文章编号:1000-5862(2019)02-0128-07

双因子模型下 CAT 测验优化设计及其效果验证

刘馨婷,彭思韦,涂冬波*
(江西师范大学心理学院,江西 南昌 330022)

摘要: 在 2 种传统的 BCAT 测验设计的基础上,提出了 4 种新的 BCAT 测验设计,并采用国际上通用的 Monte Carlo 模拟实验的方式,从被试能力参数估计精度、题库使用的曝光率及测验的效率等 3 大指标来验证新开发的 4 种 BCAT 测验设计,再与传统的 BCAT 测验设计进行比较,以验证该文提出的 4 种新的 BCAT 测验设计的科学性、效果及优势. 最后,对 BCAT 测验设计在实际应用中的选用提出了具体的意见与建议,以供实际应用者参考及借鉴.

关键词: 双因子模型; 计算机化自适应测验; 双因子模型计算机化自适应测验; 多级评分
中图分类号: B 841 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2019.02.03

0 引言

因素分析方法(Factor Analysis,FA)是国内外心理学研究中经常使用的一种统计分析方法,它主要用于探明或验证人类心理特质的结构与维度,对人们进一步认清心理学本质有着的重要影响. 双因子模型(Bifactor Model)是一种特殊的因素分析方法,又被称为全局-局部因子模型(general-specific factor model)或是嵌套模型(nested model)[1]. 双因子模型起始于斯皮尔曼的智力二因素理论,双因子模型假定所有项目均测量了一般因子(general factor),也称为 G 因子;但同时每题最多只能测量 1 个特殊因子(specific domain factor),也称为 S 因子;且假设所有因子间(G 因子与 S 因子间以及 S 因子间)正交,即相互独立. 大量研究表明双因子模型符合认知能力、心理特质、精神病理等多类测验的结构特征.

计算机化自适应测验(Computerized Adaptive Testing,CAT)采用自适应的测量方式,即电脑根据被试的特点智能化采用适合测量被试的题目进行测量,从而突破了以往测量中所有被试作答完全相同测验或量表的缺陷,真正实现了因人施测、减少了测验长度并同时提高了测量的精度. 正因 CAT 具有以上多重优点,目前 CAT 被很多大型的国际测评采用,如美国研究生入学考试(GRE)、经企管理研究生入学考试(GMAT)、军队职业倾向测验(ASV-AB)、注册护士执照考试(NCLEX-RN)等.

为了充分发挥全息项目双因子分析模型及计算机化自适应测验(CAT)的优点,有研究者提出将两者结合,提出了全息项目双因子模型的计算机化自适应测验(BCAT),并成功将 BCAT 的思想用于心理测评实践. 如 R. D. Gibbons 等[2-3] 将 BCAT 应用于抑郁症(depression)和焦虑症的测评,其中 R. D. Gibbons 等[2] 的研究表明:在基于 BCAT 的抑郁症测评系统(CAT-D)评估中,CAT-D 平均每个患者只需用 12 题就可以达纸笔测试中被试做 389 题的效果(两者能力估计值间相关高达 0.95,即 CAT-D 只需 12 题就可以达到传统纸笔测验 389 题的效果),大大减轻了患者的测试负担,这充分表明 BCAT 在实践中具有较强的应用价值和应用前景.

虽然 BCAT 在实践中的应用价值不断突显出来,但目前国际上针对 BCAT 方法本身及相应的算法的研究仍有许多有待进一步深入的地方,尤其是 BCAT 测验设计的研究有待进一步深入与探讨. 目前国际上仅提出了 2 种 BCAT 测验设计(详见下文):一种为单维视角的 BCAT 测验设计[4],另一种是基于多维视角的 BCAT 测验设计[5-6]. 然而,在单维视角的 BCAT 测验设计中,一般先施测 G 因子,待 G 因子测试完后再测试 S 因子,直至所有 S 因子测试完毕. 即用多个单维 CAT 测试模式来处理 1 个多

收稿日期:2018-10-21
基金项目:国家自然科学基金(31660278,31760288)资助项目.
通信作者:涂冬波(1978-),男,江西南昌人,教授,博士,博士生导师,主要从事心理统计与测量的研究. E-mail: tudongbo@aliyun.com

维的 BCAT,这种测验设计模式最大的不足是没有充分利用项目的多维性特点,而且使得测验的长度更长. 更为重要的是在这种测验设计中,题目仅仅使用了 1 个维度上的信息,例如在测量 G 因子时,仅仅使用了项目在 G 因子上的信息(即仅仅利用了在 G 因子上的区分度参数) 而没有充分利用该项目同时也包含了 S 因子的信息(即没利用该项目在 S 因子上的区分度参数) ,造成了浪费. 而在多维视角的 BCAT 设计中,往往需要计算信息量矩阵的逆矩阵,在一些情况下可能会出现逆矩阵不存在的情况,因此这种方法依然存在一定的局限. 为了充分利用全息双因子模型的多维性以及各因子间的正交性(即相互独立性) ,本文拟针对当前 BCAT 测验设计的不足,分别在单维视角 BCAT 和多维视角 BCAT 上,提出 4 种新的 BCAT 测验设计,一方面进一步优化当前 BCAT 测验设计,另一方面提升 BCAT 对被试能力参数的估计精度,并为实际应用者提供新的方法支持.

1　传统的 BCAT 测验设计

目前国际上关于 BCAT 的测验设计主要有 2 种: 一是单维视角的 BCAT(Unidimensional BCAT, UBCAT) ,另一种是多维视角的 BCAT(Multidimensional BCAT,MBCAT) .

1.1　传统的基于单维视角的 BCAT 设计(UBCAT)

双因子模型最大的特点是所有因子间(一般因子 G 因子和特殊因子 S 因子) 相互独立,即 G 因子与 S 因子间、S 因子间均是相互独立的,因而有学者提出基于单维视角的 BCAT(UBCAT) [4] . 即将每 1 个维度当成是 1 个独立的维度,分别进行单维的选题和估计,并按照单维的终止策略终止测验.

在 UBCAT 中,一般因子与特殊因子的施测过程是分开序列进行的,首先施测一般因子(G 因子) ,当一般因子测试精度达到要求后,接着逐个施测特殊因子(S 因子) . 由于在施测一般因子的项目同时测量了 1 个特殊因子,因此 UBCAT 会根据在某个特殊因子上项目的作答估计被试在特征因子上的能力值,并将该特殊因子的能力值作为 UBCA 的能力初始值进行该特殊因子单维的 CAT 选题.

UBCAT 在选题时,每次只考虑 1 个因子维度,被试每完成 1 个题目,当即估计被试在当前施测维度上的潜在特质水平(θ) ,而且这时使用到的区分度参数仅仅是该项目在该维度上的区分度参数,而不使用该项目在其他维度的区分度参数,即单维的算法. 在整个 BCAT 过程,由于每次只进行 1 个维度的自适应,因此 BCAT 自始至终都是使用传统的单维 CAT 的单维能力估计、单维选题策略及单维终止策略.

限于篇幅,关于传统的基于单维视角的 BCAT 设计(UBCAT) 的详细介绍,感兴趣的读者可参考文献 [4] .

1.2　传统的基于多维视角的 BCAT 设计(MBCAT)

传统的基于多维视角的 BCAT 设计(MBCAT) [5-6] 充分考虑了双因子模型的多维特征. 与 UBCAT 的单维思路不同,MBCAT 使用了多维 CAT 的思路来完成 MBCAT,即采用多维能力估计、多维选题策略和多维终止策略.

在 MBCAT 中,一般因子(G 因子) 与特殊因子(S 因子) 的施测过程是同时进行的,选题时同时考虑一般因子和多个特殊因子,被试每完成 1 个题目,当即估计被试在一般因素与所有特殊因素上的潜在特质水平(θ) . MBCAT 测验设计本质上是沿用多维 CAT(MCAT) 的方法.

2　BCAT 的优化设计

在 BCAT 的测验设计中,涉及一般因子与特殊因子能力估计的先后顺序、选题策略、能力估计方法与终止策略等算法. 本研究中的 BCAT 优化设计包括以上算法的设计与优化,具体如下.

2.1　基于单维视角的 BCAT 优化设计(UBCAT_optimality)

2.1.1　基于单维视角的优化设计 1　UBCAT_optimality1

UBCAT_optimality1 方法是在传统 UBCAT 方法的基础上,被试测试完后,最后一次采用多维 IRT 的方法同时估计被试的 G 因子和 S 因子上的能力特质水平. 它一方面综合了所有题目的信息,另一方面充分利用了每题测量 2 个维度(G 因子和 S 因子) 的信息,而传统单维能力估计仅仅利用了每题测量 1 个维度的信息,因而 UBCAT_optimality1 有望进一步提高 UBCAT 的 G 因子能力和 S 因子能力的参数估计精度.

2.1.2　基于单维视角的优化设计 2　UBCAT_optimality2

设计建立在 UBCAT_optimality1 基础上,UBCAT_optimality2 不是在整个 UBCAT 结束时而是在 UBCAT 的整个过程中自始至终都采用多维能力的估计方法.

2.2 基于多维视角的 BCAT 优化设计(MBCAT_optimality)

双因子模型因不同维度间相互独立,因此双因子模型中的每个维度均具有单维性特点,但同时双因子模型中的每个项目一般同时测量了 2 个因子(1 个 G 因子和 1 个 S 因子),因此它又具有项目内多维的特征,故也可以考虑从多维的角度进一步优化 MBCAT.

传统 MBCAT 的测验设计实质上是沿用了多维 CAT 的思路. 一般情况下,在施测多维 CAT 的过程中,希望每 1 个维度的精度均能达到理想的标准,但是在多维 CAT 中常用的一些多维终止策略(如 T 规则[7]),即采用方差协方差矩阵的迹小于事先界定的标准来终止测验. 这种方法是保证整体达到标准,但是并不能保证在测验终止时每个维度的估计精度均能达到指标. 为了避免出现这样的问题,Wang Chang 等[7]提出采用信息矩阵逆矩阵对脚线元素 $I_{ii}^{-1}(\hat{\theta}_j)$ 最大值小于标准的方式终止测验:

$$T=\inf\{j\geqslant 1:\max(I_{kk}^{-1}(\hat{\theta}_j))\leqslant d\},k=1,2,\cdots,m,$$

即让每个维度的估计方差都小于预先设定的标准 d,这样可以保证每个维度的精度都能达到标准. 因此,在本研究中 MBCAT 的终止策略采用上述方法作为 MBCAT 的终止策略.

在 Wang Chun 等[7]研究中同时还指出,在 MIRT 模型下,某一维度 θ_k 的信度可以采用下式定义:

$$\rho_{\theta_k}=1-I_{kk}^{-1}(\theta)/\sigma_{\theta_k}^2,$$

其中 $I_{kk}^{-1}(\theta)$ 表示信息矩阵的逆矩阵中对角线元素的第 k 个元素,$\sigma_{\theta_k}^2$ 表示维度 k 的方差,因为在研究中假设各维度服从均值为 0,方差为 1 的多元正态分布,所以 $\sigma_{\theta_k}^2=1$.

根据公式 $r_{xx}=1-(S_{E_x}/\sigma_x)^2$,其中 $\sigma_x=1$,可以推出,当为单维、估计标准误 $S_{E_x}=1/\sqrt{x}$ 时,$I_{xx}^{-1}(\theta)=1/x$,所以,在多维终止策略中,d 的设置标准为 $1/x$.

采用上述方式作为终止策略,因为需要计算信息量矩阵的逆矩阵,在一些情况下可能会出现逆矩阵不存在的情况,因此上述方法依然存在一定的局限性. 同时,采用方差协方差矩阵的对脚线元素小于标准终止测验的方法并不能保证其结果能够与 UBCAT 下的方法进行比较. 因此,根据双因子的单维特性,可以考虑在施测过程中,采用单维的终止策略来结束测验,与 UBCAT 方法下的终止策略保持一致. 基于传统 MBCAT 方法,提出以下 2 种优化的设计.

2.2.1 基于多维视角的优化设计 1 MBCAT_optimality1 与传统 MBCAT 过程相似,不同点在于每选出 1 个题目估计出被试的当前能力之后,还需要计算每个能力维度的单维信息量,采用单维的信息量作为终止策略,若所有维度的单维测验信息量都达到要求,就终止测验.

在多维的 CAT 中,要求所有的维度都要达到设定的标准,才会停止选题. 因为选题的过程中会综合考虑待选题目在所有维度上的信息,在这个过程中,有可能出现某些维度已经满足了精度标准,但还会继续选用该维度的题目,从而导致施测总题数变长. 为了防止这种情况,需要对已经满足精度标准的维度进行控制,一旦维度满足标准,就让这个维度的剩余题目退出选题. 在接下来的施测过程中,只关注那些剩余的还未满足的维度. 由此基于传统 MBCAT 方法以及单维信息量终止的 MBCAT 方法,提出以下第 2 种新的 MBCAT 优化设计.

2.2.2 基于多维视角的优化设计 2 MBCAT_optimality2 施测的过程与单维信息量终止的 MBCAT 相似. 采用多维选题、多维估计,每选出 1 个题目,估计完被试在各维度上的特质水平,就计算各个维度上的单维测验信息量. 与方法 7 相似,不需要等所有维度都达到精度标准才能终止测验. 如果有维度的测验信息量已经满足标准,在接下来的选题过程中,就不再选择这个维度的题目.

3 实验研究

采用 Monte Carlo 模拟的方法进行实验研究,验证本研究中新开发的 5 种 BCAT 优化设计方法的科学性与合理性,并与 2 种传统的 BCAT 方法进行比较.

3.1 研究设计

研究采用单因素(即 BCAT 设计)的实验设计,探讨并比较 6 种 BCAT 设计(详见表 1)的效果,主要比较 6 种 BCAT 设计的能力估计精度指标、BCAT 测试效率指标、题库曝光率指标.

无关变量的控制:(i)在 UBCAT 框架下,一般因子初始题的选取采用随机选题,单维的选题策略采用最大 Fisher 信息量选题,单维的估计方法采用的是单维 EAP 估计(每个维度选取 $[-3,3]$ 上的 31 个积点),终止策略采用的是计算单维的测验信息量来终止测验. 关于测验信息量终止的标准选取,主要是参考了 R. D. Gibbons 等[2-3]的标准. 当维度的精度 $S_E\leqslant 0.3$ 时,就终止测验,相当于在其研究中,终

止时的信息量约为 11. 11. 在本研究中采用与之相近的信息量标准,即保证每个维度的测验信息量大于 12($I \geqslant 12$) ,就是每个维度的测验标准误 $S_E \leqslant 1/\sqrt{12}$. (ii) 在 MBCAT 框架下,多维的选题策略采用 D. O. Segall 等[5-6]提出的 D 优化方法,能力估计方法为多维 EAP. 因为各维度的估计标准误 S_E 设置为 $1/\sqrt{12}$,根据前文中的推导,多维的终止策略按照信息矩阵的逆矩阵对角线元素 $I_{ii}^{-1}(\hat{\theta}_j) \leqslant 1/12$,单维的终止策略为单维测验信息量≥12. (iii) 同时,为了防止因题库和被试差异导致的误差,本研究中所有 BCAT 设计下的题库参数及被试参数相同.

表 1　BCAT 及其几种优化的测验设计

类型	BCAT 设计	维度测试设计	选题策略	终止策略	能力估计方法
UBCAT	传统 UBCAT	序列测试 $G \rightarrow S_1 \rightarrow S_2 \cdots$	单维信息量选题策略	每个维度均达到了事先界定的信息量	单维能力估计
	UBCAT–optimality1	序列测试 $G \rightarrow S_1 \rightarrow S_2 \cdots$	单维信息量选题策略	每个维度均达到了事先界定的信息量	单维能力估计 + 多维能力估计: 测试过程中使用单维能力估计,但测试结束后,使用所有题目及多维能力估计方法估计所有能力维度
	UBCAT–optimality2	序列测试 $G \rightarrow S_1 \rightarrow S_2 \cdots$	单维信息量选题策略	每个维度均达到了事先界定的信息量	多维能力估计方法
MBCAT	传统 MBCAT	G 与 S 同时兼顾	D 优化法	多维终止策略,多维信息矩阵逆矩阵对角线元素最小值达到事先界定的要求	多维能力估计方法
	MBCAT–optimality1	G 与 S 同时兼顾	D 优化法	单维终止策略: 每个维度均达到了事先界定的信息量	多维能力估计方法
	MBCAT–optimality1	G 与 S 同时兼顾	D 优化法,但当某个维度达到事先界定要求,则随后不选择测量了该维度的题	单维终止策略: 每个维度均达到了事先界定的信息量量	多维能力估计方法

3.2　评价指标

1) 能力估计精度指标. 均方根误差(RMSE) : 能力估计值和真值之间均方根误差,其值差异越小,估计精度越高.

$$R_{MSE} = \sqrt{\sum_{j=1}^{N} (\hat{\theta}_k - \theta_k)^2 / N}.$$

2) BCAT 测试效率指标. 根据以往研究,本文主要采用被试使用的平均题长(Max_ Length) 作为测验效率(Test Efficiency,TE) 指标,即在相同精度下,平均使用的题目量.

3) 题库曝光率指标. 采用卡方指标(χ^2) 和测验重叠率(TOR) 来反应题库曝光率,前者越大或后者越小说明题库的曝光率越高.

3.3　蒙特卡洛模拟

在本研究中模拟的题库大小为 300 题,题库的结构为双因子模型,其中特殊因子 5 个,一般因子 1 个,共 6 个能力维度. 所有项目测量了一般因子(G) ,但每题只测量了 5 个特殊因子中的 1 个. 共模拟产生 300 题,每个特殊因子均被 60 题测量,项目计分方式为 0-3 的 4 级评分. 采用 F. Samejima[8] 的多维等级反应模型(MGRM) ,其项目反应函数为 $P_{it} = P_{it}^* - P_{it-1}^*$, $P_{it}^* = 1/(1 + \exp(-D\sum_{k=1}^{m} a_{ik}(\theta_{jk} - b_{il})))$, $b_i \sim N(0,1)$,且 $b_{i1} < b_{i2} < b_{i3}$,题目区分度对数标准正态分布生成,即 $a_i \sim \log N(0,1)$. 被试能力真值从独立的多元标准正态分布中生成.

3.4　选题策略

在 UBCAT 中,选题策略采用最大 Fisher 信息量法,在多维等级反应模型下的计算公式为

$$I_i(\theta) = 1.702^2 a_i^2 \sum_{t=0}^{n} (P_{it}^*(\theta_j) - P_{i,t+1}^*(\theta_j))(1 - P_{it}^*(\theta_j) - P_{i,t+1}^*(\theta_j))^2.$$

在 MBCAT 中,借鉴 D. G. Seo 等[6]的做法,选题策略采用常用的 D-优化方法,即选择那些使测验的 Fisher 信息量矩阵行列式达到最大的题目,计算公式为

$$i_n \equiv \arg\max_j \{\det(I_{sn-1}(\hat{\boldsymbol{\theta}}^{n-1}) + I_j(\hat{\boldsymbol{\theta}}^{n-1}))\}, j \in R_n\},$$

其中 $\hat{\boldsymbol{\theta}}^{n-1}$ 为根据已经施测过的 $n-1$ 题估计出的特质水平向量; $I_{sn-1}(\hat{\boldsymbol{\theta}}^{n-1})$ 为已经施测的 $n-1$ 个题目在 $\hat{\boldsymbol{\theta}}^{n-1}$ 处的信息量; $I_j(\hat{\boldsymbol{\theta}}^{n-1})$ 为剩余题库中题目在 $\hat{\boldsymbol{\theta}}^{n-1}$ 处的信息量.

4 研究结果

4.1 不同 BCAT 设计下被试能力估计精度比较

表 2 是不同 BCAT 设计下能力参数估计精度指标(RMSE). 由表 2 可以看出,本文提出的 4 种 BCAT 设计,不论是一般能力因子 G 还是特殊能力因子 S,能力参数估计精度均高于传统的 UBCAT 和传统的 MBCAT 设计. 本文提出的 4 种优化设计中,能力参数估计精度最高的是单维信息量终止的 MBCAT(MBCAT_optimality1),其次是带维度约束单维信息量终止的 MBCAT(MBCAT_optimality2),再次是重新多维估计的 UBCAT(UBCAT_optimality1)以及单维选题多维估计的 UBCAT(UBCAT_optimality2). 这说明基于多维的 BCAT(MBCAT)设计比基于单维的 BCAT(UBCAT)设计在参数估计精度上更具优势. 在 UBCAT 的 3 种设计中,相比于只进行单维估计(传统 UBCAT 设计),采用多维估计(UBCAT_optimality1 和 UBCAT_optimality2)不仅能够提高一般因子上的参数估计精度,同时还能够提高特殊因子上的能力估计精度. 同时,在传统 UBCAT 设计与 UBCAT_optimality1 设计中,2 者使用的题目是完全一样的,唯一不同的是 UBCAT_optimality1 设计只是在传统 UBCAT 设计的基础上,用已经选出的题目重新再估计一次,因此不存在题目长度不同而导致的精度不同的情况. UBCAT_optimality1 和 UBCAT_optimality2 都是属于单维选题、多维估计的类型,但是不同之处在于 UBCAT_optimality2 是在自适应过程一开始就采用了多维估计,而 UBCAT_optimality1 则是在自适应过程中采用单维估计,等所有的题目满足了标准之后才采用多维估计,2 种方法的能力估计精度都比较接近,这说明无论是在自适应过程中还是在自适应结束之后采用多维估计方法都能够提高能力估计精度.

表 2 不同 BCAT 设计下下 RMSE 指标比较

类型	条件方法	G	S_1	S_2	S_3	S_4	S_5	S 平均	G 和 S 平均
BCAT	传统 UBCAT	0.501	0.622	0.551	0.568	0.531	0.556	0.566	0.555
	UBCAT_optimality1	0.392	0.449	0.396	0.406	0.399	0.411	0.411	0.409
	UBCAT_optimality2	0.403	0.453	0.401	0.417	0.402	0.415	0.417	0.415
MCAT	传统 MBCAT	0.389	0.461	0.449	0.449	0.452	0.448	0.452	0.441
	MBCAT_optimality1	0.357	0.418	0.376	0.380	0.377	0.385	0.386	0.382
	MBCAT_optimality2	0.375	0.437	0.394	0.402	0.398	0.400	0.405	0.401

4.2 不同 BCAT 设计下题库曝光率比较

题库的曝光率结果见表 3. 从表 3 可看出,在 UBCAT 下,由于传统 UBCAT 设计与 UBCAT_optimality1 使用的测验项目是完全一样的,因此,这 2 种方法在题库使用情况上是完全相同的. 就其他 BCAT 设计而言,传统 MBCAT 设计的题库使用中具有最小的验重叠率(TOR)和 χ^2 等曝光指标,相比较而对题库的使用最为均匀,其余 BCAT 设计的题库使用情况指标基本接近,但总体来讲基于 MBCAT 的设计在题库的使用上略优于基于 UBCAT 的设计.

表 3 不同 BCAT 设计下题库使用指标比较

类型	条件方法	χ^2	T_{OR}	$E_{R\_\min}$	$E_{R\_\max}$
MCAT	传统 UBCAT	107.185	0.446	0.005	0.914
	UBCAT_optimality1	107.185	0.446	0.005	0.914
	UBCAT_optimality2	112.826	0.466	0.003	0.932
MCAT	传统 MBCAT	80.512	0.411	0.016	0.920
	MBCAT_optimality1	96.187	0.457	0.007	0.959
	MBCAT_optimality2	108.644	0.455	0.002	0.950

4.3 不同 BCAT 设计下测验效率的比较

评价 CAT 效率的 1 个重要的指标就是被试平均使用的测验长度,即 TE 指标,结果如表 4. 从表 4 可知,对于 UBCAT 3 种设计,无论是否采用多维估计各维度特质水平,测验效率基本相当且都比较高(TE 指标低),即平均题目数量最少;而对于 MBCAT 的 3 种设计,带有维度约束单维信息量终止的 MBCAT(MBCAT_optimality2)平均题目长度也比较短,

和 UBCAT 的平均题非常接近. 主要原因在于 MBCAT_optimality2 限定,凡是某个维度达到事先界定的信息量则随后的不再选择含有该维度的项目,这一点与 UBCAT 的设计是相同的,因此与 UBCAT 设计在题目使用数量上比较接近. 而传统的 BMCAT 设计与 MBCAT_optimality1 没有"凡是某个维度达到事先界定的信息量则随后的不再选择含有该维度的项目"这一限定,即已满足精度条件的维度的题目不进行控制,那么就有可能会使得一部分已经满足条件的维度的题目被继续选择进行测试,从而使得测验的长度变长,测验效率降低,因此就出现表 4 中传统的 BMCAT 设计与 MBCAT_optimality1 的测验效率较低. 同时总体来看,传统的 MBCAT 设计是所有 6 种 BCAT 设计中平均使用题目量最大,因此测验效率相对最低.

表 4　不同 BCAT 设计下测验效率指标比较

类型	条件方法	平均题长(N)	N_{min}	N_{max}
UBCAT	传统 UBCAT	26.810	19	136
	UBCAT_optimality1	26.810	19	136
	UBCAT_optimality2	26.891	20	86
MBCAT	传统 MBCAT	42.782	5	300
	MBCAT_optimality1	40.937	20	300
	MBCAT_optimality2	28.049	18	92

5　结论与讨论

本研究在传统 BCAT 2 种测验设计的基础上,提出了 4 种新的 BCAT 设计,并采用国际上通用的 Monte Carlo 模拟实验的方式,从能力参数估计精度、题库使用的曝光率及测验的效率等 3 大指标来验证新提出的 4 种 BCAT 设计,并同时与传统的 BCAT 2 种设计进行比较. 模拟研究与实证应用研究结果表明: 本研究新提出的 4 种 BCAT 设计在能力参数估计精度普遍优于 2 种传统的 BCAT 设计,体现新方法的优越性. 在题库使用率或曝光率方面,基于 MBCAT 的设计在题库的使用上略优于基于 UBCAT 的设计,整体来看,传统的 MBCAT 及本文提出的 MBCAT_optimality1 在曝光控制上最优; 在测验效率方面,基于 UBCAT 3 种设计的平均使用题长基本相当,而基于 MBCAT 的设计中本文提出的 2 种新 MBCAT 设计优于传统的 MBCAT,整体比较而言,基于 UBCAT 的测验效率优于基于 MBCAT 的测验效率.

1) 在 UBCAT 设计下,不同 BCAT 测验设计的选用. 在 UBCAT 设计中: 本文提出的 UBCAT_optimality1 方法拥有最高的能力估计精度、最优的曝光控制和最优的测验效率,因此整体上是 UBCAT 设计中最优的设计,也是首推实际使用者使用的设计. 而考虑到传统的 BCAT 设计是所有设计中能力参数估计精度最差,虽然这 2 种方法下过度曝光题目数量不多,测验效率上也有一定的可取之处. 但在 BCAT 的实际应用中,需要的是既能够高效地评估,更要能够准确评估的方法,因此,不推荐实际运用者选用传统的 BCAT 设计. 当然,这也从另一个侧面说明本研究的必要性与重要性.

2) 在 MBCAT 设计下,不同 BCAT 测验设计的选用. 在 MBCAT 设计中: 本文提出的 MBCAT_optimality1 方法拥有最高的能力估计精度、次高的曝光控制和次高的测验效率,整体上是 MBCAT 设计中最优的设计,也是首推实际使用者使用的 BCAT 设计. 而传统的 MBCAT 与本文提出的 MBCAT_optimality2 各有优劣,前者最大的优点是曝光控制比较理想,但缺点是能力参数估计的精度稍差; MBCAT_optimality2 具有最优的测验效率,但缺点是曝光控制稍差.

限于时间及研究精力,本研究还有很多值得进一步研究及探讨的地方. 如本文未探讨 D. G. Seo 等[6] 在其研究中指出的不同因子结构下,本文新开发的 4 种测验设计的效果; 同时在 MBCAT 测验设计中,选题策略采用的是 D-优化法,未来还可以进一步探讨其他选题策略的效果,如基于贝叶斯的 D-优化方法[9]、互信息法[10] 等方法; 同时本研究 BCAT 的终止策略为不定长 CAT,定长的 BCAT 以后还有待深入.

6　参考文献

[1] Chen Fangfang,West S G,Sousa K H. A comparison of bifactor and second-order models of quality of life [J]. Multivariate Behavioral Research,2006,41(2):189-225.

[2] Gibbons R D,Weiss D J,Pilkonis P A,et al. Development

of a computerized adaptive test for depression [J]. American Journal of Psychiatry,2013,69(11):1104-1112.

[3] Gibbons R D,Weiss D J,Pilkonis P A,et al. Development of the cat-anx: a computerized adaptive test for anxiety [J]. American Journal of Psychiatry,2014,171(2):187-194.

[4] Weiss D J,Gibbons R D. Computerized adaptive testing with the bifactormodel [EB/OL]. [2018-06-12]. http://publicdocs. iacat. org/cat2010/cat07weiss&gibbons. pdf

[5] Segall D O. Multidimensional adaptive testing [J]. Psychometrika,1996,61(2):331-354.

[6] Seo D G,Weiss D J. Best design for multidimensional computerized adaptive testing with the bifactor model [J]. Educational & Psychological Measurement,2015,75(6):954-978.

[7] Wang Chun,Chang Huahua,Boughton K A. Deriving stopping rules for multidimensional computerized adaptive testing [J]. Applied Psychological Measurement,2013,37(37):99-122.

[8] Samejima F. Graded response model [M] // van der Linden W J,Hambleton R K. Handbook of modern item response theory. New York: Springer-New York Press,1997:85-100.

[9] Mulder J,van der Linden W J. Multidimensional adaptive testing with optimal design criteria for item selection [J]. Psychometrika,2009,74:273-296.

[10] Mulder J,van der Linden W J. Multidimensional adaptive testing with Kullback-Leibler information item selection [EB/OL]. [2018-09-16]. doi: 10. 1007/978-0-387-85461-8.

The Optimization of Testing Design for CAT with Bifactor Model and Its Application

LIU Xinting,PENG Siwei,TU Dongbo*

(College of Psychology,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Four new type of testing designs of computerized adaptive testing with bifactor model (BCAT) has been proposed on the basis of two traditional testing designs for BCAT. Two proposed optimality testing designs belong to the unidimensional BCAT,which are called as UBCAT_optimality1 and UBCAT_optimality2,respectively. Another two proposed optimality testing designs belongs to the multidimensional BCAT,which are called as MBCAT_optimality1 and MBCAT_optimality2,respectively. Results showed that: (i) The proposed four optimality designs for BCAT overall had higher parameter estimation precision of both general factor and special domain factor,than two exiting designs for BCAT. (ii) As for item bank exposure rate,the MBCAT designs were better than the UBCAT designs. The proposed MBCAT_optimality1 and the exiting MBCAT performed best in item exposure control. (iii) On test efficiency,the UBCAT designs used fewer items than those of the MBCAT designs.

Key words: bifactor model; computerized adaptive testing; BCAT; polytomously score

(责任编辑: 冉小晓)