

使用 Bootstrap 方法计算认知诊断评估中的信度*

郭磊^{1,2,3,4}, 张金明⁵

(1. 西南大学心理学部, 重庆 400715; 2. 西南大学统计学博士后科研流动站, 重庆 400715;

3. 中国基础教育质量监测协同创新中心西南大学分中心, 重庆 400715; 4. 重庆市脑科学协同创新中心, 重庆 400715;

5. 伊利诺伊大学香槟分校教育心理学系, 香槟, 伊利诺伊州 61820 美国)

摘要: 测验信度是衡量测验质量的一个重要指标, 认知诊断评估中同样需要重视信度问题。现有认知诊断中计算信度的方法均有一个前提假设: 被试在前后两次测验的后验概率分布和边际概率完全相同。该假设过强, 未考虑两次测验间存在的随机误差。基于 Bootstrap 抽样, 提出了两类属性信度和模式信度的指标, 分别是积差相关法和修正的一致性法。通过模拟研究比较了新方法和现有方法在不同属性个数、属性间相关性和题目数量下的表现, 并基于英语能力认证考试 ECPE 和分数减法的实证数据验证了新方法的可行性。最后, 对信度估计的影响因素进行了讨论。

关键词: 认知诊断评估; Bootstrap; 信度

中图分类号: B841.2

文献标识码: A

文章编号: 1003-5184(2018)05-0433-07

1 引言

认知诊断评估(cognitive diagnostic assessment, CDA)已成为国内外测量学研究的关注热点。CDA 优势为不仅能获得被试能力水平, 还能诊断其在知识点上的掌握情况。通过对知识状态的估计, 可知晓强项与弱项, 指导教师开展针对性的教学补救, 实现个性化教学。由此, 认知诊断被视为新一代心理测量理论的核心(涂冬波, 蔡艳, 丁树良, 2012)。

CDA 依赖测验进行评估, 因此, 测验质量决定了评估质量。测验信度是衡量测验质量的一个重要指标(温忠麟, 叶宝娟, 2011)。一个好的测验, 首先应该保证在评价同一批被试时, 在不同时间或场合得到的测量结果是一致的。在心理与教育测验中, 常用信度来衡量测验的稳定性, 信度越高, 稳定性越强。信度向来都是心理测量学的重要研究领域, 国内外有关信度的研究数不胜数, 但大多都属于经典测验理论或项目反应理论框架内的研究。而在 CDA 中, 却很少看见信度方面的研究。因此, 对于同样依赖测验的 CDA, 对其信度的研究也就非常有必要和有价值。

目前, CDA 中的信度研究刚刚处于发展阶段, 国内外相关研究主要有: (1) Templin 等(2013)提出了属性信度的计算方法, 但未关注到模式信度的指标。本文将 Templin 的方法称作“四分相关法”。(2) Cui, Gierl 和 Chang(2012)基于后验概率分布信

息, 构建了分类一致性指标以衡量 CDA 中的模式信度, 但未提出属性信度。(3) Wang, Song, Chen, Meng 和 Ding(2015)基于前人研究, 提出了属性信度和模式信度指标, 完善了之前的研究。和 Cui 等的方法进行比较后发现新指标具有同样表现。本文将 Wang 的方法称作“一致性法”。这些研究有一个相同的基本假设: 被试在两次相同测验上估计的后验概率分布和边际分布分别相同。该假设的目的是为了构建重测信度(test-retest reliability)指标, 但该假设与现实有些许不符。但凡测量总会存在误差, 即使同一批人第二次作答同一批试题, 由于随机误差的存在, 也很难保证前后两次测验的结果完全一致。在经典测验理论中表现为观察分数不一致, 而在 CDA 中则表现为后验概率分布、边际分布不一致。因此, 在 CDA 中开发出符合测验实际情况, 能够将随机误差考虑在内的信度指标至关重要。本研究基于一次施测结果, 采用 Bootstrap 方法对后验概率及边际分布抽样, 提出了两类新的属性和模式信度指标。第一类称作积差相关法, 有两个指标: ARC(Attribute-level Reliability base on Correlation)和 PRC(Pattern-level Reliability base on Correlation); 第二类称作修正一致性法, 有两个指标: ARM(Attribute-level Reliability base on Multiplication)和 PRM(Pattern-level Reliability base on Multiplication)。新指标同样是通过计算两次测验结果的一

* 基金项目: 教育部人文社会科学研究青年基金项目(15YJC190003), 2017 年重庆市社会科学规划项目(2017PY20), 中央高校基本科研业务费专项资金(SWU1809106)。

通讯作者: 郭磊, E-mail: happygl1229@swu.edu.cn。

致性来反映重测信度,不同之处在于构造第二次测验结果的方式。四分相关法以及一致性法直接假设第二次测验结果恒等于第一次测验结果,而新方法将随机误差考虑在内,通过 Bootstrap 方法合理构造第二次测验结果。为探查新指标在模拟和实证研究中的表现,本研究将与四分相关法和一致性法进行比较。

文章按如下方式组织:第二部分分别介绍四分相关法、一致性法、基于 Bootstrap 抽样构建的新指标,并给出计算步骤。第三部分是模拟研究。第四部分是实证研究。最后一部分是结论与讨论。

2 属性和模式信度的计算方法

定义本文使用的符号:属性为 $k(k = 1, 2, \dots, K)$ 。被试为 $i(i = 1, 2, \dots, N)$ 。题目为 $j(j = 1, 2, \dots, J)$ 。被试 i 的知识状态为 $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})$, 其中 α_{ik} 为二分变量, $\alpha_{ik} = 1$ 表示被试 i 掌握第 k 个属性, $\alpha_{ik} = 0$ 为未掌握。知识状态空间为 $\Omega = (\alpha_1^*, \alpha_2^*, \dots, \alpha_L^*)$, 在独立结构中, $L = 2^K$ 。Q 矩阵为 $J \times K$ 的矩阵, 元素 $q_{jk} = 1$ 表示第 j 题考察了第 k 个属性, $q_{jk} = 0$ 表示未考察。反应向量 $x_i = (x_{i1}, x_{i2}, \dots, x_{iJ})$, $x_{ij} = 1$ 表示答对, $x_{ij} = 0$ 表示答错。被试后验概率分布为 $P_{i, \text{post}} = (P_{i1}^*, P_{i2}^*, \dots, P_{iL}^*)$, 其中 $P_{i1}^* = P\{\alpha_1^* | x_i\}$, 有 $\sum_{l=1}^L P_{il}^* = 1$ 。属性 k 的边际概率为 $P_{ik} = \sum_{l=1}^L P_{il}^* (\alpha_{lk}^* = 1)$, 其中 $(\alpha_{lk}^* = 1)$ 为指示函数, 当 $\alpha_{lk}^* = 1$ 时取值为 1, 否则取值为 0。

2.1 四分相关法

Templin 等(2013)认为 CDA 中的属性信度是前后两次施测后被试在第 k 个属性上掌握情况的一致性程度。由于知识状态 α 是二分变量, 故使用四分相关计算重测信度, 其步骤为:

根据估计结果计算属性 k 的边际概率估计值 \hat{P}_{ik} ;

创建四格表。四格分别为 $\alpha_{ik}^{(1)} = 1 \ \& \ \alpha_{ik}^{(2)} = 1$, $\alpha_{ik}^{(1)} = 1 \ \& \ \alpha_{ik}^{(2)} = 0$, $\alpha_{ik}^{(1)} = 0 \ \& \ \alpha_{ik}^{(2)} = 1$, $\alpha_{ik}^{(1)} = 0 \ \& \ \alpha_{ik}^{(2)} = 0$ 。上角标(1)和(2)表示两次施测; N 个被试的属性边际概率估计值分别计算四个格子的联合分布:

$$\begin{cases} \hat{P}(\alpha_{ik}^{(1)} = 1, \alpha_{ik}^{(2)} = 1) = \sum_{i=1}^N \hat{P}_{ik} * \hat{P}_{ik} / N \\ \hat{P}(\alpha_{ik}^{(1)} = 1, \alpha_{ik}^{(2)} = 0) = \sum_{i=1}^N \hat{P}_{ik} * (1 - \hat{P}_{ik}) / N \\ \hat{P}(\alpha_{ik}^{(1)} = 0, \alpha_{ik}^{(2)} = 1) = \sum_{i=1}^N (1 - \hat{P}_{ik}) * \hat{P}_{ik} / N \\ \hat{P}(\alpha_{ik}^{(1)} = 0, \alpha_{ik}^{(2)} = 0) = \sum_{i=1}^N (1 - \hat{P}_{ik}) * (1 - \hat{P}_{ik}) / N \end{cases} \quad (1)$$

基于四格表计算四分相关, 即得到属性 k 的重测信度。

从公式(1)中可以看出, Templin 等创建了第二次施测结果(实际并未施测), 并假设第二次估计结果恒等于第一次结果。经典测验理论模型为 $X = T + E$, X 为观测分数, T 为真分数, E 表示随机误差。该模型认为真实能力和观察分数之间呈线性关系, 并相差一个随机误差部分。尽管 CDA 测量模型与经典测验理论不同, 但基于同样道理, 即使是同一批被试作答同一份测验, 也很难保证两次测验的边际概率完全一致。因此, 四分相关法的前提假设较强, 在现实中不太容易满足, 会得到误差较大的信度估计值。

2.2 一致性方法

Wang 等延续了 Templin 等对 CDA 中重测信度定义的思想, 提出了属性信度的计算方法:

$$\hat{\gamma}_k = \sum_{i=1}^N \sum_{l=1}^2 (\hat{P}_{N \times 2} * \hat{P}_{N \times 2}) / N \quad (2)$$

和模式信度的计算方法:

$$\hat{\gamma} = \sum_i \sum_l (\hat{P}_{N \times L} * \hat{P}_{N \times L}) / N \quad (3)$$

矩阵 $\hat{P}_{N \times 2}$ 的第 i 行向量为: $\hat{P}_{i \times 2} = (1 - \hat{p}_{ik}, \hat{p}_{ik})$, 表示被试 i 未掌握(或掌握)属性 k 的边际概率。例如, 当 $N = 2$ 时, 假设被试 1 在属性 k 上的边际概率估计值为 $\hat{P}_{1k} = 0.9$, 被试 2 在属性 k 上的边际概率估计值为 $\hat{P}_{2k} = 0.7$, 则 $\hat{P}_{1 \times 2} = (0.1, 0.9)$, $\hat{P}_{2 \times 2} = (0.3, 0.7)$, $\hat{\gamma}_k = \frac{(0.1 \times 0.1 + 0.9 \times 0.9) + (0.3 \times 0.3 + 0.7 \times 0.7)}{2} = 0.7$ 。

$\hat{P}_{N \times L} = (P_{il}^*)$ 为后验概率分布, 行表示人数, 列表示所有可能的知识状态。符号“ $*$ ”表示矩阵对应元素相乘。

由公式(2)和(3)可以看出, 这两个指标的计算仍然假设第二次测验的后验概率分布和边际概率恒等于第一次测验的结果。该假设和 Templin 等一样, 偏于理想化。

2.3 基于 Bootstrap 的新方法

Bootstrap 是以样本来代表总体, 在该样本中进行放回抽样, 直至抽取 n 个数据组成一个样本。这样的程序反复进行多次, 即可产生多个样本, 基于每个样本数据就可以进行统计计算(江程铭, 李纾, 2015)。

2.3.1 属性信度的计算

使用 Bootstrap 方法计算 CDA 的属性信度步骤如下:

根据估计结果计算被试 K 个属性的边际概率

估计值 $\hat{P}_{i1}, \hat{P}_{i2}, \dots, \hat{P}_{ik}$;

单个属性为二分变量,掌握某个属性的事件服从伯努利分布,因此 \hat{P}_{ik} 为该分布的均值, $SE_i(k) = \sqrt{\hat{P}_{ik}(1 - \hat{P}_{ik})}$ 为其标准误 (Rupp, Templin, & Henson, 2010; 郭磊, 郑蝉金, 边玉芳, 2015)。从样本中随机抽取一个被试 i , 其掌握属性 k 的均值为 \hat{P}_{ik} , 标准差为 $SE_i(k)$ 。

从正态分布 $N(\hat{P}_{ik}, SE_i(k))$ 中抽取一个值, 记作 \hat{P}_{ik} , 当作被试 i 第二次测验得到的边际概率;

重复步骤 ② 和 ③ M 次, 构造联合分布 $(\hat{P}_{ik,m}, \hat{P}'_{ik,m}), m = 1, 2, \dots, M$;

分别计算属性信度 ARC 和 ARM 指标:

$$ARC = cor(\hat{P}_{ik,m}, \hat{P}'_{ik,m}) \quad (4)$$

$$ARM = \sum_i \sum_l (\hat{P}_{ik,m}^{\#} * \hat{P}_{il,m}^{\#}) \quad (5)$$

cor 表示计算两列数据的皮尔逊积差相关。 $\hat{P}_{ik,m}^{\#}$ 和公式 (2) 中的 $\hat{P}_{N \times 2}$ 区别在于, 放松了被试在前后两次测验的边际概率完全相同的假设, 允许 $\hat{P}_{ik} \neq \hat{P}'_{ik}$, 更加符合测验的现实情景。

2.3.2 模式信度的计算

Bootstrap 方法计算模式信度, 区别在于步骤 (1) 是根据参数估计结果计算出被试的后验概率分布 $P_{i,post} = (P_{i1}^*, P_{i2}^*, \dots, P_{iL}^*)$ (若使用后验众数方法 MAP, 将会把被试的知识状态判归到最大后验概率 $\hat{P}_{i,max} \in \{P_{i,post}\}$ 所对应的类别中)。步骤 (2) 计算的模式标准误为: $SE_i(pattern) = \sqrt{\hat{P}_{i,max}(1 - \hat{P}_{i,max})}$ 。其余步骤同前。两个模式信度指标: PRC 和 PRM 为:

$$PRC = cor(\hat{P}_{i,max,m}, \hat{P}'_{i,max,m}), m = 1, 2, \dots, m \quad (6)$$

$$PRM = \sum_i \sum_l (\hat{P}_{ik,m}^{\#} * \hat{P}_{il,m}^{\#}) / N \quad (7)$$

使用 Bootstrap 计算 CDA 信度有以下优势: ① 贴近实际。新的方法突破了“假设被试在前后两次测验的后验概率分布和边际概率完全相同”的局限。② 抽样有据可依。被试只参加一次测验, 需要模拟出被试在参加第二次相同测验时的结果, 方可计算重测信度。四分相关法和一致性法均假设两次结果的后验概率和边际概率完全相同, 而本研究提出的新指标允许第二次测验结果加入随机误差。在此, 需要厘清一个关键问题: 若胡乱加入随机误差, 会与被试自身情况相违背。例如, 在保证测验信度较高的时候, 一个能力为 90 分的被试, 即使在第二次作

答相同测验存在误差时, 也不可能得 60 分, 而是更有可能得 88 分。本研究采用 Bootstrap 方法基于被试作答完测验的估计结果所计算出来的标准误 $SE_i(k)$ 和 $SE_i(pattern) = \sqrt{\hat{P}_{i,max}(1 - \hat{P}_{i,max})}$ 进行抽样, 该标准误正好体现了被试能力的合理波动范围。③ 步骤清晰, 计算简单。

下面将分别通过模拟研究和实证研究比较四种方法在不同实验条件下的表现。

3 模拟研究

3.1 研究设计

本研究以 DINA 模型 (Culpepper, 2015; de la Torre, 2009; Junker & Sijtsma, 2001) 为例, 但不局限于该模型。 s 和 g 参数均从 $U(0.15, 0.25)$ 中抽取。考察 3 个变量对信度估计的影响: (1) 属性个数 K : 3 个和 5 个。(2) 题目数量 J : 5 题、10 题、20 题。 Q 矩阵如附录表 1 和表 2 所示, 行代表属性数, 列代表题目; 1 表示题目考察到该属性, 0 表示未考察。 $K = 3$ 时, 将 Q_{10} 重复即可得 20 题的 Q 矩阵。(3) 协方差矩阵 Σ 的非对角线元素 ρ : 0.2 (低相关)、0.5 (中相关)、0.8 (高相关)。

1000 名被试知识状态的生成方式如下: 依据多元正态分布 $MVN_K(0, \Sigma)$ 生成 K 维连续变量矩阵, 设定各连续变量满足标准正态分布, 用 0 为切点对各连续变量进行两段切割, 并且可以通过设定 Σ 矩阵的非对角线元素 ρ 来调控各属性之间的四分相关 (詹沛达, 陈平, 边玉芳, 2016)。

Bootstrap 取样次数 M 设置为 30000 次。本研究为 $2 \times 3 \times 3$ 的完全交叉设计, 每个实验条件重复 30 次, 以减小随机误差。

3.2 信度真值的产生

固定被试的知识状态、以及题目参数, 使用 DINA 模型重复生成 H 次被试的作答数据, 将这 H 次作答数据看作多次重测 (test - retest) 的结果。计算所有作答数据两两配对 [$H * (H - 1) / 2$ 对] 的估计一致性值, 然后将这些一致性值的均值作为信度的真值 r_T , 当重复数量足够大时, 均值可以逼近信度的真值, 本研究中 H 取 200 次, 该做法可参见 Wang 等 (2015) 的研究。其中, 一致性值的计算方法采用 Wang 等 (2015) 文中的指标:

$$PTRCR_{1,2} = \frac{1}{N} \sum_{i=1}^N I(\hat{\alpha}_i^{(1)} = \hat{\alpha}_i^{(2)}) \quad (8)$$

$$ATRCR_{k,1,2} = \frac{1}{N} \sum_{i=1}^N I(ik^{(1)} = ik^{(2)}) \quad (9)$$

$PTRCR_{1,2}$ 表示模式重测一致性指标, 下角标 1 和 2 表示第一次和第二次施测。 $ATRCR_{k,1,2}$ 表示属性 k 的重测一致性指标。

3.3 评价指标

①平均偏差

$$bias = \sum_{i=1}^{30} (r_i - r_T) / 30 \quad (10)$$

其中, r_T 为信度的真值, r_i 为每次实验的信度估计值。该值越接近于 0 越好。

②误差均方根:

$$RMSE = \sqrt{\sum_{i=1}^{30} (r_i - r_T)^2 / 30} \quad (11)$$

3.4 研究结果

表 1 和表 2 是各个指标在不同实验条件下属性和模式信度估计结果的 bias 和 RMSE 值。图 1 和图 2 为对应的 bias(A) 和 RMSE(B) 折线图。由于信度真值是 H 次作答估计一致性的均值, 因此, bias 和 RMSE 的本质是“离均差的和”与“离均差平方和的算术平方根”, 两者反映的是估计值与均值的波动大小。从整体上看, 属性信度的估计比模式信度稳定, 偏差值更小。

就属性信度来说, 新方法对属性信度的估计精确度更高。表现最好的是 ARC 方法, bias 的绝对值

离 0 最近, RMSE 在大部分实验条件下是最小的, 从图 1 中也可看出, 其 bias 的趋势线在 0 周围波动最小, RMSE 的趋势线位于最下方。ARM 的结果与一致性法表现基本相当, bias 和 RMSE 与一致性法非常接近, ARM 与一致性法的趋势线基本重合。四分相关法表现最差, bias 在 0 周围波动最大, RMSE 最大、趋势线最高。属性间相关性 ρ 对属性信度的影响并未呈现一致性趋势; 随属性个数增加, 估计偏差在整体上呈现不断增大趋势, bias 波动变大, 但 ARC 的表现仍最好; 随题目数量增多, 估计偏差在整体上呈不断减小趋势。就模式信度来说, PRC、PRM 的估计精度与一致性法相当, 三种方法的 bias 和 RMSE 值非常接近。而在有些实验条件下, PRC 的精确性要比 ARM 和一致性法要高(表 2 的第 2 至第 4 行结果)。由图 2 可知, PRC、PRM 与一致性法的 bias 趋势线波动幅度较一致, RMSE 趋势线也基本重合。除此之外, 属性间的相关性、属性个数以及题目数量对模式信度的影响与属性信度的结果基本一致。

表 1 不同方法的属性信度估计精度结果

K	ρ	J	ARC		ARM		一致性法		四分相关法		属性信度 真值
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	
3	0.2	5	-0.022	0.041	0.019	0.030	0.021	0.030	-0.048	0.059	0.684
		10	0.013	0.019	0.028	0.034	0.033	0.035	0.041	0.057	0.771
		20	0.014	0.021	0.020	0.026	0.016	0.026	0.053	0.078	0.863
	0.5	5	-0.016	0.030	0.036	0.042	0.036	0.046	-0.028	0.049	0.710
		10	0.010	0.020	0.025	0.039	0.032	0.039	0.049	0.058	0.773
		20	0.003	0.009	0.007	0.012	0.007	0.012	0.043	0.060	0.891
	0.8	5	0.015	0.037	0.032	0.055	0.033	0.055	0.022	0.031	0.739
		10	-0.016	0.037	0.019	0.039	0.017	0.039	0.064	0.082	0.832
		20	0.002	0.010	0.009	0.013	0.009	0.013	0.040	0.058	0.915
5	0.2	5	0.013	0.037	0.056	0.064	0.057	0.078	0.021	0.043	0.605
		10	0.020	0.036	0.046	0.073	-0.041	0.073	0.058	0.071	0.692
		20	-0.021	0.043	0.038	0.053	0.037	0.053	-0.042	0.061	0.753
	0.5	5	0.035	0.061	0.052	0.093	0.051	0.093	0.061	0.078	0.657
		10	0.024	0.046	0.041	0.061	0.039	0.061	0.050	0.064	0.729
		20	0.012	0.024	0.023	0.045	0.024	0.045	0.060	0.077	0.790
	0.8	5	-0.047	0.094	0.068	0.117	0.063	0.120	0.090	0.127	0.686
		10	0.018	0.034	0.027	0.046	0.024	0.046	0.056	0.071	0.785
		20	0.008	0.027	0.017	0.035	0.016	0.036	0.060	0.074	0.846

表 2 不同方法的模式信度估计精度结果

K	ρ	J	PRC		PRM		一致性法		模式信度 真值
			Bias	RMSE	Bias	RMSE	Bias	RMSE	
3	0.2	5	-0.026	0.045	0.036	0.047	0.037	0.046	0.347
		10	-0.029	0.044	0.058	0.062	0.054	0.066	0.519
		20	0.031	0.045	0.046	0.059	0.048	0.059	0.676
	0.5	5	0.016	0.029	0.065	0.079	0.062	0.080	0.397
		10	-0.046	0.075	0.062	0.071	0.061	0.073	0.531

续表 2

K	ρ	J	PRC		PRM		一致性法		模式信度 真值
			Bias	RMSE	Bias	RMSE	Bias	RMSE	
5	0.8	20	0.035	0.062	0.009	0.015	0.010	0.015	0.734
		5	-0.027	0.048	0.065	0.082	0.067	0.082	0.484
		10	0.044	0.073	0.046	0.059	0.052	0.061	0.644
	0.2	20	0.018	0.024	0.014	0.020	0.017	0.023	0.795
		5	0.063	0.117	0.060	0.107	0.059	0.107	0.239
		10	-0.066	0.116	0.064	0.113	0.064	0.113	0.269
	0.5	20	0.058	0.104	-0.047	0.085	0.051	0.086	0.348
		5	-0.107	0.158	0.118	0.156	0.120	0.156	0.285
		10	0.069	0.108	0.063	0.093	0.061	0.093	0.313
	0.8	20	0.052	0.094	0.057	0.087	0.055	0.086	0.428
		5	-0.107	0.164	-0.130	0.177	0.133	0.181	0.334
		10	0.043	0.069	0.038	0.051	0.039	0.052	0.403
		20	0.014	0.021	0.023	0.037	0.023	0.038	0.574

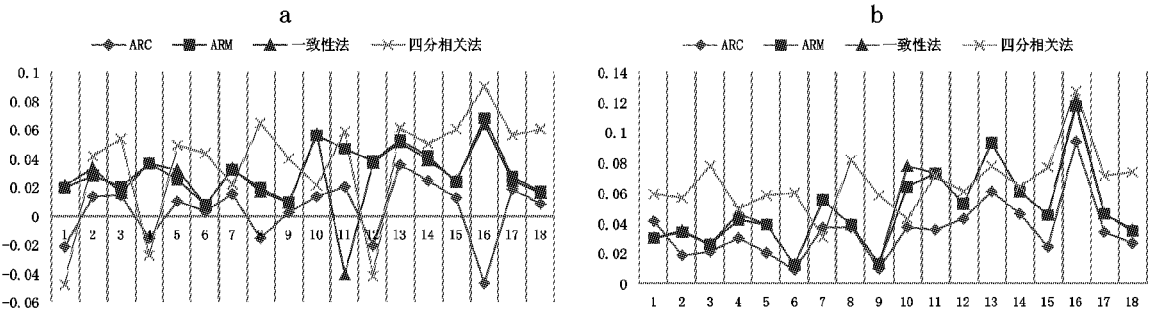


图 1 属性信度的 bias (A) 和 RMSE (B) 折线图

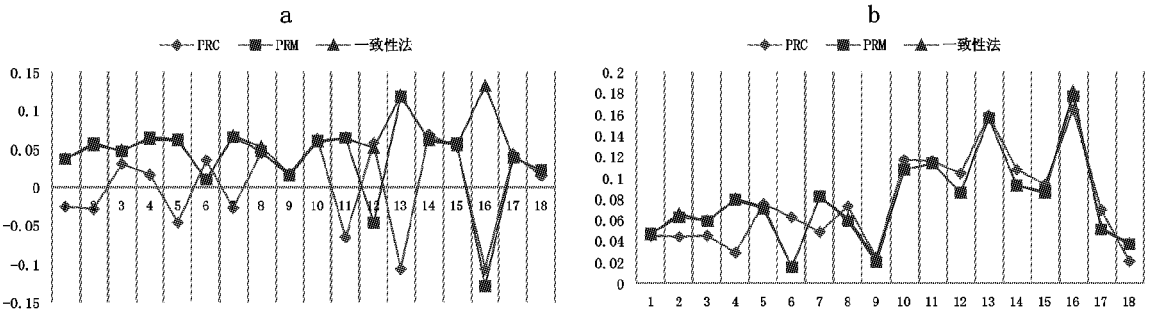


图 2 模式信度的 bias (A) 和 RMSE (B) 折线图

4 实证研究

4.1 ECPE 数据

该数据来自于 R 软件 CDM 程序包中英语能力认证考试,包含 2922 人在 28 道题目上的作答数据,考察了 3 个属性:构词规则 (Morphosyntactic rules)、

衔接规则 (Cohesive rules)、词汇规则 (Lexical rules)。作答矩阵和 Q 矩阵可分别由 data. ecpe \$ data[, -1] 和 data. ecpe \$ q. matrix 进行调用。

使用四种方法估计该数据的属性和模式信度,结果见表 3。

表 3 ECPE 信度估计结果

		ARC	ARM	一致性法	四分相关法	模式信度	PRC	PRM	一致性法
属性信度	A1	0.863	0.879	0.880	0.917	0.616	0.687	0.684	
	A2	0.760	0.809	0.808	0.834				
	A3	0.852	0.881	0.882	0.914				
	均值	0.825	0.856	0.857	0.888				

对于属性信度,模拟研究结果表明,当属性个数增大时,ARC 的估计精确度最高,之后是 ARM 和一

致性法,四分相关法表现较差。结合表 3 结果可知,使用四分相关法会高估 ECPE 的属性信度(均值为

0.888), ARM 和一致性法的属性信度均值基本接近(0.86 左右), ARC 估计的属性信度均值为 0.825。对于模式信度, 模拟研究结果表明, PRC 的表现较好, 计算得到 ECPE 模式信度为 0.616, 而 PRM 和一致性法基本相当为 0.685 左右。有趣的发现是, 不论使用何种指标, 属性 A2 的信度是最低的, 通过表 5 的 Q 矩阵分析, A1 考察了 13 次, A3 考察了 18 次, 而 A2 只考察了 6 次, 说明考察次数会影响属性信度。其原因可能有: ①当属性考察次数较少时, 该属性估计的准确性自然会降低, 导致其稳定性降低;

表 4 分数减法测验的信度估计结果

	ARC	ARM	一致性法	四分相关法		PRC	PRM	一致性法
属性信度	A1	0.959	0.966	0.968	0.992	模式信度	0.578	0.612
	A2	0.730	0.850	0.853	0.891			
	A3	0.845	0.928	0.927	0.960			
	A4	0.832	0.853	0.850	0.885			
	A5	0.722	0.711	0.703	0.654			
	均值	0.818	0.862	0.860	0.876			

模拟研究表明 ARC 表现最好, 表现最差为四分相关法。结合表 4 结果可知, 四分相关法仍高估属性信度(均值为 0.876), ARM 和一致性法估计的属性信度均值接近(均值为 0.86 左右), ARC 估计的属性信度均值为 0.818。对于模式信度, 模拟研究结果表明当属性个数增加后, PRC、PRM 和一致性法基本相当, 模式信度约为 0.6 左右。同样, 属性 A5 的信度最低, 其次是 A2 和 A4。这是因为 A5 只考察了 3 次, A2 和 A4 分别考察了 8 次和 9 次, A1 和 A3 分别考察了 14 次和 12 次。

5 结论与讨论

信度是衡量测验质量的一个重要指标, CDA 同样需要重视信度问题。本文基于 Bootstrap 抽样思想, 提出了两类计算属性和模式信度指标。新指标更加符合现实, 突破了“假设被试两次测验的后验概率和边际概率完全相同”的局限。通过模拟和实证研究, 与四分相关法和一致性法进行比较, 验证了新指标的优越性, 得到了以下主要的结论:

(1) 整体上, 属性信度的估计比模式信度稳定, 且偏差更小;

(2) 对属性信度而言, ARC 表现最优, 其次是 ARM 和一致性法, 四分相关法表现最差。属性个数增加会增大估计偏差, 题目数量增加则会减小其估计偏差;

(3) 对模式信度而言, PRC、PRM 估计精度与一致性法相当。属性间相关性、属性个数、题目数量对模式信度的影响与属性信度基本一致;

(4) 实证研究可知, 每种方法均能报告属性和

②影响信度的因素之一为测验长度, 在认知诊断中表现为属性考察次数, 当次数较少时, 信度理应不会太高。

4.2 分数减法数据

分数减法数据同样来自 CDM 程序包, 包含 536 人在 15 道题上的作答数据, 考察了 5 个属性。作答矩阵和 Q 矩阵可分别由 `data.fraction1 $ data` 和 `data.fraction1 $ q.matrix` 进行调用。使用四种方法估计该批数据的属性和模式信度, 结果见表 4。

模式信度。结合模拟研究结果, 积差相关包括的两个指标(ARC 和 PRC)表现较好。想要提高属性信度, 可适当增加该属性考察次数。

综上所述, 计算属性信度时, 综合排名为: ARC > ARM ≈ 一致性法 > 四分相关法, 推荐使用 ARC。计算模式信度时, 综合排名为: PRC > PRM ≈ 一致性法, 推荐使用 PRC。

本文结合模拟和实证研究结果, 拟探讨以下几个问题:

5.1 ARC 与 PRC 相关说明

模拟研究表明 ARC 优于其他指标, PRC 与已有指标相当。原因在于: 掌握(或未掌握)属性的边际概率值 \hat{p}_{ik} 通常要大于后验概率值 $\hat{P}_{i, \max} \in \{P_{i, \text{post}}\}$ 。例如, $K=2$ 时, 记四种知识状态 $(0,0)$ 、 $(0,1)$ 、 $(1,0)$ 和 $(1,1)$ 的后验概率分别为 P_{i1}^* 、 P_{i2}^* 、 P_{i3}^* 和 P_{i4}^* , $\sum_{l=1}^4 P_{il}^* = 1$ 。则有 $\hat{p}_{i1} = P_{i3}^* + P_{i4}^*$, $\hat{p}_{i2} = P_{i2}^* + P_{i4}^*$ 。在保证 CDA 的估计精确性前提下, 属性标准误差 $SE_i(k) = \sqrt{\hat{p}_{ik}(1 - \hat{p}_{ik})}$ 小于模式标准误差 $SE_i(\text{pattern}) = \sqrt{\hat{P}_{i, \max}(1 - \hat{P}_{i, \max})}$ 。因此, 对于属性而言, 使用 Bootstrap 抽样波动较小。

5.2 不同参数估计方法对信度的影响

Huebner 和 Wang (2011) 比较了三种参数估计方法: 后验众数法 MAP、后验期望法 EAP、极大似然估计 MLE。不同的估计方法影响后验概率分布和属性边际概率, 进而影响标准误, 导致 Bootstrap 抽样范围发生变化。本文基于 MAP 得到的结果计算

的信度,未来需探讨不同参数估计方法对信度的影响。

5.3 不同认知诊断信度指标的开发

在经典测验理论中,除重测信度,还有复本信度、内部一致性信度等。不同信度指标,其关注点不同,应用场景也不同。在报告信度时,需指出是何种信度。目前关于 CDA 中信度的研究,均从重测角度出发,这是因为该方法易于理解、指标容易构建。未来应考虑如何将其余信度指标拓展至 CDA 中,丰富 CDA 的信度指标体系。

除上述问题之外,不同的属性层级结构可能会对信度的估计带来影响,未来研究可以尝试在不同的属性层级结构下,以及不同认知诊断模型下探讨本文所提出新指标的表现。

参考文献

- 郭磊,郑蝉金,边玉芳. (2015). 变长 CD - CAT 中的曝光控制与终止规则. *心理学报*, 47(1), 129 - 140.
- 江程铭,李纾. (2015). 中介分析和自举(Bootstrap)程序应用. *心理学探新*, 35(5), 458 - 463.
- 涂冬波,蔡艳,丁树良. (2012). *认知诊断理论、方法与应用*. 北京:北京师范大学出版社.
- 温忠麟,叶宝娟. (2011). 从 α 系数到内部一致性信度. *心理学报*, 43(7), 821 - 829.
- 詹沛达,陈平,边玉芳. (2016). 使用验证性补偿多维 IRT 模型进行认知诊断评估. *心理学报*, 48(10), 1347 - 1356.
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic as-

- essment. *Journal of Educational Measurement*, 49, 19 - 38.
- Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs Sampling. *Journal of Educational and Behavioral Statistics*, 40(5), 454 - 476.
- DeCarlo, L. T. (2012). Recognizing Uncertainty in the Q - Matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement*, 36(6), 447 - 468.
- dela, T. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115 - 130.
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71, 407 - 419.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258 - 272.
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30(2), 251 - 275.
- Wang, W. Y., Song, L. H., Chen, P., Meng, Y. R., & Ding, S. L. (2015). Attribute - level and pattern - level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52(4), 457 - 476.

Using Bootstrap to Calculate Reliability in Cognitive Diagnostic Assessment

Guo Lei^{1,2,3,4}, Zhang Jinming⁵

(1. Faculty of Psychology, Southwest University, Chongqing 400715;

2. Postdoctoral Research Center for Statistics, Southwest University, Chongqing 400715; 3. Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing 400715;

4. Chongqing Collaborative Innovation Center For Brain Science, Chongqing 400715;

5. Department of Educational Psychology, University of Illinois at Urbana - Champaign, Champaign, IL, 61820 USA)

Abstract: The reliability of cognitive diagnostic assessments (CDAs) is of great importance to any cognitive diagnostic model applications. The current study proposes two attribute - level and two pattern - level reliability indices based on Bootstrap method. Simulation study is conducted to evaluate the performance of the new indices compared to existing methods under three different factors. The results indicate that: (1) overall, the estimation results of attribute - level reliability are more stable than those of pattern - level reliability when comparing the two new types of indices; (2) for attribute - level reliability, the performance of the ARC index is the best; and (3) for pattern - level reliability, the results of PRC, PRM and Wang's method are comparable. Two real data sets were also used to compare the performance of all the indices. Finally, the authors discuss several factors that may influence the degree of reliability of CDAs, and indicate some research interests for the future.

Key words: cognitive diagnostic assessment; Bootstrap; reliability