UCLA
Department of Statistics Papers

Title

Model Close Match as a Criterion for Structured Model Comparison and Its Robust Statistical Tests (June 2008 Revision)

Permalink

https://escholarship.org/uc/item/4d2438fv

Authors

Li, Libo
Bentler, Peter M.

Publication Date

2008-08-11

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Model Close Match as a Criterion for Structured Model Comparison and Its Robust Statistical Tests

Libo Li and Peter M. Bentler*

University of California, Los Angeles

June 25, 2008

*Research supported in part by grants DA00017 and DA01070 from the National Institute on Drug Abuse.

Model Close Match as a Criterion for Structured Model Comparison and Its Robust Statistical Tests

Abstract

Traditional model comparison procedure selects nested structured models by evaluating the feasibility of the equality constraints that differentiate the models. We propose instead to evaluate model close match, using the distance between two models, either as important supplementary information or as a criterion for nested model comparison. Based on MacCallum, Browne and Cai (2006) and the results of Vuong (1989) and Yuan, Hayashi and Bentler (2007), we develop a reasonable cutoff value and some ADF-like tests for inference on model closeness. Simulation studies show that several of our proposed tests have robust and desirable performance in spite of severe nonnormality when sample size is as large as 150. Consequently, a two-stage procedure which combines the traditional nested model comparison and the additional inferential information regarding model close match is further suggested to improve the typical practice of model modification.

Key words: Likelihood ratio statistic, model close match, asymptotics

1. Introduction

In structural equation modeling (SEM), a set of statistics is used for evaluating the overall exact fit of the model in terms of their type I and type II errors, including the classical normal theory based likelihood ratio (NTLR) test, Browne's asymptotically distribution free test (Browne, 1984), the Satorra-Bentler scaled test (Satorra & Bentler, 1988, 1994) or the more recent residual-based tests (Yuan & Bentler, 1997, 1998, 1999). The distribution, and hence performance, of these statistics depends on meeting the various assumptions underlying these statistics. One of the assumptions is that the strict null hypothesis holds, namely, that the model is exactly correct in the population.

Another set of statistics in SEM involves the comparison of alternative nested models that contain additional restrictions beyond those of the more general model. They include the chi-square difference test which is often an NTLR test (e.g., Jöreskog, 1971; Steiger, Shapiro, & Browne, 1985), the Satorra-Bentler scaled difference test for greater robustness (e.g., Satorra, 2000; Satorra & Bentler, 2001), or the Lagrange Multiplier (LM) and Wald (W) tests (e.g., Chou & Bentler, 1990; Lee & Bentler, 1980; Lee, 1985; Sörbom, 1989) for convenience of working with only the more general model or only the more restricted model. Similarly, as when alternative models are evaluated, the distribution, and hence performance, of this set of statistics depends on meeting various assumptions (Satorra, 1989) one of which is the strict correctness of the null hypothesis, namely, that the restrictions that differentiate the general and restricted models are exactly true in the population.

In standard model modification, the significance of each statistic in this branch is determined with reference to the assumed distribution under the null. While such a procedure may not work perfectly in practice (e.g., Yuan & Bentler, 2004), especially when such a model comparison is post hoc rather than a priori (e.g., MacCallum, Roznowski & Necowitz, 1992), some type of model comparison can not be avoided in practice. Most a priori models are incorrect in some way, and the process of model modification to yield improved models remains an inevitable and important part in the application of SEM (Jöreskog, 1993). One rationale for imposing constraints on a general model is that the estimates in the more restricted and parsimonious model will be more precise (Bentler & Mooijaart, 1989).

Ever since Jöreskog (1969) developed confirmatory factor analysis, the two sets of statistical tests discussed above have been embraced in SEM because they provide scientific rigor to testing hypotheses with nonexperimental data. After some limitations were raised on the role of testing in exploratory factor analysis (Tucker & Lewis, 1973), Bentler and Bonett (1980) noted that tests of exact fit in general SEM can not on their own provide a sufficient basis for evaluation of models, especially in large samples where any restrictive null hypothesis is liable to be rejected. They proposed that a model also needs to be evaluated in terms of the extent to which it explains covariances better than a most restricted model of uncorrelated variables which explains no covariances. They provided several so-called fit indices to evaluate such an increment in fit, and also proposed to evaluate differences in model fit between two nested models by evaluating the associated increment in fit. In the meantime, additional fit indices such as the root mean square error of approximation (RMSEA, Steiger & Lind, 1980), comparative fit index (CFI, Bentler, 1990), goodness of fit index (GFI, Jöreskog & Sörbom, 1981) etc. have been devised to provide a measure of the extent of approximate or close fit of a model.

Critiques of tests of exact fit were also made from two other perspectives, namely from a rejection of the basic null hypothesis, and from the point of view of statistical theory. It does not make sense to test a specific model null hypothesis if one does not in the first place believe that a specific model might exist in the population. Any particular model may be nothing more than an approximation to reality, and it may be said that the modeling enterprise should mainly aim to provide information about the relative performance of alternative plausible models, none of which may be precisely true (e.g., Bentler & Bonett, 1980; de Leeuw, 1988; Browne & Cudeck, 1993; MacCallum, 2003). From the point of view of statistical theory, questions have been raised on whether the distribution of a test statistic under the null hypothesis provides the most meaningful possible model evaluation when such a null hypothesis may not make a priori sense. To provide an alternative, recently researchers such as Ogasawara (in press) and Yuan, Hayashi, and Bentler (2007) investigated the general distribution of the NTLR test under model misspecification and weak distributional assumptions on the data. In addition, some asymptotically robust model close fit tests implemented

via the sample RMSEA also have been introduced and studied by Li and Bentler (2006).

These critiques of hypothesis testing on exact fit of a given model apply directly to the comparison of nested models, but little statistical development has been done to provide an alternative approach for comparing such models. In this paper, we first review some relevant statistical theories and propose a framework of close match between two competing models based on MacCallum, Browne and Cai (2006). Under this framework, the measures, the cutoff values and the corresponding estimators of model close match will be given. Then, using the results of Vuong (1989) and Yuan, Hayashi and Bentler (2007), the asymptotic distribution of these estimators will be derived and some asymptotic robust tests of close match between competing models will be defined. Finally, numerical examples will be given.

2. Theoretical Background

In classical single population SEM, the relationship of p-observed variables in a $p \times 1$ random vector $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$ and m-unobserved factors may have many different specifications. Without loss of generality, we only consider two such model specifications at one time for simplicity. In one parameterization, M_1 has q free unknown parameters which are included in a $q \times 1$ parameter vector θ, while another competing parameterization M_2 has r free unknown parameters which are included in an $r \times 1$ parameter vector γ. As a result, the hypothesized model M_1 leads to the model-implied mean $\mu(\theta)$ and covariance matrix $\mathbf{\Sigma}(\theta)$ and M_2 leads to $\mu(\gamma)$ and $\mathbf{\Sigma}(\gamma)$.

For simplicity, we assume that sampling yields a complete data set. Now let $\mu = E(X)$, $\mathbf{\Sigma} = \mathrm{cov}(X)$, $\bar{X}$ and $\mathbf{S}$ be the corresponding mean and unbiased sample estimator. Let $\beta \equiv (\mu', \mathrm{vech}(\mathbf{\Sigma})')'$ and its unbiased estimator $\hat{\beta} = (\bar{X}', \mathrm{vech}(\mathbf{S})')'$, where $\mathrm{vech}(\cdot)$ is an operator which transforms a symmetric matrix into a vector by stacking the nonduplicated elements of the matrix. Suppose that the data $X_i = (x_{i1}, \ldots, x_{ip})', i = 1, \ldots, n = N + 1$ are identically and independently drawn from X. The normal theory based log likelihood function of the observations is then given by

$$l_n(\beta) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} \log f(X_i; \beta) = \text{constant} - \frac{n}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)'\mathbf{\Sigma}^{-1}(X_i - \mu)$$

where $f(X_i;\beta)$ is the density function of the multivariate normal distribution for individual observation X_i.

Let μ_0, $\mathbf{\Sigma}_0$ denote the population counterparts to μ, $\mathbf{\Sigma}$ and $\beta_0 \equiv (\mu_0', \text{vech}(\mathbf{\Sigma}_0)')$. Let $\mathbf{\Gamma}$ be the asymptotic covariance matrix of $\hat{\beta}$, then under some standard regularity conditions (e.g., Kano, 1986; Shapiro, 1984), $\hat{\beta}$ will be strongly consistent and asymptotically normally distributed, that is, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{L} N(0, \mathbf{\Gamma})$, where $\mathbf{\Gamma}$ can be shown to be equal to $A_{\beta_0}^{-1} B_{\beta_0} A_{\beta_0}^{-1}$ (e.g., Vuong, 1989; Yuan & Jennrich, 1998) with

$$A_{\beta_0} = -E\left[\frac{\partial^2 l_i(\beta_0)}{\partial \beta_0 \partial \beta_0'}\right] \qquad B_{\beta_0} = E\left[\frac{\partial l_i(\beta_0)}{\partial \beta_0}\frac{\partial l_i(\beta_0)}{\partial \beta_0'}\right]$$

where $E(\cdot)$ denotes the expectation with respect to the true distribution of X. When μ and $\mathbf{\Sigma}$ are parameterized as M_1 and M_2 as mentioned before, the corresponding log likelihood functions become $l_n(\theta)$ and $l_n(\gamma)$ separately. In SEM, the maximum likelihood estimators of θ and γ, $\hat{\theta}_{ML}$ and $\hat{\gamma}_{ML}$, are estimated by minimizing

$$F_{ML}(\bar{X}, \mathbf{S}; \theta) = (\bar{X} - \mu(\theta))' \mathbf{\Sigma}^{-1}(\theta)(\bar{X} - \mu(\theta)) + \log|\mathbf{\Sigma}(\theta)| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}(\theta)) - \log|\mathbf{S}| - p$$

and

$$F_{ML}(\bar{X}, \mathbf{S}; \gamma) = (\bar{X} - \mu(\gamma))' \mathbf{\Sigma}^{-1}(\gamma)(\bar{X} - \mu(\gamma)) + \log|\mathbf{\Sigma}(\gamma)| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}(\gamma)) - \log|\mathbf{S}| - p$$

respectively. Let θ_* and γ_* be the minimizer of $F_{ML}(\mu_0, \mathbf{\Sigma}_0; \theta)$ and $F_{ML}(\mu_0, \mathbf{\Sigma}_0; \gamma)$ respectively. Then $\hat{\theta}$ and $\hat{\gamma}$ will be strongly consistent and asymptotically normally distributed (e.g., Vuong, 1989; Yuan & Jennrich, 1998), that is, $\sqrt{n}(\hat{\theta}_{ML} - \theta_*) \xrightarrow{L} N(0, A_{\theta_*}^{-1} B_{\theta_*} A_{\theta_*}^{-1})$ and $\sqrt{n}(\hat{\gamma}_{ML} - \gamma_*) \xrightarrow{L} N(0, A_{\gamma_*}^{-1} B_{\gamma_*} A_{\gamma_*}^{-1})$.

Let $F_1 = F_{ML}(\mu_0, \mathbf{\Sigma}_0; \theta_*)$ and $F_2 = F_{ML}(\mu_0, \mathbf{\Sigma}_0; \gamma_*)$. Then NF_1 and NF_2 are the so-called noncentrality parameters of M_1 and M_2 respectively. In the model close fit literature, the noncentrality parameter is a measure of the distance between the specified model and the saturated one and plays a key role in defining many so-called fit indices. In this article, we only focus on one of these fit indices, that is, RMSEA (Browne & Cudeck, 1993; Steiger & Lind, 1980). Let $df_1 = p^* - q$ and $df_2 = p^* - r$ denote the degrees of freedom of M_1 and M_2 respectively, where $p^* = p + p(p+1)/2$. The true RMSEAs corresponding to these models

are defined as

$$r_{10} = \sqrt{\frac{F_1}{df_1}} \qquad\qquad r_{20} = \sqrt{\frac{F_2}{df_2}} \tag{1}$$

for M_1 and M_2 respectively.

3. A Close Match Framework

For convenience of illustration, we introduce our idea of model close match by using an example by Curran, Bollen, Chen, Paxton and Kirby (2003) (see their population model 2). In this example, the population model underlying the data is as follows,

$$\mathbf{y} = \mathbf{\Pi}\eta + \epsilon \qquad\qquad \eta = \mathbf{B}\eta + \zeta$$

where ϵ and ζ are independent to each other with $E(\epsilon) = 0$, Cov(ϵ)=$\mathbf{\Psi}$, $E(\zeta) = 0$, Cov(ζ)=$\mathbf{\Xi}$. Moreover, $\mathbf{\Psi}$ = diag(.51, .51, .51, .51, .51, .2895, .51, .51, .51, .2895, .2895, .51, .51, .51, .51), $\mathbf{\Xi}$ = diag(.49, .3136, .3136),

$$\mathbf{\Pi} = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & .30 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & .30 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & .30 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}'$$

$$\text{and} \qquad \mathbf{B} = \begin{bmatrix} .00 & .00 & .00 \\ .60 & .00 & .00 \\ .00 & .60 & .00 \end{bmatrix}.$$

For our illustration, we focus on four specifications used by Curran et al. (2003). They are: Specification 1 is properly specified, Specification 2 sets $\pi_{11,2}$ as zero, Specification 3 sets $\pi_{11,2}$ and $\pi_{10,3}$ as zero, and Specification 4 sets $\pi_{11,2}$, $\pi_{10,3}$ and $\pi_{6,1}$ as zero. During the model fitting of each specification, we set $\pi_{1,1}$, $\pi_{7,2}$ and $\pi_{12,3}$ to 1.0 for identification while all other nonzero parameters in the population model are set free. As a result, the models by Specification 1, 2, 3, and 4 have degrees of freedom equal to 85, 86, 87 and 88, respectively.

For model comparison, any pair of these four specifications can be a comparison pair. Let us denote the general model of a comparison pair as M_1 while the restricted one of the pair which is nested in M_1 as M_2. Let $F_{12} = F_2 - F_1$. Then following MacCallum, Browne and Cai (2006), we reparameterize the equality constraints bridging M_1 and M_2 such as $\pi_{6,1} = 0$

for the comparison pair of Specification 3 vs. 4 as the null hypothesis $H_0^E : F_{12} = 0$ and consider $H_0^E : F_{12} = 0$ as the null hypothesis of equality or exact match of the models M_1 and M_2.

Since the omitted paths (cross loadings), $\pi_{11,2}$, $\pi_{10,3}$ and $\pi_{6,1}$, are equal to .30 in the population when compared with all other loadings that are 1.0, we can consider these as minor cross loadings. In general, a simple cluster structure such as Specification 4 in this example will be very desirable from a theoretical perspective. However, real data hardly allow such a simple cluster structure, and typically may require a more complex factor loading structure, like the population model in this example where the cluster structure is compounded by some minor cross-loadings. As a result, the null hypothesis $H_0^E : F_{12} = 0$ for any pair of the four specifications above is false, and related test statistics such as the NTLR statistic will reject this hypothesis if the sample size is large enough. Then the unwanted minor paths will be included in the final model, perhaps making it less interpretable.

What we illustrated here is a typical model comparison paradigm by the traditional approach. Like exact fit tests in model overall evaluation, the traditional approach to the model comparison involves choosing between the better fit of the general model M_1 and the parsimony or meaningfulness of the restricted model M_2 by examining a statistic assessing the equality or exact match of the nested models. Even though this approach is valuable, it may not be a complete one. In practice, $H_0^E : F_{12} = 0$ may not hold because of some minor differences between two models, e.g., unexpected minor cross loadings as illustrated above. Even though these differences may be minor or unmeaningful substantively, the traditional exact match testing procedure would inevitably favor M_1 (especially in a large sample) because of the infeasibility of exact model equality. In practice, a more realistic approach to model comparison would decide between the better fit of M_1 and the parsimony or meaningfulness of M_2, using as a criterion the degree of close match instead of exact match between two models[1]. In other words, like the concept of close fit in overall model evaluation,

[1]Theoretically, imposing an inequality constraint on unwanted or unnecessary minor paths such as $\pi_{6,1} \leq .4$ and testing this by the likelihood ratio test (see Dijkstra, 1992; Shapiro, 1985) is also a possible way of handling minor paths. Unfortunately to our knowledge, there is no development of an appropriate methodology for such purpose, with existing approaches to inequality constraints requiring a correctly specified fitting function. This requirement limits their application to close fitting models, especially when the

the concept of close match between M_1 and M_2 may yield an appropriate comparison of two models. As an additional approach to model comparison, this may yield a more practical criterion for model modification in substantive research.

Actually, our idea of close match is not completely new. MacCallum et al. (2006) realized the infeasibility of the null hypothesis $H_0^E : F_{12} = 0$ and advocated *the good-enough principle*, as presented by Serlin and Lapsley (1985), for testing for a small difference between nested models. They further suggested the following null hypothesis

$$H_0 : F_{12} \leq \delta_C \tag{2}$$

for such testing, where δ_C is some chosen small number. In this article, we consider (2) as the null hypothesis of model close match. In practice, the competing models vary in each study. In order to choose an appropriate value for δ_C across situations, MacCallum et al. (2006) expressed δ_C in terms of overall fit of M_1 and M_2, that is,

$$\delta_C = df_2 r_{2C}^2 - df_1 r_{1C}^2 \tag{3}$$

where r_{1C} and r_{2C} are some choice of true RMSEA values for M_1 and M_2. MacCallum et al. (2006) suggest to use a range of reasonable pairs of r_{1C} and r_{2C} for power analysis.

In this article, we further extend the idea of (2) and (3). Let us combine the identity $F_{12} = df_2 r_{20}^2 - df_1 r_{10}^2$ implied by (1) with (2) and (3). Then we obtain a null hypothesis of close match equivalent to (2), that is,

$$H_0 : df_2 r_{20}^2 - df_1 r_{10}^2 \leq df_2 r_{2C}^2 - df_1 r_{1C}^2. \tag{4}$$

Now let us assume $df_1 = 85$ and $df_2 = 86$ as in Specification 1 vs. 2 of our illustrative example, then the reasonable area of a r_{10} and r_{20} pair is above Line 0 in Figure 1, where Line 0 is the line $r_{20} = \sqrt{df_1/df_2} \cdot r_{10}$ where $F_{12} = 0$. Clearly Line 0 is lower than the diagonal line since $\sqrt{df_1/df_2} < 1$. When $H_0^E : F_{12} = 0$ holds, $r_{20} - r_{10} \leq 0$ and varies along Line 0 while Line 0 varies with df_1 and df_2. So the value of $r_{20} - r_{10}$ does not accurately reflect the exact or close match of two competing models. We prefer to use F_{12} for close match as in (2) instead of $r_{20} - r_{10}$.

data is nonnormal.

Now let p_1 denote the point (.06, .08) in Figure 1 and as our choice of (r_{1C}, r_{2C}) in (4). Then there should exist one line, the Line A in Figure 1, which consists of a series of points $(r_{1C'}, r_{2C'})$ in Figure 1 and satisfies the following equality

$$df_2 r_{2C'}^2 - df_1 r_{1C'}^2 = df_2 r_{2C}^2 - df_1 r_{1C}^2. \tag{5}$$

Similarly, when the point $p_2 = (0.07, 0.10)$ is chosen for (4), then the Line B exists in Figure 1 by (5). In fact, Line 0 is also a line satisfying (5). So no matter what the reasonably chosen point (r_{1C}, r_{2C}) is, there will always exist a line in the figure consisting of points satisfying (5). Let us call this line the equi-discrepancy line. Then Line 0 can be called the equi-discrepancy line of exact match while any other line above it will be the equi-discrepancy line of close match. Further, by (4), the meaning of the null hypothesis of close match can be redefined as testing if the true RMSEA pair (r_{10}, r_{20}) which we don't know exactly at hand is inside the area between the equi-discrepancy line of close match defined by the chosen RMSEA pair (r_{1C}, r_{2C}) and the equi-descrpancy line of exact match, the line 0, or not. Clearly, when the chosen RMSEA pair (r_{1C}, r_{2C}) represents a higher equi-discrepancy line of close match, then the area of testing is larger, δ_C is larger and greater discrepancy is allowed between nested models. In reverse, when (r_{1C}, r_{2C}) represents a lower line, then the area of testing is smaller, δ_C is smaller and less discrepancy is allowed.

More importantly, any equi-discrepancy line in Figure 1 by (5) will cross the vertical axis and has a point $(r_{1C_0'}, r_{2C_0'})$ with $r_{1C_0'} = 0$. For example, $(r_{1C_0'}, r_{2C_0'}) = (0, 0.053)$ for Line A while $(r_{1C_0'}, r_{2C_0'}) = (0, 0.072)$ for Line B. Combining $(r_{1C_0'}, r_{2C_0'})$ with (3) and (5), we obtain

$$\delta_C = df_2 r_{2C}^2 - df_1 r_{1C}^2 = df_2 r_{2C_0'}^2. \tag{6}$$

Notice that by (6), any chosen discrepancy defined by δ_C and $df_2 r_{2C}^2 - df_1 r_{1C}^2$ for model comparison is equivalent to a discrepancy between the saturated model and a close fitted model with df_2 degrees of freedom and the true RMSEA $= r_{2C_0'}$. For example, p_1 on Line A can be considered to represent an overall discrepancy of a model with $df = 86$ and the true RMSEA=0.053 against the saturated model while for p_2 on Line B this close fitted model has $df = 86$ and a true RMSEA=0.072. Thus, by (6), we translate the choice of δ_C or (r_{1C}, r_{2C})

for the model close match into an equivalent choice of $r_{2C_0'}$ for a close fitted reference model with its degrees of freedom equal to df_2. The advantage of such translation is obvious. The choice of $r_{2C_0'}$ sets us free from a two dimensional choice of (r_{1C}, r_{2C}) and also allows us to utilize the cutoff criteria of RMSEA which have been established in SEM for a close fitted reference model. It is commonly believed that an RMSEA equal to 0.053 or 0.072 means a discrepancy of some mild misspecification. So p_1 and p_2 by their reference modes with $r_{2C_0'} = 0.053$ and $r_{2C_0'} = 0.072$ respectively can be considered to allow too much discrepancy between two competing models. In SEM, 0.05 is widely accepted RMSEA cutoff value for model close fit and can be used as $r_{2C_0'}$ to define δ_C. However, the equi-discrepancy line for the point (0,0.05) lies above the line $r_{20} = 0.05$. This means that by such cutoff point r_{20} is allowed to be greater than 0.05 even when r_{10} is as little as between 0 and 0.01. So such a cutoff point is reasonable in terms of model overall fit on one side, but the corresponding $(r_{1C'}, r_{2C'})$ makes less sense in terms of model comparison on another side and may allow too much discrepancy. In addition, if (0,0.01) is used as cutoff point, the tolerable discrepancy may be too little. So as a compromise we decide to choose $p_3 = (0, 0.03)$ as our cutoff point and denote the corresponding cutoff value δ_C as $\delta_{C,df_2,0.03}$. For Specification 1 vs. 2 in our example, $\delta_{C,86,0.03} = 86 \cdot 0.03^2 = 0.0774$ while $\delta_{C,87,0.03} = 0.0783$ and $\delta_{C,88,0.03} = 0.0792$ are the cutoff values when the restricted models in the comparison are Specification 3 and 4 respectively.

It is widely accepted that the population RMSEA is a model and sample-size independent measure of model overall fit. So our cutoff point p_3 doesn't need to change with df_1 or df_2. It holds in general and remains the same meaning no matter what the competing models are, even though the value of $\delta_{C,df_2,0.03}$ and the shape of the corresponding equi-discrepancy line, Line C, may change in each case by (5) and (6).

Another interesting point we should mention here is that the interval between an equi-discrepancy line of close match and Line 0 always shrinks as r_{10} increases as in Figure 1. Substantively, this means that under the same tolerable model discrepancy, our close match approach by (4) allows more restriction or parsimony (larger $r_{20} - r_{10}$) when the general model contains less misspecification (smaller r_{10}), or equivalently, less restriction (smaller

$r_{20} - r_{10}$) when the general model becomes less trustable (larger r_{10}). Although we don't know what the values of r_{10} and r_{20} are, our close match approach by (4) automatically set a corresponding standard for r_{10} and r_{20} by the equi-discrepancy lines.

Now let $R_{12} = \sqrt{F_{12}}$. In this article, we treat R_{12} as another measure of model close match. Then the null hypothesis (2) can be redefined as

$$H_0 : R_{12} \leq \sqrt{\delta_C} \tag{7}$$

while the same meaning remains. When R_{12} is divided by $\sqrt{df_{12}}$, it becomes the RDR index (the root discrepancy per restriction) proposed by Browne and du Toit (1992). Although Browne now recommends against the use of RDR index (see MacCallum et al. 2006), it is possible to test such an index meaningfully by our close match framework above. Of course, R_{12} can also be divided by $\sqrt{df_1}$ or $\sqrt{df_2}$ to form a RMSEA-like index that has some special meaning. In this article, we feel that the discussion in this direction is out of our scope.

Now let us calculate F_{12} and R_{12} for all possible specification pairs in our illustrative example. The results are presented in Table A. By our cutoff values $\delta_{C,86,0.03}$, $\delta_{C,87,0.03}$ and $\delta_{C,88,0.03}$ calculated before, all specification pairs on the diagonal line are acceptable while the off-diagonal specification pairs are unacceptable.

4. General distribution of likelihood ratio statistics

Let $\hat{F}_1 = F_{ML}(\bar{X}, \mathbf{S}; \hat{\theta}_{ML})$ and $\hat{F}_2 = NF_{ML}(\bar{X}, \mathbf{S}; \hat{\gamma}_{ML})$. Then $T_{ML\_12} = N\hat{F}_2 - N\hat{F}_1 \xrightarrow{L} \chi^2_{df_{12}}$ under normality is the well-known NTLR test statistic that is used to test $H_0^E : F_{12} = 0$ in a nested model comparison. When H_0^E doesn't hold, the inequality of two nested models becomes true and T_{ML_12} follows $\chi^2_{df_{12}}(NF_{12})$ under normality and the population drift assumption which is

$$\mu_0 - \mu(\theta_*) = O(1/\sqrt{n}) \text{ and } \mathbf{\Sigma}_0 - \mathbf{\Sigma}(\theta_*) = O(1/\sqrt{n}) \tag{8}$$

$$\mu_0 - \mu(\gamma_*) = O(1/\sqrt{n}) \text{ and } \mathbf{\Sigma}_0 - \mathbf{\Sigma}(\gamma_*) = O(1/\sqrt{n}) \tag{9}$$

(e.g., Satorra, 1989; Satorra & Saris, 1985; Steiger, Shapiro, & Browne, 1985). Although this noncentral chi-square distribution of T_{ML_12} can be used for testing H_0 in (2) and its

equivalence (4) and (7) under the inequality of two nested models, the assumptions of normality and population drift are hard to satisfy or verify in practice. These limitations prevent T_{ML_12} from being the appropriate statistic to use in such practical testing situations. Satorra (1989) further proposed a generalized score test and generalized wald test which drop the assumption of normality and are asymptotically noncentral chi-square distributed under the inequality of two nested models. However, the noncentrality parameters of their noncentral chi-square distributions contain $\mathbf{\Gamma}$ which is based on the distribution of the data and varies with its nonnormality. Such distributional dependence of the noncentrality parameters, along with the requisite population drift assumption, similarly raise questions about the appropriateness of using these statistics for testing model close match in (2).

Given the inadequacy of existing methods for testing of model close match, we turn our attention to some results of Vuong (1989) and Yuan, Hayashi and Bentler (2007). Now let

$$\omega^2 = E\left[\log\left[\frac{f(X_i;\theta_*)}{f(X_i;\gamma_*)}\right]\right]^2 - \left[E\left[\log\left[\frac{f(X_i;\theta_*)}{f(X_i;\gamma_*)}\right]\right]\right]^2. \tag{10}$$

Then by Yuan, Hayashi and Bentler (2007), we obtain the following corollary, that is,

Corollary 1. Under standard regularity conditions as in Yuan and Bentler (1997),

$$\sqrt{n}\left(\hat{F}_{12} - F_{12}\right) \quad \xrightarrow{L} \quad N(0, 4\omega^2) \tag{11}$$

if $\omega^2 \neq 0$ or equivalently if $F_{12} \neq 0$ when two models are nested.

One point which should be mentioned is that this asymptotic approximation holds only when $\omega^2 \neq 0$. Vuong (1989) pointed out that the equivalence between $\omega^2 = 0$ and $f(X_i;\theta_*) = f(X_i;\gamma_*)$ holds in general (see Lemma 4.1 by Vuong). For nested models, it can be showed that $f(X_i;\theta_*) = f(X_i;\gamma_*)$ and $F_{12} = 0$ are equivalent to each other under standard regularity conditions (see Lemma 7.1 by Vuong, 1989). So a rejection of the equality or exact match of two nested models: $F_{12} = 0$, which is equivalent to $f(X_i;\theta_*) = f(X_i;\gamma_*)$, is a way to establish $\omega^2 \neq 0$ and should be conducted before the use of (11).

In fact, $\hat{F}_{12}$ has some asymptotic bias as an estimator of F_{12}. It can be shown (e.g., Li & Bentler, 2006; Vuong 1989) that

$$\mathbf{AE}(T_{ML\_12}) = NF_{12} + d_2 - d_1 \tag{12}$$

where $d_1 = \text{tr}(A_{\beta_0}^{-1}B_{\beta_0} - A_{\theta_*}^{-1}B_{\theta_*})$ and $d_2 = \text{tr}(A_{\beta_0}^{-1}B_{\beta_0} - A_{\gamma_*}^{-1}B_{\gamma_*})$, and $\mathbf{AE}$ represents the asymptotic expectation with respect to the true distribution of X.

Let $\sigma_{1*} = (\mu(\theta_*)', \text{vech}(\mathbf{\Sigma}(\theta_*))')'$, $\sigma_{2*} = (\mu(\gamma_*)', \text{vech}(\mathbf{\Sigma}(\gamma_*))')'$, $\dot{\sigma}_{1*} = \partial\sigma_{1*}/\partial\theta_*$ and $\dot{\sigma}_{2*} = \partial\sigma_{2*}/\partial\gamma_*$ respectively and let $\mathbf{D}_p$ be the duplication matrix as defined by Magnus and Neudecker (1988). Then we define

$$\begin{aligned}
\mathbf{W}_{1*} &\equiv \text{diag}\left[\mathbf{\Sigma}^{-1}(\theta_*), 2^{-1}\mathbf{D}_p'(\mathbf{\Sigma}^{-1}(\theta_*) \otimes \mathbf{\Sigma}^{-1}(\theta_*))\mathbf{D}_p\right] \\
\mathbf{W}_{2*} &\equiv \text{diag}\left[\mathbf{\Sigma}^{-1}(\gamma_*), 2^{-1}\mathbf{D}_p'(\mathbf{\Sigma}^{-1}(\gamma_*) \otimes \mathbf{\Sigma}^{-1}(\gamma_*))\mathbf{D}_p\right] \\
\mathbf{U}_1 &\equiv \mathbf{W}_{1*} - \mathbf{W}_{1*}\dot{\sigma}_{1*}(\dot{\sigma}_{1*}'\mathbf{W}_{1*}\dot{\sigma}_{1*})^{-1}\dot{\sigma}_{1*}'\mathbf{W}_{1*} \\
\mathbf{U}_2 &\equiv \mathbf{W}_{2*} - \mathbf{W}_{2*}\dot{\sigma}_{2*}(\dot{\sigma}_{2*}'\mathbf{W}_{2*}\dot{\sigma}_{2*})^{-1}\dot{\sigma}_{2*}'\mathbf{W}_{2*}
\end{aligned}$$

Then it can be shown (e.g., Li & Bentler, 2006; Yuan & Marshall 2004) that under the population drift assumption (8) and (9),

$$\mathbf{AE}(T_{ML\_12}) = NF_{12} + \text{tr}(\mathbf{U}_2\mathbf{\Gamma}) - \text{tr}(\mathbf{U}_1\mathbf{\Gamma}). \tag{13}$$

When normality is assumed also, this reduces to

$$\mathbf{AE}(T_{ML\_12}) = NF_{12} + df_{12} \tag{14}$$

Combining (12), (13), (14) and Corollary 1, we obtain the following Corollary,

Corollary 2. Under standard regularity conditions as in Yuan and Bentler (1997),

$$\sqrt{n}\left(\hat{F}_{12} - F_{12} - \frac{d_2}{n} + \frac{d_1}{n}\right) \xrightarrow{L} N(0, 4\omega^2)$$

if $\omega^2 \neq 0$ or equivalently if $F_{12} \neq 0$ when two models are nested. Under the population drift assumption, this reduces to

$$\sqrt{n}\left(\hat{F}_{12} - F_{12} - \frac{\text{tr}(\mathbf{U}_2\mathbf{\Gamma})}{n} + \frac{\text{tr}(\mathbf{U}_1\mathbf{\Gamma})}{n}\right) \xrightarrow{L} N(0, 4\omega^2)$$

When normality is assumed also, it reduces to

$$\sqrt{n}\left(\hat{F}_{12} - F_{12} - \frac{df_2}{n} + \frac{df_1}{n}\right) \xrightarrow{L} N(0, 4\omega^2)$$

Notice that Corollary 2 has no conflict with Corollary 1 because the extra terms d_1/n, d_2/n, $\text{tr}(\mathbf{U}_1\boldsymbol{\Gamma})/n$, $\text{tr}(\mathbf{U}_2\boldsymbol{\Gamma})/n$, df_1/n and df_2/n in Corollary 2 approach zero as n goes to infinity.

In the last section, we proposed R_{12} as another measure of model close match. Following the sample RMSEA definition in SEM, we define a sample estimate of R_{12},

$$\hat{R}_{12} = \sqrt{\max\left(\hat{F}_{12} - \frac{df_2}{n} + \frac{df_1}{n}, 0\right)}$$

and also define a relatively robust one by Corollary 2, that is,

$$\tilde{R}_{12} = \sqrt{\max\left(\hat{F}_{12} - \frac{\text{tr}(\hat{\mathbf{U}}_2\hat{\boldsymbol{\Gamma}})}{n} + \frac{\text{tr}(\hat{\mathbf{U}}_1\hat{\boldsymbol{\Gamma}})}{n}, 0\right)}.$$

where $\hat{\boldsymbol{\Gamma}}$ is the consistent estimator of $\boldsymbol{\Gamma}$ (e.g., Bentler, 2006), and $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ are consistent estimators of $\mathbf{U}_1$ and $\mathbf{U}_2$ obtained by replacing θ_* and γ_* by $\hat{\theta}_{ML}$ and $\hat{\gamma}_{ML}$ respectively. Then by Corollary 1 and the Delta method, we obtain the following corollary.

Corollary 3. Given $\omega^2 \neq 0$ or equivalently if $F_{12} \neq 0$ when two models are nested, then under some standard regularity conditions as in Yuan and Bentler (1997)

$$\sqrt{n}\left(\hat{R}_{12} - R_{12}\right) \xrightarrow{L} N\left(0, \frac{\omega^2}{F_{12}}\right)$$

and

$$\sqrt{n}\left(\tilde{R}_{12} - R_{12}\right) \xrightarrow{L} N\left(0, \frac{\omega^2}{F_{12}}\right)$$

Proof.

$$\sqrt{n}\left(\hat{R}_{12} - R_{12}\right) \stackrel{a}{=} \sqrt{n}\left(\sqrt{\hat{F}_{12}} - \sqrt{F_{12}}\right) \xrightarrow{L} N\left(0, \frac{\omega^2}{F_{12}}\right) \qquad \text{(Delta method)}$$

The distribution of $\tilde{R}_{12}$ can be proved in the same way.

Now let us plug $\hat{\theta}_{ML}$ and $\hat{\gamma}_{ML}$ into (10). Then we obtain an consistent estimator of ω^2, that is,

$$\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left[\frac{f(X_i;\hat{\theta}_{ML})}{f(X_i;\hat{\gamma}_{ML})}\right]\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{f(X_i;\hat{\theta}_{ML})}{f(X_i;\hat{\gamma}_{ML})}\right]\right]^2. \tag{15}$$

Yuan, Hayashi and Bentler (2007) further derived an explicit form for ω^2 under various conditions and gave the corresponding estimators. Although their work is valuable, our

preliminary results from a simulation study of normal data show that there is no big difference in performance between their estimators and $\hat{\omega}^2$ in (15). More importantly, their estimators are limited to single group mean and covariance structure analysis and are not as general as $\hat{\omega}^2$. So, in this article, we use $\hat{\omega}^2$ for the tests that follow.

5. Tests of model close match

In section 3, we proposed the null hypothesis $H_0 : \ F_{12} \leq \delta_C$ against its alternative $H_1 : \ F_{12} > \delta_C$ as the way to test model close match. Suppose now for two nested models, H_0 is true and $T_{ML\_12} \xrightarrow{L} \chi^2_{df_{12}}(NF_{12})$, then $\Pr\{T_{ML\_12} > \chi^2_{df_{12},.95}(N\delta_C)\}$ is asymptotically no more than .05. So a test of close match can be proposed as follow: reject H_0 in favor of H_1 if T_{ML_12} is greater than $\chi^2_{df_{12},.95}(N\delta_C)$. Otherwise, H_0 can not be rejected.

The last section gave some asymptotic results for $\hat{F}_{12}$, and a consistent estimator of ω^2 was given in (15). Based on these results, we propose two test statistics for model close match. The first is Yuan-Hayashi-Bentler test statistic (T_1), which is

$$T_1 = \frac{\sqrt{n}\left(\hat{F}_{12} - \delta_C - df_2/n + df_1/n\right)}{2\hat{\omega}}$$

as well as a robust version of Yuan-Hayashi-Bentler test statistic[2] (T_2), which is

$$T_2 = \frac{\sqrt{n}\left(\hat{F}_{12} - \delta_C - \mathrm{tr}(\hat{\mathbf{U}}_2\hat{\mathbf{\Gamma}})/n + \mathrm{tr}(\hat{\mathbf{U}}_1\hat{\mathbf{\Gamma}})/n\right)}{2\hat{\omega}}$$

Corollary 4. Given $\omega^2 \neq 0$ or equivalently $F_{12} \neq 0$ when two models are nested, then under some standard regularity conditions as in Yuan and Bentler (1997)

$$T_1 \overset{a}{=} T_2 \xrightarrow{L} N(\sqrt{n}\delta_1, 1) \qquad \text{where} \qquad \delta_1 = \frac{F_{12} - \delta_C}{2\omega}$$

and

1. When $F_{12} = \delta_C$, then $\delta_1 = 0$ and $T_1 \overset{a}{=} T_2 \xrightarrow{L} N(0, 1)$.

[2] $\mathbf{U}_1$ and $\mathbf{U}_2$ defined in this article and used for T_2 is not equal to and less general than $\mathbf{U}_1$ and $\mathbf{U}_2$ defined by Yuan, Hayashi & Bentler (2007). However, when the population drift assumption holds, two types of terms are equivalent, and T_2 and its counterpart in Yuan, Hayashi & Bentler (2007) should perform closely.

2. When $F_{12} > \delta_C$, then $\delta_1 > 0$ and $T_1 \stackrel{a}{=} T_2 \longrightarrow +\infty$ as $n \longrightarrow +\infty$.

3. When $F_{12} < \delta_C$, then $\delta_1 < 0$ and $T_1 \stackrel{a}{=} T_2 \longrightarrow -\infty$ as $n \longrightarrow +\infty$.

Let $\lambda_{.95}$ be 95 percent quantile of the standard normal distribution, then $\Pr\{T_1 \text{ or } T_2 > \lambda_{.95}\}$ is asymptotically no more than .05 under H_0. Clearly, T_1 and T_2 can be used to test H_0 in (2). For each of them, H_0 will be rejected if it is greater than $\lambda_{.95}$. Otherwise, H_0 can not be rejected.

In last section, we also get some results for $\hat{R}_{12}$ and $\tilde{R}_{12}$. So by Corollary 3, we define a test statistic T_3 for testing H_0 in (7) that is equivalent to H_0 in (2), which is

$$T_3 = \frac{\sqrt{n}\left(\hat{R}_{12} - \sqrt{\delta_C}\right)}{\hat{\omega}\Big/\sqrt{\hat{F}_{12} - df_2/n + df_1/n}}$$

and a relatively robust version, T_4, which is

$$T_4 = \frac{\sqrt{n}\left(\tilde{R}_{12} - \sqrt{\delta_C}\right)}{\hat{\omega}\Big/\sqrt{\hat{F}_{12} - df_2/n + df_1/n}}$$

Let $\hat{c} = (\mathrm{tr}(\hat{\mathbf{U}}_2\hat{\mathbf{\Gamma}}) - \mathrm{tr}(\hat{\mathbf{U}}_1\hat{\mathbf{\Gamma}}) - df_2 + df_1)/n$. Following Li and Bentler (2006), we further define another two robust test statistics T_5 and T_6 as

$$T_5 = \frac{\sqrt{n}\left(\tilde{R}_{12} - \sqrt{\delta_C}\right)}{\sqrt{\hat{\omega}^2 - \hat{c}}\Big/\sqrt{\hat{F}_{12} - df_2/n + df_1/n + \hat{c}}}$$

and

$$T_6 = \frac{\sqrt{n}\left(\tilde{R}_{12} - \sqrt{\delta_C}\right)}{\sqrt{\hat{\omega}^2 - 2.5\cdot\hat{c}}\Big/\sqrt{\hat{F}_{12} - df_2/n + df_1/n + \hat{c}}}$$

Clearly, $\hat{c}$ is an estimator of $c_0 = (\mathrm{tr}(\mathbf{U}_2\mathbf{\Gamma}) - \mathrm{tr}(\mathbf{U}_1\mathbf{\Gamma}) - df_2 + df_1)/n$ and converges to c_0 in the order of $O_p(n^{-3/2})$. When the data is normal, c_0 is equal to zero and $\hat{c}$ will converge to zero in the order of $O_p(n^{-3/2})$. So in this condition, T_3, T_4, T_5 and T_6 should have similar performance. When the data is nonnormal, c_0 and thus $\hat{c}$ carry information on nonnormality of the data. So compared to T_3, T_4 has a correction in the numerator while T_5 and T_6 have a correction both in numerator and denominator. Even though such corrections should not matter asymptotically, they may make a difference in performance with small samples.

Another point which should be mentioned here is that when two models are nested, one or several quantities among $\hat{F}_{12} - df_2/n + df_1/n$, $\hat{F}_{12} - df_2/n + df_1/n + \hat{c}$, $\hat{\omega}^2 - \hat{c}$ and $\hat{\omega}^2 - 2.5 \cdot \hat{c}$ can be less than or equal to zero especially in a small sample. Then the corresponding test statistics T_3, T_4, T_5 or T_6 will be undefined respectively. So during the simulations below, replications with such a problem will be discarded.

Corollary 5. Given $\omega^2 \neq 0$ or equivalently if $F_{12} \neq 0$ when two models are nested, then under some standard regularity conditions as in Yuan and Bentler (1997)

$$T_3 \stackrel{a}{=} T_4 \stackrel{a}{=} T_5 \stackrel{a}{=} T_6 \stackrel{L}{\longrightarrow} N(\sqrt{n}\delta_2, 1) \qquad \text{where} \qquad \delta_2 = \frac{F_{12} - \sqrt{F_{12} \cdot \delta_C}}{\omega}$$

and

1. When $R_{12} = \sqrt{\delta_C}$, then $\delta_2 = 0$ and $T_3 \stackrel{a}{=} T_4 \stackrel{a}{=} T_5 \stackrel{a}{=} T_6 \stackrel{L}{\longrightarrow} N(0, 1)$.
2. When $R_{12} > \sqrt{\delta_C}$, then $\delta_2 > 0$ and $T_3 \stackrel{a}{=} T_4 \stackrel{a}{=} T_5 \stackrel{a}{=} T_6 \longrightarrow +\infty$ as $n \longrightarrow +\infty$.
3. When $R_{12} < \sqrt{\delta_C}$, then $\delta_2 < 0$ and $T_3 \stackrel{a}{=} T_4 \stackrel{a}{=} T_5 \stackrel{a}{=} T_6 \longrightarrow -\infty$ as $n \longrightarrow +\infty$.

Clearly, after a rejection of exact match, like T_1 and T_2 discussed before, T_3, T_4, T_5 or T_6 also can be used to test the null hypothesis of close match H_0 in (7). H_0 will be rejected for each statistic if its estimate is greater than $\lambda_{.95}$. Otherwise, it can not be rejected.

Corollary 6. Under $H_1 : F_{12} > \delta_C$ and some standard regularity conditions as in Yuan and Bentler (1997), then T_3, T_4, T_5 and T_6 have more asymptotic power than T_1 and T_2 to reject the null hypothesis of model close match.

Proof. By Corollary 4 and 5,

$$T_1 \stackrel{a}{=} T_2 \stackrel{L}{\longrightarrow} N(\sqrt{n}\delta_1, 1) \qquad\qquad T_3 \stackrel{a}{=} T_4 \stackrel{a}{=} T_5 \stackrel{a}{=} T_6 \stackrel{L}{\longrightarrow} N(\sqrt{n}\delta_2, 1)$$

where

$$\delta_1 = \frac{F_{12} - \delta_C}{2\omega} = \frac{\sqrt{F_{12}} - \sqrt{\delta_C}}{\omega} \cdot \frac{\sqrt{F_{12}} + \sqrt{\delta_C}}{2}, \qquad \delta_2 = \frac{F_{12} - \sqrt{F_{12} \cdot \delta_C}}{\omega} = \frac{\sqrt{F_{12}} - \sqrt{\delta_C}}{\omega} \cdot \sqrt{F_{12}}$$

Clearly, $(\sqrt{F_{12}} + \sqrt{\delta_C})/2 < \sqrt{F_{12}}$ and thus $\delta_1 < \delta_2$ under H_1.

6. Examples

To evaluate whether the seven proposed statistics in the previous section are reliable tools for testing $H_0 : F_{12} \leq \delta_C$, we first look at the asymptotic approximation and one-sided type I errors of these statistics when $F_{12} = \delta_C$. It is hard to manipulate the level of F_{12} to a specific value δ_C such as the suggested cutoff value $\delta_{C,df_2,0.03}$. So in our two examples below, we first set $\delta_C = F_{12}$ for all statistics since F_{12} is known in a simulation study. Thus, for each statistic, a desirable approximation to its theoretical distribution and a desirable rejection rate close to .05 across conditions will suggest it as a reliable close match test. Otherwise, it should not be used. After this first step, we set $\delta_C = \delta_{C,df_2,0.03}$ in each statistics. Then we can examine the acceptance or rejection performance of our statistics when $F_{12} < \delta_C = \delta_{C,df_2,0.03}$ or $F_{12} > \delta_C = \delta_{C,df_2,0.03}$. Since only our second example has both a comparison pair with $F_{12} < \delta_{C,df_2,0.03}$ and a pair with $F_{12} > \delta_{C,df_2,0.03}$ (see the comparison pairs used below), and thus may have a differential performance for the statistics across pairs, we present the performance of our statistics only for the second example.

We generated normal, mild nonnormal and severe nonnormal data for each of our examples. In the mild nonnormal condition, the skewness and kurtosis of each observed variable is set to 1.0 and 3.0 during data generation. In the severe nonnormal condition, they are set to 2.0 and 7.0. For all examples, the sample size levels are set to 150, 300, 500 and 1000. So there are $3 \times 4 = 12$ data conditions for each example. The number of replications is set to 2000 under each data condition.

The whole data generation and model fitting were conducted by using EQS 6.1 (Bentler, 2006). In addition, we specified SE=OBS during the analysis. Thus, the term $(\hat{\dot{\sigma}}_*{}' \hat{\mathbf{W}}_* \hat{\dot{\sigma}}_*)$, the Fisher information estimator, in $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$ is replaced by the estimator of the Hessian or observed information matrix.

Example 1: Unwanted Paths. The example in section 3 is our first example. It contains some unwanted paths, which occurs frequently in SEM practice. For simulation, the population covariance matrix is generated from the model in section 3 and the replications are generated under each of 12 data conditions. Then different specifications are fitted to each replication. For this example, we only study the performance of the comparison pairs on the diagonal line of Table A. For all these pairs, we set $\delta_C = F_{12}$ in all statistics as mentioned

before. Then the rejection rates of all statistics for Specification 1 vs. 2, Specification 2 vs. 3 and Specification 3 vs. 4 are presented in Table 1A-1C, Table 2A-2C and Table 3A-3C respectively. In addition, as to Specification 1 vs. 2, in Figures 2 and 3 we also present the QQ-plots of T_1 to T_6 against $N(0,1)$ with $n = 150$ for normal and severe nonnormal data respectively.[3] As we mentioned before, T_3, T_4, T_5 and T_6 can be undefined if some elements in the denominators of their definitions are less than or equal to zero, and then will be discarded from the analysis. We put the number of undefined replications on the corresponding QQ-plots in Figures 2 and 3 as well as in parenthesis after the rejection rates in each cell of the tables.

In Table 1A, under the normal condition, T_{ML_12} performs well across the sample sizes. However, in Tables 1B and 1C, across the sample sizes, the inflation of the rejection rates of T_{ML_12} increases as the nonnormality of the data increases. Its performance is poor, especially with severe nonnormal data. In Figures 2 and 3, the QQ-plots of T_1 and T_2 show that they have a poor normal approximation. In Table 1A-1C, T_1 and T_2 perform poorly across all sample sizes. They overaccept in all conditions. In Figures 2 and 3, the QQ-plots show that T_3, T_4, T_5 and T_6 are well approximated by $N(0,1)$ at the upper tail, while the corresponding lower tails are not similarly well approximated. By examining all other QQ-plots (which are not presented here), we find that such contrasting performance between the upper and lower tails also occurs with T_3, T_4, T_5 and T_6 in all other comparison pairs in this article when the sample sizes are small to medium and especially when the data is severely nonnormal. Since the test of the hypothesis of close match is a one-sided test, the upper tail approximation of the test statistics is crucial to test the hypothesis of close match. A reliable upper tail convergence will be very valuable for testing purposes and could be considered as the major criterion for validating the close match test statistics, especially with small or medium samples where the overall performance of a statistic sometimes may not be good. In Tables 1A-1C, T_3, T_4, T_5 and T_6 are undefined over ten percent of the time when $n = 150$. Their failures increase as nonnormality increases, but failure reduces dramatically when n

[3]Since the number of data conditions for our two examples is too many to present here, we present Figures 2 and 3 for Example 1 only. The rest will be available upon request.

increases to 300 for all data conditions. Their rejection performance is consistent across the sample sizes in the three tables. The rejection rates are close to the target .05 for all conditions, even though they, and especially T_4, have a slight tendency to overaccept across the tables.

The results on seven test statistics for Specification 2 vs. 3 in Tables 2A-2C, and Specification 3 vs. 4 in Tables 3A-3C, are very similar. When the data is normal, T_{ML_12} performs well for Specification 2 vs. 3 and Specification 3 vs. 4 as long as the sample size is 150. However, for both specification pairs, overrejections of the null hypothesis occurs for all sample sizes as the nonnormality of the data increases. Its performance becomes especially poor when the data is severely nonnormal. As before, T_1 and T_2 perform poorly across the sample sizes in all six tables. With the same data, the rejection rates of T_3, T_4, T_5 and T_6 for Specification 2 vs. 3 and Specification 3 vs. 4 reach the target level of .05 more closely than in Specification 1 vs. 2 in most conditions across the six tables. This may be due to the increased F_{12} values of these two specification pairs (see Table A). Also in three tables of each specification pair in this example, the number of undefined cases for the four statistics generally decreases in the corresponding cells along three specification pairs. This also may be due to the increased F_{12} values along these pairs.

Example 2: Model Uncertainty. In our previous example researchers may adopt our close match based methods because they may have a strong a priori reason to favor a clean structure and to reject unwanted paths in spite of the lack of support from exact match based test statistics such as the NTLR test. However, perhaps a more typical situation occurs when a researcher does not have a strong substantive preference for a specific model. In SEM practice, there often may be many models that can be considered meaningful substantively for a single data set. This is certainly true in exploratory factor analysis. For example, both a two factor model and a three factor model may be interpretable for some psychological data. Typically, the NTLR test will give support to the model with more parameters if the extra factor in the three factor model can capture some extra characteristics of the population. On the other hand, methods like AIC and BIC, due to their assigning a penalty for more parameters in the model, sometimes may yield the opposite result. Similarly,

other indicators such as χ^2/df ratio or various fit indices may indicate only a small distance between the models. Even though these supplementary criteria are valuable, they are not probabilistic criteria and hence they do not optimally allow inference on the discrepancy between models in the population. As a remedy, our close match based test statistics are probabilistic decision criteria like the traditional exact testing or NTLR tests. By their very definition, like AIC and BIC, our close match approach already includes a tradeoff between model goodness of fit and model parsimony.

In this example, we illustrate our close match approach to solving model uncertainty by using an example from a TOEFL® iBT test[4] developed by the Educational Testing Service. For this TOEFL® iBT test, the variables in the original data (n=774) were grouped within each of four test sections: Speaking, Writing, Reading and Listening. After some parceling, there are 22 variables: six variables for Speaking, two variables for Writing, eight variables for Reading and six variables for Listening. In the language assessment area, there is not a consensus on the number of factors underlying data such as this. As a result, models with a different number of factors have been hypothesized and studied (Bachman, Davidson, Ryan, & Choi, 1995; Carroll, 1983; Hale, Rock, & Jirele, 1989; Kunnan, 1995; Swinton & Powers, 1980). In our example, we focus on only three specifications. In Specification 1, there are three factors: one factor is for all variables in the Speaking section, one factor for all variables in the Writing section, and one factor for all variables in the Reading and Listening section. In Specification 2, there are two factors: one factor is for all variables in the Speaking section and another for all other variables in the test. In Specification 3, there are also two factors: one factor is for all variables in the Speaking and Writing sections and another for all variables in the Reading and Listening sections. In each specification above, all factors are hypothesized to be correlated with each other.

In traditional simulation studies, the data is generated by a predefined true model, much as we did in Example 1. In reality, a true model may not exist, or if it does, it may be disturbed or distorted by many other sources. Thus many models can be closely fitted to

[4]TOEFL® is a registered trademark of Educational Testing Service (ETS), which provided the data for this study. This publication is not endorsed or approved by ETS.

such a population. Hence, in this example we treat the sample covariance matrix of the 22 variables from the TOEFL® iBT test data as the population matrix. We do not know its true structure, but whatever its structure, normal, mild and severely nonnormal samples with different sample sizes are generated from this matrix. The rationale behind this research paradigm is that the sample is a representative of the population underlying the TOEFL® iBT test and the samples generated from the sample covariance matrix actually represent something like parametric bootstrap-like samples. Then, as in typical bootstrap analysis, fitting the three specifications in last paragraph to the sample and the obtained bootstrap-like samples mimics the fitting to the unknown population and its many possible samples. However, unlike the bootstrap, we are able to control the distribution of the variables in our samples.

These three specifications have 206, 208, and 208 degrees of freedoms, respectively, and their population RMSEAs are .061, 0.067 and 0.072 respectively. By the widely-used cutoff values for the population RMSEA, they all have some mild misspecifications. However, the differences among these three true RMSEAs are minor. Presumably, the sample RMSEAs (or other indices mentioned before) also will imply minor misspecification, and model uncertainty will result if there is not a strong substantive preference for a particular parameterization.

Clearly, Specifications 2 and 3 are nested in Specification 1. We present the F_{12} and R_{12} values of each nested pair in Table B. Although three specifications in terms of close fit are not very distinguishable, in terms of F_{12} one could choose between models using our cutoff value $\delta_{C,df_2,0.03} = .1872$ in this case for model close match. Using that cutoff, the difference between Specification 1 vs. 2 can be considered as minor while the difference between Specification 1 vs. 3 is not ignorable. Although the Specification 1 will have a better fit due to more parameters, the ignorable difference between Specifications 1 and 2 makes Specification 2 a good candidate to replace Specification 1, while the nonignorable difference between Specification 1 and 3 would propose a rejection of Specification 3. Like Example 1, we now examine how our proposed statistics would evaluate two nested pairs in terms of type I error when confronted with a sample from this population. The simulation results regarding these statistics are presented in Tables 4A-5C.

One surprising result in Tables 4A-5C is that, in contrast to Example 1, T_{ML_12} comparing two nested pairs now performs poorly under both normal and nonnormal conditions. One possible reason may be a violation of the population drift assumption in this example. As before, T_1 and T_2 perform poorly most of the time in all six tables. When the data is normal or mildly nonnormal, T_3, T_4, T_5 and T_6 have similar rejection patterns that are close to the target one, even when $n = 150$. However, when the data is severely nonnormal, T_3 still performs very well at all sample sizes while T_4 overaccepts the null hypothesis exclusively. The performance of T_5 and T_6 under the severe nonnormality is somewhere between those of T_3 and T_4. They are better than T_4 but have some general tendency to overaccept.

So far, we set $\delta_C = F_{12}$ in all statistics for the specification pairs. As stated, in the next step we set δ_C in all statistics to the cutoff value $\delta_{C,df_2,0.03}$ for two specification pairs in this example and look at their acceptance or rejection performance. Let $T_{ML\_12,0.03}$, $T_{1,0.03}$, $T_{2,0.03}$, $T_{3,0.03}$, $T_{4,0.03}$, $T_{5,0.03}$, and $T_{6,0.03}$ denote the corresponding test statistics when δ_C is set to $\delta_{C,df_2,0.03}$. Ideally, we expect the close match hypothesis to be always accepted or rejected, i.e., hardly ever rejected or accepted, depending on if the F_{12} value of the specification pair is less than or greater than $\delta_{C,df_2,0.03}$. The rejection rates of these statistics in the simulation are presented in Tables 6A-6C (Specification 1 vs. 2) and Tables 7A-7C (Specification 1 vs. 3).

The results, shown in Tables 6A-7C, basically match our expectation. Overall, all statistics intend to accept Specification 1 vs. 2 completely, while rejecting Specification 1 vs. 3 completely as n increases. Compared to the other six statistics, $T_{ML\_12,0.03}$ performs poorly for acceptance in Tables 6A-6C, while it does better for rejection in Tables 7A-7C. For Specification 1 vs. 2, $T_{1,0.03}$ and $T_{2,0.03}$ perform better than $T_{3,0.03}$, $T_{4,0.03}$, $T_{5,0.03}$ and $T_{6,0.03}$ in general. But $T_{3,0.03}$, $T_{4,0.03}$, $T_{5,0.03}$ and $T_{6,0.03}$ have more power to reject than $T_{1,0.03}$ and $T_{2,0.03}$, as expected when comparing Specification 1 vs. 3, although they reach rejection rates above 90% only in the normal condition with $n = 1000$.

7. Discussion

Our close match approach and proposed cutoff value and statistics provides a new ap-

proach to model comparison. We believe that the methods provide an additional tool for evaluating a preferred model which may be rejected by a traditional exact match based test. In addition to avoiding limitations of the traditional exact match approach, they provide a new alternative in SEM to such common model comparison methods as AIC, BIC, χ^2/df ratio and the difference between fit indices. Moreover, our idea of equi-discrepancy lines and selection of the appropriate cutoff values for close match is also applicable to some other theoretical issues such as power analysis (see MacCallum et al., 2006).

It is well known that simulation studies always have their limitations, and our study is no exception. Based on our limited simulation results in two examples we found that T_3, T_5 and T_6 perform well in terms of type I error rates across different data conditions when n is as large as 150. As to the power to reject, the simulation results are consistent with Corollaries 3, 5 and 6, although sometimes a large sample size is needed to achieve complete rejection (e.g., Tables 7A-7C). Our simulation results also show that our statistics have a contrasting convergence performance on the lower and upper tails. Athough a good convergence on the upper tail can justify our statistics for the close match testing purpose, a poor convergence on the other tail may limit the application of our statistics to confidence interval and power analysis especially when the sample size is small. In addition, for our statistics and simulation, $\text{tr}(\mathbf{U}_1\mathbf{\Gamma})$ and $\text{tr}(\mathbf{U}_2\mathbf{\Gamma})$ instead of the more general terms d_1 and d_2 are used as the asymptotic bias. Although our examples show a desirable performance with these bias terms, some further studies may need to check the performance of these statistics with more general bias terms especially when the population drift assumption is violated.

One important point we want to emphasize again is that $F_{12} = 0$, i.e., model exact match, must be rejected in order to appropriately use T_3, T_5 and T_6 for comparing specification pairs. Since in practice one does not know exactly whether this requirement is satisfied or not, it makes sense that in a specific study one should first conduct some evaluation of the exact match hypothesisis. Of course there are many possible tests for evaluating exact match, including the NTLR test, LM test or Wald test under normality, or an asymptotically distribution free test such as the Satorra-Bentler scaled difference test (Satorra, 2000; Satorra & Bentler, 2001) or generalized score or Wald tests (Satorra, 1989). Thus we propose to use

a sequential two-stage procedure for overall nested model comparison: accept the restricted model if it satisfies an exact match test; or, accept the model if it is rejected by the exact match test but still satisfies one of the close match tests such as $T_{3,0.03}$, $T_{5,0.03}$ and $T_{6,0.03}$.

One potential problem of the two-stage procedure above is its significance level during overall nested model comparison. Notice that $H_0 : F_{12} \leq \delta_C$ is a composite of H_0^E and $H_0 - H_0^E$. Let T_E denote some reliable exact match test statistic such as the Satorra-Bentler scaled difference test, and let T_C denote some reliable close match test statistic such as $T_{3,0.03}$, $T_{5,0.03}$ and $T_{6,0.03}$. Further, let $A \equiv \{ T_E > \chi^2_{df_{12},\alpha}\}$ and $B \equiv \{ T_C > \lambda_\alpha\}$. Then Pr[reject $H_0|H_0$]=Pr$[A \cap B|H_0] \leq$max$\{$ Pr$(A \cap B|H_0^E)$, Pr$(A \cap B|H_0 - H_0^E)\} \leq$max$\{$ Pr$(A|H_0^E)$, Pr$(B|H_0 - H_0^E)\}$. Let α_E and α_C be the asymptotic significance levels of T_E and T_C respectively, then Pr$(A|H_0^E) \to \alpha_E$ and Pr$(B|H_0 - H_0^E) \to \alpha_C$. So the significance level of the two-stage strategy is asymptotically bounded above by the maximum of the asymptotic significance levels α_E and α_C.

Our theory is based on likelihood ratio principles. The asymptotic bias terms tr$(\mathbf{U}_1\mathbf{\Gamma})$ and tr$(\mathbf{U}_2\mathbf{\Gamma})$, which are widely used in the Satorra-Bentler procedure and our test statistics in this article, are special cases of more general terms d_1 and d_2 based on the likelihood ratio principle (see Li & Bentler, 2006). Given this result and the generality of the likelihood ratio, it seems that our close match test statistics, and hence the two-stage procedure of nested model comparison, may be extendable to a wide variety of situations in addition to our two illustrated examples. Clearly, this would tremendously increase the scope of application of the proposed methodology.

References

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I-C. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge, England: Cambridge University Press.

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

Bentler, P. M. (2006). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software.

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

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106*(2), 315-317.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and statistical Psychology, 37*, 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-62). Newbury Park, CA: Sage.

Browne, M. W., & du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, 27, 269-300.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research, 25,* 115-136.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. (2003). The finite sampling properties of the RMSEA: Point estimates and confidence intervals. *Sociological Methods and Research, 32,* 208-252.

De Leeuw, J. (1988). Model selection in multinomial experiments. In T. K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (pp. 118-138). Berlin: Springer.

Dijkstra, T. K. (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *British Journal of Statistical and Mathematical Psychology, 45,* 289-309.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Rep. No. RR-32; ETS RR-89-42). Princeton, NJ: ETS.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago, IL: National Educational Resources.

Kano, Y. (1986). Conditions on consistency of estimators in covariance structure model. *Journal of the Japan Statistical Society, 16,* 75-80.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach.* Cambridge, England: Cambridge University Press

Lee, S. Y. (1985). Analysis of covariance and correlation structure. *Computational Statistics and Data Analysis, 2,* 279-295.

Lee, S. Y. & Bentler, P. M. (1980). Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal, 14,* 121-136.

Li, L. & Bentler, P. M. (2006). Robust Statistical Tests for Evaluating the Hypothesis of Close Fit of Misspecified Mean and Covariance Structural Models. UCLA Statistics Preprint #494.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113-139.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing Differences between Nested Covariance Structure Models: Power Analysis and Null Hypotheses. *Psychological Methods, 11*, 19-35.

MacCallum, R., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics.* New York: Wiley.

Ogasawara, H. (in press). Approximations to the distributions of fit indexes under fixed alternatives in normal and nonnormal samples. *Psychometrika.*

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54,* 131-151.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In Heijmans, R.D.H., Pollock, D.S.G. & Satorra, A. (eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp.233-247). London: Kluwer Academic Publishers.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association,* 308-313.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507-514.

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

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73-83.

Shapiro, A. (1984). A note on the consistency of estimators in the analysis of moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 84-88.

Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrika, 72,* 133-140.

Sörbom, D. (1989). Model modification. *Psychometrika, 54,* 371-384.

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors.* Paper presented at the annual meeting of the Psychonomic Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253-264.

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups* (TOEFL Research Rep. No. RR-06; ETS RR-80-32). Princeton, NJ: ETS.

Tucker, L. R, & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica, 57,* 307-333.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92,* 767-774.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51,* 289-309.

Yuan, K.-H., & Bentler, P. M. (1999). F-tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 24,* 225-243.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64,* 737-757.

Yuan, K.-H., Hayashi, K., & Bentler, P. M. (2007). Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypothesis. *Journal of*

Multivariate Analysis, 98, 1262-1282.

Yuan, K.-H., & Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis, 65,* 245-260.

Yuan, K.-H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika, 31,* 67-90.

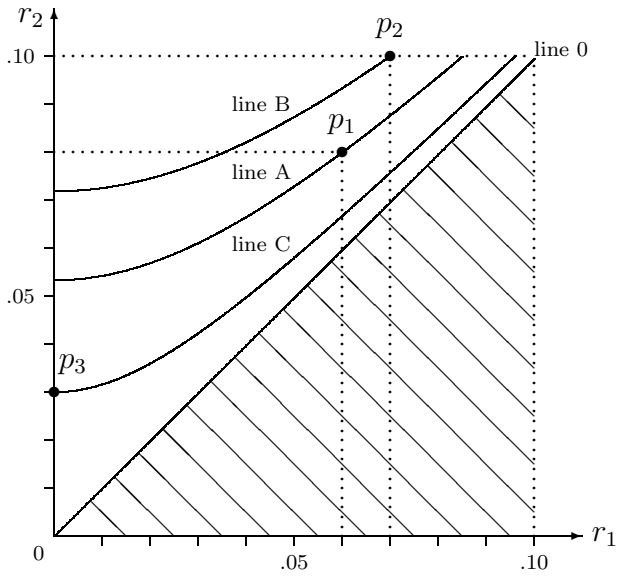

Figure 1. The equi-discrepancy lines with $df_1 = 85$ and $df_2 = 86$

Table A. F_{12} and R_{12} (in parenthesis) by specification pairs in the example by Curran et al. (2003)

Specification of restricted model	Specification of unrestricted model		
	Specification 1 ($r_{10} = 0.000$)	Specification 2 ($r_{10} = 0.021$)	Specification 3 ($r_{10} = 0.031$)
Specification 2 ($r_{20} = 0.021$)	.0396(.199)	-	-
Specification 3 ($r_{20} = 0.031$)	.0828(.288)	.0431(.208)	-
Specification 4 ($r_{20} = 0.040$)	.1408(.375)	.1012(.318)	.0581(.241)

Table B. F_{12} and R_{12} (in parenthesis) by specification pairs in Example 2

Specification of restricted model	Specification of unrestricted model
	Specification 1 ($r_{10} = 0.061$)
Specification 2 ($r_{20} = 0.067$)	.1598(.400)
Specification 3 ($r_{20} = 0.072$)	.2933(.542)

Figure 2. QQ plots of T_1 to T_6 against $N(0,1)$
for Misspecification 1 vs. 2 under the normal condition with $n = 150$

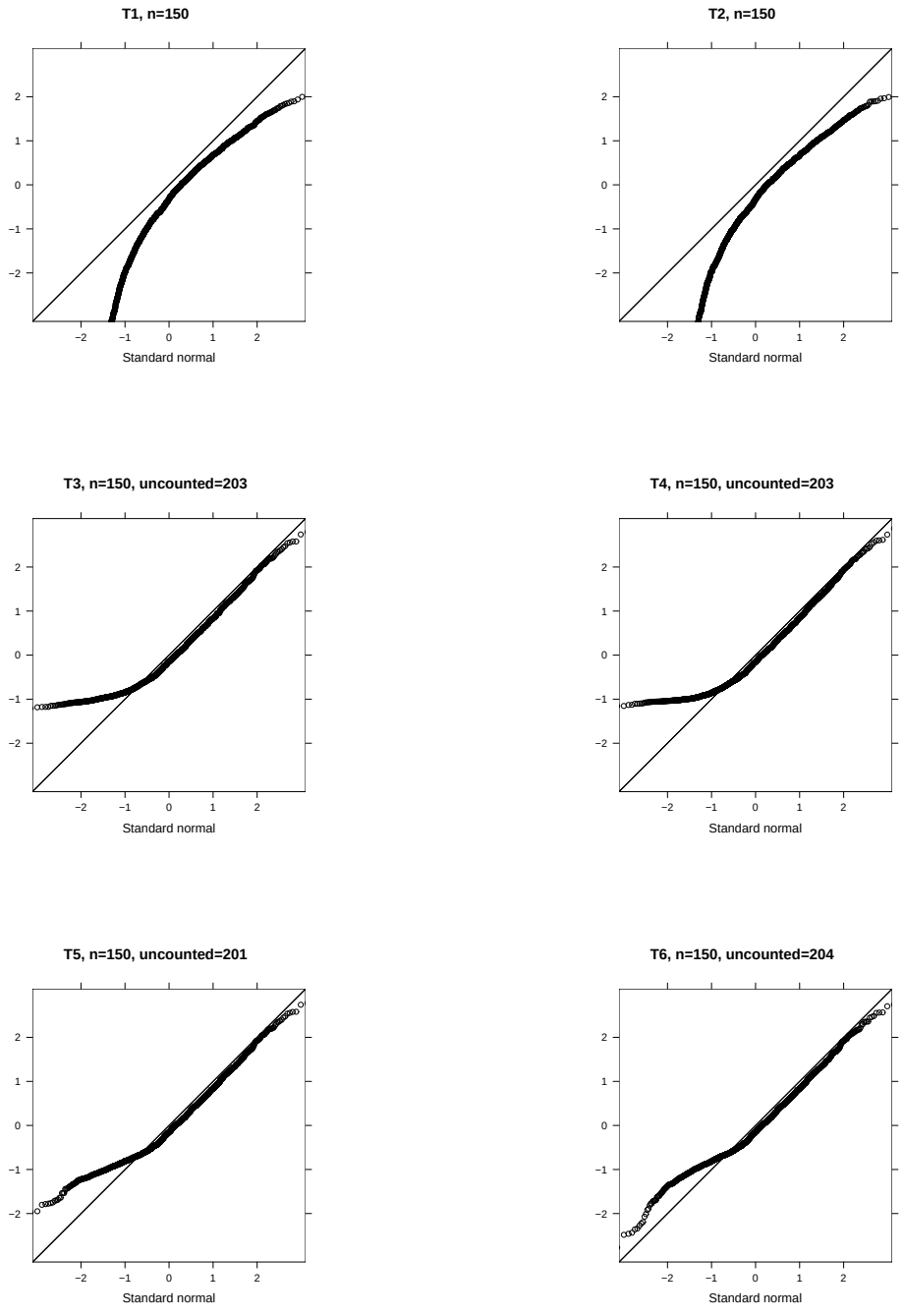

Figure 3. QQ plots of T_1 to T_6 against $N(0,1)$
for Misspecification 1 vs. 2 under the severe nonnormal condition with $n = 150$

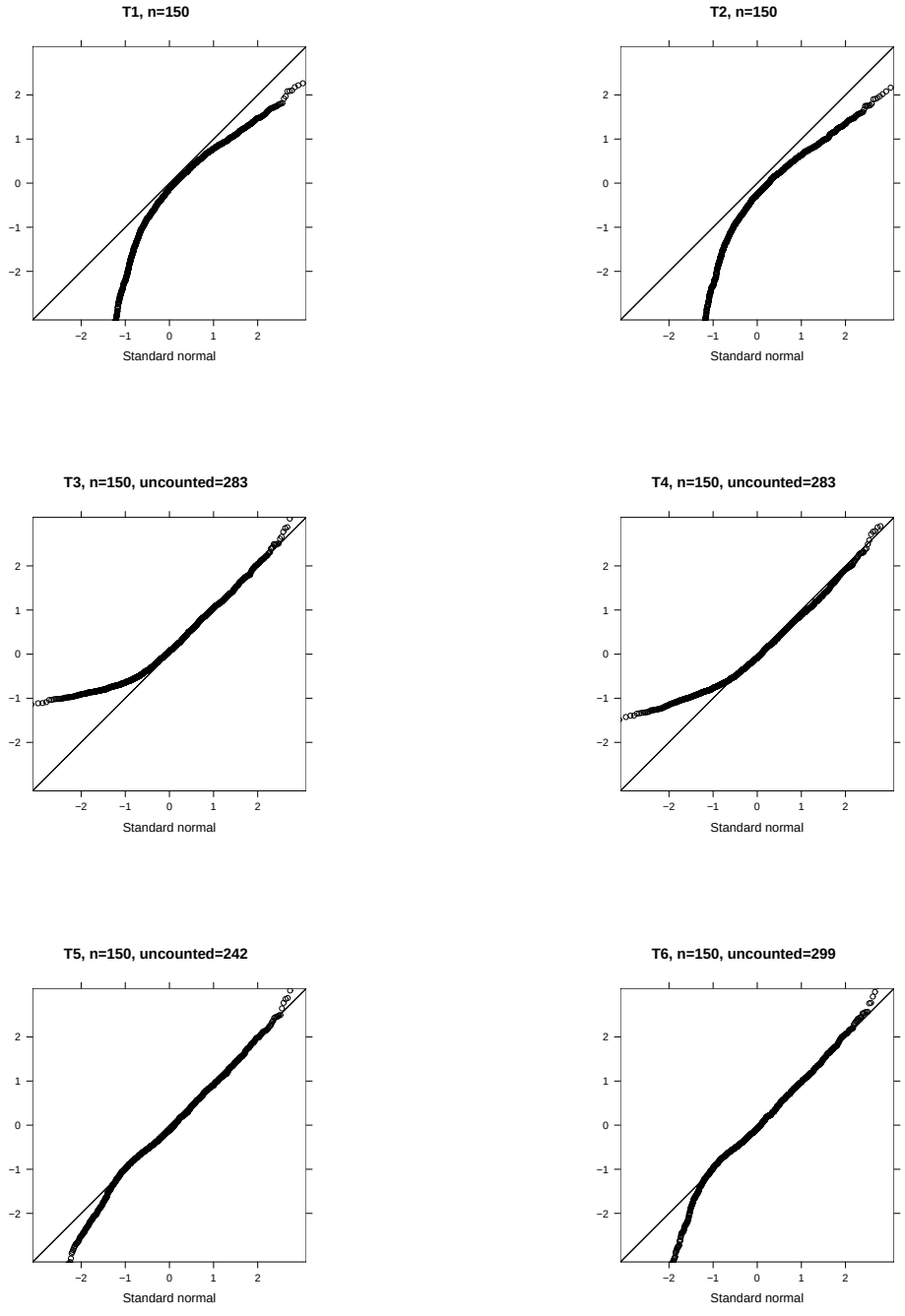

Table 1A. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 1, normal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.051	0.061	0.051	0.061
T_1	0.011	0.021	0.022	0.026
T_2	0.012	0.021	0.022	0.028
T_3	0.040(203)	0.045(23)	0.037(1)	0.045(0)
T_4	0.041(203)	0.045(23)	0.037(1)	0.045(0)
T_5	0.040(201)	0.044(22)	0.037(1)	0.045(0)
T_6	0.038(204)	0.045(22)	0.037(1)	0.045(0)

Table 1B. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 1, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.072	0.074	0.077	0.073
T_1	0.011	0.016	0.017	0.021
T_2	0.009	0.013	0.013	0.021
T_3	0.042(220)	0.042(29)	0.037(5)	0.037(0)
T_4	0.038(220)	0.037(29)	0.033(5)	0.035(0)
T_5	0.040(202)	0.039(29)	0.034(3)	0.035(0)
T_6	0.042(223)	0.041(31)	0.036(3)	0.036(0)

Table 1C. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 1, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.125	0.112	0.124	0.115
T_1	0.013	0.013	0.014	0.019
T_2	0.008	0.008	0.011	0.013
T_3	0.054(283)	0.036(77)	0.044(8)	0.043(0)
T_4	0.038(283)	0.028(77)	0.036(8)	0.037(0)
T_5	0.044(242)	0.033(59)	0.039(2)	0.038(0)
T_6	0.051(299)	0.036(76)	0.041(6)	0.039(0)

Table 2A. Rejection rate of different statistics with $\alpha = .05$
for Specification 2 vs. 3 in Example 1, normal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.060	0.053	0.054	0.059
T_1	0.013	0.015	0.021	0.029
T_2	0.013	0.016	0.022	0.029
T_3	0.045(135)	0.038(21)	0.040(1)	0.045(0)
T_4	0.047(135)	0.037(21)	0.040(1)	0.045(0)
T_5	0.046(134)	0.038(22)	0.040(1)	0.045(0)
T_6	0.043(134)	0.038(22)	0.040(1)	0.045(0)

Table 2B. Rejection rate of different statistics with $\alpha = .05$
for Specification 2 vs. 3 in Example 1, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.082	0.085	0.083	0.076
T_1	0.008	0.018	0.026	0.029
T_2	0.004	0.017	0.025	0.025
T_3	0.050(170)	0.047(32)	0.052(4)	0.048(0)
T_4	0.042(170)	0.043(32)	0.048(4)	0.045(0)
T_5	0.047(165)	0.044(30)	0.051(4)	0.046(0)
T_6	0.050(176)	0.047(31)	0.052(4)	0.048(0)

Table 2C. Rejection rate of different statistics with $\alpha = .05$
for Specification 2 vs. 3 in Example 1, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.118	0.113	0.11	0.118
T_1	0.012	0.012	0.015	0.026
T_2	0.006	0.006	0.009	0.020
T_3	0.053(271)	0.041(64)	0.037(13)	0.05(0)
T_4	0.037(271)	0.030(64)	0.032(13)	0.043(0)
T_5	0.045(231)	0.034(47)	0.035(6)	0.046(0)
T_6	0.053(298)	0.040(58)	0.036(10)	0.048(0)

Table 3A. Rejection rate of different statistics with $\alpha = .05$
for Specification 3 vs. 4 in Example 1, normal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.053	0.056	0.054	0.046
T_1	0.016	0.024	0.026	0.026
T_2	0.017	0.025	0.026	0.026
T_3	0.039(67)	0.043(2)	0.042(0)	0.036(0)
T_4	0.039(67)	0.044(2)	0.041(0)	0.037(0)
T_5	0.039(64)	0.043(2)	0.041(0)	0.036(0)
T_6	0.039(66)	0.043(2)	0.041(0)	0.036(0)

Table 3B. Rejection rate of different statistics with $\alpha = .05$
for Specification 3 vs. 4 in Example 1, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.07	0.079	0.073	0.075
T_1	0.018	0.028	0.027	0.030
T_2	0.015	0.021	0.021	0.028
T_3	0.047(83)	0.056(3)	0.045(0)	0.048(0)
T_4	0.040(83)	0.049(3)	0.043(0)	0.041(0)
T_5	0.044(83)	0.051(3)	0.043(0)	0.041(0)
T_6	0.046(88)	0.055(3)	0.045(0)	0.043(0)

Table 3C. Rejection rate of different statistics with $\alpha = .05$
for Specification 3 vs. 4 in Example 1, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.119	0.122	0.111	0.117
T_1	0.020	0.023	0.022	0.026
T_2	0.013	0.013	0.013	0.021
T_3	0.056(156)	0.056(17)	0.048(0)	0.048(0)
T_4	0.039(156)	0.040(17)	0.036(0)	0.039(0)
T_5	0.046(146)	0.050(12)	0.040(0)	0.042(0)
T_6	0.054(173)	0.054(15)	0.044(0)	0.044(0)

Table 4A. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 2, normal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.068	0.069	0.067	0.073
T_1	0.032	0.024	0.025	0.039
T_2	0.031	0.024	0.026	0.037
T_3	0.051(3)	0.044(0)	0.043(0)	0.054(0)
T_4	0.049(3)	0.043(0)	0.041(0)	0.053(0)
T_5	0.050(2)	0.044(0)	0.042(0)	0.053(0)
T_6	0.050(2)	0.043(0)	0.042(0)	0.054(0)

Table 4B. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 2, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.099	0.09	0.1	0.098
T_1	0.025	0.022	0.036	0.036
T_2	0.015	0.015	0.029	0.034
T_3	0.052(3)	0.042(0)	0.054(0)	0.048(0)
T_4	0.041(3)	0.032(0)	0.046(0)	0.044(0)
T_5	0.047(2)	0.035(0)	0.047(0)	0.045(0)
T_6	0.05(2)	0.037(0)	0.049(0)	0.045(0)

Table 4C. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 2 in Example 2, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.146	0.162	0.142	0.149
T_1	0.025	0.027	0.029	0.03
T_2	0.01	0.013	0.017	0.022
T_3	0.054(10)	0.053(0)	0.045(0)	0.045(0)
T_4	0.031(10)	0.035(0)	0.032(0)	0.032(0)
T_5	0.036(9)	0.041(0)	0.035(0)	0.034(0)
T_6	0.042(15)	0.044(0)	0.036(0)	0.035(0)

Table 5A. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 3 in Example 2, normal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.077	0.074	0.069	0.065
T_1	0.029	0.037	0.034	0.040
T_2	0.029	0.037	0.034	0.040
T_3	0.051(0)	0.050(0)	0.045(0)	0.048(0)
T_4	0.048(0)	0.049(0)	0.045(0)	0.048(0)
T_5	0.050(0)	0.049(0)	0.045(0)	0.048(0)
T_6	0.050(0)	0.049(0)	0.045(0)	0.048(0)

Table 5B. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 3 in Example 2, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.095	0.09	0.092	0.101
T_1	0.034	0.029	0.033	0.036
T_2	0.025	0.022	0.026	0.029
T_3	0.052(0)	0.046(0)	0.044(0)	0.048(0)
T_4	0.042(0)	0.035(0)	0.038(0)	0.041(0)
T_5	0.045(0)	0.038(0)	0.039(0)	0.041(0)
T_6	0.047(0)	0.04(0)	0.04(0)	0.042(0)

Table 5C. Rejection rate of different statistics with $\alpha = .05$
for Specification 1 vs. 3 in Example 2, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
T_{ML}	0.14	0.148	0.133	0.153
T_1	0.028	0.037	0.034	0.035
T_2	0.013	0.021	0.018	0.022
T_3	0.052(0)	0.058(0)	0.047(0)	0.046(0)
T_4	0.026(0)	0.033(0)	0.035(0)	0.034(0)
T_5	0.033(0)	0.04(0)	0.036(0)	0.037(0)
T_6	0.038(0)	0.044(0)	0.036(0)	0.037(0)

Table 6A. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 2 in Example 2, normal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.039	0.017	0.013	0.009
$T_{1,0.03}$	0.015	0.007	0.005	0.002
$T_{2,0.03}$	0.013	0.006	0.004	0.002
$T_{3,0.03}$	0.026(3)	0.013(0)	0.008(0)	0.003(0)
$T_{4,0.03}$	0.025(3)	0.013(0)	0.008(0)	0.003(0)
$T_{5,0.03}$	0.025(2)	0.012(0)	0.008(0)	0.003(0)
$T_{6,0.03}$	0.025(2)	0.012(0)	0.008(0)	0.003(0)

Table 6B. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 2 in Example 2, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.056	0.037	0.036	0.021
$T_{1,0.03}$	0.011	0.007	0.009	0.004
$T_{2,0.03}$	0.006	0.005	0.007	0.003
$T_{3,0.03}$	0.026(3)	0.013(0)	0.015(0)	0.006(0)
$T_{4,0.03}$	0.016(3)	0.009(0)	0.012(0)	0.004(0)
$T_{5,0.03}$	0.018(2)	0.009(0)	0.013(0)	0.005(0)
$T_{6,0.03}$	0.021(2)	0.01(0)	0.013(0)	0.005(0)

Table 6C. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 2 in Example 2, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.092	0.093	0.052	0.043
$T_{1,0.03}$	0.014	0.014	0.005	0.006
$T_{2,0.03}$	0.004	0.004	0.004	0.004
$T_{3,0.03}$	0.031(10)	0.023(0)	0.013(0)	0.009(0)
$T_{4,0.03}$	0.015(10)	0.012(0)	0.007(0)	0.006(0)
$T_{5,0.03}$	0.02(9)	0.017(0)	0.007(0)	0.007(0)
$T_{6,0.03}$	0.024(15)	0.018(0)	0.007(0)	0.007(0)

Table 7A. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 3 in Example 2, normal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.389	0.568	0.762	0.946
$T_{1,0.03}$	0.215	0.415	0.617	0.893
$T_{2,0.03}$	0.216	0.411	0.616	0.892
$T_{3,0.03}$	0.312(0)	0.488(0)	0.677(0)	0.915(0)
$T_{4,0.03}$	0.305(0)	0.485(0)	0.675(0)	0.914(0)
$T_{5,0.03}$	0.307(0)	0.486(0)	0.676(0)	0.914(0)
$T_{6,0.03}$	0.309(0)	0.486(0)	0.677(0)	0.914(0)

Table 7B. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 3 in Example 2, mild nonnormal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.415	0.566	0.723	0.918
$T_{1,0.03}$	0.193	0.343	0.538	0.834
$T_{2,0.03}$	0.154	0.311	0.515	0.824
$T_{3,0.03}$	0.283(0)	0.427(0)	0.607(0)	0.858(0)
$T_{4,0.03}$	0.249(0)	0.402(0)	0.589(0)	0.85(0)
$T_{5,0.03}$	0.261(0)	0.41(0)	0.591(0)	0.851(0)
$T_{6,0.03}$	0.269(0)	0.416(0)	0.593(0)	0.852(0)

Table 7C. Rejection rate of $T_{ML,0.03}$, $T_{1,0.03}$ to $T_{6,0.03}$
for Specification 1 vs. 3 in Example 2, severe nonnormal condition, NR=2000

Sample Size	150	300	500	1000
$T_{ML,0.03}$	0.42	0.554	0.692	0.877
$T_{1,0.03}$	0.151	0.267	0.397	0.675
$T_{2,0.03}$	0.088	0.199	0.34	0.644
$T_{3,0.03}$	0.245(0)	0.353(0)	0.489(0)	0.723(0)
$T_{4,0.03}$	0.174(0)	0.296(0)	0.435(0)	0.7(0)
$T_{5,0.03}$	0.201(0)	0.312(0)	0.453(0)	0.704(0)
$T_{6,0.03}$	0.213(0)	0.319(0)	0.46(0)	0.708(0)