我国教育考试现代化面临的五个方面挑战

杨志明 陈一龙 徐庆树
（湖南师范大学，长沙 410081）

摘要：教育考试现代化是建设教育强国、推进教育现代化的重要组成部分。实现教育考试现代化面临诸多方面的挑战。结合国际考试行业的现代化经验和我国考试实践，着重讨论考试公平、理论创新、技术应用、行业标准与运作机制、基础学科建设与人才培养5个方面问题，提出应更新考试公平的认知观念以妥善应对公平性疑问、构建权威行业标准以确保考试质量、建立激励机制以推动理论创新和技术应用、加强心理计量学学科建设以促进考试专业人才培养等路径建议。

关键词：考试现代化；考试公平性；考试行业；考试技术；心理计量学

【中图分类号】G405 **【文献标识码】**A **【文章编号】**1005-8427(2023)02-0019-6

DOI: 10.19360/j.cnki.11-3303/g4.2023.02.003

党的二十大报告提出，以中国式现代化全面推进中华民族伟大复兴，实施科教兴国战略，强化现代化建设人才支撑[1]。《中国教育现代化2035》阐明了未来一段时期我国教育现代化建设的战略目标与实施路径。教育考试现代化是教育现代化的重要组成部分，推进教育考试现代化意义重大、影响深远。然而，就如何实现教育考试的现代化，各界人士从不同立场出发提出不同意见观点，尚难达成共识。一般说来，社会民众主要关注考试的公平性问题，教育工作者关心考试的人才选拔和教学导向问题，而考试工作者考虑更多的是考试安全性、规范性和可行性问题，信息技术专家对考试的数字化革新尤其感兴趣，考试公司则更在乎考试服务的投入与回报问题。这种情况下，如何准确把握考试现代化的内涵，清楚认识考试评价改革中的一些根本性问题，深入了解考试领域的发展趋势与面临的挑战，就显得十分重要。本文就如下5个方面问题进行讨论，以期为我国教育考试现代化路径探索提供思路建议。

1 考试公平观念需要拓宽视野

考试是现代社会中一种常用的人才评价手段，公平性是其基本属性。只要考试中存在不公平问题，则无论其现代化程度多高，它的存在价值都会受到质疑。在某种程度上，考试的存废取决于其公平与否；然而，目前国内关于考试公平性的理解和观念需要进一步拓宽视野，更新认识。

首先，把公平与效率作为对立面来讨论容易导致误解[2]。事实上，“分数面前人人平等”并不能解决“寒门难出贵子”的现实问题。不同社会

收稿日期：2023-01-06
基金项目：国家社会科学基金“十三五”规划2020年教育学重点课题“中西部地区推进高考综合改革研究”（AEA200013）
作者简介：杨志明（1963— ），男，湖南师范大学测评研究中心主任，教授；
陈一龙（1979— ），男，湖南师范大学外国语学院在读博士生；
徐庆树（1987— ），男，湖南师范大学外国语学院在读博士生。

阶层的学子所获得的教育资源本来就相去甚远，条件艰苦者若要取得与条件优良者相同的分数，往往需要加倍的付出，古今中外莫不如是。相反，开辟专项培养计划、实施多元录取通道反而能够更好地兼顾各阶层学子的升学需求，实现实质上的公平。在美国，一所常春藤高校从来不会从某一所超级中学招收几十个甚至上百名高中生，而是会给贫困地区学校留有一定招生指标，尽管来自贫困地区学生的考试分数可能要低很多。试想，如果我国所有一流高校的招生指标都能够分配到市县，并禁止其在任何一所超级中学招收超过20名（或某一限额）学生，则可能会更好地促进教育均衡发展及缓解“寒门难出贵子”问题，从而达到整体上的教育公平效果。至于多元录取通道中可能产生的徇私舞弊问题，则可以通过其他方式给予解决。因此，目前过于强调“分数面前人人平等”的考试公平观念实质上阻碍了考试改革创新和现代化推进，造成了只认卷面分数、只比拼高考总分的窘境。换句话说，过分强调公平与效率的矛盾很容易造成考试现代化改革中的畏手畏脚。这不仅可能影响考试理论的创新和技术的进步，还可能导致怯于尝试国际上成熟的考试技术等后果。

其次，仅做DIF分析满足不了公平性检测的需要[3]。从技术角度而言，考试公平是指在考试设计、考试命题、考试实施、阅卷评分、分数表达、结果使用等环节，均没有出现歧视某个特定群体的情况。检测考试公平性的国际通用算法，主要是DIF（differential item functioning）方法。其基本逻辑是，在考生能力水平相近的情况下，若某个弱势群体（如农村学生）在某道题目上的得分显著地低于强势群体（如城市学生）的得分，则说明这道试题存在DIF，即存在对弱势群体不公平问题，就需要在考试之前修改试题或在考后分数处理过程中删除这道试题得分。不过，国际流行的DIF分析方法并不能解决我国考试实践中遇到的公平性问题。例如，我国的高考科目设置方案是否存在着性别和城乡差异问题，就比较复杂，很难通过DIF分析来回答。有研究表明，采用“语数外+文综或理综”高考方案时，报考理工类的女生高考总分要比男生平均高出9.45~13.69分，报考文史类的女生高考总分要比男生平均高出61.86~73.24分[4]。那么，能够由此得出女生的知识和能力水平比男生普遍高的结论吗？文综/理综模式的高考总分方案是否对男生不利？类似结果在思维能力测试中也存在吗？采用这种高考方案会对社会产生哪些影响？这些问题都不是DIF分析所能回答的。

再次，杜绝舞弊才是考试公平的最大保障。在大规模高利害考试中，考试作弊（特别是有组织的团伙作弊）会对考试公平造成巨大危害。然而，由于我国许多高利害考试存在题量偏少、主观题偏多等特点，国外现成的作弊甄别方法很难使用。比如，在甄别答案抄袭方面所采用的错同率方法、Kappa方法、g2方法、K指数方法和ω方法等，以及在甄别考生作答反应与数据拟合度方面的KL散度和l_z指数方法等，在检测国内考试作弊时就很吃力。显然，有必要加大考试作弊的成本，加强考试舞弊甄别方法的深入研究，以减少作弊带来的考试不公平危害。

最后，考试服务机构需要及时公布公平性分析报告。目前，无论是大规模的高利害考试还是小范围的低利害考试，很少看到考试机构主动公布公平性分析报告，这显然不利于考试的现代化建设。解决好这个问题并不容易，需要在观念认识、政策规定、技术操作、人才储备、社会宣传等方面做大量工作。

2 考试理论创新需要关注我国国情特点

中国是考试的故乡。我国科举考试历经1 300年，积累了丰富的考试命题和组织管理经验。然而，这些实践经验一直没能上升到理论高

度,或许这也是导致科举考试被废止的原因之一。即便在当下,忽视考试理论或误解考试科学的情况也屡屡出现。例如,强行要求某次考试的通过率为90%的行政规定,以为取消考试或不使用分数而改用等级制就能破解“唯分数”,以及使用未经等值处理的题库制作试卷等,都是缺乏理论认知的表现。完全凭经验或不计成本的考试管理操作很容易重蹈科举考试的历史覆辙。具体来说,在理论创新方面存在以下一些挑战。

首先,由于国情不同,不少国际教育测量理论无法直接用于国内。纵观国际考试行业的发展历程,可以发现数学建模是考试现代化的必要条件之一,因为它决定了考试活动背后的理论基础和各种算法。缺乏理论依据或在理论上遇到瓶颈,考试工作的科学性就得不到保障。目前,最有影响的心理计量学(psychometric)理论主要有4类:经典测验理论(classical testing theory, CTT),概化理论(generalizability theory, GT),题目反应理论(item response theory, IRT),以及认知诊断模型(cognitive diagnostic models, CDMs)。其中:CTT在指导标准化考试方面最为有用,而GT在处理非标准化考试方面能更好地估算测量误差的多种来源;IRT具有题目和考生能力参数不变性特点,即题目和能力参数定义在考生总体而不依赖于考生样本,这为计算机化自适应测验(computerized adaptive testing,CAT)的发展提供了理论基础;CDMs则能够提供考试分数之外的关于考生认知发展特点的诊断报告。不过,这些国际流行的考试理论在我国遇到了水土不服问题。由于我国的大部分考试不做试测,不做常模和等值,只报告原始分数,这就使得大部分教育测量理论失去了用武之地。

其次,国外量表翻译或考务代理的做法存在重大隐患。目前,大量翻译或照搬国外测验项目的行为不仅存在版权纠纷等问题,更存在着科学性隐患。事实上,使用一些没有根据中国国情进行调整的考试与测评项目或缺乏中国常模的翻译量表,常常会得出一些骇人听闻的结论。比如,有人使用《贝利婴幼儿发展量表》测得“中国农村儿童认知发展滞后的比例高达63%”的结论[5],就明显有违常识。此外,仅仅停留在引进或参与国际著名考试项目,如PISA、PIRLS和TIMSS,离真正拥有自主知识产权的考试项目研发和理论与技术创新还存在相当差距。

最后,针对中国问题的考试理论亟待创新和建立。目前,亟待解决的考试理论创新问题有很多,包括:我国需要构建什么样的心理计量学模型,才能更好地拟合题量少、部分题目分值较高的考试数据;在不能提前试测的情况下,如何准确估计题目参数;在主观题分值较高情况下,如何进行测验等值;在缺乏测验常模和等值处理的条件下,如何解决考试分数的可比性问题;等等。这些问题都是我国考试现代化在理论建设方面所面临的挑战。

3 技术应用既要继承又要发展

自20世纪90年代开始,国际上先后出现了计算机化考试(computer-based testing,CBT)、在线考试(internet-based testing,IBT)、计算机化自适应考试(computerized adaptive testing, CAT)、计算机化自适应多阶段考试(computerized adaptive multi-stage testing、ca-MST或MST),以及计算机化自适应认知诊断考试(cognitive diagnostic computerized adaptive testing,cd-CAT)等考试数字化形态。对这些新模式,既要消化吸收,又要努力发展。

首先,要尽快消化和吸收相对成熟的考试技术。常规的考试技术主要包括命题技术、题目和能力参数估计方法、信度估计方法、效度证据收集与解读方法、阅卷评分技术、分数报告方法(包括常模研发技术、标准设定技术、测验等值技术等)、题库构建与管理技术、考试的实施与管理方法等。遗憾的是,目前国内不少考试项目仅停留

在试卷扫描、随机组卷、计算机施测等方面，对这些工作背后的算法问题很少有深入研究，不少成熟的考试技术（如测验等值）尚未得到充分利用。

其次，考试数字化水平亟待提高。目前，大多数的数字化考试主要采用CBT或IBT形式，即把纸笔测验变成了电子版本的测验。这些做法除了实现计算机化考试，还增加了考生信息管理、题目编写与修改、试卷编辑与制卷、计算机作答、计算机阅卷评分、简单的数据分析、结果报告和信息存储等功能[6]。对于CAT、ca-MST以及cd-CAT等模式，大多数考试服务机构则掌握得不够准确，而且与考试理论脱节的情况较为普遍。2022年底，笔者对我国长期从事考试科学研究的20位教授或考试服务企业总裁做过一个在线调查：在您的印象中，国内哪些公司实施了真正意义上的CAT或ca-MST？得到的回答几乎都是否定的。事实上，要实现数字化考试，面临着很高的技术门槛，至少需要解决好以下几个突出问题：如何满足考试的时间（单次或多次）和空间（机位数量）要求，并确保网络的流畅性与安全性；如何根据考试理论研发数字化考试平台；如何解决一年多考分数之间的可比性问题；实现cd-CAT考试的前提条件是什么；等等。

再次，使用国外成功技术需要做好本土化处理。由于我国不少重要考试大量使用主观题，且无法通过提前试测估计题目参数，因此，要实现CAT或ca-MST，许多技术都必须创新。特别地，诸如题库建设、参数估计与等值、在线标定（online calibration）、选题策略、终止策略、题目曝光度控制、多维度题目反应理论的模型选择及其参数估计方法等技术，目前使用起来还存在着很多困难。此外，若要使用cd-CAT等数字化考试模式，则需要对各大学科的必备知识、关键能力、学科素养和核心价值作出操作性界定，研发更加细致的学科知识和能力标准。显然，这项工作也极具挑战性。

最后，主观题的自动评分技术亟待优化。目前，主观题自动评分的方法主要有基于文体特征和文本内容2类：前者是提取作文的词汇复杂度、句子复杂度、句子长度、衔接和连贯程度以及作文文本与题目要求的关联程度等特征，并基于这些特征建模，通过机器学习（如回归方法、分类方法等）或者深度学习的方法（如CNN、LSTM、BERT等）训练模型进行评分；后者是基于文本对主题思想的阐释程度、文本内容知识性描述的准确程度等进行评分。不过，这方面的工作仍然面临3个挑战：一是评分模型很难像自然人一样开展评分思考，更无法向考生提供有效反馈；二是评分模型对文本思想性、启示性等维度的关注较少；三是存在泄露评分模型逻辑算法的风险，考生可能会针对评分模型进行套作，骗取高分。

4 考试运作要重视行业标准和创新激励机制建设

国外考试实行市场化运作机制，建立起了权威考试行业标准和“创新-回报”机制[7]。其中，行业标准的建立确保了考试产品的基本质量，而市场化运行机制不仅吸引了大量社会资源投入，也鼓励了考试理论创新和考试技术研发。

首先，考试行业的健康运行需要有权威性行业标准。美国教育研究协会（American Educational Research Association, AERA）、全美教育测量协会（National Council on Measurement in Education, NCME）和美国心理学会（American Psychological Association, APA）联合发布了《教育与心理测评标准》（Standards for Educational and Psychological Testing），从理论基础、测评研发和测评应用3个方面设置标准，对于规范美国考试行业的实践活动发挥了重要作用。这一成功做法值得借鉴，我国目前亟须建立一套符合中国国情的考试行业标准。

其次，建立“创新-回报”机制有利于促进考

试理论和技术创新。国外的考试机构多为非营利性组织,它们通过承接政府或教育机构的考试服务项目获得收益,再把收益投入到考试项目的研发和科研之中。专家学者们或者被高薪聘用到考试机构,或者把自己的创新成果卖给考试机构获得利益回报,从而形成了“创新-回报”的良性循环。这种关系类似于爱迪生与资本家J.P.摩根的关系,前者专注于电灯等的发明创造,后者专注于电网建设与电灯生产。资本家把创新成果及时转化为财富,并部分回报给发明家,从而保障了发明家专心于发明创造。假若国内专家们或考试机构的大量投入可以获得良好回报,则一定会出现大量拥有自主知识产权的考试项目。目前国内最大的问题是,考试机构或专家们辛苦研发的知识产品一直被以“白菜价”售卖,形成了考试几乎免费、试题随便抄用的市场窘况,忽视考试知识产权保护的后果,导致了技术创新乏力。这或许是目前许多考试机构不愿创新,仍停留在试卷扫描和简单CBT水平的深层次原因。

最后,有必要区别对待高利害考试与低利害考试。在当前形势下,国内许多高利害考试(如高考和公务员考试等)不宜走市场化道路,因为这些考试是国家政策的体现,是社会公正公平的象征,对社会有着重大的影响,必须确保万无一失。然而,大量低风险、低利害性考试项目则可以尝试市场化的运行方式。国家有关部门负责制定有关政策和颁发行业标准,扮演好考试行业的引导者和裁判员角色,鼓励非营利性考试机构提供专业化、社会化的服务。如此,不仅能保护考试行业的多样化运行,还可避免让政府相关部门不得不承担无限责任与风险的局面。

5 考试现代化需要有一流学科支撑

考试现代化需要有高水平的心理计量学作为基础,需要有相关的研究生培养项目,还需要有一流学术期刊支撑。否则,考试行业的现代化发展就会成为无源之水或空中楼阁。要解决考试评价领域的学术研究和人才培养问题,目前还面临着诸多挑战。

首先,要加强心理计量学的学科基础建设。国外的心理计量学学科一般包括3大课程模块:应用统计学,教育测量学,教育心理学、教育评价和研究方法。应用统计学模块有7门必修课程,包括:中等统计方法,相关与回归,实验设计,非参数统计方法,因素分析和结构方程模型,多元统计方法导论,教育统计与测量研究专题。教育测量学模块有8门必修课程,包括:评价工具的研发与应用,教育测量与评价,教育测量的理论与技术,量表标定(scaling)方法,题目反应理论,教育测量与评价专题研讨,教育考试中的等值与标定,概化理论。教育心理学、教育评价和研究方法模块有4门课程,包括:教育心理学,量化教育研究法,项目评估,教育心理学专题研讨。美国提供心理计量学博士课程的高校有斯坦福大学等十几所名校。相比于国外成熟发达的学科建设,我国目前仅有个别高校设置了较为完备的心理计量学课程,其他为数不多的高校仅设置了心理测量、教育评价、语言测试等与心理计量学关系密切的博士课程,而且每年测量学方向博士毕业生数量稀少。这与我国教育考试领域庞大的实践活动和用人需求形成了反差。由此,我国的心理计量学学科建设和人才培养亟待加强。

其次,要重视文理兼通的心理计量学人才培养。心理计量学是考试现代化的理论基石,这方面的人才要求文理兼通。其中,大量专业课程涉及数学和数理统计学内容,同时还需要具备教育心理学知识,最好还能具备计算机编程的能力,符合这样高标准要求的教师和学生为数甚少。而且,具备这种知识结构的人才很容易被金融行业挖走,很难留在考试行业。这无疑会影响考试专业化人才队伍建设。

最后,要建设一批高水平的学术刊物。国外

教育测量领域的一流学术期刊很多，如Psychometrika, Journal of Educational Measurement, Applied Psychological Measurement, Intelligence, Journal of Educational and Behavioral Statistics等，都拥有很高学术地位，影响广泛。而我国目前仅有《中国考试》《教育测量与评价》《教育与考试》和《考试研究》等少数学术期刊，尚无一家刊物进入CSSCI来源期刊目录（简称C刊），仅《中国考试》进入C刊扩展版。根据大多数高校规定博士毕业之前要在C刊发表1~3篇论文要求，我国心理计量学领域的博士研究生要么不研究考试评价问题，要么把论文发表到国外期刊上，这种情况阻碍了考试专业化人才培养。

总的来说，推进我国教育考试现代化进程中，尚有诸多问题值得深入思考，还有多方面的挑战需要应对。在实现考试现代化的道路上，不仅要确保考试的公平公正，还要建立创新激励机制，促进考试理论的发展和考试技术进步，更要通过相关学科建设与学术研究，大力培养高水平的人才队伍。

参考文献：

[1] 习近平：高举中国特色社会主义伟大旗帜 为全面建设社会主义现代化国家而团结奋斗：在中国共产党第二十次全国代表大会上的报告[EB/OL].（2022-10-25）[2022-12-22]. http://www.gov.cn/xinwen/2022-10/25/content_5721685.htm.

[2] 刘海峰. 高考改革：公平为首还是效率优先[J]. 高等教育研究, 2011, 32(5): 1-6.

[3] American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing[M]. Washington, DC: American Educational Research Association, 2014: 49-74.

[4] 杨志明, 李沛, 刘湘艺. 学业成就测试和高阶思维能力测试的性别差异分析[J]. 教育测量与评价, 2021(3): 3-10.

[5] 周群峰. 农村儿童发展现状调查：认知落后最高占比63%[N/OL]. 中国新闻周刊, 2017-07-07[2023-01-01]. http://country.cnr.cn/gundong/20170707/t20170707_523838837.shtm.

[6] 杨志明, 夏胜俊, 李希. 教育考试数字化：模式、特点与启示[J]. 教育测量与评价, 2022(6): 3-12.

[7] 杨志明. 海外考试服务的特点及其启示[J]. 教育测量与评价, 2016(9): 4-8.

Five Challenges in Educational Test Modernization in China

YANG Zhiming, CHEN Yilong, XU Qingshu

（Hunan Normal University, Changsha 410081, China）

Abstract: Educational test modernization plays an essential role in strengthening and modernizing China's education system. The path to test modernization, however, still need to be clarified. Considering the progress of international test modernization and the on-the-ground situation in China, we acknowledge both lessons and challenges regarding testing fairness, theoretical advancements, technology improvement, industry operating standards, and college psychometric curriculum development. Expanding and refreshing the cognitive concepts regarding testing fairness could enrich its studies. Establishing an "innovation-based return" rule may benefit the advancement of testing theories and technologies. Setting national academic standards for the testing industry could improve testing quality. Developing high-quality psychometric programs in colleges could increase the number of skilled psychometrician and research scientists for the testing industry.

Keywords: test modernization; test fairness; test industry; test technology; psychometrics

（责任编辑：王海东）