心理学报 2011, Vol. 43, No.4, 453−461
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2011.00453

单维测验合成信度三种区间估计的比较*

叶宝娟[1] 温忠麟[1,2]

([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)

摘 要 已有许多研究建议使用合成信度来估计测验信度，并报告其置信区间。有三种方法或途径可以计算单维测验合成信度的置信区间，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。本文通过模拟研究进行比较，发现 Delta 法与 Bootstrap 法得到的置信区间相当接近，但用 LISREL 输出的标准误计算的与 Bootstrap 法得到的结果相差很大。推荐用 Delta 法估计合成信度的置信区间(使用 Mplus 容易实现)，但不能直接用 LISREL 输出的标准误来计算。举例说明了如何计算单维测验的合成信度以及用 Delta 法计算其置信区间。

关键词 合成信度；置信区间; Bootstrap 法; Delta 法; LISREL

分类号 B841

心理与教育测验中，信度(reliability)是衡量测验可靠性的指标，人们习惯用α系数作为测验信度的估计。如果测验满足下面两个条件：(1)误差不相关；(2)测验是基本τ等价测验(即任意两个题目的真分数只相差一个常数，Graham, 2006)，则α系数等于测验信度(Novick & Lewis, 1967)。但实际上很少有测验能满足第二个条件。如果条件不成立，α系数通常是低估测验信度(Biemer, Christ, & Wiesen, 2009; 刘红云, 2008; Sijtsma, 2009a, 2009b; 屠金路，王庭照，金瑜，2010)，但有些情况下也可能高估测验信度(Green & Yang, 2009; Raykov, Dimitrov, & Asparouhov, 2010; Revelle & Zinbarg, 2009; Yang & Green, 2010)。比较好的方法是，利用验证性因子模型，用合成信度(composite reliability, 即合成分数的信度，有的文献译为组合信度)来估计测验信度(Bentler, 2009; Brown, 2006; 邱皓政，林碧芳, 2009; Raykov, 1998, 2002)。同时，信度的置信区间(confidence interval)也受到重视(Duhachek & Iacobucci, 2004)，因为通过置信区间可以了解估计的误差范围。本文的目的是比较单维测验合成信度的三种区间估计。首先，简单介绍合成信度及其计算公式，和合成信度三种区间估计；然后用 4(题目个数)×3(因子负荷)×4(样本容量)×3(估计方法)的实验设计进行模拟比较；接着，根据比较结果推荐 Delta 法，并给出用 Delta 法估计合成信度的置信区间的例子和程序；最后对有关的问题进行讨论并得出结论。

1 合成信度

一个单维测验由 p 个题目 $x_1, x_2, \cdots, x_p$ 组成，测量了潜变量 ξ，$\delta_1, \delta_2, \cdots, \delta_p$ 为 $x_1, x_2, \cdots, x_p$ 的测量误差，则有

$$x_i = \lambda_i \xi + \delta_i,\ i = 1, 2 \cdots, p \qquad (1)$$

其中，λ_i 表示题目 i 在潜变量 ξ 上的负荷，δ_i 是误差项。整个测验分数 $X = x_1 + x_2 + \cdots + x_p$ 的合成信度为(Brown, 2006; Fornell & Larcker, 1981; Raykov, 1998, 2002)

$$\rho = \mathrm{var}(\sum_{i=1}^{p} \lambda_i \xi) \Big/ \left[\mathrm{var}(\sum_{i=1}^{p} \lambda_i \xi) + \sum_{i=1}^{p} \mathrm{var}(\delta_i) \right] \qquad (2)$$

公式(2)得到的只是信度的点估计。和通常的参数估计一样，对于要评价的总体参数，点估计提供的信息量有限并且不能给出估计的偏差(Raykov et al., 2010)。要知道信度估计的误差范围，需要置信区间。在评价一个测验质量时，最好用信度的区间估

收稿日期: 2010−11−23

* 国家自然科学基金项目(30870784)资助。

通讯作者: 温忠麟, E-mail: wenzl@scnu.edu.cn

计来补充信度点估计得到的信息(Raykov & Shrout, 2002; Zinbarg, Yovel, Revelle, & McDonald, 2006)。在编制测验或评价测验时，如果发现可接受的信度大小(例如 0.7)包含在信度的置信区间(例如(0.65, 0.73))中，那么对此测验做出质量评价、或做出接受还是拒绝此测验的决策时应谨慎(Raykov & Shrout, 2002)。

目前估计单维测验合成信度的标准误(SE)有两种方法：Bootstrap 法和 Delta 法，还有一种途径是直接用结构方程建模(Structural Equation Modeling, SEM)软件输出的标准误。用这些标准误就可以计算合成信度的置信区间。

Romano, Kromrey 和 Hibbard(2010)比较了各种估计α系数置信区间的方法，但还没有研究比较各种估计合成信度置信区间的方法。相比α 系数，合成信度是一种更好的估计信度的指标，所以很有必要比较估计合成信度置信区间的各种方法。

2 估计合成信度的置信区间的方法和途径

Bootstrap 法经常被用于参数估计的标准误难以用公式简单计算的场合。用 Bootstrap 法可以估计合成信度的标准误，进而计算置信区间(Raykov, 1998; 屠金路，金瑜，王庭照，2005)，这种方法需要对一个固定的样本(当作总体)进行重复取样，比较麻烦，但计算的结果是一种实证结果，通常作为真值的反映，可以用来比较其他方法计算结果是否合理。Bootstrap 法有多种取样方案，本文所说的 Bootstrap 法是从一个给定的样本中有放回地重复取样以产生出许多样本，即将原始的样本当作总体，从这个总体中重复取样以得到类似于原始样本的 Bootstrap 样本(Wen, Marsh, & Hau, 2010)。大量重复样本的内在变异性为模型参数置信区间的估计提供了实证基础，用 Bootstrap 法得到的参数分布能够完全获得取样的变异性，即使数据非正态，用 Bootstrap 法得到的参数区间估计也优于其它方法得到的区间估计(Chan, 2009; Little, Card, Preacher, & Mcconnell, 2009)。

另一种估计合成信度置信区间的方法是 Delta 法(Raykov, 2002)，近年来有不少研究(例如，Laenen, Alonso, & Molenberghs, 2007; Laenen, Alonso, Molenberghs, & Vangeneugden, 2009a, 2009b; Raykov & Marcoulides, 2004; Raykov & Penev, 2009; 2010)。这种方法简单易行，但只是一种近似算法，其精度需要权衡。Delta 法先用验证性因子分析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)。

用 Delta 法估计单维测验合成信度的标准误的公式(Raykov, 2002)如下

$$\mathrm{SE}=\sqrt{D_1^2\,\mathrm{var}(u)+D_2^2\,\mathrm{var}(v)+2D_1D_2\,\mathrm{cov}(u,v)} \quad (3)$$

其中，SE 是标准误，u 是标准化因子负荷的和，v 是误差方差的和：

$$u=\sum_{i=1}^{p}\lambda_i,\ v=\sum_{i=1}^{p}\theta_{ii} \quad (4)$$

D_1 和 D_2 由下面公式计算得到：

$$D_1=\frac{2uv}{(u^2+v)^2},\ D_2=\frac{-u^2}{(u^2+v)^2} \quad (5)$$

除了上述 Bootstrap 法和 Delta 法可以计算合成信度的标准误、进而计算置信区间外，在一些 SEM 软件中添加额外参数估计合成信度时，会像平常参数那样，在结果文件中给出合成信度的点估计值的同时，也给出点估计值的标准误，因而也可以计算合成信度的置信区间。这里以常用的 LISREL 软件输出的标准误进行比较，但讨论部分也涉及另一个常用的 SEM 软件 Mplus。

Bootstrap 法得到的结果最为可信，但需要数据模拟技术，计算过程相当麻烦。Delta 法是一种近似计算，可以根据计算公式在 SEM 软件(本文用 LISREL)中编程，在做验证性因子分析的同时给出了计算标准误所需要的参数，将这些参数带入公式(5)和(3)进行简单计算即可得标准误。直接使用 LISREL 输出的标准误比用 Delta 法还要简单。问题是，后面两种结果精度如何？下面通过模拟研究进行比较。

3 模拟研究方法

3.1 研究设计

对于单维测验，模型只有一个潜变量，考虑的因素有：题目个数，因子负荷，样本容量，估计方法。(1)题目个数，用少于 3 个题目计算的信度没有什么意义(Bacon, Sauer, & Young, 1995)，因此设置的题目个数应不少于 3 个，本研究设置的题目个数为 3, 6, 10, 15。(2)因子负荷，在一个测验中，因子负荷小于 0.3 或大于 0.9 的题目罕见(Bacon et al., 1995)，因此设置的因子负荷在 0.3–0.9 之间，分为 3 种情况：高负荷(0.7~0.9)，中负荷(0.5~0.7)，低负

荷(0.3~0.5)。(3)样本容量，对心理学研究而言，样本容量为 1000 已经相当大，因子分析的许多研究样本容量为 100 至 300 (Zhang, Preacher, & Luo, 2010),并且，心理学还会用到类似 500 的中等容量的样本，所以本研究设置的样本容量分别为 100, 300, 500, 1000。(4)估计方法或途径，包括 Bootstrap 法、Delta 法和 LISREL 软件的结果文件输出。本模拟实验是一个 4×3×4×3 的设计，前面三个因素是被试间的，最后一个因素是被试内的。

在每种处理(即水平组合)中，模拟一批正态数据，然后随机取一个样本(前提是该样本较好地拟合了一个单维模型)：RMSEA 和 SRMR 小于 0.08, NNFI 和 CFI 大于 0.9 (温忠麟，侯杰泰，Marsh, 2004)，用 Delta 法和 Bootstrap 法计算合成信度的标准误, LISREL 也会输出一个标准误。比较三个标准误的差异，也就相当于比较了合成信度的置信区间的差异。还可以了解当测验题目、因子负荷和样本容量变化时，这些差异是如何变化的。

3.2　用 Bootstrap 法估计合成信度的置信区间

本文用 Bootstrap 法计算合成信度的置信区间有 3 个步骤：第一步，从原始样本中重复取样 1000 次，得到 1000 个 Bootstrap 样本，容量与原始样本的容量相同。这一步可以用大多数的结构方程软件来实现，本文使用 LISREL 8.72; 第二步，计算 1000 个 Bootstrap 样本的合成信度(点估计); 第三步，计算第二步得到的 1000 个 Bootstrap 样本合成信度的标准差，这个标准差就是用 Bootstrap 法计算得到的合成信度的标准误，进而计算合成信度的置信区间，区间中点是原始样本的合成信度，区间半径是 Bootstrap 法得到的标准误的 1.96 倍。

3.3　用 Delta 法估计合成信度的置信区间

Delta 法计算合成信度的置信区间时，一种做法是先用附录一的 LISREL 程序计算合成信度点估计值和区间估计所需要的参数，再根据公式(5)和(3)计算合成信度的标准误，进而计算置信区间，用这种方法比用 Bootstrap 法估计合成信度的置信区间要方便得多。

3.4　用 LISREL 软件结果文件输出的标准误计算合成信度的置信区间

在 LISREL 程序中添加额外参数计算合成信度时，其结果文件在输出合成信度的点估计值的同时，也给出点估计值的标准误，直接用来计算置信区间，用这种方法计算合成信度的置信区间是三种方法中最简单的一种方法。

4　模拟研究结果

合成信度的三种标准误的比较见表 1。在 Bootstrap 重复取样的 1000 个样本中，有些样本的模型是不收敛的，在收敛的模型中，有些模型的解是不恰当的(improper)，比如方差(或标准误)的估计值是负值。本研究所设计的 48 个处理条件中，有 41 个处理条件对应的 Bootstrap 样本的恰当解比例为 100%，其余 7 个处理条件的恰当解的比例在 97%以上。计算用 Bootstrap 法得到的标准误时只使用收敛到恰当解的样本的结果。因为 Bootstrap 法赖以计算的样本不少于 970 个，所以计算结果有效。

因为 Bootstrap 法得到的标准误是一种实证结果，可以看作是真值，将其他方法计算的标准误与其比较来计算偏差。因为置信区间半径是标准误的 2 倍(精确点说是 1.96 倍)，所以如果标准误相差 0.01，信度的下限就会相差 0.02 (上限没有必要关注)。从应用角度看，如果信度下限相差 0.01 (相应的标准误相差 0.005)，只是微小差别; 如果信度下限相差 0.02 (相应的标准误相差 0.01)，是有点差别; 如果信度下限相差 0.05(相应的标准误相差 0.025)，是有实质差别。

4.1　Delta 法的标准误

比较 Delta 法和 Bootstrap 法的结果(见表 1 的 Bias_D 列)。Delta 法的标准误偏差(绝对值)普遍很小，在设计的 48 个处理中，只有 4 个处理的标准误偏差(绝对值)超过 0.005。这 4 个处理是：N=100，负荷为低，题目数为 3、6 和 10; N=100，负荷为中，题目数为 3。标准误偏差(绝对值)超过 0.01 的只有 2 个处理：N=100，负荷为低，题目数为 3、6。仔细检视表 1 中 Delta 法的标准误偏差，有正有负，但都很小，特别是题目数量多时。所以，可以认为 Delta 法估计的标准误是近似无偏的。

在 Delta 法中，只有小 N(不超过 200)会引起小偏差，所以，除了小 N 要谨慎外，都可以使用 Delta 法估计标准误，进而计算信度的置信区间。

4.2　LISREL 软件输出的标准误

比较 LISREL 软件的输出结果和 Bootstrap 法的结果(见表 1 的 Bias_L 列)。LISREL 软件输出的标准误全部偏差为正，在设计的 48 个处理中，全部偏差都超过 0.025，半数以上还超过了 0.05。这说明，LISREL 软件输出的标准误(记为 SE_L)严重高估了标准误，特别是当样本容量较少的时候。例如，N=100 时, SE_L 偏差几乎全部超过 0.1。

表 1 单维测验合成信度的三种标准误比较

题目个数	负荷	*N*	信度	SE_B	SE_D	SE_L	Bias_D	Bias_L
3	低	100	0.504	0.070	0.086	0.180	0.016	0.110
		300	0.504	0.051	0.048	0.082	–0.003	0.031
		500	0.486	0.038	0.040	0.079	0.002	0.041
		1000	0.399	0.032	0.033	0.076	0.001	0.044
	中	100	0.654	0.068	0.060	0.164	–0.008	0.096
		300	0.621	0.037	0.038	0.098	0.001	0.061
		500	0.629	0.030	0.029	0.077	–0.001	0.047
		1000	0.664	0.017	0.018	0.054	0.001	0.037
	高	100	0.807	0.036	0.034	0.149	–0.002	0.113
		300	0.831	0.019	0.017	0.086	–0.002	0.067
		500	0.842	0.013	0.012	0.066	–0.001	0.053
		1000	0.858	0.007	0.008	0.047	0.001	0.040
6	低	100	0.598	0.075	0.063	0.177	–0.012	0.102
		300	0.573	0.039	0.038	0.108	–0.001	0.069
		500	0.558	0.031	0.031	0.082	0.000	0.051
		1000	0.541	0.022	0.023	0.060	0.001	0.038
	中	100	0.762	0.038	0.037	0.156	–0.001	0.118
		300	0.767	0.022	0.021	0.090	–0.001	0.068
		500	0.768	0.015	0.016	0.068	0.001	0.053
		1000	0.767	0.012	0.011	0.050	–0.001	0.038
	高	100	0.907	0.017	0.014	0.138	–0.003	0.121
		300	0.917	0.008	0.007	0.086	–0.001	0.078
		500	0.901	0.007	0.007	0.061	0.000	0.054
		1000	0.917	0.004	0.004	0.046	0.000	0.042
10	低	100	0.702	0.052	0.045	0.161	–0.007	0.109
		300	0.677	0.026	0.028	0.096	0.002	0.070
		500	0.665	0.023	0.022	0.076	–0.001	0.053
		1000	0.634	0.017	0.017	0.055	0.000	0.038
	中	100	0.868	0.019	0.020	0.148	0.001	0.129
		300	0.866	0.011	0.012	0.091	0.001	0.080
		500	0.856	0.010	0.010	0.066	0.000	0.056
		1000	0.862	0.006	0.006	0.048	0.000	0.042
	高	100	0.925	0.012	0.011	0.141	–0.001	0.129
		300	0.940	0.005	0.005	0.081	0.000	0.076
		500	0.943	0.004	0.004	0.062	0.000	0.058
		1000	0.948	0.003	0.002	0.046	–0.001	0.043
15	低	100	0.776	0.035	0.033	0.158	–0.002	0.123
		300	0.769	0.019	0.020	0.091	0.001	0.072
		500	0.745	0.016	0.017	0.076	0.001	0.060
		1000	0.732	0.012	0.012	0.053	0.000	0.041
	中	100	0.889	0.014	0.016	0.172	0.002	0.158
		300	0.890	0.009	0.009	0.088	0.000	0.079
		500	0.882	0.007	0.008	0.063	0.001	0.056
		1000	0.902	0.004	0.005	0.047	0.001	0.043
	高	100	0.974	0.003	0.004	0.134	0.001	0.131
		300	0.970	0.002	0.003	0.072	0.001	0.070
		500	0.962	0.002	0.003	0.062	0.001	0.060
		1000	0.966	0.001	0.002	0.045	0.001	0.044

注：SE_B 和 SE_D 分别表示 Bootstrap 法和 Delta 法得到的标准误。SE_L 表示 LISREL 软件输出的标准误。将 Bootstrap 法估计的标准误看作真值, Bias_D 表示 Delta 法的标准误的偏差, Bias_L 表示 LISREL 软件输出的标准误的偏差。

利用 LISREL 软件输出的 SE_L 计算合成信度的置信区间, 下限会偏低, 并且幅度还较大。所以, 不能直接使用 LISREL 输出的标准误计算合成信度的置信区间。

4.3　合成信度的标准误变化情况

因为 Bootstrap 法的结果是一种实证结果, 所以检视 Bootstrap 法的标准误的变化情况可以知道不同条件下合成信度估计的精度情况。

保持其他条件不变, 随着题目个数的增加, 标准误会变小, 唯一的例外是 N=100、因子负荷低的时候, 3 个题目的标准误比 6 题目的标准误略小。这可能与在这两种情况下, Bootstrap 法的恰当解的百分比不同有关。保持其他条件不变, 随着负荷的升高, 用 Bootstrap 法估计的标准误减小。保持其他条件不变, 随着样本容量的增加, 用 Bootstrap 法估计的标准误减小。

总之, 题目越多、负荷越高或者样本容量越大, 合成信度的标准误越小, 从而合成信度估计越精确。容易看出, 用 Delta 法得到的标准误与 Bootstrap 法得到的标准误变化情况一致。因为 LISREL 输出的标准误不能使用, 所以没有必要讨论其变化情况。

5　用 Delta 法计算单维测验合成信度的置信区间示例

上面的研究结果是, 可以使用 Delta 法的标准误计算合成信度的置信区间。下面以一个 6 个题目的单维测验(N=300)为例, 说明如何用 Delta 法计算合成信度及其置信区间。附录一的 LISREL 程序可以用来计算合成信度点估计值以及区间估计所需要的参数, 这个程序与普通的 CFA 程序几乎相同, 只不过多了 5 个额外参数(additional parameter, AP), 参照程序说明, 读者很容易理解。这个模型的拟合指数为: RMSEA=0.023, NNFI=0.995, CFI=0.997, SRMR = 0.026, 模型拟合很好。

合成信度的点估计值, 公式(3)中的参数 $\mathrm{var}(u), \mathrm{var}(v), \mathrm{cov}(u,v)$ 及公式(4)中的参数 u,v 都可以在 LISREL 的输出结果中找到, 见附录一的注释。将 LISREL 的输出结果代入公式(5)和(3)很容易求得此例合成信度的标准误, 进而计算其置信区间, 此例计算得到的合成信度的点估计为 0.767, 标准误为 0.021, 合成信度的 95%的置信区间为(0.726, 0.808)。假设测验信度达到 0.7 才可接受, 因为合成信度置信区间的下限(0.726)超过 0.7, 所以此测验信度可以接受。

在此例中, 用 Bootstrap 法估计合成信度的置信区间得到的标准误是 0.022, 与用 Delta 法得到的标准误几乎相同, 得到的合成信度的 95%的置信区间为(0.725, 0.809)。而 LISREL 结果文件输出的标准误是 0.090(见附录一的注释), 高估了好几倍, 用此标准误估计的置信区间为(0.590, 0.944), 其下限在可接受的信度水平之下, 测验的信度变成不可接受, 而这与实际不符。

6　讨论和结论

在评价一个测验质量的优劣时, 不仅需要信度的点估计, 区间估计也很重要, 通过区间估计可以衡量信度估计的精确性。如果信度的标准误大, 则置信区间较长, 信度估计的精确性不高; 反之, 如果信度的标准误小, 置信区间较短, 信度估计的精确性就高。

如果测验信度的置信区间的下限比预先设定的可接受的信度还大, 那么此测验的信度可以接受, 即可以接受此测验。如果测验信度的置信区间的上限比可接受的信度还小, 那么此测验的信度不高, 不应当使用此测验。如果可接受的信度包含在测验信度的置信区间中, 那么此测验的信度值得怀疑, 在做出接受或者拒绝此测验的决定时应谨慎, 应当做更多的研究。

本文比较了三种估计单维测验合成信度的标准误的方法或途径: Bootstrap 法、Delta 法和 LISREL 输出的标准误。结果发现 Delta 法的标准误与 Bootstrap 法的标准误差异很小, 而 LISREL 输出的标准误远远大于 Bootstrap 法的标准误。Bootstrap 法的结果是一种实证结果, 可以作为真值看待, 但 Bootstrap 法计算过程相当麻烦。Delta 法通过近似计算得到标准误, 与 Bootstrap 法的结果差别很小, 而且比 Bootstrap 法简单得多, 所以推荐使用 Delta 法。但当 N 较小(不足 200)而且得到的区间下限勉强超过可接受的信度界值时, 要谨慎。虽然使用 LISREL 输出的标准误计算合成信度置信区间最简单, 但结果不可靠, 所以不能使用。

随着题目个数的增加、负荷的升高或样本容量的增多, 用 Bootstrap 法(还有 Delta 法)估计的合成信度的标准误倾向于减小。这提示我们, 在编制测验时, 为了确保测验的质量、提高信度估计的精度, 应适当增加测验题目, 选用因子负荷高的题目, 并尽可能多施测一些被试。

值得留意的是，与本文所应用的 LISREL 一样，Mplus 也可以在程序中添加额外参数，计算合成信度；不同的是，Mplus 不是按平常的参数给出其标准误，而是用 Delta 法计算其标准误(Muthén & Muthén, 2010)，并且直接给出置信区间。因此 Mplus 软件结果文件输出合成信度的标准误与本文用 Delta 法求得的合成信度标准误是一样的(计算误差除外)，附录二给出了 Mplus 计算本文所举例子的合成信度的程序，用程序中 OUTPUT 部分的 CINTERVAL 命令可以直接得到参数的置信区间。用此程序可以直接得到合成信度的点估计值、Delta 法的标准误，以及相应的合成信度置信区间。参照程序中的说明，读者容易理解。相比借助于 LISREL 软件使用 Delta 法，借助于 Mplus 软件方便很多，一个简单的 Mplus 程序(如附录二)即可求得合成信度及其置信区间，因此推荐用 Mplus 软件计算合成信度的置信区间。

求合成信度的置信区间，其实可以只考虑使用单侧概率。例如，前面计算得到的合成信度的 95% 的置信区间为(0.726, 0.808)，则合成信度低于 0.726 的可能性为 2.5%。实际应用中，我们不需要有 95%的把握说合成信度在置信区间中，只需要有 95%的把握说，合成信度不低于置信区间下限。这样，当置信区间的半径是标准误的 1.65 倍(而不是 1.96 倍)时，就有 95%的把握说，合成信度不低于置信区间下限了。

本文在数据正态，误差不相关的情形之下，用模拟研究比较了三种估计单维测验合成信度的置信区间的方法，并由此得出了结论。通常，假设误差不相关是合理的，但非正态数据却经常碰到。当数据非正态时，本文结论是否同样适用呢？还有，本文对单维测验得到的结论是否适用于多维测验呢？这些都有待进一步研究。

参 考 文 献

Bacon, D. R., Sauer, P., & Young, M. (1995). Composite reliability in structural equations modeling. *Educational and Psychological Measurement, 55*, 394–406.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143.

Biemer, P. P., Christ, S. L., & Wiesen, C. A. (2009). A general approach for estimating scale score reliability for panel survey data. *Psychological Methods, 14*, 400–412.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (p.339). New York: Guilford Press.

Chan,W. (2009). Bootstrap standard error and confidence intervals for the difference between two squared multiple correlation coefficients. *Educational and Psychological Measurement, 69*, 566–584.

Chiou, H. J., & Lin, B. F. (2009). *The principle of structural equation modeling and its applications* (p.103). Beijing: China Light Industry Press.

[邱皓政，林碧芳．(2009). *结构方程模型的原理与应用* (p.103). 北京：中国轻工业出版社.]

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology, 89*, 792–808.

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*, 930–944.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121–135.

Laenen, A., Alonso, A., & Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika, 72*, 443–448.

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., & Vangeneugden, T. (2009b). Reliability of a longitudinal sequence of scale ratings. *Psychometrika, 74*, 49–64.

Liu, H. (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.]

Little, T. D., Card, N. A., Preacher, K. J., & Mcconnell. E. (2009). Modeling longitudinal data from research on adolescence. In R. M. Lerner & L. D. Steinberg (Eds.), *Handbook of adolescent psychology* (3rd ed., Vol. 1, pp. 15–54). Hoboken, NJ: John Wiley & Sons.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed., pp.641–644). Los Angeles: Muthén & Muthén.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13.

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22*, 369–374.

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., & 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., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable

analysis. *Structural Equation Modeling, 17*, 629–641.
Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*, 195–212.
Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154.
Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A monte carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement, 70*, 376–393.
Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*, 107–120.
Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*, 169–173.
Tu, J., Jin, Y., & Wang, T. (2005). Study of using bootstrap to estimate confidence interval of composite reliability. *Psychological Science, 28*, 1199–1200.
[屠金路, 金瑜, 王庭照. (2005). Bootstrap 法在合成分数信度区间估计中的应用. *心理科学*, 28, 1199–1200.]
Tu, J., Wang, T., & Jin, Y. (2010). Using the structural equation modeling approach to estimate multi-factor non- congeneric composite reliability. *Psychological Science, 33*, 666–669.
[屠金路, 王庭照, 金瑜. (2010). 结构方程模型下多因子非同质测量合成分数的信度估计. *心理科学, 33*, 666–669.]
Wen, Z., 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. (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., Marsh, H. W., & Hau, K. T. (2010). Structural equation model of latent interactions: An appropriate standardized solution and its scale-free properties. *Structural Equation Model, 17*, 1–22.
Yang, Y., & Green, S. B. (2010). A Note on structural equation modeling estimates of reliability. *Structure Equation Modeling, 17*, 66–81.
Zhang, G., Preacher, K. J., & Luo, S. (2010). Bootstrap confidence intervals for ordinary least squares factor loadings and correlations in exploratory factor analysis. *Multivariate Behavioral Research, 45*, 104–134.
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 Measuremen, 30*, 121–144.

附录一　计算单维测验的合成信度的 LISREL 程序。

```
DA NI=6 NO=300 MA=CM
CM SY
1.095
0.347    0.954
0.408    0.308    1.019
0.299    0.252    0.382    0.887
0.403    0.380    0.370    0.341    1.097
0.422    0.351    0.335    0.356    0.485    1.121
MO NX=6 NK=1 TD=DI, FR AP=5
  ! AP=5  表示增加 5 个额外参数
PA LX
6(1)
CO PAR(1)=LX(1 1)+LX(2 1)+LX(3 1)+LX(4 1)+LX(5 1)+LX(6 1)
  ! PAR(1)  等于所有题目因子负荷的和, 即等于公式(4)中的 u
CO PAR(2)=TD(1 1)+TD(2 2)+TD(3 3)+TD(4 4)+TD(5 5)+TD(6 6)
  ! PAR(2)  等于所有题目误差方差的和, 即等于公式(4)中的 v
CO PAR(3)=PAR(1)^2
CO PAR(4)=PAR(2)+PAR(3)
CO PAR(5)=PAR(3)*PAR(4)^–1
  ! PAR(5)  等于合成信度点估计值
OU ME=ML ALL ND=3
```

注释：合成信度的点估计值及 LISREL 结果文件输出的标准误, 对应于 LISREL 的输出结果“LISREL Estimates”部分的“ADDITIONAL PARAMETERS”中的 AP(5)及其标准误, 分别为 0.767 和 0.090。u,v 对应于“ADDITIONAL PARAMETERS”中的 AP(1)与 AP(2), 分别为 3.616 和 3.974。$\mathrm{var}(u), \mathrm{var}(v), \mathrm{cov}(u, v)$ 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 0.038, 0.021 和 –0.003。

附录二 计算单维测验的合成信度及其置信区间的 Mplus 程序。

```
DATA:       FILE IS p.dat;
VARIABLE: NAMES ARE x1-x6;
MODEL:     F1 BY x1-x6*(p1-p6);
            x1-x6 (a1-a6);
            F1@1;
MODEL CONSTRAINT:
            new(H1-H4);
            !H1-H4 为新增参数
            H1=(p1+p2+p3+ p4+p5+p6)**2;
            !H1 等于所有题目因子负荷的和的平方
            H2=a1+a2+a3+ a4+a5+a6;
            !H2 等于所有题目误差方差的和
            H3=H1+H2;
            H4=H1/H3;
            !H4 等于合成信度点估计值
OUTPUT:
            CINTERVAL;
            !输出参数的置信区间
```

注释：合成信度的点估计值及 Mplus 软件给出的点估计值的标准误，对应于 Mplus 输出结果中的“MODEL RESULTS”部分的“New/Additional Parameters”中的 H4 及其标准误，分别为 0.767 和 0.021。合成信度的 95%置信区间的下限和上限，对应于“CONFIDENCE INTERVALS OF MODEL RESULTS”部分的“New/Additional Parameters”中的 H4 的“Lower 2.5%”和“Upper 2.5%”，分别为 0.726 和 0.808。

A Comparison of Three Confidence Intervals of Composite Reliability of A Unidimensional Test

YE Bao-Juan[1]; WEN Zhong-Lin[1,2]

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

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

Abstract

The widely used coefficient α may underestimate or overestimate reliability when its premise assumption is violated and therefore is not a good index to evaluate reliability. Composite reliability can better estimate reliability by using confirmatory factor analysis (see e.g., Bentler, 2009; Green & Yang, 2009). As is well known, point estimate contains limited information about a population parameter and could not give how far it could be from the population parameter. The confidence interval of the parameter could provide more information. In evaluating the quality of a test, the confidence interval of composite reliability has received more and more attention in recent years.

There are three approaches to estimate the confidence interval of composite reliability of a unidimensional test: Bootstrap method, Delta method and directly using the standard error in the output of an SEM software (e.g., LISREL). Each of the three approaches produces a standard error of composite reliability. Then the confidence interval can be easily formed based on the standard error. Bootstrap method provides an empirical result of the standard error of composite reliability and is the most credible, but the method needs data simulation technique and is not be easily mastered by general applied researchers. Delta method computes the standard error of composite reliability by approximate calculation, and the method is much simpler than Bootstrap method. LISREL software can directly give the standard error of composite reliability, and this method is the simplest among the three methods.

To evaluate the standard errors of composite reliability obtained by Delta method and LISREL software, we compared them with that obtained by Bootstrap method, because the latter can be treated as the true value in theory. A simulation study was conducted to the comparison. Four factors were considered in the simulation design: (a) the number of items on each test (k=3, 6, 10, and 15); (b) factor loading (high, medium and low); (c) sample size (N=100, 300, 500, and 1000); (d) the method for calculating the standard error of composite reliability (Bootstrap, Delta, and LISREL). Totally, 48 treatment conditions were generated in terms of the above 4-factor simulation design (i.e., 48=4×3×4×3).

The simulation results indicated that the difference between the standard errors obtained by Delta method and Bootstrap method was ignorable under each designed condition, except when sample size was small (less than 200)and standardized factor loadings were not high (less than 0.7). However, there was substantial difference between the standard errors obtained by LISREL software directly and Bootstrap method under each designed condition. Noting that the result from Bootstrap method can be treated as the true value, we recommended that Delta method could be adopted to estimate the confidence interval of composite reliability of a unidimensional test. At the same time we revealed that the standard error directly obtained by LISREL software is severely biased.

We used an example of a unidimensional test to illustrate how to calculate composite reliability and its confidence interval by using Delta method based on LISREL output. We also showed that the same results could be directly obtained by using SEM software Mplus that automatically calculates the confidence interval with Delta method and presents the confidence interval.

Key words composite reliability; confidence interval; Bootstrap method; Delta method; LISREL