心理学报 2012, Vol. 44, No.12, 1687−1694
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2012.01687

测验同质性系数及其区间估计*

叶宝娟[1] 温忠麟[2]

([1]江西师范大学心理学院, 南昌 330022) ([2]华南师范大学心理应用研究中心, 广州 510631)

摘 要 在决定将多维测验分数合并成测验总分时, 应当考虑测验同质性。如果同质性太低, 合成总分没有什么意义。同质性高低可以用同质性系数来衡量。用来计算同质性系数的模型是近年来受到关注的双因子模型(既有全局因子又有局部因子), 测验的同质性系数定义为测验分数方差中全局因子分数方差所占的比例。本文用 Delta 法推导出计算同质性系数的标准误公式, 进而计算其置信区间。提供了简单的计算同质性系数及其置信区间的程序。用一个例子说明如何估计同质性系数及其置信区间, 通过模拟比较了用 Delta 法和用 Bootstrap 法计算的置信区间, 发现两者差异很小。

关键词 同质性系数; Delta 法; 置信区间

分类号 B841

1 引言

在心理、教育、社会和管理等研究领域中, 经常会碰到多维测验, 即一个测验包含了多个分测验。在决定将多个维度的测验分数合成测验总分时, 应当考虑测验同质性(homogeneity)的高低。同质性是指所有题目都测量了相同的特质, 如果所有题目之间的相关都高, 则同质性高(Revelle & Zinbarg, 2009; 温忠麟, 叶宝娟, 2011)。如果测验同质性高, 合成总分是有意义的; 如果同质性太低, 合成总分没有什么意义, 以合成总分为基础进行的统计分析也就没有什么意义。举例来说, 常见的人格测验都是多维测验, 不同维度之间相关性低, 也即同质性低, 因此合成总分是没有意义的。但是, 许多应用工作者在合成多维测验的总分时, 直接将各个题目的分数相加得到总分, 而没有考虑这样做是否合适, 即没有考虑测验是否同质。有的则报告整份测验的 α 系数, 以为 α 系数高就表明测验是同质的。已有许多研究发现, α 系数高不代表测验是同质的(刘红云, 2008; Revelle & Zinbarg, 2009), 应当使用同质性系数(homogeneity coefficient)衡量测验的同质性。考虑到同质性高低与通常的信度高低有很大不同, 所以本文使用“同质性系数”而不用“同质性信度” (参见温忠麟, 叶宝娟, 2011)。

近年来, 诸多学者提倡用置信区间来报告参数估计结果, 因而信度系数的置信区间也受到重视(例如, Bonett, 2010; Raykov, Dimitrov, & Asparouhov, 2010; Raykov & Penev, 2009;Woods, 2007; Zou, 2007)。报告测验的同质性系数时也一样, 最好同时报告其点估计值和置信区间, 以了解同质性系数估计的误差范围。

目前还未见同质性系数区间估计的方法, 我们不妨借鉴合成信度(composite reliability)的区间估计方法。有三种方法或途径估计单维测验合成信度的置信区间(叶宝娟, 温忠麟, 2011), 包括 Bootstrap 法、Delta 法和直接用统计软件(如 LISREL)输出的标准误进行计算。Bootstrap 法得到的结果是实证结果(Wen, Marsh, & Hau, 2010), 最为可信, 但需要数据模拟技术, 很麻烦, 一般的应用工作者不容易掌握。Delta 法是一种近似计算参数估计标准误的方法(Raykov, 2002; Raykov & Marcoulides, 2011), 可以在 SEM 软件中添加额外参数, 根据结果文件

收稿日期: 2011−11−12

* 国家自然科学基金项目(31271116)、教育部人文社会科学重点研究基地项目(11JJD190005)和教育部人文社会科学研究青年基金项目(12YJC190031)资助。

通讯作者: 温忠麟, E-mail: wenzl@scnu.edu.cn

给出参数值, 经简单计算就可得到标准误, 进而计算置信区间, 比 Bootstrap 法简单。SEM 软件添加额外参数估计合成信度时, 结果文件中会直接给出其标准误, 比 Bootstrap 法和 Delta 法都要简单。叶宝娟和温忠麟(2011)通过模拟研究, 发现 Delta 法与 Bootstrap 法得到的置信区间非常接近, 而用 LISREL 输出的标准误计算的置信区间与 Bootstrap 法得到的置信区间相差很大, 推荐用 Delta 法估计合成信度的置信区间。

能否用 Delta 法估计同质性系数的置信区间呢? 应当如何建模和如何计算呢? 本文将进行这方面的探讨。首先, 简单介绍了测验同质性系数以及用于分析同质性的模型; 接着, 用 Delta 法推导出同质性系数估计的标准误, 进而计算置信区间; 然后, 编写了 LISREL 程序和 Mplus 程序, 用一个例子说明如何用本文推导的公式进行计算; 最后对相关的问题进行了讨论。

2 同质性系数及其计算模型

设一个测验由 p 个题目 $x_1, x_2, \cdots, x_p$ 组成, 测量了一个全局因子 G (general factor)和 n 个局部因子 $F_1, F_2, \cdots, F_n$ (local factor), 全局因子和局部因子都是因子分析中所谓的公共因子。此外, 每个题目 x_i 还有自己的特殊因子 s_i 和误差 e_i, 这样, x_i 有下面的分解式(Revelle & Zinbarg, 2009; Zinbarg, Revelle, & Yovel, 2007; Zinbarg, Revelle, Yovel, & Li, 2005, Zinbarg, Yovel, Revelle, & McDonald, 2006):

$$x_i = a_i G + \sum_{j=1}^{n} \lambda_{ij} F_j + s_i + e_i, \quad i = 1, 2, \cdots, p \qquad (1)$$

其中, a_i 是题目 i 在全局因子 G 上的负荷, λ_{ij} 表示题目 i 在局部因子 F_j 上的负荷, 公共因子、特殊因子与误差三者之间不相关(无论是同一个题目还是不同题目都成立), 局部因子与全局因子不相关, 但当局部因子多于两个时允许部分局部因子之间相关。对于每个题目, 除了全局因子和一个局部因子外, 其余负荷设定为零; 对于每个局部因子(对应于一个维度), 只有部分题目在其上有负荷, 其余题目的负荷设定为零。测验的同质性系数定义为测验分数方差中, 全局因子分数方差所占的比例(温忠麟, 叶宝娟, 2011):

$$\rho_{\text{hom}} = \frac{(\sum_{i=1}^{p} a_i)^2 \operatorname{var}(G)}{\operatorname{var}(\sum_{i=1}^{p} x_i)} \qquad (2)$$

公式(2)定义的同质性系数 ρ_{hom} 在许多文献中称为信度 ω_h (Revelle, & Zinbarg, 2009; Zinbarg et al., 2007; Zinbarg et al., 2005; Zinbarg et al., 2006), 最先是 McDonald (1985)定义的, 但他用的是别的符号(也见 McDonald, 1999)。过去文献上只是将 ω_h 作为信度系数的一种进行研究, 但温忠麟和叶宝娟(2011)为 ω_h 指出了明确的统计背景, 即 ω_h 是同质性系数。

在验证性因子分析中, 特殊因子和误差合在一起作为测量误差, 即 $\delta_i = s_i + e_i$, 并且假设误差之间不相关, 则公式(2)变为:

$$\rho_{\text{hom}} = \frac{(\sum_{i=1}^{p} a_i)^2 \operatorname{var}(G)}{(\sum_{i=1}^{p} a_i)^2 \operatorname{var}(G) + \operatorname{var}(\sum_{i=1}^{p}\sum_{j=1}^{n} \lambda_{ij} F_j) + \sum_{i=1}^{p} \operatorname{var}(\delta_i)} \qquad (3)$$

公式(1)表示的模型就是近年来受到关注的双因子模型(bifactor model), 最早可追溯到 Holzinger 和 Swineford (1937)的研究, Zinbarg 等人(2005)使用了公式(1)的四项分解方式。双因子模型可以检验测验的结构, 研究者通过实证研究发现很多的多维测验是双因子结构, 既有全局因子也有局部因子(例如, Brouwer, Mejer, Weekers, & Baneke, 2008; Ebesutani et al., 2011; Edwards, Cheavens, Heiy, & Cukrowicz, 2011; Joseph, Mary, & Dave, 2011; Patrick, Hicks, Nichol, & Krueger, 2007)。双因子模型可以通过局部因子的负荷判断其作用大小, 也可用于探讨局部因子对效标的独特预测作用(Chen, West, & Sousa, 2006)。显然, 任何一个可以用传统因子分析建模的多维测验, 都可以尝试建立双因子模型进行分析, 看看是否有全局因子, 全局因子分数的方差贡献有多大。换一个角度说就是, 任何一个多维测验, 都可以尝试建立双因子模型计算同质性系数, 只要模型拟合可以接受, 就可以得到一个同质性系数的估计。

Raykov 和 Zinbarg (2011)基于二阶因子模型(其中只有一个二阶因子)推导出二阶因子分数方差所占的比例(也用 ω_h 表示)。不难证明, 任何一个这样的二阶因子模型, 都可以转化为一个双因子模型, 就是说, 二阶因子模型嵌套于双因子模型之中(Chen et al., 2006; Yung, Thissen, & McLeod, 1999)。二阶模型中的二阶因子对应于双因子模型中的全局因子, 因此, Raykov 和 Zinbarg (2011)定义的 ω_h 就是用二阶因子模型推导出的测验同质性系数。

对加入了反向题目(negatively worded item)的单维测验也可以建立双因子模型。平衡使用正向题和反向题的单维测验通常当作单维分析, 但越来越多的研究者发现这样的测验存在项目表述方法效应(method effect), 即由项目表述引起的变异, 应当加以控制(DiStefano & Motl, 2009; Marsh, Scalas, & Nagengast, 2010; Vautier & Pohl, 2009; Ye, 2009)。检验测验是否存在方法效应的一种做法是建立相关特质相关方法(Correlated-Trait Correlated-Method, CTCM)模型, 把项目表述效应看做是方法效应, 通过验证性因子分析将其从特质效应和误差效应中分离出来。这样建立的模型就是一种特殊的双因子模型, 包括全局因子(所测特质 G, 影响全部题目)、两个局部因子(正向题目效应因子 F_1, 反向题目效应因子 F_2)和测量误差。有的测验可能仅包括影响正向题的方法因子, 有的测验可能仅包括影响反向题的方法因子, 有的测验则可能同时包括这两类方法因子。评价这类测验的同质性系数可以了解排除了方法效应和测验误差引起的变异之后, 由所测特质的变异占总变异的比例, 进而评价合成总分是否有意义。当同质性低而测验的合成信度可以接受的时候, 说明必须建立包含方法因子的双因子模型进行进一步的统计分析, 而不能简单地使用测验总分进行统计分析。

3　用 Delta 法计算测验同质性系数的置信区间

公式(3)的点估计值没有提供一个实证研究的样本同质性系数与实际感兴趣的总体同质性系数接近程度的信息, 而置信区间提供了同质性系数的一个可能范围, 可帮助研究者对全局因子变异占观测分数变异的程度做出更可靠的结论。

Delta 法是一种近似计算参数估计的标准误的方法。Browne (1982)用 Delta 法计算协方差的标准误, Ogasawara (1999)用 Delta 法计算(矩阵)相关系数的标准误, Raykov (2002)将 Delta 法引入信度系数的区间估计中, 用来计算信度系数的标准误。后来有许多研究者用 Delta 法估计各种信度系数的置信区间, 例如追踪研究测验信度(Laenen, Alonso, Molenberghs, & Vangeneugden, 2009a, 2009b), 二分测验信度(Raykov et al., 2010), 两水平研究中测验信度(Raykov & Penev, 2009)。应用 Delta 法的前提是可以得到模型参数及其方差和协方差, 进而得到参数的光滑函数(smooth functions)的近似标准误(Raykov & Marcoulides, 2004)。

下面我们用 Delta 法推导计算测验同质性系数的置信区间。对公式(3), 设

$$s=(\sum_{i=1}^{p}a_i)^2\ \mathrm{var}(G),\quad t=\mathrm{var}(\sum_{i=1}^{p}\sum_{j=1}^{n}\lambda_{ij}F_j)+\sum_{i=1}^{p}\mathrm{var}(\delta_i)\ (4)$$

则(3)式变为

$$\rho_{\mathrm{hom}}=f(s,t)=\frac{s}{s+t}\qquad(5)$$

为了对同质性系数应用 Delta 法, 考虑公式(5)的估计,

$$\hat{\rho}_{\mathrm{hom}}=f(\hat{s},\hat{t})=\frac{\hat{s}}{\hat{s}+\hat{t}}\qquad(6)$$

上式的一阶 Taylor 展开式为:

$$\hat{\rho}_{\mathrm{hom}}=f(\hat{s},\hat{t})\approx f(s_0,t_0)+(\hat{s}-s_0)f_s'(s_0,t_0)+(\hat{t}-t_0)f_t'(s_0,t_0)\qquad(7)$$

其中, $\approx$表示近似, s_0,t_0是 s,t 的总体值, $f_s'(s_0,t_0)$ 和 $f_t'(s_0,t_0)$ 表示 $f(s,t)$ 在(s_0,t_0)的偏导数, 分别记为 D_1 和 D_2, 则公式(7)变为:

$$\hat{\rho}_{\mathrm{hom}}\approx f(s_0,t_0)+(\hat{s}-s_0)D_1+(\hat{t}-t_0)D_2\qquad(8)$$

算出偏导数:

$$D_1=\frac{t_0}{(s_0+t_0)^2},\quad D_2=-\frac{s_0}{(s_0+t_0)^2}\qquad(9)$$

对公式(8)的两边同时计算方差, 就得到同质性系数方差的近似计算公式(方便起见使用等号):

$$\mathrm{var}(\hat{\rho}_{\mathrm{hom}})=D_1^2\ \mathrm{var}(\hat{s})+D_2^2\ \mathrm{var}(\hat{t})+2D_1D_2\ \mathrm{cov}(\hat{s},\hat{t})\ (10)$$

从而同质性系数的标准误为:

$$SE(\hat{\rho}_{\mathrm{hom}})=\sqrt{D_1^2\ \mathrm{var}(\hat{s})+D_2^2\ \mathrm{var}(\hat{t})+2D_1D_2\ \mathrm{cov}(\hat{s},\hat{t})}\ (11)$$

实际计算时, D_1 和 D_2 中的 s_0,t_0 是未知的, 分别用它们的估计值代入:

$$D_1=\frac{\hat{t}}{(\hat{s}+\hat{t})^2},\quad D_2=-\frac{\hat{s}}{(\hat{s}+\hat{t})^2}$$

同质性系数的置信度为$1-\alpha$ 的置信区间为:

$$[\,\hat{\rho}_{\mathrm{hom}}-Z_{\alpha/2}\cdot SE(\hat{\rho}_{\mathrm{hom}}),\hat{\rho}_{\mathrm{hom}}+Z_{\alpha/2}\cdot SE(\hat{\rho}_{\mathrm{hom}})\,]\qquad(12)$$

其中, $Z_{\alpha/2}$是标准正态分布的双侧α分位点。公式(12)就是 Delta 法得到的同质性系数的置信区间, 半径为 $Z_{\alpha/2}\cdot SE(\hat{\rho}_{\mathrm{hom}})$, 表示同质性系数估计的误差范围。区间长度越长(即标准误越大), 估计的同质性系数精确度越低, 反之, 精确度越高。

4　用 Delta 法估计同质性系数置信区间示例

用一个例子说明如何用 Delta 法来计算测验同质性系数的置信区间。假设一个测验有 8 个题目测量自信, 其中 4 个题目是正向题, 4 个题目是反向题,

有一个全局因子(即自信 G, 影响全部 8 个题目)、两个局部因子 F_1 和 F_2 (分别影响正向题和反向题), 通常假定影响正向题和反向题的方法因子不相关(Marsh et al., 2010)。测量方程如下:

$x_1 = a_1G + \lambda_{11}F_1 + \delta_1$, …, $x_4 = a_4G + \lambda_{41}F_1 + \delta_4$,

$x_5 = a_5G + \lambda_{52}F_2 + \delta_5$, …, $x_8 = a_8G + \lambda_{82}F_2 + \delta_8$

其中, a_1 表示第 1 题在因子 G 上的负荷, λ_{11} 表示第 1 题在因子 F_1 上的负荷, δ_1 表示第 1 题的误差, 其余符号类推。

如所知, 因子分析中需要通过固定负荷或者固定方差给因子指定测量单位(侯杰泰, 温忠麟, 成子娟, 2004)。虽然从理论上说, 无论是固定负荷还是固定方差, 得到的误差方差以及要计算的同质性系数不会改变(微小的计算误差除外) (温忠麟, 叶宝娟, 2011), 但双因子模型固定方差比固定负荷好。固定每个因子的方差, 因子的单位就已经指定了; 但如果固定负荷, 因为双因子模型中每个指标从属于两个因子(一个全局因子, 一个局部因子), 需要联合两个指标的负荷才能间接得到两个因子的单位(相当于解方程组), 这给模型的收敛性带来不确定因素。这就可以理解, Zinbarg 及其合作者(Zinbarg et al., 2007; Zinbarg et al., 2005, Zinbarg et al., 2006)的做法都是固定全局因子和局部因子的方差为 1。

因此, 本文也使用固定方差的方法指定模型, 即设定 G, F_1 和 F_2 的方差为 1, 由公式(3)得到同质性系数为:

$$\rho_{\text{hom}} = \frac{(\sum_{i=1}^{8} a_i)^2}{(\sum_{i=1}^{8} a_i)^2 + (\sum_{j=1}^{4} \lambda_{j2})^2 + (\sum_{j=5}^{8} \lambda_{j3})^2 + \sum_{j=1}^{8} \theta_j}$$

用 LISREL 8.8 软件求测验的同质性系数及其置信区间所需要的参数的程序见附录 1, 这个程序与普通的验证性因子分析程序差不多, 但增加了几个额外参数。新增加的额外参数不会影响原有参数的估计值和模型的拟合程度。模型的拟合指数为: $\chi^2(12)=8.318$, RMSEA= 0.000, NNFI = 1.001, CFI = 1.000, SRMR = 0.005, 模型拟合很好(温忠麟, 侯杰泰, Marsh, 2004)。

LISREL 的输出结果可以给出公式(3)中的同质性系数的点估计值, 以及公式(11)中的所有参数(但 D_1 和 D_2 需要计算), 见附录 1 的注释。将 LISREL 的输出结果代入公式(11)易求得同质性系数的标准误, 进而计算其置信区间。本例同质性系数的点估计值为 0.734, 用 Delta 法求得的同质性系数的标准误为 0.029, 同质性系数 95%的置信区间为(0.677, 0.791)。如果在 LISREL 中用 Bootstrap 法抽样 1000 次进行计算, 本测验同质性系数的标准误是 0.030, 95%的置信区间为(0.675, 0.793), 与用 Delta 法得到的结果几乎相同, 因为 Bootstrap 法得到的结果是实证结果, 可以当作真值, 因此本例用 Delta 法计算的置信区间结果可靠。

本例用 Delta 法估计的同质性系数的置信区间的长度为 0.114, 估计的同质性系数精确度不算高, 同质性系数的误差范围大。假定同质性系数为 0.7 才可接受, 此测验的同质性系数置信区间的下限 0.677 在 0.7 之下, 上限 0.791 在 0.7 之上, 不能确定测验的同质性系数是否可以接受, 需要做更多的研究来判断测验是否可合成总分。如果仅用点估计的信息做出判断, 因为同质性系数的点估计值 0.734 大于 0.7, 认为测验同质性系数可以接受, 测验可以合成总分, 与区间估计得出的结果并不一致。因此, 很有必要估计和报告同质性系数的置信区间, 以便对测验的质量做一个比较客观的评价。

如果忽略方法因子, 将测验当作单维处理, 此时同质性系数等于合成信度(温忠麟, 叶宝娟, 2011), 此例测验的同质性系数(合成信度)为 0.883, 用 Delta 法估计测验的同质性系数(合成信度)的标准误为 0.006, 95%的置信区间为(0.872, 0.894)。由此可见, 如果不考虑方法因子, 将测验当作单维处理, 会高估测验的同质性系数(高估 0.149), 低估其标准误(低估了 0.023)。通过本例可以看出, 如果测验题目存在方法因子, 而在计算测验的同质性系数时忽略方法因子, 计算的同质性系数会偏高, 置信区间也不准确。

5 讨论

如果将全部局部因子和全局因子引起的变异都当作真分数变异, 计算信度就得到合成信度; 如果仅将全局因子引起的变异当作真分数变异, 计算信度就得到同质性系数(温忠麟, 叶宝娟, 2011)。

对多维的测验分数合成总分时, 应当考虑测验同质性的高低, 同质性高低可以用同质性系数来衡量。同质性系数越大, 测验总分越强地受到所有题目所测量的全局因子的影响, 测验总分能越强地概括到所有题目测量的共同特质上, 即所有题目的共性越多, 因此可以将所有题目的测验分数合成总分。

双因子模型有许多用场, 但本文是将其用来估计同质性系数, 前提是数据适合传统的因子分析和模型拟合良好。虽然运行验证性因子分析程序后通常都可以得到一个同质性系数, 但只有当模型可以接受时, 得到的同质性系数才有意义。因此, 估计同质性系数之前一定要检验模型。在模型不可接受的情况下, 不可用本文介绍的方法确定测验是否可以合成总分, 应当用其他方法作出判断。验证性因子分析的参数估计和检验是建立在渐近理论(asymptotic theory)上的, 因此, 估计同质性系数时样本容量应当比较大(比如 200 以上), 否则估计的结果可能很不精确(置信区间可以反映出来)。和合成信度一样, 同质性系数与因子模型有关, 相同的量表和实测数据, 建立的模型不同得到的同质性系数也可能不同, 因此, 模型中的测量关系要正确指定, 否则, 估计的结果可能不准确。此外, 如果数据不适合传统的因子分析, 比如数据为二分变量数据, 则用公式(2)和(3)计算的同质性系数往往不准确。

测验的同质性系数是未知的总体参数, 可以用样本的同质性系数来估计。同其它参数的点估计一样, 样本的同质性系数会在真值(总体的同质性系数)附近波动。最好用同质性系数的置信区间补充点估计得到的信息, 以此来评价测验同质性系数。如果测验同质性系数的置信区间的下限大于设定的系数值(如 0.6), 测验可以合成总分; 如果同质性系数的置信区间的上限小于设定的系数值, 说明合成总分意义不大; 如果同质性系数的置信区间包含了设定的系数值, 则还不能判断测验是否可以合成总分, 需要获取更多的数据做进一步分析。

Delta 法是计算测验信度系数的置信区间的主要方法, 近年来被用于各种信度系数置信区间的估计(Laenen et al., 2009a, 2009b; Raykov & Penev, 2009, 2010)。用这种方法计算信度系数的置信区间简单而实用。本文推导出用 Delta 法估计测验同质性系数的标准误公式, 进而求得同质性系数的置信区间。用例子说明了如何用 Delta 法计算测验同质性系数置信区间, 并给出了其 LISREL 计算程序, 应用工作者可以直接套用。

使用 Mplus 软件的新版本(如 Mplus6.0), 编写适当的程序, 可以很容易得到 Delta 法的标准误, 并直接输出同质性系数的置信区间。附录 2 给出了用 Mplus 6.11 软件(Muthén & Muthén, 2010)求本文示例中的同质性系数的点估计值、Delta 法的标准误及其置信区间的程序。在程序中 OUTPUT 部分添加 CINTERVAL 命令可以直接得到同质性系数的置信区间。在计算测验同质性系数的置信区间时, 使用 Mplus 软件的读者可以套用附录 2 的程序进行计算。使用 LISREL 软件的读者可以套用附录 1 给出的 LISREL 程序, 根据其结果输出, 再按公式(11)计算标准误, 进而求得置信区间。

当使用测验总分进行统计分析的时候, 建议先报告同质性系数及其置信区间, 让读者了解统计结果的可信程度。当同质性系数不可接受的时候, 将题目得分相加成测验总分没有意义, 基于测验总分的统计也就没有什么意义。不过, 只要测验的合成信度可以接受, 还是可以继续进行结构方程建模分析, 只是不宜使用测验总分进行显变量分析。

参 考 文 献

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*, 368–385.

Brouwer, D., Mejer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the dispositional hope scale. *Psychological Assessment, 20*, 310–315.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in Applied Multivariate Analysis* (pp. 77–141). Cambridge, UK: Cambridge University Press.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189−225.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem scale. *Personality and Individual Differences, 46*, 309–313.

Ebesutani, C., Smith, A., Bernstein, A., Chorpita, B. F., Higa-McMillan, C., & Nakamura, B. (2011). A bifactor model of negative affectivity: Fear and distress components among younger and older youth. *Psychological Assessment, 23*, 679–691.

Edwards, M. C., Cheavens, J. S., Heiy, J. E., & Cukrowicz, K. C. (2011). A reexamination of the factor structure of the center for epidemiologic studies depression scale: Is a one-factor model plausible? *Psychological Assessment, 22*, 711–715.

Hau, K. T., Wen, Z. L., & Cheng, Z. J. (2004). *Structural equation model and its applications*. Beijing: Educational Science Publishing House.

[侯杰泰, 温忠麟, 成子娟. (2004). *结构方程模型及其应用*. 北京: 教育科学出版社.]

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54.

Joseph, B., Mary, P., & Dave, H. (2011). Investigating early literacy numeracy: Exploring the utility of the bifactor model. *School Psychology Quarterly, 26*, 97–107.

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009a). A family of measures to evaluate scale reliability in a longitudinal setting. *Journal of the Royal Statistical Society, 172*, 237–253.

Laenen, A., Alonso, A., Molenberghs, G., & Vangenugden, T. (2009b). Reliability of a longitudinal sequence of scale ratings. *Psychometrika, 74*, 49–64.

Liu, H. Y. (2008). Alpha coefficient and congeneric test. *Psychological Science, 31*, 185–188.

[刘红云. (2008). α系数与测验的同质性. *心理科学, 31*, 185–188.]

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*, 366–381.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

Ogasawara, H. (1999). Standard errors for matrix correlations. *Multivariate Behavioral Research, 34*, 103−122.

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, P. J. (2007). A bifactor approach to modeling the structure of the Psychopathy Checklist-Revised. *Journal of Personality Disorders, 21*, 118−141.

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37*, 89–103.

Raykov, T., Dimitrov, D, M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17*, 265–289.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling, 11*, 621–637.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Taylor Francis Group, LIC.

Raykov, T., & Penev, S. (2009). Estimation of maximal reliability for multiple-component instruments in multilevel designs. *British Journal of Mathematical and Statistical Psychology, 62*, 129–142.

Raykov, T., & Zinbarg, R. E. (2011). Proportion of general factor variance in a hierarchical multiple-component measuring instrument: A note on a confidence interval estimation procedure. *British Journal of Mathematical and Statistical Psychology, 64*, 193–207.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154.

Vautier, S., & Pohl, S. (2009). Do balanced scales assess bipolar constructs? The case of the STAI scales. *Psychological Assessment, 21*, 187–193.

Wen, Z. L., Hau, K. T., & Marsh, H. W. (2004). Structural equation model testing: Cutoff criteria for goodness of fit indices and chi-square test. *Acta Psychologica Sinica, 36*, 186–194.

[温忠麟, 侯杰泰, Marsh, H. W. (2004). 结构方程模型检验: 拟合指数与卡方准则. *心理学报, 36*, 186–194.]

Wen, Z. L., Marsh, H. W., & Hau, K. T. (2010). Structural equation models of latent interactions: An appropriate standardized solution and its scale-free properties. *Structural Equation Modeling: A Multidisciplinary Journal, 17*, 1–22.

Wen, Z. L., & Ye, B. J. (2011). Evaluating test reliability: From coefficient alpha to internal consistency reliability. *Acta Psychologica Sinica, 43*, 821–829.

[温忠麟, 叶宝娟. (2011). 测验信度估计: 从α系数到内部一致性信度. *心理学报, 43*, 821–829.]

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods, 12*, 185–204.

Ye, B. J., & Wen, Z. L. (2011). A comparison of three confidence intervals of composite reliability of a unidimensional test. *Acta Psychologica Sinica, 43*, 453–461.

[叶宝娟, 温忠麟. (2011). 单维测验合成信度三种区间估计的比较. *心理学报, 43*, 453–461.]

Ye, S. Q. (2009). Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences, 46*, 197–201.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128.

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ω_h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement, 31*, 135–157.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald'sω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement, 30*, 121–144.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*, 399–413.

附录 1：用 Delta 法求测验同质性系数置信区间的 LISREL 程序[1]

```
DA NI=8 NO=1000 MA=CM
CM SY
1.0627
0.6438  1.0097
```

[1]程序所用的软件 LISREL8.8 可以识别公式中的括号，使用较早版本时，需要将 PAR (1)与 PAR (2)的平方展开，并逐一相加。

```
0.6590  0.6553  1.1127
0.6619  0.6453  0.6606  1.0563
0.4014  0.3715  0.3853  0.3869  1.0431
0.4346  0.3982  0.4254  0.4354  0.8037  1.1356
0.4362  0.4068  0.4310  0.4144  0.8094  0.8381  1.1021
0.4333  0.3735  0.3980  0.4218  0.8129  0.8329  0.8278  1.0938
MO NX=8 NK=3 LX=FU,FR TD=DI,FR PH=SY,FI AP=4
PA LX
4(1 1 0)
4(1 0 1)
VA 1 PH (1 1) PH (2 2) PH (3 3)
CO PAR (1)=(LX (1 1)+LX (2 1)+LX (3 1)+LX (4 1)+LX (5 1)+ LX (6 1)+LX (7 1)+LX (8 1))^2
! PAR (1) 等于公式(4)中的 s 的估计值
CO PAR (2)=TD (1 1)+TD (2 2)+TD (3 3)+TD (4 4)+TD (5 5)+TD (6 6)+TD (7 7)+TD (8 8)+
(LX (1 2)+LX (2 2)+LX (3 2)+LX (4 2))^2+(LX (5 3)+LX (6 3)+LX (7 3)+LX (8 3))^2
! PAR (2)等于公式(4)中的 t 的估计值
CO PAR (3)=PAR (1)+PAR (2)
CO PAR (4)=PAR (1)*PAR (3)^-1
! PAR (4)等于同质性系数点估计值
OU ALL ND=3
```

注释：同质性系数的点估计值，对应于 LISREL 的输出结果“LISREL Estimates”部分的“ADDITIONAL PARAMETERS”中的 AP (4)，其值为 0.734。s,t 的估计值对应于“ADDITIONAL PARAMETERS”中的 AP (1)与 AP (2)，分别为 28.922 和 10.504。var(s)、var(t)、cov(s,t)估计值对应于 “Covariance Matrix of Parameter Estimates”中的 AP (1)与 AP (2)的方差及协方差，分别为 4.086, 1.148 和−1.061。

附录 2：用 Delta 法求测验同质性系数置信区间的 Mplus 程序

```
DATA: FILE IS p.dat;
VARIABLE: NAMES ARE y1-y8;
MODEL: f1 BY y1-y8*(p1-p8);
        f2 BY y1-y4*(p9-p12);
        f3 BY y5-y8*(p13-p16);
        y1-y8 (a1-a8);
        f1@1;
        f2@1;
        f3@1;
        f1 with f2 @0;
        f1 with f3 @0;
        f2 with f3 @0;
MODE CONSTRAINT:
        new (H1-H4);
        H1=(p1+p2+p3+p4+p5+p6+p7+p8)**2;
         !H1 等于公式(4)中的 s 的估计值
        H2=a1+a2+a3+a4+a5+a6+a7+a8+(p9+p10+p11+p12)**2+(p13+p14+p15+p16)**2;
         !H2 等于公式(4)中的 t 的估计值
        H3=H1+H2;
```

```
    H4=H1/H3;
     !H4 等于同质性系数点估计值
OUTPUT:
    CINTERVAL;
```

注释：同质性系数的点估计值及用 Delta 法计算的同质性系数的标准误，对应于 Mplus 输出结果中的“MODEL RESULTS”部分中的“New/Additional Parameters”H4 的参数估计值及其标准误，其值为 0.734 和 0.032。同质性系数的 95%置信区间的下限和上限，对应于 Mplus 输出结果中的“CONFIDENCE INTERVALS OF MODEL RESULTS”部分中的“New/Additional Parameters”H4 的“Lower 2.5%”和“Upper 2.5%”的值，其值为 0.672 和 0.795。

Estimating Homogeneity Coefficient and Its Confidence Interval

YE Baojuan[1]; WEN Zhonglin[2]

([1] *School of Psychology*, *Jiangxi Normal University*, *Nanchang* 330022, *China*)

([2]*Center for Studies of Psychological Application*, *South China Normal University*, *Guangzhou* 510631, *China*)

Abstract

Multidimensional tests are frequently applied to the studies of psychology, education, society and management. Before aggregating all item scores to form a composite score of a multidimensional test, we should consider the homogeneity of the test. Homogeneity coefficient which reflects the extent that all test items measure the same trait can be employed to evaluate test homogeneity. If homogeneity coefficient is low, the composite score is meaningless and cannot be used for further analyses.

Homogeneity coefficient is the proportion of variability in composite score that is accounted for by the general factor, which is viewed as common to all items. Any multidimensional test can be represented by a bifactor model that contains a general factor and local factors. Hence homogeneity coefficient can be calculated based on a bifactor model. A unidimensional test with positively worded items and negatively worded items can also be represented by a bifactor model, where the assessed construct is the general factor and method factors are local factors.

The confidence interval of homogeneity coefficient provides more information than its point estimate. There are three approaches to estimate the confidence interval of composite reliability: Bootstrap method, Delta method and direct use of the standard error generated from an SEM software output (e.g., LISREL). It has been found that the interval estimates that obtained by Delta method and Bootstrap method were almost the same, whereas the results obtained by LISREL software and by Bootstrap method had large differences. Delta method was recommended when estimating the confidence interval of composite reliability.

In order to compute the confidence interval of homogeneity coefficient, we deduced a formula by using Delta method for computing the standard error of homogeneity coefficient. Based on the standard error, the confidence interval can be obtained easily.

We used an example to illustrate how to calculate homogeneity coefficient and its confidence interval by using the proposed Delta method with LISREL software. We also illustrated how to get the same result with Mplus software that automatically calculates the standard error with Delta method and presents the confidence interval.

Before composite scores of a test are aggregated for further statistical analysis, it is recommended to report homogeneity coefficient so that readers could evaluate the extent that the statistical results are reliable.

Key words homogeneity coefficient; Delta method; confidence interval