for effective model differentiation.

Table 3. Cross-classification of all replications in the main simulation study based on their $RMSEA_0$ and CFI value relative to the corresponding thresholds.

	$RMSEA_0$	
CFI	$< .158$	$\geq .158$
$\geq .95$	24.91% [0-100%]	67.17% [0-100%]
$< .95$	6.29% [0-84.2%]	1.64% [0-21.1%]

Note. $RMSEA_0$ = RMSEA values for the null baseline model; CFI = CFI values for the estimated true model. For the proposed rule of thumb to work, $RMSEA_0$ values below .158 ought to co-occur with CFI values below .95. Each cell in the cross-classification contains the overall average percentage and range of average percentages of replications across conditions in the main simulation study that is consistent with its thresholds-requirements.

In sum, despite its relatively good average positive predictive value, the proposed .158 rule of thumb does not fully meet its purpose. In its current form it is too general and ignores the role of one of the key components of CFI (cf. data correlation). In light of the wide range of values and variation in performance, it does not seem advisable to utilize a fixed general RMSEA threshold as the conclusive answer for assessing whether or not to apply the CFI for fit assessment.

Discussion

If we would desire not mere mindless binary rule-following but more deliberate practice when assessing model fit, we need to better clarify what type of fit each of the different indices stand for and to provide a better insight in their inner workings to understand why fit

Table 4. Attainability of the threshold: Sensitivity of the null baseline noncentrality and CFI to data correlation r in relation to the constant $RMSEA_0$ rule of thumb and corresponding threshold in terms of the null baseline noncentrality λ_0^{158}.

r	threshold		$\lambda_0^{(\Sigma)}$	$\lambda_0^{(S)}$			CFI		
	$RMSEA_0$	λ_0^{158}		M	MIN	MAX	M	MIN	MAX
.1	.158	139.099	13	43	0	106	.95	.55	1.00
.2	.158	139.099	109	139	53	265	.98	.78	1.00
.3	.158	139.099	245	275	152	453	.99	.92	1.00
.5	.158	139.099	642	667	397	970	1.00	.96	1.00
.7	.158	139.099	1303	1324	1019	1655	1.00	.98	1.00
.9	.158	139.099	2798	2832	2409	3326	1.00	.99	1.00

Note. The results stem from the main simulation study and show an example for the conditions where the sample size n = 200 and the number of variables p = 8. $RMSEA_0$ = RMSEA threshold of the null baseline model; λ_0^{158} = .158 threshold for $RMSEA_0$ translated in terms of null baseline noncentrality; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\lambda_0^{(S)}$ = sample value of the null baseline noncentrality; CFI = CFI value for the estimated true model.

indices behave like they do. In this study, we started with such endeavour for the Comparative Fit Index.

CFI is a relative model fit measure expressed as a ratio of the noncentrality of the model of interest to that of a baseline comparison model. In essence this implies that the CFI is in fact a standardized statistic where the standard of comparison is typically provided by the noncentrality of the null model that is by default chosen as comparison model. This does mean that one CFI is not the other because the baseline standard, the noncentrality of the null model, is determined by data dimensions (i.e., $n \times p$) and amount of multivariate dependence in the data (i.e., $|\mathbf{R}|$). This is important as the implications of absolute value judgement of good fit according to CFI might not correspond to the relative improvement CFI stands for. With a small CFI metric space, low relative improvement does not necessarily imply that a model is not good in terms of absolute fit, while a high relative fit given a large metric space can still be associated with a large amount of absolute misspecification. The broader the baseline, the less strict the CFI ≥ .95 rule of thumb becomes as more absolute misspecification is allowed for a model that is considered to adequately fit. This natural feature of a standardized/relative measure such as CFI, brings Moshagen and Auerswald (2018) to caution strongly against CFI's use for evaluating absolute fit of a single model.

However, such decontextualized assessment of fit of a single model is unfortunately quite commonplace in practice with the default application of the binary rule of thumb: CFI ≥ .95 means "good fit" whatever that might mean. If we formalize the latter as correctly identifying the true model as a good fitting model, with a binary decision rule that works at least 95% of the time, our simulation results show that the rule of thumb needs to be adjusted based on data characteristics or only be applied under certain qualifications.

Qualifications for use of CFI's rule of thumb. Our results illustrate the theoretically derived principle that a wider basis for model differentiation is provided by increasing the three core components of the null baseline noncentrality – sample size n, number of variables p, and multivariate dependence as reflected by $|\mathbf{R}|$, the determinant of the data correlation matrix. This results

in high rates of qualifying the correctly specified model as having good fit in high signal to noise conditions, that is high correlation with added high sample size regardless of the number of variables. In contrast, in low signal to noise conditions, that is low sample size and low correlation, the CFI ≥ .95 rule was too strict and an increase of the number of variables made matters even worse. In the latter conditions, the null baseline model is already quite close in absolute fit to the correctly specified model, hence it is less likely to observe a huge relative change of 95% of that small distance even for a correctly specified model. Consequently, a word of caution for the current binary use of the CFI ≥ .95 rule of thumb in such conditions is in order. Sample sizes below 200 are unfortunately not uncommon (Jackson et al., 2009; MacCallum & Austin, 2000) and the prevailing pragmatic idea that standardized factor loadings of .3 (r = .09) and .4 (r = .16) are sufficient for meaningful interpretation (Brown, 2015) seems too optimistic.

The CFI ≥ .95 rule of thumb would approximately work in this 95% correct sense as a function of sample size and correlation: for n = 1000, a correlation of at least r = .1, for n = 500, a correlation of at least r = .2 is required, for n = 200 a correlation of at least r = .3, and for n = 100 a correlation of at least r = .5. Based on our simulation results, a conjecture could be put forward that a baseline noncentrality of $\lambda_0^{(S)} \geq 1400$ provides a sufficient broad metric space for fine-grained model differentiation using the CFI (e.g., conditions in line with this requirement had very narrow CFI range for the true model and far above the .95 rule of thumb). This is a conservative guideline as things do not necessarily look bad in all smaller baseline conditions. Although the general CFI metric-space principle holds, the specific values suggested here are of course based on the limited set of levels of factors considered in the small simulation study, and would be somewhat adjusted with availability of results for more factor levels (e.g., extra sample size conditions) or even other design factors such as the data-generating model. Yet, the general identified patterns related to the CFI baseline are mostly data driven and core points and non-value specific recommendations can in that sense be trusted to generalize quite well.

We already mentioned that these type of additional qualifications, on when the CFI rule of thumb can be used, are not something new. Specifically, we looked into the recommendation not to use CFI if the RMSEA of the null model is less than .158 (Kenny, 2015). Even though this qualification does attempt to provide a more nuanced reporting of CFI, the simulation results showed that in light of its wide variation in performance across conditions, it is not advisable to use this specific qualification without careful deliberation. Yet, the underlying idea does contain merit as it essentially intends to filter out cases where there is a lack of covariance and high levels of noise in the data. Perhaps, we should not even consider SEM in such cases in the first place (e.g., Barrett, 2007) or at the minimum realize that it's not reasonable to expect a large relative fit difference from a null baseline model that itself is already very closely fitting to the data in an absolute parsimony fit sense.

Adjusting CFI's rule of thumb. Alternatively, instead of including additional qualifications on when to use CFI's rule of thumb, we could also adjust the rule of thumb depending on data characteristics. The general pattern of results shows that the CFI threshold should even become stricter in the more optimal situations (high correlation r, high sample size n: CFI 5% quantiles as high as .99), while it needs to be reduced considerably in the less optimal situations (low correlation r, low sample size n). The latter could even result in setting a threshold value as low as CFI ≥ .57 for a specific condition (n = 100, p = 24, r = .1). When realistic CFI values for a true model cover such a broad range, CFI loses its informativeness for absolute model fit assessment.

Effect size. Another more drastic, but likely preferable alternative to including additional qualifications on when to use CFI's rule of thumb or adjusting its threshold value as a function of data characteristics, would be to actually interpret CFI's value. In this respect, it is useful to see CFI as an extension of the linear regression model's R-square effect size measure to the broader SEM field. Both measures have indeed a similar setup:

$$r^2_{Y|X} = 1 - \frac{SS_{error}}{SS_{total}}$$

$$\text{CFI}_{(m,0)} = 1 - \frac{\lambda_m}{\lambda_0}$$

$$\text{effect size} = 1 - \frac{\text{misspecification target model vs. saturated model}}{\text{misspecification null model vs. saturated model}}$$

This further clarifies that in essence, CFI is, like the R-square, a standardized effect size measure and hence all reservations with respect to interpretations of standardized effect size measures (e.g., Baguley, 2009)

transfer to the interpretation of CFI. Such a realization has two major implications.

Firstly, CFI can be a useful benchmark metric for interpreting the relative magnitude of the effects within the same application dataset. Having a set of competing models, CFI can be used to quantify the effect size of the paths in which the models differ. In other words, we are using CFI as intended as an incremental comparative fit index among a set of models for the same dataset and interpreting its value in terms of relative magnitude.

Secondly, comparing CFI's across different datasets is not straightforward as given their standardized nature, a value of .95 is indeed similar in relative magnitude, but not necessarily in absolute magnitude. The latter would require that the denominator in CFI's formula remains constant across datasets. Where R-square is a relative reduction in variance not accounted for, and the denominator is a proxy for total variance in the outcome variable, CFI is a relative reduction in model non-centrality, and – when the baseline model is the null model – the denominator can be seen as a proxy for the amount of generalized variance in the manifest variables of the model, the determinant of the observed correlation matrix $|\mathbf{R}|$. An interpretation of CFI in terms of absolute magnitude would require an interpretation of the amount of generalized variance, that is the value of this determinant. The determinant of a correlation matrix can be seen geometrically as the volume of the swarm of standardized data points, with $|\mathbf{R}| = 1$ in case of all zero-correlations (corresponding to a 'ball' in a multidimensional plane) and with $|\mathbf{R}| = 0$ for a matrix with perfect linear dependence (a ball flattened along at least one dimension). Whereas people in practice often already find it hard to interpret the absolute magnitude of a variance, it is fair to say that even less people have a good intuition about what a large or small generalized variance or determinant is for their dataset. The current lack of straightforward interpretability of CFI in terms of absolute magnitude essentially disqualifies it in practice for assessing the absolute fit of a single model or for comparing model fit between different datasets.

Nevertheless, the central role of this determinant should revive some interest in understanding classic measures of multivariate statistics (e.g., Anderson, 1958) to further our understanding of more modern SEM practices. In the meantime, we recommend implementing a reporting standard where next to the CFI also its denominator, the baseline model's noncentrality λ_0 is reported to provide some context for interpretation. These quantities are generally available or easy to request in common SEM software such as Mplus or R:lavaan. If the default null model is chosen as baseline, explicit reporting of its three key components – sample size n, number of manifest variables p, determinant of the observed correlation matrix $|\mathbf{R}|$ – would help in gaining some intuition on common reference values for these data characteristics[2] in your field of application and eventually allow for a better interpretation of relative and absolute magnitude of CFI even across datasets.

Other Considerations. One limitation of the current study is that we only considered the default null model in which all observed variables are uncorrelated while looking at the performance of CFI. However, it was already discussed by Bentler and Bonett (1980, p. 604) that "the incremental fit indices depend critically on the availability of a suitable framed null model". Widaman and Thompson (2003) argue that there are numerous situations in which the default null model would be an improper choice. Different alternatives for specification of a proper baseline model can be found in the literature (e.g., Little, 2013; Widaman & Thompson, 2003). While Widaman and Thompson (2003) already touched upon it, going forward it is important to systematically evaluate the potential influence of the chosen null model on performance evaluation of the different comparative fit indices under different circumstances, as well as the substantive consequences of comparing a model of interest to a more meaningful baseline model.

In this study, we focused on the typical maximum likelihood estimator used in structural equation modelling, yet it would be of interest to expand the study to other estimators in particular for the categorical data case, both including limited-information estimators based on the polychoric correlation matrix or bivariate contingency tables as well as full-information estimators

[2] In a linear model, it is similarly good practice to report next to the R-square also the total variance of the outcome variable (or alternatively the residual standard deviation) to contextualize the percentage.

based on the item response patterns (cf. item response theory tradition). A move to the categorical case might also essentially call for a different baseline model; for categorical data, correlations are strongly constrained by their marginal distributions as mean and variance are intertwined.

Another avenue for further research would be to explore the impact of transitioning from classic estimates for the two noncentrality parameters in the CFI to bias-corrected estimates as for instance suggested by Raykov (2005). Raykov did add caution as for instance a bias-correction bootstrap estimate of noncentrality is feasible, but the properties of the approach for this particular case have not been fully studied. Yet deflating differential sampling bias in both numerator and denominator of CFI could potentially ensure that its sampling behavior is even more systematic and in line with the driving components of the baseline.

Conclusion

To conclude, the CFI does what it is supposed to do, but we haven't been using it in a smart fashion. The CFI is a relative fit measure where the standard for comparison is provided by the noncentrality of the (null) baseline model. The common CFI ≥ .95 rule of thumb implies that regardless of context we are happy with a reduction of 95% of the misspecification by the null model. Current practices make us prone to hunting down this magic CFI ≥ .95 value as a pseudo absolute fit measure disregarding the existence of the baseline. CFI as an absolute but meaningless criterion that needs to be fulfilled to achieve an adequate model that can serve as starting point for further analysis. To help remedy this, we recommend that at a minimum a dual reporting standard is followed where both model of interest and the (null) baseline model are evaluated to provide proper context for interpretation of the CFI value. By making the presence of the baseline (and its core components) explicit in the reporting, the need to take it into account when interpreting fit indices also becomes explicit and non-ignorable. Even more optimal would be if CFI is not simply used as a mere number in a search for model adequacy but used as a real relative fit index intended to evaluate the relevance of cumulative theoretically motivated model restrictions in terms of % reduction in misspecification as measured by the baseline model (Bentler & Bonett, 1980).

References

Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242.

Brown, T. A. (2015). Confirmatory Factor Analysis for Applied Research (2nd). Guilford Press.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*(1), 1–36.

Graybill, F. A. (1983). *Matrices with Applications in Statistics* (2nd). Wadsworth International Group.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (6th). Pearson.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336.

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

Jackson, D. L., Gillapsy, J. A., & Purch-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23.

Kenny, D. A. (2015). Measuring model fit. http://davidakenny.net/cm/fit.htm

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assess- ment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98.

Moshagen, M., & Auerswald, M. (2018). On congruence and incongruence of measures of fit in strucutral equation modeling. *Psychological Methods*, *23*(2), 318–336.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620.

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148–1172.

R Core Team. (2020). R: A language and environment for statistical computing.

Raykov, T. (2005). Bias-corrected estimation of noncentrality parameters of covariance structure models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(1), 120–129.

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, Article 1715.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310–334.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21–40.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, *50*(3), 253–263.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37.

Citation:

van Laar, S., & Braeken, J. (2021). Understanding the Comparative Fit Index: It's All About the Base! *Practical Assessment, Research & Evaluation*, 26(26). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/26/

Corresponding Author

Saskia van Laar
Centre for Educational Measurement at the University of Oslo (CEMO)
Oslo, Norway

email: s.van.laar [at] cemo.uio.no

Appendix A: Noncentrality λ_0 of the null model

$$\lambda_m = \chi^2_m - \mathrm{df}_m \tag{1}$$

$$= F_m(n-1) - \mathrm{df}_m \tag{2}$$

$$= (\log|\hat{\boldsymbol{\Sigma}}_m| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\hat{\boldsymbol{\Sigma}}_m^{-1}) - p)(n-1) - \mathrm{df}_m \tag{3}$$

Equations 1-3 outline how the noncentrality parameter of any model would be estimated as the difference between the model's chisquare against the saturated model and the model's degrees of freedom. The model's chisquare value is based on the product of the sample size n and the minimum value F_m of the used fit function. Under maximum likelihood estimation, F_m is a function of the discrepancy between the model-implied variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ and the observed variance-covariance matrix $\mathbf{S}$ (e.g., Bollen, 1989), where p represents the number of observed variables and $\mathrm{tr}(\mathbf{X})$ and $|\mathbf{X}|$ are respectively the trace and determinant of a matrix $\mathbf{X}$.

Key in getting to the expression for the noncentrality λ_0 for the null model (Equation 2 in the main text) is that the minimal fit value F_0 for the null model can be further simplified using the fact that the model-implied covariance matrix under the null model comes down to a diagonal matrix $\mathbf{diag}(\mathbf{S})$ with the observed variances on the diagonal (cf. Equation 5). This results in $\mathbf{S}\hat{\boldsymbol{\Sigma}}_0^{-1}$ leading to a matrix with all ones on the diagonal such that the trace equals the number of observed variables p and cancels out the subsequent $-p$ term in the expression for F_0 (cf. Equation 6).

$$F_0 = \log|\hat{\boldsymbol{\Sigma}}_0| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\hat{\boldsymbol{\Sigma}}_0^{-1}) - p \tag{4}$$

$$= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{diag}(\mathbf{S})^{-1}) - p \tag{5}$$

$$= \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| + p - p \tag{6}$$

Using the fact that the determinant of a matrix product can be split into products of determinants, each of the remaining two log determinants can be written out given that a variance-covariance matrix $\mathbf{S}$ is a multiplicative function of a corresponding

correlation matrix $\mathbf{R}$ and an inverse diagonal matrix with standard deviations on the diagonal. Thus we have

$$\log|\mathbf{S}| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\,\mathbf{R}\,\sqrt{\mathbf{diag}(\mathbf{S})}| \tag{7}$$

$$= \log|\sqrt{\mathbf{diag}(\mathbf{S})}| + \log|\mathbf{R}| + \log|\sqrt{\mathbf{diag}(\mathbf{S})}| \tag{8}$$

$$= \log\prod_{j=1}^{p}\sqrt{S_{jj}} + \log|\mathbf{R}| + \log\prod_{j=1}^{p}\sqrt{S_{jj}} \tag{9}$$

$$= \log\prod_{j=1}^{p} S_{jj} + \log|\mathbf{R}| \tag{10}$$

and

$$\log|\mathbf{diag}(\mathbf{S})| = \log|\sqrt{\mathbf{diag}(\mathbf{S})}\,\mathbf{I}\,\sqrt{\mathbf{diag}(\mathbf{S})}| \tag{11}$$

$$= \log\prod_{j=1}^{p} S_{jj} + 0 \tag{12}$$

where Equation 12 makes use of the fact that the correlation matrix of a diagonal variance-covariance matrix is an identity matrix $\mathbf{I}$ which determinant is exactly equal to 1.

The re-expressions of the log determinant terms in Equations 10 and 12 allow to simplify the expression for F_0 further by elimination

$$F_0 = \log|\mathbf{diag}(\mathbf{S})| - \log|\mathbf{S}| \tag{13}$$

$$= \log\prod_{j=1}^{p} S_{jj} - \log\prod_{j=1}^{p} S_{jj} - \log|\mathbf{R}| \tag{14}$$

$$= -\log|\mathbf{R}| \tag{15}$$

such that the estimated noncentrality of the null model comes down to

$$\lambda_0 = F_0(n-1) - \mathrm{df}_0 = -\log|\mathbf{R}|(n-1) - p(p-1)/2$$

where $p(p-1)/2$ is the degrees of freedom of the null model.

Table B1. CFI and its underlying noncentrality measures in numerator and denominator as a function of data correlation **R**, sample size n, and number of variables p.

		$p = 4$							$p = 8$						
n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05
100	0.1	0.948	0.888	0.0	6.3	0.3	18.8%	0.74	0.813	0.608	0.0	22.7	2.7	41.6%	0.62
	0.2	0.819	0.777	13.9	19.7	0.6	20.1%	0.81	0.503	0.388	40.7	69.0	3.1	34.3%	0.83
	0.3	0.652	0.615	36.8	43.7	0.7	11.3%	0.91	0.255	0.203	108.5	137.0	3.4	19.2%	0.90
	0.5	0.312	0.304	110.3	116.2	0.8	1.6%	0.96	0.035	0.031	306.8	334.6	3.0	1.8%	0.96
	0.7	0.084	0.083	242.1	249.1	0.8	0.0%	0.98	0.001	0.001	637.3	667.0	3.1	0.0%	0.98
	0.9	0.004	0.004	553.9	560.8	0.9	0.0%	0.99	0.000	0.000	1385.0	1409.1	3.1	0.0%	0.99
		$p = 12$							$p = 24$						
	0.1	0.659	0.331	0.0	46.7	6.0	51.7%	0.61	0.292	0.015	0.0	152.6	32.2	84.2%	0.57
	0.2	0.275	0.145	63.1	133.0	5.7	34.4%	0.85	0.033	0.002	65.0	371.9	31.4	68.2%	0.81
	0.3	0.085	0.048	180.5	250.6	6.8	21.6%	0.91	0.002	0.000	337.7	653.9	32.3	45.8%	0.89
	0.5	0.003	0.002	509.3	583.2	6.6	2.6%	0.96	0.000	0.000	1065.7	1372.4	32.3	7.6%	0.95
	0.7	0.000	0.000	1042.0	1105.5	6.8	0.0%	0.98	0.000	0.000	2209.2	2517.1	32.3	0.0%	0.97
	0.9	0.000	0.000	2228.0	2294.2	6.4	0.0%	0.99	0.000	0.000	4712.2	5005.5	32.1	0.0%	0.98

Note. $\lvert\mathbf{R}_\Sigma\rvert$= determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$= average sample noncentrality for the estimated true model; < .95= model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

— Table B1 continued —

n	r	$p = 4$							$p = 8$						
		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	<.95	Q.05
200	0.1	0.948	0.919	4.7	11.2	0.5	20.6%	0.76	0.813	0.704	13.4	43.0	2.7	31.8%	0.76
	0.2	0.819	0.795	33.9	40.5	0.7	12.9%	0.91	0.503	0.439	109.3	139.4	2.7	14.8%	0.91
	0.3	0.652	0.637	79.6	85.1	0.8	4.4%	0.95	0.255	0.225	245.1	275.3	2.7	3.0%	0.96
	0.5	0.312	0.308	226.6	232.3	0.7	0.1%	0.98	0.035	0.033	641.6	667.1	2.9	0.0%	0.98
	0.7	0.084	0.083	490.1	498.0	0.8	0.0%	0.99	0.001	0.001	1302.6	1323.8	2.8	0.0%	0.99
	0.9	0.004	0.004	1113.9	1119.3	0.7	0.0%	1.00	0.000	0.000	2798.0	2831.5	2.8	0.0%	1.00
		$p = 12$							$p = 24$						
	0.1	0.659	0.472	17.4	85.8	5.2	37.6%	0.78	0.292	0.071	0.0	260.1	17.6	47.4%	0.80
	0.2	0.275	0.202	192.3	258.9	4.9	12.4%	0.93	0.033	0.009	405.9	696.4	18.1	17.6%	0.92
	0.3	0.085	0.064	427.0	494.8	5.6	3.3%	0.96	0.002	0.001	951.3	1251.3	19.3	2.7%	0.96
	0.5	0.003	0.003	1084.6	1149.8	5.4	0.1%	0.98	0.000	0.000	2407.3	2692.8	19.0	0.0%	0.98
	0.7	0.000	0.000	2150.1	2222.3	5.4	0.0%	0.99	0.000	0.000	4694.5	4991.7	18.4	0.0%	0.99
	0.9	0.000	0.000	4521.9	4586.6	4.8	0.0%	1.00	0.000	0.000	9700.4	9991.0	17.9	0.0%	0.99

Note. $\lvert\mathbf{R}_\Sigma\rvert$= determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\bar{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\bar{\lambda}_m^{(S)}$= average sample noncentrality for the estimated true model; < .95= model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

— Table B1 continued —

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\bar{\lambda}_0^{(S)}$	$\bar{\lambda}_m^{(S)}$	<.95	Q.05
		$p=4$							$p=8$						
	0.1	0.948	0.937	20.9	26.8	0.7	17.5%	0.87	0.813	0.770	75.4	103.4	2.7	20.1%	0.89
	0.2	0.819	0.811	93.7	99.3	0.8	3.1%	0.96	0.503	0.477	315.3	344.2	2.6	0.7%	0.97
	0.3	0.652	0.647	208.1	212.8	0.8	0.2%	0.98	0.255	0.243	654.7	684.9	2.7	0.0%	0.98
	0.5	0.312	0.311	575.6	581.9	0.7	0.0%	0.99	0.035	0.034	1646.0	1679.0	2.7	0.0%	0.99
	0.7	0.084	0.084	1234.3	1241.0	0.7	0.0%	1.00	0.001	0.001	3298.4	3323.2	2.7	0.0%	1.00
	0.9	0.004	0.004	2793.7	2803.4	0.7	0.0%	1.00	0.000	0.000	7037.1	7060.3	2.8	0.0%	1.00
500		$p=12$							$p=24$						
	0.1	0.659	0.577	142.5	210.9	4.7	16.7%	0.91	0.292	0.168	338.7	620.6	11.2	10.8%	0.94
	0.2	0.275	0.243	579.7	646.1	4.3	0.2%	0.97	0.033	0.020	1428.8	1708.3	12.1	0.0%	0.97
	0.3	0.085	0.076	1166.4	1233.0	4.3	0.0%	0.98	0.002	0.001	2792.3	3073.9	11.7	0.0%	0.99
	0.5	0.003	0.003	2810.4	2879.2	4.0	0.0%	0.99	0.000	0.000	6432.3	6714.4	13.2	0.0%	0.99
	0.7	0.000	0.000	5474.2	5549.3	4.3	0.0%	1.00	0.000	0.000	12150.1	12454.3	11.5	0.0%	1.00
	0.9	0.000	0.000	11403.8	11487.5	4.5	0.0%	1.00	0.000	0.000	24665.1	24938.3	12.3	0.0%	1.00

Note. $\lvert\mathbf{R}_\Sigma\rvert$= determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\bar{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\bar{\lambda}_m^{(S)}$= average sample noncentrality for the estimated true model; < .95= model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.

— Table B1 continued —

n	r	$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI		$\lvert\mathbf{R}_\Sigma\rvert$	$\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$	noncentrality			CFI	
				$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05			$\lambda_0^{(\Sigma)}$	$\overline{\lambda}_0^{(S)}$	$\overline{\lambda}_m^{(S)}$	<.95	Q.05
		$p = 4$							$p = 8$						
1000	0.1	0.948	0.941	47.7	55.0	0.7	8.9%	0.93	0.813	0.790	178.9	208.4	2.5	5.3%	0.95
	0.2	0.819	0.815	193.4	198.6	0.7	0.1%	0.98	0.503	0.491	658.5	686.5	2.7	0.0%	0.98
	0.3	0.652	0.648	422.2	428.4	0.7	0.0%	0.99	0.255	0.249	1337.3	1365.9	2.8	0.0%	0.99
	0.5	0.312	0.311	1157.2	1164.5	0.7	0.0%	1.00	0.035	0.035	3320.0	3351.0	2.7	0.0%	1.00
	0.7	0.084	0.084	2474.5	2475.0	0.7	0.0%	1.00	0.001	0.001	6624.9	6634.5	2.6	0.0%	1.00
	0.9	0.004	0.004	5593.4	5596.0	0.7	0.0%	1.00	0.000	0.000	14102.2	14120.4	2.7	0.0%	1.00
		$p = 12$							$p = 24$						
	0.1	0.659	0.617	351.0	419.0	4.5	2.8%	0.95	0.292	0.222	953.4	1236.3	10.0	0.6%	0.97
	0.2	0.275	0.258	1225.4	1292.3	4.0	0.0%	0.99	0.033	0.025	3133.5	3416.4	10.7	0.0%	0.99
	0.3	0.085	0.081	2398.8	2458.1	4.0	0.0%	0.99	0.002	0.002	5860.7	6151.5	11.1	0.0%	0.99
	0.5	0.003	0.003	5686.8	5755.9	4.4	0.0%	1.00	0.000	0.000	13140.7	13412.6	9.7	0.0%	1.00
	0.7	0.000	0.000	11014.4	11058.4	4.2	0.0%	1.00	0.000	0.000	24576.3	24861.0	11.3	0.0%	1.00
	0.9	0.000	0.000	22873.7	22938.7	4.3	0.0%	1.00	0.000	0.000	49606.1	49921.5	10.4	0.0%	1.00

Note. $\lvert\mathbf{R}_\Sigma\rvert$= determinant of the population correlation matrix as expression of the degree of multivariate dependence; $\lvert\overline{\mathbf{R}}_\mathbf{S}\rvert$ = average determinant of the sample correlation matrices; $\lambda_0^{(\Sigma)}$ = population value of the null baseline noncentrality; $\overline{\lambda}_0^{(S)}$ = average sample value of the null baseline noncentrality; $\overline{\lambda}_m^{(S)}$= average sample noncentrality for the estimated true model; < .95= model rejection rate or percentage of replications where the sample CFI value for the estimated true model is below .95; Q.05 = 5% quantile of the CFI sample values for the estimated true model.