心 理 学 报 2008,**40**(1):119 ~ 124
*Acta Psychologica Sinica*

# 检验的临界值:真伪差距多大才能辨别?
## ——评《不同条件下拟合指数的表现及临界值的选择》

温忠麟[1,2] 侯杰泰[3]

([1]华南师范大学心理应用研究中心,广州 510631)
([2] 香港考试及评核局,香港) ([3] 香港中文大学教育学院,香港)

**摘 要** Hu 和 Bentler(1998,1999)通过模拟研究推荐结构方程模型拟合指数临界值后,受到不少批评和质疑。此后有关拟合指数的研究重点不再是推出新的临界值标准。郭庆科等人的文章《不同条件下拟合指数的表现及临界值的选择》,仿照 Hu 和 Bentler 的做法,通过模拟研究推荐新的拟合指数临界值标准。本文旨在揭示这种做法的错误所在。用简单的 $Z$ 检验,说明检验的临界值是不能通过模拟研究确定的。通过将一个特定真模型的众多错误模型分类,说明结构方程分析中真模型与错误模型差距的多样性,无法通过模拟一对真伪模型来代表。讨论了统计检验的本质和确定临界值的逻辑,还谈到应当从哪些角度检验和评价结构方程模型。

**关键词** 结构方程,拟合指数,检验,临界值。
**分类号** B841.2

郭庆科先生等人的文章《不同条件下拟合指数的表现及临界值的选择》[1](以下简称文[1]),主要目的之一是仿照 Hu 和 Bentler 的做法[2,3],试图用统计模拟的方法,找出结构方程模型(structural equation models, SEM)拟合指数(goodness of fit indices)的临界值(包括推出 2 界值策略)。然而,我们不妨检视文章所用的思路,从研究方法的角度宏观审视,就可以发现这种试图通过模拟研究确定拟合指数临界值的做法,从出发点开始就有问题。

在用统计方法分析之前,先看看一个简单、怪诞但与寻找检验的最佳临界值问题很相似的例子,以说明文[1 ~ 3]的方法不能找出所谓最佳临界值。某综合大学有研究者希望用学生的高考数学分数来推测(相当于检验)他们是否在读数学系(相当于真模型),该研究者要定一个临界分数,达到这个分数以上的学生就推测他在读数学系,否则就推测他不是在读数学系。因为不是所有数学高分的人都在读数学系,所以无论将这个临界分数定在哪里,推测时都可能犯错误。该研究者明白,如果将临界分数定得太高(如满分 150 时定在 140 分),许多数学系学生会被错误地推测为不是在读数学系(拒真);如果将临界分数定得太低(如 80 分),许多非数学系学生会被错误地推测为在读数学系(受伪)。于是,他找来数学系和外语系(外语系相当于一个错误模型)全体学生的数学高考分数,然后尝试了许多临界分数,发现将临界值定在 105 分时,两类错误率之和最小,因此将 105 分定为最佳临界分数。有另一位研究者不服气,他不要外语系的学生成绩,而是找来数学系和物理系(物理系相当于另一个错误模型)全体学生的数学高考分数,发现如果用 105 分做临界分数,物理系的学生差不多都被推测到数学系了(即第二类错误率高),他也尝试了许多临界分数,发现将临界分数定在 120 分时,两类错误率之和最小,因此他将 120 分定为最佳临界分数。

上面例子确定最佳临界值的方法,显然是不妥的,因为最后得到的临界分数与选择哪个系来与数学系比较有关。如果把这样得到的临界分数用到其他的综合大学,就更加没有道理。然而,文[1 ~ 3]使用的就是这样的方法,只是复杂的模型分析细节,影响了人们对方法本质的理解。Hu 和 Bentler 的文章[2,3]发表后,温忠麟、侯杰泰和 Marsh 提出质疑,换一个角度部分地重复 Hu 和 Bentler 的模拟研究,来说明他们的临界值标准(包括 2 指数准则)是不妥的[4]。同一组作者不仅通过模拟研究,而且从理论

收稿日期:2007 - 10 - 09
通讯作者:温忠麟, E-mail: wenzl@ scnu. edu. cn

上说明了他们的临界值标准为何不妥[5]。文[5]引入“可以接受的错误模型”(acceptable misspecified model)的概念,说明错误模型与真模型的差距会影响临界值的确定。此后,其他一些研究者也对 Hu 和 Bentler 的临界值标准提出质疑[6~8]。Fan 所参与的两篇文章立场明确,在文献综述中质疑 Hu 和 Bentler 所提出的临界值未足可信[6,7]。他们也指出 Marsh, Hau 和 Wen 的文章[5],已提供更详细分析及评论。虽然 Yuan 认为拟合指数仍有一定存在的价值,但他指出:指数的均值及分布,受样本容量、数据分布及所选取统计量的影响,拟合指数除了反映模型拟合程度外,还反映这些因素。故拟合指数的临界值、置信区间、检验力(即拒绝错误模型的概率),都备受质疑[8]。

可以说,上述多位研究者对 Hu 和 Bentler 的研究方法和推荐的指数临界值标准的批评和质疑,阻止了其流行。要不然,按作者之一 Bentler(流行的 SEM 软件 EQS 的作者)的名气和发文刊物《Psychological Methods》以及《Structural Equation Modeling》的影响(2004 年的影响因子分别是 5.53 和 1.92),他们的临界值标准可能早就成为国际流行的“金标准”(rule of thumb)了。通过 PsychINFO (2000 ~ 2006)检索关键词“confirmatory factor analysis”(验证性因子分析)和“CFA”,在美国心理学会(APA)主办的刊物中有 154 篇论文报告了拟合指数,共有 18 篇采用了 Hu 和 Bentler 推荐的新指数准则,仅占 11.7%。当然,即使再多人使用,也不能说明他们的研究方法和推荐的指数临界值标准就是合适的。

不仅 Hu 和 Bentler 没有对质疑提出反驳,而且后来的研究重心都不是提出新的临界值标准。文[4]虽然提出了卡方准则,但该文的重心是对 Hu 和 Bentler 的临界值标准进行质疑,最后仍然认为应当使用指数的传统临界值。用模拟研究得到的卡方准则也是不妥的,同一组作者后来的文章[5]已不再提卡方准则。Yuan 认为拟合指数的临界值标准是很难确定的。他提出了几个构造拟合指数的新方法,得到的拟合指数比文献中已有的拟合指数要好,还系统地提出了研究和发展拟合指数的一般规则[8]。Fan 和 Sivo 主要是分析各指数的特性,而非寻找最佳临界值[6]。文[1]仍然采用 Hu 和 Bentler 的做法,通过设计一个非常特殊的真模型和错误模型来确定拟合指数的临界值(包括推出 2 界值策略),如果用其结果来指导 SEM 应用中拟合指数临界值的选择,极易造成误导。下面从统计角度详细揭示文[1]所用方法的问题本质。

## 1 拟合指数临界值是不能用模拟研究确定的

先简单介绍一下文[1]的做法:首先,设计一个真模型,用于产生不同条件下的样本数据。共考虑了 2(分布形态)×4(评分等级数)×6(负荷量)×6(样本容量)=288 种条件,用统计软件对每种条件模拟出 200 个样本。然后,设定一个错误模型,用样本数据分别拟合真模型和错误模型。对所考察的拟合指数,预先设定若干临界值,这里以 CFI 和临界值 0.9 为例来说明。如果一个样本拟合真模型而 CFI <0.9,检验结果是拒绝真模型,则犯了第一类错误(拒真),由 200 个样本的拟合结果就可以算出第一类错误率 $\alpha$;如果一个样本拟合错误模型而 CFI ≥ 0.9,检验结果是接受错误模型,则犯了第二类错误(受伪),由 200 个样本的拟合结果就可以算出第二类错误率 $\beta$。这样就可得到临界值为 0.9 时 CFI 的两类错误率之和。如果在候选的若干个临界值中,临界值为 0.95 时两类错误率之和最小,则 0.95 就是 CFI 的最佳临界值。

这样确定的临界值只对所设计的真模型和错误模型而言有效。这是因为,即使真模型不变,换一个错误模型,确定的临界值就可能很不一样了。所以,如果使用类似的做法,不同的研究者设计不同的模型,就会得到不同的结果[1~3,6,7]。为了不纠缠于 SEM 复杂的模型设定等与本质无关的细节,下面用人们熟知的 $Z$ 检验来说明,设计不同的错误模型会得到不同的最佳临界值,因而临界值是不能用模拟研究确定的。

仿照上面的模拟方法,设计真模型为正态分布 $N(70,10^2)$,错误模型为正态分布 $N(68,10^2)$。拟合真模型相当于检验原假设 $H_0:\mu=70$;拟合错误模型相当于检验原假设 $H_0:\mu=68$。不失一般性,这里以单侧(右侧)检验为例。设从真模型 $N(70,10^2)$ 获得容量为 100 的样本,样本均值为 $\bar{X}$,则 $\bar{X}$ ~ $N(70, 1)$。如所知,检验统计量是 $Z=\frac{\bar{X}-\mu_0}{\sigma_0}\sqrt{N}$。检验 $H_0:\mu=70$ 时,检验统计量(此时 $\mu_0=70$, $\sigma_0=10$, $N=100$)记为 $Z_T$,$Z_T=\bar{X}-70$;检验 $H_0:\mu=68$ 时,检验统计量(此时 $\mu_0=68$, $\sigma_0=10$, $N=100$)记为 $Z_F$,$Z_F=\bar{X}-68$。用通常的单侧检验的临界值 1.65(对应的显着性水平是 0.05),检验 $H_0:\mu=70$(拟合真模型)时,如果由样本计算的 $Z_T>1.65$,检

验结果是拒绝真模型,则犯了第一类错误。检验 $H_0:\mu=68$(拟合错误模型)时,如果由样本计算的 $Z_F\leqslant 1.65$,检验结果是接受错误模型,则犯了第二类错误。因为 $\bar{X}\sim N(70,1)$,所以 $Z_T=\bar{X}-70\sim N(0,1)$,而 $Z_F=\bar{X}-68\sim N(2,1)$,两类错误率之和如图 1(a)阴影部分所示。显然,取临界值 1.0,得到的两类错误率之和最小,如图 1(b)阴影部分所示。这样,同时考虑两类错误率之和时,最佳临界值是 1.0,而不是传统的 1.65。因为这里的问题比较简单,可以知道检验统计量的分布,所以不用通过数据模拟就可以从分布密度函数图直观得到最佳的临界值。而在 SEM 中,不知道拟合指数(即统计量)的分布,需要模拟才能粗略地确定最佳临界值。

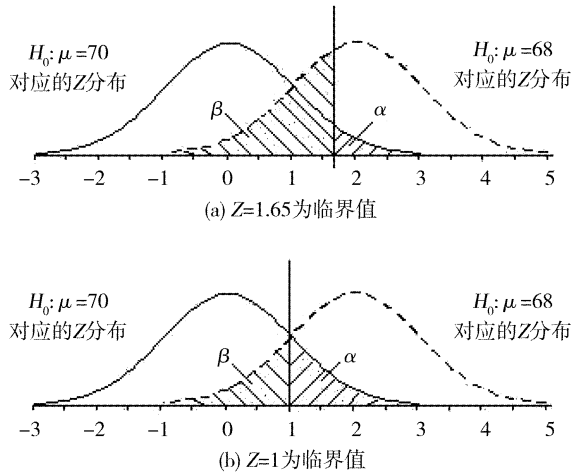


图 1　不同临界值的两类错误率之和

然而,如果设定错误模型为正态分布 $N(66,10^2)$,则最佳的临界值不是 1.0,而是 2.0,见图 2 所示,读者不难按上面的方法推出来。所以,如果同时考虑两类错误率,根本就不存在一个最佳的临界值,通过一、两个错误模型确定的临界值是没有应用价值的。

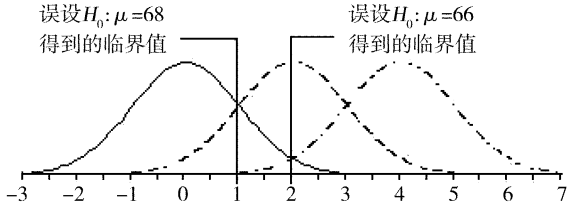


图 2　不同错误模型得到的临界值

上面的例子可以说是模型检验最简单的情形了,尚且不存在检验的最佳临界值,在包含了各种常见统计方法的 SEM 中,更加不可能存在检验的最佳临界值了。如果有人对通常 $Z$ 检验或 $t$ 检验推荐最佳临界值,恐怕没有人会买账。然而,SEM 中复杂的模型往往就把人们的注意力引导到具体的细节上,影响了对事物本质的理解。Hu 和 Bentler 分别用两个真模型经过模拟研究[2,3],推荐了一套指数临界值(包括 2 指数准则),但随后引起不少批评和质疑[4~8]。

可是,文[1]还在通过设计一个非常特殊的真模型和错误模型来确定拟合指数的临界值(包括推出 2 界值策略),设计的真模型包含 3 个相互独立的因子,各有 5 个指标(见图 3)。通常情况下因子之间有或多或少的相关,3 个因子相互独立是一种极端的模型,所以文[1]的真模型不要说在 SEM 中没有代表性,即使在 CFA 中也缺乏代表性。其实,真模型有代表性也没有用,下面会看到,错误模型与真模型的差距才是关键。

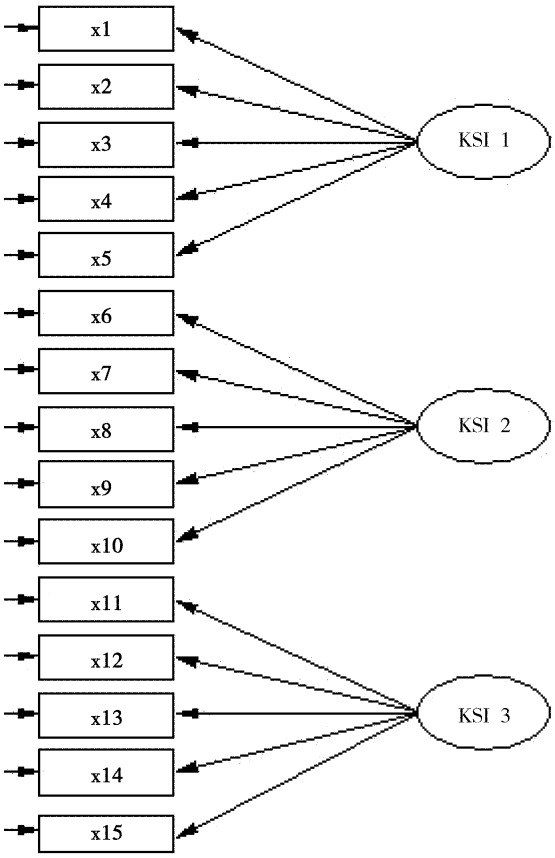


图 3　文[1]设计的真模型

## 2　结构方程模型中错误模型与真模型差距的多样性

上面的 $Z$ 检验例子,真模型与错误模型的差距是非常单纯的,就是真模型正态分布的均值与错误模型正态分布的均值之差。在样本容量确定的情况

下,这个均值之差决定了最佳临界值。然而,在SEM 中,错误模型与真模型差距是复杂多样的。

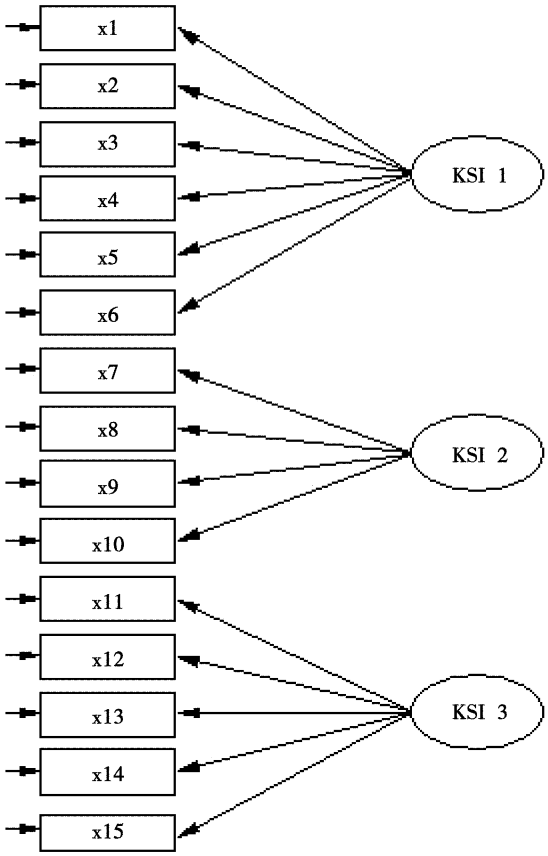


图 4 文[1]设计的错误模型

不要说 CFA(或 SEM)的各种模型,就以文[1]设计的那个真模型为例,足以说明错误模型与真模型差距的多样性。错误模型可以分为两种:第一种是因子结构和真模型的相同,即错误模型和真模型的因子个数相同;第二种是因子结构和真模型的不同,即错误模型和真模型的因子个数不同。第一种错误模型可以进一步分成三类,一类是参数过少(under-parameterized)的错误模型,删除了一条或多条真模型中原有的路径,相当于将一个或多个自由估计的参数固定为零。Hu 和 Bentler 设计的错误模型就是参数过少的模型[2,3]。一类是参数过多(over-parameterized)的错误模型,增加了一条或多条真模型中没有的路径,相当于将一个或多个固定为零的参数自由估计。还有一类是参数误置(wrong-parameterized)的错误模型,如文[1]设计的错误模型,指标 - 因子关系误置就属于这一类(见图4)。第二种错误模型可以分成两类,一类是因子过多,即错误模型的因子比真模型的多。另一类是因子过少,即错误模型的因子比真模型的少(见图5)。无论是因子过多还是过少,肯定是有某些参数误置了。

检验和评价一个模型,可以从几个方面入手,一是看参数估计值,如符号是否合适,参数是否显着等;二是看每个方程的 $R^2$,可以知道负荷是否太低;三是检查残差矩阵;四是看修正指数(modification index, MI);五是看拟合指数。拟合指数是对模型整体拟合好坏的度量,但许多时候通过前面的局部检验就可以发现错误模型了。对上述的第一种错误模型,用拟合指数以外的方法有更高的检验力。

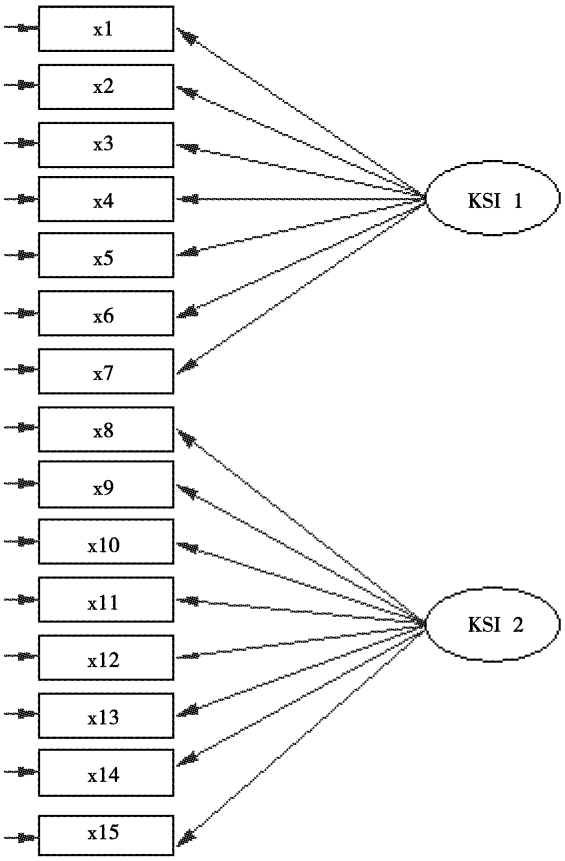


图 5 因子过少的错误模型

先看参数过多的错误模型,因为一个本来没有的路径(即参数值等于零)让其自由估计,结果通常是不显着,所以容易发现参数过多的错误模型。参数过少的错误模型也容易发现,通过查看修正指数,可以找到被忽略的路径,如果增加该路径后卡方值显着减少,可以认为原来拟合的模型参数过少。在参数误置的错误模型中,一个题目本来属于因子 $A$,被归到因子 $B$ 了。该题目被归到了本来不属于的因子 $B$,会表现为参数过多;该题目没有归到本来属于的因子 $A$,会表现为参数过少。所以,参数误置是参数过多和参数过少的叠加,这类错误模型更加容易被发现。如文[1]设计的模型,正态情形且 $N$ =

150 时,如果用文[1]考虑的那些拟合指数做检验,两类错误率之和的最小值在 0.45 上下,最低的是 0.35(见文[1]表 5),但如果用常规做法:先对参数进行 $t$ 检验,如果参数都显着才看拟合指数,并且以传统的临界值(如 CFI = 0.9, LISREL8.5 以上版本大约为 CFI = 0.92),则两类错误率之和不会超过 0.15。还可以查看负荷大小,例如标准化负荷小于 0.3 认为模型需要改进(因为此时对应的指标的信度不足 0.1),这种做法也足以拒绝文[1]设计的错误模型。这说明,对于第一种错误模型,不需要考虑第二类错误率去确定最佳的指数临界值,用传统的做法更好:一方面,通过查看负荷大小、参数的显着性等可以很好地发现错误模型(第二类错误率低);另一方面,使用传统的指数临界值(如 CFI = 0.9)第一类错误率较低。所以综合使用各种方法评价和检验模型,两类错误率之和较低。

对于第二种错误模型,由于错误模型和真模型结构上的差异,通过参数检验、查看负荷大小等可能发现不了,有时候只能指望拟合指数来把关了。即使通过模拟研究确定最佳临界值是可行的,也应当设计第二种错误模型,特别是因子过少的错误模型,因为因子分析的一个主要目的是简化数据结构,所以研究者往往将待检验的理论模型设置得比较简单(即因子过少)。可惜,以往的模拟研究连这方面都没有注意到。

前面 $Z$ 检验的例子已经表明,错误模型与真模型的差距不同,会导致不同的最佳临界值。而 SEM 中错误模型与真模型的差距复杂多样,仅仅通过设计一个真模型和一个错误模型就想得到临界值标准,显然是行不通的,所以这样的研究应当停止了。

## 3　检验的本质和临界值的确定

不要说复杂的 SEM 中的各种拟合指数临界值不能通过所谓的研究来“科学地”确定,就说两组均值差异的 $t$ 检验的临界值,也不能通过研究来确定。通常的做法是不管第二类错误率,只控制第一类错误率。依据所谓的“小概率原理”,将 $\alpha$ 取值为 0.05,就确定了 $t$ 检验的临界值。为什么 $\alpha$ 取值为 0.05,不取 0.04 或 0.06? 原因是在没有统计软件的时代为了方便制作统计表查找临界值。也有许多人干脆就取 2 为 $t$ 检验的临界值,而不管自由度是多少。有了统计软件,检验结果会给出一个显着性概率,现在的习惯做法是:如果显着性概率不超过 0.05, 则认为两组差异显着;如果显着性概率在 0.05 ~ 0.10 之间, 则认为两组差异边缘显着(marginally significant);如果显着性概率超过 0.10, 则认为两组差异不显着。但这只是习惯做法,不是通过研究得出来的,应用工作者心里应当明白这一点。

还是以两组均值差异的 $t$ 检验为例。两个总体的均值要么有差异,要么没有差异。然而,有时候虽然有差异,但根据所得的样本无法把它们的差异分辨出来。因而统计检验能解决的问题不是判断两个总体的均值是否有差异,而是判断差异是否大到可以分辨出来的程度。如果差异在统计上能分辨,就是差异显着;相反,如果不能分辨,未必就没有差异,而是差异不显着。显然,多大的差异才是可以分辨的,在样本容量一定的情况下,取决于临界值的选取。

同样道理,在 SEM 分析中,如果一个模型与真模型不能分辨,则认为该模型是真模型;如果能分辨,则认为该模型是错误模型。在样本容量一定的情况下,错误模型与真模型差距多大才能分辨,取决于使用的拟合指数和临界值的选取。以 CFI 为例,传统上取临界值 0.9(以零模型为基准,所拟合的模型如果“改进”了 90%,可以接受),约定这个 0.9(即 90%)也是源于习惯,和约定 $t$ 检验的 $\alpha$ 取值为 0.05 道理一样。诚然,也可以约定 CFI 的临界值为 0.95,即增加第一类错误率以减少第二类错误率,就像可以约定 $t$ 检验的 $\alpha$ 取值为 0.10 一样。但这个 CFI 的临界值取 0.95 不是“科学地”研究出来的,仍然是人为约定。

在 SEM 分析中,有不少线索可以发现错误模型,如参数估计不显着,负荷太低,残差矩阵有异常的元素,有比较大的修正指数等,所以我们对第二类错误率可以放心一些,关注点仍然是第一类错误率,如果将 CFI 那样的指数临界值定得太高,或将 SRMR 那样的指数临界值定得太低,都会增加第一类错误率。从这个角度看,我们认为传统的临界值没有理由改变。不过,不要将临界值看得太绝对,特别是当拟合指数在临界值附近时,不能让临界值一锤定音。

## 4　结语

除非已知真模型和错误模型,否则不可能找到使两类错误率之和达到最小的检验临界值。不要说 SEM 中的模型检验,就是简单的 $Z$ 检验,也不存在使两类错误率之和达到最小的临界值。所以,企图通过模拟研究来确定拟合指数的临界值,出发点就

是错误的。

在 SEM 分析中,检验和评价一个模型需要从多个方面进行,拟合指数只是其中的一个方面。在发现错误模型方面,检验路径参数、检视负荷或 $R^2$、检视残差矩阵、检视修正指数等等,往往比检验拟合指数有更高的检验力。

通常所用的临界值是根据常识和经验约定俗成的,不是研究出来的。像文[1]那样"研究"出来的两个界值,只是对该设计所用的某一对真模型和错误模型有效,和单个界值一样,不能推广。是可以将拟合指数临界值模糊到一个范围,就像 $t$ 检验时将显着性水平在 0.05 ~ 0.10 之间说成边缘显着一样,但这个模糊的范围仍然是约定的。当拟合指数在临界值附近时,先不要根据拟合指数下明确的结论,应当综合多种方法对模型做出检验和修正,结合模型的理论背景和可解释程度,对最终拟合的模型进行多角度的讨论。

## 参 考 文 献


1 Guo Q, Li F, Wang W, Chen X, Meng Q. Performance of fit indices in different conditions and the selection of cut - off values. Acta Psychologica Sinica, 2008, 40(1): 109 ~ 118
(郭庆科,李芳,王炜丽,陈雪霞,孟庆茂. 不同条件下拟合指数的表现及临界值的选择. 心理学报, 2008, 40(1): 109 ~ 118)

2 Hu L, Bentler P M. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 1998, 3(4): 424 ~ 453

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

4 Wen Z, Hau K T, Marsh H W. Structural equation model testing: Cutoff criteria for goodness of fit indices and Chi-square test (in Chinese). Acta Psychologica Sinica, 2004, 36(2): 186 ~ 194
(温忠麟, 侯杰泰, 马什赫伯特. 结构方程模型检验:拟合指数与卡方准则. 心理学报, 2004, 36(2): 186 ~ 194)

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

6 Fan X, Sivo S A. Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. Structural Equation Modeling, 2005, 12(3): 343 ~ 367

7 Sivo S A, Fan X, Witta E L, Willse J T. The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. The Journal of Experimental Education, 2006, 74(3): 267 ~ 288

8 Yuan K-H. Fit indices versus test statistics. Multivariate Behavioral Research, 2005, 40(1): 115 ~ 148


# Cutoff Values for Testing: How Great the Difference between the True and the False Makes Them Distinguishable?


WEN Zhong-Lin[1,2], HAU Kit-Tai[2]

([1] *Faculty of Educational Science, South China Normal University, Guangzhou* 510631, *China*)

([2] *Research Division, Hong Kong Examinations and Assessment Authority, Hong Kong, China*)

([3] *Faculty of Education, The Chinese University of Hong Kong, Hong Kong, China*)



**Abstract**

Subsequent to Hu and Bentler's (1998, 1999) simulation studies and proposed cutoff criteria for goodness of fit indices in structural equation analyses, several critiques have been published challenging their research design and results. No more new cutoff criteria for fit indices have been proposed since then. However, the recent paper in this journal titled "Performance of fit indices in different conditions and the selection of cut-off values" (in Chinese) imitated Hu and Bentler 's procedures in search for new cutoff values for goodness of fit indices. The purpose of this paper is to explain why this kind of research design is wrong. By using the simple Z-test analogy, we showed that the cutoff values for testing should never be determined through simulation studies. Classifications were proposed for the various misspecified models against a certain true model in structural equation analyses to demonstrate the variety of differences between the true model and the misspecified models. It is obvious that the cutoff values obtained through simulation studies depend on the magnitude of the difference between the true and the misspecified models being chosen, ignoring the variety of the differences involved. The rationales of statistical testing and cutoff value setting were discussed. Guidelines on testing and evaluating a fitted model or alternative models were deliberated.



**Key words** structural equation model, fit indices, testing, cutoff value.