

结合 a 分层的兼具项目曝光和广义测验重叠率控制的选题策略*

郭 磊¹ 王卓然¹ 王 丰¹ 边玉芳^{1,2}

(¹北京师范大学认知神经科学与学习国家重点实验室;

²中国基础教育质量评价与提升协同创新中心, 北京 100875)

摘 要 测验安全和题库使用率在计算机化自适应测验中十分重要, 特别是高风险测验。传统的 SHGT 法兼具同时控制项目曝光率和广义测验重叠率的功能, 但题库使用率较差。a 分层法能够提高题库使用率, 但对过度曝光的项目控制不足。本研究将 a 分层法的思想与 SHGT 法相结合, 各取所长, 提出了 3 种新的选题方法: SHGT_a 法, SHGT_b 法和 SHGT_c 法。研究结果表明: (1)与 SHGT 法相比, 新方法均可以在有效地控制项目曝光率和广义测验重叠率同时, 极大地提高题库使用率; (2)随着预设项目曝光率($r_{\max}$)和广义测验重叠率($\bar{\Omega}_{\max}$)取值的增大以及共享人数 α 的减小, 新方法对被试能力估计的精度呈上升趋势。比起 SHGT 法, 新方法仍能保持很高的题库使用率; (3)当区分度和难度的相关(r_{ab})较大时, SHGT_b 和 SHGT_c 法在能力估计精度方面优于 SHGT_a 法; (4)在不同的测验考察内容比例下, 3 种新方法对被试能力估计的精度均较好; (5)与 SHGT 法相比, 新方法能够有效地控制项目曝光率过度控制的问题。

关键词 项目曝光率; 广义测验重叠率; 计算机化自适应测验; a 分层法; 选题策略

分类号 B841

1 引言

计算机化自适应测验(Computerized Adaptive Testing, CAT)在过去几十年里备受关注, 已经成为了许多大规模教育测量项目的测验模式(陈平, 2011)。与传统的纸笔测验相比, CAT 的最大优势表现在测试更少项目的同时, 能够快速获取更加精确的能力估计值(Weiss, 1982), 并且施测更加灵活。由于近年来网络的快速发展, CAT 的测验功效发挥到了极致, 像 GRE、ASVAB、GMAT、美国护士资格考试等大型考试都采用了 CAT (唐小娟, 丁树良, 俞宗火, 2012)。

在 CAT 测验中, 特别是高风险(high stake)测验, 考试的安全性十分重要(程小扬, 丁树良, 严深海, 朱隆尹, 2011)。为了保证测验及题库安全性, 主要做法是控制题库中项目的曝光率, 使其低于预先设

定的曝光率最大值。在 CAT 测验中, 研究者提出了许多控制项目曝光率的方法, 根据 Georgiadou, Triantafillou 和 Economides (2007)的总结, 当前主要有 5 类项目曝光控制方法: (1)随机化方法; (2)条件选择方法; (3)分层方法; (4)结合前三者的综合方法; (5)多阶段自适应设计。然而, 以上的方法均只关注了项目曝光率的控制。

Way (1998)认为在 CAT 中, 项目曝光率和测验重叠率是两个反映测验安全性的指标。因此, 在 CAT 测验中只考虑曝光率控制是不够的。Chen, Ankenmann 和 Spray (2003)对项目曝光和测验重叠之间的关系进行了研究, 推导出了测验重叠率和项目曝光率之间的关系, 即当被试量无穷大时, 测验重叠率等于曝光率方差与曝光率均值平方的和与曝光率均值的比值, 即 $\bar{T} = (S_r^2 + \bar{r}^2) / \bar{r}$ ($\bar{T}$ 为测验重叠率, S_r^2 为曝光率方差, $\bar{r}$ 为曝光率均值)。随后,

收稿日期: 2013-06-14

* 高等学校博士学科点专项科研基金资助课题(20120003110002)的资助。

通讯作者: 边玉芳, E-mail: bianyufang66@126.com

Chen 和 Lei (2005)对 Sympson - Hetter (SH)选题策略(Sympson & Hetter, 1985)进行了改进, 提出了兼具控制项目曝光率和测验重叠率的 SHT 法。但 SHT 法在使用之前, 仍需要进行大量迭代模拟以获取稳定的项目曝光控制参数, 模拟过程非常耗时, 偶尔还会出现不能满足所有曝光率都符合目标值的情况(van der Linden, 2003), 并且一旦题库中的项目或被试群体稍有变化, 便要重新模拟。由此, Chen, Lei 和 Liao (2008)提出了一种在线(on the fly)更新项目曝光控制参数的 SHTOR 法。虽然 SHTOR 法弥补了 SHT 法的不足, 但该方法所定义的测验重叠率是指所有两两被试之间(pairwise)共享项目数量的平均值(Chen & Lei, 2005; Way, 1998)。该定义下的测验重叠率可能与现实不符, 因为在实际中, 某个被试可能会接收到不止一个被试分享的项目, 这时就需要考虑一组被试之间的测验重叠率问题, 即广义测验重叠率(general test overlap; Chang & Zhang, 2002; Chen, 2010)。可以看出, 两两被试间的测验重叠率是广义测验重叠率的特例, 当分享项目的被试数量为 1 时(记作 $\alpha = 1$), 两者相近。基于此, Chen (2010)提出了 SHGT 法, 解决了对多个被试(通常为 1 人~5 人)之间共享项目的广义测验重叠率控制的问题。尽管该研究能够同时实现对项目曝光率和广义测验重叠率的控制, 但当项目曝光率和广义测验重叠率控制在稍大值(例如 0.3), 或分享项目人数较少(例如 2 人及以下)时, 题库中未使用的项目数量高达 1/3 以上, 题库使用率较差。对题库使用不均匀会降低测验及题库安全性, 增大经济耗损 (Revuelta & Ponsoda, 1998; Stocking & Swanson, 1993; Weiss, 1985)。因此, 如何能够在同时控制曝光率和测验重叠率的基础上, 达到均匀使用题库的目的就显得十分重要。

那么, 究竟该如何提高题库使用的均匀性呢? 受 Chang 和 Ying (1999)提出的 a 分层方法的启发, 我们认为将 a 分层方法的思想 and SHGT 法相结合可以提高后者题库的使用率。a 分层的优势在于, 能够提高未使用或较少使用项目的曝光率, 使得项目曝光率和题库使用率更加均衡。但根据 Parshall, Harmes 和 Kromrey (2000)的研究表明, a 分层对于某些项目仍然有较大的曝光率。并且在实际题库中, a 与 b 通常都是正相关的(Lord, 1975), 如果某一层内难度 b 的范围不足以覆盖被试能力水平时, 将会导致某些项目过度选择。很明显, SHGT 法和 a 分层的缺陷可以相互弥补, 前者能够有效的控制过度

曝光项目的出现, 但不能提高题库使用均匀性, 后者虽能提高题库使用率, 但未能有效地控制过度曝光的项目。因此, 本研究尝试将两种方法相结合, 以实现既能控制项目曝光率和广义测验重叠率, 又能提高题库使用率的目的。

查阅国内外文献, 尚未见到能够在同时控制项目曝光率和广义测验重叠率的基础上, 提高题库使用率的研究, 并且没有研究过测验考察的内容比例对不同的选题策略有何影响。实际中, 不同测验所考察的内容比例是根据具体的测验目的而设置的, 而题库的内容比例是相对稳定不变的。因此, 研究测验考察的内容比例对不同选题策略的影响很有必要。本研究将 a 分层、按 b 分块的 a 分层(Chang, Qian, & Ying, 2001)以及按内容分块的 a 分层方法(Yi & Chang, 2003)与 SHGT 法相结合, 分别记作 SHGT_a 法、SHGT_b 法和 SHGT_c 法, 意在实现上述目标。本文拟采用蒙特卡洛方法进行模拟研究, 意在探讨: (1)在不同的项目曝光率和广义测验重叠率水平下, 不同选题策略之间的表现有何差异; (2)在不同区分度和难度的相关水平下, 不同选题策略之间的表现有何差异; (3)在不同的内容考察比例下, 不同选题策略之间的表现有何差异。

2 相关选题策略简介

2.1 SHGT 法

SHGT 法是一个比较复杂的选题策略, 它融合了 SH 法的思想, 同时控制了广义测验重叠率, 采用在线更新项目曝光控制参数的方法而成。该方法有几大优势: (1)可以同时控制项目曝光率和广义测验重叠率; (2)在线更新曝光控制参数, 无需迭代模拟, 大大节省了 CAT 的时间; (3)能够适用于题库中项目和被试群体发生变化的情况; (4)可以和其他选题策略相结合使用。

基于广义测验重叠率的概念, Chen (2010)给出了其计算公式:

$$\bar{\Omega}_{\alpha} = \frac{\sum_{t=\alpha+1}^p \sum_{i=1}^n \sum_{j=1}^{\alpha} [m_{it} - m_{i(t-1)}] \binom{m_{it}-1}{j} \binom{t-m_{it}}{\alpha-j}}{h \binom{p}{\alpha+1}} \quad (1)$$

其中, $\bar{\Omega}_{\alpha}$ 为某个被试与已参加测验当中的 α 个被试之间共享项目数量的平均值, p 为已经参加测验的总人数, n 为题库数量, h 为测验长度, m_{it} 为第 i 题被已施测的前 t 个被试使用的次数。规定, 若

$p \leq \alpha$, 则 $\bar{\Omega}_{\alpha} = 0$ 。详细推导过程请参见 Chen 和 Lei (2010)。

为了计算和编程的便捷性, 需要对广义测验重叠率进行重构, 可以根据递归算法进行计算, 公式如下:

$$\bar{\Omega}_{\alpha,t} = \left(\frac{t-\alpha-1}{t} \right) \bar{\Omega}_{\alpha,(t-1)} + \left(\frac{\alpha+1}{t} \right) - \frac{\sum_{i=1}^n [m_{it} - m_{i(t-1)}] \binom{t-m_{it}}{\alpha}}{h \binom{t}{\alpha+1}} \quad (2)$$

其中, $\bar{\Omega}_{\alpha,t}$ 表示当测验已施测了 t 个被试, 在下一个被试施测之前, 有 α 个被试给该被试分享的项目数量的平均比例, 其余符号同前。若 $\alpha \geq t$, 则 $\bar{\Omega}_{\alpha,t} = 0$ 。

很容易知道, 某道项目使用的频率越高, 广义测验重叠率越大。因此, 为了将实际施测中得到的广义测验重叠率控制在允许的最大值 $\bar{\Omega}_{\max}$ 之下, 需要定义两个指标: 一个是项目贡献率指标:

$\eta_i = \binom{t-m_{it}}{\alpha}$, 另一个是项目贡献临界值指标:

δ/h 。其中, 临界值指标的算法如下:

$$\frac{\delta}{h} = \frac{h \binom{t}{\alpha+1} \left[\left(\frac{t-\alpha-1}{t} \right) \bar{\Omega}_{\alpha,(t-1)} + \left(\frac{\alpha+1}{t} \right) - \bar{\Omega}_{\max} \right]}{h} \quad (3)$$

在定义了两个指标之后, SHGT 法的具体操作分为以下几个步骤:

(1) 将题库中所有项目的起始曝光参数(记为 k_i , i 为项目)设置为 1, 设定测验允许的最大曝光率 $r_{\max}$ 、最大广义测验重叠率 $\bar{\Omega}_{\max}$ 、测验长度 h 、题库大小 n 以及共享项目的被试数量 α ;

(2) 对第 t 个被试实施 CAT 测验。根据公式(3)计算出第一题的临界值 δ/h , 在给定的选题策略(在 SHGT 法中, 本文采用最大费歇信息量法; 在新方法中, 采用 b-matching 法)下, 若选出了题库中的第 i 题, 则将第 i 题的贡献率 η_i 和临界值 δ/h 进行比较, 并且从均匀分布 $U(0,1)$ 中产生一个随机数 x 。如果满足条件 $\eta_i \geq \delta/h$ 且 $x \leq k_i$, 那么施测第 i 题, 否则将此题从题库中删除, 不再对该被试施测。如此往复;

(3) 在第一题 i 施测后, 将前一步的临界值 δ/h 更新为 $(\delta - \eta_i)/(h-1)$, 作为选择下一题比较的条件。将施测的第二题记作项目 j , 当第二题 j 施测后, 继续更新临界值为 $(\delta - \eta_i - \eta_j)/(h-2)$, 即每做完一道项目就更新一次临界值, 如此往复;

(4) 基于已施测的 t 个被试的测验情况, 计算出每个项目的项目选择概率(记为 $P(S)$)和项目曝光概率(记为 $P(A)$), 然后对 k 值进行更新:

如果 $P(A) > r_{\max}$, 那么 $k = 0$;

如果 $P(A) \leq r_{\max}$ 并且 $P(S) > r_{\max}$, 那么 $k = r_{\max} / P(S)$;

如果 $P(A) \leq r_{\max}$ 并且 $P(S) \leq r_{\max}$, 那么 $k = 1$

(5) 更新完 k 值后, 为了保证被试均能顺利完成 CAT 测验, 需要设置 h 个 k 值等于 1, 具体做法是令最接近 1 的那些 k 值等于 1, 直到有 h 个 1 为止;

(6) 在得到所有 k 值后, 返回到第 2 步对第 $t+1$ 个被试施测 CAT 测验。重复以上步骤直到所有被试参加完 CAT 测验。

需要强调的是, 当设定的 $\bar{\Omega}_{\max}$ 趋近其下限值时, 会出现无题可选的情况, 此时应在步骤(2)后加上一个补救措施, 详细描述请参见 Chen (2010)。

2.2 a 分层法, 按 b 分块的 a 分层法, 按内容分块的 a 分层法

2.2.1 a 分层法 (STR_a)

Chang 和 Ying (1999) 最早提出了 a 分层方法来提高题库安全性。经研究表明, 尽管高 a 的项目很有价值, 但当高 a 项目施测于能力真值与项目难度不接近的被试时, 该项目的价值并不能得到充分体现。当对被试能力真值认识有限时, 应尽量避免使用高 a 的项目, 并且在 CAT 早期阶段使用低 a 项目, 后期使用高 a 项目的方式是合理的。由此, 他们提出了著名的 a 分层选题策略, 步骤为: (1) 根据区分度 a 将题库分成 K 个水平; (2) 相应地, 将测验也分成 K 个阶段; (3) 在测验的第 k 个阶段, 选择项目难度与能力估计值最接近的 h_k 个项目施测, 保证 $h_1 + h_2 + \dots + h_k = h$; (4) 重复步骤 3。

2.2.2 按 b 分块的 a 分层法 (STR_b)

正如 Chang 和 Ying (1999) 的研究表明, a 分层有较好表现的一个前提是 a 和 b 之间没有相关。但是在实际题库中, a 与 b 通常都是正相关的 (Lord, 1975)。如果某一层内的 b 范围不足以覆盖被试能力水平时, 就会导致某些项目过度选择。而且在高 a 层中, 高 a 低 b 的项目很少, 将会导致这些项目过度曝光。于是, Chang, Qian 和 Ying (2001) 提出按难度 b 分块的 a 分层法, 步骤为: (1) 基于难度 b 将题库分成 M 块。所有组块中项目数量相同。将这些组块按照升序排列; (2) 在每个组块中, 按照 a 值分成 K 个水平; (3) 将同一水平的不同组块重新组合,

形成 K 个水平的题库, 这样在同一水平内的难度 b 也覆盖了整个能力范围; (4)按照 a 分层的步骤进行 CAT 测验。若 a 和 b 的相关为 0 时, STR_b 和 STR_a 是一样的。

2.2.3 按内容分块的 a 分层法(STR_c)

Van der Linden (2000)认为, CAT 只有将统计性能和非统计要求相结合才能被接受, 即要在实际中考虑内容平衡等非统计属性, 以便 CAT 测验有较高的内容效度以及被试的测验分数可以比较。于是, Yi 和 Chang (2003)提出了按内容分块的 a 分层法, 步骤为: (1)根据内容领域将题库分成若干个组; (2)在每个组里实施 STR_b。若内容领域为一个时, STR_c 和 STR_b 是一样的。

3 研究设计

本研究要比较的选题策略包括: 随机选题法(RN)、SHGT 法、SHGT_a 法、SHGT_b 法以及 SHGT_c 法。结合方式为, 首先运用 STR_a (b 或 c) 法将题库分层, 在每一层内使用 SHGT 法。其中, RN 作为比较的基线。采用 Matlab 2011b 自编以上所有选题策略。

3.1 题库及被试

首先生成 360 题的题库($n=360$)。区分度 a, 难度 b 以及猜测度 c 按如下先验分布生成: $a \sim U(0.5, 1.5)$, $b \sim N(0, 1)$, $c \sim U(0, 0.4)$ 。其次, 本研究中固定题库所考察的内容领域数量 $g=3$, 并且规定内容领域的项目数量比例为 1:1:1, 各 120 题, 由此生成模拟题库, 用于进行所有的实验。按照先验分布 $\theta \sim N(0, 1)$ 生成 3000 名被试。

3.2 实验条件及说明

本文考察几种选题策略的表现, 并考虑 5 种可能的影响因素: (1)最大项目曝光率 $r_{\max}$: 0.2 和 0.3; (2)最大广义测验重叠率 $\bar{\Omega}_{\max}$: 0.3 和 0.4; (3)共享人数 α : 1 人、2 人和 3 人; (4)a 和 b 之间的相关水平 r_{ab} : 0.2 和 0.8; (5)CAT 测验考察的内容比例¹: 1:1:1 和 1:2:3。

本研究中 CAT 测验的终止规则选取定长 CAT, 这也是大多数 CAT 研究采取的方法。固定测验长度 $h=30$, 这是因为 Stocking (1994)建议题库大小至少应该是测验长度的 12 倍。若使用 SHGT_a 法, 令层数 $K=4$ 。施测顺序为先施测低 a 层, 最后施测高

a 层。每层内项目数量固定为 7 题, 7 题, 8 题和 8 题; 若使用 SHGT_b 法, 令块数 $M=3$, 再令层数 $K=4$, 其他同 SHGT_a 法; 若使用 SHGT_c 法, 先按照内容领域将题库分成若干个组, 随后在每个组里实施 SHGT_b 法; 利用 EAP 法对被试能力进行更新。

5 种选题策略均采用多项式模型(multinomial model)进行内容平衡控制, 详细过程描述请参见 Chen 和 Ankenman (2004)。在考虑内容平衡时, Chen (2010)给出了 $\bar{\Omega}_{\max}$ 的下限值, 计算公式如下:

$$\bar{\Omega}_L \approx 1 - \frac{\sum_{c=1}^g h_c (1 - \frac{h_c}{n_c})^\alpha}{h} \quad (4)$$

其中, $\bar{\Omega}_L$ 为 $\bar{\Omega}_{\max}$ 的下限值, g 为内容领域的数量, h_c 为 CAT 测验中考察了内容领域 c 中的项目数量, n_c 为题库中属于内容领域 c 的项目数量。根据本研究中选取的实验条件, 运用公式(4)计算出每种实验条件下 $\bar{\Omega}_{\max}$ 的下限值。其中最大下限值等于

$$\frac{0.2616(\bar{\Omega}_L \approx 1 - \frac{5(1-5/120)^3 + 10(1-10/120)^3 + 15(1-15/120)^3}{30})}{0.2616} \approx 0.2616$$

), 因此, 影响因素(2)选取 0.3 和 0.4 两个水平, 意在说明其影响趋势。

3.3 评价指标

(1)广义测验重叠率

$$\bar{\Omega}_\alpha = \frac{\sum_{t=\alpha+1}^p \sum_{i=1}^n \sum_{j=1}^\alpha [m_{it} - m_{i(t-1)}] \binom{m_{it}-1}{j} \binom{t-m_{it}}{\alpha-j}}{h \binom{p}{\alpha+1}}$$

(2)误差均方根

$$RMSE = \sqrt{\sum_{n=1}^N (\theta_n - \hat{\theta}_n)^2 / N}$$

其中, N 为被试数量, θ 和 $\hat{\theta}$ 分别为能力真值和估计值。RMSE 反映了参数真值与估计值之间的平均偏差大小, 其值越小越好。

除此之外, 程序还记录了最大项目曝光率和使用过的项目数量, 以此考察各选题策略的性能。

4 研究结果

总体来看, 根据表 1 至表 4 的结果, 不论共享人数为多少, SHGT 及 3 种新方法均能很好地控制项目曝光率和广义测验重叠率。例如, 根据表 1 结

¹ CAT 测验考察的内容比例是指被试参加的测验所规定的考察内容比例, 和题库本身的内容比例无关, 即题库的内容比例始终为 1:1:1, 但 CAT 测验的内容比例分别为 1:1:1 和 1:2:3。

表 1 $r_{ab}=0.2$, 测验内容比例为 1:1:1 时, 5 种选题策略的结果

ISS	$r_{\max}$	$\bar{\Omega}_{\max}$	Max_ER	$\bar{\Omega}_1$	$\bar{\Omega}_2$	$\bar{\Omega}_3$	#used	Per.(%)	RMSE
SHGT1	0.2	0.3	0.200	0.177	0.323	0.443	215	59.72	0.280
	0.2	0.4	0.200	0.172	0.315	0.432	233	64.72	0.271
	0.3	0.3	0.300	0.246	0.428	0.561	179	49.72	0.274
	0.3	0.4	0.300	0.255	0.443	0.580	178	49.44	0.264
SHGT2	0.2	0.3	0.186	0.164	0.300	0.413	232	64.44	0.284
	0.2	0.4	0.200	0.173	0.317	0.434	227	63.06	0.274
	0.3	0.3	0.192	0.164	0.300	0.413	240	66.67	0.277
	0.3	0.4	0.278	0.227	0.400	0.530	204	56.67	0.268
SHGT3	0.2	0.3	0.128	0.112	0.212	0.300	306	85.00	0.323
	0.2	0.4	0.186	0.157	0.290	0.400	248	68.89	0.306
	0.3	0.3	0.128	0.112	0.212	0.300	304	84.44	0.311
	0.3	0.4	0.181	0.157	0.289	0.400	238	66.11	0.286
SHGT_a1	0.2	0.3	0.200	0.116	0.216	0.303	356	98.89	0.289
	0.2	0.4	0.200	0.117	0.219	0.307	349	96.94	0.272
	0.3	0.3	0.300	0.122	0.224	0.310	353	98.06	0.273
	0.3	0.4	0.296	0.120	0.223	0.310	353	98.06	0.266
SHGT_a2	0.2	0.3	0.198	0.116	0.217	0.304	356	98.89	0.297
	0.2	0.4	0.200	0.111	0.209	0.294	358	99.44	0.287
	0.3	0.3	0.300	0.119	0.220	0.306	359	99.72	0.285
	0.3	0.4	0.300	0.120	0.221	0.308	354	98.33	0.271
SHGT_a3	0.2	0.3	0.192	0.114	0.212	0.299	354	98.33	0.312
	0.2	0.4	0.200	0.115	0.216	0.302	349	96.94	0.304
	0.3	0.3	0.282	0.114	0.213	0.299	356	98.89	0.307
	0.3	0.4	0.297	0.123	0.226	0.312	357	99.17	0.284
SHGT_b1	0.2	0.3	0.200	0.111	0.209	0.293	353	98.06	0.279
	0.2	0.4	0.200	0.113	0.212	0.299	353	98.06	0.273
	0.3	0.3	0.300	0.126	0.231	0.319	356	98.89	0.277
	0.3	0.4	0.297	0.117	0.217	0.303	353	98.06	0.266
SHGT_b2	0.2	0.3	0.200	0.115	0.216	0.303	357	99.17	0.283
	0.2	0.4	0.200	0.117	0.218	0.305	349	96.94	0.274
	0.3	0.3	0.296	0.124	0.228	0.316	359	99.72	0.279
	0.3	0.4	0.300	0.120	0.221	0.307	356	98.89	0.271
SHGT_b3	0.2	0.3	0.200	0.112	0.209	0.294	353	98.06	0.301
	0.2	0.4	0.200	0.113	0.210	0.296	352	97.78	0.292
	0.3	0.3	0.293	0.113	0.211	0.296	356	98.89	0.289
	0.3	0.4	0.297	0.117	0.217	0.304	356	98.89	0.277
SHGT_c1	0.2	0.3	0.200	0.109	0.205	0.290	355	98.61	0.272
	0.2	0.4	0.200	0.119	0.210	0.295	355	98.61	0.274
	0.3	0.3	0.295	0.112	0.210	0.294	354	98.33	0.278
	0.3	0.4	0.296	0.113	0.211	0.296	355	98.61	0.261
SHGT_c2	0.2	0.3	0.200	0.107	0.200	0.282	357	99.17	0.293
	0.2	0.4	0.200	0.112	0.210	0.296	351	97.50	0.280
	0.3	0.3	0.296	0.115	0.214	0.299	352	97.78	0.272
	0.3	0.4	0.300	0.117	0.217	0.303	352	97.78	0.265
SHGT_c3	0.2	0.3	0.200	0.110	0.208	0.293	347	96.39	0.301
	0.2	0.4	0.200	0.111	0.208	0.293	352	97.78	0.288
	0.3	0.3	0.282	0.114	0.213	0.299	358	99.44	0.281
	0.3	0.4	0.296	0.117	0.217	0.302	352	97.78	0.274
RN	—	—	0.098	0.083	0.160	0.230	360	100	0.372

注：ISS 为选题策略(Item Selection Strategy), Max_ER 为最大曝光率, SHGT1-SHGT3 分别表示共享人数为 1 个、2 个和 3 个时的 SHGT 程序, $\bar{\Omega}_1-\bar{\Omega}_3$ 分别表示 1 个、2 个和 3 个被试分享项目的广义测验重叠率, #used 表示使用到的项目数量, Per.(%)表示使用到的项目百分比

表 2 $r_{ab}=0.8$, 测验内容比例为 1:1:1 时, 5 种选题策略的结果

ISS	$r_{\max}$	$\bar{\Omega}_{\max}$	Max_ER	$\bar{\Omega}_1$	$\bar{\Omega}_2$	$\bar{\Omega}_3$	#used	Per.(%)	RMSE
SHGT1	0.2	0.3	0.200	0.170	0.313	0.430	232	64.44	0.303
	0.2	0.4	0.200	0.173	0.316	0.433	233	64.72	0.285
	0.3	0.3	0.300	0.250	0.435	0.572	188	52.22	0.295
	0.3	0.4	0.297	0.249	0.434	0.569	191	53.06	0.276
SHGT2	0.2	0.3	0.188	0.164	0.300	0.413	235	65.28	0.307
	0.2	0.4	0.200	0.176	0.321	0.441	220	61.11	0.294
	0.3	0.3	0.189	0.164	0.300	0.413	236	65.56	0.296
	0.3	0.4	0.280	0.228	0.400	0.530	194	53.89	0.288
SHGT3	0.2	0.3	0.125	0.112	0.212	0.300	296	82.22	0.314
	0.2	0.4	0.182	0.157	0.289	0.399	246	68.33	0.311
	0.3	0.3	0.124	0.112	0.212	0.300	300	83.33	0.314
	0.3	0.4	0.184	0.158	0.290	0.400	232	64.44	0.293
SHGT_a1	0.2	0.3	0.200	0.140	0.259	0.357	345	95.83	0.307
	0.2	0.4	0.200	0.137	0.254	0.352	339	94.17	0.301
	0.3	0.3	0.300	0.165	0.294	0.397	339	94.17	0.300
	0.3	0.4	0.300	0.169	0.300	0.402	344	95.56	0.295
SHGT_a2	0.2	0.3	0.200	0.136	0.252	0.349	347	96.39	0.316
	0.2	0.4	0.200	0.134	0.247	0.343	347	96.39	0.305
	0.3	0.3	0.300	0.168	0.300	0.402	345	95.83	0.307
	0.3	0.4	0.300	0.170	0.303	0.408	349	96.94	0.302
SHGT_a3	0.2	0.3	0.200	0.113	0.213	0.300	346	96.11	0.326
	0.2	0.4	0.200	0.135	0.250	0.345	344	95.56	0.309
	0.3	0.3	0.282	0.114	0.213	0.300	352	97.78	0.313
	0.3	0.4	0.298	0.165	0.295	0.397	345	95.83	0.306
SHGT_b1	0.2	0.3	0.200	0.117	0.217	0.304	344	95.56	0.287
	0.2	0.4	0.200	0.116	0.216	0.303	352	97.78	0.280
	0.3	0.3	0.297	0.126	0.232	0.321	349	96.94	0.283
	0.3	0.4	0.297	0.122	0.225	0.313	348	96.67	0.265
SHGT_b2	0.2	0.3	0.196	0.119	0.222	0.311	350	97.22	0.299
	0.2	0.4	0.200	0.118	0.221	0.309	352	97.78	0.283
	0.3	0.3	0.300	0.125	0.229	0.317	346	96.11	0.289
	0.3	0.4	0.297	0.126	0.233	0.323	351	97.50	0.274
SHGT_b3	0.2	0.3	0.198	0.114	0.213	0.300	347	96.39	0.308
	0.2	0.4	0.200	0.114	0.214	0.300	346	96.11	0.289
	0.3	0.3	0.294	0.118	0.224	0.300	348	96.67	0.291
	0.3	0.4	0.300	0.120	0.223	0.311	353	98.06	0.285
SHGT_c1	0.2	0.3	0.198	0.114	0.213	0.300	351	97.50	0.287
	0.2	0.4	0.200	0.110	0.206	0.290	354	98.33	0.282
	0.3	0.3	0.300	0.121	0.224	0.311	352	97.78	0.281
	0.3	0.4	0.300	0.120	0.222	0.309	349	96.94	0.268
SHGT_c2	0.2	0.3	0.200	0.113	0.212	0.298	352	97.78	0.300
	0.2	0.4	0.200	0.117	0.218	0.306	355	98.61	0.287
	0.3	0.3	0.300	0.116	0.215	0.300	351	97.50	0.289
	0.3	0.4	0.297	0.115	0.214	0.299	350	97.22	0.277
SHGT_c3	0.2	0.3	0.193	0.113	0.212	0.297	351	97.50	0.303
	0.2	0.4	0.198	0.107	0.201	0.282	351	97.50	0.291
	0.3	0.3	0.282	0.113	0.212	0.298	348	96.67	0.292
	0.3	0.4	0.300	0.125	0.232	0.323	358	99.44	0.287
RN	—	—	0.106	0.083	0.160	0.230	360	100	0.392

注: 同表 1

表 3 $r_{ab}=0.2$, 测验内容比例为 1:2:3 时, 5 种选题策略的结果

ISS	$r_{\max}$	$\bar{\Omega}_{\max}$	Max_ER	$\bar{\Omega}_1$	$\bar{\Omega}_2$	$\bar{\Omega}_3$	#used	Per.(%)	RMSE
SHGT1	0.2	0.3	0.200	0.175	0.319	0.438	225	62.50	0.282
	0.2	0.4	0.200	0.176	0.322	0.441	221	61.39	0.273
	0.3	0.3	0.300	0.254	0.440	0.575	190	52.78	0.270
	0.3	0.4	0.300	0.250	0.435	0.571	181	50.28	0.263
SHGT2	0.2	0.3	0.186	0.164	0.300	0.413	228	63.33	0.283
	0.2	0.4	0.200	0.173	0.316	0.434	221	61.39	0.276
	0.3	0.3	0.190	0.164	0.300	0.413	231	64.17	0.278
	0.3	0.4	0.266	0.227	0.400	0.530	199	55.28	0.269
SHGT3	0.2	0.3	0.200	0.114	0.213	0.299	352	97.78	0.344
	0.2	0.4	0.181	0.157	0.290	0.400	244	67.78	0.299
	0.3	0.3	0.300	0.116	0.214	0.300	351	97.50	0.324
	0.3	0.4	0.181	0.157	0.289	0.400	238	66.11	0.289
SHGT_a1	0.2	0.3	0.200	0.126	0.234	0.327	356	98.89	0.295
	0.2	0.4	0.200	0.127	0.236	0.329	351	97.50	0.291
	0.3	0.3	0.297	0.137	0.250	0.343	356	98.89	0.293
	0.3	0.4	0.300	0.139	0.253	0.347	353	98.06	0.284
SHGT_a2	0.2	0.3	0.200	0.126	0.234	0.325	356	98.89	0.311
	0.2	0.4	0.200	0.128	0.237	0.330	348	96.67	0.304
	0.3	0.3	0.300	0.142	0.258	0.354	356	98.89	0.307
	0.3	0.4	0.300	0.140	0.254	0.348	353	98.06	0.293
SHGT_a3	0.2	0.3	0.197	0.114	0.213	0.299	359	99.72	0.323
	0.2	0.4	0.200	0.128	0.238	0.331	352	97.78	0.314
	0.3	0.3	0.297	0.133	0.212	0.299	346	96.11	0.312
	0.3	0.4	0.296	0.137	0.251	0.345	353	98.06	0.307
SHGT_b1	0.2	0.3	0.200	0.120	0.223	0.312	354	98.33	0.296
	0.2	0.4	0.200	0.127	0.236	0.330	352	97.78	0.288
	0.3	0.3	0.296	0.130	0.240	0.332	352	97.78	0.285
	0.3	0.4	0.292	0.131	0.239	0.330	351	97.50	0.273
SHGT_b2	0.2	0.3	0.200	0.124	0.230	0.321	353	98.06	0.308
	0.2	0.4	0.200	0.126	0.234	0.325	357	99.17	0.298
	0.3	0.3	0.300	0.136	0.249	0.343	350	97.22	0.302
	0.3	0.4	0.289	0.137	0.250	0.344	355	98.61	0.287
SHGT_b3	0.2	0.3	0.198	0.114	0.213	0.300	359	99.72	0.320
	0.2	0.4	0.200	0.125	0.232	0.324	353	98.06	0.307
	0.3	0.3	0.295	0.134	0.213	0.300	355	98.61	0.311
	0.3	0.4	0.297	0.136	0.248	0.341	354	98.33	0.285
SHGT_c1	0.2	0.3	0.200	0.124	0.230	0.321	352	97.78	0.282
	0.2	0.4	0.200	0.122	0.227	0.317	355	98.61	0.280
	0.3	0.3	0.298	0.129	0.237	0.328	353	98.06	0.277
	0.3	0.4	0.300	0.128	0.236	0.326	351	97.50	0.264
SHGT_c2	0.2	0.3	0.200	0.122	0.227	0.315	353	98.06	0.291
	0.2	0.4	0.200	0.123	0.229	0.319	353	98.06	0.285
	0.3	0.3	0.289	0.132	0.240	0.331	348	96.67	0.279
	0.3	0.4	0.300	0.131	0.239	0.329	356	98.89	0.272
SHGT_c3	0.2	0.3	0.198	0.114	0.213	0.300	356	98.89	0.317
	0.2	0.4	0.200	0.120	0.223	0.312	353	98.06	0.294
	0.3	0.3	0.298	0.131	0.243	0.343	356	98.89	0.292
	0.3	0.4	0.292	0.136	0.247	0.340	352	97.78	0.277
RN	—	—	0.104	0.083	0.160	0.230	360	100	0.387

注：同表 1

表 4 $r_{ab}=0.8$, 测验内容比例为 1:2:3 时, 5 种选题策略的结果

ISS	$r_{\max}$	$\bar{\Omega}_{\max}$	Max_ER	$\bar{\Omega}_1$	$\bar{\Omega}_2$	$\bar{\Omega}_3$	#used	Per.(%)	RMSE
SHGT1	0.2	0.3	0.200	0.176	0.321	0.441	215	59.72	0.307
	0.2	0.4	0.200	0.173	0.318	0.436	216	60.00	0.294
	0.3	0.3	0.300	0.254	0.443	0.581	173	48.06	0.301
	0.3	0.4	0.300	0.254	0.441	0.578	176	48.89	0.279
SHGT2	0.2	0.3	0.184	0.164	0.300	0.413	227	63.06	0.344
	0.2	0.4	0.200	0.174	0.319	0.437	219	60.83	0.305
	0.3	0.3	0.184	0.163	0.300	0.413	222	61.67	0.299
	0.3	0.4	0.281	0.227	0.399	0.529	196	54.44	0.286
SHGT3	0.2	0.3	0.200	0.116	0.217	0.305	356	98.89	0.369
	0.2	0.4	0.179	0.157	0.289	0.400	233	64.72	0.307
	0.3	0.3	0.300	0.143	0.265	0.364	353	98.06	0.329
	0.3	0.4	0.181	0.157	0.289	0.400	234	65.00	0.308
SHGT_a1	0.2	0.3	0.200	0.145	0.267	0.369	347	96.39	0.316
	0.2	0.4	0.200	0.146	0.269	0.372	346	96.11	0.313
	0.3	0.3	0.300	0.182	0.323	0.431	343	95.28	0.298
	0.3	0.4	0.298	0.177	0.314	0.420	353	98.06	0.286
SHGT_a2	0.2	0.3	0.200	0.144	0.265	0.366	340	94.44	0.322
	0.2	0.4	0.200	0.142	0.262	0.363	350	97.22	0.317
	0.3	0.3	0.297	0.169	0.299	0.405	344	95.56	0.311
	0.3	0.4	0.300	0.176	0.312	0.418	351	97.50	0.301
SHGT_a3	0.2	0.3	0.198	0.113	0.212	0.299	351	97.50	0.346
	0.2	0.4	0.200	0.150	0.275	0.379	341	94.72	0.331
	0.3	0.3	0.284	0.114	0.213	0.299	354	98.33	0.323
	0.3	0.4	0.300	0.163	0.294	0.397	343	95.28	0.314
SHGT_b1	0.2	0.3	0.200	0.124	0.230	0.321	345	95.83	0.300
	0.2	0.4	0.200	0.127	0.236	0.328	351	97.50	0.290
	0.3	0.3	0.300	0.140	0.255	0.351	355	98.61	0.291
	0.3	0.4	0.293	0.136	0.248	0.341	352	97.78	0.279
SHGT_b2	0.2	0.3	0.200	0.120	0.224	0.314	351	97.50	0.305
	0.2	0.4	0.200	0.124	0.231	0.322	353	98.06	0.285
	0.3	0.3	0.295	0.129	0.237	0.329	353	98.06	0.293
	0.3	0.4	0.300	0.137	0.251	0.345	353	98.06	0.282
SHGT_b3	0.2	0.3	0.194	0.114	0.213	0.299	353	98.06	0.311
	0.2	0.4	0.200	0.127	0.237	0.330	347	96.39	0.295
	0.3	0.3	0.285	0.114	0.213	0.300	357	99.17	0.291
	0.3	0.4	0.296	0.145	0.263	0.360	353	98.06	0.291
SHGT_c1	0.2	0.3	0.200	0.126	0.234	0.327	353	98.06	0.285
	0.2	0.4	0.200	0.124	0.230	0.321	349	96.94	0.274
	0.3	0.3	0.300	0.131	0.241	0.332	343	95.28	0.273
	0.3	0.4	0.300	0.133	0.243	0.335	352	97.78	0.268
SHGT_c2	0.2	0.3	0.200	0.122	0.227	0.316	347	96.39	0.291
	0.2	0.4	0.198	0.123	0.229	0.319	351	97.50	0.285
	0.3	0.3	0.300	0.132	0.241	0.333	348	96.67	0.279
	0.3	0.4	0.298	0.131	0.241	0.334	354	98.33	0.271
SHGT_c3	0.2	0.3	0.192	0.114	0.213	0.300	355	98.61	0.308
	0.2	0.4	0.200	0.123	0.228	0.319	354	98.33	0.302
	0.3	0.3	0.297	0.114	0.213	0.300	350	97.22	0.293
	0.3	0.4	0.296	0.130	0.240	0.332	343	95.28	0.278
RN	—	—	0.101	0.083	0.160	0.230	360	100	0.386

注：同表 1

果, 在 SHGT1 下, 当 $r_{\max} = 0.2$, $\bar{\Omega}_{\max} = 0.3$ 时, 项目最大曝光率等于 0.2 , $\bar{\Omega}_1 = 0.177$, 远小于 $\bar{\Omega}_{\max}$, 但 $\bar{\Omega}_2$ 和 $\bar{\Omega}_3$ 的值超过了 $\bar{\Omega}_{\max}$, 这是因为此时共享人数为 1 人, 因此对 $\bar{\Omega}_1$ 的控制比 $\bar{\Omega}_2$ 和 $\bar{\Omega}_3$ 要好得多。同理, 当共享人数为 2 人时, 对 $\bar{\Omega}_1$ 和 $\bar{\Omega}_2$ 的控制较为理想。特别地, 不论实验条件如何, 当共享人数为 2 人或 3 人时, SHGT 法对项目曝光率均会出现过度控制(over control)现象, 即项目最大曝光率会远远低于 $r_{\max}$, 导致 RMSE 值增大, 并且可以看出, 该方法对题库的使用率并不高, 尤其是当共享人数为 1 人或 2 人时, 最高题库使用率只有 2/3 左右, 意味着有 1/3 的题量未得到运用, 我们知道, 编制一个项目的实际费用是相当昂贵的, 尤其是当题库中项目数量较大时(例如上千题), 这一比例将十分严重。这一研究结果与 Chen (2010)研究结果一致。由于 3 种新方法采用了分层技术, 使得最大曝光率得以提升, 在任何实验条件下均未出现对项目曝光率的过度控制现象, 从而提升了能力估计的精度(RMSE 值略有下降)。相较于 SHGT 法, 不论实验条件如何, 3 种新方法均能极大地提高题库使用率(均为 94%以上), 从而保证了测验及题库安全性, 减小了题库建设的经济耗损, 这一提升对实际应用的意义不言而喻。作为基线的随机选题策略(RN), 对测验及题库安全性的控制是最强的, 题库使用率高达 100%, 不论实验条件如何变化, 项目最大曝光率大约为 0.1, 接近其下限值 0.083。和其余 4 种方法相比, RN 选题策略对广义测验重叠率的控制最好, 但对被试能力估计的精度是最差的。

具体来说, 表 1 和表 2 是固定测验内容比例为 1:1:1 时, 5 种选题策略的不同表现。从实验结果可以看出, 随着 $r_{\max}$ 和 $\bar{\Omega}_{\max}$ 的增大, 不论采用何种方法, RMSE 值均会逐渐降低, 能力估计精度逐渐提高, 但题库使用率逐渐下降; 随着共享人数由 1 人增加至 3 人时, 不论采用何种方法, RMSE 值均会逐渐增大, 能力估计精度逐渐降低, 这跟实际情况相符, 因为在现实中, 若分享项目数量人数较多, 必然会影响到被试能力估计的精度。在 SHGT 方法下, 题库使用率逐渐上升, 而在新方法下, 题库使用率均很高; 当 $r_{ab} = 0.2$ 时, SHGT_b 和 SHGT_c 方法的 RMSE 值均较低, 稍好于 SHGT_a 方法, 但当 r_{ab} 增大至 0.8 时, SHGT_a 方法在能力估计精度方面的表现不如其在 $r_{ab} = 0.2$ 时的表现, SHGT_b 和 SHGT_c 方法在能力估计精度方面的表现与 $r_{ab} = 0.2$ 时相差无几, 且均好于 SHGT_a 方法。这

表明随着区分度和难度相关水平的增加, 按照 b 分块思想得到的结果更加精确, 这也是 SHGT_b 和 SHGT_c 法相较于 SHGT_a 法的优势所在。

表 3 和表 4 是固定测验内容比例为 1:2:3 时, 5 种选题策略的不同表现。此时能够得到与测验内容比例为 1:1:1 时相似的结果, 即随着 $r_{\max}$ 和 $\bar{\Omega}_{\max}$ 的增大以及共享人数 α 的减小, 采用 SHGT 法或 3 种新方法, 得到的 RMSE 值均比较接近, 且呈下降趋势, 能力估计精度逐渐提高; 题库使用率方面, 在 SHGT 方法下呈下降趋势, 而在新方法中仍保持了很高的题库使用率, 显示出了融合分层技术后, 新的选题策略的优势: 即能够在同时控制项目曝光率及广义测验重叠率的同时, 极大地提高题库的使用率。当 r_{ab} 增大时, SHGT_b 和 SHGT_c 法在能力估计精度方面均要优于 SHGT_a 法。当测验考察的内容比例发生变化时, 3 种新方法以及 SHGT 法的 RMSE 值均有下降, 但变化不显著。其中, SHGT_c 相较于其他方法而言, RMSE 值的下降幅度最小。

5 小结与讨论

测验安全性与题库使用均匀性是 CAT 非常关注的问题, 不同的选题策略各具特点。CAT 中发展较早, 比较经典的选题策略为极大费歇信息量法(Maximum Fisher Information, MFI), 该策略能够最精确地估计被试的能力水平, 但其题库使用率非常差(Yi & Chang, 2003), 它是一种对题库和测验毫无保护措施的选题策略。为了能够解决 MFI 带来的问题, 研究者们提出了许多控制项目曝光率的方法, 如引言中提到的 5 类项目曝光控制方法。但只控制曝光率是远远不够的。正如 Way (1998)认为的, 在 CAT 中, 项目曝光率和测验重叠率是两个反映测验安全性的指标。基于此, Chen (2010)提出了兼具同时控制项目曝光率和广义测验重叠率功能的 SHGT 法, 但该方法的题库使用率较差, 特别是当 α 较小, 或 $r_{\max}$ 和 $\bar{\Omega}_{\max}$ 较大时。

本研究借鉴 a 分层方法的思想, 成功地将 SHGT 法与不同形式的 a 分层法相结合, 在保留各自优势的前提下, 相互弥补了缺陷。SHGT 法在控制广义测验重叠率的同时, 解决了项目过度曝光问题, a 分层法可以有效提高题库使用率, 保证了测验安全性。

模拟研究结果表明, SHGT 法及 3 种新方法均能很好地实现对项目曝光率和广义测验重叠率的控制, 并极大地提高题库使用率。特别地, 3 种新方

法在任何实验条件下均未出现对项目曝光率过度控制的现象,从而提升了能力估计的精度,比起 SHGT 法更具优势。具体来说,当 $r_{\max}$ 和 $\bar{\Omega}_{\max}$ 增大以及共享人数 a 减小时,SHGT 法以及 3 种新方法得到的 RMSE 值均比较接近,呈下降趋势,能力估计精度逐渐提高,并且在 SHGT 方法下题库使用率呈下降趋势,而在新方法下中仍然保持了很高的题库使用率,最大可提升近 1/3,充分显示了新方法的优越性。当区分度和难度相关较低时,SHGT_b 和 SHGT_c 法在能力估计精度方面均稍好于 SHGT_a 法,但不明显;当区分度和难度相关较高时,SHGT_b 和 SHGT_c 法在能力估计精度方面均优于 SHGT_a 法,提高较明显。如果某一题库中的 r_{ab} 较高时,建议使用 SHGT_b 和 SHGT_c 法。Chang 和 Ying (1999)认为, a 分层方法有较好应用的重要前提是参数 a 和 b 之间没有相关,因此,当 a 和 b 之间具有高相关时,SHGT_b 和 SHGT_c 法表现更好也在意料之中。在实际测验中,施测者往往会根据测验目的设置不同的测验考察内容比例。在测验内容比例发生变化时,4 种方法在能力估计精度方面均有较稳定的表现,其中以 SHGT_c 法的 RMSE 值最小。因此,在测验内容比例不均衡时,建议选用 SHGT_c 法。本文虽然很好的将不同的 a 分层方法与 SHGT 法结合,取得了令人满意的结果,但仍有继续可以研究的地方:

(1)本研究只采取了在 a 分层的每一层内选取近似相等的项目数量,没有考察升序的实验条件。根据已有研究表明,采用升序的 a 分层效果更佳 (Chang & Ying, 1999; Chang & Ying, 1996, 2008; Hau & Chang, 2001),这在以后研究中可以进行探讨;

(2)本研究发现,在测验考察内容比例不均衡条件下,新方法均有较稳定的表现,尤以 SHGT_c 法表现最好。但这是在题库及内容数量相对较小,测验长度固定为 30 题时的结果。今后可以研究在不同题库容量、不同内容领域数量及比例条件下,新方法的表現;

(3) CAT 的优势在于可以对每个被试的能力估计精度进行控制,这时就需要采用变长的 CAT。具体做法可以根据每层内达到的信息量值作为变长 CAT 的标准 (Wen, Chang, & Hau, 2000; 戴海琦, 陈德枝, 丁树良, 邓太萍, 2006);

(4)程小扬等人 (2011)提出了引入曝光因子的 CAT 选题策略,该方法使题库中项目的调用更加均匀,曝光率指标明显降低,能力估计精度也较高。

将该选题策略与本文提出的方法进行比较也是值得研究的方向。

参 考 文 献

- Chang, H. H., Qian, J. H., & Ying, Z. L. (2001). A-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25(4), 333-341.
- Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3), 213-229.
- Chang, H. H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222.
- Chang, H. H., & Ying, Z. L. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73(3), 441-450.
- Chang, H. H., & Zhang, J. M. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67(3), 387-398.
- Chen, P. (2011). *Item replenishing cognitive diagnostic computerized adaptive testing—based on DINA model*. Unpublished doctoral thesis, Beijing Normal University.
- [陈平. (2011). 认知诊断计算机化自适应测验的项目增补——以 DINA 模型为例. 博士学位论文, 北京师范大学.]
- Chen, S. Y. (2010). A procedure for controlling general test overlap in computerized adaptive testing. *Applied Psychological Measurement*, 34(6), 393-409.
- Chen, S. Y., & Ankenman, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, 41(2), 149-174.
- Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40(2), 129-145.
- Chen, S. Y., & Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, 29(3), 204-217.
- Chen, S. Y., & Lei, P. W. (2010). Investigating the relationship between item exposure and test overlap: Item sharing and item pooling. *British Journal of Mathematical and Statistical Psychology*, 63(1), 205-226.
- Chen, S. Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 61(2), 471-492.
- Cheng, X. Y., Ding, S. L., Yan, S. H., & Zhu, L. Y. (2011). New item selection criteria of computerized adaptive testing with exposure-control factor. *Acta Psychologica Sinica*, 43(2), 203-212.
- [程小扬, 丁树良, 严深海, 朱隆尹. (2011). 引入曝光因子的计算机化自适应测验选题策略. *心理学报*, 43(2), 203-212.]
- Dai, H. Q., Chen, D. Z., Ding, S. L., & Deng, T. P. (2006). The comparison among item selection strategies of CAT with multiple-choice items. *Acta Psychologica Sinica*, 38(5), 778-783.
- [戴海琦, 陈德枝, 丁树良, 邓太萍. (2006). 多级评分题计算机自适应测验选题策略比较. *心理学报*, 38(5), 778-783.]

- Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8), 4–37.
- Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, 38(3), 249–266.
- Lord, F. M. (1975). The ‘ability’ scale in item characteristic curve theory. *Psychometrika*, 40(2), 205–217.
- Parshall, C., Harmes, J. C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research*, 40(1), 28–52.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(4), 311–327.
- Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94–5). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292.
- Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the Proceedings of the 27th annual meeting of the Military Testing Association. San Diego.
- Tang, X. J., Ding, S. L., & Yu, Z. H. (2012). Application of computerized adaptive testing in cognitive diagnosis. *Advances in Psychological Science*, 20(4), 616–626.
- [唐小娟, 丁树良, 俞宗火. (2012). 计算机化自适应测验在认知诊断中的应用. *心理科学进展*, 20(4), 616–626.]
- van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 27–52). Norwell MA: Kluwer.
- van der Linden, W. J. (2003). Some alternatives to Sympon–Hetter item–exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 28(3), 249–265.
- Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17(4), 17–27.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53(6), 774–789.
- Wen, J. B., Chang, H. H., & Hau, K. T. (2000). *Adaptation of a-stratified method in variable length computerized adaptive testing*. Paper presented at the American Educational Research Association Annual Meeting. Seattle.
- Yi, Q., & Chang, H. H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56(2), 359–378.

a-Stratified Methods Combining Item Exposure Control and General Test Overlap in Computerized Adaptive Testing

GUO Lei¹; WANG Zhuoran¹; WANG Feng¹; BIAN Yufang^{1,2}

(¹ National Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China)

(² National Cooperative Innovation Center for Assessment and Improvement of Basic Education Quality, Beijing Normal University, Beijing 100875, China)

Abstract

Test security and item pool utilization rate are very important in computerized adaptive testing (CAT), especially in the high-stakes tests. Most existing methods only focus on the item exposure rate, but rarely control the test overlap rate. Way (1998) suggested that the item exposure and the test overlap rate be two indices of test security. Following this reasoning, Chen (2010) proposed an on-line version of the Sympon-Hetter procedure with general test overlap control (SHGT) that didn't need iterative simulations.

Although the SHGT method could control item exposure and general test overlap simultaneously without iterative simulations, the item pool utilization rate was not very ideal when the item exposure or test overlap rate was slightly high or the number of examinees who shared the information with another examinee was small. Thus, the test security was threatened. To address the limitation of the SHGT method, we combined the a-stratified method with the SHGT method, and proposed three new methods: SHGT_a method, SHGT_b method, and SHGT_c method.

Simulation results indicated that: (1) Compared to the SHGT method, these three new methods could not

only improve the item pool utilization rate, but also maintain a very high precision of ability estimate in the same experiment condition; (2) With the increase of maximum item exposure rate ($r_{\max}$) and maximum general test overlap rate ($\bar{\Omega}_{\max}$), and the decrease of α , the precision of ability estimation increased. Compared to the SHGT method, these three new methods could maintain a higher item pool utilization rate; (3) SHGT_b and SHGT_c outperformed SHGT_a in the aspect of the precision of ability estimation when r_{ab} was high; (4) The three new methods had a good performance in respect of the precision of ability estimation; (5) Compared to the SHGT method, these three new methods could also effectively solve the problem of item exposure over-control.

Some future directions of study were suggested at the end of this paper.

Key words item exposure rate; general test overlap rate; computerized adaptive testing; a-stratified method; selection strategy