

计算机化自适应测验及应用于规模化考试的主要问题分析

郑旭东¹ 高守林¹ 任友群²

- (1. 华东师范大学 教育学部教育信息技术学系, 上海 200062;
2. 华东师范大学 教育学部课程与教学研究所, 教育信息技术学系, 上海 200062)

[摘要] 计算机化自适应测验是评估学生个性化能力的有效测评方式, 广泛应用于国外的 K-12 教育、升学考试和资格认证中。以考试技术改革为突破口, 在规模化考试甚至是未来高考中实施 CAT, 是促进教学测试评估和招考改革的有效措施。为此, 本研究首先对 CAT 的技术部件、测试流程、不同类型以及与纸笔考试的异同进行梳理, 并对其在国外教育中的典型应用和存在问题进行深入分析; 继而在调研与分析国内中学教育“机考”现状及改革期望的基础上, 对未来规模化考试中实施 CAT 可能出现的社会问题、主体与资金投入、系统与题库建设、应用与保障机制等问题展开剖析, 以期为国内规模化招考方式的改革提供借鉴。

[关键词] 计算机化自适应测验; 自适应技术; 高考改革; 能力评估

[中图分类号] G434

[文献标识码] A

[文章编号] 1007-2179(2016)04-0040-10

考试招生制度是教育综合改革“牵一发而动全身”的重点领域和关键环节(袁贵仁, 2013)。2010 年颁布的《国家中长期教育改革和发展规划纲要(2010—2020 年)》^①明确提出“要改革教育质量评价制度, 完善教育质量监测评估体系, 定期发布测评结果, 建立学生发展指导制度”。《中共中央关于全面深化改革若干重大问题的决定》^②提出“对考试招生制度改革, 从根本上解决一考定终身的弊端”, 以新高考为核心的教育考试招生改革已然成为国家和社会关注的焦点。此外, 党的十八大报告^③指出“把立德树人作为教育的根本任务”, 《教育部关于全面深化课程改革落实立德树人根本任务的意见》^④也提出“加强考试招生和评价的育人导向”, 并将“协同推进教材编写、教学实施、评价方式、考试命题等环节的改革”作为立德树人的主要任务之一。国内教学评估和招生考试方式的改革, 必将影响到未来人才的能力发

展和评估选拔。

基于国内外先行的成功经验, 以考试技术改革为突破口, 能有效推进考试招生的改革(肖龙海, 2014)。值得注意的是, 俗称“美国高考”的美国学术能力测验(Scholastic Aptitude Test, 简称 SAT)从 2016 年起将采用计算机进行考试, 改变之前只用纸笔考试的方式。而计算机化自适应测验将科学的测评理念和先进的信息技术相结合, 具有差异化、个性化和自适应的测试特征, 在国内外考试招生中已被广泛应用, 如美国研究生入学考试(简称 GRE)、托福以及中国汉语水平考试(Hanyu Shuiping Kaoshi, 简称 HSK)。旨在对考生的能力水平进行个性化评估的计算机化自适应测验, 不仅可对学校教与学活动效果进行测试评估, 也是推进高考招生改革、实行一年多考和科学测评学业能力水平的有效考试方式。

[收稿日期] 2016-04-17

[修回日期] 2016-06-20

[DOI 编码] 10.13966/j.cnki.kfjyyj.2016.04.006

[基金项目] 上海高校“立德树人”人文社会科学重点研究基地——上海市信息科技教育教学研究基地。

[作者简介] 郑旭东, 华东师范大学教育信息技术学系博士生, 研究方向: 学习科学与技术设计、教育信息化(zhengxudong1102@163.com); 高守林, 华东师范大学教育信息技术学系硕士生, 研究方向: 学习科学与技术设计; 任友群, 博士, 教授, 博士生导师, 华东师范大学课程与教学研究所, 教育信息技术学系, 研究方向: 教育信息化、学习科学、教育政策。

一、计算机化自适应测验概述

(一) CAT 的技术部件与测试流程

Computerized Adaptive Testing 和 Computer Adaptive Test 的缩写形式均为 CAT, 两者同指基于计算机的个性化自适应测验, 只是用词不同。因而, 国内将 CAT 也一般翻译为计算机化自适应测验或计算机自适应测验, 且常混合使用。CAT 并没有确定的定义, 一般指以计算机为媒介的个性化测验, 目的是为每个被试构造最优的测验, 从而对被试的能力水平进行精确的测度(张华华, 2013)。

在采用 CAT 测评的过程中, 被测者考前不知道题目, 且题目是根据被测者的作答情况从巨大的题库中随机抽取, 以尽可能符合其能力水平, 所以此测试方式具有自适应性。CAT 的技术部件和测试流程如图 1 所示(Weiss et al., 1984; Thompson et al., 2007; Thompson et al., 2011)。

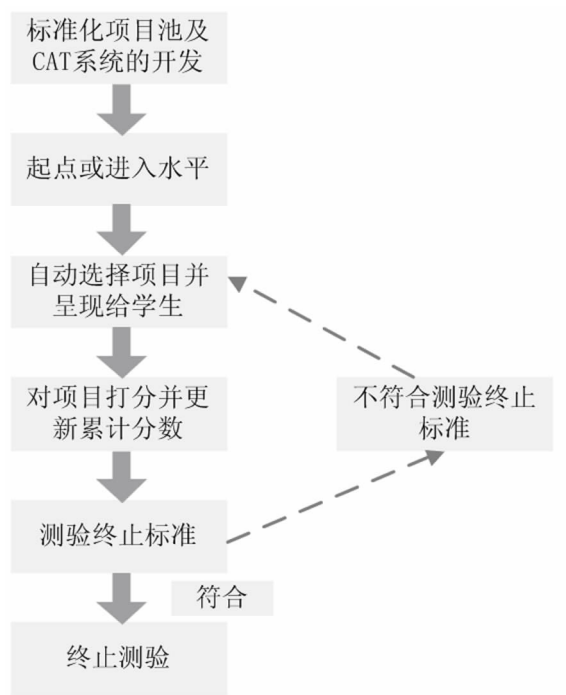


图1 CAT 技术部件与测试流程

1. 标准化项目池。标准化项目池也称标准化试题库, 是 CAT 系统的基础, 也是实施整个测试流程的基础。其中所有项目(试题)的参数必须都用测量模型标准化, 项目难度级别需要覆盖所有测试者的能力水平。

2. 起点或进入水平。CAT 开始时, 项目自动选

择基本的进入水平。有两种方式确定考生的进入水平: 1) 如果考生之前的基本信息可为系统所用, 则可较精确地确定其测验起点; 2) 通常情况下, CAT 将所有考生的平均能力水平假定为进入水平开始测验, 所以第一个项目往往为中等难度。

3. 项目选择算法。根据考生作答情况, 计算机自适应地为考生选择算法和选题策略支持, 如最大费希尔信息量、最小期望后验标准差、最大全局信息量、最大加权信息量、贝叶斯准则等方法(李铭勇等, 2010)。

4. 打分算法。考生完成一个项目后, CAT 更新对其能力水平的估计(如果项目回答正确, 评估其能力水平较高, 反之较低)。这个打分的评估过程一般通过最大似然估计和贝叶斯估计等方法完成。

5. 测试终止标准。理想的 CAT 测试终止标准精确测量出考生的能力水平。实际测验中, 终止标准取决于测验目的, 一般由测验时间决定。

(二) CAT 的类型

计算机化自适应测验有多种不同的类型, 依据测试理论和采用的组卷技术可以分为两大类。

1. 基于测试理论的分类

1) 基于项目反应理论的单维 CAT

项目反应理论(Item Response Theory, 简称 IRT) 也称为潜在特质理论或潜在特质模型, 是针对经典测量理论而提出的现代心理测量理论。单维 CAT(Unidimensional CAT, 简称 UCAT) 是以项目反应理论为基础的单维计算机化自适应测验。项目反应理论提出三个假设(DeMars, 2010): 1) 能力单维性, 即组成某个测试的所有项目都测量同一潜在特质; 2) 局部独立性, 即对某个被试而言, 项目间无相关存在; 3) 项目特征曲线, 即对被试某项目的正确反映概率与其能力之间的函数关系所作的模型。基于以上三个假设, 单维 CAT 根据被测者正确与否的反应, 不断选择与其能力水平相匹配的项目, 最终测出其在某一维度的真实能力水平。

2) 基于多维项目反应理论的多维 CAT

多维 CAT(Multidimensional CAT, 简称 MCAT) 是基于多维项目反应理论(Multidimensional Item Response Theory, 简称 MIRT) 的计算机化自适应测验。由于人在思考和完成测试任务时往往需要多种能力的配合, 加之项目反应理论是能力单维性

假设理论,所以单维 CAT 在多维能力测验时存在技术和理论上的问题。因而在项目反应理论的基础上演化出了多维项目反应理论,引入多维能力、多维项目区分度参数及多个步骤难度参数,进而模拟测验项目和被测者之间的交互作用,采用概率模型表征被测者利用多维能力水平答对特定项目的概率(Reckase,2009)。所以通过多维 CAT 可以对被测者多个维度的能力水平分别推断和测量,并进行认知诊断。

3) 基于认知诊断理论的 CD-CAT

认知诊断计算机化自适应测验(Cognitive Diagnostic CAT, 简称 CD-CAT)是认知诊断理论(Cognitive Diagnostic Theory)和 CAT 的结合。目的是通过 CAT 对被测者个体知识结构、加工技能或认知过程等属性进行认知诊断,从而了被试的知识点掌握情况,为教育教学工作提供针对性的指导,促进学生个性化发展(郭磊等,2015)。CD-CAT 的自适应体现在对被测者认知属性层面上的知识掌握水平诊断上。此外,认知诊断 CAT 比多维 CAT 提供的信息更丰富和准确。

2. 基于组卷技术视角的分类

根据采用的组卷技术,CAT 又可分为全部序列型、分组序列型和随机组卷型三种(张华华,2013)。

1) 全部序列型

采用全部序列型(Fully Sequential)组卷技术的 CAT,实质上完全体现了自适应测试的思想:除第一道试题外,被测者回答的所有题目是根据前一道题作

答情况随机抽取。对不同的被测者而言,这种全部随机抽取题目并组成的试卷也是不同的。

2) 分组序列型

分组序列型(Group Sequential)组卷技术在目前的 CAT 设计中应用十分广泛,GRE 考试就是典型的分组序列型 CAT。这种类型的 CAT 将每个难点设计为一个模块,在考前对每个难点组卷,组成高中低不同难度的题组(一组题)。测试开始时,先给被测者中等难度的题组,之后根据被测者对题组的作答情况,抽取并呈现更难或更容易的题组。分组序列型 CAT 可以在考前对每个模块先组好题,并事先让专家评价每份考卷,把控考卷质量。但是,如果每个模块的组题工作量大,让专家把控所有的考卷质量不现实。此外,人的能力是连续的,单独在几个难点上组卷的设计也不合理。

3) 随机组卷型

随机组卷型 CAT 采用的技术结合了全部序列和分组序列两种技术的优势。其设计思路是:测试开始时使用题组,当被试的能力水平估计出来后,用全部序列技术为下一阶段测选题组卷(题组形式)。根据这一阶段测出的能力水平,再进行更下一阶段的选题组卷。

(三) CAT 与传统纸笔考试比较

目前实施的 CAT 中,测试时间、测试内容、问题形式等与传统纸笔考试相同(Virginia Department of Education,2012),只是两种考试的测试目的、学生回答的问题及顺序等差异较大(见表一)。

表一 CAT 与传统纸笔考试比较

比较维度	相同	不同	解 释
测试目的		✓	CAT 针对不同学生的能力水平进行个性化准确评估;纸笔考试依据所有学生的考试分数进行排名。
问题形式	✓		CAT 与纸笔考试的问题形式基本相同,如选择题、填空题等。
学生回答试题数量	✓		理论上,不同能力水平的学生在 CAT 中答题数量不同,但在实际考试中与纸笔考试一样,每个学生答题数量相同。
学生回答的问题		✓	CAT 题目是按学生的能力水平从题库中随机抽取,故不同学生作答题目不同;在纸笔考试中,所有学生作答的问题相同。
答题顺序		✓	CAT 中学生需按照假设顺序作答,且不能回看修改,若连续稳定答对试题,获得高分的机会也会增加;纸笔考试可任意顺序答题并修改。
测试内容	✓		CAT 与纸笔考试内容均按课程标准或考纲设计,基本相同。
完成测试的时间	✓		在 CAT 与纸笔考试中,学生完成测试的时间相同
成绩量化		✓	CAT 的成绩由答题正确率和难度共同决定,测量出能力水平等级;纸笔考试成绩仅由得分决定,难以测出能力水平等级。

二、CAT 在国外教育中的应用与存在问题

放眼全球,美国对 CAT 的研究最领先,因而其推广应用也较广泛和深入。从 1994 年 4 月起,美国注册护士执照考试就从传统的纸笔方式转变为 CAT 方式。CAT 也应用于美国注册管理会计师考试、军事服役职业能力倾向等资格考试。

CAT 在教育领域也应用广泛,已较大规模地应用于美国 K-12 教育测评。多州评价联盟测评如智慧平衡评价联盟(Smarter Balanced Assessment Consortium,简称 SBAC),单个州测评如犹他州的学生成长与卓越评价(Student Assessment of Growth and Excellence)等都采用 CAT 方式。教育测评机构和互联网教育企业也纷纷投入到 CAT 领域,为自适应教/学、测评等提供支持服务,如美国最大的教育测评机构西北测评协会(Northwest Evaluation Association)提供基于 CAT 的学业进展测评(Measures of Academic Progress)。在线教育企业 Knewton (<https://www.knewton.com/>)的业务核心目标是测量学习者对知识点的掌握程度,并提供预测性分析及个性化推荐,适应性地满足学习者个性化需求。

欧洲的研究机构、教育企业和组织机构也积极探索 CAT 的研究和教育教学测评实践。如英国剑桥大学心理测试中心开发了开源测试平台 Concerto (<http://www.psychometrics.cam.ac.uk/newconcerto/options>),用户通过 Concerto 可以创建多种在线评估,包括从简单的调查到复杂的基于项目反应理论的 CAT 在线评估。

(一)CAT 应用典型案例

智慧平衡评价联盟是美国多州共建的合作共同体。它源于 2009 年奥巴马政府宣布的“竞争卓越”(Race to the Top)教育改革计划,并投入 43.5 亿美元用于联邦和各州的教育改革。该联盟建立旨在多州间合作开发基于美国《共同核心标准》(Common Core State Standards,简称 CCSS)的一流考试评估系统—智慧平衡评价系统(Smarter Balanced Assessment System,简称 SBAS)(见图 2)。该系统测评的决策、标准、部件、部署、支持和实施,都是为了改善教师教学和提高学生学习,让所有学生为大学和就业做准备。

智慧平衡评价系统的主要功能是向州联盟的每位学生提供 CAT 服务:对 3~8 年级和 11 年级学生的英语语言艺术/素养和数学能力进行测试,包括总结性评价、可选择的临时评价和形成性评价三种,及一个安全在线报告系统。

总结性评价在每学年的最后 12 个星期进行,测评包括两部分:CAT 和在计算机上进行但非自适应的实际操作表现测评。总结性评价旨在准确测评学生在学习和成绩方面的进步,并作为对学校、地区、州问责制度中的部分评估项目;同时,也为学生提供了大学及职业准备所需知识与技能的指导方向和进步措施。

可选择的临时评价(Optional interim assessments)的实施时间可以自由选择。该评价用于向教育者提供学生全年的进步信息。与总结性评价类似,临时评价也由 CAT 和实际操作表现两部分组

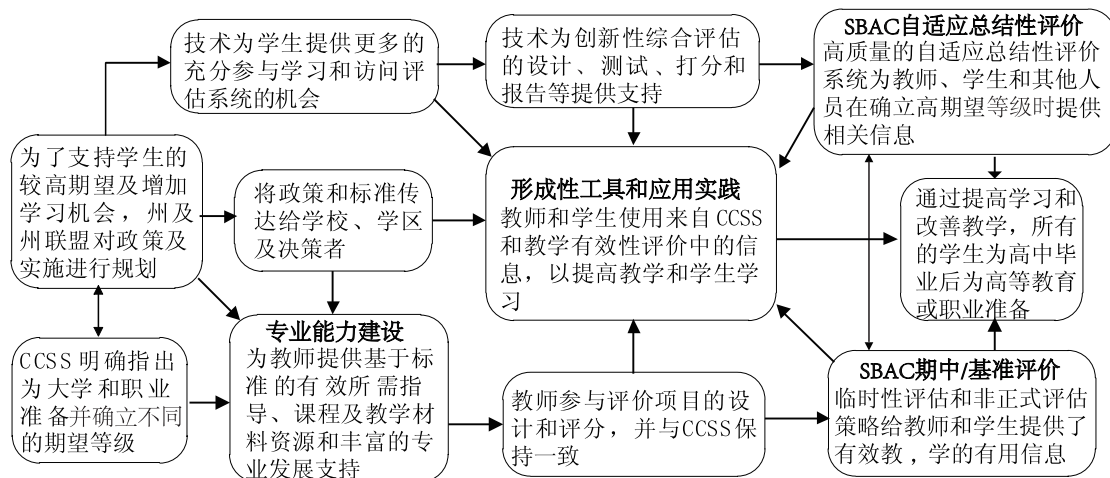


图2 智慧平衡评价联盟及测评实施概要

成。它将帮助教师、学生和家长了解学生的学习是否步入正轨,了解学生在共同核心标准下的优势和不足,也有助于教师教学、专业能力提升及课程考试的发展。

形成性评价的实施是智慧平衡评价系统数字资源库中,遵循《共同核心标准》和智慧平衡评价联盟评价目标的专业发展材料、资源和工具的基础。基于研究的教学工具将按需满足和帮助教学解决学习中的挑战和不同的教学问题。该数字资源库还包含了与评价系统相关的所有组件的发展材料,如实际操作表现评分规则。

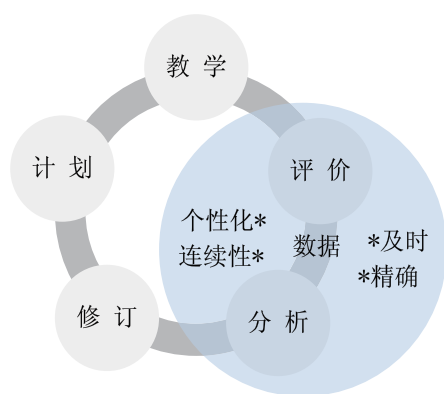


图3 智慧平衡评价系统学习分析应用策略模型

值得注意的是,运用智慧平衡评价系统对学生进行评价不是最终目的。对评价数据信息进行深度挖掘和分析,改进教学活动方式,提升教与学的效果以及为大学和职业做准备,才是开发和应用该系统的最终目标。因此,如图3所示,这一系统还提供了学习分析功能,可以对测评的实时数据进行精确、个性化和连续性的挖掘与分析,为针对性的实施教学干预、修订教学计划和改变教学方式提供有用的参考信息。

(二)存在问题

相较于传统纸笔考试而言,CAT有诸多优势,但随着CAT在国外教育教学和测试中的广泛应用,其存在问题也逐步显现。

1. 项目曝光率不均衡。在计算机化自适应测验中,区分度大的项目往往被随机抽取的概率和频率也高,区分度小的项目在测验中被抽取的可能性较低,这就造成了项目曝光率不均衡。项目曝光率一般通过选题策略进行控制,如随机化方法、曝光率参数控制、分层选题策略等。但测量精度和项目曝

光控制又往往相互制约(毛秀珍等,2013),提升项目曝光率的均衡性,会造成测验精度的降低。

2. 题库安全问题。CAT题库的安全问题涉及测试的有效性和公平性。CAT测验人数庞大,高区分度的项目不断曝光,题库的安全性会降低。所以实施大规模的CAT要求试题库足够大。而建设巨大的试题库或多个平行题库提升安全,又会带来巨大的财力和物力消耗,教育部门难以承担。另外,受选题算法和测验精度的限制,以扩大项目曝光率提高题库的利用率和安全性的方式也有技术难题。

3. 测量的能力结构简单。CAT目前以较为传统和易测量的能力测量为主,难以对被测者高阶、难以测量的能力结构(Hard-to-Measure Constructs),如视觉空间处理、协作、好奇、创造、问题解决等能力进行测量和评估。另外,被测者使用的能力并非是某一方面的能力,而是综合使用了多种能力。因此,测量结果并非完全准确,会存在偏差。

4. 设计理念问题。以适应性方式对被测者的能力进行准确评估是CAT秉持的基本理念。但在CAT题库的设计、开发和测验使用中,为了追求适应性效果和测量学生的真实能力,测验容易脱离一定原则和框架的约束,如题库应按照课程标准设计。美国爱达荷州在CAT的最初实践中,就因一味追求自适应,测验系统偏离课程标准,遭遇失败(REL West at WestEd,2008)。

三、规模化实施“机考”的现状与改革期望

在普通中学教育中,大规模使用计算机进行考试的只有信息技术科目。高中信息技术学业水平考试也是国内中学教育较早大规模采用“机考”考试之一,部分省市开展的“机考”已有十余年。但是,这种“机考”不具有自适应特征,考试也存在诸多问题,考试方式改革诉求十分迫切。本文以高中信息技术学业水平考试为例,对国内各省市自治区省级信息技术教研员进行访谈和问卷调查,了解目前中学教育中规模化实施“机考”的基本现状,以及未来考试改革中采用CAT的期望与可行性。

(一)调查设计与实施

本次调查包括四方面内容:我国大陆31个省市自治区的高中信息技术学业水平考试方式;不同考

试方式的现状与问题;各省市自治区改革高中信息技术学业水平考试方式的未来走向;各省级教研员实施 CAT 的态度。

调查对象是我国大陆 4 个直辖市、5 个自治区和 22 个省份的省级信息技术教研员,共计 31 名。调查时间为 2016 年 3~4 月,期间通过网络向 31 个省、市、自治区的 31 名教研员发放问卷 31 份。由于 7 个省市自治区未统一或实施高中信息技术学业水平考试,所以回收问卷 24 份,有效问卷 24 份。调查对象中,女性 11 人,男性 13 人;年龄最大者 58 岁,最小者 30 岁,平均年龄 50 岁。从调查对象年龄看,各省市自治区的省级信息技术教研员在“机考”研究和实施方面经验和专业知识丰富。此外,调查还随机选取 5 名教研员,通过现场、邮件和微信三种形式进行访谈。

(二) 调查结论

1. 高中信息技术学业水平考试方式

调查发现,17 个省市自治区的高中信息技术学业水平考试采用传统“机考”方式,占 70.8%;7 个省市自治区采用纸笔考试,尚无省市自治区采用 CAT 方式。

2. 不同考试方式的现状与问题

在实施“机考”方面,17 个省市自治区中,88.2% 的省级信息技术教研员认为,“计算机化自动测评相对便捷,是‘机考’的最大优势”。88.2% 的教研员认为,难以真实测量出考生的能力水平是目前“机考”的最大问题。此外,76.4% 的教研员认为“机考”不能避免应试“刷题”和“题海战术”,这也是教研员访谈中普遍反映的问题。由于“机考”是非个性化和非适应性考试,加之题库容量小和题目曝光控制技术落后,通过题库“刷题”早已成为考生准备考试的主要方式。

在实施纸笔考试的 7 个省市自治区中,85.7% 的省级信息技术教研员认为,本省市自治区的高中信息技术学业水平考试至今采用纸笔考试的最主要原因是考试政策。接受访谈的教研员均认为,信息技术课应注重学生的基本操作能力,而纸笔考试难以对考生的操作技能进行测评,且纸笔考试的测评效率低。这也是 7 个省市自治区教研员认为纸笔考试有必要为基于计算机的考试取代的关键原因(见图 4)。

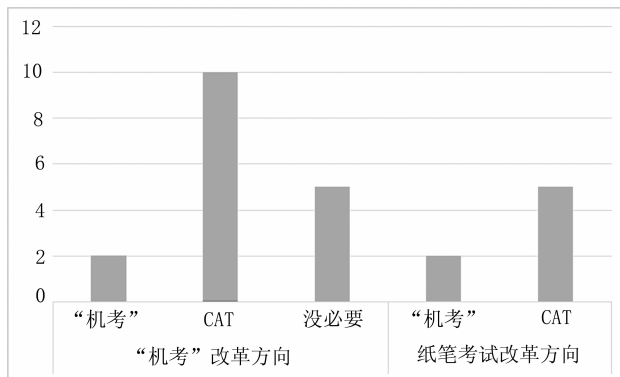


图4 “机考”和纸笔考试改革

3. 考试方式的改革期望与采用 CAT 的态度

调查发现,除高中信息技术学业水平考试采用“机考”和纸笔考试存在诸多问题外,各省市自治区的省级信息技术教研员均不同程度地表达了对现行考试方式的改革期望(见图 4)。在“机考”改革方面,来自 17 个“机考”省市自治区的 10 名省级教研员期望未来采用 CAT;来自 7 个纸笔考试省市自治区的 7 名教研员,全部期望未来的信息技术学业水平考试不再使用纸笔考试,而认为应采用 CAT 的有 5 人。

综上,期望未来改革两种考试并采用 CAT 的教研员为 15 人,占 62.5%。这表明高中信息技术学业水平考试改革方面,多数省级研究员对 CAT 态度积极。问卷调查和访谈还发现,教研员对考试方式向 CAT 转变的重要原因在于,“CAT 能准确评估考生的能力水平,并可为升学及发展提供真实性参考”,说明省级信息技术教研员对 CAT 表现出较高的认知。

但是,在更大规模考试(如高考)中实施 CAT 的可行性方面,多数教研员谨慎(见图 5)。41.6% 的省级教研员认为高考采用 CAT 可行,认为不可行和不知道的各占 29.2%。国内高考对考生的影响以及考生和社会对高考的重视程度远远高于其他考试,这也是多数教研员对高考采用 CAT 态度谨慎的关键原因。值得注意的是,俗称“美国高考”的 SAT 将在今年采用计算机考试。

日本也将对现行的“大学入学中心考试”(即日本高考)进行改革,从 2020 年起采用基于计算机的考试(日本文部科学省,2015),即“机考”。美日对高考方式改革取向,对国内高考等规模化考试采用

和实施 CAT 有借鉴意义。

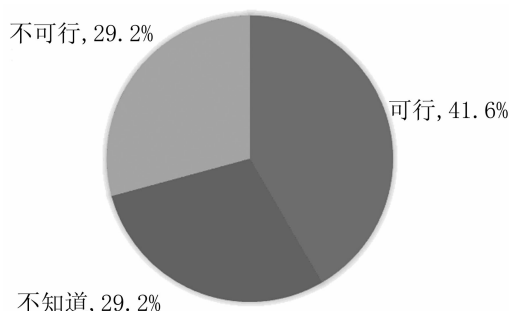


图5 未来高考中实施 CAT 的可行性

四、国内规模化考试实施 CAT 主要问题

随着信息技术的快速发展和测评算法模型的不断突破, CAT 在不久的将来必成为主要的教学诊断和招生考试方式。我国目前正在全面推进招生考试改革, 计划到 2020 年基本建立中国特色现代教育考试招生制度。但若未来的日常教学诊断、规模化考试评估(如高中信息技术学业水平考试), 甚至是“新高考”中实施 CAT, 也必将面临以下问题。

(一) 对 CAT 认知偏差引发的不良社会影响

国内现行的考试制度多年来深受考生、家长、学校和社会诟病, 考生的“高分低能”和综合能力素质不高早已为社会和国家关注。尽管 CAT 是从考生的能力水平角度进行评估, 能较大程度地避免应试带来的问题。但采用 CAT 也可能在社会上带来如下三方面问题:

1. “机考”和 CAT 的认知混淆

“机考”在规模化考试中已较常见, 如高中信息技术学业水平考试普遍采用“机考”。尽管 CAT 也采用计算机进行考试评估, 但与传统“机考”不同的是, 其具有自适应和个性化评估的特征, 对每个考试者的能力进行评估。而“机考”是数字化纸笔考试, 难以对不同考试者的能力水平进行测度, 且“机考”中的“刷题”等题海战术不适用于 CAT。

此外, 目前采用的“机考”并未应用于语文、数学和外语等所谓的“主科”中, 中国上千年来采用的纸笔考试仍是重大考试中采用的方式。在所有考试科目中, 不分科目的主次与轻重而采用 CAT, 必须突破考生和家长对采用 CAT 会带来招考不公平和不公正的认识, 必须跨过社会大众对 CAT 的偏见。

2. CAT 组卷方式可能带来的问题

在计算机化自适应测验中, 除了第一道(组)试题外, 考生作答的题目均是计算机依据上一道(组)题的对错随机抽题和组卷, 所以每个考生的作答试题及难度不一。加之考生不能对已作答考题进行查改, 习惯了使用统一试卷和反复查改试题答案的考生, CAT 的这种组卷和作答方式可能会给他们以“不公平”和“不平等”的借口。

3. 计算机及应用带来的问题

尽管 CAT 对考生的计算机使用能力几乎没有要求, 但误操作、电脑死机和网络传输等问题会影响考生的心理和成绩, 降低测评准确性; 计算机性能和新旧差异也会引发考生和家长对考试公平和公正的质疑。

传统的纸笔考试不会发生上述问题, 采用 CAT 不仅可能会出现以上问题, 甚至更多。当这些问题不断发酵, 考生和家长对考试的公平和平等产生质疑时, 其不良影响会为教育行政和招考部门带来压力。

(二) 主体与资金的投入问题

在美国, 考试像其他诸多领域一样完全市场化, 许多知名企业和组织(如 Pearson、OAT、NWEA)投入大量的人员、资金和技术研发和提供 CAT 服务。美国的 K-12 教育测评一般也以州/区联盟或单个州的形式购买 CAT 服务, 并在其下辖的学校中使用。如果州/地区对 CAT 及相关服务不满意或是其他原因, 也可退出联盟或停止购买测评服务。上文提及的智慧平衡评价联盟就是最初得到了美国联邦“竞争卓越”项目和“公共教学主管部门”的资助。

另一种形式是单个州向企业提供资金, 建设属于本州的计算机化自适应系统。如犹他州通过州议会批准经费和招标获得企业支持, 最终建成该州的 CAT 系统——SAGE 题库项目的建设依照犹他州课程标准, 全州教育工作者编制或选择其他州的试题, 最终通过犹他州教育办公室、全州教师, 甚至是家长的力量来完成(Kate, 2015)。

我国体制内各种大型考试, 如高考一般由政府部门主导, 很少受市场化影响。在大规模甚至是全国性考试测评中, 采用 CAT 必然需要一个包含大量项目的庞大题库和先进的自适应系统, 其中涉及各种复杂测量模型和先进的规则算法等技术, 并需要

对大量项目进行编制和质量把控。如此庞大和复杂的系统性工程建设应由谁主导? 建设主体又是谁? 各个环节所需的大量资金应当由谁投入?

以高考为例,它由教育部统一组织和调度,未来高考中所采用的 CAT 系统理应也必须由教育部统筹主导建设。但先进的 CAT 系统和题库并非是一两个部门就能建成的,应借鉴美国 CAT 建设和应用的成功经验,引入和联合相关测评企业、高校、科研机构、测评机构及教育质量监测中心,形成专业的研究、设计、开发、运维、测评、监控和服务团队共同开发。CAT 系统和题库项目的初期建设阶段,资金需求很大,主要应由政府财政投入,次之则应为企业以赞助支持等形式入资,企业可在后续的测评服务中获得一定比例的盈利。

(三) 系统与题库的建设问题

国内高考目前涉及省自主命题、全国统一命题和教育部考试中心命题(海南省)三种,未来高考究竟需要一个 CAT 系统还是多个? 如何建设 CAT 系统与题库?

全国统一命题考试,可以借鉴美国州联盟的做法,参加全国统一命题的省市联合建设和使用同一个 CAT。而对于自主命题考试,经济和教育发达省份可独立建设本省的 CAT,甚至可通过租赁方式供其他地区有偿使用。美国犹他州的 SAGE 也是类似做法,且有三个州已与犹他州达成协议,以租赁方式购买和使用 SAGE 的测评服务(Wood, 2014),其中佛罗里达州以 540 万美元购买 SAGE 的年终测试服务(Burton, 2014)。其实,以上 CAT 系统的建设和应用思路在我国教育信息化建设中早已采用,如教育信息化资源通过“自主建设”“共建共享”“公建众享”三种建设模式(陈琳, 2012),迅速提升了数字化学习资源在校际、区域和全国的共享,扩大了优质教育资源的覆盖面。

进一步而言,国内目前是否有先进的测评技术和丰富的经验?

计算机测评技术在我国大型考试中的应用不乏先例。2012 年,湖北省为了配合英语高中新课程改革,全省 87 所高中的 7253 名高考生参加了高三英语网络调考(赵莉, 2012)。这种所谓的“机考”尽管并非是自适应的,但也不失为未来计算机化自适应测验的路演和尝试。值得注意的是,计算机化自适

应测验技术在国内大型考试中也有应用,如国家汉办面向全球考生举办的中国汉语水平考试;大连三万名学生在 2011 年就以在线 CAT 的方式,参加了三天的持续性评估(Liu et al., 2013)。并且国内学者和测评机构在 CAT 的项目曝光控制算法、选题策略、评分规则和终止规则等核心技术层面有着深入的研究(郭磊等, 2015;喻晓锋等, 2015;汪文义等, 2015;毛秀珍等, 2015)。可见,尽管 CAT 开发技术在国内应用缓慢,但也在不断地尝试和发展,积累了一定的技术和经验。

(四) 应用与保障的机制问题

高考目前已不是一项形式简单的考试,实际上动员和使用了政府部门、教育系统及社会许多类型的资源,以尽可能保障这种全国性考试的诚信、公平、公正和安全。如果高考采用 CAT,如何保障考试的诚信与安全? 我国是否具备制定和实施保障机制的经验?

实际上, CAT 相较于传统纸笔形式的考试更加安全和公平。一方面, CAT 拥有相当庞大的题库,且试题是根据考生的当下能力随机抽取,与纸质试卷相比在源头上杜绝了泄题,提升了考试的安全性。另一方面, CAT 是对学生达到的能力水平进行评估,备考常见的“题海战术”在这种考试中不可取,也行不通。这将有利于促进基础教育教学制度和考试制度的改革。

除中国汉语水平考试采用 CAT 之外,我国教育部考试中心从 2006 年开始就负责和实施中国大陆地区的基于互联网传输的托福和 GRE 考试,包括网上报名、考点管理、考务人员培训和公布考试成绩等环节。所以,无论是 CAT 考务人员的培训,还是实施环节和相关服务的保障,国内都有制定和实施相关保障制度的经验。且经过几十年纸质高考的累积经验、可行制度及部门间的合作机制,也为 CAT 高考保障机制的制定和实施提供了借鉴。另外, CAT 是先进的评价理念、测评算法和信息技术结合,其实施和保障的各种制度机制也必须经过实践不断验证、修订和完善。

五、结语

CAT 是信息技术在教育教学和考试评估中的创新应用。无论是国际考试测评的趋势,还是我国

招生考试制度的改革,科学的测评理念和信息技术相结合的考试评估方式,必将成为未来包括高考在内的各类考试评估的首要选择。尽管 CAT 在我国规模化的实施会存在诸多问题,但关键是如何迈出通过技术改变测评目的和方式的第一步。

信息技术课程恰为国内学校教育中率先实施 CAT 提供了“试验田”,主要体现于以下三方面:1)着重培养学生技术能力和计算思维的高中信息技术课程新标准即将完成修订(任友群等,2016),实施与其更为适应的 CAT 测评方式具有必要性。2)国内多数省市自治区的高中信息技术学业水平考试采用“机考”,为同样是基于计算机的 CAT 测评方式提供了可行性。3)多数地区(如上海)的信息技术课程还未成为“新高考”考试科目,改革测评方式的阻力较小,为实施 CAT 测评提供可能。

信息技术课程在率先实施 CAT 的过程中,前期的探索性应用应注意以下问题:1)不可偏离信息技术课程标准。CAT 题库的设计和题目的动态“进化”应紧紧围绕信息技术课程标准,而非一味追求适应学生的水平。2)CAT 应充分应用于教与学的形成性评价。不仅信息技术学业水平考试可实施 CAT,信息技术教与学的形成性评价中更要应用,即对学生在教与学活动中产生的海量数据信息进行挖掘和分析,提供适应性、发展性和针对性的干预与建议。3)与其他评价方式的灵活结合。信息技术课程不仅是一门数值计算和工具操作的课程,更是一门以培养学生创造能力和计算思维的课程(任友群等,2016)。故面向基于任务、作品和表现的评价(如简单 App 的开发)应灵活采用学生互评和自评等方式,而不盲目采用 CAT。

总之,有理由相信,信息技术对国内基础教育的影响不会仅停留在教与学的表面,而会更深层次地渗透、影响和改变各类考试的目的与方式。以能力发展和综合素质为标准的人才选拔将因此更公平公正,对人性与科学发展和现代教育考试招生制度的建立将起巨大的推动作用,也必将极大地推进我国教育领域的综合改革和立德树人这一教育根本任务的实现。

[注释]

①国家中长期教育改革和发展规划纲要(2010-2020年). ht-

tp://www. moe. edu. cn/publicfiles/business/htmlfiles/moe/moe_838/201008/93704. html.

②中共中央关于全面深化改革若干重大问题的决定. http://news. xinhuanet. com/2013-11/15/c_118164235. htm.

③十八大报告(全文). http://www. xj. xinhuanet. com/2012-11/19/c_113722546. htm.

④教育部关于全面深化课程改革落实立德树人根本任务的意见. http://www. moe. edu. cn/publicfiles/business/htmlfiles/moe/s7054/201404/167226. html.

[参考文献]

[1] Burton, D. (2014). Utah rents year-end test to Florida for \$5. 4M [EB/OL]. [2015-12-20]. http://utahpolicohub. com/usoe-negotiated-5-4m-sage-contract-with-florida-without-board-knowledge-or-legal-counsel-review/.

[2] 陈琳,王鑫,李凡,蒋艳红,陈耀华(2012). 创建数字化学习资源共建众享模式研究[J]. 中国电化教育,(1):73-77.

[3] DeMars, C. (2010). Item response theory[M]. USA: Oxford University Press:38-60.

[4] 郭磊,郑焱金,边玉芳(2015). 变长 CD-CAT 中的曝光控制与终止规则[J]. 心理学报,(1):129-140.

[5] Kate, W. R. (2015). A brief History of SAGE assessments in Utah [EB/OL]. [2015-12-20]. http://www. utahparentsineducation. com/a-brief-history-of-sage-assessments-in-utah/.

[6] 李铭勇,张敏强,简小珠(2010). 计算机自适应测验中测验安全控制方法评述[J]. 心理科学进展,(8):1339-1348.

[7] Liu, H. Y., You, X. F., Wang, W. Y. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English Achievement Test in China [J]. Journal of Classification, (2): 152-172.

[8] 毛秀珍,辛涛(2013). 认知诊断 CAT 中项目曝光控制方法的比较[J]. 心理学报,(6):694-703.

[9] 毛秀珍,辛涛(2015). 多维计算机化自适应测验:模型、技术 and 方法[J]. 心理科学进展,(5):907-918.

[10] Reckase, M. (2009). Multidimensional item response theory [M]. New York: Springer: 65-66.

[11] REL West at WestEd(2008) [EB/OL]. [2016-3-5]. http://relwest-archive. wested. org/system/memo_questions/11/attachments/original/Computer_20adaptive_20testing_20June_202008_1_. pdf.

[12] 日本文部科学省(2015). 新大学入試、コンピュータ一面面で出題解 [EB/OL]. [2016-3-5]. http://www. yomiuri. co. jp/teen/news/20150716-OYT8T50036. html.

[13] 任友群,隋丰蔚,李锋(2016). 数字土著何以可能?——也谈计算思维进入中小学信息技术教育的必要性和可能性[J]. 中国电化教育,(1):1-8.

[14] Thompson, N. A., & Prometric, T. (2007). A practitioner's guide for variable-length computerized classification testing[J]. Practical Assessment Research & Evaluation, (1): 1-13.

[15] Thompson, N. A., & Weiss, D. J. (2011). A framework

for the development of computerized adaptive tests[J]. *Practical Assessment, Research & Evaluation*, (1): 1-9.

[16] Virginia Department of Education (2012). Comparison of a Computer Adaptive Test and a traditional test[EB/OL]. [2015-12-5]. http://www.doe.virginia.gov/testing/test_administration/cat/comparison_cat_traditional.pdf.

[17] Wood, B. (2014). Cashing in: Three states to pay \$10M for rights to study Utah's SAGE tests[EB/OL]. [2015-12-5]. <http://www.slttrib.com/news/1912812-155/three-states-paying-10m-for-rights>.

[18] 汪文义,丁树良,宋丽红(2015). 认知诊断中基于条件期望的距离判别方法[J]. *心理学报*, (12): 1499-1510.

[19] Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems[J]. *Journal of Educational Measurement*, (4): 361-375

[20] 肖龙海(2014). 网络高考:技术创新推进高考改革[J]. *教育研究*, (2): 123-128.

[21] 袁贵仁(2013). 深化教育领域综合改革[EB/OL]. [2015-12-26]. http://www.moe.gov.cn/publicfiles/business/htmlfiles/moe/moe_176/201311/159604.html.

[22] 喻晓锋,罗照盛,高椿雷,李喻骏,王睿,王钰彤(2015). 使用似然比 D~2 统计量的题目属性定义方法[J]. *心理学报*, (3): 417-426.

[23] 张华华(2013). 计算机化考试与中国教育评估[J]. *心理学探新*, (5): 387-391.

[24] 赵莉(2012). 七千高中生尝鲜英语机考[EB/OL]. [2015-12-26]. <http://edu.people.com.cn/GB/116076/120067/17501690.html>.

(编辑:李学书)

Computerized Adaptive Testing: A Review and Main Problems in Large-Scale Testing

ZHENG Xudong¹, GAO Shoulin¹ & REN Youqun²

(1. Department of Education Information Technology, East China Normal University, Shanghai 200062, China;

2. The Institute of Curriculum and Instruction, Department of Education Information Technology, East China Normal University, Shanghai 200062, China)

Abstract: Computerized adaptive testing (CAT) is an evaluation method to assess students' individual ability and is widely applied in K-12 education, Entrance Examination and Certification Exam in foreign countries. If the reformation of the examination technology can be taken as a breakthrough for examination reform, CAT will be an effective way to promote the assessment and enrollment reform for a large-scale testing (e.g. the new College Entrance Examination). This paper explored the technical components, testing procedure and various classes of CAT, and analyzed its typical applications and problems in foreign education. And then the CBT(Computer Based Testing) conditions and reform expectations to CAT in China middle school were investigated and analyzed. Based on the above analyses, five problems of CAT when applied in large-scale examination were further analyzed, such as potential negative social impact caused by CAT, the problem of construction bodies, funding and investment, testing pool construction, and guarantee mechanism.

Keywords: computerized adaptive testing; adaptive technology; college entrance examination reform; capability evaluation