İSTANBUL ÜNİVERSİTESİ YAYINEVİ

Istanbul Management Journal

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Istanbul Management Journal, 88: 1-20

DOI: 10.26650/imj.2020.88.0001
http://imj.istanbul.edu.tr

Submitted: 14.10.2019
Revision Requested: 01.11.2019
Last Revision Received: 18.11.2019
Accepted: 06.05.2020

How and When to Use Which Fit Indices? A Practical and Critical Review of the Methodology

Muhsin Murat Yaşlıoğlu[1] **Duygu Toplu Yaşlıoğlu**[2]

Abstract

With the help of statistical software programs, such as AMOS, Lisrel, R, Matlab, and many equivalents, most of the complicated research models have become more computable and easily understandable. Even the most complicated and complex models with various relationships can be easily computed with the help of software. Although with slight differences, outputs are consistent, and tables are mostly comprehensible. However, with the increasing curiosity and amount of knowledge about the research methodology, these simple looking outputs start to become more complicated and deeper. Even though aforementioned statements seem contradictory, what we imply here is very sound to a mid-level researcher because, as knowledge and understanding of statistics deepens, questions and doubts about from where, how, and why these numbers are calculated increase. Curiosity about the fit indices, chi-square and degrees of freedom, modification indices, covariances, and residuals begin to arouse.

In this review and commentary, we focus on the infamous CMIN (or chi-square), different model definitions, and calculation of fit indices by the help of these models while avoiding statistical jargon as much as possible. With the aim of putting an end to a decade long debate, when and how to use which fit indices, what they really indicate, and which numbers refer to good or bad fit is also discussed.

Keywords

Fit Indices, CFA, SEM, Chi-Square, Discrepancy Functions, Modification Indices, Maximum Likelihood

1 **Corresponding author:** Muhsin Murat Yaşlıoğlu (Doç. Dr.), İstanbul Üniversitesi, İşletme Fakültesi, İşletme Bölümü, İstanbul, Türkiye. E-posta: muratyas@istanbul.edu.tr ORCID: 0000-0003-2464-5439

2 Duygu Toplu Yaşlıoğlu (Dr. Öğr. Üyesi), İstanbul Üniversitesi, İşletme Fakültesi, İşletme Bölümü, İstanbul, Türkiye. E-posta: duygut@istanbul.edu.tr ORCID: 0000-0002-5637-8999

To cite this article: Yaslioglu, M. M., & Toplu-Yaslioglu, D. (2020). How and when to use which fit indices? A practical and critical review of the methodology. *Istanbul Management Journal,* 88: 1-20. http://doi.org/10.26650/imj.2020.88.0001

Introduction

In order to be able to comprehend structural equation modelling, confirmatory factor analysis, and fit concepts, some literature and definitions are necessary to clarify. First of all, we have to begin with the definition of the models mentioned in the software. These are most of the time confusing, not only because of their nature but because they are never really described anywhere in the process. Secondly, the estimation methodology is to be defined. Most SEM users, no matter what software they use, are accustomed to the "Maximum Likelihood Estimation," but a great deal of these researchers have no idea what it is and what it does. Finally, there are several concepts which need to be clarified before discussing fit of the models. Some of these concepts are CMIN, Chi-Square, Log-Likelihood (see also Maximum Likelihood), C and F values, NPAR, p and P, fit and index, and PRATIO. Despite sounding familiar, most of the time they are misleading, confusing, bewildering, and even confounding. We see them and think we know them, but we never think about what they really are or where they come from. Alongside the discussion of several concepts and terminology, the necessary values and key-points will be discussed throughout the paper.

Models Definitions

Despite slight naming differences among statistical software, there are three main models essential for the calculation of SEM and CFA. First is the one that researchers want to investigate, which is called "default model," "structural model," or "measurement model". Following is the one in which every measured variable is accounted as independent of each other and any latent variable. This is called "independence model;" it also is the "baseline/null model" for CFA and SEM – we will discuss this confusion further below. Final is "saturated model" in which all variables covary with every each other.

Since its name will be mentioned several times here, before we begin defining models, an introduction to a fairly common concept called parsimony is also essential. Parsimony means simplicity, so the parsimonious models are simple models with less parameters to be estimated. Of course, parsimony of a model can only be judged relatively, often comparing nested models.

Nested models are the models in which one of the models contains all the variables, parameters, and interactions of the other and at least one extra term (parameter, constraint e.g.). Extended model is called the full (or complete) model, and abridged is called the restricted (or reduced) model; hence, saturated, default, and independent models are all nested models, where saturated is the full model and default is the restricted version of it (so is the independent).

Besides, there is a ratio called PRATIO (parsimony ratio) which compares the degrees of freedom for default model (df) and independence model (df_i). Its formula is simply PRATIO=df/df_i. This ratio is also used to calculate "Parsimony adjusted measures of fit" or namely PNFI and PCFI. These will be discussed later in this paper.

Aforementioned models are:

Saturated model: This is the fully explanatory model in which per every degree of freedom, there are as many parameter estimates; therefore, df_s=0. That is to say every variable in the model co-varies with every each other. This is the most general model possible. Goodness of fit measures are "1.0" for this model. Besides, some measures such as RMSEA cannot be computed for saturated model, and because saturated model, by its nature, is the most un-parsimonious model possible, parsimony-based fit measures (PNFI, PGFI etc.) will be 0. It is an inane and illogical model in the sense that it is guaranteed to fit perfectly to any set of data collected. Any other model in the same research (that also implies the same dataset) is a nested (constrained) version of the saturated model.

Null or baseline model (AKA independence model in AMOS and some other software): The comparison model is frequently used as the "baseline model," differences from which must be significant if a proposed structural model (the one with straight arrows connecting some latent variables – also called the default model in AMOS) is to be investigated further; however, the term "baseline model" implies comparison with an alternative that is more complex than a no-effect hypothesis. The terms "naïve model" and "null model" better indicate the kinds of models that researchers have used as baselines so far (Schwab, A., & Starbuck, W. H., 2013).

In the SEM or CFA baseline model, the covariances in the covariance matrix among the latent variables are all assumed to be zero. Despite its official name, AMOS and several other statistical software name "null/baseline" model as "independence model". It makes sense because the independence model is the one which assumes all relationships among measured variables are "0." Independence model is an uncorrelated variables model, and for computation, many fit measures, such as TLI=NNFI, RFI, IFI, NFI, CFI, PNFI, and PCFI, necessitate a "null/baseline" model in comparison with researchers' measurement model. This model assumes that variables or latent factors of a construct are uncorrelated. Unlike the saturated model which have a parsimony ratio of "0," the independence model has a parsimony ratio of "1."[1] Most of the fit measures will have a value of "0" since this is the worst model possible, whether parsimony-adjusted or not. In rare occasions, some fit indices, such as RMSEA and GFI, may have a non-zero value depending on the data (Schermelleh-Engel, K., et.al., 2003).

1 Please refer to PRATIO in this paper.

Default (structural or measurement) model: This is the researcher's measurement or structural model (AMOS calls it the "default model"). In comparison to saturated model, this model is always more parsimonious, and it is always better fitting than the independence model when compared using fit indices. Thus, the default model will have a goodness of fit between the perfect fitting "saturated model" and worst possible model with lowest explanatory power, "the independence model."

Estimation and Maximum Likelihood Estimation

Even though there is no simple way to describe Maximum Likelihood Estimation (MLE), it is essential to say this method is the default for many statistical software in order to be able to calculate many of the fit indices. Its complexity should not be taken for granted; however, some concepts about the estimation process and routines can be elaborated.

There are several estimation techniques, most of them perform one of three things (Templin, J. 2015):

1. Minimize some function: If the estimation process includes the word "least" in its name, then minimization should be expected. Most of these techniques minimize the squares of the error terms (or std. deviations). Types of least squares techniques include ordinary, generalized, weighted, WLSMV, iteratively re-weighted, and diagonally weighted. It is usually conducted as a last resort.

2. Maximize some function: Mostly, this gold standard of estimation techniques comes with the name "maximum" in it, such as maximum likelihood, residual maximum likelihood, and robust maximum likelihood.

3. Usage of simulation for sampling from data: These use recent advanced techniques of re-sampling through the help of recent simulation methods. Some of these include Gibbs sampling, Metropolis-Hastings algorithm, Monte Carlo simulation, and Bayesian Markov Chain Monte Carlo. These are typically used for complex models where maximum likelihood is not applicable or in which some prior values are necessary.

Simply;

(1) MLE is a procedure to determine best model parameters (reality) that fit the given data with maximizing log-likelihood function to estimate parameters. The formulas here, while being quite mathematical, are familiar to most statisticians'. But one can immediately ask: "Why not likelihood function but log-likelihood?". Simply put, mathematically its asymptotes meet at the same values, and it is way easier to find a maximum of log-likelihood since it includes "sums" rather than

"products" as likelihood function does. Additionally, one can easily understand that maximization of products is harder than sums. Since we need derivatives of functions to find out asymptotes, it is easier to take derivatives of sums.

(2) MLE also helps compare different models with the same data using some information criteria. This is mathematically even more advanced. There are formulas called information theory techniques. The most common one is Kullback-Leibler information criterion, which quantifies the distance between two given models. Since depending on full probability density functions, it is very hard to calculate (Burnham, K. P., & Anderson, D. R., 2001). Japanese statistician Hirotugu Akaike (1987) proved that K-L information could be estimated based on maximum log-likelihood and created AIC (Akaike Information Criterion). Its formula is:

$$AIC = -2(\ln(\text{ø}|x)) + 2K$$

It actually is "-2" times log-likelihood added by "2" times the number of parameters. Both log-likelihood and AIC are only meaningful when compared to other models with the same data (they are relative not absolute). They have no meaning by themselves, so the higher or lower the values mean nothing without comparison. Moreover, if you are comparing two "bad" models, they can only mean one is better than the other but cannot say anything about how bad/good they are. AMOS reports several similar model comparison values such as AIC, BCC, BIC, CAIC, ECVI, and MCVI. Keep in mind that these values are only for models' comparison and relative . They do not indicate a fit for models. Simply put, if you are to compare two nested models[2] among each other, they are handy. If not, just ignore them. Complicated, poorly fitting models get high scores. For comparison purposes, this means the lower the values the better[3].

Some Other "sine qua non" Concepts

Since we now are aware of maximum likelihood estimation and log-likelihoods, we can talk about chi-square (χ^2) values calculated per model in AMOS. It is named as "CMIN" which allegedly stands for "chi-square minimum." If one is accustomed to basic statistics, then he or she should also know about chi-square test and that it stands for "independence." This means, without terminologically using definition of hypotheses, if a χ^2 value is statistically significant ($p<0.05$) then these two observations are "independent" from each other. In CFA and SEM, it is potentially unwanted. We want our measurement model (default model in AMOS) to be "not independent" from the data of observations.

2 Two models are nested if one contains all the terms of the other, and at least one extra term.

3 Also see the "Model comparison" section below.

The problem is that it is not easy to comprehend how CMIN is calculated. As one googles chi-square, he or she will most probably end up with what we call "Pearson Chi-Square" formula saying something like, "If you subtract expected values from observed values and square them, then divide them by expected values, you end up with chi-squared for each observation." If you add them all, you find a summed chi-square value. This is what confuses most people because we have observed values on one side of the arrow since factors are unobserved (latent) variables.

Moreover, this CMIN is referred to as a fit index; therefore, it should be comparing two models, not observations. What are these two models? To evaluate the fit of the factor model, its "function of log-likelihood value" has to be compared to that of some less constrained model, such as the saturated model. The chi-square test compares the model (default model) to the saturated model (it should fit about the same). Many fit indices compare the model to the null/baseline model instead (baseline model should fit much worse than measurement model). AMOS uses function of log-likelihood to report CMIN. Chi-square is calculated through multiplying the number of samples and F_{ML} (function of ML); therefore, $C=n(F_{ML})$. C value is derived from F, and this value is also called "minimum discrepancy function."

As discussed earlier in the model definitions section, saturated and default models are nested models, where saturated is the full and default is the restricted. Difference between function of log-likelihood of two nested models also gives the chi-square. If one simply calculates function of log-likelihood for saturated and default models and takes the difference, they end up with the chi-square for default model. The number of parameters to be estimated are also subtracted (of course, saturated model has more NPAR) to end up with "df" for default model. Eventually, chi-square distribution table can be used to calculate probability and test the null hypothesis of independence.

The number of parameters to be estimated defines the complexity of the model. Models with many parameters to estimate are called complex. Less parameters means the model is simple. In AMOS and other programs, *number of distinct parameters to be estimated* is called "NPAR." The word "distinct" is also important here. For instance, if two or more parameters are required to be equal to each other, then these count as one, not two.

This leads us to another important concept in statistics, degrees of freedom (df). Degrees of freedom is the NPAR (q) subtracted from the number of sample moments (p), so the formula is (df=p-q).

One of the main fit measures (perhaps it should be called "THE" fit measure) is CMIN. It is the minimum value of C of the discrepancy, otherwise called chi-Square of likelihood ratio test. Since chi-square statistics all require a significance value,

"p value" is marked as "P" for testing the hypotheses that the model fits perfectly in population. As discussed earlier in this paper, it is the discrepancy between perfectly fit model (saturated model) and default model.

Increase in NPAR (also implying decrease in df), declines log-likelihood for the nested models using the same sample. This means, saturated model always has lower value for function of log-likelihood. Sample size increases the likelihood functions; despite the sample size being the same in the nested models, this does not mean the difference stays the same with smaller sample sizes. Chi-square test value increases as the sample size increases, and this makes the values significant since the (df) stays the same. It sounds complicated, but think of it as a test statistic of independence getting larger as the number of samples increases, which makes it more significant at a time. If two models (in our case, it is saturated and default models) are independent of each other, then they simply are not fit to each other. This is true but not necessarily correct, and this is the reason that we need more indices to be able to look at.

Here are some quotes directly from respected statisticians/researchers:

> "The power of the test to detect an underlying disagreement between theory and data is controlled largely by the size of the sample. With a small sample an alternative hypothesis which departs violently from the null hypothesis may still have a small probability of yielding a significant value of. In a very large sample, small and unimportant departures from the null hypothesis are almost certain to be detected." (Cochran, 1952)

> "If the sample is small, then the test will show that the data are 'not significantly different from' quite a wide range of very different theories, while if the sample is large, the test will show that the data are significantly different from those expected on a given theory even though the difference may be so very slight as to be negligible or unimportant on other criteria." (Gulliksen and Tukey, 1958, pp. 95–96)

> "Such a hypothesis [of perfect fit] may be quite unrealistic in most empirical work with test data. If a sufficiently large sample were obtained this statistic would, no doubt, indicate that any such non-trivial hypothesis is statistically untenable." (Jöreskog, 1969, p. 200)

"Do they mean that we should limit the sample size? Despite they sound in that manner, one should also know that "Significant properties of maximum likelihood (ML) estimate are consistency, normality, and efficiency. However, it has been proven that these properties are valid when the sample size approaches infinity. Many researches warn that a behavior of ML estimator working with the small sample size is largely unknown. (Psutka, J. V. and Psutka J., 2015)"

One logical way to assess fit is to find the discrepancy value (CMIN) per degrees of freedom, given that it tends to increase with number of sample moments. CMIN/df value can give the researcher an absolute value for fit. Arguments begin just here, because various researchers have suggested various acceptable values for this value. Wheaton and colleagues (1977) suggested 5 or less, some suggested as low as "2," or as high as "5." Byrne et.al. (1989) puts forward that $\chi^2/df > 2$ indicates bad fit.

Values less than "1" will probably require insignificant CMIN values and will therefore not be even necessary to calculate. Anything close to "1" should be very good fit, but how far apart could it fall from "1?" Let's remember the calculation of degrees of freedom (df=Sample moments - number of distinct parameters); thus, as df increases with sample size so does X2. Here we should first look at NPAR. The default model's chi-square calculation, not by chance, is the difference of NPAR between saturated model and measurement (default) model. If "df" for default model is calculated taking the number of parameters into account, this means we can ignore it simply because it is already taken into account. Sample size should be the only variable here to decide the value for CMIN/df cut point. Here we can use common sense:

(1) If the commonly accepted minimum sample size in a factor analysis is at least 50 and also 5 times the number of variables. This means minimum sample for a decent number of variables as around 150 (there is no real calculation here but merely observation).

(2) If minimum number is around 150, doubling this number seems fair for a cut point. Let's say 300 here is a cut point for sample size to categorize CMIN/df value.

(3) Then we can say, looking to our commonly mentioned cut points of CMIN/df, if sample size is between 150-300, then 3.5 (median of 2-5) can be taken as cut point to assess the fit. If sample size is above 300, then "5" can be taken as the criterion. More than 5 χ^2 per degrees of freedom indicates a bad fit regardless. This value should be less. Please read further.

(4) To decide whether a CMIN/df is good enough, one should also compare the worst model's (independence model) CMIN/df value. These values should be significantly different from each other because if worst model is fit enough, this requires measurement model to be even much fitter. Luckily, we have fit indices comparing these values.[4]

4 Please refer to relative fit indices (NFI,RFI, CFI and TLI)

Indices: Fit and Others

Before beginning to discuss anything about fit, we have to make a short list of things often confused by researchers. Researchers MUST keep in mind that:

(1) Fit has very little to do with validity: Most researchers confuse fit with validity. Validity is a much broader concept to begin with.

(2) If model is fit, this means your data is consistent with what you want to measure.

(3) If model is fit, then it is useful model.

(4) If model is fit, then it will probably be able to be replicated in other researches.

(5) If model is fit, the researcher can stop adding covariances among residual error terms.

(6) If model is fit, then the researcher can proceed with further evaluation of construct and other validities.

(7) If model is fit, it is NOT necessarily correct or valid.

(8) A good fitting model is ONLY "reasonably consistent with the data."

Strictly keeping the list above in mind, there are several indices to measure the fit of the proposed measurement model (default model). Also, there is even more debate about what to use and when to use it. Mostly, simple models, with a moderate number of sample observations, have good fit. As the models get complicated and sample size increases, these fit indices start to drop. Frequently, researchers face the dilemma of choosing between fit indices because while some are above cut points, others are below expected values. Here are some problems: what are the cut points for indices? Is there a commonly accepted value for each? What index is best for models with many variables? After being able to answer all these questions, another problem may rise: what if some of them are above expected values and some are not, who tells us which to go for, and finally, if one can solve all these issues, how are two or more similar models with the same data compared. In this section we will try to answer these questions with avoiding complicated, sophisticated jargon of statistics. This does not mean we will leave things out; this implies we will keep it as "simple and stupid" as possible.

Fit indices (measures) in AMOS are categorized into sub groups. These are: absolute fit indices, relative/incremental fit indices, parsimony (check above) fit indices, non-central chi-square distribution (population discrepancy based) fit indices, information theoretic fit indices, and fit measure based on sample size.

Absolute Fit Indices

Absolute fit indices indicate fit without comparing the default model to anything for the best fit model. Despite there being a comparison with the best fit model (saturated model), the indices indicate the model fit themselves. CMIN and CMIN/df are the basis of absolute fit indices discussed above. Other absolute fit indices include RMR and GFI.

RMR and GFI

It is the Root Mean Squared Residuals, therefore also called RMSR or SRMR. This value is simply what it says. It squares the amount by which the sample (measurement) covariances differ from their estimates. It is much like average of sum of squared errors (or residuals) in regression, yet as measurement units differ from each other, it is more relevant to carry out the calculation based on residual correlation matrix. Usually an RMR value (based on correlations) less than 0.05 indicates a good fit. This unfortunately is not a part of AMOS, but a script or manual calculation will sort out this problem. The smaller the value is the better.

Thanks to AMOS and LISREL, a more advanced version of RMR is calculated under the name of GFI (Goodness of fit). GFI compares by dividing squared weighted sum of the variances of measurement and estimation, where weighting depends on estimation method. Much like R2 in regression, it takes a value between "0-1." It is not suggested to use this index since it is affected by sample size. There also is a "df" adjusted version called AGFI, if one wants to use it, this one should be preferred. A "GFI" value larger than 0.95 can be accepted as good fit, preferably larger in small sample sizes and less parameters. GFI is greatly affected by sample size, so simply do not use this index (Kenny, D.A., 2005).

Incremental Fit Indices

These fit indices are also called relative or comparative indices because these indices or measures are based on the idea that things may be worse. There always (hopefully always; if not, do not even bother testing the model) is a worse model than default model, where each observation is taken into account as independent. Independent model is also called, due to its nature for comparison, baseline or null model.

Researchers may immediately ask why use the worst model but not the best. The answer is hidden in the calculation. As defined earlier, C (in Amos CMIN or in some cases F[5]) value is calculated with the help of perfectly fit model, which is also the "saturated model" namely. This model is the best fit model to the data. Please remember, fit and validity are two different things!

5 Discussed under title "Fit measures based on population discrepancy"

NFI and TLI

Relative fit measures are NFI (Normed Fit Index), RFI (Relative Fit Index), IFI (Incremental Fit Index), TLI (Tucker Lewis Index), and CFI (Comparative Fit Index). NFI is calculated using minimum discrepancy (CMIN – Chi-Square) of default model with CMIN of independent model. NFI gets a value between "0-1," where a value of "1" represents perfect fit to data. The higher the difference between model and worst fit results in a bigger value. A value of 0.90 and above is accepted to represent acceptable fit. The fit can be overestimated if the number of parameters is increased. RFI is the "degrees of freedom" corrected version of NFI; therefore, it solves the issue of parameter increase. It gets a value between 0 and 1 like NFI, and values above 0.90 is acceptable. For both NFI and RFI, smaller sample size tends to inflate the values; therefore, it is mostly suitable for larger samples. For smaller sample sizes, 0.95 is acceptable.

CFI

CFI is also "df" corrected versions of NFI. This time, it is not divided but rather subtracted. For every parameter estimated, there is just one penalty. With larger samples and low number of parameters change, values tend to be very close to NFI. CFI may get values larger than "1" but "1" is always reported as maximum. Value of "1" does not indicate perfect fit but simply means "df" of default model is larger than chi-square of the default model.

TLI, also called Non-normed fit index, is very similar to RFI. Lower "chi-square to df ratios" indicates a better fit. TLI and CFI depend on the average size of correlations in the model. If the average correlation among variables is low, values are also low. That being said, if several experimental variables (uncorrelated) are added to the default model, then this decreases the value of TLI (also CFI). A suggestion here can be that if the research model has several experimental or control variables, then TLI and CFI are not to be suggested. Values above 0.90 are acceptable, and 0.95 indicates good fit. If the model has very strongly or very weakly correlated variables, then the suggestion is to ignore these indices.

Fit Measures Based on Population Discrepancy

F0 and RMSEA

As discussed earlier, the function of discrepancy or log-likelihood, in Amos is presented as chi-square, "n" value being sample size minus number of groups (n=N-g ; g is mostly 1 in our cases) Steiger, Shapiro, and Browne (1985) proved ($C=n.F_0$) under certain conditions has a noncentral chi-square distribution with df degrees of freedom and non-centrality parameter Delta=(C -df) =nF_0. This results in F_0= [(C-df) / n] (or simply and generally; F_0= [(C-df)/ (N-1)]. Non centrality parameter is then used to

compare two nested model, such as default and saturated models. The problem here is that F_0 always favors complex models and will never favor the simpler model, or in other words, parsimonious model. Steiger and Lind (1980) suggested compensating for the effect of model complexity by dividing F_0 by the number of degrees of freedom for testing the model. This ratio then gives us "mean square error of approximation" (this makes sense since discrepancy function is a square). Taking the square root of the resulting ratio gives the population "root mean square error of approximation," or simply RMSEA. The calculation, in mathematical terms, favors larger sample size or df. Just like TLI, if chi-square equals to df, then the value becomes "0." One can simply expand the calculation by rewriting F_0 as "(χ^2-df) /n." Thr formula becomes "[(χ^2-df) / (df.n)]," and size effect of "df" will be more obvious. The smaller the "df" is the larger the RMSEA is, even with very small chi-square.[6] This may indicate a "bad fit" since RMSEA values below 0.08 indicates an acceptable, and 0.05 indicates a good fit. The suggestion is to use RMSEA in high df values and not even compute with low values or to at least be very cautious when you have low df.

PCLOSE

PCLOSE is actually a "p" value, something we are familiar seeing in almost every statistical analysis; however, this time it should not be confused with the p value of chi-square (where H_0; RMSEA=0) which stands for exact fit. This makes sense because it stands for a "close fit." Browne and Cudeck (1993), based on experience with SEM and RMSEA, argue that a RMSEA of 0.05 or less points to a good (close) fit; hence it calculates p value for null hypothesis of H_0; RMSEA<=0.05. When PCLOSE is significant, null hypothesis is rejected, indicating lack of close fit. PCLOSE should be insignificant to indicate good fit.

Parsimony Adjusted Fit Indices

James and colleagues (1982) and Mulaik and colleagues (1989) suggest adjusting NFI and GFI by multiplying indices with a ratio called PRATIO. PRATIO, as mentioned earlier in the related section of this paper, compares the degrees of freedom for default model (df) and independence model (df_i). The formula is simply PRATIO=df/df_i. AMOS also calculates PGFI by using the same method. Usually and debatably, values above 0.80 indicate a good fit. The quotation below clarifies the use of parsimony indices:

> "Although many researchers believe that parsimony adjustments are important, there is some debate about whether or not they are appropriate. I see relative fit indices used infrequently in the literature, so I suspect most researchers do not favor them. My own perspective is that researchers should evaluate

6 For instance a chi-square value of 2 (obviously not significant) with 1 df and 90 samples will give out an RMSEA of 0.106. sqrt((2-1)/1.(90-1)) = 0.106

> model fit independent of parsimony considerations, but evaluate alternative theories favoring parsimony. With such an approach, we would not penalize models for having more parameters, but if simpler alternative models seem to be as good, we might want to favor the simpler model."(Newsom, J. T., 2018)

Modification Indices

Modification indices show us how much chi-square (test statistics) will decrease if covariance is added among error terms of mentioned variables. It is only informational for CFA or SEM. Given a poorly-fitting model, you may want to know what path(s) you could add to make it better. If you change something according to MIs, then it is exploratory in nature. Be alert. This will be further evaluated below.

Also, adding paths looking to MIs makes the consecutive models nested to each other; therefore, one can use the model comparisons based on chi-square as mentioned below.

How much MI value is worth intervention? Actually, there is no certain limit to this. MI values show the test statistics (chi-square or CMIN) change since models are nested by nature. Change in CMIN may not mean much if it does not change the fit. Researchers may individually calculate a rough estimate for CMIN/df change by dividing the highest MI value with the "df." If the decrease in CMIN/df seems significant, then the covariance or path may be added. If not, then it seems negligible. This can be done as many times as the model is re-estimated; however, the user should be cautious in their use of MIs. If new models are developed with the help of MIs, then it must be reported. Do not pretend that you have a theoretical reason for part of a model that was put there because it was suggested by MI indices table! This is simply fraud. Using MIs makes the analysis exploratory by nature. This means if you are to use MI to correct the model, then this should be reported as exploratory SEM. The second option is that you reserve a part of the data to first explore, then use the remaining part to confirm (lesser evil).

Model Comparisons

Comparing two good models among each other is a nice comparison. If you are comparing two bad models, then it is a burden, and moreover, it leads to nothing but choosing the lesser evil. How good your model is is not described in this paper because it not only depends on fit indices or other values, such as AVE, MSV, or ASV (also not described here), but also theoretical background and other validity questions. Model comparisons only and simply compare two or more models. Do not assign more value to them, and do not fall into the mistake of calling a better model valid!

If one wishes to compare models, there are few criteria. Some of these information criteria are also reported with AMOS:

- The model with lower AIC (mentioned before) or BIC (Bayesian information criteria-not mentioned in this paper) is better, but, again, these are relative numbers. They do not indicate an absolute fit. Simply note down the models' AIC and BIC values, and compare them.
- If models are significantly different from each other, then a complicated version is better
- If models are not significantly different, then a simpler version is preferable.

If models are nested (such as default and saturated models mentioned earlier), then:

- Log-likelihood functions can be calculated, and difference among them with df can be used in chi-square distribution to test their difference. Added paths or deleted paths on a model make them nested to each other, so, one can compare their log-likelihoods. (This is not in AMOS by default, but R, Matlab, or AMOS scripts can be used to calculate).

As a rule of thumb, CFA is used to "confirm" a factor structure or a measurement model. Therefore, any changes made to this model will take it apart from confirmation and will make it exploratory in nature. Model comparisons are mostly suggested for exploratory SEM or path model comparisons.

Conclusion and Notes on Fit Indices

Several researchers and statisticians suggest different values and cut-points for different so-called useful fit indices. Individual researchers should keep in mind some notes about fit indices:

- Normality affects absolute fit indices. Non-normal data inflates chi-square and, therefore, decreases absolute fit values. Incremental and population discrepancy measures are less affected (Kenny, D. A., 2015).
- Number of variables affect fit. Increasing the variables decreases the fit. RMSEA, especially, increases (we do not want this) as more variables are added. Indices such as NFI, TLI, and CFI are relatively more stable but also declines slightly in such case, which is all probably because of an inflated chi-square.
- BIC, RMSEA, and TLI requires parsimony the most (also respectively among each other), and NFI and CFI requires it the least.
- NFI does not adjust for sample size. Increasing sample size decreases the fit value. TLI and CFI are relatively stable with sample size, and variation decreases between larger sample sizes. RMSEA, however, declines with sample size. Larger sample researches favor RMSEA.

- While testing for exact fit, a researcher should go for insignificant CMIN, which is almost always impossible (Unless with very few variables and a small sample size).
- To assess a good or close fit, researchers may go for different values;
    - RMSEA (below 0.05 to 0.08): If the model is parsimonious and sample size is large, then below 0.05 or closer values; otherwise, 0.08 or below.
    - CFI, RNI, NFI, TLI, RFI, IFI (above 0,90 to 0,95): Depending on variable size, variables below 10-12 require 0.95 for close fit, variables above 12 may require 0.90 as cut-point. The higher is always the better.
    - RMR below 0.05 or 0.08 for larger samples and GFI, preferably 0.90 or above. It is preferable not to use these indices.
    - For comparing models (almost always nested models), information criteria, such as (AIC, BIC e.g.), are useful.
    - For gradual comparisons and model refining, Modification Indices are very beneficial.
    - Assigning names to nested models in AMOS and using these to calculate likelihood ratios is the best way for model comparisons. (This requires an advanced knowledge and expertise in AMOS)

After all discussions, some essential fit indices to take into account are CMIN and CMIN/df, F0, RMSEA, and PCLOSE. Optionally, NFI, TLI, and CFI can be used. Researchers must determine a rationale for fit criteria, mention those rationale in their papers, and, perhaps, regard reporting several different types of fit indices. There is no one set of rules which to use, but a researcher can take into account the size of the sample, number of variables, and fit indices' pros and cons. Finally, at least referring to one index from every different group of indices that we mentioned earlier in this text may reduce the criticism for the fit of the model.

Genişletilmiş Özet

AMOS, Lisrel, R, Matlab ve birçok benzer istatistiksel yazılım programlarının yardımıyla, karmaşık araştırma modellerinin çoğu daha hesaplanabilir ve kolayca (?) anlaşılabilir hale gelmiştir. Hatta birçok farklı ilişkilere sahip, karmaşık modeller bile yazılımlar yardımıyla kolayca hesaplanabilmektedir. Aralarında küçük farklılıklar olmasına rağmen, çıktılar genellikle tutarlıdır ve oluşan tablolar çoğunlukla anlaşılabilirdir. Bununla beraber, araştırma metodolojisi hakkında artan merak ve bilgi miktarı dolayısıyla, bu basit görünümlü çıktılar daha da karmaşıklaşmaya ve derinleşmeye başlamıştır. Sözü edilen ifadeler çelişkili gözükse de bu noktada ima edilen durum orta seviye bir araştırmacı için oldukça tanıdık gelecektir, çünkü bir araştırmacının istatistik bilgisi ve anlayışı derinleştikçe, bu rakamların nereden, nasıl ve neden hesaplandığına dair sorular ve şüpheler artmaktadır. Bu sorular ve şüpheler uyum indeksleri, ki-kare ve serbestlik dereceleri (Degrees of Freedom), değişiklik indeksleri (Modification Indices), kovaryanslar ve artıklar (residuals) hakkında merak uyandırmaya başlamaktadır. Bu doğrultuda, istatistiksel jargondan mümkün olduğunca kaçınarak CMIN (ya da ki-kare), farklı model tanımları ve bu modellerin yardımıyla uyum indeksi hesaplamalarına odaklanılmaktadır. Tüm bunlarla birlikte bu çalışmada, on yıllık bir tartışmaya da son vermek amacıyla; hangi uyum indekslerinin ne zaman ve nasıl kullanılacağı, tam olarak neyi belirttikleri ve hangi değerlerin iyi veya kötü uyum anlamına geldiği tartışılmaktadır.

Bu çalışmada tartışılmakta olan, yapısal eşitlik modellemesi, doğrulayıcı faktör analizi ve uyum kavramlarını kavrayabilmek literatürde olan bazı tanımların netleştirilmesi gerekmektedir. Öncelikle çoğu zaman kafa karıştırıcı olabilen, araştırma sürecinin birçok noktasında yeterince açıklanmayan ve istatistik işlemlerin yapılması için kullanılan yazılımlarda bulunan modellerin tanımlanması ve daha sonra da tahmin yöntemlerinin açıklanması yerinde olacaktır. Çoğu "yapısal eşitlik modellemesi (SEM)" yöntemi kullanan araştırmacı hangi yazılımı kullanırsa kullansın "Maximum Likelihood" yöntemine alışır ancak büyük bir kısmının bu yöntemin gerçekte ne olduğu ve ne yaptığı hakkında hiçbir fikri yoktur. Ayrıca modellerin uyumunu tartışmadan önce açıklığa kavuşturulması gereken birkaç kavram vardır. Bunlar; CMIN, Ki-kare, Log-Likelihood (Maximum Likelihood), C ve F değerleri, NPAR, p ve P, uyum ve indeks, PRATIO. Bu kavramların çoğu tanıdık gelmelerine rağmen, çoğu zaman yanıltıcı, kafa karıştırıcı, şaşırtıcı ve çelişkili olabilmektedir. Genellikle bu kavramlar, çeşitli araştırmalarda görülmekte ve bilindiği düşünülmektedir ancak gerçekte ne olduklarını ve nereden geldikleri üzerinde düşünülmemektedir. Bu sebeple bu çalışmada birçok kavram ve terminolojinin tartışılmasının yanı sıra, gerekli değerler ve önemli noktalar ele alınmıştır.

Ayrıca Türkçe genişletilmiş özette yer verilemeyen ancak makalede İngilizce olarak ayrıntılandırılmış konular:

- Yuvalanmış modeller (nested models), araştırma modeli, doymuş (saturated) model, bağımsızlık (independence) modeli gibi kavramlar ve bu modellerin uygunluk değerlerini hesaplarken nasıl kullanıldığı.

- Maximum Likelihood yönteminin faktör analizinde ve uygunluk değerlerini hesaplamada niçin önemli olduğu ve tam olarak ne yaptığı.

- Ki-Kare kavramının ayrıntılı olarak incelenmesi ve neden uygunluk değerlerinin en önemlisi olduğunun tartışılması.

- Tüm uygunluk istatistiklerinin ayrıntılı açıklaması, benzerlik ve farkları, güçlü ve zayıf yönleri.

- Düzeltme indislerinin (Modification Indices) ne olduğu ne şekilde kullanması gerektiği.

Bazı araştırmacılar ve istatistikçiler, uyum indeksleri için farklı değerler ve sınırlılıklar belirlemektedir. Bu nedenle araştırmacılar uyum indeksleri için şu noktaları akılda tutmalıdır:

- Normallik mutlak uyum indekslerini etkilemektedir. Normal olmayan veriler ki-kareyi arttırır ve böylece mutlak uyum değerlerini azaltır (Kenny, D.A., 2015).

- Değişken sayısı uyum indekslerini etkilemektedir. Değişkenlerin artması uyumu azaltır. Yeni değişkenlerin eklenmesi yoluyla, istenmeyen bir durum olan RMSEA'nın yükselmesi de mümkündür. NFI, TLI ve CFI gibi indeksler nispeten daha kararlı bir yapıda olsa da değişken sayısına göre ufak azalmalar gösterebilir. Bu durumun sebebinin de ki-karenin artması olduğu tahmin edilmektedir.

- BIC, RMSEA ve TLI modelde sıkılığın (parsimony) karşılığını verirken, NFI ve CFI bunu en az ödüllendiren indekslerdir.

- NFI örneklem büyüklüğüne göre kendini ayarlamaz, artan örneklem büyüklüğü uyum değerini azaltır.

- TLI ve CFI değerleri örneklem büyüklüğü ile nispeten daha kararlı bir ilişki içerisindedir ve örneklem büyüklüğü arttıkça değişkenlik azalır. RMSEA da örneklem büyüklüğü ile düşüş göstermekte, büyük örneklem büyüklükleri RMSEA'nın lehine bir durum ortaya koymaktadır.

- Kesin bir uyumluluk için, araştırmacı CMIN değerinin anlamsız olmasını beklemelidir. Ancak bu durum neredeyse her zaman anlamsızdır. (Çok az değişken ve çok küçük bir örneklem büyüklüğü olmadığı sürece)

- İyi ya da tam uygunluğun olup olmadığını değerlendirmek için araştırmacılar farklı değerler kabul edebilmektedir;

    - RMSEA (0,05 ve 0,08'in altında): Model sıkı (parsimonious) ise ve örneklem sayısı fazla ise 0,05'in altında ya da ona yakın değerler olması, aksi takdirde ise 0,08'in altında olması beklenir.

    - CFI, RNI, NFI, TLI, RFI, IFI (0,90 ve 0,95'in üstünde): Değişken büyüklüğüne bağlı olarak değişmektedir. 10-12 değişkenden az olan durumlar 0,95 ya da ona yakın bir uyum gerektirmekte iken 12'den fazla değişkeni olan durumlar ise için sınır nokta 0,90'dır. Bu değer için daha yüksek olması her zaman daha iyidir.

    - Daha büyük örneklemler için RMR'nin 0,05 ya da 0,08'den küçük ve tercihen GFI'nin 0,90 ya da üstünde olması beklenir. Ancak bu indekslerin tercihen kullanılmaması öngörülmektedir.

    - Modelleri karşılaştırmak için AIC, BIC gibi kriterler daha yararlıdır.

    - Aşamalı karşılaştırmalar ve model arındırma için Modifikasyon İndeksleri (Modification Indices) çok faydalıdır.

    - AMOS'ta iç içe geçmiş modellere isim atamak ve onların olasılık oranlarını hesaplamak model karşılaştırmaları için en iyi yöntemdir. (AMOS konusunda ileri düzeyde bilgi ve uzmanlık gerektirir.)

Tüm bu tartışmalardan sonra dikkate alınması gereken bazı temel uyum indeksleri; CMIN ve CMIN/df, F_0, RMSEA ve PSCLOSE'dir. İsteğe bağlı olarak NFI, TLI ve CFI de kullanılabilir. Bunlar doğrultusunda araştırmacıların uyum kriterleri için mantıklı gerekçeler belirlemeleri, bu gerekçeleri makalelerinde belirtmeleri ve birkaç farklı uyum indeksi ile karşılaştırmalar yapmaları gerekmektedir. Bu noktada kullanılması gereken kurallar bütünü bulunmamaktadır ancak araştırmacı örneklem büyüklüğünü, değişken sayısını, uygun endekslerin artılarını ve eksilerini dikkate alarak karar vermelidir. Son olarak, bu çalışmada bahsedilen her farklı indeks grubundan bir indekse atıfta bulunmak, modelin uyumuna yönelik eleştirileri azaltacaktır.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The authors has no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

References

Akaike, H. (1987). Factor analysis and AIC. In Selected papers of Hirotugu Akaike (pp. 371-386). Springer, New York, NY.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Testing structural equation models, 154,* 136.

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife research, 28*(2), 111–119.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin, 105*(3), 456.

Cochran, W. G. (1952). The $\chi 2$ test of goodness of fit. *The Annals of Mathematical Statistics,* 315–345.

Gulliksen, H., & Tukey, J. W. (1958). Reliability for the law of comparative judgment. *Psychometrika, 23*(2), 95–110.

James, L. R., Mulaik, S. A., & Brett, J. (1982). C*ausalanalysis: Models, assumptions and data.* Beverly Hills, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 200.

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

Mulaik, S.A., James, L.R., Van Alstine, J., Bennet, N., Lind, S., and Stilwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430–445.

Newsom, J. T. (2018). Minimum sample size recommendations (Psy 523/623 structural equation modeling, Spring 2018). Manuscript Retrieved from upa.pdx.edu/IOA/newsom/semrefs.htm.

Psutka, J. V., & Psutka, J. (2015, September). Sample size for maximum likelihood estimates of Gaussian model. In International Conference on Computer Analysis of Images and Patterns (pp. 462-469). Springer, Cham.

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*(2), 23–74.

Schwab, A., & Starbuck, W. H. (2013). Why Baseline Modelling is Better than Null-Hypothesis Testing: Examples from International Business Research. Philosophy of Science and Meta-Knowledge in International Business and Management, 171.

Steiger, J. H., & Lind, J. (1980). Paper presented at the annual meeting of the Psychometric Society. Statistically-based tests for the number of common factors.

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

Templin, J. (2015). Maximum Likelihood Estimation; Robust Maximum Likelihood; Missing Data with Maximum Likelihood [PowerPoint slides]. Retrieved from https://jonathantemplin.com/files/sem/sem15pre906/sem15pre906_lecture03.pdf on 9-10-2019.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological methodology, 8*, 84–136.