心理科学进展 2013, Vol. 21, No. 10, 1883–1892
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2013.01883

在线标定技术在计算机化自适应测验中的应用*

陈 平[1] 张佳慧[2] 辛 涛[2]

([1]北京师范大学认知神经科学与学习国家重点实验室, 北京 100875)
([2]北京师范大学发展心理研究所, 北京 100875)

摘 要 计算机化自适应测验 (*Computerized Adaptive Testing*, CAT) 近年来得到迅猛发展, 题目增补对 CAT 的题库建设与维护至关重要。新题标定作为题目增补过程中的技术难点, 它的精度直接影响被试能力估计的准确性, 目前在线标定技术经常用于标定新题。从在线标定设计和在线标定方法两个方面, 对两类 CATs (以项目反应理论为基础的传统 CAT 以及以认知诊断理论为基础的认知诊断 CAT (*Cognitive Diagnostic CAT*, CD-CAT)) 中的在线标定相关研究进行述评。传统 CAT 领域有着较丰富的研究, 而 CD-CAT 的在线标定研究则刚刚起步。未来研究应进一步探讨在线标定设计/方法之间的比较与结合, 以及 CD-CAT 和多维 CAT 的在线标定研究。

关键词 在线标定; 计算机化自适应测验; 题库建设; 认知诊断计算机化自适应测验
分类号 B841

1 引言

近年来, 随着教育测量理论、计算机编程技术与网络通讯技术的飞速发展, 计算机化自适应测验(*Computerized Adaptive Testing*, CAT)已经成为很多大规模评价项目(比如, 美国医生护士资格考试 NCLEX、美国商学院研究生入学考试 GMAT 和美国军事服役职业能力测验倾向成套测验 ASVAB)的首选。CAT 以项目反应理论(*Item Response Theory*, IRT)为指导依靠大型题库, 自行去适应被试水平, 灵活施测难度最恰当而且性能最优的题目, 从而实现对被试的高效测量 (漆书青, 戴海崎, 丁树良, 2002)。CAT 最主要的优点是它的能力估计效率明显高于传统的纸笔测验, 即采用相同数量的题目, CAT 可以达到更高的精度(Wainer et al., 1990; Weiss, 1982)。CAT 的其他优点还包括：公平、公正且高效的评分与报告呈现、与多媒体技术(如音频和视频)结合可进行听力测试和口语测试、题库得到良好维护时可提供连续测验(continuous testing) (Cheng, 2008)。

CAT 由题库、初始题目选择方法、能力估计方法、选题策略以及终止规则 5 个重要部分组成, 其中题库是 CAT 使用的前提, 因为每名被试作答的题目均来自题库(陈平, 辛涛, 2011a)。题库本身并不是对大量题目的简单堆积, 而需要依托专业的团队使用科学的方法进行构建。Flaugher (2000) 将 CAT 的题库建设归纳为 5 个步骤：(1)基于测验双向细目表或测验细则(test specification), 在每个内容领域开发足够多的、难度分布较广的题目; (2)对题目质量进行专业性检查(specialist review)和敏感性检查(sensitivity review), 以保证题目的高质量并且避免出现对某些被试子样本(如少数民族)有偏或有冒犯性的内容; (3)对新开发的题目进行预试, 以收集预试数据并检查题目是否测量它们想要测的内容; (4)基于预试数据采用经典测验理论(*Classical Test Theory*, CTT)和 IRT 对题目进行分析, 并根据相关标准修改或删除不合格的题目; (5)将题目转换成计算机呈现形式。使用大型数据库软件(如 SQL Server、MySQL)存储题干(item stem)、题目选项(item option)、题目正确答案(item key)和题目参数值等重要信息。

需要强调的是, CAT 在使用一段时间后, 研

收稿日期：2012-12-01
* 北京师范大学青年教师科学基金项目资助。
通讯作者：陈平, E-mail: pchen@bnu.edu.cn

究者和实践者往往需要考虑一个新的问题，即题库中的某些题目可能会因为过度曝光、存在缺陷或者过时等原因不再适合被继续使用(Wainer & Mislevy, 1990)，游晓锋、丁树良和刘红云(2010)建议可以淘汰质量不高的题目、让过度曝光的题目“休眠”等；另一方面，测验管理者需要不断地向题库补充新鲜“血液”，对题库进行题目增补(item replenishing)。这就需要邀请专家不断编制新题[1]，并对新题进行标定，然后才能将其添加到题库当中(Guo & Wang, 2003)。其中，对新题的标定既是重点也是难点，标定的精度将直接影响到被试能力估计的准确性，因为来自题目标定的误差会直接传递到对被试的评分过程中 (Cheng & Yuan, 2010)。另外，新题的标定具体包括两方面的含义：一是估计新题的题目参数，二是将新题参数置于旧题的参数量尺上(陈平，辛涛, 2011a)。

Wainer 和 Mislevy (1990)概括了 CAT 中两种标定新题的策略：(1)传统标定策略。该策略在旧题和新题之间设置一些共同题(也称为锚题)，然后使用单独的标定样本 (calibration sample) 估计新题，接着基于锚题的新旧两套参数估计值使用 Stocking-Lord (Stocking & Lord, 1983)等方法计算线性转换系数 (如 A 和 B)，最后将新题和旧题置于同一参数量尺上。这种策略的思路简单清晰，但缺点也比较明显：对新题使用单独的标定样本意味着需要耗费额外的资源 (如钱和时间)，而且为了减小链接误差(linking error)需要设置较多的共同题；(2)在线标定策略(online calibration strategy)。在线标定是在被试自适应作答旧题的过程中，将新题呈现给被试作答以收集被试在新题上的作答反应，并估计新题题目参数的技术(Wainer & Mislevy, 1990)。测验开始前，主试会正式告知被试他们在某些题目(新题)上的作答反应不参与分数计算或能力估计，而不是将新题暗中散布到测验当中以获得可靠的数据。从某种意义上讲，既作答旧题又作答新题的被试实际上起到的是锚人(anchor person)设计中锚人的作用，所以不需要再单独进行等值。相对于传统标定策略，在线标定策略具有“不需要额外的标定研究即可在估计被试能力的同时也标定新题、不需要复杂的等值方法或设计即可将新题和旧题参数置于同一量尺、被试在作答新题和旧题时具有相同的考试动机”等优点(陈平，辛涛，2011b)，目前广泛应用于传统 CAT 的新题标定中(e.g., Chang & Lu, 2010; Makransky, 2009; 游晓锋等, 2010)。

Chang (2012)以及唐小娟、丁树良和俞宗火(2012)将 CAT 大致分为两类：除了以 IRT 为基础的传统 CAT，还有一类是以认知诊断理论(*Cognitive Diagnostic Theory*, CDT)为基础的认知诊断 CAT (*Cognitive Diagnostic CAT*, CD-CAT)。CD-CAT 将认知诊断和 CAT 两者相结合，不仅可以提供关于被试掌握/未掌握哪些属性的诊断反馈，还可以提高诊断测量的准确性与效率，近年来在教育测量领域得到愈来愈广泛的关注。类似于传统 CAT，题库也是 CD-CAT 使用的前提。而且随着时间的推移，CD-CAT 的题库维护与管理显得愈发重要，比如, CD-CAT 也存在着对题库进行题目增补和新题标定(即准确估计新题的题目参数并将它们置于旧题参数量尺上)的问题。只不过不同于传统 CAT 的单维结构，CD-CAT 具有多维结构(需要估计的被试属性掌握模式或知识状态是多维离散变量)，这使得 CD-CAT 中的题目增补较传统 CAT 更为复杂，除了标定新题，还需要标识新题对应的 Q 矩阵。Chen, Xin, Wang 和 Chang (2010, 2012)将在线标定技术引入 CD-CAT 中，并将传统 CAT 中 3 种在线标定方法推广到 CD-CAT, 这说明在 CD-CAT 领域中在线标定技术也是一种值得推广的好办法，特别是当认知诊断等值方法和技术的发展还不够成熟的时候。

其实除了传统 CAT 和 CD-CAT，基于多维 IRT (*Multidimensional IRT*, MIRT)的多维 CAT (*Multidimensional CAT*, MCAT) 近年来已经成为 CAT 发展的一个新方向(陈平, 2011)。MCAT 将自适应测验与多维潜在特质(multitrait)估计相结合，在发展形成性评价(formative assessment)方面具有很大潜力(Wang & Chang, 2011)。MCAT 传承了传统 CAT 的很多优点，比如较传统的非自适应测验，它用较少的题目就可以得到更加准确的能力估计值；另一方面，MCAT 可以提供被试在一系列分量表(subscale)上的信息，这些信息有助于标识被试在测验所测属性上的优缺点(Wang & Chang, 2011)。与传统 CAT 和 CD-CAT 一样，

[1] 将题库中参数已标定 (calibration) 的题目称为旧题，新题是相对于旧题而言。

Reckase (2009)将题库视为 MCAT 使用的重要前提。尽管目前还没有研究探讨 MCAT 题库的设计(Reckase, 2009), 但是可以预见的是, 新题的增补和新题的标定对 MCAT 题库的维护与管理同样至关重要。目前, 在 MCAT 领域还没有看到公开发表的论文探讨在线标定。

因此, 本文对在线标定技术在 CAT (包括传统 CAT 和 CD-CAT)中的研究进展进行述评。由于在线标定设计(online calibration design)与在线标定方法(online calibration method)是在线标定的两个重要环节, 所以第 2 节详细讨论 CAT 中的在线标定设计, 第 3 节对 CAT 中的在线标定方法展开讨论, 最后一节呈现已有研究存在的重要问题并展望今后的研究方向。

2　CAT 中的在线标定设计

作为在线标定的重要环节, 在线标定设计关注的是被试在自适应作答旧题的过程中, 如何将新题分配给被试作答可以获得更为精确的标定结果：是以随机的方式将新题分配给被试作答, 还是以自适应的方式分配, 还是采用其他的方式。Wainer 和 Mislevy (1990)以及游晓锋等人(2010)描述的新题植入(seeding new items)方法对应于这里的在线标定设计。接下来, 分别对传统 CAT 和 CD-CAT 中的在线标定设计按提出的时间顺序进行述评。

2.1　传统 CAT 中的在线标定设计

Wainer 和 Mislevy (1990)认为在进行在线标定时, 可通过 2 种设计方式将新题植入被试的 CAT 测验过程中：(1)随机设计。即对每位被试, 从新题题集中随机选择固定数量的新题, 然后将选中的新题植入被试 CAT 测验中的随机位置。随机设计有两个特点, 一是随机选题, 二是植入测验的随机位置。这种设计实施起来虽然比较简单和方便, 但是没有充分反映 CAT 测验在线标定过程中的选“人”逻辑, 也没有充分体现 CAT“自适应”的特点。而且, 被试在作答过程中很可能会突然遇到非常容易或非常难的新题, 这样被试就有可能区分出哪些是新题、哪些是旧题 (Jones & Jin, 1994), 从而导致被试有可能不尽全力作答新题, 收集到的“被试在新题上的作答反应”有可能会不真实。另外, 游晓锋等人(2010)在探讨传统 CAT 中新题 (原文中称为原始题) 题目参数的在线标定时, 采用的新题植入方法也属于随机设计; (2)自适应设计。Lord (1980)指出在自适应测验中, 为了高效估计被试能力, 选题策略应该基于被试在已作答题目上的表现选择最适合被试作答的题目。类似地, 为了更为高效地估计新题的题目参数并充分利用在线标定技术的特点, Wainer 和 Mislevy (1990)、Jones 和 Jin (1994)及 Chang 和 Lu (2010)都建议参与标定过程的被试也应该自适应地进行选择, 或者说将新题以自适应的方式呈现给被试作答。但是在具体实施自适应设计时, 有一个难点是需要知道新题的题目参数。Wainer 和 Mislevy (1990)给出的建议是, 可以基于出题者对题目的主观判断(subjective judgment)给出粗略的题目参数估计值。

Makransky (2009)考虑到在真实职业测验(occupational testing)的测验开发阶段中题目标定所需的资源(如被试样本)很难获得, 于是在事先没有 CAT 题库的情形下, 提出 3 种自动在线标定设计用于在估计被试能力的同时对题目进行标定, 它们分别是两阶段策略(*Two-Phase Strategy*, P2)、多阶段策略(*Multi-Phase Strategy*, M)以及连续更新策略(*Continuous Updating Strategy*, C)。P2 包括随机和自适应两个阶段。在随机阶段, 题目以随机的形式分配给固定数量的被试作答, 然后在估计这批被试的能力水平时假设所有题目的难度参数值都等于预设值 0。而且, 在每个阶段结束后都会对题目进行标定。在自适应阶段, 题目以自适应的方式呈现给剩余被试作答, 然后基于随机阶段得到的题目参数估计值对被试能力进行估计。当所有题目的平均被作答次数(或平均曝光次数)超过某个预设值(如 50 次) 时, P2 由随机阶段过渡到自适应阶段。M 策略是在 P2 的基础上由多于 2 个以上的阶段组成, 当收集到越来越多的数据使得自适应算法可用时, 测验中各部分由全部是随机选择逐渐过渡到全部是自适应选择(M 策略的一个实例详见表 1)。C 策略的第一个阶段与 P2 和 M 类似, 其连续更新的特点主要体现在最后一个阶段(此时所有测验部分都是自适应选择) :题目每曝光一次, 其题目参数就更新一次, 然后基于最新的题目参数值对被试能力估计一次。Makransky (2009)的研究表明 C 策略在所有模拟条件下都一致优于其他两种策略。值得注意的是, Makransky (2009)提出的 3 种自动在线标定设计都是以随机

设计和自适应设计为基础，但是在“事先没有 CAT 题库、没有任何旧题信息可以利用、所有题目都被视为新题”的前提下提出的，所以严格意义上讲，这 3 种设计并不满足在线标定的原始定义，因为在线标定是指在“已有 CAT 题库”的前提下将新题植入被试 CAT 测验过程中然后估计新题题目参数的过程。另外，在随机阶段“为估计被试能力水平，而假设所有题目的难度参数值都等于 0”的做法是否合理也还值得商榷。

表 1 多阶段策略实例示意表

阶段	测验部分 1	测验部分 2	测验部分 3
1	随机选择	随机选择	随机选择
2	随机选择	随机选择	自适应选择
3	随机选择	自适应选择	自适应选择
4	自适应选择	自适应选择	自适应选择

Chang 和 Lu (2010)在不定长 CAT 的情境中提出在线标定的序贯设计(sequential design)，这种设计选择最合适的被试[2]参与新题题目参数的估计。具体而言，首先根据被试在旧题上的作答反应估计被试的能力水平，接着基于 D-最优设计(Fedorov, 1972; Silvey, 1980)序贯地、自适应地选择最合适的被试参与新题的作答(称为设计阶段)，然后基于已选被试在新题上的作答反应以及他们的能力值估计新题的题目参数(称为参数估计阶段)，最后不断重复设计阶段和参数估计阶段直到新题的参数估计值满足预先设定的精度。按照序贯设计的逻辑，很容易知道它并不适用于真实情境下的 CAT 测验。因为被试参加完 CAT 测验并不允许立即离开，他们需要等待能力估计的结果并确定是否被选为设计点。如果被选为设计点，他们接下来需要对新题进行作答。很明显，如果被试在 CAT 测验过程中作答所有的新题，那么上述问题将不会存在，因为在参数估计阶段前所有被试对所有新题的作答反应都是已知的。但是，如果新题的数量相对较大时，要求被试除了作答旧题还要作答所有的新题是不太可能的，因为还需要考虑疲劳效应以及其它的一些效应。

尽管传统 CAT 领域已有不少在线标定设计，但遗憾的是，目前还没有看到将以上各种在线标定设计进行比较的研究文献。

2.2 CD-CAT 中的在线标定设计

不同于传统 CAT，目前 CD-CAT 中关于在线标定设计的研究还不多，只涉及随机和自适应两种设计。汪文义、丁树良和游晓锋(2011)在研究 CD-CAT 新题 (原文中称为原始题) 属性标定的实验过程中，将从新题题集中随机选择的新题植入被试 CD-CAT 的随机位置，所以他们的设计属于随机设计。随机设计的优缺点在传统 CAT 的在线标定设计中已经提及，这里不再赘述。

Chen 等人(2010, 2012)受 Makransky (2009)所提出的标定策略的启发，采用自适应设计的思路标定新题，但在具体实施时又与 Makransky (2009)的设计方案稍有不同。为了能够为每位被试自适应地选择新题，他们采用基于数据(data-based)的方法确定新题的初始参数估计值：首先将新题随机分配给被试的子样本(如前 25%的被试) 作答，并使用在线标定方法(详见“3 CAT 中的在线标定方法”) 对新题进行预标定(称为预标定阶段，记为 PC)；然后对于剩余被试(如后 75%的被试)，CD-CAT 测验基于新题的题目参数预估值自适应地选择新题给被试作答；最后基于剩余被试在新题上的作答反应对新题进行重新标定(称为重新标定阶段，记为 RC)。由于参与预标定阶段和重新标定阶段的被试量都会影响最后的标定结果，所以如何有效地从预标定阶段过渡到重新标定阶段是自适应标定设计中的一个关键问题。Chen 等人 (2010, 2012) 通过对参与预标定阶段和重新标定阶段的被试量设置 3 种不同的比例(1:3, 1:1 和 3:1)，实现 3 种在线自适应标定设计方案(详见图 1)。这 3 种设计方案的表现优劣，文中并未给出结论。但是通过文中提供的模拟数据可以发现：对于这 3 种方案，孰优孰劣并没有明显的规律，因为不同的在线标定方法以及不同的实验条件(不同题目参数范围与不同属性掌握概率的组合)会产生不一致的结果。另外，从理论上讲，相对于随机设计，自适应设计应该能够产生更加精确的标定结果。但是他们的模拟研究表明：在绝大多数模拟情境下，特别是当“确定性输入、噪音‘与’门” (*the Deterministic Inputs, Noisy “and” Gate*, DINA)模型的题目参数(猜测参数 g 和失误参数 s)值较大($g,s \in (0.25,0.45)$)时，自适应设计并未像

[2] 实验设计 (experimental design) 中的术语是设计点 (design points)。

预期一样可以改善题目参数的返真性。对于这个问题，从在线标定设计和实验设计两方面都需要进行更加深入的分析与探讨，详见“4 问题与展望”部分。

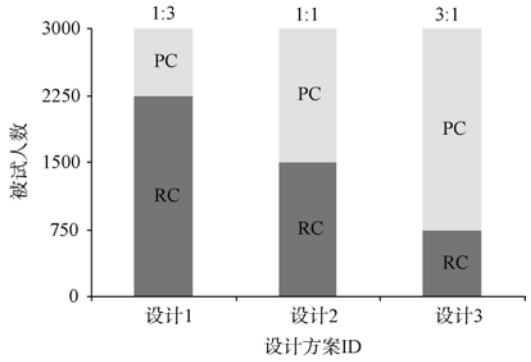

图 1　3 种在线自适应标定设计方案

2.3　CAT 中在线标定设计的优劣对比

表 2 对目前 CAT (传统 CAT 和 CD-CAT)中已有在线标定设计的优缺点进行了对比，可为测量研究者和实践者在选择在线标定设计时提供依据。根据表 2，很容易知道各种在线标定设计的适用情境。比如，对于传统 CAT，如果是在事先没有 CAT 题库的前提下对题目进行标定，建议借鉴 Makransky (2009)的 3 种自动在线标定设计中表现最好的连续更新策略(C 策略)。如果事先存在 CAT 题库，建设采用 Wainer 和 Mislevy (1990)提出的自适应设计对新题进行在线标定；当新题题目数量相对较小(比如 5 题或 10 题)时，要求被试在 CAT 测验过程中作答所有的新题是可以接受的，这时能够严格满足预设标定精度的序贯设计是一个非常好的选择。另外，对于 CD-CAT，如果选择 DINA 模型作为认知诊断模型，建议使用实施起来更为简单、更为方便的随机设计，因为 Chen 等人(2010，2012)的研究表明随机设计已经可以获得较高的标定精度。

3　CAT 中的在线标定方法

不同于在线标定设计，在线标定方法的主要任务是在收集完被试参加 CAT 测验的作答反应之后估计新题的题目参数，并将它们置于旧题的参数量尺上。类似地，接下来分别对传统 CAT 和 CD-CAT 中的在线标定方法进行述评。

3.1　传统 CAT 中的在线标定方法

在过去的 20 多年里，研究者提出多种在线标定方法，根据标定思路的不同可以将它们归为以下三类：

3.1.1　条件极大似然估计方法

Stocking (1988)提出方法 A (Method A)和方法 B (Method B)两种在线标定方法，本质上，它们在标定新题时都使用了两次条件极大似然估计(*Conditional Maximum Likelihood Estimation*, CMLE)。方法 A 首先基于被试在旧题上的作答反应使用 CMLE 估计被试的能力值；其次，将被试

表 2　CAT 中各种在线标定设计优劣对比

CAT 类别	在线标定设计	优点	缺点	备注
传统 CAT	随机设计 (Wainer & Mislevy, 1990)	实施方便和简单	没有充分体现 CAT“自适应”的特点	目前尚未看到研究文献对这 4 种在线标定设计进行比较
	自适应设计 (Wainer & Mislevy, 1990)	充分体现 CAT 在线标定过程中的“选人”逻辑	为了为被试自适应地选题，对题目参数初始值的提供相对粗糙	
	自动在线标定设计 (Makransky, 2009)	结合随机设计和自适应设计，且适用于事先没有 CAT 题库的情形	随机阶段“假设所有题目的难度参数值都等于 0”的做法有待改进	
	序贯设计 (Chang & Lu, 2010)	可以严格满足预先设定的标定精度	不太适用于真实情境下的 CAT 测验	
CD-CAT	随机设计 (汪文义等, 2011)	实施方便和简单	无法体现 CAT 在线标定过程中的“选人”逻辑	Chen 等人(2010，2012)的研究表明：相对于随机设计，自适应设计并未显著改善题目参数的返真性
	自适应设计 (Chen et al., 2010, 2012)	在重新标定阶段，为部分被试实现了自适应选择新题	在预标定阶段，仍有部分被试是以随机方式作答新题	

能力估计值固定(即看成是能力真值)，然后结合被试在新题上的作答反应再次使用 CMLE 估计新题的题目参数。在具体实施 CMLE 时，可使用牛顿-拉夫逊(*Newton-Raphson*, N-R)迭代方法或者二分法(bisection method)或者两者的结合(先使用二分法再使用 N-R，然后再重复使用这两种方法可提高迭代速度) 求解非线性对数似然方程。注意方法 A 有一个强假设—— 将能力估计值看成是能力真值，这样处理可以将固定的能力值作为“桥梁”将旧题和新题参数置于同一量尺。但方法 A 的缺点也很明显，比如能力估计的偏差会直接传递到对新题的标定过程中，还有可能会产生不希望出现的量尺漂移(scale drift)。而方法 B 在一定程度上可以克服方法 A 的理论缺陷，因为它借助参数已经标定的锚题对量尺漂移现象进行校正。方法 B 中，每名被试除了作答旧题和新题外还需要作答一些锚题，不同于方法 A，它将固定的能力值既用于标定新题又用于标定锚题。于是每个锚题都有两套参数值(一套与旧题参数在同一量尺上，另一套与新题参数在同一量尺上)，然后基于锚题的新旧两套参数使用 Stocking-Lord 方法(Stocking & Lord, 1983)可以将新题参数置于旧题参数量尺上。方法 B 由于需要使用锚题，所以往往需要更大的样本量或更长的测验长度。

游晓锋等人(2010)提出夹逼平均法求取题目的难度，还提出双参数 CMLE 方法和多重迭代 CMLE 方法在线标定新题。夹逼平均法基于“当作答人数较多时，正确作答与错误作答题目的被试中总是存在一些被试的能力值与题目的难度值非常接近”的思路求取题目难度值 b。双参数 CMLE 的在线标定方法首先基于被试在自适应阶段的作答反应估计被试的能力值 θ，接着使用夹逼平均法获得新题的难度估计值 $\hat{b}$，然后将被试能力和新题难度估计值固定(看成已知)，最后使用 CMLE 方法仅仅估计新题的区分度参数 a。这种方法不同于传统双参数 CMLE 方法，传统双参数 CMLE 方法的思路是首先通过夹逼平均法以及“先将新题的通过率 p 转换为标准正态分数 z ($\frac{1}{\sqrt{2\pi}}\int_z^{+\infty} e^{-\frac{t^2}{2}}dt = p$)，然后计算二列相关系数 r_b ($r_b = z/b^{(0)}$)，再进行转换 ($a^{(0)} = r_b \Big/ \sqrt{1-r_b^2}$) ”的方式依次获得新题的难度初值($b^{(0)}$)和区分度初值($a^{(0)}$)，然后在被试能力 $\hat{\theta}$ 已经获得的前提下使用 CMLE 方法同时估计新题的难度 b 和区分度 a。而游晓锋等人(2010)将由夹逼平均法得到的难度值固定(注意不是作为难度初值 $b^{(0)}$，而是最终的难度估计值 $\hat{b}$)，然后仅对区分度参数 a 进行 CMLE 估计。尽管文中呈现的模拟结果较好，但是这种“将题目的两个参数分开估计”的做法是否合理、是否具有理论依据，还有待进一步的论证。另外，仅由夹逼平均法得到的难度估计值是否比传统做法得到的难度估计值 (经过多步迭代) 精度更高呢？这也需要进行更为深入的研究与探讨。多重迭代 CMLE 方法是对双参数 CMLE 方法的扩展，所以上述疑问同样存在于多重迭代 CMLE 方法当中。

3.1.2 EM 方法

Wainer 和 Mislevy (1990) 提出经典的只有一个 EM 循环的边际极大似然估计 (*Marginal Maximum Likelihood Estimate with one EM cycle*, OEM) 方法。OEM 在实施 MMLE/EM 算法时仅包含一个 EM 循环：在 E 步(期望步)中，仅基于被试在旧题上的作答反应计算能力的后验分布；在 M 步中(最大步)，仅基于被试在新题上的作答反应通过最大化对数边际似然函数来估计新题的题目参数。因为 E 步中的能力后验分布是根据被试在旧题上的作答反应计算得来，所以理论上新题与旧题的题目参数会在同一量尺上。OEM 具有 EM 算法的优点并允许新题一个一个被标定，其不足之处在于，它在计算能力后验分布时没有充分吸收来自新题的信息。

Ban, Hanson, Wang, Yi 和 Harris (2001)在 OEM 的基础上提出有多个 EM 循环的边际极大似然估计(*Marginal Maximum Likelihood Estimate with Multiple EM cycles*, MEM)方法。MEM 是对 OEM 的扩展，它包括多个 EM 循环。当 EM 循环次数等于 1 时, MEM 等价于 OEM，注意从第一个 EM 循环得到的新题题目参数估计值可以作为第二个 EM 循环的新题题目参数初始值。从 MEM 的第二个 EM 循环开始，被试在旧题和新题上的作答反应都用于计算能力的后验分布(E 步)，并且将旧题的题目参数固定为常数，然后不断通过最大化对数边际似然函数来更新新题的题目参数(M 步)，重复 E 步和 M 步直到满足预先设定的迭代精度。因为在 M 步中固定了旧题的题目参数，

所以理论上新题的题目参数与旧题的题目参数会在同一量尺上。MEM 方法的优点是它在 OEM 的基础上充分利用来自新题的信息。

3.1.3　IRT 软件方法

Ban 等人 (2001) 还提出基于 IRT 参数估计软件 BILOG (Mislevy & Bock, 1990)的 BILOG/Prior 方法, 该方法仅使用 BILOG 软件就能完成整个的新题标定过程。不同于方法 A 和方法 B 是通过固定能力估计值来标定新题, BILOG/Prior 方法本质上是通过对旧题设置较强的先验分布、对新题设置默认的先验分布(如 $\ln a \sim N(0,1)$ 和 $b \sim N(0,1)$)来近似固定旧题的题目参数。对旧题设置较强先验分布的做法是：将旧题先验分布的均值设置为旧题的参数估计值, 然后设置很小的先验方差。这种方法同时对旧题和新题进行估计, 于是每个旧题都有两套参数估计值, 通过对旧题先验分布设置很小的方差可以使旧题和新题近似地处在同一量尺上。值得注意的是, 这种方法并不适用于小样本, 因为被试对旧题的作答反应一般构成稀疏矩阵(sparse matrix), 而且当每个题上的作答反应较少(e.g., 50)时运行 BILOG 软件常常报错。

Ban 等人(2001)对 5 种在线标定方法(方法 A、方法 B、OEM 方法、MEM 方法以及 BILOG/ Prior 方法)在 3 种不同样本大小(300、1000 和 3000)下的表现进行比较, 并将题目参数返真性作为评价指标。他们的研究发现, MEM 方法的表现最优, 因为在所有的样本情境下, MEM 产生的参数估计误差都最小; 方法 A 由于其理论缺陷, 具有最大的参数估计误差。

3.2　CD-CAT 中的在线标定方法

在 CD-CAT 领域中, 研究在线标定方法的文献不多, Chen 等人(2010, 2012)将传统 CAT 中 3 种有代表性的在线标定方法 (Method A、OEM 和 MEM) 推广到 CD-CAT 领域, 分别记为 CD-Method A、CD-OEM 和 CD-MEM。在推广过程中, 各种方法的标定思路并没有发生改变, 变化的只是模型以及相应的计算公式。他们的实验结果表明：所有 3 种方法都能够比较准确地标定参数, 而且当 DINA 模型的猜测参数 g 和失误参数 s 值都较小(如 $g, s \in (0.05, 0.25)$)时, CD-Method A 在题目参数返真性方面的表现优于 CD-OEM 和 CD-MEM。需要强调的是, 类似于方法 A, CD-Method A 方法也具有明显的理论缺陷, 即将被试知识状态估计值看成是知识状态真值, 这样知识状态的估计误差会传递到新题的标定过程中。所以当 DINA 模型的 g 和 s 值都较大 (如 $g, s \in (0.25, 0.45)$) 导致模式判准率较低 (即知识状态估计误差较大) 时, CD-Method A 的表现不如其他两种方法。

3.3　CAT 中在线标定方法的特点比较

表 3 对目前 CAT 中已有在线标定方法的特点进行比较, 旨在为将来测量研究者与实践者在选择在线标定方法时提供借鉴。从表 3 中, 可以很容易了解各种在线标定方法的适用情境。比如, 对于传统 CAT, 建议使用标定精度最高且对样本量无特殊要求的 MEM 方法; 但是如果在编写程序方面存在困难的话, 也可以选用 BILOG/Prior 方法, 因为该方法仅仅操作 BILOG 软件就能够达到与 MEM 接近的标定精度, 但必须保证每个新题上的作答反应人数大于 100。如果担心 MEM 在迭代计算时所花时间太多的话, 也可以退而求其次使用标定精度相对较高且对样本量无特殊要求的 OEM 方法。此外, 对于 CD-CAT, 如果将 DINA 模型选作认知诊断模型, 在线标定方法的选择较大程度上依赖于题目参数的大小。当题目参数都较小(如小于 0.25)时, 建议使用更为简单的 CD-Method A 方法; 而当题目参数都较大 (如大于 0.25) 时, 可以选用算法较简单但标定精度很高的 CD-OEM 方法。

4　问题与展望

目前对 CAT 中在线标定技术的研究尚属起步阶段, 已有研究也还不多, 而且在在线标定的两个重要环节(在线标定设计和在线标定方法)中都存在没有解决的问题。第一, 如前所述, 方法 A (Stocking, 1988)和 CD-Method A (Chen et al., 2012)都具有明显的理论缺陷, 比如, 它们都将估计的能力值或知识状态看成是真实的能力值或知识状态, 然后将它们用于标定新题。这样的话, 在整个的新题标定过程中就存在着两个误差来源：一个是标定过程本身所产生的误差, 另一个是由能力或知识状态的估计误差传递到题目标定过程中所产生的误差。为了校正能力估计过程中产生的估计误差, Jones 和 Jin (1994)建议使用全功能极大似然估计 (*Full Functional Maximum Likelihood*

Estimation, FFMLE)方法估计题目参数，而且 Stefanski 和 Carroll (1985)认为 FFMLE 估计量具有一致性，在样本量较大的条件下(如有效样本大小是 60)比传统 MLE 估计量的表现更优。因此，另一个值得考虑的研究方向是将 FFMLE 方法与方法 A 或 CD-Method A 方法结合起来使用，以在题目标定过程中对能力或知识状态的估计误差进行校正。

第二，在 CD-CAT 中的在线标定设计方面，Chen 等人(2010, 2012)的研究仅考虑 3 种“参与预标定阶段和参与重新标定阶段”的被试人数比例(1:3, 1:1 和 3:1)，得到的结果与结论可能没有完全真实地揭示其内部规律。今后很有必要对更多、更细的被试人数比例(如 1:3, 1:2, 2:3, 4:5, 1:1, 5:4, 3:2, 2:1 和 3:1)进行讨论，以更加准确地确定参与预标定阶段和重新标定阶段的最佳人数比例。另外，他们的研究仅考虑最简单的认知诊断模型(DINA 模型)以及最简单的属性层级结构(所有属性相互独立)，因此所得结果与结论的适用性和外推性比较有限。今后有必要观察随机标定设计与自适应标定设计在更复杂认知诊断模型(如 Fusion 模型、NIDA 模型和 DINO 模型)以及更复杂属性层级结构(线型、收敛型、发散型和无结构型)下的表现。

第三，在 CD-CAT 中的在线标定方法方面，Chen 等人(2010, 2012)的研究主要是以推广为主，并没有基于 CD-CAT 与在线标定技术自身的结构与特点，开发能够满足预设精度的在线标定方法;

表 3 CAT 中各种在线标定方法的特点

CAT 类别	在线标定方法	标定精度	算法复杂度	使用时对样本量有无特殊要求
传统 CAT	方法 A (Stocking, 1988)	最低	最简单	无
	方法 B (Stocking, 1988)	较高	较简单	被试的能力值用于同时对新题和锚题进行标定，所以较方法 A 需要更大的样本量
	双参数 CMLE 方法和多重迭代 CMLE 方法 (游晓锋等, 2010)	—	较简单	无
	OEM 方法 (Wainer & Mislevy, 1990)	较高	较复杂	无
	MEM 方法 (Ban et al., 2001)	最高	最复杂	无
	BILOG/Prior 方法 (Ban et al., 2001)	很高[3]	较简单	BILOG 软件要正常运行，每个题目上的作答反应人数最好大于 100
CD-CAT	CD-Method A (Chen et al., 2010, 2012)	当 DINA 模型参数值较小时，标定精度优于其它两种方法	最简单	无
	CD-OEM (Chen et al., 2010, 2012)	当 DINA 模型参数值较大时，标定精度优于 CD-Method A	较复杂	无
	CD-MEM (Chen et al., 2010, 2012)	与 CD-OEM 接近	最复杂	无

注：表中“—”说明：双参数和多重迭代 CMLE 方法并未与其他 5 种方法进行过比较，所以无法提供它的相对标定精度。

[3] Ban 等人 (2001) 的研究表明当有 10 个新题且样本量达到 3000 时, BILOG/Prior 方法的标定精度与 MEM 方法的结果很接近。

而且他们在研究中假设专家预先构建好新题 Q 矩阵并且假设它完全正确，这种做法过于理想。类似地，Chen 等人(2010，2012)仅考虑最为简单的 DINA 模型并假设所测属性相互独立，对于其它的认知诊断模型以及线型、收敛型和发散型等更可能存在的属性层级结构，这些在线标定方法的表现同样值得关注。

第四，目前常见的 IRT 模型大多是单维度(unidimensional)的模型，然而在实际的心理与教育测量问题中，传统 IRT 中的单维性假设往往是不成立的。而且测验资料的多维性与被试在完成测验任务时需要多种能力的共同配合是相符的(康春花，辛涛，2010)，因此从传统的单维 IRT 拓展为多维度 IRT (*Multidimensional IRT*，MIRT)就显得十分必要(涂冬波，蔡艳，戴海琦，丁树良，2011)。MIRT 模型不仅更加适合目前的测验资料，而且能够从多个维度(多个角度)表征题目与被试能力之间的关系(Reckase，2009)。所以，基于 MIRT 的 MCAT 得到愈来愈广泛的应用。目前，在传统 CAT 和 CD-CAT 中，在线标定技术经常用于标定新题，而在 MCAT 领域还没有看到关于在线标定的论文公开发表。因此，很有必要研究适用于 MCAT 的在线标定设计和在线标定方法。在线标定设计方面，传统 CAT 与 CD-CAT 中随机设计和自适应设计的思想可以很容易推广到 MCAT，至于还有没有其他的在线标定设计方式可以获得更为精确的标定结果，还有待进一步的研究与探讨；而在在线标定方法方面，传统 CAT 中几种经典的在线标定方法，如 Method A、OEM 和 MEM 方法是否能够直接推广至 MCAT，是否可以基于 MCAT 的多维结构以及在线标定的特点开发其他的在线标定方法，这也是今后值得研究的新方向。可以肯定的是，由于多维结构的引入，MCAT 中的被试能力估计方法以及在线标定算法都会更加复杂。

参考文献

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.

陈平，辛涛. (2011a). 认知诊断计算机化自适应测验中在线标定方法的开发. *心理学报, 43*, 710–724.

陈平，辛涛. (2011b). 认知诊断计算机化自适应测验中的项目增补. *心理学报, 43*, 836–850.

康春花，辛涛. (2010). 测验理论的新发展：多维项目反应理论. *心理科学进展, 18*, 530–536.

漆书青，戴海崎，丁树良. (2002). *现代教育与心理测量学原理*. 北京：高等教育出版社.

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*, 616–626.

涂冬波，蔡艳，戴海琦，丁树良. (2011). 多维项目反应理论：参数估计及其在心理测验中的应用. *心理学报, 43*, 1329–1340.

汪文义，丁树良，游晓锋. (2011). 计算机化自适应诊断测验中原始题的属性标定. *心理学报, 43*, 964–976.

游晓锋，丁树良，刘红云. (2010). 计算机化自适应测验中原始题项目参数的估计. *心理学报, 42*, 813–820.

Ban, J.-C., Hanson, B. A., Wang, T. Y., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191–212.

Chang, H. H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195–226). Charlotte, NC: Information Age.

Chang, Y. -C. I., & Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika, 75*, 140–157.

Chen, P., Xin, T., Wang, C., & Chang, H. (2010). *A comparative study on on-line calibration methods in cognitive diagnostic computerized adaptive testing*. Paper presented at the 75th meeting of the Psychometric Society. Athen, Georgia.

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*, 201–222.

Cheng, Y. (2008). *Computerized adaptive testing: New developments and applications*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

Cheng, Y., & Yuan, K. H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75*, 280–291.

Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J.Mislevy (Eds.), *Computerized adaptive testing: A primer* (Chap. 3, 2nd ed., pp. 37–59). Mahwah, NJ: Erlbaum.

Fedorov, V. V. (1972). *Theory of optimal design*. New York: Academic Press.

Guo, F. M., & Wang, L. (2003). *Online calibration and scale stability of a CAT program*. Paper presented at the annual meeting of National Council on Measurement in Education. Chicago, IL.

Jones, D. H., & Jin, Z. Y. (1994). Optimal sequential designs

for on-line item estimation. *Psychometrika, 59*, 59–75.

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

Makransky, G. (2009). *An automatic online calibration design in adaptive testing*. Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing, McLean, USA.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG3: Item analysis and test scoring with binary logistic models* (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software.

Reckase, M. D. (2009). *Multidimensional item response theory* (Chap. 10, pp. 311–339). New York: Springer.

Silvey, S. D. (1980). *Optimal design: An introduction to the theory for parameter estimation.* London: Chapman and Hall.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics, 13*, 1335–1351.

Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Rep. 88-28). Princeton, NJ: ETS.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (Chap. 4, pp. 65–102). Hillsdale, NJ: Erlbaum.

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika, 76*, 363–384.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.

Application of Online Calibration Technique in Computerized Adaptive Testing

CHEN Ping[1]; ZHANG Jiahui[2]; XIN Tao[2]

([1] *State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing* 100875, *China*)

([2] *Institute of Developmental Psychology, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Computerized Adaptive Testing (CAT) has developed rapidly in recent years, and item replenishing is essential for item bank construction and maintenance in CAT. Calibration of new items is a technical challenge in item replenishing, and the precision of which directly impacts the accuracy of the estimation of examinees' abilities. Now the online calibration technique has been commonly used to calibrate the new items. Studies on online calibration for two kinds of CATs (traditional CAT based on item response theory and cognitive diagnostic CAT (CD-CAT) based on cognitive diagnostic theory) are reviewed in the aspects of online calibration design and online calibration method. There have been relatively abundant research results for the traditional CAT, while online calibration studies of CD-CAT have just started. Future studies could further explore the comparison and combination of different online calibration designs/methods, as well as the online calibration for CD-CAT and multidimensional CAT.

Key words: online calibration; computerized adaptive testing; item bank construction; cognitive diagnostic computerized adaptive testing