Equivalence Testing and Its Applications
Fall 2018 Lecture

Ke-Hai Yuan

(ref: Yuan, Chan, Marcoulides & Bentler 2016. Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330.
Marcoulides & Yuan 2017. New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling, 24*(1), 148–153.)

Introduction and motivation

- Conventional null hypothesis testing (NHT) is a very important tool if the ultimate goal is to find a difference or to reject a model or hypothesis
- NHT is clumsy if the purpose is to use a model or hypothesis for further data analysis
- Let $H_0 : P = M(\theta)$ represent that the population (P) can be fitted by a model $M(\theta)$, and T be the test statistic for H_0 (e.g., the likelihood ratio statistic), and $(T|H_0) \sim \chi^2$
    - If T is statistically significant, then we reject H_0 so that the model cannot be used
    - If T is not statistically significant, we do not know whether the model is properly formulated or how bad the model is
- Example: When using a t-statistic to compare the means of two groups, we need to assume equal variance in order for the t-statistic to follow a Student t-distribution. However, there is no effective way of confirming $H_\sigma : \sigma_1^2 = \sigma_2^2$. A significant F-statistic implies H_σ does not hold. A non-significant F-statistic does not imply that H_σ holds.
- The conventional confidence interval (CI) has the same problem. For example, if a CI for $\delta = \mu_2 - \mu_1$ is given by [0, 1.3], it only means that you cannot reject $H_0 : \mu_1 = \mu_2$, but it does not mean you can claim $\delta = 0$. Since another person can claim $\delta = 1.3$.
- A lot of problems in statistics are to find a good model and then use it to account for the relationships among the observed variables (regression, time series, generalized linear model, growth curve model, factor model, item response model, etc.)
- Only a small proportion of research interest is to reject a hypothesis

- For researchers who need to use a model (rather than to reject a model), a more proper methodology is equivalent testing.
- In NHT, the setup is to reject $H_0 : \mu_2 = \mu_1$
- In equivalence testing, the setup is to reject $H_{0a} : |\mu_2 - \mu_1| > \epsilon_0$ or a bad model
- A key component of equivalence testing is the selection of ϵ_0, which is a tolerable boundary for departure from the target or model misspecification
- The method of equivalence testing has been used in:
    - establishing the equivalence of different treatment programs (Dunnett & Gent, 1977; Rogers, Howard & Vessey, 1993);
    - equivalence of confidence intervals for means (Seaman & Serlin, 1998; Tryon, 2001; Tryon & Lewis, 2008);
    - bioequivalence or equivalence of different drugs in biostatistics (Barker et al., 2002; Ocaña et al., 2008);
    - testing for lack of associations among variables (Goertzen & Cribbie, 2010);
    - Wellek (2010) gave a systematic description of the method and illustrated its applications

  We will describe its application in structural equation modeling, which has many application in psychology, education, organization research, health and policy

Outlines of the development

- Test of *not-close fit* in power analysis (MacCallum, Browne & Sugawara et al, 1996); Multiple-group analysis (Yuan & Chan, 2016)
- Our study includes contrasting type I error and power between equivalence testing and NHT
- We discuss how to set tolerable size of misspecification in SEM
- We also define a concept of minimum tolerable size (T-size) of model misspecification, which is parallel to p-value in NHT
- Connecting T-size with existing measures of model misspecification in SEM
- The single most notable property of equivalence testing is that it allows a researcher to confidently claim that the size of misspecification in the current model is below the T-size
- We also have R code for conducting equivalence testing in SEM

An example of SEM/factor analysis

- Holzinger and Swineford (1939) reported a data set with cognitive test scores on 26 items and 145 students from the Grant-White school
- Nine of the 26 variables have been used in illustrating various new developments in SEM since Jöreskog (1969).
- We will also use the 9-variables: visual perception, cubes, lozenges; paragraph comprehension, sentence completion, word meaning; addition, counting dots, and straight-curved capitals.

Figure 1. A confirmatory factor model, data from Holzinger & Swineford (1939)

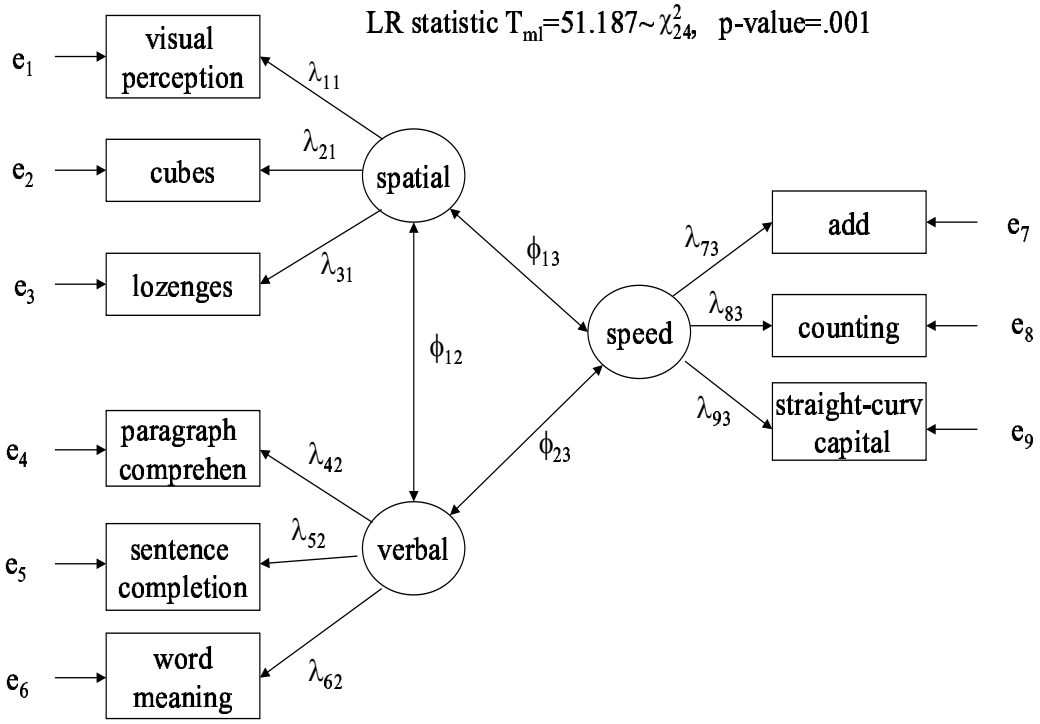
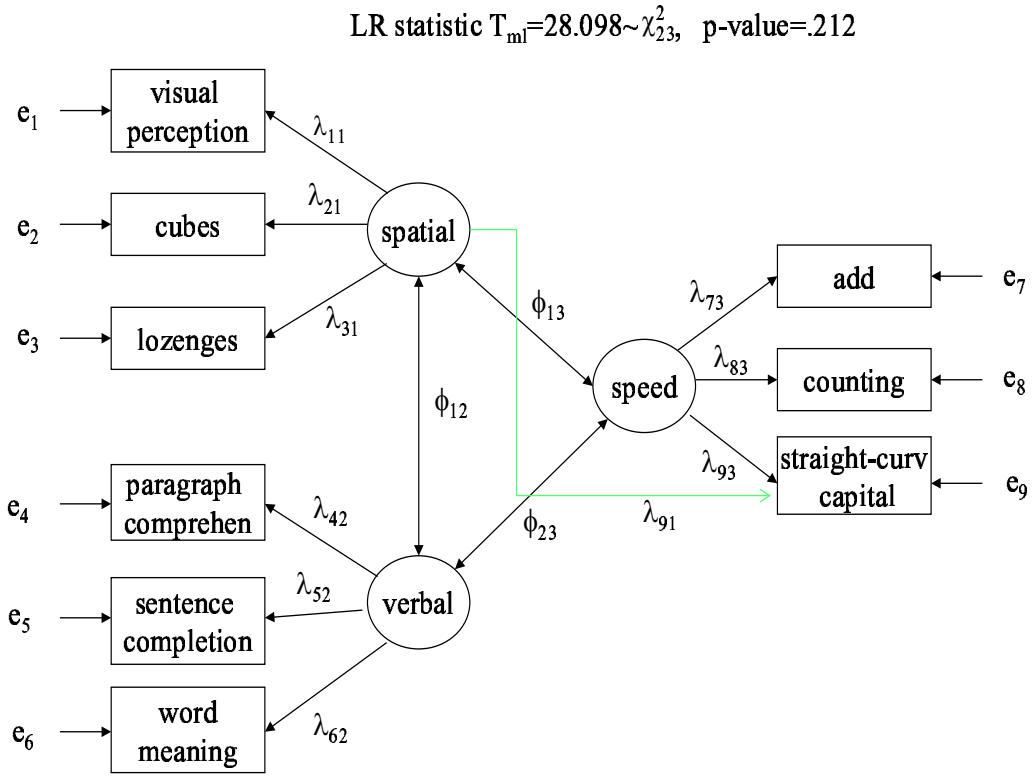


Figure 2. Another confirmatory factor model, same data



- By separating measurement errors from latent true variables, SEM allows us to obtain unbiased estimates of the correlations among the theoretical latent variables
- Similarly, measurement errors lead to biased estimates of regression coefficients, and SEM removes the bias by separating measurement errors for latent variables

Likelihood ratio statistic

- For the first confirmatory factor model, $T_{ml} = 51.187$, with p-value=.001 when referred to χ^2_{24}
- For the second model, $T_{ml} = 28.098$, corresponding to a p-value=.212 when referred to χ^2_{23}.
- Can we claim the the 2nd model is the correct model?
- Can we make a claim about the quality (size of misspecification) of the 2nd model?
- The likelihood ratio statistic $T_{ml} = nF_{ml}$, where

$$F_{ml}[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})] = \text{tr}[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})| - p,$$

  with $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ being the covariance matrix derived from the model, and p being the number of variables
- Under standard regularity condition, $T_{ml} \sim \chi^2_{df}(\delta)$ with

$$\delta = n \min_{\theta} F_{ml}[\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$$

- In the literature, $\epsilon_0 = \min_\theta F_{ml}[\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$ is used to measure the discrepancy between data and model
- More parameters tend to make ϵ_0 smaller, but the model also becomes less interesting
- Steiger and Lind (1980) proposed root mean square error of approximation (RMSEA)

$$\text{RMSEA}_0 = (\epsilon_0/df)^{1/2} \text{ and } \text{RMSEA} = \{\max(T_m l - df, 0)/[n \times df]\}^{1/2}$$

- In the literature of SEM, it is generally agreed that every model is wrong, but some are useful, and established nominal labeling for RMSEA is

RMSEA	$\leq .01$	(.01, .05]	(.05, .08]	(.08, .10]	$> .10$
label of fit	excellent	close	fair	mediocre	poor fit

- People also propose to use confidence interval for the population RMSEA_0 $[c_{lower}, c_{upper}]$, which can be obtained from the CI for noncentrality parameter with $T_{ml} \sim \chi^2_{df}(\delta)$ (Venables, 1975).

Back to the example

- For the 2nd confirmatory factor model, $T_{ml} = 28.098$, corresponding to a p-value=.212 when referred to χ^2_{23}; RMSEA = .039, and the 95% CI for the population RMSEA is $[0, .090]$
- Most researchers would claim that the model is perfect, since CI contains 0.
- However, because the CI contains .90, a reviewer can also claim that the model is simply mediocre
- The conflict is simply because the conventional NHT is a clumsy tool for model construction
- We need to use equivalence testing for structural equation modeling (SEM)

Equivalence testing for parameters

- The null hypothesis

$$H_{0a} : \textit{the difference between } \boldsymbol{\theta} \textit{ and } \boldsymbol{\theta}_0 \textit{ is greater than } \epsilon_0, \tag{1}$$

  where the difference can be the Euclidean distance, the standardized Mahalanobis distance, or the sum of absolute differences between the coordinates of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$

- The value ϵ_0 is a small positive number up to our tolerance on the size of difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ or any meaningful standard

- Each meaningful measure of difference will correspond to a statistic by which the test is performed, and the statistic stochastically increases with the difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$

- The hypothesis H_{0a} is rejected when the statistic is **smaller than** a critical value determined by the distribution of the statistic and the significance level (e.g., $\alpha = .05$).

- Rejection of H_{0a} implies that the difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ is within a tolerable size.

- If $\epsilon_0 = 0$, then, except the possibility of type I error, rejection of H_{0a} implies that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

- In particular, we are $(1 - \alpha)$ confident that the unknown parameter $\boldsymbol{\theta}$ is close enough to the target $\boldsymbol{\theta}_0$.

- In contrast, rejecting H_0 in conventional null hypothesis testing implies that $\boldsymbol{\theta}$ does not equal $\boldsymbol{\theta}_0$, while not rejecting H_0 does not allow us to endorse $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ either.

Equivalence testing for overall model fit

- In SEM or many other disciplines of statistics, overall model evaluation is more fundamental than parameter evaluation
- Conventional null hypothesis is

$$H_0 : F_{ml0} = 0 \text{ or equivalently } H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0),$$

  and T_{ml} is compared against χ^2_{df} for inference.
- The hypothesis in equivalence testing is

$$H_{0a} : F_{ml0} > \epsilon_0, \tag{3}$$

  where ϵ_0 is a small positive number that one can tolerate for the size of misspecification.
- Via the relationship of T_{ml} and F_{ml}, the hypothesis in (3) can be expressed in terms of the population noncentrality parameter, $H_{0a} : \delta > \delta_0$, where $\delta_0 = (N-1)\epsilon_0$.
- Let $c_\alpha(\epsilon_0)$ be the *left-tail critical value* of $\chi^2_{df}(\delta_0)$ corresponding to cumulative probability α.
- We reject the hypothesis H_{0a} in (3) if $T_{ml} \leq c_\alpha(\epsilon_0)$ and, under the assumptions $T_{ml} \sim \chi^2_{df}(\delta)$ with $\delta = (N-1)F_{ml0}$, type I error is controlled at level α.
- When H_{0a} is rejected, we can conclude that the size of misspecification in the current model, as measured by F_{ml0}, is smaller than or equal to ϵ_0.

Equivalence testing for overall model fit

- We can also specify the tolerable size of misspecification through $\epsilon_0 = df(\text{RMSEA}_0)^2$ or $\delta_0 = n \times df(\text{RMSEA}_0)^2$, and use the conventional cutoff values of RMSEA (.01, .05, .08, and .10) to distinguish between excellent, close, fair, mediocre, and poor models, respectively.
- However, as we shall see, such conventional cutoff values make equivalence testing much more stringent on the size of misspecification than they are used with the conventional point estimate of RMSEA (Steiger & Lind, 1980).

Equivalence testing vs NHT

- Equivalence testing is a special case of general statistical hypothesis testing. Thus, it also suffers from type I and type II errors.
- Both equivalence testing and NHT use the same test statistic T_{ml}. However, the rejection regions are different.
  - Equivalence testing aims to reject a model that is not qualified for further consideration, and it happens when T_{ml} falls in the interval $[0, c_\alpha(\epsilon_0)]$.
  - NHT is to reject a correct model that we ideally would like to have, and it happens when T_{ml} falls in the interval $(c_{1-\alpha}, \infty)$, where $c_{1-\alpha}$ is the right-tail critical value of χ^2_{df}
  - A better (less misspecified) model corresponds to more power in equivalence testing whereas a worse model corresponds to more power in NHT.
- The implication of type I error in equivalence testing is different from that in NHT.
  - With equivalence testing, type I error occurs when claiming a not-acceptable model ($F_{ml0} > \epsilon_0$) as acceptable, and the probability of committing such errors is controlled at level α.
  - With NHT, type I error occurs when claiming a correctly specified model ($F_{ml0} = 0$) as a misspecified one ($F_{ml0} > 0$), and minimizing such errors might seem a right thing to do. However, the error of treating a misspecified model as a correctly specified one is left to chance.
- When $c_\alpha(\epsilon_0)$ is smaller than $c_{1-\alpha}$ and the observed statistic T_{ml} falls between these two numbers, we are unable to reject the null hypothesis in NHT and cannot claim the hypothesis or model as acceptable either in equivalence testing. Such a scenario indicates that, while not rejecting the conventional null hypothesis, we cannot even prove that the size of model misspecification is below a tolerable threshold.

- When $c_\alpha(\epsilon_0)$ is greater than $c_{1-\alpha}$ and T_{ml} falls between these two numbers, we reject the conventional null hypothesis with NHT but at the same time, according to ϵ_0, we can tolerate the degree of approximation/misspecification in the model.
- Because $c_\alpha(\epsilon_0)$ increases with ϵ_0, there exists an ϵ_0 such that $c_\alpha(\epsilon_0) = c_{1-\alpha}$. Then, equivalence testing may seem to yield an identical conclusion with NHT. But the results or implications are different.
    - For equivalence testing at $c_\alpha(\epsilon_0) = c_{1-\alpha}$, we have agreed that a misspecification at size $\varepsilon_0 = F_{ml0}$ in the model is acceptable
    - Under NHT we have no information on the degree of misspecification in the current model.

Minimum tolerable size (T-size) of misspecification

- Comparing an observed statistic to the critical value at a given level of significance is a key step of conventional statistical inference
- A more informative element in conventional statistical inference is the p-value, which is the area above T_{ml} under the probability-density curve of χ^2_{df}.
- The concept of p-value in equivalence testing is defined similarly, it is the area below the observed T_{ml} under the density curve of the noncentral chi-square distribution $\chi^2_{df}(\delta_0)$.
- Like the p-value in NHT, with all other factors given, the p-value in equivalence testing becomes smaller (more significant) as sample size increases.
- In addition, the p-value in equivalence testing also depends on ϵ_0, the tolerable size of misspecification.
- An even more informative element in equivalence testing is the tolerable size of misspecification according to which the current model is deemed as acceptable for explaining the relationship among the observed variables.
- Clearly, if ϵ_0 is tolerable, then any value below ϵ_0 is also tolerable. However, a smaller ϵ_0 may render the current model as not acceptable.

- We define the *minimum tolerable size* (T-size) of misspecification corresponding to the observed T_{ml} as the ϵ_t that satisfies

$$T_{ml} = c_\alpha(\epsilon_t), \tag{4}$$

  where $\epsilon_t = 0$ if $T_{ml} \leq c_\alpha$.

- With the ϵ_t in (4), for any tolerable size ϵ_0 that is greater than ϵ_t we can reject the H_{0a} in (3) and the probability of committing an error is less than α.

- If ϵ_0 is less than ϵ_t, then we will have to tolerate a larger type I error or to reformulate the model for research to proceed.

- The T-size in equivalence testing plays essentially the same rule as that of p-value in NHT.

- We might relate the T-size ϵ_t to RMSEA or other fit indices, and also call the resulting fit indices T-size.

Continuation of the example

- For the 2nd confirmatory factor model, $T_{ml} = 28.098$, corresponding to a p-value=.212 when referred to χ^2_{23}; RMSEA = .039, and the 95% CI for the population RMSEA is $[0, .090]$
- With equivalence testing, letting $\alpha = .05$ and solving equation (4) with $T_{ml} = 28.098$ yields T-size $\epsilon_t = .158$ and $\text{RMSEA}_t = .083$, corresponding to ncp $\delta_t = (N-1)\epsilon_t = 22.705$.
- We are 95% confident that the size of misspecification is no more than .158 as measured by F_{ml0}, or equivalently no more than .083 as measured by RMSEA, or no more than 22.705 as measured by ncp.
- If we can tolerate a misspecification of $\text{RMSEA}_t = .083$, then we can proceed with the 2nd factor model with 95% confidence.

Continuation of the example

- The T-size ϵ_t is simply the upper limit of the CI for $F_{ml0} = \delta/(N-1)$ with confidence level $(1-2\alpha)$ (see e.g., Venables, 1975).
- It follows from the correspondence between confidence interval and null hypothesis testing, the T-size $\text{RMSEA}_t = .083$ is simply the upper limit of the .90 confidence interval for RMSEA.
- We have an R program that calculates the T-size RMSEA based on the solution of ϵ_t in equation (4) for any $0 < \alpha < 1$.
- The R program also calculates the T-size comparative fit index (CFI) to be introduced in the next section. The code of the program can be downloaded at
  `http://www3.nd.edu/~kyuan/EquivalenceTesting/T-size_RMSEA_CFI.R`,
  and we briefly introduce the application of the program in an appendix.

Equivalence Testing with comparative fit index (CFI)

- Bentler (1990) defined CFI, which involves comparing a substantively interesting model with a base model, and the goodness of the model should be judged relatively.
- The most widely used base model is the independence model (all observed variables are independent)
- Let F_{mli0} be the discrepancy between the population and the independence model, and F_{ml0} be the discrepancy for the interesting model
- The population CFI is defined as

$$\text{CFI} = 1 - \frac{\delta}{\delta_i} = 1 - \frac{F_{ml0}}{F_{mli0}},$$

  and the agreed upon value for good model is CFI $\geq .95$
- Thus, in equivalence testing, the default hypothesis is

$$H_{0a} : \text{CFI} < \text{CFI}_0. \tag{5}$$

- Instead of working out a procedure of testing the hypothesis in (5), we will develop an estimate for T-size CFI_t that satisfies

$$P(\text{CFI}_t \leq \text{CFI}) \geq 1 - \alpha, \tag{6}$$

- With a CFI_t that satisfies (6), we will reject the H_{0a} in (5) for any value of CFI_0 that is smaller than CFI_t, and type I error is controlled at level α.
- Alternatively, we should not accept the current model if we cannot tolerate a misspecification with size CFI_t.

Equivalence Testing with CFI

- We will use the idea of Bonferroni correction to obtain a CFI_t that satisfies the probability specified in (6).
- Let ϵ_t and ϵ_{it} be defined by

$$T_{ml} = c_{\alpha/2}(\epsilon_t) \text{ and } T_{mli} = c_{1-\alpha/2}(\epsilon_{it}), \tag{7}$$

- It follows from the definitions of ϵ_t and ϵ_{it} that (see e.g., Venables, 1975)

$$P(F_{ml0} \leq \epsilon_t) = 1 - \alpha/2 \text{ and } P(F_{mli0} \geq \epsilon_{it}) = 1 - \alpha/2. \tag{8}$$

- Thus,

$$P(\{F_{ml0} \leq \epsilon_t\} \cap \{F_{mli0} \geq \epsilon_{it}\}) \geq 1 - \alpha. \tag{9}$$

- When $\epsilon_t \leq \epsilon_{it}$ and ϵ_{it} is positive, define

$$\text{CFI}_t = 1 - \frac{\epsilon_t}{\epsilon_{it}}, \tag{10}$$

  then the CFI_t in (10) can serve as the T-size CFI.
- Although both ϵ_t and ϵ_{it} are non-negative, it is possible that $\epsilon_{it} < \epsilon_t$. For the purpose of $\text{CFI}_t \in [0, 1]$, a modified definition of CFI_t is

$$\text{CFI}_t = 1 - \frac{\epsilon_t}{\max(\epsilon_{it}, \epsilon_t)}.$$

- In contrast, the conventional sample CFI is defined as

$$\text{CFI}_c = 1 - \frac{\max(T_{ml} - df, 0)}{\max(T_{mli} - df_i, 0)},$$

  which is simply a descriptive statistic

- For the 2nd confirmatory factor model, $T_{ml} = 28.098$, corresponding to a p-value=.212 when referred to χ^2_{23}; RMSEA = .039, and the 95% CI for the population RMSEA is $[0, .090]$
- With equivalence testing, letting $\alpha = .05$ and solving equation (4) with $T_{ml} = 28.098$ yields T-size $\epsilon_t = .158$ and $\text{RMSEA}_t = .083$
- The sample CFI is given by $\text{CFI}_c = .989$, and at $\alpha = .05$, the T-size CFI is $\text{CFI}_t = .931$.
- We are 95% confident that the population CFI is above .931.
- If using the cutoff values established for CFI_c (e.g., Hu & Bentler, 1999) to judge the size of CFI_t, then the 2nd factor model may not be deemed as achieving an acceptable level of fit
- We need to have new rules for evaluating the goodness of models according to T-size

Adjusted Cutoff Values of RMSEA with Equivalence Testing

- We denote the conventional cutoff values of RMSEA as RMSEA_c and those in equivalence testing as RMSEA_e, which are intended norms when judging the size of RMSEA_t
- Of course, any norms or cutoff values cannot avoid arbitrariness. They simply facilitate researchers to communicate their findings.
- Since .01, .05, .08 and .10. are widely accepted cutoff values for RMSEA_c to distinguish between excellent, close, fair, mediocre, or poor fit, we will examine the values of RMSEA_t at each of these values of RMSEA_c to obtain the corresponding RMSEA_e.
- Notice that at the sample level, the RMSEA_t corresponding to equation (4) is determined by T_{ml}, which is further determined by RMSEA_c according to

$$T_{ml} = (N-1)df(\text{RMSEA}_c)^2 + df. \tag{11}$$

  In our study, at each value of $\text{RMSEA}_c = .01, .05, .08$ and .10 we generated T_{ml} according to equation (11) for $df = 1, 2, \ldots, 100$; and for 24 conditions on sample size N: 30, 40, 50, 60, 80, 100, 120, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000. So, for each value of RMSEA_c, there are 2400 values of RMSEA_t, and we use these 2400 values as the dependent variable to explore the functional relationship between RMSEA_t and RMSEA_c. The RMSEA_e is simply the predicted value of RMSEA_t by this functional relationship.
- Since $\text{RMSEA} = [\delta/(n \times df)]^{1/2}$, we choose to linearly predict $\ln(\text{RMSEA}_t)$ by at each value of $\text{RMSEA}_c = .01, .05, .08$ and .10, with 11 predictors $\ln(df)$, $[\ln(df)]^2$, $df^{1/5}$, $df^{1/2}$, df, $\ln(n)$, $[\ln(n)]^2$, $n^{1/5}$, $n^{1/2}$, n, and $\ln(df)\ln(n)$.
- When all the 11 predictors are included, the R-squares of linearly predicting $\ln(\text{RMSEA}_t)$ are $R^2 = .9997, .9996, .9978$, and .9958 at $\text{RMSEA}_c = .01, .05, .08$, and .10, respectively.
- Because the 11 predictors are correlated and they may not all be needed, we next used the best subset regression to select the most relevant predictors.

Table 1. The estimated regression coefficients for best subset of predictors of $\ln(\text{RMSEA}_t)$, with $\text{RMSEA}_e = \exp(y_e)$.

predictors	y_{e01}	y_{e05}	y_{e08}	y_{e10}
intercept	1.34863	2.06034	2.84129	2.36352
$\ln(df)$	-.51999	-.62974	-.54809	-.49440
$[\ln(df)]^2$	.01925	.02512	.02296	.02131
$\ln(n)$	-.59811	-.98388	-.76005	-.64445
$[\ln(n)]^2$		.05442	.10229	.09043
$n^{1/5}$			-1.11167	-1.01634
$n^{1/2}$	.00902			
n		-.00005188		
$\ln(df)\ln(n)$	.01796	.05260	.04845	.04422
R^2	.9997	.9996	.9977	.9955

Adjusted Cutoff Values of RMSEA with Equivalence Testing

- Let y_e be the predicted value by the linear combination of predictors according to Table 1, we can obtain the value of $\text{RMSEA}_e = \exp(y_e)$ corresponding to $\text{RMSEA}_c = .01$, .05, .08, and .10, respectively.
- We will refer to these values as adjusted cutoff values.
- To facilitate applications, the formulas for evaluating the adjusted cutoff values are implemented in R code, which can be downloaded at `http://www3.nd.edu/~kyuan/EquivalenceTesting/RMSEA_e.R`, where the needed inputs are the degrees of freedom df and sample size N.
- Continuation of the example: For the 2nd factor model, we have $\text{RMSEA}_c = .039$, and $\text{RMSEA}_t = .083$ at $\alpha = .05$. At $n = 144$ and $df = 23$, the values of $\text{RMSEA}_e = \exp(y_e)$ are .069, .091, .116, and .135, respectively. Thus, according to the adjusted cutoff values, $\text{RMSEA}_t = .083$ also indicates that the modified model achieves close fit.

Adjusted Cutoff Values of CFI with Equivalence Testing

- Cutoff values for CFI are not as finely defined as RMSEA, and most researchers only use .95 as the cutoff values
- We propose to use $\text{CFI}_c = 1 - \text{RMSEA}_c = .99$, .95, .92 and .90, and call the corresponding models as achieving excellent, close, fair, mediocre, and poor fit.
- With similar design, we obtained the cutoff values for the T-size CFI corresponding to $\text{CFI}_c = .99$, .95, .92 and .90.

Table 2. The estimated regression coefficients for best subset of predictors of $\ln(1 - \mathrm{CFI}_t)$, with $\mathrm{CFI}_e = 1 - \exp(y_e)$.

predictors	y_{e99}	y_{e95}	y_{e92}	y_{e90}
intercept	4.67603	4.12132	6.31234	5.96633
$\ln(df)$	-.50827	-.46285	-.41762	-.40425
$[\ln(df)]^2$			.01554	.01384
$df^{1/5}$	.87087	.52478		
$[\ln(df_i)]^2$			-.00563	-.00411
$df_i^{1/5}$	-.59613	-.31832		
$\ln(n)$	-1.89602	-1.74422	-1.30229	-1.20242
$[\ln(n)]^2$	.10190	.13042	.19999	.18763
$n^{1/5}$			-2.17429	-2.06704
$n^{1/2}$		-.02360		
$\ln(df)\ln(n)$	.03729	.04215	.05342	.05245
$\ln(df_i)\ln(n)$			-.01520	-.01533
R^2	.9836	.9748	.9724	.9713

Adjusted Cutoff Values of CFI with Equivalence Testing

- Let y_e be the predicted value by the linear combination of predictors according to Table 2, we can obtain the value of $\text{CFI}_e = 1 - \exp(y_e)$ corresponding to $\text{CFI}_c = .99$, .95, .92, and .90, respectively.
- We will refer to these values as adjusted cutoff values, and they allow us to nominally judge the tolerance level of CFI_t using established norms for judging the value of CFI_c.
- To facilitate applications, the formulas for evaluating the adjusted cutoff values are implemented in R code, which can be downloaded at `http://www3.nd.edu/~kyuan/EquivalenceTesting/CFI_e.R`, where the needed inputs are the degrees of freedom df, sample size N and the number of variables.
- Continuation of the example: For the 2nd factor model, we have $\text{CFI}_c = .989$ and $\text{CFI}_t = .931$ at $\alpha = .05$.
- At $N = 144$ and $df = 23$, the adjusted cutoff values of CFI_e are .941, .874, .828, and .798, respectively.
- Thus, $\text{CFI}_t = .931$ also indicates that the modified model achieves close fit.
- It is important to emphasize once again that the correspondence between CFI_c and CFI_e or between RMSEA_c and RMSEA_e is simply to facilitate the communication of the goodness of the model as measured by T-size (RMSEA_t or CFI_t).
- Even when the model is excellent, the tolerable size of misspecification is much greater in the case of RMSEA and smaller in the case of CFI

Conclusion

- When researchers choose SEM for data analysis, they aim to use the model to account for the relationship among the observed variables, and to further elaborate on values of the parameter estimates.
- However, conventional null hypothesis test is developed to reject the model under the null hypothesis rather than accept it.
- Equivalence testing allows a researcher to accept a model for data analysis
- The most important feature of equivalence testing is that it gives us the desired confidence for the current model with a misspecification being smaller or greater than the observed T-size (RMSEA_t or CFI_t).
- In summary, equivalence testing gives SEM the needed property to be a scientific methodology, and we thus propose that conventional null hypothesis testing be replaced by equivalence testing and recommend that researchers start routinely reporting the T-size in order to convey the goodness of the model.
- We have focused mainly on using equivalence testing to endorse SEM models, equivalence testing can replace conventional null hypothesis testing when evaluating all types of models that are further used for data analysis (e.g., times series models, generalized linear models, item response models, etc.).