文章编号:1000-5862(2016)05-0441-15

“互联网 +”测评:自适应学习之路

张华华[1,2],汪文义[3*]

(1. 伊利诺伊大学香槟分校心理系,伊利诺伊州 香槟 61820;2. 华东师范大学教育学部,上海 200063;
3. 江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:在“互联网 +”催生 11 项重要领域的顶层设计背景下,介绍传统测评模式,如纸笔测试和以考试中心提供测验的测评模式之后,主要从学习的角度,率先提出“互联网 +”测评及其框架和设计思路. 简要介绍“互联网 +”的概念,概述最新的测评理论、方法和技术,探讨如何把云计算和大数据等技术与测量理论、方法和技术深度融合问题及如何解决测评中急需解决的问题,促进测评的全面发展和大规模应用,使之更好地为自适应学习提供服务,全面提升教与学的质量,促进学习者的发展.

关键词 “互联网 +”;测评;自适应测验;自适应学习;教育质量监测

中图分类号:B 841.7 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2016.05.01

0 传统测评

教育测量与评价(measurement and evaluation)主要是研究对教育现象进行测量和价值判断的理论、方法和技术,是当今现代教育科学研究的三大领域之一. 它不仅在教育科学体系中占有重要的地位,而且在教育教学过程中具有重要而广泛的应用价值. 教育测量与评价更是教育改革与发展的风向标和助推器,它贯穿、引领和改造教育实践[1-2]. 教育测量是依据一定的法则,对教育活动中某一现象给予数量化的描述. 教育评价是根据一定的教育目标,运用可行的科学手段,通过系统地收集信息资料和分析整理,对教育活动、教育过程和教育结果进行价值判断,从而为教育决策提供依据的过程. 教育测量是为了取得数据,教育评价是分析和解释测量数据,对教育价值进行判断. 教育测量是教育评价的基础,并且通过教育评价才能获得实际意义. 本文将测量与评价简称为测评.

在当代教育体系中,教育测评几乎涉及所有核心教学环节. 简言之,如何进行教学管理乃至教育政策制定,如何判断课程体系是否符合课程标准、教学效果是否达到目标,如何了解学生能力水平、掌握了什么知识或技能,如何让老师更了解学生、因材施教,引导学生个性化学习,如何指导学生录取工作等,这些都是教育测评的最直接应用. 没有教育测评,这些问题都无法解决[3]. 测评不仅有教育测评(如能力测评、诊断测评),还有心理测评(如人格测评、兴趣测评)、语言测评(如汉语测评、英语测评)、人才测评(如胜任力测评)、医学测评(如计算机模拟病例考试)等.

绝大多数测评项目或考试,如中考、高考、研究生考试、公务员考试、英语四六级考试等高风险考试,仍采用纸笔考试. 高风险考试是指考试分数会对个人或集体造成重大影响的考试,像证书、资格、毕业、升学考试都可以归类为高风险考试. 在一次考试中,除少数考试为了防止作弊、试卷备用或提高抽样效率会使用 AB 卷和题册,且多份试卷多为平行卷(难度、考试、题型等基本一致)之外,许多纸笔考试只使用一张相同试卷. 纸笔考试具有悠久历史,符合人们传统观念,只需试卷和场所,并且考生具备读写能力就可以完成考试. 纸笔考试具有一定的不足:1)经济成本相对较高,如需要多次印刷、分发、运输、保管和批改试卷等;2)通常不能即时报告分数,

收稿日期:2016-06-30

基金项目:中国国家汉办 HSK 研究项目,美国国家科学基金(NSF-DRL1252389),国家自然科学基金(31500909,31360237,31160203,30860084),国家留学基金(201509470001)和教育部人文社会科学研究青年基金(13YJC880060)资助项目.

作者简介:张华华(1953-),男,上海出生,教授,博士生导师,长江学者,主要从事计算机化自适应测验研究.

通信作者:汪文义(1983-),男,湖南衡山人,副教授,博士,主要从事教育测量和信息处理研究.

影响考生的体验,多数不具备辅助学习功能,测验结果对教学的帮助十分小;3)纸笔考试时间相对较长,易产生疲劳效应;4)纸笔考试效率较低,用一张试卷去测试所有学生,这可能对一些学生显得太简单,而对另外一些学生又太难,这样就较难测准学生真实水平,而不确定真实水平就较难为学生提供个性化的指导,尤其是获得零分的差生,将对这个学生一无所知;5)纸笔考试手段上落后,无法呈现形象而具体的真实测验情境,测试内容静态,更多强调陈述性知识的回忆等[4];6)考试安全性较难保证,中国每年都在考试安全上花很大力气,如入闱式命题,有的考试保密室设三道铁门,可还是会出问题. 网上和报纸常报导,如高考、研究生入学考试、国家一级建造师执业资格考试泄题. 国外同样存在考试作弊的情况,比如日本"高考"和美国"高考"作弊或泄题.《华盛顿邮报》2016 年 6 月 10 日报导,被称为美国"高考"的主办方 ACT 公司(American College Test Inc.),距离开考仅剩几小时,突然宣布取消原定 2016 年 6 月 11 日在韩国、中国香港两地考点举行的考试,理由是泄题. 不少中国学生远赴韩国和中国香港地区参加 ACT 考试,却在考试前几个小时,才得知此消息. 中国已经花大力气整治考试作弊,见:2016 年 6 月 1 日开始实行的新修订的《中华人民共和国教育法》第七十九条、第八十条和第八十一条;在全国人大常委会通过的刑法(九)修订案中,已经明确考试作弊入刑. 尽管如此,依然有人作弊,而且从考场内作弊到考场外作弊,从个人"夹带"的原始作弊方式发展到集团方式的高科技作弊,比如携带微型无线耳机、信号接收器、具备通信功能的智能手表等高科技"作弊神器"等.

随着个人电脑普及,许多考试通过计算机实施(机考),以弥补纸笔测验的不足. Thomson Prometric 公司(普尔文公司)是全球最大的计算机化认证考试服务公司,20 多年来为世界知名公司提供了许多成功的测验开发和实施解决方案,方便雇主聘用最胜任的人才,为才能展示提供更好的机会. 2007 年 10 月,美国教育考试中心(ETS)收购 Prometric 公司. Prometric 公司拥有数千员工,平均每年实施 9 百万次考试,涉及学分、职前、驾照、餐饮、金融、医疗、IT 认证等考试. 在测验开发方面,提供任务分析、学科专家征募、项目开发、题库系统、测验设计、测验发布、标准设定、心理计量分析等服务. 提供的 MyItemWriter 可用于建立、编辑、审阅、上传测验项目到题库,支持开发单选题、多选题等,仅支持十分简单的图文编辑功能. 在测验实施方面,全球分布的考试中心(8 000 个考试中心分布在 160 多个国家)提供各种形式的测验,如纸笔测验、计算机化测验等,题型涉及可能含音频和视频的选择题、情景判断题、案例分析等. 还在测验安全性控制(如考试作弊监控和识别)、测验注册和安排、测验付费、网络管理和认证管理方法提供服务. 在国内,全国计算机等级考试(National Computer Rank Examination,NCRE)采用全国统一命题,统一考试时间的形式,所有级别/科目全部实行上机考试,目前全国大部分高校都是 NCRE 考点.

基于客户机和服务器(Client/Server,C/S)和考试中心的机考模式,不能较好地适应未来大规模或常态化考试的需求. 如 NCRE 系统采用 C/S 模式,每一个考场(机房)仅需一个服务器,数量少,软件安装与环境设计简单,但是考试机数量多,通常数百台,考试机安装与环境设计较为繁琐、考试机的安装和部署工作量巨大[5]. 每次考试需要导入和导出测试数据,如果要更新客户机和服务器系统,增加的工作量不容小视. ETS 主办的美国研究生入学考试(Graduate Record Examination,GRE)考试,网上频频报导由 Prometric 公司实施的 GRE 考试考位不够或 GRE 考试报名考位暂满等消息,中国大陆只有 41 个大中城市设立了 GRE 考点(Prometric 考试中心). 随着考试规模的扩大,Prometric 公司需要花费巨大的人力和财力,用于增设考试中心,维护和更新相关设备和系统等. 这种单纯以资格认证为主的考试公司,并不能为学习者提供多少益处,因为其认证考试结果往往是通过与不通过,并且考试地点只能是考试中心,并不能真正实现"随时随地"进行测验,更不能实现以测试促进学习的目的.

在"互联网 +"催生 11 项重要领域的顶层设计背景下,本文主要从学习的角度,率先提出"互联网 +"测评及其框架和设计思路,一方面,推动测量理论、方法和技术的全面发展和大规模应用;另一方面,希望测评能够促进学习,更好地服务于自适应学习.

1 "互联网 +"测评

1.1 什么是"互联网 +"

在中国,"互联网 +"这一概念由易观国际董事长兼首席执行官于扬在 2012 年第五届移动互联网博览会首次提出,他认为"'互联网 +'这一公式应该是所有行业产品和服务与多屏全网跨平台用户场

景结合之后产生的一种化学公式”. 在 2015 年全国两会上，马化腾提交了一份《关于以“互联网 + ”为驱动，推进我国经济社会创新发展的建议》的议案，呼吁把“互联网 + ”提升为国家战略[6]. 2015 年 3 月 5 日，在十二届全国人大三次会议上，李克强总理在政府工作报告中首次提出“互联网 + ”行动计划. 2015 年 7 月 1 日国务院正式颁布《国务院关于积极推进“互联网 + ”行动的指导意见》(国发〔2015〕40 号，下简称意见). 意见中指出，“互联网 + ”是把互联网的创新成果(云计算、大数据技术和移动互联网等)与经济社会各领域深度融合，推动技术进步、效率提升和组织变革，提升实体经济创新力和生产力，形成更广泛的以互联网为基础设施和创新要素的经济社会发展新形态. 意见中明确了发展目标，到 2025 年，网络化、智能化、服务化、协同化的“互联网 + ”产业生态体系基本完善，“互联网 + ”新经济形态初步形成，“互联网 + ”成为经济社会创新发展的重要驱动力量. “互联网 + ”的本质特征是实现传统产业的在线化、数据化，实现产业升级或产生新的产业链[7].

这里所说的经济社会各领域理所当然地包括教育领域，而且“互联网 + 经济社会各领域”带来的变革毫无疑问地会深刻影响到教育的变革[8]. “互联网 + ”迅速发展，标志着教育正在走向大数据时代，充分发挥大数据对教育的引领作用，建构“互联网 + ”教育[8-9]的深度整合平台，已成为教育研究的一个热点问题[7]. “互联网 + ”给教育发展带来的机遇[6-7]，可以更好地实现教育公平、大规模教育、个性化学习. 2015 年 9 月举行的第十四届教育技术国际论坛以“技术、学习、教育创新: 教育技术的机遇与挑战”为主题，来自多个国家和地区的专家、学者针对“互联网 + ”时代下的教育信息化与教育变革、有效学习、技术支持下的创新学习以及信息技术与教育的深度融合等议题进行了深入交流和探讨[10]. “国家推进教育信息化，加快教育信息基础设施建设，利用信息技术促进优质教育资源普及共享，提高教育教学水平和教育管理水平”已经写入新修订的《中华人民共和国教育法》第六十六条.

伴随着云技术、大数据、移动互联网和物联网的发展，出现了翻转课堂、微课、MOOCs(慕课)、手机课堂、教育 APP、电子书包、创客运动、教育云等一系列新技术、新理念、新模式[6-9,11-14]，在线教育保持高速增长，中国在线教育市场空间巨大[15-16]. 各类在线教育平台模式逐渐形成，如 K12 教育平台(一起作业网、提分网、阿凡题、猿题库、学霸君、Triumph Learning)、高等教育平台(万门大学与啄木鸟教育)、职业教育平台(多贝、51CTO、沪江网、开课吧与无忧英语)和综合平台(网易云课堂、传课网、YY 教育、Udacity、Coursera、edX、学堂在线、好大学在线、华文慕课、Knewton).

教育或在线教育离不开测评. “互联网 + ”式的在线教育平台很大程度上能够克服传统教育评价难以收集评价依据和评价信息单一化、片段化的问题，可以全过程、全方位采集教育数据[17]. 平台记录、存储了学习者的一切结构化、半结构化和非结构化的学习行为数据. 以规模性(volume)、多样性(variety)、高速性(velocity)和价值性(value)等特点的大数据，只生成信息和知识的原生素材. 数据本身不会说话，只有对数据进行专业化的分析之后，作为生产要素的数据的大价值才会充分显现[18]. 实现大数据的价值增值，挑战是巨大的. 高质量的数据更是大数据发挥效能的前提和基础，强大、高端的数据分析技术是大数据发挥效能的重要手段.

在大数据情形下，保证数据质量面临挑战[19-21]. 对大数据进行有效分析的前提是必须要保证数据的质量，专业的数据分析工具只有在高质量的大数据环境中才能提取出隐含的、准确的、有用的信息; 否则，即使数据分析工具再先进，在充满“垃圾”的大数据环境中也只能提取出毫无意义的“垃圾”信息. 因此数据质量在大数据环境下显得尤其重要[20]. 数据质量对于数据分析结果的准确性有着决定性作用，直接影响教育评价的可信度与可靠性. 美国在整体上非常关注教育数据质量[22]. 如何保障教育评价的客观性、全面性与可靠性，对于教育教学质量的提升、教育的全面发展有着重要意义[22]. 中国的在线课程、在线教学、在线测评等面临着较为严重的质量管控或标准缺失问题[9,23-25]. 其中许多问题的解决都需要专业测评理论、方法和技术.

1.2　测评理论、方法和技术

1.2.1　**计算机化自适应测验**　计算机化自适应测验(Computerized Adaptive Testing，CAT)是一种适应被试能力水平的测验[26]. CAT 根据被试已经作答题目上的表现，从题库中序贯选择适合被试潜在能力水平的题目给被试作答. 避免能力高的被试作答太多容易的题目，能力低的被试作答太多难题. 相对于纸笔或非自适应的机考，CAT 具有如下优势: 1) 被试只需作答更少(一半)的题目，花费更短测试时间，就可以获得同样的测量精度; 2) 在计算机自动

评分技术的支持下,可以即时报告学生分数,并提供关于学生能力、知识和技能等丰富诊断信息,有助于辅助教学;3)使用多媒体技术甚至虚拟化技术让题型更新颖,使测验情景更具真实性,能够测量纸笔测验难以测量的多方面能力;4)建立在客观测量理论基础之上,结合最新的选题和组卷等算法,使测验质量和安全性更高.

CAT 已经广泛用于"美国士兵职业倾向成套测验"(Armed Services Vocational Aptitude Battery,ASVAB)、"美国研究生入学考试"(Graduate Record Examinations,GRE)、"(工商)管理类研究生入学考试"(Graduate Management Admission Test,GMAT)、"美国教师资格考试"(Praxis)、"美国护士执照或资格系列考试"(National Council Licensure Examination,NCLEX)、"美国全国教育进展评估"(National Assessment of Educational Progress,NAEP)[26-29],其中 NAEP 正在开展较大规模的自适应测验前期研究. 另外,K12 评价也将从纸笔测验转向计算机化在线考试[30]. 在中国,CAT 也运用到军队入伍考试中,通过自适应考试淘汰一部分有心理缺陷的人,每年淘汰一个师左右的人员. 测评需要建立在相应的测量理论、方法和技术之上,才能保证测评数据及结果的客观性、全面性与有效性. 下面较为详细地叙述 CAT 中测量理论、方法和技术.

1.2.2 CAT 中测量理论 CAT 常以现代测量理论之项目反应理论为基础. 与经典测量理论相比,项目反应理论具有以下优点[31-33]:项目反应理论深入测验的微观领域,将被试特质水平及其项目反应关联起来并且将其参数化、模型化,这是自适应测评客观化和量化的前提;项目反应理论模型的项目参数估计独立于被试样本,这是自适应测评建立大型题库的重要理论依据;项目难度参数与能力参数定义在同一个量表上,这一特点为自适应测评奠定了基础;Fisher 局部信息量和相对熵(Kullback-Leibler,KL)全局信息量,可以度量被试能力点估计测量误差和区间信息量,这是自适应测评构建选题算法的理论基础.

随着测量学研究的深入,众多研究表明,许多教育或心理测验,如 NAEP、国际学生评估项目(PISA)、国际数学和科学成就趋势研究(TIMSS)、中国国家基础教育质量监测(NAEQ)和西方五因素人格问卷(如 NEO-PI-R)等都是多维测验[34-38]. 然而单维项目反应理论难以处理. 因此,研究者针对用于多维测验分析的多维项目反应理论展开了大量的研究,涉及多维项目反应理论模型、参数估计、测验等值、计算机化自适应测验、临床结局评价、高考数据分析、心理测验数据分析等诸多方面[39-49]. 多维项目反应理论模型可以更好地互借各相关能力之间信息,提高测量精度. 下面给出一个用于 0-1 评分的多维项目反应理论模型:

$$P_j(\boldsymbol{\theta}) = P(u_j = 1 \mid \boldsymbol{\theta}) = c_j + (1 - c_j)/(1 + \exp(-(\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{\theta} - b_j))), \quad (1)$$

其中 $P_j(\boldsymbol{\theta})$ 表示多维能力向量 $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$ 条件下,被试在项目 j 上的正确作答概率,$Q_j(\boldsymbol{\theta}) = 1 - P_j(\boldsymbol{\theta})$ 表示错误作答概率,$\boldsymbol{a}_j$、b_j 和 c_j 分别为项目 j 的区分度、难度和猜测参数.

除多维项目反应理论模型之外,还有带反应时间(response time)心理计量模型[50-53]、高阶项目反应理论模型[53-55]、多水平项目反应理论模型[53,56]等. 多维项目反应理论模型相对较为复杂,并且分数报告粒度较粗. 由于社会各方面对测量学的要求不仅局限于宏观层面,还需要了解学生认知加工方面的信息,教育和心理测量需要从认知的角度来评估和诊断学生水平,以便更好地进行教学指导和促进发展. 因此,在教育测量中认知诊断评估应运而生. 认知诊断评估是新一代测验理论的典型代表或核心. 认知诊断评估旨在测量学生特定的知识和加工技能,为学生提供认知强项和弱项等信息,即报告学生掌握了哪些知识或技能,未掌握哪些知识或技能. 教育认知诊断评估的目标是成为"促进学习的评价". 近来,认知诊断评估理论、认知诊断模型、Q 矩阵(项目与所考查属性的关联矩阵)标定、Q 矩阵理论(如测验设计)、参数估计、信度和效度、等值方法、自适应测验等取得了迅速的发展[57-76].

下面给出一个应用广泛、便于解释、十分简单的认知诊断模型,确定性输入噪音与门模型(deterministic inputs,noisy "and" gate model,DINA)[77],又名约束潜在分类模型(Restricted latent class model,RLCM)[78]:

$$P_j(\boldsymbol{\alpha}) = P(u_j = 1 \mid \boldsymbol{\alpha}) = g_j^{1-\eta_j}(1 - s_j)^{\eta_j},$$

其中 $\boldsymbol{\alpha}$ 为离散型 0-1 属性向量(知识状态),$P_j(\boldsymbol{\alpha})$ 表示给定知识状态 $\boldsymbol{\alpha}$ 条件下,被试在项目 j 上的正确作答概率,$Q_j(\boldsymbol{\alpha}) = 1 - P_j(\boldsymbol{\alpha})$ 表示错误作答概率,$\eta_{ij} = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$ 表示知识状态 $\boldsymbol{\alpha}$ 条件下项目 j 上的理想反应,s_j、g_j 和 q_j 为项目 j 的失误参数、猜测参数和所测的属性向量.

在认知诊断评估或认知诊断模型中,需要指定

测验项目与考查的知识、技能等属性的关联关系，这种关系的形式化表示就是 Q 矩阵. Q 矩阵理论的概念是由认知诊断评估的开创者之一 Tatsuoka 提出的[59,79]. Q 矩阵的行表示项目，列表示属性. 一个 n 行 K 列的 Q 矩阵描述了 n 个项目与 K 个属性之间的关系. 矩阵的元素描述某个项目是否考查了某个属性. 如项目 j 考察了属性 k，则 $Q_{jk}=1$，否则为 0. 认知研究发现，认知属性和技能的掌握往往存在着一定的逻辑或心理先后顺序或认知结构，有研究者用属性层级结构或认知模型来描述[80]，并借助于图论中可达矩阵来表示. 可达矩阵是一类非常重要的矩阵，它在测验蓝图的编制中十分重要[63-64,81]. 梳理领域知识的前序和后续逻辑关系，对于学生学习和理解领域知识，具有重要意义[82]，对于教材的组织与编写也具有指导作用. 比如，在计算机学科中，知识的前序和后续逻辑关系(或演化关系)使用知识图谱[82]来表示.

1.2.3　CAT 中测量方法　“智慧评测”以促成个体化的“自适应”学习，如果没有“自适应”就谈不上“智慧评测”. 智慧评测的最关键部分就是自适应引擎，而自适应引擎具有智能关键在于算法，以解决如何适应性选题、如何对被试作答反应进行“记分”(score)和如何保证测评的安全性等问题. 算法和引擎可由研究开发团队通过不断地建模与测试来优化. 如果使用多维项目反应理论模型，自适应测验通常是根据被试当前所有作答反应，采用能力估计方法[83-84]得到最新的被试能力估计值，然后根据能力估计值按照一定的选题方法[48,85-87]选择试题给被试作答. 下面以多维连续型能力为例，简要介绍自适应引擎中的两类算法，单维连续型能力模型下的选题策略只是其特例. 对于认知诊断模型，需要更改替换 $P_j(\boldsymbol{\theta})$ 为 $P_j(\boldsymbol{\alpha})$、积分变成求和；$\boldsymbol{\alpha}$ 估计相对直接，无需迭代；对 $\boldsymbol{\alpha}$ 不满足一切求导的条件，没有 Fisher 信息量等概念，也没有 Robbins-Monro 式方法和 Fisher 信息量选题方法等；$\boldsymbol{\alpha}$ 估计方法可参考文献[88]，选题算法可参考相关文献[73-74,89-93].

(Ⅰ)能力估计方法. 极大似然估计方法：若被试作答 n 个项目，得分向量为 $\boldsymbol{u}$，给定项目反应理论模型和局部独立性假设，可计算似然函数和对数似然函数分别为

$$L(\boldsymbol{\theta}\mid\boldsymbol{u})=\prod_{j=1}^{n}P_j(\boldsymbol{\theta})^{u_j}Q_j(\boldsymbol{\theta})^{1-u_j},$$

$$l(\boldsymbol{\theta}\mid\boldsymbol{u})=\sum_{j=1}^{n}\left(u_j\ln P_j(\boldsymbol{\theta})+(1-u_j)\ln Q_j(\boldsymbol{\theta})\right).$$

对数似然函数对能力 $\boldsymbol{\theta}$ 的 1 阶微商为含能力 $\boldsymbol{\theta}$ 的非线性方程组，需要使用迭代算法估计能力 $\boldsymbol{\theta}$. 给定对数似然函数 2 阶微商(Hessian 矩阵)：

$$H(\boldsymbol{\theta})=\left(\frac{\partial^2 l(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right).$$

牛顿-拉夫逊算法(Newton-Raphson algorithm)的迭代公式为

$$\hat{\boldsymbol{\theta}}^{(t+1)}=\hat{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\delta}^{(t)}=\hat{\boldsymbol{\theta}}^{(t)}-(H(\hat{\boldsymbol{\theta}}^{(t)}))^{-1}\frac{\partial l(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}}\Bigg|_{\hat{\boldsymbol{\theta}}}(t).$$

如果使用信息矩阵 $I(\boldsymbol{\theta})=-E(H(\boldsymbol{\theta}))$ 替代 Hessian 矩阵，就可以得到费歇尔迭代方法(Fisher scoring approach)的迭代公式为

$$\hat{\boldsymbol{\theta}}^{(t+1)}=\hat{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\delta}^{(t)}=\hat{\boldsymbol{\theta}}^{(t)}+(I\hat{\boldsymbol{\theta}}^{(t)})^{-1}\frac{\partial l(\boldsymbol{\theta})\mid\boldsymbol{u}}{\partial\boldsymbol{\theta}}\Bigg|_{\hat{\boldsymbol{\theta}}}(t),$$

能力的极大似然估计量的渐近服从多元正态分布 $N(\hat{\boldsymbol{\theta}},(I(\hat{\boldsymbol{\theta}}))^{-1})$.

极大似然估计量 $\hat{\boldsymbol{\theta}}$ 的偏差为 $B(\hat{\boldsymbol{\theta}})=(B(\hat{\boldsymbol{\theta}}_1),B(\hat{\boldsymbol{\theta}}_2),\cdots,B(\hat{\boldsymbol{\theta}}_K))^{\mathrm{T}}$，其中元素 $B(\boldsymbol{\theta}_t)$ 计算公式为

$$B(\boldsymbol{\theta}_t)=\frac{1}{2}\sum_{k,p,q=1}^{K}I(\boldsymbol{\theta})^{pt}I(\boldsymbol{\theta})^{qk}E\left(\frac{\partial^3 l(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_q\partial\boldsymbol{\theta}_k}\right),$$

其中 $I(\boldsymbol{\theta})^{pt}=-1/E(\partial^2 l(\boldsymbol{\theta}\mid\boldsymbol{u})/\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}_t)$，$I(\boldsymbol{\theta})^{qk}=-1/E(\partial^2 l(\boldsymbol{\theta}\mid\boldsymbol{u})/\partial\boldsymbol{\theta}_q\partial\boldsymbol{\theta}_k)$，即表示信息矩阵中第 p 行 t 列、第 q 行 k 列元素的倒数.

贝叶斯估计方法：若给定能力的先验分布 $f(\boldsymbol{\theta})$，通常假设为多元标准正态分布. 观察到得分数据之后，更新的能力后验分布为

$$f(\boldsymbol{\theta}\mid\boldsymbol{u})=f(\boldsymbol{u},\boldsymbol{\theta})/f(\boldsymbol{u})=f(\boldsymbol{u}\mid\boldsymbol{\theta})f(\boldsymbol{\theta})/f(\boldsymbol{u})=L(\boldsymbol{\theta}\mid\boldsymbol{u})f(\boldsymbol{\theta})/f(\boldsymbol{u}),$$

其中 $f(\boldsymbol{u})$ 为 $\boldsymbol{u}$ 的边际分布. 按照能力后验分布，对能力 $\boldsymbol{\theta}$ 求期望，可得能力 $\boldsymbol{\theta}$ 的期望后验估计为

$$\hat{\boldsymbol{\theta}}=E(\boldsymbol{\theta}\mid\boldsymbol{u})=\int\cdots\int\boldsymbol{\theta}f(\boldsymbol{\theta}\mid\boldsymbol{u})\,\mathrm{d}\boldsymbol{\theta}.$$

按最大后验估计方法估计能力，牛顿-拉夫逊算法的迭代公式为

$$\hat{\boldsymbol{\theta}}^{(t+1)}=\hat{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\delta}^{(t)}=\hat{\boldsymbol{\theta}}^{(t)}-\left(\frac{\partial^2\ln f(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}^2}\right)^{-1}\Bigg|_{\hat{\boldsymbol{\theta}}}(t)\ \frac{\partial\ln f(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}}\Bigg|_{\hat{\boldsymbol{\theta}}}(t).$$

若能力的先验分布为 $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，此时有

$$\frac{\partial^2\ln f(\boldsymbol{\theta}\mid\boldsymbol{u})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}=H(\boldsymbol{\theta})-\boldsymbol{\Sigma}^{-1},$$

对应的信息矩阵 $I(\boldsymbol{\theta})=-E(H(\boldsymbol{\theta}))+\boldsymbol{\Sigma}^{-1}$.

加权似然估计方法：由于极大似然估计方法得到的能力估计有偏，为了得到无偏估计量，有 2 种方

法[83,94-96],一种是直接修正性(corrective)方法:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{MLE} - B(\hat{\boldsymbol{\theta}}_{MLE}).$$

另一种方法是预防性(preventive)方法,其牛顿-拉夫逊算法的迭代公式为

$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - \delta^{(t)} = \hat{\boldsymbol{\theta}}^{(t)} - (\partial S(\hat{\boldsymbol{\theta}}^{(t)})/\partial \boldsymbol{\theta})^{-1} S(\hat{\boldsymbol{\theta}}^{(t)})$,其中 $S(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta} \mid \boldsymbol{u})/\partial \boldsymbol{\theta} - I(\boldsymbol{\theta})B(\boldsymbol{\theta})$,相比极大似然估计方法,将得分函数(score function)由对数似然函数修改成为 $S(\boldsymbol{\theta})$.

(Ⅱ)选题方法. 兼顾单维和多维项目反应理论模型,描述了 CAT 中常用的选题方法,如以单维项目反应理论模型为例的 Robbins-Monro 式方法和 a 分层方法,和以多维项目反应理论模型为例的 Fisher 信息量、KL 信息量、最大优先级指标选题方法.

CAT 根据被试已经作答题目上反应,从题库中序贯选择适合被试能力水平的题目. 不妨假设被试作答的题目难度依次为 $b_1, b_2, \cdots, b_n, \cdots$(下标表示 CAT 题目顺序). 根据 Lord 提出的 CAT 基本想法:"上一题作答正确,增加下一题难度;反之,降低下一题难度",且下一题的难度依赖于当前题目上的反应(对应被试能力),而被试在题目上的反应为随机变量,概率分布由(1)式确定,因此 $b_j(j = 1,2,\cdots)$ 为随机变量. 下一题的难度仅依赖于当前题目的难度. 随机变量序列 $b_1, b_2, \cdots, b_n, \cdots$,满足马尔柯夫性(Markov property),随机变量序列形成随机过程(stochastic process).

(i) Robbins-Monro 式方法:考虑到序列 $b_1, b_2, \cdots, b_j, \cdots$ 的随机性,Lord 将随机逼近(stochastic approximation)理论中 Robbins-Monro 方法应用于自适应测验选题[97-100]:$b_{n+1} = b_n + d_n(u_n - \gamma)$,其中 d_n 为预先设定的数值为正且递减的数列,如 $d_n = d_1/n(n = 2,3,\cdots)$,$d_1$ 和 γ 为设定的正数,如 $d_1 = \sqrt{2\pi}/(a(1-c))$ 和 $\gamma = 0.5(1+c)$(多维情形需要适当变化). 根据 Robbins-Monro 理论[101],在一定条件下适当选择 d_n 和 γ,最终的难度(能力和难度位于同一量尺)为能力的一致估计. 值得注意的是,Lord 从难度序列$\{b_n\}$的角度构建了随机过程,用于构建选题方法;Chang Hua-Hua 认为 CAT 中作答反应随机变量序列$\{u_n\}$存在相依关系,将杜布(Joseph Leo Doob)创立的鞅论(Martingale Theory)应用于 CAT,设计自适应过程并推导出了能力估计的大样本渐近性质[99-100]. 鞅论是随机过程的一个重要分支.

(ii) Fisher 信息量选题方法:由于能力的极大似然估计量的渐近分布为 $N(\hat{\boldsymbol{\theta}}, (I(\hat{\boldsymbol{\theta}}))^{-1})$[83,102-103],即均值为能力的极大似然估值 $\hat{\boldsymbol{\theta}}$,而协方差矩阵是在能力 $\hat{\boldsymbol{\theta}}$ 处的 Fisher 信息量矩阵的逆. 因此要准确估计 $\boldsymbol{\theta}$,就要使 $\hat{\boldsymbol{\theta}}$ 的误差小,即能力为 $\hat{\boldsymbol{\theta}}$ 的测验 Fisher 信息矩阵 $(I(\hat{\boldsymbol{\theta}}))^{-1}$ 的某些测度(如行列式)越小越好. 若被试已经作答了 n 个项目,对应的能力估计值为 $\hat{\boldsymbol{\theta}}_n$,最大信息量选题方法所选择的题目为

$$j_{n+1} = \arg\max_j \{\det((I(\hat{\boldsymbol{\theta}}_n) + I_j(\hat{\boldsymbol{\theta}}_n))), j \in R\},$$

其中 R 表示题库中某被试尚未作答的题目集合(剩余题库),$I(\hat{\boldsymbol{\theta}}_n)$ 表示被试已作答的所有题目的信息量之和,$I_j(\hat{\boldsymbol{\theta}}_n)$ 表示题目 j 上的信息量.

(iii) a 分层方法:基于局部信息量的 Fisher 信息量选题方法,趋于选择区分度较高的项目,一方面会造成低区分度项目很少使用[104-105],另一方面,在测验长度较短的 CAT 中,若被试在前几个题连续做错,极易造成能力低估[100,106]. 张华华等提出了 a 分层选题方法[104-105,107-110],涉及 a 分层方法、b 分块 a 分层方法、层优化方法、a 分层方法与条件概率方法(SH)或优先级指标方法结合等. a 分层方法是在测验前期能力估计不准的条件下,使用区分度小的项目,而在测验后期使用区分度大的项目,又称"升 a 方法". 若题库容量为 N,欲将题库分成 L 层,下面简要给出 a 分层方法和 b 分块 a 分层方法的步骤.

a 分层方法的步骤:先将所有项目依区分度大小升序排列,将前 L 个区分度最小的项目作成第 1 组;将区分度为第 $L+1, L+2, \cdots, 2L$ 小的项目作成第 2 组,依次类推;然后将各组中区分度最小的项目抽取出来,放在第 1 层子题库中;将各组中区分度次小的项目抽取出来放入第 2 层子题库;依次类推,最后将各组中区分度最大的项目放入第 L 层子题库. 这样,各层项目在区分度的平均值是上升的. 进入 CAT 测验时,各个被试是在第 1 层题库中选择与当前能力估计值相匹配的项目;到一定时候,则进入第 2 层题库选题,最后才在第 L 层题库选题.

b 分块 a 分层方法的步骤:先将题库中所有项目按 b 从小到大排列,相邻 L 个构成 1 个块(block),然后每块中又按 a 升序排列,各块中 a 值最小者放在第 1 层;次小者放在第 2 层;依次类推,最大者放在第 L 层. 于是每一层中项目难度分布与整个题库难度分布相似,但各层区分度平均值从小到大变化.

近年来，有许多研究者对 a 分层方法做了进一步优化或推广. 如为了进一步减小卡方值、平衡曝光率、提高题库安全性方面，丁树良等[33] 对 0-1 和多级评分模型下各层项目数分配方案、层跳转规则、动态平滑分层方面开展了研究. 还有研究者将 a 分层方法推广到多维项目反应理论模型[111].

KL 信息量选题方法: 在测验长度较长时，Fisher 信息量才能用于度量接近能力真值的能力估计处的测量误差，因此 Fisher 信息量被称为局部信息量(local information)[112]. 而在测验初期，能力估计值与真值相差比较大时，能力估计值处 Fisher 信息量用处不大，并不能用于衡量能力真值处的测量误差. 因此，Chang Hua-Hua 等提出基于 KL 的全局信息量(global information)[112] 选题方法，该方法所选择的题目为

$$j_{n+1} = \arg\max_j \{KL_j(\hat{\boldsymbol{\theta}}_n), j \in R\},$$

式中题目 j 上的全局信息量 $KL_j(\hat{\boldsymbol{\theta}}_n)$ 及涉及的相对熵计算公式分别为

$$KL_j(\hat{\boldsymbol{\theta}}_n) = \int_{\theta_{10}-\delta_n}^{\theta_{10}+\delta_n} \cdots \int_{\theta_{10}-\delta_n}^{\theta_{10}+\delta_n} KL_j(\hat{\boldsymbol{\theta}}_n \parallel \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta},$$

$$KL_j(\hat{\boldsymbol{\theta}}_n \parallel \boldsymbol{\theta}) = P_j(\hat{\boldsymbol{\theta}}_n) \ln(P_j(\hat{\boldsymbol{\theta}}_n) / P_j(\boldsymbol{\theta})) + Q_j(\hat{\boldsymbol{\theta}}_n) \ln(Q_j(\hat{\boldsymbol{\theta}}_n) / Q_j(\boldsymbol{\theta})),$$

其中 δ_n 可取 $3/\sqrt{n}$[48,99,112]. 除了 Fisher 和 KL 信息量选题方法外，还有简化 KL、香农熵、互信息、最小化误差、最大最小特征值(E-Optimality) 等选题方法[48,84-87].

最大优先级指标: 有研究[85,113] 将最大优先指标方法[114] 应用于多维项目反应理论模型，优先级指标(priority index) 计算公式为

$$PI_j = \prod_{d=1}^{D} (w_d f_{jd})^{c_{jd}},$$

其中 c_{jd} 为约束矩阵中的元素，c_{jd} 指示题目 j 上考察了某个能力维度测量精度、内容领域、答案选项、题型、是否被选中(用于曝光控制) 等约束指标; $f_{jd} = (X_k - x_k)/X_k$ 为缺额(quota left) 比率; w_d 为权重.

1.2.4　CAT 中测量技术

(Ⅰ) 等值技术 测评离不开题库维护和建设，网上许多题库充其量只能算“题堆”，这些题库仅仅只是把题目累积起来. 题库中的项目参数必须具有可比性，否则测验结果就不可以比较. 这就要求题库建立在项目反应理论基础之上，借助联接设计(linking design)，将题库中题目进行组卷、施测，然后估计项目参数，并通过等值方法(如项目特征曲线等值法) 进行项目参数等值. 获得统一量尺上的项目参数，才能进行 CAT 施测. 为了进行年级增值评估，还需要垂直等值技术.

(Ⅱ) 组卷方法 在等值设计中，涉及到自动化生成试卷方法或技术. 比如，有一张原始卷和一个题库，可能需要按照原始卷从题库中自动抽卷，并组成很多试卷，这些试卷要求与原始卷的质量及难易度一样. 题库平台中必要开发相应的智能组卷算法，在各种各样约束条件下，快速生成满足要求的试卷. 张华华团队已经成功开发了应用于汉语语言考试(HSK) 的快速智能组卷算法和程序[115-116]，以及能在测试时根据被试能力水平实时而动态生成试卷的算法[30]. 这些算法或技术可提高测验质量和安全性.

(Ⅲ) CAT 终止策略 根据测验长度不同，分为定长和变长 CAT. 定长 CAT 只需以固定的测验长度为终止策略. 定长 CAT 的终止测验主要可以分为 2 类，一类是以测验能力估计标准误的绝对度量标准. 有研究者[117] 基于 Fisher 信息矩阵 $I(\hat{\boldsymbol{\theta}})$ 给出了最小化行列式规则(D 规则)、最小化特征值规则(E 规则)、最小化迹规则(T 规则) 等绝对标准型终止策略，其中 D 规则和 E 规则是基于卡方分布及假设检验相关结论而得到，T 规则直接以能力估计的总方差或各维度能力的最大方差小于指定值的方式终止测验; 另一类是题库中剩余项目所能减少测量误差的边际标准. 因为 CAT 的长度与选题策略和题库有关，为了解决作答太多的项目而精度提高幅度较小的问题，有研究者提出了预测标准误减少量(predicted standard error reduction，PSER) 终止规则[85].

(Ⅳ) 在线估计或标定方法 除了采用等值方式建立题库，还可以通过 CAT 在线估计或标定方式获得项目参数和 Q 矩阵等. 起初可以根据老师的认识，由老师标注试题参数和所考查的知识，标注好了之后，结合数据和数学模型对试题参数进行修正并进行新题批量标注. 在确定模型之后，开始进行新题标注以自动扩充题库，即在自适应测试中，在学生作题时不知不觉的把新题放上去，学生做完之后，把这些题收回来，不纳入学生计分，而是对这些题目的参数进行估计. 使用恰当的数学模型和自适应测验就可以使标定工作智能化地运行，进行题库增量建设[118-119].

(Ⅴ) Q 矩阵设计与标定 认知模型确定之后，要充分发挥认知诊断评估的作用，首要的前提是有合理的测验编制. 事实上测验蓝图的编制，即测验的设计十分重要，因为测验蓝图直接关系到测验是否能

够为每一个被试提供充分详细的信息. 丁树良等在 Q 矩阵理论、认知诊断测验蓝图设计、Q 矩阵标定、Q 矩阵理论应用等方面取得了丰富的研究成果,并且发现在属性之间不存在补偿条件下,采用 0-1 评分时可达阵设计测验蓝图,可显著提高测验或 CAT 的分类准确率[63,120]. 哥伦比亚大学应志良等[121-122]开展了 Q 矩阵标定的数据驱动方法,陈平[71,123,124]、汪文义[125-126]、喻晓锋等[127-128]对 Q 矩阵修正或在线标定方法进行了研究.

1.3 "互联网 +"测评

"互联网 +"与测评两者如何深度融合? 笔者认为,关键是如何借助云计算、大数据、移动互联网和物联网的最新成果,并结合测量理论、方法和技术,解决测评中急需解决的众多问题,更好地拓展测量理论、方法和技术的全面发展和大规模应用,服务于个性化学习,提升教育教学质量和促进教育的全面发展. 下面主要从 4 个方面进行叙述.

1.3.1 考试安全性 CAT 需要从题库中选择难度最适合被试的试题,对于相同或不同时间参加测验的相同或不同被试,有的选题方法常会频繁选择和使用某些项目,CAT 的安全性曾受到过严重危机. Kaplan 教育中心曾雇人参加 GRE 测验并记忆题目. 在 2002 年网上曝出 GRE 的 CAT 真题后,中国大陆、香港地区、台湾地区和韩国重新启用纸笔考试[98,129]. 随着最新的选题策略和题库自动扩充技术的发展,加上许多大公司,如甲骨文、微软、雅虎、英特尔、威睿(Vmware)、谷歌等提供多种云计算下安全技术,可较好地满足题库、测验和试题的安全性. 考试时间、考试试题和试卷的随机性,将大幅提高测试安全性. 如随着题库的日益扩充,试题组合随着试题数呈指数式增长,CAT 每个考生试卷并不相同,要想通过其他人偷题作弊几乎不可能. 故,大容量题库下的计算机化考试,自动化和个性化生成考生试题或试卷,考试安全问题就能得以有效解决. 背一套试卷没有多大意义,要偷整个题库并记住则不太可能.

1.3.2 教育大数据分析技术 教育大数据主要来源于教育过程中过程性、即时性的行为与微观表现,可以分析微观、个体的学生的特征,发现个性,当然个性数据中也可以发掘学生群体的共性. 教育大数据虽然大,但是价值大、质量高的数据仍有待挖掘,这需要大数据分析技术,且需要借助基于测评理论的大数据分析技术. 基于测评理论、方法和技术收集的高质量测评数据,是教育大数据中非常重要的一部分,它既是较好的结构化或半结构化数据,便于数据分析; 还是重要的教学效果的客观化标准,结合教育大数据和大数据分析技术,可以进行相关影响因素分析甚至进行因果分析等,可验证教育或在线教育的效果,并为优化课程、教学、学习资源推荐等提供客观依据.

例如,使用带反应时间或可修改答案的心理计量模型[50-52,130],对系统记录的每道题作答时长、修改情况等进行分析,可以侦测考试作弊或随机作答行为,可辅助或更准确估计能力和项目参数,或可用于侦测考试作弊行为、设计更有效的测验[50]和用于选题以平衡项目曝光[51]等. 在线教育平台可以用来收集学生的学习进度、有效的学习时长、章节或单元的阅读频次、练习和复习频次等数据[9],结合测评结果可以进行溯因分析,如识别学习不积极且未学好的学生. 测评不仅是对学习者的能力排序,更是为了促进学习成绩上升和能力提高; 基于教育大数据的全面性,可进行追踪式测评,借助于垂直等值技术和纵向数据分析技术,有望解决增值评估问题.

美国目前最大自适应学习平台 Knewton,依托亚马逊的大数据和云平台服务(AWS),测评核心技术明显不够,Knewton 系统中测评还不够智慧. 比如 Knewton 平台中提供的数学乘方、开方或根式运算(operations with radical expressions)测试,测试内容涉及小学、中学甚至大学(复杂的多项式乘法)课本中的相关知识. 在一个测验中试题难度有较大的跨越,自适应程度不高,学生一直在做题,并不会适时终止测验并报告掌握和未掌握的知识,影响学习者的积极性. Knewton 主要通过测试学生,得到学生知道什么、学生学习方式等,根据知识图谱和概率图模型推荐个性化的学习路径[131]. 在这一系列过程中,Knewton 不断的挖掘学生表现数据. 根据一个给定活动的完成情况,系统指引学生进入下一个活动. 2 个学生的学习路径图可能大相径庭,一个学生朝左走,另一个学生朝右走,这就得因材施教,如果一个同学在某些方向遇阻,那应该帮助他导向另一个方向. 在美国较多中学和大学都在使用 Knewton 系统,美国时代周刊上刊登介绍 Knewton 的文章,接受采访的学生们对 Knewton 的评价较高.

Knewton 公司只是"互联网 +"测评融合的一个序曲. 至少目前为止,两者并未达到深度融合. 开展教学大数据挖掘、网络资源收集和评价、嵌入学习分析技术、个性化学习内容推荐、自动问答系统、作文自动评分、虚拟化技术、多模态交互等研究,实现技

术、资源、数据、测评和推荐有机结合，充分发挥在线教育的优势，仍有相当长的路要走.

1.3.3　云测评平台(平台即服务)　基于云服务提供商提供的云计算服务或开源云计算平台，开发可供用户定制的云测评平台. 建立在云测评平台上的测评系统，具有较高的安全性、可靠性和可扩展性，部署和维护也更为方便; 实行按需付费，可以降低测评成本，如各考试机构可以“解放”内部的 IT 资源，如不必增置和维护某些硬件，也无需自行开发测评软件; 可以获得专业化的测评软件和享受专业化咨询服务. 分析数据作为教师实施个性化教学和分层教学的科学依据，面向大量学习者的个性化教学，在互联网技术、大数据分析技术和云技术的支持下已经成为现实. 现在所要做的是如何优化算法、改进分析方法，使其对学习、教学等相关数据的分析更加准确、丰富、有价值，使学习更有效率[132]. 教育质量评估可准确地向政府机构报告教育质量的现状，为教育决策提供信息、依据和建议，具有重要意义. “互联网 + ” 教育质量评估，将是“互联网 + ”测评优先发展方向，将更好地实现评估、诊断和服务功能，促进教育公平和均衡发展，同时更好地进行数据共享.

1.3.4　云应用组件(软件即服务)　教育测量中数据分析的商业软件，软件多样、功能单一但使用并不简单，完全不能满足“互联网 + ”测评的实时性和常态化的需求. 为推动“互联网 + ”测评的广泛应用，充分利用云平台、云计算、大数据技术、富媒体和自媒体技术、虚拟化技术等，开发用于教育测评的云应用组件，涉及项目反应理论参数估计、题库系统、等值算法、自适应测验算法、组卷算法、多阶段测验算法、诊断模块等网络化模块开发，并提供相应服务. 笔者及团队已经开发了真正应用的 B/S 架构下的计算机化自适应测验系统[70]、计算机自动组卷系统[115].

2　“互联网 + ”测评的发展方向

技术已经渗透进教育的每个可以改变的环节，技术平台能够有效降低过去在个性化教学实施上所消耗的大量成本，一个全新的教育生态系统正在逐渐建立. 从纸质版教材到数字化教学材料，从按年龄分班到混龄教学，从班级教学到个性化教学，从简单的学生成绩到教育大数据(教学过程全程技术跟踪) ，从成绩排名到是否达到评估标准(标准和教学目标挂钩) ，从统一的教材到个性化学习资料推荐，从以学习资源为中心的学习模式转变为以学习者为中心的学习模式，以课堂教学到基于项目的学习方法等等，这些转变都在悄然进行. 科技与教育的深度融合，或将带来一次测评变革和学习变革.

2.1　“互联网 + ”测评的标准化

《国家中长期教育改革和发展规划纲要(2010—2020 年) 》指出“提高义务教育质量. 建立国家义务教育质量基本标准和监测制度，严格执行义务教育国家课程标准、教师资格标准. ”，还提到“整合国家教育质量监测评估机构及资源，完善监测评估体系，定期发布监测评估报告”. 《国务院关于积极推进“互联网 + ”行动的指导意见》强调“互联网 + ”益民服务，明确指出“探索新型教育服务供给方式. 鼓励互联网企业与社会教育机构根据市场需求开发数字教育资源，提供网络化教育服务. ”

在借鉴美国 NAEP 基础之上[133]，中国国家基础教育质量监测(NAEQ) 开发的监测工具采用了标准参照测验. 标准参照测验关注学生具体知识或技能的掌握情况及达到的水平. 标准参照测验有助于发挥考试的诊断功能和促进学生发展，从而对教育评价产生了深刻影响[32]. 标准参照测验的广泛应用或需求，较好地体现了其在教育评价中的重要性: 美国的“力争上游”教改计划中强调“采用新型标准和评价，促使学生在大学或工作岗位上取得成功，在全球范围内具备更好的人才竞争力”; 美国前教育部长阿恩·邓肯(Arne Duncan) 曾表示“一旦建立和采用新的标准，就需要创建新的测试，测量学生是否满足这些标准”[134].

因此，在提供优质数字教育资源的同时，并建立教学资源、测评与课程标准挂钩，采用云端提供各年级各学科的标准参照测验形式的测评，以及实现 NAEQ 自适应化，这无疑是中国未来测评发展的一个重要方向. 老师和学生利用终端设备(如平板电脑、手机等) 接入云端，进行随堂自适应测评、自适应指导和自适应学习，老师将从编制试卷、批改试卷或作业、数据分析等繁重工作中得到解放. 将来，当“互联网 + ”基于标准的测评标准化、常态化和普遍化，周期性的国家基础教育质量监测甚至变得不再需要，只要通过大数据分析日常的教育大数据(包括标准测评大数据) ，就可以实现全覆盖的国家基础教育质量监测宏伟愿景.

2.2　“互联网 + ”测评的服务化

随着《中国基础教育大数据发展蓝皮书(2015) 》在国内正式公布，各方开始关注教育大数

据的同时,“互联网 +”测评将更好地推动教育大数据的发展. 根据云计算的按需服务和消费模式,“互联网 +”测评也将采用服务提供模式. 云计算等新的计算机技术为构建面向服务的“互联网 +”测评提供技术保障,提供通用考试服务平台,将极大地降低重复开发系统的成本、维护和管理成本. “互联网 +”测评将更好地追踪老师和学生等的状态变化,丰富测评的抽样数据,甚至不再是样本数据而是总体数据. “互联网 +”测评借助于大数据分析技术,可以更好地进行测评结果的相关因素分析甚至因果分析. “互联网 +”测评,尤其是移动互联网,将促使“以测促学”和在测试中运用知识和内化知识,让学习无处不在.

2.3 “互联网 +”测评的学习化

随着“互联网 +”测评的服务化,不再局限于纸笔考试和考试中心模式,各种学业、职业、认证、资格考试将重返学校. 人类的测评观念也将发生根本性转变,测评不是终结,不再是一考定终生,而是终生学习,测评是为了更好地促进学习,让学习高效化、个性化或自适化等. 基于教育大数据,建立促进个性发展的教育体系,是未来学校发展的基本趋势,如实现因人施测、个性化学习(学习能力的匹配、查漏补缺、最佳学习方式或路径的推荐)、因材施教,并通过数据分析建立学习者个人知识地图等[135]. 互联网 +”测评将更好地利用大数据改变人类认知、帮助发现真正的学生潜在特质(个性、能力、知识地图). 教育大数据首先可以建立自适应学习和个性化教育产品,建立个性化教育体验,为每一个学生的学习路径进行优化,并通过数据来实证教育效果[3].

前已叙及,Knewton 是一个提供个性化教育平台的公司. 创立于 2008 年的 Knewton,被誉为“全球最领先的自适应学习平台”,其基于大数据分析和推荐系统,能够根据学生特点和学习习惯,即时调整内容供应,使教学更加个性化,从而提高学习效果. 它主要通过 3 项核心服务优化学习过程: 为学生提供内容推荐服务、为教师提供学情分析服务、为内容提供商提供内容洞察和分析服务. 在 2011 年 1 月,亚利桑那州立大学开始使用由 Knewton 提供个性化学习体验的课程[136]: “学生的退课率由 16% 降至 7%,通过率由 64% 升至 75%,45% 的学生提前 4 星期完成课程学习,并进行下一阶段学习”. Knewton 目前的测试中覆盖了 6 ~ 12 年级数学、9 ~ 11 年级生物、4 ~ 6 年级英语等. 采用测验题目与短视频相结合的学习方式,短视频来源于 YouTube(有的含有广告),还会推荐作业和教学材料. Knewton 不仅学生可用于学习,教师还可以创建课程并添加作业和邀请学生; 还允许用户创建测验题目,如添加内容、题目、选项、答案、解释、适合年级水平等,甚至还可创建教学材料,如添加内容、教学内容、适合年级水平等. 众多公司与 Knewton 合作,开发新一代产品,如 Triumph Learning 公司与 Knewton 合作开发了 Waggle,实现州立核心标准(Common Core State Standards,CCSS) 与 K-12 教学材料结合. 2016 年 1 月 20 日,中国好未来集团与 Knewton 签署全方位的合作协议,将 Knewton 的自适应平台用于在线教学.

如何从智慧测评到智慧学习[137],笔者认为自适应技术只是一个实现“智慧学习”的辅助工具,不会也不能把教师赶出教室,而是利用先进技术帮助老师能够更好地进行教学. 如何用自适应技术以及互联网技术来帮助一线教师,将这一理念运用在具体教学中,中国大连的教师进行了有效尝试,并已有了实践案例. 教师继续上课,在上课的过程中把课堂作业发给学生,把课堂作业收上来之后并用高速扫描仪上传到云端,通过云计算对学生的掌握情况进行分析,把学生的诊断报告打出来,教师可以马上把这些报告发给学生[99]. 这不是一个虚无缥缈的幻想,而是互联网技术和计算机化考试普及后智慧学习的必然结果,期望融合测量学理论和大数据技术的自适应考试和学习模式能很好地应用于教学过程,真正实现对广大学习者因材施教这一古老而伟大的目标.

3 参考文献

[1] 邹卓鹏.《教育测量与评价》发刊词 [J]. 教育测量与评价,2008(1):1.

[2] 侯光文. 教育测量与评价的基本原理 [J]. 山东教育科研,1991(1):73-76.

[3] 李子. “教育技术”3 大巨头如何改变教育 [EB/OL]. [2016-04-21]. http://toutiao.com/i6275894062935441921/.

[4] Miller M D,Linn R L,Gronlund N E. Measurement and assessment in teaching [M]. New Jersery: Pearson Education,Inc,2009.

[5] 胡海斌,周智勇,李青,等. 全国计算机等级考试系统环境的自动部署 [J]. 计算机应用,2014,34(S2):361-363.

[6] 平和光,杜亚丽. “互联网 + 教育”: 机遇、挑战与对策 [J]. 现代教育管理,2016(1):13-18.

[7] 张忠华，周萍. “互联网 + ”背景下的教育变革 [J]. 教育学术月刊，2015(12)：39-43.
[8] 周荣斌. “互联网 + ”国家行动的教育际遇 [J]. 思想政治课教学，2015(12)：9-12.
[9] 胡永斌. “互联网 + ”背景下美国 K-12 教育转型分析 [J]. 中国电化教育，2016(3)：33-38.
[10] 乜勇，杨玉玉. “互联网 + ”时代下的技术与教育创新：“第十四届教育技术国际论坛”述评 [J]. 中国远程教育，2016(1)：68-72.
[11] 朱月翠，张文德. “互联网 + 教育”基本模型探析 [J]. 中国教育信息化，2015(19)：12-15.
[12] 王乔峰，曹效英，路璐. “互联网 + 教育”模式的发展情况分析 [J]. 中国教育信息化，2015(15)：9-11.
[13] 祝智庭. 电子书包标准与应用对接“人人通” [J]. 中国现代教育装备，2014，197(13)：5-10.
[14] 祝智庭，刘名卓. “后 MOOC”时期的在线学习新样式 [J]. 开放教育研究，2014，20(3)：36-43.
[15] 刘东梅. 在线教育二十年：从“教育 + 互联网”到“互联网 + 教育” [J]. 互联网经济，2015(7)：90-97.
[16] 胡梅，马斌. “互联网 + 高等职业教育”的现实可能与当代变革 [J]. 现代教育管理，2016(1)：19-24.
[17] 张岩. “互联网 + 教育”理念及模式探析 [J]. 中国高教研究，2016(2)：70-73.
[18] 吴瑜，刘欢，任友群. “互联网 + ”校园：高校智慧校园建设的新阶段 [J]. 远程教育杂志，2015(4)：8-13.
[19] 宫学庆，金澈清，王晓玲，等. 数据密集型科学与工程：需求和挑战 [J]. 计算机学报，2012，35(8)：1563-1578.
[20] 宗威，吴锋. 大数据时代下数据质量的挑战 [J]. 西安交通大学学报：社会科学版，2013，33(5)：38-43.
[21] 余伟，李石君，杨莎，等. Web 大数据环境下的不一致跨源数据发现 [J]. 计算机研究与发展，2015，52(2)：295-308.
[22] 郑燕林，柳海民. 大数据在美国教育评价中的应用路径分析 [J]. 中国电化教育，2015(7)：25-31.
[23] 王丽莉，孙宝芝. 互联网 + 时代背景下网络教育发展新趋势 “2015 国际远程教育发展论坛”综述 [J]. 中国远程教育，2015(12)：12-17.
[24] 杨现民，王榴卉，唐斯斯. 教育大数据的应用模式与政策建议 [J]. 电化教育研究，2015(9)：54-61，69.
[25] 陈丽，林世员，郑勤华. “互联网 + ”时代中国远程教育的机遇和挑战 [J]. 现代远程教育研究，2016(1)：3-10.
[26] Chang Hua-Hua，Ying Zhiliang. Computerized adaptive testing [A] // Salkind N J，Rasmussen K. Encyclopedia of measurement and statistics [M]. Thousand Oaks，CA：SAGE Publications，Inc，2007：170-173.
[27] Quellmalz E S，Pellegrino J W. Technology and testing [J]. Science，2009，323(5910)：75-79.
[28] Yan Duanli，von Davier A A，Lewis C. Computerized multi-stage testing theory and applications [M]. Boca Raton：CRC Press，2014.
[29] van der Linden W J，Glas G A W. Computerized adaptive testing：theory and practice [M]. New York：Kluwer Academic Publishers，2000.
[30] Zheng Yi，Chang Hua-Hua. On-the-fly assembled multi-stage adaptive testing [J]. Applied Psychological Measurement，2015，39(2)：104-118.
[31] 漆书青，戴海崎，丁树良. 现代教育与心理测量学原理 [M]. 北京：高等教育出版社，2002.
[32] 戴海琦. 心理测量学 [M]. 北京：高等教育出版社，2010.
[33] 丁树良，罗芬，涂冬波. 项目反应理论新进展专题研究 [M]. 北京：北京师范大学出版社，2012.
[34] Debeer D，Buchholz J，Hartig J，et al. Student，school，and country differences in sustained test-taking effort in the 2009 PISA reading assessment [J]. Journal of Educational and Behavioral Statistics，2014，39(6)：502-523.
[35] Makransky G，Mortensen E L，Glas C A W. Improving personality facet scores with multidimensional computer adaptive testing：An illustration with the Neo Pi-R [J]. Assessment，2012，20(1)：3-13.
[36] Rijmen F，Jeon M，von Davier M，et al. A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys [J]. Journal of Educational and Behavioral Statistics，2014，39(4)：235-256.
[37] Yao Lihua，Boughton K A. A multidimensional item response modeling approach for improving subscale proficiency estimation and classification [J]. Applied Psychological Measurement，2007，31(2)：83-105.
[38] Zhang Jinming. Calibration of response data using MIRT models with simple and mixed structures [J]. Applied Psychological Measurement，2012，36(5)：375-398.
[39] Cai Li. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm [J]. Psychometrika，2010，75(1)：33-57.
[40] Reckase M D. Multidimensional item response theory [M]. New York：Springer，2009.
[41] 刘红云，骆方，王玥，等. 多维测验项目参数的估计：基于 SEM 与 MIRT 方法的比较 [J]. 心理学报，2012，44(11)：121-132.
[42] 杜文久，肖涵敏. 多维项目反应理论等级反应模型 [J]. 心理学报，2012，44(10)：1402-1407.
[43] 康春花，辛涛. 测验理论的新发展：多维项目反应理论 [J]. 心理科学进展，2010，18(3)：530-536
[44] 毛秀珍，辛涛. 多维计算机化自适应测验：模型、技术和方法 [J]. 心理科学进展，2015，23(5)：907-918.

[45] 涂冬波,蔡艳,戴海琦,等. 多维项目反应理论:参数估计及其在心理测验中的应用[J]. 心理学报,2011,43(11):1329-1340

[46] 许志勇,丁树良,钟君. 高考数学试卷多维项目反应理论的分析及应用[J]. 心理学探新,2013,33(5):438-443.

[47] 詹沛达,王文中,王立君,等. 多维题组效应 Rasch 模型[J]. 心理学报,2014,46(8):1208-1222.

[48] Wang Chun, Chang Hua-Hua. Item selection in multidimensional computerized adaptive testing: gaining information from different angles [J]. Psychometrika, 2011, 76(3): 363-384.

[49] Zheng Yi, Chang Chi-Hung, Chang Hua-Hua. Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement [J]. Quality of Life Research, 2013, 22(3): 491-499.

[50] Wang Chun, Fan Zhewen, Chang Hua-Hua, et al. A semiparametric model for jointly analyzing response times and accuracy in computerized testing [J]. Journal of Educational and Behavioral Statistics, 2013, 38(4): 381-417.

[51] Fan Zhewen, Wang Chun, Chang Hua-Hua, et al. Utilizing response time distributions for item selection in CAT [J]. Journal of Educational and Behavioral Statistics, 2012, 37(5): 655-670.

[52] Wang Chun, Chang Hua-Hua, Douglas J A. The linear transformation model with frailties for the analysis of item response times [J]. British Journal of Mathematical and Statistical Psychology, 2013, 66(1): 144-168.

[53] Fox J P. Bayesian item response modeling theory and applications [A]//Fienberg S E, van der Linden W J. Statistics for social and behavioral sciences [C], New York: Springer, 2010.

[54] Huang Hung Yu, Wang Wen Chung, Chen Po-Hsi, et al. Higher-order item response models for hierarchical latent traits [J]. Applied Psychological Measurement, 2013, 37(8): 619-637.

[55] De la Torre J, Song Hao. Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach [J]. Applied Measurement in Education, 2009, 33(8): 620-639.

[56] Huang Hung-Yu. A multilevel higher order item response theory model for measuring latent growth in longitudinal data [J]. Applied Psychological Measurement, 2015, 39(5): 362-372.

[57] Leighton J P, Gierl M J. Cognitive diagnostic assessment for education: Theory and applications [M]. New York: Cambridge University Press, 2007.

[58] Rupp A A, Templin J L, Henson R A. Diagnostic measurement: Theory, methods, and applications [M]. New York: The Guilford Press, 2010.

[59] Tatsuoka K K. Cognitive assessment: an introduction to the rule space method [M]. New York: Taylor & Francis Group, 2009.

[60] 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.

[61] 汪文义,宋丽红. 教育认知诊断评估理论与技术研究[M]. 北京:北京师范大学出版社,2015.

[62] de la Torre J. The generalized DINA model framework [J]. Psychometrika, 2011, 76(2): 179-199.

[63] 丁树良,汪文义,杨淑群. 认知诊断测验蓝图的设计[J]. 心理科学,2011,34(2):258-265.

[64] 丁树良,杨淑群,汪文义. 可达矩阵在认知诊断测验编制中的重要作用[J]. 江西师范大学学报:自然科学版,2010,34(5):490-494.

[65] Cui Ying, Gierl M J, Chang Hua-Hua. Estimating classification consistency and accuracy for cognitive diagnostic assessment [J]. Journal of Educational Measurement, 2012, 49(1): 19-38.

[66] Wang Wenyi, Song Lihong, Chen Ping, et al. Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment [J]. Journal of Educational Measurement, 2015, 52(4): 457-476.

[67] Xin Tao, Zhang Jiahui. Local equating of cognitively diagnostic modeled observed scores [J]. Applied Psychological Measurement, 2015, 39(1): 44-61.

[68] Xu Xueli, Chang Hua-Hua, Douglas J. A simulation study to compare CAT strategies for cognitive diagnosis [EB/OL]. 2003-04-18 [2016-05-07]. http://iacat.org/sites/default/files/biblio/xu03-01.pdf.

[69] Wang Chun, Zheng Chanjin, Chang Hua-Hua. An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing [J]. Journal of Educational Measurement, 2014, 51(4): 358-380.

[70] Liu Hongyun, You Xiaofeng, Wang Wenyi, et al. The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China [J]. Journal of Classification, 2013, 30(2): 152-172.

[71] Chen Ping, Xin Tao, Wang Chun, et al. On-line calibration methods for the DINA model with independent attributes in CA-CAT [J]. Psychometrika, 2012, 77(2): 201-222.

[72] Chang Hua-Hua. Making computerized adaptive testing diagnostic tools for schools [A]//Lissitz R W, Jiao H. Computers and their impact on state assessment: Recent history and predictions for the future [C]. Charlotte, NC: Information Age Publisher Inc, 2012: 195-226.

[73] Wang Chun, Chang Hua-Hua, Huebner A. Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing [J]. Journal of Educational

Measurement，2011，48(3)：255-273.

[74] Cheng Ying，Chang Hua-Hua. The modified maximum global discrimination index method for cognitive diagnostic computerized adaptive testing [EB/OL]. 2007-06-07 [2016-05-07]. http://iacat.org/sites/default/files/biblio/cat07cheng.pdf

[75] McGlohen M K，Chang Hua-Hua. Combining computer adaptive testing technology with cognitively diagnostic assessment [J]. Behavior Research Methods，2008，40(3)：808-821.

[76] McGlohen M K. The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment [M]. Austin: University of Texas at Austin，2004

[77] Junker B，Sijtsma K. Cognitive assessment models with few assumptions，and connections with nonparametric item response theory [J]. Applied Psychological Measurement，2001，25(3)：258-272.

[78] Haertel E H. Using restricted latent class models to map the skill structure of achievement items [J]. Journal of Educational Measurement，1989，26(4)：301-321.

[79] Tatsuoka K K，Toward an integration of item-response theory and cognitive error diagnosis [A] // Frederiksen N et al. Diagnostic Monitoring of Skill and Knowledge Acquisition [C]. Hillsdale，NJ: Erlbaum，1990: 453-488.

[80] Leighton J P，Gierl M J，Hunka S M. The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach [J]. Journal of Educational Measurement，2004，41(3)：205-237.

[81] 丁树良，罗芬，汪文义，等. 0-1 和多值可达矩阵的性质及应用 [J]. 江西师范大学学报: 自然科学版，2015，38(3)：265-269.

[82] 高俊平，张晖，赵旭剑，等. 面向维基百科的领域知识演化关系抽取 [J]. 计算机学报，2016(39)：1-15.

[83] Wang Chun. On latent trait estimation in multidimensional compensatory item response models [J]. Psychometrika，2015，80(2)：428-449.

[84] Segall D O. Multidimensional adaptive testing [J]. Psychometrika，1996，61(2)：331-354.

[85] Yao Lihua. Comparing theperformance of five multidimensional CAT selection procedures with different stopping rules [J]. Applied Psychological Measurement，2013，37(1)：3-23.

[86] Wang Chun，Chang Hua-Hua，Boughton K A. Kullback-Leibler information and its applications in multi-dimensional adaptive testing [J]. Psychometrika，2011，76(1)：13-39.

[87] van der Linden W J. Multidimensionaladaptive testing with a minimum error-variance criterion [J]. Journal of Educational and Behavioral Statistics，1999，24(4)：398-412.

[88] Huebner A，Wang C. A note on comparing examinee classification methods for cognitive diagnosis models [J]. Educational and Psychological Measurement，2011，71(2)：407-419.

[89] Cheng Ying. Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method [J]. Educational and Psychological Measurement，2010，70(6)：902-913.

[90] Cheng Ying. When cognitive diagnosis meets computerized adaptive testing: CD-CAT [J]. Psychometrika，2009，74(4)：619-632.

[91] Wang Chun. Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length [J]. Educational and Psychological Measurement，2013，73(6)：1017-1035.

[92] Kaplan M，De la Torre J，Barrada J R. New item selection methods for cognitive diagnosis computerized adaptive testing [J]. Applied Psychological Measurement，2015，39(3)：167-188.

[93] Wang Wenyi，Ding Shuliang，Song Lihong. New item-selection methods for balancing test efficiency against item-bank usage efficiency in CD-CAT [A] //Millsap RE. Quantitative psychology research: 78th annual meeting of the psychometric society [C] Heidelberg: Springer，2015: 133-151.

[94] Firth D. Bias reduction of maximum likelihood estimates [J]. Biometrika，1993，80(1)：27-38.

[95] Wang Shudong，Wang Tianyou. Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing [J]. Applied Psychological Measurement，2001，25(4)：317-331.

[96] Warm T A. Weighted likelihood estimation of ability in item response theory [J]. Psychometrika，1989，54(3)：427-450.

[97] Lord F M. Robbins-Monroprocedures for tailored testing [J]. Educational and Psychological Measurement，1971，31(1)：3-31.

[98] Chang Hua-Hua. Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond [A] // Kaplan D W. The SAGE handbook of quantitative methodology for the social sciences [C] Thousand Oaks，CA: Sage Publications，Inc.，2004.

[99] Chang Hua-Hua. Psychometrics behind computerized adaptive testing [J]. Psychometrika，2015，80(1)：1-20.

[100] Chang Hua-Hua，Ying Zhiliang. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests [J]. The Annals of Statistics，2009，37(3)：1466-1488.

[101] Robbins H，Monro S. A stochastic approximation method

[J]. The Annals of Mathematical Statistics, 1951, 22(3): 400-407.

[102] Chang Hua-Hua. The asymptotic posterior normality of the latent trait for polytomous IRT models [J]. Psychometrika, 1996, 61(3): 445-463.

[103] Chang Hua-Hua. The asymptotic posterior normality of the latent trait in an IRT model [J]. Psychometrika, 1993, 58(1): 37-52.

[104] Chang Hua-Hua, Qian Jiahe, Ying Zhiliang. A-stratified multistage computerized adaptive testing with b blocking [J]. Applied Psychological Measurement, 2001, 25(4): 333-341.

[105] Chang Hua-Hua, Ying Zhiliang. A-stratified multistage computerized adaptive testing [J]. Applied Psychological Measurement, 1999, 23(3): 211-222.

[106] Chang Hua-Hua, Ying Zhiliang. To weight or not to weight? Balancing influence of initial items in adaptive testing [J]. Psychometrika, 2008, 73(3): 441-450.

[107] Deng H, Ansley T, Chang Hua-Hua. Stratified and maximum information item selection procedures in computer adaptive testing [J]. Journal of Educational Measurement, 2010, 47(2): 202-226.

[108] Yi Qing, Chang Hua-Hua. a-stratified CAT design with content blocking [J]. British Journal of Mathematical and Statistical Psychology, 2003, 56(2): 359-378.

[109] Leung C K, Chang Hua-Hua, Hau K T. Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm [J]. Applied Psychological Measurement, 2002, 26(4): 376-392.

[110] Cheng Ying, Chang Hua-Hua, Douglas J, et al. Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints: balancing measurement efficiency and exposure control [J]. Educational and Psychological Measurement, 2009, 69(1): 35-49.

[111] Lee Y H, Ip E H, Fuh C D. A strategy for controlling item exposure in multidimensional computerized adaptive testing [J]. Educational and Psychological Measurement, 2007, 68(2): 215-232.

[112] Chang Hua-Hua, Ying Zhiliang. Aglobal information approach to computerized adaptive testing [J]. Applied Psychological Measurement, 1996, 20(3): 213-229.

[113] Su Yahui. A comparison of constrained item selection methods in multidimensional computerized adaptive testing [J]. Applied Psychological Measurement, 2016, 40(5): 346-360.

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

[115] Wang Shiyu, Zheng Yi, Zheng Chanjin, et al. An automated test assembly design for a large-scale chinese proficiency test [J]. Applied Psychological Measurement, 2016, 40(3): 233-237.

[116] Chen Peihua. Three-element item selection procedures for multiple forms assembly: An item matching approach [J]. Applied Psychological Measurement, 2015, 40(2): 114-127.

[117] Wang Chun, Chang Hua-Hua, Boughton K A. Deriving stopping rules for multidimensional computerized adaptive testing [J]. Applied Psychological Measurement, 2013, 37(2): 99-122.

[118] Chen Ping, Wang Chun. A new online calibration method for multidimensional computerized adaptive testing [J]. Psychometrika, 2015.

[119] Makransky G. An automatic online calibration design in adaptive testing [EB/OL]. [2016-05-07]. http://publicdocs.iacat.org/cat2010/cat09makransky.pdf.

[120] 涂冬波. 项目自动生成的小学儿童数学问题解决认知诊断 CAT 编制 [D]. 南昌: 江西师范大学, 2009.

[121] Liu Jinchen, Xu Gongjun, Ying Zhiliang. Data-driven learning of Q-matrix [J]. Applied Psychological Measurement, 2012, 36(7): 548-564.

[122] Liu Jinchen, Xu Gongjun, Ying Zhiliang. Theory of self-learning Q-matrix [J]. Bernoulli, 2013, 19(5A): 1790-1817.

[123] 陈平, 辛涛. 认知诊断计算机化自适应测验中在线标定方法的开发 [J]. 心理学报, 2011, 43(6): 710-724.

[124] 陈平, 辛涛. 认知诊断计算机化自适应测验中的项目增补 [J]. 心理学报, 2011, 43(7): 836-850.

[125] 汪文义. 认知诊断评估中项目属性辅助标定方法研究 [D]. 南昌: 江西师范大学, 2012.

[126] 汪文义, 丁树良, 游晓锋. 计算机化自适应诊断测验中原始题的属性标定 [J]. 心理学报, 2011, 43(8): 964-976.

[127] 喻晓锋, 罗照盛, 高椿雷, 等. 使用似然比 D^2 统计量的题目属性定义方法 [J]. 心理学报, 2015, 47(3): 417-426.

[128] 喻晓锋, 罗照盛, 秦春影, 等. 基于作答数据的模型参数和 Q 矩阵联合估计 [J]. 心理学报, 2015, 47(2): 273-282.

[129] Davey T, Nering M. Controlling item exposure and maintaining item security [A] // Mills C N. Computer-based testing: Building the foundation for future assessments [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2002, 165-192.

[130] Wang S, Fellouris G, Chang Hua-Hua. Computerized adaptive testing that allows for response revision: design

and asymptotic theory [J]. Statistica Sinica under Review arXiv: 1501.01366, 2015.

[31] Wilson K. Knewton technical white paper: The Knewton platform a general-purpose adaptive learning infrastructure [EB/OL]. [2016-05-07]. http://learn.knewton.com/technical-white-paper.

[32] 颜正恕，徐济惠. 线上线下一体化“互联网 + ”个性化教学模式研究 [J]. 中国职业技术教育，2016(5): 74-78.

[33] 张华华，王纯. 美国教育进展评估带给我们什么启示 [J]. 教育测量与评价，2010(2): 4-9.

[34] Duncan A. Address by the secretary of education at the 2009 governors education symposium: States will lead the way towards reform [EB/OL]. [2016-05-07]. http://www2.ed.gov/news/speeches/2009/06/06142009.pdf.

[35] 余胜泉. 大数据时代的教育创新 [EB/OL]. [2016-05-07]. http://www.wtoutiao.com/p/1f8O7Z3.html.

[36] Knewton. Knewton technology helped more Arizona State University students succeed [EB/OL]. [2016-05-07]. https://www.knewton.com/assets-v2/downloads/asu-case-study.pdf.

[37] Zhang Susu, Chang Hua-Hua. From smart testing to smart learning: how testing technology can assist the new generation of education [J]. International Journal of Smart Technology and Learning, 2016, 1(1): 67-92.

“Internet Pus” Measurement and Evaluation: A New Way for Adaptive Learning

CHANG Hua-Hua[1,2], WANG Wenyi[3*]

(1. Department of Psychdogy, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA;
2. Faculty of Education, East China Normal University, Shanghai 200062, China;
3. College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: Under the background of a top-level design of Internet plus 11 critical areas, after introducing traditional evaluation models, such as pencil and paper test, and the test mode based on test center, from the point of view of learning, the framework and design ideas about “Internet plus” measurement and evaluation were first proposed. Firstly, the concept of “Internet plus” was introduced briefly. Secondly, the latest measurement theory, methods, and techniques were elaborated. Then, how to incorporate cloud computing, big data, mobile Internet, and Internet of things into measurement and evaluation were discussed. The new evaluation mode could be used to extend applications of large-scale testing, provide personalized learning services, improve the quality of teaching, which contributes to the comprehensive development of education.

Key words: “Internet plus”; measurement and evaluation; adaptive testing; adaptive learning; assessment of education quality

(责任编辑: 冉小晓)