计算机自适应测验在美国州立 K－12 教育测评中的实践与探索*

陆　宏　高佳佳　胡一平

摘要　基于项目反应理论的计算机自适应测验有着诸多优势，国内却依然缺乏计算机自适应测验的实践应用。在美国，计算机自适应测验已经以不同形式存在于州立 K－12 教育测评中，虽然存在各种质疑之声，但计算机自适应测验技术依然在不断发展。同时，美国的发展历程也为国内实施计算机自适应测验提供了借鉴和启示。

关键词　计算机自适应测验；　美国州立 K－12 教育测评；　项目反应理论

作者简介　陆　宏/山东师范大学教育技术系教授　（济南 250014）
高佳佳/山东师范大学教育技术系硕士研究生　（济南 250014）
胡一平/山东师范大学教育技术系硕士研究生　（济南 250014）

自适应测验起源于 Alfred Binet[1] 开发的智力测验，1973 年 Weiss[2] 借助计算机技术实施了 Binet 的自适应测验，从而产生了计算机自适应测验（CAT）的雏形。而创立于 20 世纪 50 年代的项目反应理论（IRT）的不断丰富和完善，则打开了 CAT 迈向实际应用的大门，并且开辟了测验领域的新天地。

一、问题提出的背景

当前，国内基础教育的变革趋势已不可逆转，诸多亟待解决的问题当中，最为民众所关注的便是考试方式的变革，如将“一考定终生”转变为“提供多次考试机会”，提高考试的公平性等。在 2014 年 9 月出台的《国务院关于深化考试招生制度改革的实施意见》[3] 中明确指出：完善高中学业水平考试，创造条件为有需要的学生提供同一科目参加两次考试的机会，高考中的外语科目也提供两次考试机会；高中学业水平考试中各科目分为合格性和等级性考试。然而，要将这些合理的规定转变成可操作的行动，仅仅依靠传统的纸笔测验显然是不够的，因为纸笔测验难以确保两次考试难度的一致性，在合格性和等级性考试的编制上也缺乏可靠的依据。

基于项目反应理论的计算机自适应测验（IRT－CAT）恰恰可以弥补纸笔测

* 本文系山东省高等学校科技计划项目“基于项目反应理论的英语词汇自适应学习系统的研制”（项目编号：J13LN12）阶段性成果。

验的不足，IRT 是以概率函数的形式描述试题作答反应结果是如何受被试能力水平和试题特征联合作用的影响，[4] 其优势在于：首先，被试能力参数的估计不依赖于具体的试题，这一点使得同一科目一年多考成为可能；其次，IRT 所提供的根据不同试题对不同被试单独计算信息量的方法，成为区分合格性和等级性测验的技术保障。因此，在笔者看来，将 IRT－CAT 引入国内基础教育正是恰逢其时（本文中以下所指 CAT 均为 IRT－CAT）。

二、研究综述

（一）国外研究综述

虽然 CAT 涉及复杂的概率与数理统计知识，但由于欧美国家有着侧重量化研究的传统，因而部分心理与教育测量的专业人士对 CAT 抱有浓厚的研究兴趣。如美国伊利诺伊大学香槟校区 Chang Hua Hua 的团队长期致力于 CAT 中曝光度与内容平衡的研究，以及如何将其应用于认知诊断之中。美国明尼苏达大学的 David J. Weiss 等人还成立了国际计算机自适应测验协会（IACAT），以便研究者间交流和共享信息。另外，大量 CAT 与传统纸笔测验间的比较表明，尽管两者在对被试个别心理特质的影响上存在差异，但两者的测验结果往往呈现出很高的一致性。[5]

国外的 CAT 研究不仅理论成果丰硕，更显示出应用范围广泛的特点。特别是在美国，CAT 不仅应用于诸如研究生资格考试（GRE、GMAT）、医生护士资格考试、注册会计师考试、军事服役职业能力倾向等面向全国、甚至世界范围的大型考试，也以不同形式存在于美国州立 K－12 教育测评之中。

（二）国内研究综述

梳理国内的研究成果，可以发现以下特点：

1. 偏向完善 CAT 各技术环节的理论研究

文献中尤以控制试题曝光度和内容平衡的选题策略居多，如“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”、[6]“引入曝光因子的计算机化自适应测验选题策略”。[7] 这些研究的共同点在于被试和题库都采用蒙特卡洛模拟的方法产生，缺乏实践的检验与应用。

2. CAT 与语言测试有一定的结合

早在 1989 年，桂诗春[8] 就撰文介绍了 CAT 在欧美国家语言测试中的应用，此后，教育部考试中心就一直致力于 CAT 的研究和开发，王蕾等[9] 还提出了构建我国少儿英语远程计算机自适应测验题库的设想，刘红云等[10] 也曾开发了用于认知诊断的国家英语二级自适应测验系统，并在大连、北京等地进行了大规模施测。但总体来看，我国的计算机自适应语言测试（CALT）尚未超越介绍多实验少、思考多实践少的初始阶段。

3. 缺少 CAT 的实际应用

就技术层面而言，依据项目反应理论所涉及的各种经典算法开发常规的