文章编号:1000-5862(2017)05-0462-08

计算机多阶段自适应测验的组卷方法

李贵玉,涂冬波*,戴步云,宗一涛,高旭亮,苗 莹
(江西师范大学心理学院,江西 南昌 330022)

摘要: 计算机多阶段自适应测验(MST)实施的关键是成功组建多个满足测验规范(即统计和非统计约束)的平行测验(或称测验面板),自动组卷(ATA)为实现测验平行提供了可能. 现有的 MST 组卷方法研究主要包括以下几种:1)基于线性规范算法的组卷方法;2)基于启发式算法的组卷方法;3)基于蒙特卡洛算法的组卷方法;4)基于在线组卷的方法. 该文讨论这几种方法的优缺点并进行比较,同时指出未来可进一步改进这种方法并开发基于认知诊断测验的自动组卷方法.

关键词: 计算机多阶段自适应测验;自动组卷;组卷方法;测验规范
中图分类号: B 841 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2017.05.04

0 引言

20 世纪 70 年代,项目反应理论的诞生为计算机自适应测验(Computerized Adaptive Testing,CAT)的发展提供了可能. 计算机自适应测验已成为国内外心理测量领域的研究热点,并在教育及心理测量领域得到广泛的应用. 随着测量理论及实际应用的深入,CAT 因自身特性开始暴露出一些问题:如曝光控制不足、不可修改答案与真实测验情景不符、难以满足内容平衡等. 为了弥补 CAT 的不足,计算机多阶段自适应测验(Multistage Adaptive Test,MST)应运而生. 计算机多阶段自适应测验以项目集(即模块,module)来管理题目,在阶段(stage)水平上自适应,允许被试在阶段内修改题目以减轻被试测验焦虑并提高测量精度;同时 MST 的组卷方法可以实现测验间的平行. 相比较 CAT 而言,MST 有许多固有优势:1)CAT 要求题目必须满足局部独立性假设[1],而 MST 对于共用题干的题目(即题组,testlet)按照多级计分来分析,不被 CAT 的题目假设所限制;2)传统的 CAT 不允许考生修改答案,导致考生对测验掌控感不足,进而产生考试焦虑、被试偶尔失误不能得到改正,从而不能正确测量被试真实能力水平影响测验精度,MST 基于阶段的自适应可以允许被试在阶段内修改题目,一定程度上缓解被试焦虑同时,避免失误带来的测量误差,提高测验精度,也更符合传统作答情景;3)CAT 组卷时追求信息量最大易造成题库内高信息量的题目过度曝光,而 MST 通过构建平行面板提升题库使用率从而控制曝光;4)CAT 忽视了测验的非统计特性,追求单纯的信息量,而 MST 可以提前对测验的非统计特性,如内容平衡、题型平衡、敌对题目(enemy item,即题目间存在题目答案相互提示的线索)、字数等方面进行管理,更大限度地提升测验精度与测验之间的平行程度;5)相较于 CAT 施测过程中在线组卷无法对题目进行管理,MST 预先组卷可以帮助测验开发者更好地管理测验;6)在 MST 中被试作答题目数相等,同样能提供良好的测验精度[2]. 计算机多阶段自适应测验(Multistage Adaptive Test,MST)进入研究者的视野并在近年来越趋受到关注. MST 作为纸笔测验(Paper And Pencil Based Test,P&P)和 CAT 的"折中"在 2011 年被应用于美国研究生入学考试中(GRE),2004 年被应用于美国注册会计师考试(CPA),同时还应用于美国国家教育进展评估(NAEP)、美国 K-12 评估、国际调查评估(PAICC)、美国医学执照考试(USMLE)等大型考试中,并取得了良好的效果.

收稿日期: 2017-05-22
基金项目: 国家自然科学基金(31660278,31760288),教育部人文社科项目(11YJC190002),高等院校博士点基金(20123604120001),江西省社会科学规划重点项目(13JY01),江西省教育科学规划项目(12YB088,13YB029)和江西师范大学青年英才培育计划资助项目.
通信作者: 涂冬波(1978-),男,江西南昌人,教授,博士生导师,主要从事心理统计与测量的研究. E-mail: tudongbo@aligun.com

事实上，MST 并不是近年提出的测验方式. 早在 1990 年 J. J. Adema [3] 就提出 2 阶段测验理论，同年 C. Lewis 等 [4] 提出计算机优势测验（computerized mastery testing，CMT），在此基础上 R. M. Luecht 等 [5] 提出计算机自适应顺序测验（computer-adaptive sequential testing）；R. M. Luecht 等 [6-7] 提出捆绑多阶段自适应测验（bundled multistage adaptive test）及 R. D. Armstrong 等 [8] 提出的多形式测验（multiple form structures），这些测验形式虽名称各异但与 MST 在实质上相同. 国际上关于 MST 的理论、技术及应用深受研究者的欢迎. 近年来计算机多阶段自适应测验逐渐在文章中被广泛使用 [7,11,13-15,18-20,23-27].

MST 建立在若干个平行的面板（panel）基础上，一个完整的面板包含以下几个元素：模块（module）、阶段（stage）、路径（pathway），具体详见图 1. 一定数量的项目组成项目集，即模块（module）. 不同难度水平的模块混合组成阶段（stage），若干个自适应阶段组成面板，MST 的题库由若干平行的面板组成. 在图 1 中，1M 表示第 1 阶段内题目难度中等（Moderate）；2E、2M、2H 分别表示第 2 阶段内容题目难度容易（Easy）、中等和较难（Hard）；3E、3M、3H 依此类推. Panel#1、Panel#2、Panel#3 表示相互平行的测验面板. 测验开始后将被试随机分配到预先组好某个测试面板，随后根据考生在第 1 阶段的作答将其自适应到下一阶段中的与其当前能力值相匹配的模块；考生作答的一系列模块组成路径. 图 1 中每个 Panel 有 7 种测试路径，详见图 1 中的箭头走向. 其中，实线箭头为主要路径共 3 种，即考生最有可能被自适应的路径，虚线为次要路径共 4 种.

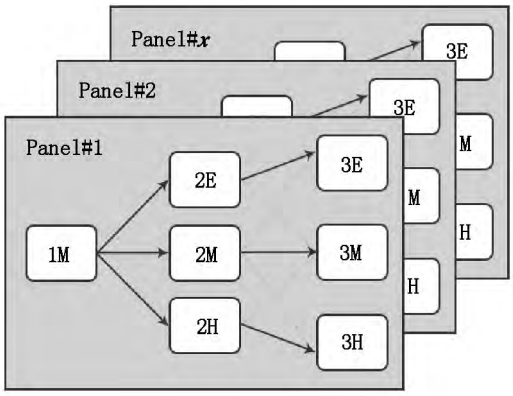

图 1　三阶段自适应测验的多个平行面板

平行测试面板的构建是 MST 实施的前提，也是 MST 的核心部分. 构建 MST 面板需满足测验规范的要求，即同时兼顾测验统计学特征（如测验信息量等）和非统计学特征（如测验内容平衡等），以保证每个测试面板具有较好的信效度. 在 MST 中，统计与非统计特征主要涉及目标测验信息量（target test information function，TTIFs）、测验长度、阶段与模块的数量、内容平衡和曝光控制等. 然而，在构建面板时这些因素并不是相互独立的，而是紧密结合在 MST 的架构之中 [28]. 同线性测验（linear test）一样，为了确保测验的安全性、题目的使用率等，MST 的研究者希望可以组建多个平行的面板 [25]. 在线性测验中，题目被预先组建成固定的测验形式，当测验的信息量及其他测量条件足够相似时则可以认为这些预先组建的测验形式是平行的. 在 MST 中，路径等同于线性测验的固定形式，然而在 MST 中由于模块的难度水平不同，面板内的路径常常是不平行的. 但是当 2 个面板直接相对应的路径相互平行时，则可以认为这 2 个面板相互平行. 需要注意的是，组建 MST 的平行面板时，出于测验规范而组建了平行路径时，则不必要求模块间的平行 [27]. 测验规范不同所需组建的平行面板也不同，因此满足测验规范的面板必须由良好的组卷方法来保证. 若是组卷方法不能满足测验规范的要求，即不能构建平行的面板，则 MST 无法实施. 因此，组卷是否成功是 MST 实施的关键.

尽管 MST 结合了 CAT 与传统测验的优势，既实现了依被试能力施测又相对符合传统测验的作答习惯，但是由于测验规范的增加及组建大量平行的面板的要求，MST 的组卷仍面临较大的难度. 良好的 MST 组卷应同时满足以下 3 个目标：（i）各阶段内模块有明确的信息曲线足以区分测验的不同路径；（ii）面板间相对应的路径间的信息曲线足够相似从而确保面板间的平行；（iii）每个面板的每条路径满足非统计约束 [28].

组卷的成败关系到 MST 的具体实施. 目前国内外研究者已对 MST 的组卷做了大量的研究并已取得较好的结果，主要集中在对 MST 组卷方法的研究. 早期组卷是基于经典测量理论的题目难度与区分度随机组卷 [29]. 随着项目反应理论的出现，基于信息函数（如 Fisher 信息量）的自动组卷方法开始被研究者采用 [23]. F. M. Lord [30] 提出将启发式算法（Heuristic methods）用于 MST 的自动组卷，L. Swanson 等 [31] 在启发式算法的基础上加入非统计约束提出加权离差模型（the weighted deviation mode）用于 MST 的组卷，但该组卷方法并未被应用于实际的 MST 组卷中. R. M. Luchent [5] 将所有测验规范按一定的权重进行标准化加权，提出标准化加权离差启发算法（normalized weighted absolute deviation heuristic），受到目前研究者的广泛应用 [28]. 除此之外，

T. J. J. M. Theunissen[32]将线性规划算法用于 MST 的自动组卷,并被研究者广泛采用. 现在 0-1 线性规范是研究者在 MST 自动组卷中使用较多的线性规范算法[33]. D. I. Belov 等[12]提出将蒙特卡洛算法用于 MST 的自动组卷; Zheng Ying[27]等提出一种在线(on-the-fly)组卷的 MST 组卷方法,这种方法不同以往的自动组卷算法.

总之,组卷是 MST 的重中之重,是 MST 成功组建多个平行面板的关键. 因此,文章中对 MST 组卷的基本思路进行了阐述,重点介绍了当前国内外 MST 组卷的主要方法并比较了各类组卷方法的优劣; 最后,对 MST 的组卷方法做了总结,并在分析当前国内外研究的基础上进一步阐述了未来发展趋势和研究方向.

1 MST 的组卷

MST 的组卷包含面板的数量、面板内阶段的数量、阶段内模块的数量、模块所包含的题目数、题目难度水平等因素. MST 组卷涉及组卷策略和组卷方法两部分: 组卷策略决定题目难度水平、内容平衡等因素在模块还是在路径水平平行上,统计约束与非统计约束必须包含在模块或路径内; 而组卷方法是决定这些因素能否平行的关键. MST 的组卷方法是为了确保统计和非统计约束在面板内或面板间平行. 测验的统计约束最初由经典测量理论的难度与区分度参数来确定[29],随着项目反应理论的发展,目标测验信息函数(testinformation function,TIF)逐渐成为统计约束的主要表现形式. 目标测验信息函数通常采用 Fisher 信息量,R. M. Luecht[23]对目标 TIF 的设置做了详细的描述. 此外,统计约束的目标 TIF 的设置必须考虑题库是否支持,如题库质量一般却要求提供较好的 TIF 等,往往有限的题库质量本身正是 MST 组卷的最大限制[28].

2 MST 组卷策略

2.1 MST 的 2 种组卷策略

MST 面板的设计完成后,组建多个平行的 MST 面板通常分为 2 步进行: (i)从题库选取题目组建模块; (ii)以模块来组建面板. R. M. Luecht[6]提出 2 种策略用来组建平行的 MST 面板,分别是自上而下策略(top-down)和自下而上策略(bottom-up).

自上而下策略: 构建路径水平的平行. 从题库中构建多个平行的面板,同时面板的路径也相互平行,包括统计约束(目标 TIF)和非统计约束两者都平行. 路径的平行有 2 种,以图 1 为例,路径的第 1 种平行构建方式只需保证主要路径平行,即图 1 实线箭头所指示的路径 1M-2M-3M、1M-2H-3H、1M-2E-3E. 因为这 3 条主要路径基本代表了大多数被试可能的作答路径,所以面板内部只需要保证这 3 条路径平行即可. 其他路径可以根据模块难度组装,必要时也可以加相应限制. 路径的第 2 种平行构建方式是保证所有可能的路径平行. 采用自上而下策略组建平行的 MST 面板时,要为整个测验及路径设置目标 TIF,并将非统计约束分配在路径内.

自下而上的策略: 构建模块水平的平行,包括统计约束(目标 TIF)和非统计约束两者都平行,即组建平行的模块. 模块平行后,可以随意混合相互平行的模块来组建多个平行的 MST 面板. 由于模块之间相互平行,相应地由模块组成的面板的路径相互间也是平行. 采用自下而上策略组建平行的 MST 面板时要为模块设置目标 TIF,模块难度不同,目标 TIF 不同. 同时将非统计约束分配在各模块内.

van der Linden 等[34]提出了一种混合策略(mixed),既有模块水平的平行也有路径水平的平行. 即在满足统计约束时为各路径设置目标 TIF 使统计约束以自上而下的方式平行,同时将非统计约束分配至各个阶段(也可以分配至模块即自下而上)保持平行.

2.2 MST 组卷策略的评价

自上而下策略较适用于短测验,因为测验较短时难以对每个模块平均分配统计和非统计约束,因此只能将其平均分配在路径内,组建路径水平平行的面板. 当测验规范所要求的约束条件难以在阶段或模块水平平均分配时,也可采用自上而下策略. 在这种情况下,统计约束与非统计约束将不均匀地分配到不同的模块以生成初始模块集合,并且每个模块的替代形式允许在测量特性上不同,只要最终路径是平行的并且满足必要的约束. 同时自上而下策略能更好地控制测验面板属性,如防止存在跨模块的敌对项目等. 当题库及统计与非统计约束较易满足测验规范要求时,自下而上策略更加简单快速. 当测验蓝图较复杂,统计约束与非统计约束难以在同一水平实现平行时,可考虑采用混合策略.

3 MST 的自动组卷方法

MST 的组卷建立在面板的水平上,对于大规模测验而言,同时组建多个平行面板是优先的选

择[36]. 因此组建 MST 的平行面板时,可以将其视为多目标函数及约束的优化问题. 组装多个平行的 MST 面板实际是一个系统的过程,手动快速组建这些面板相对困难,因此需要使用自动测验组卷(Automated Test Assemble,ATA)算法来帮助完成面板的组建. 自动组卷以优化算法或启发法从满足特定目标的题库中选择题目,由所选题目组成模块,模块组成面板. 受制于解决方案的各种约束,大多数 ATA 目标可以由最小化或最大化的目标函数组成的数学优化模型表述[5,35]. 如目标是选择用于模块的题目,其总体上满足目标 IRT 测验信息函数,受制于测试级别和模块级别的要求、各种题目内容的频率限制,其他分类特征、字数限制和其他定量变量也可以并入约束或目标函数中. 一旦开发了数学优化模型(即约束的目标函数和非统计约束被正式指定),就可以使用计算机程序来组建面板. 大多数可用的自动组卷计算机程序使用线性规划算法、网络流算法或启发式算法等从题库中选择本地或全域满足所述目标的题目. 同时有商业可用的计算机软件包用于自动组卷[36]. 目前国内外普遍使用的 MST 测验自动组卷算法有线性规划算法(linear programming methods)和启发式算法(heuristic methods)[26].

3.1　基于启发式算法的自动组卷方法

基于启发式算法的自动组卷方法,将测验分解为一系列局部优化问题,每个局部优化问题选择一个单独的题目添加到测验中[37]. 其以统计信息量为标准函数(如 TIF),同时考虑非统计约束(如内容平衡). 由于启发式算法选取题目时是顺序式的,总是选择那些测量特性最优的题目,这将导致测验早期选取的题目质量高于测验后期. 因此启发式算法是一种贪婪的算法[30].

为了平衡启发式算法的不足,更加均匀合理地利用题库,T. Ackerman[37] 提出合并策略; van der Linden[36] 提出在迭代选取项目时,一次只选择一个项目进入模块而不是一次性完成模块的组建. 所选项目的顺序可以是螺旋形,随机化或根据距离当前目标 TIF 的偏离程度确定的[28]; T. Ackerman[37]、L. Swanson 等[31] 提出允许初始组卷选择那些测量特性最优的题目,随后以“交换”步骤在模块之间交换项目,以实现较小的模块差异.

常用启发式算法的测验组卷方法主要有以下 3 种: 加权离差模型(weighted deviation model, WDM)[31]、标准化加权绝对离差启发法(normalized weighted absolute deviation heuristic,NWADH)[5] 和最大优先指标(maximum priority index,MPI)[17]. 原则上,这些方法都可以用于 MST 的自动组卷. 但实际上,只有 NWADH 被应用在实际的 MST 组卷研究中. R. M. Luecht 等[6]、L. N. Patsula[24]、R. K. Hambleton 等[18]、M. G. Jodoin 等[19] 都使用了 NWADH 法来实现 MST 的自动组卷.

3.1.1　标准化加权绝对离差启发法(NWADH)

标准化加权绝对离差启发法(normalized weighted absolute deviation heuristic,NWADH)采用约束目标的加权离差,但是其对每个约束的离差进行标准化,使得它们处于同一度量标准上[36]. 它能同时兼容多个内容或分类维度、多个定量目标、多个测验模块以及其他复杂的测验组卷问题,如敌对题目等[5]. NWADH 属于局部优化模型,它一题接一题地依照序列来组卷. 所有的统计及非统计约束联合起来设置为目标函数来满足当前的选题要求. 伴随着每一个项目的选取,目标函数依照已选题目的测量特性来进行更新,如此循环直至测验组卷完成[5]. R. M. Luecht[5] 提出的 NWADH 法是使用最多的启发式组卷算法,其 NWADH 的算法如下:

$$\left| T - \sum_{i=1}^{I} x_i u_i \right|, \tag{1}$$

其中 T 表示测验要求的总信息量; $x_i \in \{0,1\}$,表示测验是否包含项目 i,其值为 1 时表示包含,为 0 时表示不包含; $\sum_{i=1}^{I} x_i = n$,即测验长度; u_i 表示项目的统计约束,通常采用 Fisher 信息量.

在组卷时,选取(1)式最小的项目集,同时为了优化 NWADH 算法(1)式中的绝对离差函数最小化问题可以转换为最大化问题,形式如下:

$$\mathrm{MAX} \sum_{i=1}^{I} e_i x_i, \tag{2}$$

在剩余题库中,所选项目 i 的标准绝对离差为

$$e_i = 1 - d_i / \sum_{i \in R_{j-1}} d_i, i \in R_{J-1}, \tag{3}$$

其中

$$d_i = \left| \left(\frac{T - \sum_{k=1}^{I} u_k x_k}{n - j + 1} \right) - u_i \right|, i \in R_{j-1}. \tag{4}$$

在(3)式中,j 为测验所需题目数,R_{j-1} 为选取 $j-1$ 题后剩余的题库题目数. 组卷时选取能使 e_i 最大的题目进入模块. (3) ~ (4)式是考虑统计约束(信息量)时的绝对离差法,然而完整的 MST 组卷还需要考虑非统计约束如内容平衡、题型、敌对题目、字数等. 此时需要对测验要求的非统计约束条件给予一定的权重. 一般来说,权重值取决于测验要求,可以通过预先模拟获得. 此时(3)式加入非统计约

束的标准化的权重后形式如下:

$$e_i^* = \left(1 - \frac{d_i}{\sum_{i \in R_{j-1}} d_i}\right) + \frac{c_i}{\sum_{i \in R_{j-1}} c_i}, i \in R_{j-1}, \quad (5)$$

$$c_i = v_{ig} W_g + (1 + v_{ig}) \underline{W}_g, \quad (6)$$

$$\underline{W}_g = W^{[MAX]} - \frac{1}{G} \sum_{i=1}^{G} W_g,$$

(6) 式中,c_i 表示在剩余题库的 R_{j-1} 题中第 i 题的权重,v_{ig} 值为 0 时则题目 i 不包含非统计约束 g,值为 1 时包含. W_g 表示每个内容限制的权重,$\underline{W}_g$ 表示各权重的均值,$W^{[MAX]}$ 表示权重的最大值. 因此当测验包含非统计约束时用公式(5) 中的 e^* 代替公式(2) 中的 e.

3.1.2 加权离差模型(weighted deviation mode, WDM) 加权离差法是对于连续变量的约束,例如基于信息量的约束,离差是以相应约束中的数值来计算. 对于类别变量约束,例如内容平衡,根据项目特性指标计算离差,其表达式为

$$\sum_{j=1}^{J} w_j d_{Lj} + \sum_{j=1}^{J} w_j d_{Uj},$$

其中 d_{Uj} 表示组装的测验模块超过约束 j 的上限时,和约束 j 的上限之间的差; d_{Lj} 表示组装的测验模块不满足约束 j 的下限时,和约束 j 的下限之间的差; w_j 表示约束 j 的权重.

假设测验长度为 n,测验中已经有 $k-1$ 个项目,那么剩余题库中的候选题目 t 计算方式为

$$\sum_{i=1}^{I} a_{ij} x_i + (n - k)\ \nu_j + a_{tj},$$

$a_{ij} \in \{0,1\}$ 表示题目 i 是否包含约束 j,其值为 1 代表题目 i 包含约束 j,否则反之; $x_i \in \{0,1\}$ 表示题目 i 是否包含在测验内,其值为 1 代表题目 i 包含在测验内,否则反之; ν_j 表示当前剩余题库中约束 j 的平均出现次数.

3.1.3 最大优先指标(maximum priority index, MPI) MPI 法最初是用于有约束的 CAT 的选题策略,用核心标准(如 Fisher 信息量) 乘以因子,因子是由每个约束所允许的剩余题目的数量来计算. 约束矩阵为相关矩阵 $\boldsymbol{C}(J \times K)$. 项目 j 的优先指标计算方式为

$$PI_j = I_j \prod_{k=1}^{K} (w_k f_k)^{c_{jk}},$$

其中 $C_{jk} \in \{0,1\}$ 表示题目 j 是否与约束 k 相关,其值为 1 代表题目 j 与约束 k 相关,否则反之; w_k 表示约束 k 的权重; f_k 为测量约束 k 最左端的值(quota left). 在 2 阶段组卷框架中,首先选择题目以满足所有下限,即阶段 1. 此时计算方式为

$$f_k = (l_k - x_k) / l_k,$$

l 表示约束 k 的下限; x_k 表示先前选择的与约束相关的题目的数量.

当下限值达到要求时,即进入下一阶段确保优先指标不违法上限,此时计算方式为 $f_k = (u_k - x_k) / u_k$, u_k 表示约束 k 的下限.

3.1.4 基于启发式算法的自动组卷的评价 启发式算法总能在相对更短的时间内成功地组建平行的测验且计算量更小,虽然它有时在一定程度上会违反测验规范(如测验的统计和非统计约束),但是这种违反是可接受的. 启发式算法相较于其他算法受题库及测验规范的约束条件的影响较小,这也是它能成功组建平行测验的原因之一. 因此,当研究者需要组建平行测验而对一定程度上的违反测验规范的约束条件能容忍时,启发式算法是一个较好的选择.

3.2 基于线性规划算法的自动组卷

3.2.1 基于 0-1 线性规划算法的自动组卷方法 多个平行测验的组卷可以看为 0-1 线性规划(0-1 linear programming methods) 问题,即线性规划问题的决定变量严格限制为 0、1 时[34]. ATA 中的主要优化方法之一是将特定的测验模块作为高维二进制(0-1) 空间中的点. 空间的每个轴都与一个项目相对应,该点的坐标表示是否将给定项目分配给测验模块. 0-1 线性规划算法作为线性规划算法的一种,主要用于优化 2 维空间上的目标函数,即受到多个约束的常见目标函数包括测试信息函数、目标测验的信息偏差、以及多个平行测验的差异. 如目标函数优化问题可能涉及使测验信息最大化,这取决于固定测完长度、预期测验时间、内容约束和敌对题目.

0-1 线性规划算法优化组卷可以分为 2 步: 第 1 步,先计算相对容易的线性规划问题,可以不考虑决定变量的 0-1 限制,先从常规的线性规划模型入手使用相对简单的算法; 第 2 步,优化 0-1 线性规范问题,此时可以使用分流捆绑法(branch-and-bound). 分流捆绑法是一种从相对容易状态为起点的树状搜索方法,可以解决多个线性规划问题[34]. 0-1 线性规划算法的常见形式为

$$Maximize: \sum_{i=1}^{I} \sum_{k=1}^{K} I_i(\theta_k)\, x_i, \quad (7)$$

$$x_i \in \{0,1\}, i = 1, \cdots, I,$$

其中 $\theta_1, \cdots, \theta_K, k$ 表示能力量尺上取得的标准点,(7) 式受以下几个因素影响: $\sum_{i=1}^{I} x_i = n$,即总测验长度,I 为题库中总题目数; $\sum_{i=1}^{I} t_i x_i \leqslant T_u$,即总预期

测验时间; $C_r^{(l)} \leqslant \sum_{i \in V_{cr}} x_i \leqslant C_r^{(l)}$, $r = 1,\cdots,R$,即内容约束,V_{cr} 为所选题目属于内容 r; $\sum_{i \in v_e} x_i \leqslant 1, e = 1,\cdots,E$,即敌对题目; V_e 为所选题目属敌对题目.

3.2.2　基于 0-1 线性规划算法的自动组卷的评价　0-1 线性规划算法在组卷时总能成功完成组卷要求,并严格满足所有的测验组卷约束. 但是,解决0-1 线性规划问题是较为复杂的[33]. 当测验约束条件复杂度增加时,题库自身限制不能满足所有的测验约束,就会出现过度约束的优化不可行问题从而导致测验组卷失败. 因此 0-1 线性规划算法受测验约束条件和题库特性的影响较大,不适用于测验约束条件较多较复杂且题库容量有限的组卷情况. 较适用于测验约束条件较易满足的情况,这时 0-1 线性规划算法可以组建出满足所有测验约束且完全平行的测验. 同时,0-1 线性规范法计算量大,对计算机硬件要求较高,组卷需要依赖特定商业软件(如 CPLEX2 等).

3.3　基于蒙特卡洛(Monte Carlo)算法的自动组卷

3.3.1　基于蒙特卡洛算法的自动组卷方法　D. I. Belov 等[12]提出将蒙特卡洛(Monte Carlo,MC)算法用于 CAT 的组卷中取得了较好的效果,随后他们又将这一算法用于 MST 的组卷同样取得了满意的结果. 其基本思路是将 MST 中的各个模块看作各个节点(node),每个节点由一定量的题目组成. 不同的节点之间组合形成不同的路径,每个节点信息量的和即为路径信息量. 第 1 阶段只有一个节点即模块称为根结点(root node),各阶段的节点数即模块数由测验自身测验需求决定. 蒙特卡洛算法的 MST 组卷可以分为以下 3 个部分:

1) 设立 IRT 目标,即为各节点设立 TIF 值. 首先将连续变量能力分布 θ 划分为 m 个点,产生 m 个线性测验; 将给定的能力分布分成 m 个点,m 不等于 l,模拟 m 个线性测验上的 l 个被试的能力真值. 计算子测验的平均分数即 $A[x,y]$,x 属于 m 个线性测验中的一个,y 属于 MST 的 n 个阶段的一个. 构建 $T[]$,即每个节点子测验的量. 用 r 表示 MST 的节点数. 计算节点的 TIF,节点 TIF 之和构成路径的 TIF.

2) 组建 MST. 确定 TIF 后,蒙特卡洛的组卷方法类似于组建线性测验. 先前分配给路径的题目减少了下一个路径搜索区域,这种减少源于路径间公共节点的题目的共用. 当其所有路径被组装后 MST 组卷完成. 当一个路径组建失败,所有路径和部分组装 MST 的题目都返回到题库中进行再次组卷. 详细算法可参见文献[13].

3) 组建多个不重叠的 MST. D. I. Belov 等[13]提出常规的一题接一题的顺序组卷方法容易造成大量重叠题目的 MST,因此他们提出口袋法用来解决这个问题从而组建大量不重叠的 MST. 其思路是: 先顺序组建 MST,随后在大量重叠的题目里找没有重叠的题目的最大子集,并由这些题目组成题目口袋[12]. 然而由于算法的原因,导致组建的 MST 不能满足目标 TIF 值. 因此 D. I. Belov 等[13]提出可以从 3 个方面来解决这个问题: (i) 增加符合测验要求的新题; (ii) 假定倾向与题库信息函数的能力分布; (iii) 设立目标 TIF 时为模拟的模块数设一个相对较小的值[13].

3.3.2　基于蒙特卡洛算法的自动组卷评价　蒙特卡洛算法的收敛情况依赖于可行域即满足测验要求的题库数量的大小. 在组卷过程中偶尔也会在一定程度上违反测验约束尤其统计约束. 蒙特卡洛算法本质上也是一种启发式算法在可行域内寻求最优的题目. 但相较于启发式算法它受题库影响更大,易被题库与测验约束之间的关系影响且存在组卷结果达不到预期目标信息量的情况.

3.4　其他自动组卷方法

R. D. Armstrong 等[10]提出了 2 步法组建 MST; R. D. Armstrong 等[8]提出用神经网络算法组卷; Chen Peihua 等[16]提出基于随机样本和分类 2 种方法; Zheng Ying 等[27]提出在线组卷的 OMST 即“on-the-fly”MST. 上述组卷方法中只有 OMST 适用于 MST 的组卷要求. OMST 是基于模块的自适应计算机测验,由若干阶段组成,各阶段只包含一个模块. 第 1 阶段随机给被试分配若干项目,根据被试的作答在题库中自适应地为被试选择下一阶段的题目直至测验结束. 采用题目替换算法来保证测验平衡并使用 SH 法来控制题库的使用率并取得较好的结果[26].

4　总结与展望

4.1　研究总结

MST 既能像 CAT 一样自适应地依被试能力为被试选题,同时还能在测验实施前使测验开发者更好地对测验进行管理,在测验实施中 MST 在一定程度上允许被试修改答案降低被试考试焦虑更准确地测量被试能力的水平. 同时 MST 的模块可以允许共

用题干的题型,解决了 CAT 的局部独立性[1],使测验所能包含的题型更加灵活丰富. MST 相较于 P&P 更加节省人力、物力、财力,MST 无须打印派送问卷也无须对问卷进行人工计分,既避免了打印派送环节问卷泄露又避免了人工计分带来的测量误差. 同时 MST 基于平行面板的题库建设思想又最大限度地降低了被试间的测量误差,带来更精准的测验精度及被试能力值.

MST 的优点近年来得到研究者的重视并在教育与心理测量领域广泛应用. 然而,能否成功组建多个平行的测验面板是 MST 实施的关键. 因此国内外学者关于 MST 的组卷做了大量的研究和探讨,这对 MST 的应用与推广具有重要的意义. 纵观这些组卷方法,各有优势与不足,可以在一定程度上进一步地改进. 线性规划法能够完全满足测验规范但存在组卷失败的可能性,启发式算法总能组卷成功但是会在一定程度上违反测验规范. 表 1 直观、全面地简述了目前国内外主要组卷方法的特点及优缺点.

表 1 MST 组卷方法及优缺点

方法	优点	缺点
线性规划算法	完全满足测验规范,组卷的版本间完全平行,总能成功组卷	可能导致测验组卷失败、计算量大、依赖商业组卷软件、受测验规范及题库影响大
启发式算法	对题库质量依赖较小,计算量相对小	算法倾向选测验特性最优的题目导致题库使用不均衡,曝光控制不足
蒙特卡洛算法	测验组卷成功且满足测验规范	受题库质量影响较大,不满足测验统计约束
on-the-fly 法	在线组卷节约人力成本,易组卷成功,满足测验规范,曝光控制更好	测验开发者无法对测验进行管理

4.2 研究展望

国内外学者关于 MST 组卷方法的研究已经非常多,本文在已有研究基础上对未来的 MST 组卷研究提出了几点展望.

1) 目前的 MST 组卷方法大多是基于单维项目反应理论(dimensional Item Response Theory,IRT)模型,而多维项目反应理论(Multidimensional Item Response Theory,MIRT)模型将被试能力水平划分为多个维度,其模型参数更复杂,对组卷算法要求也更多. 因此,未来研究可以考虑基于 MIRT 的模型开发适用的 MST 自动组卷算法,为 MST 在 MIRT 上的测验设计、测验自动组卷等提供参考及支持.

2) 现有的组卷方法各有其优缺点,未来研究者可以着眼于开发具备现有组卷方法的优点又能弥补其不足的新的 MST 组卷方法,即成功组建多个平行的测验面板又能使其完全满足测验规范是一个值得努力的方向.

3) 0-1 线性规划算法组卷时能完全满足测验规范,最能体现 MST 构建完全平行面板的要求. 但其受题库的数量、质量及测验规范影响大,因此,未来研究可以考虑开发新的题库建设方法,使题库更合理更能满足测验规范从而成功实现自动组卷,拓展 0-1 线性规划算法的适用范围,这样便能更好地发挥 0-1 线性规范算法的优势.

4) 心理测验理论和认知心理学的发展使认知诊断理论成为研究热点[38],目前的认知诊断计算机自适应测验仍是题目水平的自适应,并面临着与基于项目反应理论的 CAT 一样的困境. 虽然高椿雷等[39]已将 MST 的在线组卷法即 on-the-fly 组卷方法应用于认知诊断并取得较好的结果. 但是,其本质上仍不具备 MST 的固有优势,如测验开发者不能在测验实施前对其进行管理、测验面板间难以完全平行、难以满足非统计约束上的测验平行等. 目前尚无研究者将 MST 的自动组卷技术与认知诊断的特性相结合开发适用于自动组卷的认知诊断多阶段自适应测验. 因此,MST 的自动组卷算法与认知诊断相结合,仍有很大的研究价值与空间.

5 参考文献

[1] 王钰彤,罗照盛,王睿. 计算机多阶段自适应测验研究评述[J]. 心理科学,2015,38(2):452-456.

[2] Kim S,Moses T,Yoo H. A comparison of IRT proficiency estimation methods under adaptive multistage testing [J]. Journal of Educational Measurement,2015,52(1):70-79.

[3] Adema J J. The construction of customized two-stage tests [J]. Journal of Educational Measurement,1990,27(3):241-253.

[4] Lewis C,Sheehan K. Using Bayesian decision theory to design a computerized mastery test [EB/OL]. [2017-01-21]. http://www.iacat.org/sites/default/files/biblio/v14n4p367.pdf.

[5] Luecht R M. Computer-assisted test assembly using optimization heuristics [J]. Applied Psychological Measurement,1998,22(3):224-236.

[6] Luecht R M,Nungester R. Some practical examples of computer-adaptive sequential testing [J]. Journal of Educational Measurement,1998,35(3):229-249.

[7] Luecht R M,Brumfield T,Breithaupt K. A testlet assembly design for adaptive multistage tests [J]. Applied Measure-

ment in Education,2006,19(3):189-202.

[8] Armstrong R D,Jones D H,Kunce C S. Study of a network-flow algorithm for test assembly [J]. Applied Psychological Measurement,1996,22:89-98.

[9] Armstrong R D,Jones D H,Koppe N B,et al. Computerized adaptive testing with multiple form structures [J]. Applied Psychological Measurement,2004,28(3):147-164.

[10] Armstrong R D,Jones D H,Wu I. An automated test development of parallel tests from a seed test [J]. Psychometrika,1992,57(2):271-288.

[11] Armstrong R D,Roussos L. A method to determine targets for multi-stage adaptive tests [R]. Newton,PA: Law school Adamission Council,2005.

[12] Belov D I,Armstrong R D. Monte Carlo test assembly for item pool analysis and extension [J]. Applied Psychological Measurement,2005,29(4):239-261.

[13] Belov D I,Armstrong R D. A Monte Carlo approach to the design,assembly,and evaluation of multistage adaptive tests [J]. Applied Psychological Measurement,2008,32(2):119-137.

[14] Breithaupt K,Hare D R. Automated simultaneous assembly of multistage testlets for a high-stakes license examination [J]. Educational and Psychological Measurement,2007,67(1):5-20.

[15] Chen Lingyin. An investigation of the optimal test designs for multi-stage test using the generalized partial credit model [D]. Austin: University of Texas at Austin,2011.

[16] Chen Peihua,Chang Huahua,Wu Haiyan. Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach [J]. Educational and Psychological Measurement,2012,72(6):933-953.

[17] Cheng Ying,Chang Huahua. The maximum priority index method for severely constrained item selection in computerized adaptive testing [J]. British Journal of Mathematical and Statiatical Psychology,2009,62(2):369-383.

[18] Hambleton R K,Xing Dehui. Optimal and nonoptimal computer-based test designs for marking pass-fail decision [J]. Applied Measurement in Education,2006,19(3):221-239.

[19] Jodoin M G,Zenisky A,Hambleton R K. Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purpose [J]. Applied Measurement in Education,2006,19(3):203-220.

[20] Keng L. A comparison of the performance of testlet-based computer adaptive tests and multistage tests [D]. Austin: University of Texas at Austin,2008.

[21] Luecht R M. Implementing the computer-adaptive sequential testing(CAST) framework to mass produce high quality computer-adaptive and mastery tests [EB/OL]. [2017-01-16]. http://files.eric.ed.gov/fulltext/ED442823.pdf.

[22] Luecht R M. Exposure control using adaptive multi-stage item bundle [EB/OL]. [2017-01-16]. http://files.eric.ed.gov/fulltext/ED475831.pdf.

[23] Luecht R M,Burgin W. Test information targeting strategies for adaptive multistage testing designs [EB/OL]. [2017-01-16]. http://files.eric.ed.gov/fulltext/ED475830.pdf.

[24] Patsula L N. A comparison of computerized adaptive testing and multistage testing [D]. Amherst: University of Massachusetts,Amherst,1999.

[25] Samejima F. A use of the information function for tailored testing [J]. Applied Psychological Measurement,1977,1(2):233-247.

[26] Zheng Ying,Nozawa Y,Gao Xiaohong,et al. Multistage adaptive testing for a Large-Scale classification test: design,heuristic assembly,and comparison with other testing modes [R]. ACT Research Report Series,2012.

[27] Zheng Yi,Chang Huahua. On-the-fly assembled multistage adaptive testing [J]. Applied Psychological Measurement,2015,39(2):104-118.

[28] Yan Duanli,Alina A,von Davier. Computerized multistage testing theory and applications [M]. New York: CRC Press,2014.

[29] Gulliksen H. Theory of mental tests [M]. New York: John Wiley,1950.

[30] Lord F M. Practical applications of item characteristic curve theory [J]. Journal of Educational Measurement,1977,14(2):117-138.

[31] Swanson L,Stocking M L. A model and heuristic for solving very large item selection problems [J]. Applied Psychological Measurement,1993,17(2):151-166.

[32] Theunissen T J J M. Binary programming and test designs [J]. Psychometrika,1985,50(4):411-420.

[33] Theunissen T J J M. Some applications of optimization algorithms in test design and adaptive test [J]. Applied Psychological Measurement,1989,10(4):381-389.

[34] Timminga E. The construction of parallel tests from IRT-Based item banks [J]. Journal of Education Statistics,1990,15(2):129-145.

[35] van der Linden W J. Optimal assembly of psychological and educational tests [J]. Applied Psychological Measurement,1998,22(22):195-211.

[36] van der Linden W J,Glas G A W. Computerized adaptive testing: Theory and practice [M]. Netherlands: Kluwer,2000.

[37] Ackerman T. An alternative methodology for creating parallel test forms using the IRT information function [EB/OL]. [2017-01-16]. http://files.eric.ed.gov/fulltext/ED306279.pdf.

(下转第 483 页)

[18] Hong Anxiang,Chen Gang,Li Junli,et al. A flower image retrieval method based on ROI feature [J]. Journal of Zhejiang University: Science A,2004,5(7):764-772.

An Improved Depth Convolutional Neural Network for Fine Image Classification

YANG Guoliang,WANG Zhiyuan,ZHANG Yu
(School of Electrical Engineering and Automation,Jiangxi University of Science and Technology,Ganzhou Jiangxi 34100,China)

Abstract: Fine image classification is different from traditional image classification. Due to the similarity between intraclass and intraclass differences of fine-grained images themselves,it is difficult to express the characteristics of fine image based on manual feature and local feature combination method. Based on the improved depth convolution neural network model,due to the large number of deep convolution neural network structure parameters and the large number of neurons,the training model is difficult,and the Gaussian distribution is used to initialize the first six parameters. The activation function is used after the correction of the Relus-Softplus function,the TOP1 accuracy rate of the flower image database OXford-102 flowers is 85.75%,and the TOP3 accuracy rate is 94.50%. The experimental results show that the model has obvious advantages over the traditional method,and the recognition rate is higher than that of the unmodified CNN model.

Key words: fine-grained image classification; deep convolutional neural network; activation function; feature extraction

(责任编辑: 冉小晓)

(上接第 469 页)

[38] 涂冬波,蔡艳,丁树良. 认知诊断理论、方法与应用 [M]. 北京: 北京师范大学出版社,2012.
[39] 高椿雷,罗照盛,郑蝉金,等. 具有认知诊断功能的多阶段自适应测验及其影响因素研究 [J]. 心理科学,2015,39(6):1492-1499.
[40] Timminga E. Solving infeasibility problems in computerized test assembly [J]. Applied Psychological Measurement,1998,22(3):280-291.

The Test Assemble Methods of Multistage Adaptive Test

LI Guiyu,TU Dongbo*,DAI Buyun,ZONG Yitao,GAO Xuliang,MIAO Ying
(School of Psychology,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: The key to implement multistage adaptive test(MST) is to build multiple parallel tests(or panels) that meet statistical and nonstatistical constraints required. Automated test assemble(ATA) provides a way to achieve parallel tests. Existing assemble methods are mainly based on linear programming,heuristic algorithm,Monte Carlo and on-the-fly method. Future studies should focus on the comparison and improvement of these methods and the development of ATA based on cognitive diagnosis tests.

Key words: multistage adaptive Test; automated test assemble; test assemble methods; test specification

(责任编辑: 冉小晓)