秦春影,吴龙月,王爱平. 计算机自适应测验中试题泄露的实时监控方法研究与应用[J]. 江西师范大学学报(自然科学版),2022,46(2):118-125.
QIN Chunying, WU Longyue, WANG Aiping. The study and application of real-time monitoring methods for item leakage in computerized adaptive tests [J]. Journal of Jiangxi Normal University(Natural Science),2022,46(2):118-125.

**文章编号**:1000-5862(2022)02-0118-08

# 计算机自适应测验中试题泄露的实时监控方法研究与应用

秦春影[1],吴龙月[1],王爱平[2*]
(1. 南昌师范学院数学与信息科学学院,江西 南昌 330032;2. 亳州学院电子与信息工程系,安徽 亳州 246800)

**摘要**:在连续施测下计算机自适应测验(CAT)中的试题被曝光的可能性急剧增加,因此需要对试题进行实时监控,当试题的参数发生显著性变化时必须将其进行强制"退休". 序贯监测程序(SMP)通过检测CAT中的试题统计特征的变化来判断试题是否泄露;然而在用SMP监控试题时会出现较大的Ⅰ类错误率,并且在一些条件下其统计检验力较低. 该文以残差的$R$指标作为考生拟合统计量(PFS),与SMP方法相结合,构建了一种新的监测方法(PFS_SMP);该方法以被试作答信息为依据判断被SMP标记的试题是否泄露,从试题和被试这2个层面保证测验的安全性和公平性. 最后,通过模拟实验和实证分析来对基于$R$的PFS_SMP的表现进行评价,实验结果表明:PFS_SMP方法能降低在SMP监测试题时的Ⅰ类错误,并能提高其统计检验力.

**关键词**:计算机自适应测验;被试拟合指标;序贯监测程序;残差;试题安全
**中图分类号**:B 814.7 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2022.02.02

## 0 引言

随着社会电子化的程度越来越高,信息安全问题与社会的每一个成员息息相关,已经渗透到日常的各个方面,对于测验领域也不例外. 计算机自适应测验(Computerized Adaptive Testing, CAT)[1-2]是测验发展的新方向,它能根据考生在试题上的作答情况,自适应按顺序安排其后的试题,真正实现了"因人施测",因而能够更精确地测量出被试的能力. 相对于纸笔测验,它有更多的优势,从效度、公平和安全等多个方面对测验进行保证,并且有着更短的测验长度和测试时间,大大提高了测验效率[3]. 自20世纪80年代起,CAT在美国教育测试中得到了广泛应用,如NCSBN(National Council of State Boards of Nursing)、GMAT(Graduate Management Admission Test)等. 20世纪90年代末,美国匹兹堡大学的研究者创建了世界上第1个自适应教学系统[4]. 而国内CAT的研究主要还是集中在理论上,在应用方面的研究相对较少,相关研究主要有杨业兵开发的中国士兵人格问卷[5]、涂冬波编制的小学儿童数学问题解决的诊断CAT系统[6]和邓远平编制的大学生人格内外向的CAT测验[7]等.

随着CAT在实际应用中的推广,试题泄露这一测验安全威胁也愈发突显. 在连续的施测(其含义是测验在多个时间点举行,比如测验在连续的一段时间内举行,考生可以选择合适的时间点参加测验)情况下,CAT中的试题被持续曝光的可能性急剧增加,对题库的安全性形成巨大的挑战. 美国的GRE考试曾出现过有组织的大规模盗题的情

**收稿日期**:2021-10-11
**基金项目**:江西省教育科学十四五规划2021课题(21YB257)资助项目.
**作者简介**:秦春影(1981—),女,安徽宿州人,副教授,主要从事教育大数据、教育统计与测量有关的研究. E-mail:qcy_qin@qq.com
**通信作者**:王爱平(1956—),女,甘肃庆阳人,教授,主要从事人工智能、信息安全等有关的研究. E-mail:710875983@qq.com

况[8-9]. 因此,需要对正在使用的试题进行实时监控,控制各个试题出现的曝光度[10-13]. 若试题的参数(如难度和区分度等)发生显著性的变化,则必须将试题进行强制“退休”. Zhang Jinming[14]提出了序贯监测程序(Sequentially Monitoring Procedure, SMP),用它来监测 CAT 中试题的统计特征在测试过程中是否产生变化,从而判断试题是否发生了泄露;但该方法会产生一定的虚报,并且统计检验力在一些条件下不高. 本文将在 SMP 的基础上引入基于残差的 $R$ 指标(Residual, $R$),并将它作为考生拟合统计量(PersonFitStatistic, PFS)[15],与 SMP 方法相结合提出了一种新的 PFS_SMP 方法,在 SMP 的基础上判断被标记的试题是否发生真正泄露. PFS_SMP 是基于自适应测验展开的,可以实现试题的在线实时检测,即在测验施测的过程中实时检测试题的统计指标,当发现异常时立即切换到安全题库下施测;该方法从试题和被试这 2 个方面去保证测验的安全性与公平性. 最后,通过模拟实验和实证分析来评价新方法的优势与不足.

## 1　序贯监测程序概述

当一个题库使用较长时间后,针对被使用多次的试题,需要对其进行监测. 记在考试过程中第 $j$ 个试题的得分序列为 $\{U_{1j},U_{2j},\cdots,U_{nj},\cdots\}$,其中 $n$ 是指作答该试题的第 $n$ 个考生,而非参加测试的第 $n$ 个考生. 若该考生答对该题,则 $U_{nj}=1$;否则 $U_{nj}=0$.

众所周知,考生在测验试卷上的作答过程是个随机过程,他在每个试题上的作答是个随机事件. 若试题的得分只有 2 种情况,则考生在该试题上的得分服从二项分布. 在试题没有被泄露的情况下,对于每一个考生,其正确作答某题的概率与考生自身的能力值和试题的难度、区分度以及猜测度相关,假定 $P(U_{ij}=1\mid\theta_i,a_j,b_j,c_j)$ 为考生 $i$ 正确作答第 $j$ 题的概率,$\theta_i$ 为该考生的能力值,$a_j$、$b_j$、$c_j$ 分别为该试题的区分度、难度、猜测度,则在施测过程中,该试题的得分序列 $\{U_{1j},U_{2j},\cdots,U_{nj},\cdots\}$ 应服从相同的分布.

若某试题在第 $n_k$ 个考生后发生泄露,则获得该试题信息的考生将会有新的正确作答概率 $P'(U_{ij}=1\mid\theta_i,a'_j,b'_j,c'_j)$,并且泄露后试题的参数会发生变化,显然 $P'(U_{ij}=1\mid\theta_i,a'_j,b'_j,c'_j)>P(U_{ij}=1\mid\theta_i,a_j,b_j,c_j)$. 而对于未获得该试题信息的考生,他们的作答概率仍服从 $P(U_{ij}=1\mid\theta_i,a_j,b_j,c_j)$. 在不引起误解的情况下,下文将 $P(U_{ij}=1\mid\theta_i,a_j,b_j,c_j)$ 和 $P'(U_{ij}=1\mid\theta_i,a'_j,b'_j,c'_j)$ 分别记为 $P(\theta)$ 和 $P'(\theta)$. 因此,对于在试题泄露后回答该试题的考生,他正确回答的概率有 2 种情况:$P(\theta)$ 或 $P'(\theta)$. 设考生得知该试题信息的可能性为 $e$,利用全概率公式可得在该试题泄露后任一考生答对题的概率为

$$P_1(\theta)=eP'(\theta)+(1-e)P(\theta),\qquad(1)$$

其中 $e$ 的大小是未知的,但它只有等于 0 和大于 0 这 2 种情况. 当 $e=0$ 时,该题没有泄露;当 $e>0$ 时,

$$P_1(\theta)-P(\theta)=e(P'(\theta)-P(\theta))>0,$$

即在试题泄露之后,该试题对于后面的所有考生都显得简单了. 而试题的泄露是未知的,泄露点 $n_k$ 的确定也很困难,不同试题的泄露点位置也不一样. 因此,监控方法很有必要对试题进行实时监控,在试题泄露后尽快将其甄别出来,以保证 CAT 的安全性和公平性.

SMP 于 2014 年被 Zhang Jinming[14] 提出,它的工作原理是:当试题泄露后,该试题的正确作答概率会有较明显提升,利用前后的正确作答概率的差异构造出假设检验的统计量,若统计量的值超过了预设的临界值 $c_\alpha$,则认为该试题已经泄露[8].

### 1.1　Ⅰ 类错误

假设被监控试题的泄露点是 $n_k$(即在第 $n_k$ 个考生后该题被泄露),而监测结果显示该题是在第 $n$ 个考生后被泄露,若 $n<n_k$(即监控程序显示在第 $n$ 个考生作答时该试题被泄漏,实际上此时该题并没有泄漏),则犯了 Ⅰ 类错误,这意味着监控程序给出了错误甄别(见图 1).

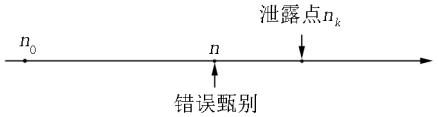


**图 1　Ⅰ 类错误**

### 1.2　Ⅱ 类错误

假设被监控试题的泄露点是 $n_k$(即在第 $n_k$ 个考生后该题被泄露),而监测结果显示该试题是在第 $n$ 个考生后被泄露,若 $n>n_k$(即监控程序显示在第 $n$ 个考生作答时该题被泄露,此时该试题确实泄露了,监控程序给出了正确甄别;但在第 $n_k$ 个考生到第 $n$ 个考生之间,该试题发生了泄露而监控程序却没有甄别出来),则犯了 Ⅱ 类错误,$[n_k,n]$ 是延迟区间(见图 2).

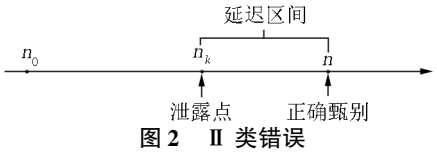


**图 2 Ⅱ 类错误**

### 1.3 统计量

在 CAT 中每道试题都有它潜在的目标考生子群体,试题难度较大的子群体能力比试题难度较容易的子群体能力更高,因此,定义 $p$ 是在某一被监控试题的考生目标子群体中某名考生的得分期望值,其计算公式如下:

$$p = E(U_n \mid \text{该试题目标考生子群体}) = \int_{-\infty}^{\infty} P(\theta) f(\theta)\,\mathrm{d}\theta,$$

其中 $p$ 是随机从考生目标子群体中选取的一个考生作答该试题的正确概率,$f(\theta)$ 是该试题的目标子群体的能力密度函数,因此可以用正确作答比例去估计 $p$ 的值. 在测试过程中,若试题没有泄露,则相应的考生作答该试题的 $p$ 值是一样的;若在第 $n_k$ 个考生后泄露,则前 $n_k$ 个考生 $p$ 值一样,第 $n_k$ 个考生后的考生的该试题得分期望值为 $p_1$,其计算公式如下:

$$p_1 = \int_{-\infty}^{\infty} P_1(\theta) f(\theta)\,\mathrm{d}\theta,$$

其中 $P_1(\theta)$ 由式(1)得到. SMP 的监控过程主要由一系列的假设检验构成,假设作答被监控试题的第 $n$ 个考生指的是当前考生,将到 $n$ 为止的所有考生的作答分为 2 个部分:前 $n-m$ 个考生作答为参考移动样本,第 $n-m+1$ 个考生到第 $n$ 个考生的作答为目标移动样本. 之所以被称为移动样本是因为监控过程是实时连续的,$n$ 会不断往后移,其中 $m$ 是移动样本量[7]. "序贯"的含义就体现在此,在一次次后移的过程中完成监测.

参考移动样本的正确作答概率估计值为

$$\hat{p}_{n-m} = \frac{1}{n-m}\sum_{j=1}^{n-m} U_j,$$

目标移动样本的正确作答概率估计值为

$$\hat{p}_{nm} = \frac{1}{m}\sum_{j=n-m+1}^{n} U_j.$$

若试题在第 $n$ 个考生作答时未泄露,则 $\hat{p}_{n-m}$ 和 $\hat{p}_{nm}$ 的差异不会很大;若试题在第 $n$ 个考生作答前已经泄露,尤其是在 $n-m$ 处泄露,则 $\hat{p}_{n-m}$ 和 $\hat{p}_{nm}$ 的差异会很大,因此可以依据这 2 个样本的正确作答概率估计值的差异来构造统计量,统计量被标准化后如下:

$$\hat{Z}_{nm} = (\hat{p}_{nm} - \hat{p}_{n-m})\sqrt{m(n-m)/n}\Big/\sqrt{\hat{p}_{n-m}(1-\hat{p}_{n-m})}. \tag{2}$$

SMP 监测方法的主要原理为假设在第 $n$ 个考生处试题未泄露,预先设定一个切分点 $c_\alpha$,利用构造的统计量与 $c_\alpha$ 作对比,若 $\hat{Z}_{nm} > c_\alpha$,则认为在第 $n$ 个考生处试题已经被泄露了.

SMP 监控试题的操作过程为:对于监控的每一个试题,当作答人数达到一定数量(如 100)时开启检测程序进行监测,起始点记作 $n_0$,当前作答的考生为第 $n$ 个考生,设置移动样本量长度为 50,利用 $\hat{p}_{n-m}$ 和 $\hat{p}_{nm}$ 构造出 $\hat{Z}_{nm}$,依据式(2)计算 $\hat{Z}_{nm}$,预先设置好切分点 $c_\alpha$ 的值,将 $\hat{Z}_{nm}$ 与 $c_\alpha$ 进行对比,若 $\hat{Z}_{nm} > c_\alpha$,则认为该试题已被泄露. SMP 由一系列的假设检验构成,对于在 CAT 中被监控的试题,当 $n(n > n_0)$ 往后每移一位时,就再一次计算 $\hat{Z}_{nm}$ 的值,将其与 $c_\alpha$ 做比较,进行显著性检验.

## 2 PFS_SMP 方法研究

### 2.1 考生拟合统计量

残差是在回归分析中的重要概念. 残差在数理分析统计中是指实际观察值与估计值(拟合值)之间的偏差;给定一个不可观测函数,它将自变量与因变量联系起来. 若对某些数据进行回归,则因变量的观测值与拟合函数值之间的偏差为残差. 残差在应用中蕴含的逻辑就是:通过对比理想情况(模型预测值)与实际情况(实际观察值)的差异来发现其中的异常情况. 预期偏差会使残差统计量膨胀,这与被试拟合检验的思想一致. 因此,这里提出基于残差的被试拟合统计量 $R$ 指标,它主要是检验被试在实际观察作答和期望作答概率之间的差异. 第 $i$ 个被试的拟合指标 $R_i$ 的计算方法为

$$R_i = \sum_{j=1}^{J} \log((u_{ij} - E(U_{ij} \mid \theta_i, a_j, b_j, c_j))/(P(u_{ij} \mid \theta_i, a_j, b_j, c_j)))^2, \tag{3}$$

其中 $J$ 为项目个数. 在实际情况下,考生的真实能力无法得到,因此本研究采用能力估计值,借助 Matlab 语言编程进行估计. 已知项目参数及被试作答数据,用期望后验估计(Expected A Posterior)[16] 方法得出能力估计值. $u_{ij}$ 表示被试 $i$ 在项目 $j$ 上的作答,$u_{ij} = 1$ 表示被试 $i$ 正确作答项目 $j$,$u_{ij} = 0$ 表示被试 $i$ 错误作答项目 $j$,$E(U_{ij} \mid \theta_i, a_j, b_j, c_j)$ 表示能力为 $\theta_i$ 的被试 $i$ 在项目 $j$ 上正确作答的期望概率,$P(u_{ij} \mid \theta_i, a_j, b_j, c_j)$

表示能力为 $\theta_i$ 的被试 $i$ 在项目 $j$ 上的正确作答概率.

在被试进行作答的过程中,并不能一直保持稳定的状态,偶尔会出现异常的状态,因此作答数据也会出现一些异常作答的情况,主要包含以下几种异常作答模式:

(i) 创造性作答(Creative Responding Behavior). 它是指由于被试以独特或具有创造性的方式作答这些试题,而错误地作答了简单的试题.

(ii) 随机作答(Random Guessing Responses). 它是指在测验动机较低下的情况下,被试凭猜测随机反应.

(iii) 疲劳(Fatigue). J. B. Sympson 等[11] 讨论的外部因素(如疲劳) 可能影响被试的反应,从而导致被试在考试快结束时的异常反应.

(iv) 睡眠(Sleeping Behavior). 根据 Zhang Jinming 等[9] 指出,睡眠是指由于被试对考试形式的混淆而在考试开始时表现不佳.

(v) 作弊(Cheating). 它是指能力较低的被试通过欺骗、抄袭等一系列手段使自己正确作答较难的试题.

(vi) 随机作弊(Cheating with Randomness). 它是指作弊的随机出现.

(vii) 预知试题 (Item Pre-knowledge). 它是指考生在考试之前了解试题的信息,这表明试题发生了泄露.

### 2.2　PFS_SMP 方法的基本原理

在了解有关残差应用的逻辑以及 $R$ 指标的含义和作用后, 再构建新的方法:PFS_SMP 方法. PFS_SMP 方法的基本原理主要为:假设题库共有 $J$ 个试题,有 $N$ 个考生,对考生作答过的试题做出相应的标记,分横纵 2 列记录数据,一列为 SMP 对试题进行的监控进程,即当试题被作答至一定次数时,可能会发生质量上的变化,用本文所介绍的 SMP 方法对试题进行检验,这一监测为题目水平的监测[8];另一列为考生作答的试题,通过结合 SMP 方法计算考生拟合统计量 $R$ 指标,检验考生实际的观察作答和期望作答概率之间的差异,以判断考生作答数据是否异常,这一监测为考生水平的监测[12]. 若某个试题被 SMP 判断为泄露,并且 $R$ 指标也显示对应考生的作答数据出现了异常,则认为该试题发生了泄露,并将其进行封存,不再用于之后的测试. 新构建的 PFS_SMP 方法是从试题和考生 2 个方面对题目进行监测,极大地保证了测验的安全性与公平性.

### 2.3　PFS_SMP 方法的操作步骤

新构建的 PFS_SMP 方法主要分为以下几个步骤:

(i) 当试题的曝光次数达到一定数量时,启动 SMP 对试题的质量进行相应监控,若试题被标记为“泄露”,则继续完成 CAT 测试;

(ii) 在测试结束后,统计该考生在作答的所有试题中被 SMP 标记为泄露的试题个数;

(iii) 用 SMP 方法与残差指标 $R$ 相结合,依据式(3) 计算考生的 $R$ 指标,即 PFS 值. 若 PFS 值超过预先设定的 $C_{\mathrm{PFS}}$(临界值),则把步骤(ii) 被标记为“泄露” 的试题进行封存,直到所有考生完成测验;

(iv) 统计被标记为泄露的题目和异常的考生数据,并计算各指标的值,对结果进行评价.

## 3　研究设计

实验研究采用的是 Monte Carlo 模拟方法,模拟被试的真实能力、测试的题库、作答数据以及泄题情况.

Monte Carlo 模拟也被称为随机抽样,是一种用随机数来解决计算问题的模拟研究方法. Monte Carlo 模拟的基本原理是:当一个问题具有概率特征时,可以通过计算机模拟生成抽样结果,并从抽样中计算统计值或参数;随着模拟次数的增加,可以通过对每个统计量或参数的估计平均值来得出稳定的结论.

### 3.1　模拟设计

为了检验 PFS_SMP 方法的表现,本文设计了实验进行对比研究:(i)SMP 方法,即用原来的序贯检测程序对题目进行监控;(ii)PFS_SMP 方法,即用 PFS_SMP 方法对被 SMP 标记为“泄露” 的试题进行判断后再决定是否封存. 通过比较这 2 种方法的表现来分析 PFS_SMP 方法的优势或劣势以及其对试题安全监控的效果.

模拟实验数据:

(i) 模拟被试,假设被试的能力参数服从标准正态分布,即 $\theta \sim N(0,1)$[11],并模拟 250 个被试.

(ii) 模拟项目,区分度 $b$ 和猜测度 $c$ 分别按 $b \sim U(0.5,1.5)$ 和 $c \sim U(0,0.25)$ 生成,低区分度题目按 $a \sim \log N(0,1)$ 模拟, 高区分度题目按 $a \sim \log N(0.5,1)$ 模拟,共模拟 100 个项目.

(iii) 模拟作答数据,共包含 2 种情况,

(a) 在题目未泄露时的正确作答概率,采用 3 参数 Logistic 模型生成,即

$$P(u = 1 \mid \theta) = c + (1 - c)/(1 + \mathrm{e}^{-Da(\theta - b)}),$$

其中 $\theta$ 是被试的能力值,$P(\theta)$ 是能力值为 $\theta$ 的被试的正答概率,$a$ 是试题的区分度参数,$b$ 是难度参数,$c$ 是试题的猜测参数. 若正确作答概率大于等于随机数,则 $u = 1$,否则 $u = 0$. $D$ 为量表常数.

(b) 接触到泄露试题后,用修正的 3 参数 Logistic 模型[17] 生成数据,即

$$P(u = 1 \mid \theta) = c' + (1 - c')/(1 + e^{-Da(\theta - b)}),$$

其中 $c'$ 为修正的猜测参数,且 $c' = p(m) + c - cp(m)$,$p(m)$ 是在考生事先了解题目信息时的作答概率,本文取 $p(m) = 0.75$[18].

(iv) 模拟各异常作答模式,包括前面介绍到的疲劳、睡眠、创造性作答、随机作答、作弊、随机作弊、预知试题,这些异常作答模式的生成方式如下.

(a) 疲劳. 测验后期的 30% 试题作答概率按比例下降 50% 生成;

(b) 睡眠. 测验后期的 30% 试题按错误作答生成;

(c) 创造性作答. 当考生的能力参数减去试题难度参数的值超过 1 时,作答概率按比例下降 50% 生成;

(d) 随机作答. 测验后期的 30% 试题随机生成作答;

(e) 作弊. 将测验后期的 30% 的错误修改成正确作答;

(f) 随机作弊. 随机将测验中的 30% 的错误试题修改成正确作答;

(g) 预知试题(试题泄露). 设置 $n_k = 200$,即在试题被作答至第 200 次时试题发生了泄露,之后启用修正的 3 参数 Logistic 模型模拟作答. 监控起始点为 $n_0 = 100$,即当试题被作答至第 100 次时,启用监测程序[13,18-20],移动样本 $m = 50$,$c_\alpha$ 取值为 2.5,$C_{\mathrm{PFS}}$ 取值为 2,每个实验条件下重复 30 次.

在本次模拟研究中,当某道试题被作答次数(即曝光率)达到 100 次($n_0 = 100$)时,就开启 SMP 来监测试题,在每一次的模拟过程中都将记录好以下数据:被监控试题数量,被错误标记的试题数量(即犯 Ⅰ 类错误的数量),犯 Ⅰ 类错误前的考生作答数量,已泄露但认为未泄露的试题数量,实际泄露的试题数量. 然后计算犯 Ⅰ 类错误的概率以及在几种异常作答模式情况下的统计检验力.

### 3.2 评价指标

本文采用Ⅰ类错误率和统计检验力作为评价指标.

(i) Ⅰ 类错误率:$P_{\mathrm{I}}$ = 未泄露但被标为泄露的试题数/(被监控的试题数 − 实际泄露的试题数);

(ii) 统计检验力:$P_{\mathrm{T}}$ = 实际已泄露且被标为泄露的试题数/实际泄露的试题数.

### 3.3 研究结果及讨论

表 1 和表 2 分别是被试拟合指标的 Ⅰ 类错误率和统计检验力的结果.

**表 1 在相应条件下 2 种方法的 Ⅰ 类错误率及其标准差**

| 题目区分度 | 题目数量 $J = 20$ | | 题目数量 $J = 40$ | |
|---|---|---|---|---|
| | SMP | PFS_SMP | SMP | PFS_SMP |
| 高 | 0.051 (0.01) | 0.052 (0.01) | 0.048 (0.01) | 0.045 (0.01) |
| 低 | 0.056 (0.02) | 0.053 (0.01) | 0.051 (0.01) | 0.052 (0.01) |

注:括号内数值表示重复实验 30 次的标准差. 下文同.

**表 2 在几种异常作答模式下 2 种方法的统计检验力及其标准差**

| 题目个数 | 题目区分度 | 方法 | 统计检验力 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 疲劳 | 睡眠 | 创造性作答 | 随机作答 | 作弊 | 随机作弊 | 预知试题 |
| 20 | 高区分度 | PFS_SMP | 0.376 (0.01) | 0.516 (0.02) | 0.530 (0.03) | 0.792 (0.01) | 0.978 (0.01) | 0.744 (0.01) | 0.983 (0.02) |
| | | SMP | 0.332 (0.02) | 0.464 (0.02) | 1.000 (0) | 0.961 (0.02) | 1.000 (0) | 0.167 (0.01) | 0.957 (0.02) |
| | 低区分度 | PFS_SMP | 0.389 (0.01) | 0.291 (0.02) | 0.802 (0.03) | 0.855 (0.01) | 0.877 (0.01) | 0.314 (0.02) | 0.898 (0.02) |
| | | SMP | 0.321 (0) | 0.279 (0.02) | 1.000 (0) | 0.811 (0.04) | 1.000 (0) | 0.153 (0.01) | 0.876 (0.03) |
| 40 | 高区分度 | PFS_SMP | 0.664 (0.01) | 0.620 (0.02) | 1.000 (0) | 0.974 (0.01) | 1.000 (0) | 0.820 (0.02) | 1.000 (0) |
| | | SMP | 0.553 (0.01) | 0.801 (0.02) | 1.000 (0) | 0.963 (0.01) | 1.000 (0) | 0.442 (0.01) | 0.965 (0.03) |
| | 低区分度 | PFS_SMP | 0.492 (0.01) | 0.431 (0.01) | 0.931 (0.01) | 0.832 (0.01) | 0.714 (0.01) | 0.384 (0.02) | 0.984 (0.01) |
| | | SMP | 0.280 (0.01) | 0.434 (0.02) | 0.889 (0.02) | 0.823 (0.04) | 1.000 (0) | 0.185 (0.01) | 0.945 (0.03) |

为了显示出PFS_SMP方法的效果,研究中固定 $c_\alpha$ 与 $C_{\mathrm{PFS}}$ 的值分别为2.5和2,然后计算并记录在相应条件下SMP组与PFS_SMP组的Ⅰ类错误率及其标准差. 在这里分别记录了当试题数量为20与40时,高低区分度下2种方法的Ⅰ类错误率. 由表1可知:随着试题数量的增加,Ⅰ类错误率会逐渐减小;试题的区分度越高,Ⅰ类错误率会越低;属性个数对Ⅰ类错误率没有较大影响. 由表2可知:随着试题数量的增加,在各个异常作答模式下的统计检验力会明显增大;试题的区分度越高,统计检验力则越高.

在表1与表2中将SMP组与PFS_SMP组作对比可以发现:在相同的试题数量、区分度与考察属性个数下,PFS_SMP组有着更低的Ⅰ类错误率和更高的统计检验力. 这说明:PFS_SMP方法能够在一定程度上降低在相同条件下的Ⅰ类错误率,并且提高统计检验力. 在预知试题的情况下,PFS_SMP方法在统计检验力方面比SMP方法更高.

## 4　实证数据分析

为了进一步考察PFS_SMP方法的应用,将它用于分析一批实际数据,该数据是某省的一次大规模标准化测量,题库中一共有360多项选择题,测验采用CAT机试,自适应地为每位考生选题,整个测验持续15 d,每天的考生是200 ~ 950人,一共有8 765名考生参加了测试,对于整个数据集,采用3参数模型进行拟合,量表常数 $D = 1$[16],得到所有试题的参数和各考生的能力参数. 难度参数和区分度参数的均值分别为0.136、0.998,它们的标准差分别为0.562、0.475. 整个考生群体的能力分布呈负偏态分布,能力参数的均值和标准差分别是0.967、1.132. 采用如下的步骤对各个测验窗口的数据进行分析:

(i) 监控每个被选中的试题,若某个试题的使用次数达到100,则启动PFS_SMP程序来监测该试题;

(ii) 对每个被监测的试题,若SMP程序将它标记为泄露题,则同时标记出施测了这个题目的考生;

(iii) 对标记的考生计算 $R$ 统计量的值,判断其是否为异常考生;

(iv) 根据上述的2个步骤,从试题和考生的水平对题目和考生的异常情况进行判断.

基于上述的过程,一共有18个题目被标记为泄露题,357名考生被标记为异常考生. 较低比例的试题被标记的主要原因可能是该测验并不是一个高风险测验,对考试分数特别重视的考生比例可能并不是太高. 图3和图4分别显示了被标记出的试题和未被标记出的试题的参数随着测验窗口的增加而变化的情况.

从图3和图4可以看出:一方面,当试题发生泄露时,其参数会有很明显的下降趋势,而对于未泄露试题的参数则在整个测验窗口中保持相对稳定的状态. 另一方面,发生泄露试题的参数随着测验窗口的增加而在持续下降. 由图3可以看出:第71题是一道区分度较大的题,根据最大信息量选题规则,应该有很多考生选到这道题,几乎可以确定在测验的第2天就发生了泄露,导致该题的参数在逐渐下降,到第15天,它已经变成了一道区分度很低的试题,这是因为此时很多考生都已经提前了解了这道题的信息.

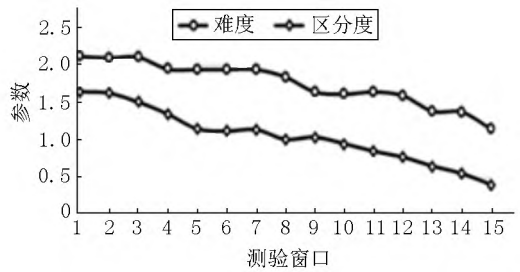


**图3　第71题的参数变化趋势(泄露题)**

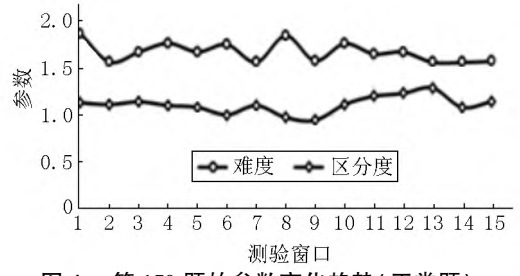


**图4　第150题的参数变化趋势(正常题)**

## 5　总结与展望

### 5.1　结论

题目的安全性一直是在所有大型测试中非常重要的问题,历年来有很多研究者在这方面做了相关的研究,针对题目的曝光提出了很多曝光控制方法. 在E. Georgiadou等[18]的研究中,他总结出了5大类曝光控制法:随机化法、分层法、条件选择法、结合前3者的综合方法和多阶段自适应设计. 但控制题目的曝光率并不能够有效控制试题泄露,有的试题

曝光率达到几百上千次都没有泄露,而有的试题可能只曝光了几次便被泄露了,因此文献[14,20]提出了 SMP 方法,运用监测程序对试题进行更有效的监控. 但 SMP 方法会发生一定的虚报,即一定的 Ⅰ 类错误率,相应的统计检验力会降低,因此在试题被 SMP 方法监测标记为泄露后,还需要其他的统计方法对其进行判断.

为了降低 SMP 方法的 Ⅰ 类错误率和提高其统计检验力,本文提出基于残差的 $R$ 指标[21],并与 SMP 方法相结合,构建了新的方法:PFS_SMP 方法. PFS_SMP 方法相较于 SMP 方法的优势为:使用了被试的作答信息,从试题和被试 2 个角度出发,核实被 SMP 监测标记为泄露的试题是否真正泄露,以此降低 Ⅰ 类错误率. 此外,本文采用了修正的 3 参数 Logistic 模型模拟作答数据,将考生能力、试题的区分度、难度及猜测度等参数考虑进去,让结果更加符合实际情况[19,22].

通过一系列的实验验证,得出了以下几个结论:

(i) 试题数量的增加与区分度的提高能够降低 Ⅰ 类错误率与提高相应的统计检验力;

(ii)PFS_SMP 方法能够在一定程度上有效降低在 SMP 方法中存在的 Ⅰ 类错误率;

(iii)PFS_SMP 方法能够在一定程度上提高在 SMP 方法中的统计检验力.

### 5.2 展望

本文对一些研究条件做了一定的限制,如 $c_\alpha$ 与 $C_{\mathrm{PFS}}$ 的取值固定为 2.5 和 2,后续研究可以取在这 2 个值附近的其他值,并进行相应的对比观察,以研究随着 $c_\alpha$ 与 $C_{\mathrm{PFS}}$ 的值发生改变实验结果是否会产生相应的变化;若试题的相关参数和考生的能力值发生相应改变,则 $c_\alpha$ 与 $C_{\mathrm{PFS}}$ 又该如何取值来进行更好的研究,这是未来可以研究的方向.

本文主要对 SMP 方法与 PFS_SMP 方法中 Ⅰ 类错误率和统计检验力的变化进行了相应的研究,而对于在 SMP 方法中的试题泄露位置却未进行详细研究. 随着试题数量或区分度等相关参数的变化,泄露位置的探查会发生什么变化以及有什么方法可以提高探查泄露位置的准确性与从统计测量视角来推动考试公平和教育公平[23],这些也是未来可以研究的方向.

## 6 参考文献


[1] VAN DER LINDEN W J, GLAS C A W. Computerized adaptive testing: theory and practice [M]. Berlin: Springer, 2000.

[2] MAGIS D, YAN Duanli, VON DAVIER A A. Computerized adaptive and multistage testing with R: using packages catR and mstR [M]. Switzerland: Springer International Publishing, 2017.

[3] 唐倩,毛秀珍,何明霜,等. 认知诊断计算机化自适应测验的选题策略 [J]. 心理科学进展, 2020, 28(12): 2160-2168.

[4] 于建芳,徐振国,刘剑. 计算机自适应测试系统研究综述 [J]. 中国教育技术装备, 2015(6): 28-29.

[5] 杨业兵. 应用项目反应理论对《中国士兵人格测验》的项目分析及计算机自适应施测方案 [D]. 西安: 第四军医大学, 2008.

[6] 涂冬波. 项目自动生成的小学儿童数学问题解决认知诊断 CAT 编制 [D]. 南昌: 江西师范大学, 2009.

[7] 邓远平. 基于展开反应机制的计算机化自适应人格测验研究 [D]. 南昌: 江西师范大学, 2014.

[8] YI Qing, ZHANG Jinming, CHANG Huahua. Severity of organized item theft in computerized adaptive testing: a simulation study [J]. Applied Psychological Measurement, 2008, 32(7): 543-558.

[9] ZHANG Jinming, CHANG Huahua, YI Qing. Comparing single-pool and multiple-pool designs regarding test security in computerized testing [J]. Behavior Research Methods, 2012, 44(3): 742-752.

[10] STOCKING M L, LEWIS C. Controlling item exposure conditional on ability in computerized adaptive testing [J]. Journal of Educational and Behavioral Statistics, 1998, 23(1): 57-75.

[11] MARTHA L S, CHARLES L. Controlling item exposure conditional on ability in computerized adaptive testing [EB/OL]. [2021-09-12]. https://onlinelibrary. wiley. com/doi/10. 1002/j. 2333 - 8504. 1995. tb01659. x.

[12] WAINER H. Rescuing computerized testing by breaking Zipf's law [J]. Journal of Educational and Behavioral Statistics, 2000, 25(2): 203-224.

[13] WAY W D. Protecting the integrity of computerized testing item pools [J]. Educational Measurement: Issues and Practice, 1998, 17(4): 17-27.

[14] ZHANG Jinming. A sequential procedure for detecting compromised items in the item pool of a CAT system [J]. Applied Psychological Measurement, 2014, 38(2): 87-104.

[15] MEIJER R R, SIJTSMA K. Methodology review: evaluating

person fit [J]. Applied Psychological Measurement,2001,25(2):107-135.
[16] BAKER F B,KIM S H. Item response theory:parameter estimation techniques [M]. 2nd ed. New York:Marcel Dekker,2004.
[17] MCLEOD L,LEWIS C,THISSEN D. A Bayesian method for the detection of item preknowledge in computerized adaptive testing [J]. Applied Psychological Measurement,2003,27(2):121-137.
[18] 郭磊,刘伟. CAT 中结合贝叶斯方法与序贯监测程序的题库质量监控技术 [J]. 心理科学,2018,41(1):189-195.
[19] GEORGIADOU E,TRIANTAFILLOU E,ECONOMIDES A A. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005 [J]. The Journal of Technology, Learning, and Assessment,2007,5(8):4-38.
[20] 张金明,曹灿兮,揭勇菁. 实时监控计算机自适应考题的两种方法及其稳健性比较 [J]. 中国考试,2017(2):20-32.
[21] YU Xiaofeng,CHENG Ying. A change-point analysis procedure based on weighted residuals to detect back random responding [J]. Psychological Methods,2019,24(5):658-674.
[22] PETRIDOU A,Williams J. Accounting for aberrant test response patterns using multilevel models [J]. Journal of Educational Measurement,2007,44(3):227-247.
[23] 汪文义,张华华. 统计测量视角下考试公平推动教育公平的对策 [J]. 江西师范大学学报(自然科学版),2017,41(4):383-393.

# The Study and Application of Real-Time Monitoring Methods for Item Leakage in Computerized Adaptive Tests

QIN Chunying[1],WU Longyue[1],WANG Aiping[2*]
(1. School of Mathematics and Information Science,Nanchang Normal University,Nanchang Jiangxi 330032,China;
2. Department of Electronic and Information Engineering,Bozhou University,Bozhou Anhui 246800,China)

**Abstract**:Computerized Adaptive Test (CAT) makes the possibility of each item being exposed increase,so the item needs to be monitored in real time. When the item parameters change significantly,it must be forced to "retire". The sequential monitoring program (SMP) is proposed in 2014 to determine whether an item is leaking by detecting changes in the statistical characteristics of the item in CAT. However,when using SMP to monitor the item,there will be a relatively high error rate of type I,and the statistical test will also have a greater impact. In this paper,based on the residual person fit statistic R,combined with the SMP method,a new monitoring method (PFS_SMP) is proposed. The PFS_SMP method can be applied to determine whether each respondent takes aberrant response behavior,and each item is known by the future respondents during the CAT,and to ensure the safety and fairness of the test. Finally,a simulation study and an empirical study are considered,and the results show that the PFS_SMP method can yield a well-controlled error rate of type I,and have a promising power as well.

**Key words**:computerized adaptive test;test fit index;sequential monitoring procedures;residual;item security

(责任编辑:冉小晓)