李佳,况天昊. 结合项目反应时间与项目区分度的 CAT 选题新策略[J]. 江西师范大学学报(自然科学版),2023,47(4):377-383.
LI Jia, KUANG Tianhao. The new item selection strategies in CAT in combination with item response time and item discrimination[J]. Journal of Jiangxi Normal University (Nature Science),2023,47(4):377-383.

文章编号:1000-5862(2023)04-0377-07

结合项目反应时间与项目区分度的 CAT 选题新策略

李 佳[1],况天昊[2]
(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 江西科技学院信息工程学院,江西 南昌 330098)

摘要:计算机化自适应测验(CAT)的测量有效性不仅在于测验的项目数量,而且还在于被试完成测验所花费的时间. 该文提出的结合项目反应时间和项目区分度动态分层的选题新策略是一种连续升 a 降 β 的选题方法,该方法在保证测验精度的同时,不仅有效地降低了被试的测验时间,而且还提高了题库的利用率. 蒙特卡罗(Monte Carlo)模拟实验结果表明:新选题方法在测验精度、测验时间有效性、题库利用率和测验安全性等评价指标中总体表现良好.

关键词:计算机化自适应测验;选题策略;项目反应时间;项目区分度;项目曝光控制
中图分类号:B 841 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2023.04.07

0 引言

计算机化自适应测验(computerized adaptive testing, CAT)是根据被试能力水平挑选合适的、信息量丰富的项目给被试作答的“量体裁衣”式的测验,其最大的优点是采用比传统测验更少的试题,花费最少的时间,但能得到对被试能力更准确的估计[1]. CAT 在美国研究生入学考试(GMAT)、美国医生护士资格考试(NCLEX)以及美国职业能力测验(ASVAB)中得到了广泛的应用[2]. 理论上,CAT 在项目挑选上单纯依赖最大化 Fisher 信息量的试题,就可以得到最大的测验精度. 但是,CAT 的测量有效性不仅和测验的项目数量有关,而且还和完成该测验所花费的总时间有关. 测验花费的总时间是指在测验中每个项目作答时间(项目反应时间)的总和. 项目反应时间(item response time, IRT)是指在测验过程中被试完成每个项目所花费的时间. 被试的项目反应时间可以为被试提供较多的测验信息:1)通过项目反应时间,在限制时间的测验中可以区分被试的解题速度和被试真实能力水平;2)通过项目反应时间,可探讨被试作答速度和作答准确率的关系;3)通过项目反应时间,在测验过程中可区分被试的异常反应模式;4)通过项目反应时间,可提高 CAT 项目选择的有效性;5)通过项目反应时间,可比较不同被试在测验中采用的答题策略[3]. 因此,近几年很多学者开始研究如何利用测验的时间信息来估计被试的能力. 在项目反应时间模型方面,根据被试的作答反应和项目反应时间之间的假设关系,可将项目反应时间模型分为 4 类:1)假设项目反应时间与被试作答反应具有条件独立性的对数正态模型和 Box-Cox 正态模型;2)假设项目反应时间和被试作答反应具有依赖性的 Thissen 模型和 4 参数逻辑斯蒂克(Logistic)模型;3)假设项目反应时间和被试作答反应兼具独立性和依赖性的层次模型、层次线性转换 IRT 模型和半参数反应模型;4)强调个体认知过程的竞赛模型、Q-Diffusion 模型和认知过程模型. 在这些模型中,因为对数正态分

收稿日期:2023-03-12
基金项目:国家自然科学基金(61967009, 62067004, 62267004, 62067005)资助项目.
作者简介:李 佳(1979—),女,江西南昌人,副教授,主要从事计算机辅助教学和心理测量方面的研究. E-mail:1276676143@qq.com

布结构简单,且具有良好的统计特性,所以对数正态项目反应时间模型(a lognormal model for response times,LM)的应用最为广泛[3]. 在结合项目反应时间的 CAT 选题策略方面,Fan Zhewan[4] 等在最大信息量选题策略(maximum information,MI)的基础上提出了单位时间内最大化信息量的选题策略(MI with time,MIT). 该方法在测验全程中都考虑了时间优化,始终选择信息量高且耗时短的项目给被试作答. 与 MI 方法相比,MIT 方法极大缩短了测验时间,但是由于选题策略的约束使得选题范围进一步缩小,所以项目曝光率比 MI 方法更高,从而导致测验安全性受到威胁. 为了应对这一问题,Fan Zhewen 等[4] 提出了按 a 分层 b 分块在单位时间内最大化信息量的 AST 选题策略,在将题库先按项目难度 b 分块再按项目区分度 a 分层之后,在每一层项目中采用 MIT 选题策略选择项目给被试作答. 该方法能较好地平衡项目曝光,但是测验精度有所下降[5]. 为了改善 MIT 方法在项目曝光方面的缺陷,E. M. Choe[6] 提出了按项目时间强度 β 分层的 BMIT 选题策略,将题库按项目时间强度从低到高进行分层,在每一层中采用 MIT 选题策略; 该方法能维持较高的测量精度,但是在分层数较少时只能略微平衡项目曝光[5]. 另外,上述 2 种固定分层方法在题库发生改变时,只有将题库重新分层才能使用,增大了测验时间,而且分层数和每层选择项目的数量都没有明确的标准,这也给实际应用带来较多的不确定性.

1 项目反应理论

1.1 3PLM 模型简介

在项目反应理论(item response theorem,IRT)框架下,最常见的是 3 参数逻辑斯蒂克模型(3PLM),能力为 θ_i 的被试正确作答第 j 个项目的概率为 $P_{ij}=c_j+(1-c_j)/(1+\exp(-Da_j(\theta_i-b_j)))$,其中 a_j 为项目 j 的区分度参数,b_j 为项目 j 的难度参数,c_j 为项目 j 的猜测度参数,$D=1.7$. 能力为 θ_i 的被试完成第 j 个项目的 Fisher 信息量为 $I_j(\theta_i)=D^2a_j^2(1-c_j)/((c_j+e^{Da_j(\theta_i-b_j)})(1+e^{-Da_j(\theta_i-b_j)})^2)$,笔者注意到,由于项目信息量与项目的区分度的平方成正比,所以项目信息量与项目区分度密切相关. 根据局部独立性假设,被试的测验信息量为 $I(\theta_i)=\sum_{j=1}^{L}I_j(\theta_i)$,其中 L 为测验长度.

1.2 对数正态项目反应时间模型(LM)简介

在众多项目反应时间模型中,因为对数正态模型[7] 不需要积分就可以估计出被试的期望反应时间,且该模型简单易懂,所以该模型得到广泛的应用. 给定被试 i 潜在的速度参数 τ_i,它和被试 i 的潜在能力 θ_i 密切相关且服从 2 维正态分布[8],定义被试 i 对项目 j 的反应时间 T_{ij} 的概率密度函数为 $f(t_{ij}|\tau_i)=\alpha_j e^{-(\alpha_j(\ln t_{ij}-\beta_j+\tau_i))^2/2}/(t_{ij}\sqrt{2\pi})$,其中 α_j 和 β_j 分别是项目 j 的时间区分度参数和时间强度参数. 按对数正态分布的标准形式重新写出 T_{ij} 概率密度函数的变形式为 $f(t_{ij}|\tau_i)=e^{-(\log t_{ij}-(\beta_j-\tau_i))^2/(2(1/\alpha_j)^2)}/(t_{ij}\cdot\sqrt{2\pi(1/\alpha_j)^2})$,显然,这是参数 $\mu=\beta_j-\tau_i$、$\sigma^2=1/\alpha_j^2$ 的对数正态分布,因此被试 i 对项目 j 的反应时间 T_{ij} 服从对数正态分布: $T_{ij}|\tau_i\sim\log(N(\beta_j-\tau_i,1/\alpha_j^2))$. 又因为对数正态随机变量的数学期望为 $e^{\mu+\sigma^2/2}$,所以被试 i 对项目 j 的期望反应时间,即平均反应时间为 $E(T_{ij}|\tau_i)=e^{\beta_j-\tau_i+1/(2\alpha_j^2)}$. 当项目 j 具有低项目时间强度 β_j 时,可以得到更低的 $E(T_{ij}|\tau_i)$ 值. 因此,项目反应时间和项目时间强度密切相关.

2 结合项目反应时间的选题策略

2.1 在单位时间内最大化 Fisher 信息量 MIT 选题策略

为了缩短测验时间和提高测量的有效性,Fan Zhewen 等[4] 将测验的 Fisher 信息量和项目期望反应时间 $E(T_{ij}|\tau_i)$ 相结合,根据项目 Fisher 信息量和被试 i 对项目 j 的期望反应时间的比值,在剩余题库中选择具有最大 $I_{MIT_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)/E(T_{ij}|\tau_i)$ 取最大值的项目给被试作答. 该方法仅牺牲一点能力估计准确性,就可以极大降低测验的平均完成时间,能够做到在减少测验时间的前提下保证能力估计的准确性. 但是很明显,MIT 方法在选题时更喜欢挑选高信息量 $I_j(\hat{\theta}_i)$ 和低 $E(T_{ij}|\tau_i)$ 值的项目,也就是说,高区分度、低时间强度的高质量项目更受欢迎. 双

重约束使得在题库中的某些项目被不同的被试频繁选用,对题库的安全性造成极大威胁. 又因为高信息量和低项目反应时间这 2 个条件的共同制约,所以 MIT 方法甚至比最大信息量选题方法(MI) 的项目曝光率更高,给整个 CAT 带来较高的风险[7].

2.2　在 a 分层 b 分块下的在单位时间内最大化 Fisher 信息量 AST 选题策略

a 分层 b 分块方法能有效控制项目的过度曝光,它先将题库按项目难度参数从小到大排列,每 K 个项目分为 1 块(其中 K 为固定参数,在各模拟实验中一般取 $K=4$) ,最后一块不要求有 K 个项目,再将在每块中的 K 个项目按项目区分度从小到大进行排列,然后在每一块中取出第 1 个项目构成第 1 层项目集,在每一块中取出第 2 个项目构成第 2 层项目集,依次类推,共得到 K 层项目集,这就完成了对题库的 a 分层 b 分块. 在分层后每一层项目的难度参数 b 的分布情况和整个题库的难度参数分布情况相一致,并且在这 K 个项目集中区分度参数 a 呈递增趋势. 在题库中的项目来自且唯一来自某一层,各项目层之间没有交集,各项目层的项目的并集正好是原题库. 因此,a 分层 b 分块是对题库项目集合的一个划分. 通过模拟实验发现: a 分层 b 分块的方法虽然可以较好平衡项目曝光,但是仍存在 50%以上过低曝光的项目. 因为在每层中仍然是调用高信息量的项目给被试作答,所以大量低区分度的项目还是不能被较好地利用,这造成在题库中低区分度项目的浪费,并且该方法的测验精度比 MIT 方法的测验精度有所下降.

2.3　按项目时间强度分层的在单位时间内最大化 Fisher 信息量 BMIT 选题策略

首先,将在题库中的项目按照项目时间强度由低到高分为 K 层(其中 K 为固定参数,在各模拟实验中一般取 $K=4$) ; 然后,在每一层内采用 MIT 选题策略选择固定数量的项目或满足固定信息量的项目给被试作答. 该分层方法平衡了在选题时的时间强度范围,降低了项目的总曝光率. 但通过模拟实验发现,依然存在大量过低曝光的项目. 这是因为在每层中总是调用低项目反应时间的项目,而大量较高项目反应时间的项目未被利用,这造成题库利用率不高.

2.4　结合项目反应时间和项目区分度的动态分层选题新策略 NMIT

事实上,在单位时间内最大化信息量既和项目的 Fisher 信息量有关,也和项目的期望反应时间有关. 和项目 Fisher 信息量密切相关的是项目区分度 a,和项目期望反应时间密切相关的是项目时间强度 β. 因此,本文提出将项目区分度和项目时间强度相结合,对题库进行动态分层的 CAT 选题新策略. 该方法通过有效控制项目区分度和项目时间强度的取值,同时兼顾到测验精度、测验时间和项目曝光等方面的 CAT 测验要求.

2.4.1　**定长测验**　设测验长度为 L,在剩余题库中挑选具有 $I_{\mathrm{NMIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)(\beta_j/a_j)^{(1-L(i)/L)}/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答(若有多个项目同时达到最大值,则随机挑选一个给被试作答) ,其中 $L(i)$ 表示被试 i 的当前测验长度. 这是一种动态的分层方法,在测验初期,被试的能力估计还不够准确,这时使用高时间强度、低区分度的项目对测验精度和测验时间的评估影响不大. 随着测验的进行,$L(i)/L$ 将逐渐变大,而 $1-L(i)/L$ 将逐渐变小. 而在测验后期,该选题策略逐渐接近 MIT 方法,挑选的是低时间强度且高区分度的高质量项目,不仅可以缩短测验时间、提高测验效率,而且还可以提高测验精度,被试能力估计也更准确. 另外,在测验初期,强行调用高时间强度、低区分度的项目,使得项目选择的范围更广,可均衡项目曝光率,提高项目利用率,题库更安全有效,而且动态分层选题策略不需要事先对题库进行分层,也可节约测验时间.

2.4.2　**不定长测验**　设测验信息量为 I_{Inf},在剩余题库中挑选具有

$$I_{\mathrm{NMIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)(\beta_j/a_j)^{(1-I_{\mathrm{Inf}}(i)/I_{\mathrm{Inf}})}/E(T_{ij}\mid\tau_i)$$

取最大值的项目给被试作答,其中 $I_{\mathrm{Inf}}(i)$ 表示被试 i 的当前测验信息量.

总体来说,新的选题策略实质上是一种升 a 降 β 的方法,并且是一种连续的升 a 降 β 的过程. 由于项目时间强度参数和项目区分度参数相互制约,所以在整个测验过程可以调用大量中等质量的项目(包括高区分度低时间强度的项目、低区分度高时间强度的项目和中等区分度和中等时间强度的项目) ,提高在题库中项目的过低曝光率,从而提高题库的利用率.

3 比较不同选题策略的 CAT 表现

3.1 被试及题库模拟

为了便于对比试验结果,本文所有试验模拟条件同文献[8].

1) 蒙特卡罗模拟产生 1 000 个被试,被试的能力值 θ_i 和潜在速度 τ_i 密切相关且服从 2 维正态分布,其参数为 $(\theta_i,\tau_i)\sim N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$,其中均值矩阵和协方差矩阵分别为

$$\boldsymbol{\mu}_2=(0,0)^{\mathrm{T}},\boldsymbol{\Sigma}_2=\begin{pmatrix}1.00 & 0.20\\ 0.20 & 0.16\end{pmatrix}.$$

2) 用蒙特卡罗方法模拟生成题库 500 个项目且满足条件 $(a_j^*,b_j,\beta_j)\sim N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$(3 元正态分布),$c_j\sim\beta(2,10)$(贝塔分布),$\alpha_j\sim U(2,4)$(均匀分布),其中 $a_j^*=\ln a_j$,均值矩阵和协方差矩阵分别为

$$\boldsymbol{\mu}_1=(0.3,0.0,0.0)^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}_1=\begin{pmatrix}0.10 & 0.15 & 0\\ 0.15 & 1.00 & 0.25\\ 0 & 0.25 & 0.25\end{pmatrix}.$$

题库的项目数据如表 1 所示.

表 1 题库的项目数据

项目数据	区分度 a_j	难度 b_j	猜测度 c_j	时间区分度 α_j	时间强度 β_j
平均值	1.349 8	0.000 4	0.166 6	3.000 2	0.999 7
标准差	1.371 9	0.996 9	0.103 3	0.577 4	0.555 6

3.2 模拟 CAT 的施测过程

本文不考虑内容平衡[9]和机会红利[10]对 CAT 的影响,并设被试的能力初值为 0,采用 NMLE 方法[11]对被试能力进行估计.

采用定长和不定长 2 种测验形式. 定长测验的测验长度为 40,在分层选题策略中将题库分为 4 层($K=4$),每层选出 10 题. 不定长测验是当被试所测项目累积信息量达到 16 时结束,当在分层选题策略中每层信息量达到 4 时退出.

3.3 参与比较的 6 种选题策略

1) 最大信息量选题策略(MI):在剩余题库中挑选被试当前能力下具有最大项目信息量 $I_j(\hat{\theta}_i)$ 的项目给被试作答. 这是各种选题策略测验准确性(或测验精度)的比较标准.

2) 随机化选题策略(RS):在剩余题库中随机挑选一个项目给被试作答. 该方法题库利用率最高,但测量精度最差,可以作为各种选题策略测验安全性的比较标准.

3) 在单位时间内最大化 Fisher 信息量的选题策略(MIT):从剩余题库中挑选具有 $I_{\mathrm{MIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答. 该方法可以作为各种选题策略测验时间的比较标准.

4) a 分层 b 分块的单位时间内最大化 Fisher 信息量的 AST 选题策略:在将题库按 a 分层 b 分块后,在每层中挑选具有 $I_{\mathrm{MIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答.

5) 基于项目时间强度的单位时间内最大化 Fisher 信息量的 BMIT 选题策略:在将题库按时间强度分层后,在每层中挑选具有 $I_{\mathrm{MITj}}(\hat{\theta}_i)=I_j(\hat{\theta}_i)/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答.

6) 结合项目反应时间和项目区分度的动态分层选题新策略 NMIT:

(a) 在定长试验中,从剩余题库中挑选具有 $I_{\mathrm{NMIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)(\beta_j/a_j)^{(1-L(i)/L)}/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答;

(b) 在不定长测验中,从剩余题库中挑选具有 $I_{\mathrm{NMIT}_j}(\hat{\theta}_i)=I_j(\hat{\theta}_i)(\beta_j/a_j)^{(1-I_{\mathrm{In}j}(i)/I_{\mathrm{In}j})}/E(T_{ij}\mid\tau_i)$ 取最大值的项目给被试作答.

3.4 评价指标

第 1 组采用被试能力估计的绝对离差和均方根误差来评价能力估计的准确性:

$$e_{\mathrm{ABS}}(\hat{\theta})=\sum_{i=1}^{n}|\theta_i-\hat{\theta}_i|/n,$$

$$e_{\mathrm{RMSE}}(\hat{\theta})=\sqrt{\sum_{i=1}^{n}(\theta_i-\hat{\theta}_i)/n}.$$

第 2 组采用卡方检验统计量和项目过低曝光率[12]来评价题库的安全性和题库的利用率:

$$\chi^2=\sum_{j=1}^{M}\frac{(e_{r_j}-L/M)^2}{L/M},\ L_E=\sum_{j=1}^{M}L_{E_j}/M.$$

第 3 组采用平均测验时间和测验时间的标准差,来测量被试测验时间的有效性和稳定性:

$$\bar{t}=\sum_{i=1}^{n}t_i/n,s_t=\sqrt{\sum_{i=1}^{n}(t_i-\bar{t})^2/(n-1)},$$

其中 n 为被试总人数,θ_i 为第 i 个被试的能力真值,$\hat{\theta}_i$ 为第 i 个被试的能力估计值,M 为在题库中的项目数,e_{r_j} 表示第 j 个项目的曝光率,其值为项目 j 在不同被试测验中出现的次数除以题库项目数 M. 在定长测验中,L 表示测验长度,在不定长测验中,L 表示平均测验长度,L_{E_j} 用于标注题库中曝光率小于 0.05 的项目,即

$$L_{E_j}=\begin{cases}1, e_{r_j}\geqslant 0.05,\\ 0, e_{r_j}<0.05,\end{cases}$$

t_i 表示被试 i 完成测验所花费的时间.

能力估计的准确性表明了测验的准确性,能力估计的绝对离差和均方根越小表示能力估计越准确. 卡方统计量和项目过低曝光率反映了题库项目的曝光情况,卡方值和项目过低曝光率越小说明项目调用越均匀,题库利用率越高,CAT 的安全性也越好. 测验时间的平均值和标准差越小表明 CAT 测验时间更短,测验更有效也更稳定. 本文用采测验精度、测验时间有效性、题库安全性和题库利用率这 3 组评价指标对以上 6 种选题策略进行综合评价比较.

4　实验结果及其分析

当实验为定长测验时,实验结果如表 2 所示; 当实验为不定长测验时,实验结果如表 3 所示.

表 2　定长测验 6 种选题策略的表现

选题策略	$e_{\mathrm{ABS}}(\hat{\theta})$	$e_{\mathrm{RMSE}}(\hat{\theta})$	χ^2	L_E	$\bar{t}$/min	s_t/min
MI	0.140 7	0.158 1	101.682	0.649	104.112	46.805
RS	0.293 1	0.304 4	0.000	0.000	161.395	72.628
MIT	0.156 8	0.167 3	120.682	0.704	70.682	33.169
AST	0.206 5	0.214 9	68.412	0.543	85.455	46.866
BMIT	0.189 1	0.200 8	54.226	0.471	83.437	44.252
NMIT	0.169 2	0.175 4	19.610	0.126	76.455	36.407

从表 2 可以看出: 模拟实验在测验精度方面的最大信息量选题策略 MI 方法能力估计准确性最高,在单位时间内最大信息量选题策略 MIT 方法和选题新策略 NMIT 方法能力估计准确性相当,只比 MI 方法测验精度稍低,2 种固定分层选题策略 AST 方法和 BMIT 方法的测验精度会更差一点,但还是比随机选题策略 RS 方法的测验精度好很多. 在题库的安全性方面,表现最好的是 RS 方法,表现最差的是 MIT 方法,MI 选题策略是全程挑选高信息量的项目,而 MIT 选题策略是全程挑选高信息量且耗时短的项目,比 MI 方法要求更高,由于高区分度低项目反应强度的项目更被频繁选出,所以 MIT 方法的卡方值比 MI 方法的卡方值更大,项目过低曝光率也更高,2 种固定分层选题策略的卡方值虽然有所下降,但是项目过低曝光率还是比较大,均在 50% 左右,这说明题库利用率较低. 合理控制项目参数的新选题策略 NMIT 方法的卡方值和项目过低曝光率明显比前面提到的 3 种方法小很多,具有更好的测验安全性和题库利用率. 在测验时间有效性方面,结合项目反应时间的选题策略明显比传统的选题方法平均测验时间更短,测验时间的标准差也更小. 这是因为: 若固定测验时间,则结合项目反应时间的选题策略可以实施更多的项目,获得更多的测验信息. 因此,结合项目反应时间的选题策略比传统的选题策略更有效,也更稳定. 在这些含有项目反应时间的方法中 MIT 方法的平均测验时间最短,2 种固定分层的选题策略由划分题库需要花费时间而导致了测验时间的增加,而选题新策略 NMIT 方法是一种动态的分层方法,不需要额外花费时间划分题库. MI 方法在每次挑选信息量最大的项目的过程中也可能会挑选到费时的项目,平均测验时间更长; AST 方法和 BMIT 方法均衡了项目信息量和做

题时间,但是项目过低曝光率过高,题库利用率不高. 而选题新策略 NMIT 方法克服了 2 者的缺点,仅仅牺牲一点点测验精度就可以更有效地缩短测验平均时间,同时保证测验的安全性,降低题库的建设成本,取得了比较满意的结果.

表 3 不定长测验 6 种方法的表现

选题策略	$e_{ABS}(\hat{\theta})$	$e_{RMSE}(\hat{\theta})$	χ^2	L_E	$\bar{t}$/min	s_t/min
MI	0.162 2	0.170 7	89.167	0.669	100.850	45.338
RS	0.308 0	0.312 7	0.000	0.000	140.291	63.116
MIT	0.179 9	0.198 1	100.850	0.754	69.167	31.152
AST	0.261 0	0.273 6	53.261	0.580	80.943	42.623
BMIT	0.243 3	0.252 5	49.526	0.497	78.425	44.176
NMIT	0.181 8	0.190 3	18.373	0.205	71.425	32.176

从表 3 可以看出: 模拟实验结果和定长测验类似. 因为不定长测验的平均测验长度短于定长测验的测验长度,所以被试所采用的项目数量更少,测量精度总体小于对应的定长测验. 当然,不定长测验的平均测验时间和卡方值也都小于对应的定长测验. 不定长测验以统一的测验累积信息量为终止标准,为每个被试的能力水平估计值设定了相同的测量误差,保证了参与测试的每个被试最后获得相同信度的得分. 因此,不定长测验的测试结果更公平,在 CAT 中也是更推荐使用的测验终止规则[1].

5 讨论

因为在题库中项目参数既有系统参数又有时间反应参数,所以将被试做题时间和被试潜在能力结合在一起,可以花费更少的测验时间而得到相同的测验信息,因此本文重点研究了如何将项目反应时间模型应用于 CAT 的选题策略中. 测验反应时间可以为计算机自适应测验提供更多的测验信息,结合项目反应时间的选题策略可以避免选择那些非常耗时的项目,提高项目选择的效率,也更能突显计算机化自适应测验的优势. 现有的结合项目反应时间的 MIT 选题策略尽管平均测验时间最短,但是项目曝光率过大,试题的安全性得不到保障. 2 类固定分层的 AST 方法和 BMIT 方法尽管可以较好地控制高质量项目的过渡曝光,但是存在题库需要事先分层,题库的分层数目和每层选题数目均不确定,且当题库中的项目发生改变时需要重新划分题库,项目的低曝光率太高,题库的利用率不高等问题. 通过以上实验结果表明,结合项目反应时间和项目区分度的动态分层选题新策略 NMIT 方法具有以下优点: 1) 不需要事先对题库进行分层,节省测验时间; 2) 通过对题库的动态分层,提高项目的过低曝光率,提高中等质量项目的使用频率,保障题库的安全性和题库的利用率; 3) 简单灵活,可以直接应用于当前大规模的 CAT 场景中.

在今后的研究中还应该考虑如何将项目反应时间和内容平衡结合在一起,提出更符合应用实践的选题策略; 另外,如何将新选题策略应用于结合项目反应时间的多维 CAT 模型(multidimensional CAT,MCAT) 中也还需要做进一步的研究.

6 参考文献

[1] 漆书青,戴海崎,丁树良. 现代教育与心理测量学原理[M]. 北京: 高等教育出版社,2002: 154-155.

[2] VELDKAMP B P. On the issue of item selection in computerized adptive testing with reponse times [J]. Journal of Education Measurement, 2016, 53 (2): 212-228.

[3] 郭磊,尚鹏丽,夏凌翔. 心理与教育测验中反应时模型应用的优势与举例 [J]. 心理科学进展,2017,25(4): 701-710.

[4] FAN Zhewen, WANG Chun, CHANG Huahua, et al.

Utilizing response time distributions for item selection in CAT [J]. Journal of Educational and Behavioral Statistics, 2012, 37(5): 655-670.

[5] 郭治辰,汪大勋,蔡艳,等. 结合题目作答时间的计算机化自适应测验选题方法 [J]. 心理科学,2021,44(5): 1241-1248.

[6] CHOE E M. Controlling item exposure for response time-informed item selection in computerized adaptive testing [D]. Champaign: University of Illionois at Urbana-Champaign,2014.

[7] LINDEN W J. A lognormal model for response times on test items [J]. Journal of Educational and Behavioral Statistics,2006,31(2): 181-204.

[8] CHOE E M,JUSTIN L K,CHANG Huahua. Optimizing the use of response times for item selection in computerized adaptive testing [J]. Journal of Educational and Behavioral Statistics,2018,43(2): 135-158.

[9] 李佳,丁树良,方剑英. 基于平均数形式的选题策略比较 [J]. 江西师范大学学报(自然科学版),2015,39(1): 17-20.

[10] 李佳,丁树良. 多种分层方法在 CAT 校准误差中的应用研究 [J]. 江西师范大学学报(自然科学版),2016,39(1): 69-72.

[11] 李佳,丁树良. 计算机化自适应测验中能力估计新方法 [J]. 江西师范大学学报(自然科学版),2019,43(2): 111-115.

[12] CHENG Ying, JEFFREY M P, SHAO Can. *a*-stratified computerized adaptive testing in the presence of calibration [J]. Educational and Psychological Measurement, 2015, 75(2): 260-283.

The New Item Selection Strategies in CAT in Combination with Item Response Time and Item Discrimination

LI Jia[1], KUANG Tianhao[2]

(1. College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China;
2. Information Engineering College, Jiangxi University of Technology, Nanchang Jiangxi 330098, China)

Abstract: Measurement efficiency of computerized adaptive testing shall not only be assessed in terms of the number of items administered but also in terms of the time it takes to complete the test. New item selection strategy based on both item response time and item discrimination are proposed in the paper. The item discrimination is increasing and the item time intensity is decreasing during the whole test process. The results of Monte Carlo study show that new method has better performances on the test estimation accuracy, testing time efficiency, testing safety and the utilization of the item bank.

Key words: CAT; item selection strategy; item response time; item discrimination; item exposure control

(责任编辑: 冉小晓)