心理学报　2019, Vol. 51, No. 9, 1057−1067
Acta Psychologica Sinica　DOI: 10.3724/SP.J.1041.2019.01057

让自适应测验更知人善选 ——基于推荐系统的选题策略*

王璞珏 [1]　刘红云 [1,2]

([1] 北京师范大学心理学部; [2] 北京师范大学心理学部应用实验心理北京市重点实验室, 北京 100875)

摘　要　基于推荐系统中协同过滤推荐的思想, 提出两种可以利用已有答题者数据的 CAT 选题策略：直接基于答题者推荐(DEBR)和间接基于答题者推荐(IEBR)。通过两个模拟研究, 在不同题库和不同长度的测验中, 比较了两种推荐选题策略与两种传统选题策略(FMI 和 BAS)在测量精度和对题目曝光率控制上的表现, 以及影响推荐选题策略表现的因素。结果发现：两种推荐选题策略对题目曝光率的控制优于两种传统选题策略, 测量精度不亚于 BAS 方法, 其中 DEBR 侧重选题精度, IEBR 对题目曝光率控制最好。已有答题者数据的特点和质量是影响推荐选题策略表现的主要因素。

关键词　选题策略; 已有答题者数据; 推荐系统; 协同过滤推荐; 模拟研究

分类号　B841

1　引言

计算机自适应测验(Computerized Adaptive Testing, CAT)基于一定的选题策略, 为不同能力的答题者提供不同难度的题目, 用一套“量身定制”的测验更准确高效地测量出每名答题者的真实能力(Weiss, 1982)。随着智慧学习和智慧测验的推广和流行, CAT 的应用范围愈加广泛(Zhang & Chang, 2016), 随之产生了大量答题者完成测验后留下的过程性数据。从数据挖掘的角度来看, 这些数据中蕴含了丰富的信息, 包括作答结果、过程中能力估计值和下一道题目之间的映射关系, 通过合适的技术手段从中可以挖掘出有用的模式, 预测未知的结果(Tan, Steinbach, & Kumar, 2006), 即抽象和建立一套新的选题规则, 既可以重现与产生数据所用策略相近的选题精度, 还可以根据现有选题数据中暴露出的问题(如常见的题库使用不均匀), 动态地调整这套规则, 弥补原有策略的不足。毛秀珍和辛涛(2011)指出 CAT 选题策略发展至今, 一个重要的改进方向是如何充分利用答题者的先验信息。对于每一个正在完成测验的答题者而言, 已有答题者数据正是一类具有重要价值但被长期被忽视的先验信息来源。由于 CAT 的提出和发展主要基于项目反应理论(Item Response Theory, IRT; Chang, 2015), 在现有的 IRT 框架下提出的选题策略大多仅利用了当前答题者的作答信息, 难以将已有答题者数据纳入 CAT 选题的考虑范围中, 也就难以灵活和直接地从他人数据中学习选题经验并改进选题策略。

如何构建更智慧的辅助学习和测验系统, 进一步实现自适应的目标, 是一个跨学科的问题, 需要心理学、教育学、统计学和机器学习等多领域专业知识和技术手段的融合(Chen, Li, Liu, & Ying, 2018; Zhang & Chang, 2016)。具体到如何使用已有答题者数据改进 CAT 选题策略, 鉴于上述传统选题策略的局限性, 同样可以尝试在 IRT 的理论基础之上引入全新的技术手段, 推荐系统是一个合适的选择。

收稿日期: 2018-06-10

* 国家自然科学基金项目(31571152)、北京市与中央在京高校共建项目(019-105812)和国家教育考试科研规划 2017 年度课题(GJK2017015)的资助。

通信作者: 刘红云, E-mail: hyliu@bnu.edu.cn

推荐系统(Recommender Systems)是一系列利用已有数据为用户进行项目推荐的算法和技术，可以根据用户的需求给出精准的匹配，是数据挖掘领域的热门研究课题(Ricci, Rokach, & Shapira, 2015)，诸多成熟的算法已在商业、文娱、社交等应用领域得了巨大的成功(Covington, Adams, & Sargin, 2016; Quijano-Sánchez, Recio-García, Díaz-Agudo, & Jiménez-Díaz, 2011; Smith & Linden, 2017)。应用于教育领域中，推荐系统可以利用大规模的已有学习数据，预测学生在新题目上的作答表现，准确率优于传统方法(Thai-Nghe, Drumond, Krohn-Grimberghe, & Schmidt-Thieme, 2010)。近年来快速发展的在线学习(e-Learning)正是借助推荐系统为数以万计的学习者设计出具有个性化的学习计划(刘淇等, 2018; Klašnja-Milićević, Ivanović, & Nanopoulos, 2015)。由此可见，推荐系统可以为如何利用已有答题者数据选题提供可行的方案。

推荐系统还可以与 IRT 相结合，构建起同样注重适应性的智能学习系统。朱天宇等人(2017)将 DINA 模型与矩阵分解技术整合为一套协同过滤的试题推荐方法，得以同时完成对知识掌握程度的估计和题目的推荐，推荐效果优于使用单一的认知诊断模型或数据挖掘算法。Chen 等人(2018)将推荐系统、多维 IRT 模型和强化学习三者结合，提出两个适应性学习系统的原型，使用该系统选择学习材料比随机选择以两种统计指标衡量都有更高的效率，并指出适应性学习的核心成分应当是一个推荐系统，依据学习成绩推测潜在的知识掌握状态，选择适合该状态的学习材料。可以发现，这与适应性测验的核心流程十分相似，即根据作答结果选择最合适答题者真实水平的测验题目。换而言之，CAT 选题策略本质上也可视为一个推荐系统。然而，将推荐系统与 CAT 选题结合尚无先例。只要找到合适的推荐技术，便可以弥补这一空白。

推荐系统中协同过滤推荐(Collaborative Filtering Recommender)正是利用大量的已有用户数据，对当前用户的喜好做出预测和推荐，与利用已有答题者数据为当前答题者选题的目标不谋而合。协同过滤推荐假设如果两个用户过往对相同的项目感兴趣，他们可能在未来仍有相似的偏好，从而过滤出最贴近用户喜好的项目进行推荐(Pirasteh, Jung, & Hwang, 2014)。协同过滤推荐简单易行，不需要训练模型，其底层假设在大量场景中经验证稳定有效，是推荐系统中最为成熟和流行的一类推荐方法(Koren & Bell, 2015)。使用协同过滤推荐完成 CAT 选题，可以避免传统选题策略复杂的计算公式和约束流程，从已有答题者数据中快速筛选出适合当前答题者作答的题目。此外，在协同过滤推荐的假设之上可以根据研究者需要加入其它规则，设计出可灵活扩展的选题策略，既可以侧重选题精度或题目曝光率控制，也可以在保证一定精度的情况下兼顾题库使用和测验安全。例如，计算已有答题者在已作答题目上的相似性，借助某种推荐算法过滤出若干道适合当前答题者的备选题目，首先满足选题精度的要求，同时使用某种曝光控制法，从备选题目池中选出最终要作答的题目，这样便兼顾了题库的均匀使用。

基于上述分析，本研究旨在将推荐系统中的协同过滤推荐用于 CAT 选题，提出可以利用已有答题者数据的全新选题策略(以下简称推荐选题策略)。然后通过蒙特卡洛模拟研究，在不同条件下考察推荐选题策略在选题精度和对题目曝光率控制方面的表现。

2 选题策略

2.1 生成第一批数据的传统选题策略

大量可靠的用户历史数据是精准推荐的前提和保障，在 CAT 中便对应着已有答题者数据。同理可知，如果过往答题者作答的题目都不符合其真实能力，数据库中积累了大量低测量精度的选题数据，则很难预期推荐选题策略可以从中找到正确的选题规律，为新答题者选出合适的题目。除了选题精度，CAT 选题策略还应注意对题目曝光率的控制。如果过往的选题策略没有充分使用整个题库，使产生的答题数据中题目曝光失衡，那么推荐系统选题策略可能会受到影响，按已有不均衡的比例选择题目。

现阶段我们首先需要使用研究成熟且特点鲜明的传统选题策略，生成特点不同的第一批已有答题者数据，以考察推荐选题策略的选题特点。第一种选用的策略是 Lord (1980)提出的基于最大 Fisher 信息量(Maximum Fisher Information, MFI)选题方法，该方法通过最大化测验信息量的方式提高选题精度，是最为流行的 CAT 选题策略，但在题目曝光率控制方面存在缺陷(Chang, 2015)。第二种策略是 Chang, Qian 和 Ying (2001)提出的按 b 分块的 a-分层策略(a-Stratified Strategy with b Blocking, BAS)，该方法在测验初期提高了低区分度题目的曝光率，同时减少了过度曝光的题目数。此外，分层方法生

成的已有答题者数据会继续保留分层的特点, 使推荐选题策略的搜索范围可以缩小在特定层之内, 可提高选题速度。

2.2　基于协同过滤推荐的新选题策略

协同过滤推荐有两种主要的实现方式：基于用户的协同过滤(User-Based Collaborative Filtering, 例如 Jia, Yang, Gao, & Chen, 2015)会寻找与当前用户喜好最相似的用户, 然后在相似用户的过往数据中寻找项目推荐给当前用户; 基于项目的协同过滤(Item-Based Collaborative Filtering, 例如 Pirasteh et al., 2014)则试图在项目库中寻找与当前用户喜好项目最相似的项目, 将其推荐给当前用户。考虑到已有答题者数据中答题者的数量一般会多于题库中的题目数量, 寻找相似答题者更加容易, 而且随着已有答题者数量增大可获得更多的参考信息, 更利于找到最合适的题目, 因此以基于用户的协同过滤推荐的思想设计选题策略, 将寻找相似答题者作为实现推荐选题的第一步。每当答题者完成一道题, 就在已有数据中寻找作答过相同题目且作答结果一致的已有答题者, 将其选定为本道题的相似答题者, 以他们为参考群体进行下一道题目的推荐。与推荐系统中常用的余弦相似度不同, 由于本研究暂仅关注 0-1 计分的题目, 相似答题者的判定仅有对或错两种结果, 也就是以简单的二分方式而非连续尺度计算答题者的相似度, 计算复杂度低, 速度更快。每次选出的相似答题者仅针对当前题目而言, 非相似答题者仍有可能在下一道题目答完后被判定为相似答题者, 这样设计可以扩大一次完整 CAT 对已有答题者数据的参考范围, 使推荐选题策略可利用的信息更多, 选题更加精准。

找到相似答题者后, 可改良协同过滤推荐的底层假设使其适用于 CAT 场景。一种改良的假设是：当前答题者可以作答与相似答题者相同的下一道题目, 这样便得到一种直接的推荐选题策略, 不借助题目参数完成选题。另一种假设是：相似答题者与当前答题者会有相似的能力值, 然后借助题目参数完成选题, 这样便得到一种间接的推荐选题策略。基于这两种假设都可能找到多道可推荐的题目, 考虑到已有答题者数据可能存在题目曝光不均匀的问题, 最终的题目将以随机选择的方式产生, 随机化操作是一类常用的可以控制题目曝光率的方法(Georgiadou, Triantafillou, & Economides, 2007)。至此形成两种推荐选题策略：直接的基于答题者推荐(Direct Examinee-Based Recommender, DEBR)将所有相似答题者回答过的下一道题目与当前答题者未作答题目的交集作为备选题目, 从中随机抽取一题作为当前答题者的下一道题目。间接的基于答题者推荐(Indirect Examinee-Based Recommender, IEBR)将统计所有相似答题者答完本题后的当前能力估计值的范围, 将当前答题者未作答题目中难度参数 b 位于该范围中的题目作为备选题目, 从中随机抽取一题作为当前答题者的下一道题目。将能力估计值与 b 参数匹配选题的操作借鉴了分层方法, 使用匹配 b 参数的方法相比于 FMI 不但运算复杂度低, 可提高选题速度, 而且在不损失估计精度的情况下对题目曝光率控制更好(Chang & Ying, 1999)。

在较少情况下, 上述两种推荐选题策略可能找不到可推荐的题目, 可称为选题失败。由于协同过滤推荐仅在选题过程中使用, CAT 中其它流程仍照常进行, 包括使用参数估计的方法得到该答题者作答每一道题后的当前能力估计值。当找不到可推荐题目时, 将使用当前答题者的能力估计值匹配 b 参数选择下一道题目。除了前文所述匹配 b 参数的优点, 如果生成已有答题者数据的策略不注重题库的均匀使用, 存在部分题目从未在过往数据中出现过, 该方法还可重新启用该题目, 提高对低曝光题目的使用。综上可见, 本文提出的两种推荐选题策略都使用了简捷快速的操作, 在保证选题精度的情况下尽可能注重对题目曝光率的控制。

3　研究 1：基于模拟题库和不同终止条件的选题策略比较

3.1　研究设计

研究 1 将探究两个常见的影响 CAT 选题和推荐系统的因素。首先, 选择不同的传统选题策略, 生成不同特点的已有答题者数据, 是否会影响两种推荐选题策略的表现？模拟条件为两种选题策略：侧重测量精度的 FMI 方法和侧重控制题目曝光的 BAS 方法。其次, 采用不同长度的测验, 生成不同数量的已有答题者数据, 是否会影响两种推荐选题策略的表现？模拟条件为定长 20 道题目和 40 道题目两种终止条件。研究一共 2×2=4 种模拟条件的组合, 每种条件组合下重复 100 次。

研究 1 使用的模拟题库为 400 道 0-1 计分的题目, 全部基于三参数 Logistic 模型(3PLM), 题目参数与常见策略比较的设定一致(Barrada, Olea, & Abad, 2010; Cheng, Patton, & Shao, 2015), 区分度参数 a 服从正态分布 $N\,(1.2, 0.25)$, 难度参数 b 服从

标准正态分布 $N(0, 1)$, 猜测参数 c 服从正态分布 $N(0.25, 0.02)$, a 参数与 b 参数存在中等程度的正相关($r = 0.45$)。答题者真实能力参数 θ 服从标准正态分布 $N(0, 1)$。研究 1 的模拟流程为：首先使用传统选题策略对第一批的 1000 名答题者进行 CAT 模拟, 生成第一批已有答题者数据, 然后使用推荐选题策略结合第一批已有答题者数据, 对第二批的 1000 名能力分布相同的答题者进行 CAT 模拟。能力估计方法均为贝叶斯后验期望法。使用 BAS 策略时, 题库分为 4 层, 每层含 100 道题, 每名答题者在每层作答 5 或 10 道题后进入下一层。在两种测验长度的条件下加入随机选择题目作为测量精度和曝光率控制的比较基线。

3.2 评价指标

本研究将使用七种 CAT 选题策略比较中常见的评价指标(He, Diao, & Hauser, 2014), 对答题者真实能力的测量精度和对题目曝光率控制的情况进行评价。同时提出一种新指标, 用于衡量推荐选题策略对已有答题者数据的使用情况。每种模拟条件下的最终结果为 100 次模拟的均值。不同评价指标的定义如下：

(1) 均方误差(Mean Square Error, MSE)：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2 \tag{1}$$

其中 $\hat{\theta}_i$ 第 i 名答题者的最终能力估计值, θ_i 为第 i 名答题者的真实能力值, N 为一批答题者的数量。

(2) 平均绝对误差(Mean Absolute Error, MAE)：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(\hat{\theta}_i - \theta_i\right)\right| \tag{2}$$

(3) 真实能力值与最终能力估计值的相关 $r_{\theta,\hat{\theta}}$：

$$r_{\theta,\hat{\theta}} = \frac{\sum_{i=1}^{N}\left(\theta_i - \bar{\theta}\right)\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)}{S_\theta S_{\hat{\theta}}} \tag{3}$$

其中 $\bar{\theta}$ 和 S_θ 为所有答题者真实能力值的均值和标准差, $\bar{\hat{\theta}}$ 和 $S_{\hat{\theta}}$ 为所有答题者最终能力估计值的均值和标准差。

(4) 题目实际曝光率与理想分布的卡方值(χ^2)：

$$\chi^2 = \sum_{i=1}^{N}\frac{\left(r_i - L/K\right)^2}{L/K} \tag{4}$$

其中 r_i 为第 i 道题目的曝光率, L 为测验长度, K 为题库容量(Chang & Ying, 1999)。

(5) 测验重叠率(Overlap Rate, OR), 定义为任意两个答题者作答题目相同的比率：

$$\mathrm{OR} = \frac{K}{L}\times S_r^2 + \frac{L}{K} \tag{5}$$

其中 S_r^2 为所有题目曝光率 r_i 的方差(Chen, Ankenmann, & Spray, 2003)。

(6) 曝光不足(Underexposed), 定义为没有使用过的题目数。

(7) 曝光过度(Overexposed), 定义为曝光率大于 20%的题目数。

(8) 答题者调用率(Utilization Rate of Examinees), 定义为推荐策略每次选题时调用的相似答题者数量占全部已有答题者的比例。

具体而言, 八种指标中测量精度的评价指标有三种：均方误差、平均绝对误差和能力估计相关; 题库使用的评价指标有两种：卡方值和曝光不足的题目数; 测验安全的评价指标有两种：测验重叠率和曝光过度的题目数; 答题者调用率则用于评价推荐选题策略可以利用多少已有答题者的作答信息为当前答题者寻找合适的题目。

3.3 研究结果

在定长 20 道题目的 CAT 中, 两种传统选题策略生成的已有答题者数据的特点与预期一致：FMI 的测量精度最高, 但题库使用不均匀; BAS 的测量精度稍低, 在测验安全和题库使用方面更好(见表 1)。使用 FMI 生成的已有答题者数据时, DEBR 的测量精度较高, 与 FMI 相比仅有小幅下降, 优于 BAS 和 IEBR, 且大幅改善了题目使用不均匀的问题。IEBR 对题目曝光率的控制最佳, 在题库使用和测验安全的四种指标上全部优于其它策略, 在保证一定精度的前提下(高于随机选择题目的测量精度)最为理想地均匀使用整个题库。由于答题者调用率不受测验阶段影响, 故计算完成一次测验所有步骤的平均值, 此时 DEBR 的答题者调用率远高于 IEBR。使用 BAS 生成的已有答题者数据时, 两种推荐选题策略与 BAS 相比精度稳定不变, 且都可以在已有基础上进一步优化测验安全和题库使用, 答题者调用率也基本相同。

在定长 40 道题目的 CAT 中, 传统选题策略和推荐选题策略表现出的特点与 20 道题目时基本一致。使用 FMI 生成的数据时, DEBR 损失较小的精度, 大幅减少了曝光不足的题目数, IEBR 的选题精度与 BAS 持平, 在测验安全和题库使用的四种指标上再次达到了该条件下的最优水平。使用 BAS 生成的数据时, 两种推荐选题策略几乎没有损失精度, 题库使用的均匀程度仍有提升, IEBR 的提升幅度

稍高于 DEBR。在更长的测验中, DEBR 和 IEBR 的答题者调用率整体升高, 呈现出的相对高低趋势不变。

由研究 1 的结果可见, 由不同传统策略生成的不同特点的已有答题者数据会直接影响推荐选题策略表现出的趋势。如果使用 FMI 生成第一批已有答题者数据, 推荐选题策略的表现为大幅启用未曝光的题目, 改善题目曝光率控制, 且产生常见的权衡损失一定精度, DEBR 权衡的幅度小于 IEBR; 如果使用 BAS 生成第一批已有答题者数据, 已有数据中题库使用较为均匀, 两种推荐选题策略都将保持精度并进一步改善题目曝光率控制, 包括答题者调用率在内的各指标十分接近。测验长度不影响新策略在特定数据下表现出的趋势, 但会影响在各指标上的绝对大小, 包括更高的精度和答题者调用率, 更少的曝光不足等。

当生成数据的传统策略相同而测验长度不同时, 同一推荐选题策略的表现可以有较大差异, 这种不一致性源自测验长度有两种作用路径, 既可能通过影响传统策略的表现改变已有答题者数据的质量(在各指标上的绝对大小), 也可能是通过生成数据的数量最终影响到推荐选题策略的表现, 于是需要控制测验长度, 用另一种增加数据量的方式分离上述影响。此外, 在研究 1 中已有答题者数据全部由传统选题策略生成, 而在现实中第二批答题者作答结束后, 推荐选题策略便可以使用自身生成的数据, 此时选题的结果是否稳定值得探究。研究 1 仅使用了模拟题库, 还需要在真实题库下进一步考察推荐选题策略的表现。上述问题将在研究 2 中进一步探讨。

4　研究 2：基于真实题库和合并数据的选题策略比较

4.1　研究设计

研究 2 将在更接近现实的情境下考察推荐选题策略的表现。首先, 换用真实题库, 当题库质量不如模拟题库理想时, 推荐选题策略的表现是否会受到影响？其次, 现实中积累数据的方式除增长测验之外, 还可以将使用同一题库的两批不同的答题者数据合并。那么使用合并后的数据, 推荐选题策略是否仍有良好的测量精度和优秀的题目曝光率控制？此时答题者数量与题库中题目数之比增加, 相当于推荐系统中用户—项目评分矩阵的形状发生显著改变, 而研究 1 中增长测验是增加每名答题者回答的题目数, 相当于仅改变了用户—项目评分矩阵的数据稀疏程度, 而不改变矩阵的形状。为了控制这一变量, 在研究 2 中仅采用 20 道题的终止规则。

研究 2 使用 TIMSS 2015 八年级科学测验的 276 道题目, 其中 125 道题基于 2PLM, 其余 151 道题基于 3PLM, 该题库中 a 参数的分布大多集中于 1 附近, 高区分度的题目所占比例较小, b 参数的分布范围小于模拟题库, 尤其 b 参数小于 0 的低难度题目不多, 3PLM 下题目的 c 参数整体较大, 可见该题

表 1　模拟题库下各选题策略的表现

选题策略	均方误差	平均绝对误差	能力估计相关	卡方值	测验重叠率	曝光不足	曝光过度	答题者调用率
定长 20 道题目								
随机选题	0.323	0.449	0.829	2.595	5.56%	0	0	
FMI	0.090	0.234	0.954	127.852	40.80%	315	41	
DEBR (FMI)	0.141	0.291	0.930	66.341	21.83%	22	29	14.12%
IEBR (FMI)	0.242	0.383	0.872	8.712	7.09%	1	2	2.53%
BAS	0.224	0.370	0.882	14.164	9.00%	46	6	
DEBR (BAS)	0.217	0.365	0.884	11.246	8.25%	44	4	4.25%
IEBR (BAS)	0.222	0.369	0.882	11.187	8.15%	42	4	4.66%
定长 40 道题目								
随机选题	0.198	0.354	0.890	4.572	11.05%	0	0	
FMI	0.052	0.178	0.974	118.335	45.72%	240	80	
DEBR (FMI)	0.089	0.228	0.956	95.045	34.38%	37	78	19.77%
IEBR (FMI)	0.126	0.277	0.937	7.571	11.80%	0	15	5.19%
BAS	0.126	0.278	0.932	18.962	15.03%	14	36	
DEBR (BAS)	0.125	0.276	0.933	15.930	14.27%	13	27	6.98%
IEBR (BAS)	0.128	0.280	0.931	12.012	13.25%	14	17	7.22%

注：括号内为生成已有答题者数据的选题策略, 下同。

库质量低于研究 1 使用的模拟题库。研究 2 的模拟流程为：首先使用传统选题策略对第一批的 1000 名答题者进行 CAT 模拟，生成第一批已有答题者数据；然后使用推荐选题策略结合第一批已有答题者数据，对第二批的 1000 名能力分布相同的答题者进行 CAT 模拟(至此与研究 1 流程相同)；最后将两批共 2000 名答题者的数据合并作为已有答题者数据，再次使用推荐选题策略对第三批的 1000 名能力分布相同的答题者进行 CAT 模拟。使用 BAS 策略时，题库分为 4 层，每层含 69 道题，每名答题者在每层作答 5 道题后进入下一层。研究 2 中生成第一批数据的传统选题策略，答题者的真实能力分布，能力估计方法，重复次数和评价指标都与研究 1 相同。

4.2 研究结果

与研究 1 中 20 道题目下的结果相比，更换题库后 FMI 和 BAS 生成数据的特点不变但数据质量变差(见表 2)。使用 FMI 生成的第一批已有答题者数据时，两种推荐选题策略表现出与研究 1 相同的特点，在大幅改善题目曝光失衡的同时，DEBR 更注重保持精度，IEBR 使用题库更为均匀，两种推荐策略调用的答题者数量比使用模拟试题库时都提升近一倍，DEBR 仍远高于 IEBR。将 FMI 与推荐选题策略生成的两批已有答题者数据合并，对第三批答题者选题时，两种推荐策略对题目曝光率控制的改善愈加明显，DEBR 的精度始终高于 IEBR 和 BAS，IEBR 对题目曝光率的控制达到最理想的水平，DEBR 和 IEBR 找到的相似答题者数量都与合并前基本不变，由于合并数据使已有答题者数量翻倍，答题者调用率相应减半，与研究 1 中 20 道题目下的结果相近。

使用 BAS 生成的第一批已有答题者数据时，两种推荐策略的选题结果相近，DEBR 在精度指标上稍有提升，IEBR 进一步降低了卡方值和测验重叠率，调用的答题者数量基本一致，低于 FMI 下的水平。两批数据合并后，DEBR 也改善了测验安全和题库使用，IEBR 的改善更加明显，测量精度的波动始终处于合理范围。值得注意的是，合并后 DEBR 找到的相似答题者数量翻倍，使得调用率基本不变，IEBR 与合并前调用的答题者数量相同，调用率则相应缩小一半。

考察真实题库中所有题目的曝光率在两轮迭代内的变化，可以更加明显地发现这一变化与答题者调用率的变化具有一致性。FMI 生成首批数据时(见图 1，红色横线表示完全均匀的理想曝光率 $r_{ideal} = \frac{L}{K} = 0.072$)，DEBR 发生精度曝光率权衡的幅度较小(图 1b)，第一轮选题结果更接近 FMI (图 1a)，更容易在已有作答数据中找到相似答题者，因而答题者调用率的值更高；而 IEBR 会选出更多不常用的题目，改善曝光的幅度较大(图 1d)，也使得选题时相似答题者数量大幅减少，答题者调用率的值较低。合并数据进行第二轮选题时，两种策略都在原有基础上改善曝光控制(图 1c 和 1e)，调用率以同等幅度降低，数值的大小与优化曝光率的最终结果相互匹配。同理可以解释 BAS 生成首批数据时的情况(见图 2)，第一轮选题两种推荐策略的权衡趋势和答题者调用率都十分接近(图 2b 和 2d)，由于 BAS 有一定的曝光控制能力(图 2a)，DEBR 和 IEBR 的调用率都位于 FMI 条件下两种推荐策略的中间水平。第二轮选题 DEBR 基本达到了其优化曝光的上限(图 2c)，调用率变化甚微，IEBR 仍在明显改善题库使用的均匀程度(图 2e)，调用率再度降低。由

表 2 模拟真实情境下各选题策略的表现

选题策略	均方误差	平均绝对误差	能力估计相关	卡方值	测验重叠率	曝光不足	曝光过度	答题者调用率
随机选题	0.320	0.440	0.830	2.551	8.02%	0	0	
FMI	0.152	0.307	0.922	150.511	58.48%	214	33	
DEBR (FMI)	0.190	0.341	0.901	101.793	40.81%	53	38	25.04%
DEBR (FMI+DEBR)	0.233	0.380	0.875	47.426	21.10%	29	35	12.69%
IEBR (FMI)	0.265	0.408	0.855	43.395	19.63%	0	24	5.24%
IEBR (FMI+IEBR)	0.274	0.414	0.852	11.830	8.19%	0	0	2.86%
BAS	0.259	0.404	0.861	42.965	19.48%	20	27	
DEBR (BAS)	0.253	0.395	0.869	43.449	19.65%	12	33	9.75%
DEBR (BAS+DEBR)	0.262	0.403	0.865	39.684	18.29%	13	26	9.51%
IEBR (BAS)	0.266	0.408	0.858	37.491	17.49%	17	24	9.96%
IEBR (BAS+IEBR)	0.267	0.407	0.855	25.305	13.07%	8	18	5.13%

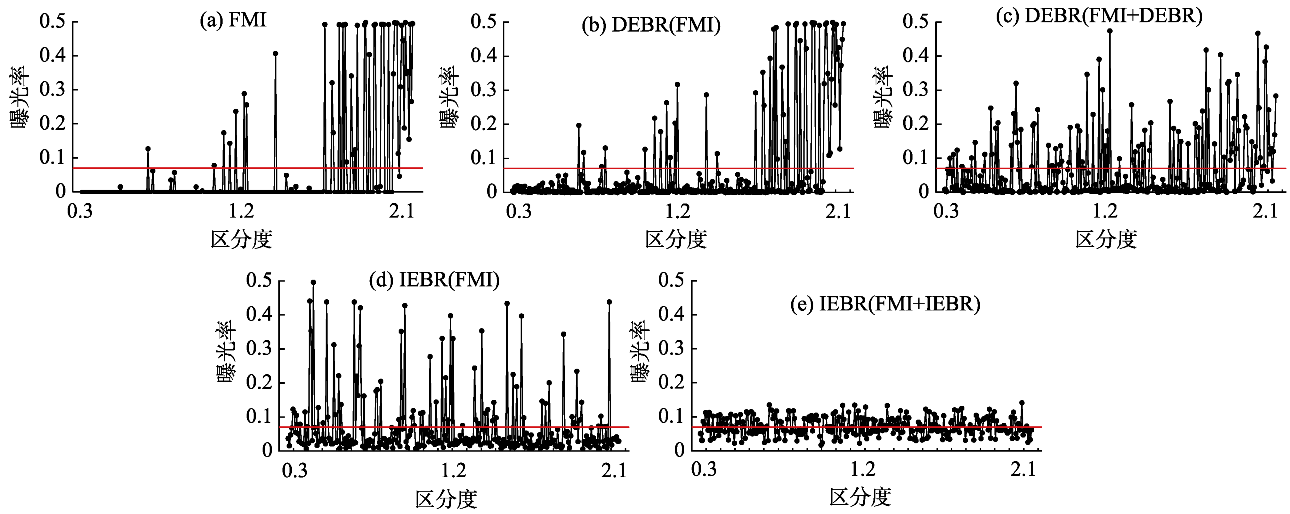

图 1 FMI 生成首批数据时两轮推荐选题的题目曝光率变化

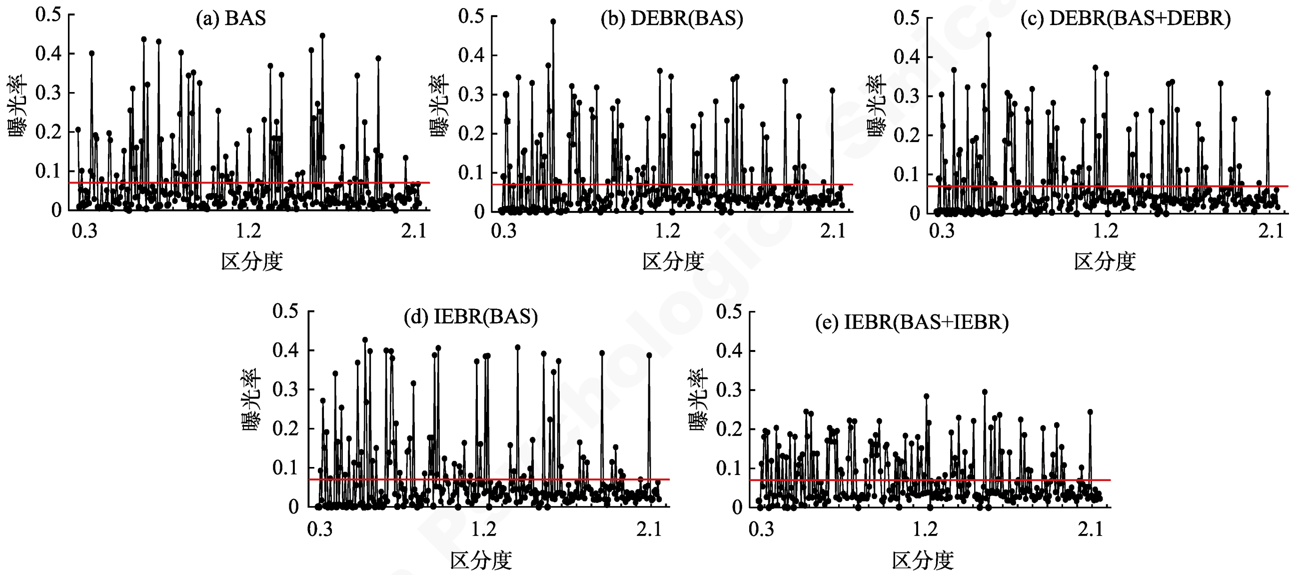

图 2 BAS 生成首批数据时两轮推荐选题的题目曝光率变化

此可见，答题者调用率可以视作推荐策略选题特点和权衡趋势的侧面衡量指标。

根据研究 2 的结果可以发现，换用质量不理想的真实题库不影响两种推荐选题策略的选题特点和良好的性质。合并传统策略和推荐选题策略自身生成的两批数据后，仅增加数据量而不改变数据的特点, DEBR 和 IEBR 的优势表现得更加明显，特点更加鲜明。

5 讨论

本研究提出了全新的基于协同过滤推荐的 CAT 选题策略，通过两个模拟研究发现：利用已有答题者数据的推荐选题策略可以保证良好的测验安全和均匀的题库使用，并有不低于分层方法的选题精度。在具体的 CAT 场景下，如果该数据中题库使用失衡，推荐选题策略会首先启用整个题库，达到该条件下选题精度和曝光率控制较好的平衡点；当已有答题者数据不存在极端的题库曝光不均时，推荐选题策略会进一步优化曝光率控制，同时不再以损失精度为代价。具体到两种新提出的策略，直接基于答题者推荐(DEBR)策略更加注重保持精度，间接基于答题者推荐(IEBR)策略改善曝光率控制的能力更强。两个模拟研究的结果都表明：由不同传统选题策略决定的已有答题者数据的特点最为主要地影响推荐选题策略表现出的选题趋势，题库质量、测验长度和答题者数量不影响该趋势，而是通过影响已有答题者数据的质量，最终一起决定了推荐选题策略在各指标上的具体大小。

本研究有两个最重要的创新之处。第一个创新在于发现了已有答题者数据作为一种先验信息对

选题的重要价值。本研究将当前答题者的数据与大量过往答题者的数据之间搭起桥梁，再次扩展了CAT 选题可参考信息的来源和数量。模拟研究的结果表明在已有的选题数据数量充足且准确可靠的情况下，借鉴他人的选题经验可以选出符合当前答题者能力的题目，同时改善过往选题数据中使用题目不均匀的问题。相比于当前答题者可以产生的数据，已有答题者数据无疑更加丰富，可挖掘的潜力巨大。本研究的另一个创新在于发现了推荐系统和CAT 选题的共通性，借鉴协同过滤推荐的技术建立了一套选题规则，并初步证明协同过滤推荐的底层假设同样适用于 CAT 的选题场景。基于该假设可以有机结合推荐系统技术和传统选题方法，设计出灵活的推荐选题策略。例如 DEBR 和 IEBR 在均匀使用题库方面有优秀的表现，得益于在基于用户推荐中加入了多种控制题目曝光率的选题操作，可见推荐选题策略是一个可不断改进的框架，未来还有使适应性测验更加精准和智能的提升空间。随着研究不断深入，尤其是推荐系统的更多引入，可能会在生成首批数据或预防选题失败等方面逐渐摆脱对传统选题策略的依赖，使推荐选题策略更少受到诸如 IRT 的底层假设不满足所产生的影响。本研究的探索也启发更多心理和教育领域的研究者，可以尝试把以推荐系统为代表的大数据技术和机器学习算法作为传统方法的结合和替换的选项。

在两个模拟研究中，推荐选题策略对各能力层次答题者的估计精度仍依赖于已有答题者数据的质量，与生成该数据的选题策略表现基本一致，对于能力居中的答题者估计精度较高，对于位于分布两端的被试估计精度较低，但不会低于已有数据中的精度水平。另一方面，推荐选题策略实际发生选题失败的概率都非常小。以 40 道题目的条件为例，对一批 1000 名答题者共需选出 40000 道题目，DEBR 发生选题失败的概率平均为 1.16% (462 道)，IEBR 发生选题失败率平均为 2.03% (812 道)，平均一名答题者发生选题失败的次数不到 1 次，这使得采用何种方法解决选题失败对测量精度和题目曝光率的影响非常微弱。在选题失败的情况中，出现找不到相似答题者的概率更低，且主要发生在使用FMI 生成的曝光不均匀的第一批答题者数据时，在其它各条件下发生的概率则小于万分之一。由此可见，仅需要一种曝光率控制较好的传统选题策略，模拟生成几千名答题者的已有答题数据，同时作为选题失败的备用策略，便可将放心地使用推荐选题策略为后续答题者选题，而后续答题者的数据还可以继续作为已有答题者数据供推荐选题策略使用，通过这种数据的迭代和积累，不断增加可参考信息的多样性，同时降低选题失败的概率。

本研究作为一种新方法的尝试和探索，尚有许多值得进一步探讨和研究的问题。第一，本研究对最可能影响推荐系统表现的已有数据质量、特点和数据量进行了探讨，但没有对自适应测试中可能影响选题策略表现的答题者能力分布特点和题库特征进行深入分析。未来可继续考察已有答题者和新答题者能力分布存在差异，题库题量和题目参数分布特点，答题者作答的模式和准确性等因素对推荐选题策略的精度和选题失败率的影响。第二，随着已有答题者数据量增大，两种推荐选题策略的测量精度反而降低，这可能是由于本研究设计推荐选题策略时十分注重解决题目曝光不均匀的问题，除相似答题者的设计之外，没有进一步提高选题精度的具体操作，限制了新策略在面对更大的数据时保持高精度，未来可针对此局限进一步改进选题策略。第三，本研究提出的推荐选题策略仅适用于单维和0-1 计分的 CAT，现实中还有大量多级评分的题目，且基于使用的 IRT 模型不同，还有更复杂的多维CAT 和认知诊断 CAT，如何在这些复杂模型中快速且高效地选题是如今研究的热点和难点(Akbay, & Kaplan, 2017; Kaplan, de la Torre, & Barrada, 2015; Zhang, & Chang, 2016; 毛秀珍, 辛涛, 2015)，因此，针对多级评分题目和基于复杂模型的 CAT 改进推荐选题策略也是一个重要的研究方向。

结合本研究的结果和针对上述值得探讨的问题提出几种改进推荐选题策略的思路：第一，继续结合传统选题策略。以 IEBR 为例，找到相似答题者后可将匹配 b 参数替换为精度更高的选题方法。第二，修改相似答题者的定义，例如考虑当前题目之前若干题目的作答结果，或是借用推荐系统中多种相似度计算公式，找到更精准的相似答题者，提高选题精度。第三，协同过滤推荐还有基于项目推荐的方式，即计算适用于 CAT 场景的题目相似度，选出与已作答题目最相似的未作答题目，这种基于题目推荐的选题策略可以更好地避免选题失败，也更易于选出新加入题库尚未使用过的题目。第四，当新用户加入，因数据稀缺对用户了解不足时，协同过滤推荐往往会难于做出推荐，这一问题被称为冷启动(cold start)，随着技术发展产生了一系列解决冷启动问题的方法(Lika, Kolomvatsos, &

Hadjiefthymiades, 2014)，可借鉴这些方法解决测验前期测量不准确和选题失败的问题。第五，除了协同过滤推荐，推荐系统中还有许多新技术可用于改进 CAT 选题策略。例如基于模型的推荐，使用机器学习的算法对用户评分数据构建复杂模型完成推荐，可用的算法十分多样(Ricci et al., 2015)，可以提高协同过滤推荐的预测力和灵活性，也为推荐系统迁移至 CAT 场景提供了更多选择。近几年，深度学习发展正热，与推荐系统相结合催生出深度推荐算法，得以解决日益增长的海量数据和愈加复杂的推荐问题(Covington, Adams, & Sargin, 2016; Wang, Wang, & Yeung, 2015)，这对于规模庞大且题目类型复杂的 CAT 选题同样有借鉴意义。

6　结论

本研究发现：(1)推荐系统中的协同过滤推荐可移植于 CAT 选题，设计出的推荐选题策略在保证一定测量精度的同时，对题目曝光率的控制更好；(2)已有答题者数据是一类对选题有价值的先验信息，该数据的特点和质量是影响推荐选题策略表现的主要因素。

参　考　文　献

Akbay, L., & Kaplan, M. (2017). Transition to multidimensional and cognitive diagnosis adaptive testing: An overview of cat. *The Online Journal of New Horizons in Education-January*, *7*(1), 206–214.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, *34*(6), 438–452.

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20.

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. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

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, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation system for adaptive learning. *Applied psychological measurement*, *42*(1), 24–41.

Cheng, Y., Patton, J. M., & Shao, C. (2015). a-stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement*, *75*(2), 260–283.

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for Youtube recommendations. *In Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198). Boston, MA: ACM.

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), 1–39.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, *74*(4), 677–696.

Jia, Z., Yang, Y., Gao, W., & Chen, X. (2015, February). User-based collaborative filtering for tourist attraction recommendations. *In 2015 IEEE International Conference on Computational Intelligence & Communication Technology* (pp. 22–25). Ghaziabad, India: IEEE.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied psychological measurement*, *39*(3), 167–188.

Klašnja-Milićević, A., Ivanović, M., & Nanopoulos, A. (2015). Recommender systems in e-learning environments: A survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, *44*(4), 571–604.

Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 77–118). Boston, MA: Springer.

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

Liu, Q., Chen, E. H., Zhu, T. Y., Huang, Z. Y., Wu, R. Z., Su, Y., & Hu, G. P. (2018). Research on educational data mining for online intelligent learning. *Pattern Recognition and Artificial Intelligence*, *31*(1), 77–90.

[刘淇, 陈恩红, 朱天宇, 黄振亚, 吴润泽, 苏喻, 胡国平. (2018). 面向在线智慧学习的教育数据挖掘技术研究. *模式识别与人工智能*, *31*(1), 77–90.]

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Mao, X. Z., & Xin, T. (2011). Item selection method in computerized adaptive testing. *Advances in Psychological Science*, *19*(10), 1552–1562.

[毛秀珍, 辛涛. (2011). 计算机化自适应测验选题策略述评. *心理科学进展*, *19*(10), 1552–1562.]

Mao, X. Z., & Xin, T. (2015). Multidimensional computerized adaptive testing: Model, techniques and methods. *Advances in Psychological Science*, *23*(5), 907–918.

[毛秀珍, 辛涛. (2015). 多维计算机化自适应测验：模型, 技术和方法. *心理科学进展*, *23*(5), 907–918.]

Pirasteh, P., Jung, J. J., & Hwang, D. (2014, April). Item-based collaborative filtering with attribute correlation: A case study on movie recommendation. In N. T. Nguyen, B. Attachoo, B. Trawiński, & K. Somboonviwat (Eds.), *In Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems* (pp. 245–252). Cham, Switzerland: Springer.

Quijano-Sánchez, L., Recio-García, J. A., Díaz-Agudo, B., & Jiménez-Díaz, G. (2011, March). Happy movie: A group recommender application in facebook. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference* (pp. 419–420). Palm Beach, FL: AAAI.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 1–34). Boston, MA: Springer.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon. com. *IEEE Internet Computing*, *21*(3),

12–18.
Tan, P. N., Steinbach, M., & Kumar., V. (2006). *Introduction to Data Mining*. New York, NY: Pearson Education.
Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science*, *1*(2), 2811–2819.
Wang, H., Wang, N., & Yeung, D. Y. (2015, August). Collaborative deep learning for recommender systems. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1235–1244). Sydney, NSW, Australia: ACM.
Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492.
Zhang, S., & Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, *1*(1), 67–92.
Zhu, T. Y., Huang, Z. Y., Chen, E. H., Liu, Q., Wu, R. Z., Wu, L., … Hu, G. P. (2017). Cognitive diagnosis based personalized question recommendation. *Chinese Journal of Computers*, *40*(1), 176–191.
[朱天宇，黄振亚，陈恩红，刘淇，吴润泽，吴乐，... 胡国平. (2017). 基于认知诊断的个性化试题推荐方法. *计算机学报*, *40*(1), 176–191.]

Make adaptive testing know examinees better: The item selection strategies based on recommender systems

WANG Pujue[1]; LIU Hongyun[1,2]

([1] *Faculty of Psychology, Beijing Normal University;* [2] *Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing,* 100875, *China*)

Abstract

Better CAT item selection strategies may be designed by making better use of information from previous examinees' responses. The past examinees' data serve as a valuable reference for selecting items more accurately and evenly for new examinees. However, most of the existing strategies proposed under the theoretical framework of IRT only use information from the current examinee and fail to take full advantage of past examinees' data. A collaborative filtering recommender approach from the recommender system literature is able to find items that best match one's preference by utilizing information from others, which shares the similar goal as the item selection strategy of CAT. Therefore, the present study adapted the underlying assumptions of collaborative filtering recommender and proposed new item selection strategies which take advantage of past examinees' data, and then investigated the potential factors that might affect the performance of new strategies.

In light of user-based collaborative filtering, we defined similar examinees as a group of examinees who uniformly answered the same items, and proposed two strategies, Direct Examinee-Based Recommender (DEBR) and Indirect Examinee-Based Recommender (IEBR). Two simulation studies were conducted to examine the measurement accuracy and item exposure control of new strategies under different conditions. In study 1, a simulated item bank was considered. The recommender-based strategies used two different types of past examinees' data generated by FMI and BAS, respectively, to select items under two fixed-length CATs. In study 2, a real item bank was used to test new strategies under a more realistic setting. The effect of combining two batches of past examinees' data from different recommender-based strategies was also investigated.

In both studies, when using past examinees' data with high accuracy but poor item exposure control (generated by FMI), the recommender-based strategies greatly remedied unbalanced item utilization with an acceptable loss of accuracy. When using past examinees' data with better tradeoff of measurement precision and test security (generated by BAS), the recommender-based strategies kept the accuracy at the same level and further improved item exposure control. More specifically, DEBR focused on maintaining the accuracy and had lower measurement error than IEBR; IEBR was good at improving the control of item exposure and made better use of the whole item bank than all the other strategies. These features of two recommender-based strategies were stable and consistent under different item banks and different length of CATs. The extent to which DEBR and IEBR demonstrated their features was influenced by the quality of item bank, test length, number of past examinees and strategy used to generate data.

In general, this research successfully combined the recommender systems with CAT item selection methods to establish a new flexible framework, which is an unprecedented innovation upon the traditional item selection strategies. This research also provided empirical evidence for the value of past examinees' data and the recommender system approach as a feasible alternative option for selecting items in CAT. Finally, suggestions for future studies were provided regarding investigating the proposed new strategies in various situations and upgrading recommender-based strategies for more CAT conditions, including finding diverse measures of similarities between examinees or items and employing more complex algorithms of recommender system to meet the demands of large-scale tests.

Key words selection strategy; past examinees' data; recommender system; collaborative filtering recommender; simulation study