

美国大规模教育评价项目对中国公民科学素质调查的借鉴与启示

姚建欣

(北京师范大学物理学系, 北京 100875)

(莱布尼茨科学与数学教育研究所, 德国基尔 24118)

[摘要] 基于心理测量与教育测量的研究成果, 以 NAEP 为代表的美国大规模教育评价项目不断改进其评价方法、评价指标和评价结论报告。这些评价项目都已经在以项目反应理论 (IRT) 为基础对测量结果进行处理和分析。文章介绍项目反应理论在准确性、样本选取和信息量上的优势, 建议在科普研究所正在进行的“基于 Rasch 模型的试测分析研究”的基础上, 进一步开展项目反应理论的应用研究, 提升中国公民科学素质调查的研究水平。

[关键词] 科学素质调查 教育评价 项目反应理论

[中图分类号] C3 **[文献标识码]** A **[文章编号]** 1673-8357 (2015) 05-0062-05

Enlightenment from US's Large-Scale Educational Assessment Projects to China's National Survey for Public Scientific Literacy

Yao Jianxin

(Department of Physics, Beijing Normal University, Beijing 100875)

(Department of Physics Education, Leibniz Institute for Science and Mathematics Education, Kiel, Germany 24118)

Abstract: Based on the achievement of psychometrika and educational measurement, the analytical methods, indexes, and reports for US's large-scale educational assessment projects were improved. The Item Response Theory (IRT) was widely used in those projects. The advantage of IRT was introduced in order to offer some suggestions which can improve China's national survey for public scientific literacy.

Keywords: survey for public scientific literacy; educational measurement; Item Response Theory

CLC Numbers: C3 **Document Code:** A **Article ID:** 1673-8357 (2015) 05-0062-05

收稿日期: 2015-08-24

基金项目: 中国科协研究生科普研究能力提升类项目 (2013KPYJD44)。

作者简介: 姚建欣, 北京师范大学和德国莱布尼茨科学与数学教育研究所联合培养博士研究生, 研究方向为科学教育、物理教育, Email: yaojx@mail.bnu.edu.cn。

中国公民科学素质调查是研究我国公民科学素质的最主要手段,这些调查影响了相关政策的制定,也为各地的科普工作提供了指导^[1]。从1992年至2010年,中国已进行了8次全国公民科学素质调查。然而,何薇、徐善衍等学者在肯定中国公民科学素质调查的贡献的同时,还指出我国的公民科学素质调查存在进一步改进的空间。徐善衍教授提出,我国公众科学素质的监测评估应该能“了解总体状况,提出具体问题”^[2]。而如何发现具体问题,不仅需要理论工作的深入,还需要测量手段的精进。

NAEP(美国国家教育进展评估)、PISA、TIMSS等大规模教育评价项目与中国公民科学素质调查的规模相当。以NAEP为例,其结论报告对教育政策的制定、选民的投票倾向等都有相当程度的影响。每年国家成绩报告单公布的时候,CNN(美国有线电视新闻网)都会在黄金时段加以报道和评论。与其社会影响力相应,NAEP项目的承担者CES(美国教育统计委员会)在教育测量的研究和实践上,也一直走在世界的前列。其研究成果和实践经验被世界各国的学者和教育管理者参考借鉴。

源于以上背景,本文介绍NAEP项目的评估方法及优点。基于对其实践经验的分析,特别是项目反应理论的应用对评估结论的改善,为中国公民科学素质调查的改进提出具体建议。

1 NAEP项目简介

20世纪60年代,进行全国性的教育普查的想法就已经开始萌芽^[3]。时任美国教育部长的弗兰西斯·凯普尔(Francis Keppel)呼吁建立一个全国性的学生学业成就评估系统,以解决美国缺乏有关学生学业成就方面信息的状况。1969年实施了第一次全国范围的实验性调查,随着取样和统计分析技术的发展,到1996年,实验性的NAEP评价项目在公立

学校变为正式实施的常规评价项目。2001年《美国中小学教育法案》(即“不让一个孩子掉队”法案, *No Child Left Behind*)通过,确立了NAEP的评估模式。至此,NAEP成为美国唯一的一个全国范围且持续监测中小学生学习业成就的测量系统。NAEP对学校教育教学质量进行监控,促进成绩较差的学校和学区改进教学方式,对确保所有群体的学生都受到优质教育具有重要的作用。

2 NAEP项目特点介绍

NAEP的评价指标,维度设定和划分等诸多方面都有可借鉴之处,这里重点讨论其基于项目反应理论(Item Response Theory, IRT),在参数估计、矩阵抽样(matrix sampling)和变量预测上的进展。

从20世纪80年代起,NAEP的测试就已经基于项目反应理论进行试卷开发和数据统计。目前,NAEP数据分析的一般流程是:(1)根据测试数据进行项目参数估计;(2)基于项目参数构建模型,根据学生的答题情况和各种预测变量(如性别、年龄、种族等)估计学生的能力值;(3)采用垂直等值(vertical linking)的方法将各年级学生的最终成绩置于同一个量尺上,以便于各年级之间的直接比较;(4)得出学区、州成绩单和国家报告单。下面将进一步解释其具体的操作方案,及如何将其应用于科学素质调查。

2.1 项目参数估计

项目参数估计是对数据分析处理的第一步。项目反应理论将项目(item)作为测量被试能力水平的基本单位^①,通过项目参数来描述项目的属性,继而反映被试的特征和能力水平。图1是一条典型的项目特征曲线,包括三个计量学参数:(1)区分度参数(a参数),对应项目特征曲线的斜率,反映项目对被试能力水平的鉴别力,类似于经典测试理论中的区分度。(2)难度参数(b参数):对应项目特征曲线位于能力量尺上的位置,反

①对于一般的研究,英文的“item”一词实际上对应的是中文的“试题”,但是由于该理论早期引入时的翻译“项目”已经约定俗成,故沿用“项目”一词对应英文的“item”一词。

映项目的难度。(3) 猜测参数 (c 参数): 对应项目特征曲线左下渐进线的高度, 代表被试仅凭猜测答对项目的可能性。

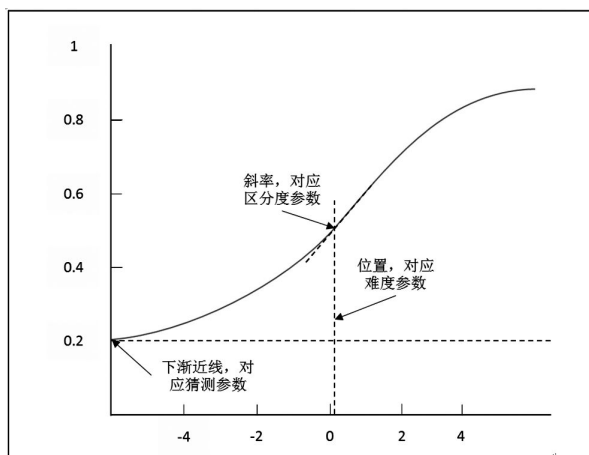


图 1 典型的项目特征曲线 (ICC 曲线)

通过对学生应答模式的统计分析, 对上述三个项目参数进行拟合。三参数模型的特征函数公式为^[4]:

基于拟合出的参数, 结合测试的维度划分和指标设计, 就可以进一步分析项目和被试的能力水平。对于中国公民科学素质调查, 参数拟合的具体算法并不是目前的关键。因为即使运行多年的 NAEP 项目, 目前也没有将精力过多的投入开发算法, 至今仍是采用商业软件 (如 BILOG 和 PARSCALE) 进行参数估计。所以本文也不再就算法进行讨论, 继续介绍采用项目反应理论带来的好处。

2.2 抽样设计

项目反应理论最明显的优势是被试能力参数与项目参数具有不变性^[5]。这为题库建设、组卷、抽样和施测带来了极大的便利。这也是 NAEP 项目坚持采用该理论作为测试设计的基础的原因。因为在传统的测试理论中, 为了保证测试的信效度, 必须保证较大的题量, 随之而来就是测试时间的增长, 这样就大大的增加了样本选择和施测的成本。基于项目反应理论进行矩阵抽样就可以解决这一问题。

这里以 NAEP 的数学评估为例, 介绍其解决方案。NAEP 的数学测试有五个维度 (数字属性与运算、测量、几何、数据分析与概率、代数)。如果每个维度有 20 道题, 则共需完成 100 道题。如此大的题量会给学生造成较大的负担和压力, 是全国普查所不能承受的。NAEP 采用矩阵取样解决这个问题。将其中题目抽出, 组成五套子试卷, 每套试卷 20 题, 则每个学生只需完成 20 题。然后, 基于项目反应理论的特点——被试能力参数与项目参数具有不变性, 就能进行统计推断, 得出整个样本的学生表现情况, 给出学区的学业成绩报告单^②。NAEP 项目十余年的实践证明, 这种方案对整体的描述可信、准确, 且节省测试成本, 取得了良好的效果。

对于 NAEP 本身, 虽然该方案效果良好且一直在采用, 实际上仍有由多维性等带来的一些技术问题正在研究和改进。但对于中国公民科学素质调查, 应用起来将更加可行。因为许多学者都指出: 由于测试结果的强相关, 在实际应用中米勒的三维指标可以逐渐向一维指标合并^[6]。这更好地符合了矩阵抽样的应用条件。

举例来解释如何在中国公民科学素质调查中进行一维指标测试的矩阵抽样。假定原本一整套问卷有 96 道试题 (64 个单选题和 32 个多选题)。若采用部分矩阵抽样设计, 则可以把所有的试题分出 7 个子集, 将同一层次样本的被试分成 7 个组群。如图 2 所示 (图中每个 X 代表 2 道题, 每个学生编号代表一组学生), 每个学生将完成一部分共同题 (24 道) 和子集题 (24 道)。然后再运行程序进行参数估计, 就可以拟合出总体样本的成绩水平。大量数据拟合研究和实证研究的结果都对这种“项目反应理论 + 矩阵抽样”的研究设计的精准度予以证明支持^[7]。这样一来, 假定原本需要约 90 分钟完成一整套问卷, 现在

②这里限于文章篇幅和侧重, 只描述了 NAEP 解决方案的流程, 没有解释如何用软件进行参数估计, 如何设计、组装平行卷等具体操作。因为操作软件进行统计分析并不困难, 项目反应理论的实际应用已经有几十年, 在这期间相关的商业软件间的竞争已经使得操作界面越发友好, 便于操作人员掌握和使用。

只需要 45 分钟即可完成。并且可以推断,由于答题时间的缩短,不认真作答、不完整作答的情况会大大减少。可以预期的结果是,不仅施测成本和施测难度下降,获得信息的精确度还可以上升。

学生 编号	单项选择题 (题数 N = 64)							多项选择题 (题数 N = 32)								
	矩阵子集							矩阵子集								
	共同题	1	2	3	4	5	6	7	共同题	1	2	3	4	5	6	7
1	XXXX	XXXX							XX	XX						
2	XXXX		XXXX						XX		XX					
3	XXXX			XXXX					XX			XX				
4	XXXX				XXXX				XX				XX			
5	XXXX					XXXX			XX					XX		
6	XXXX						XXXX		XX						XX	
7	XXXX							XXXX	XX							XX

图 2 部分矩阵抽样示意图

2.3 项目功能差异分析

完成项目参数估计后, NAEP 项目组就进行第二步工作: 基于项目参数构建模型, 根据学生的答题情况和各种预测变量(如性别、年龄、种族等)估计学生的能力值。对于中国公民科学素质调查, 则可以在完成项目参数估计后, 基于收集到的不同群体的项目参数构建模型, 然后根据学生的答题情况和各种预测变量(如性别、学历、科普活动经历等)对公民的科学素养进行预测。在初始阶段测量数据不足以建模的情况下, 进行项目功能差异分析(DIF)也能提供大量有用信息。

项目功能差异分析用以检验某一系列项目是否对某一个被试群体有某种偏好(有可能有利于某一群体, 也有可能不利于某一群体)。这一功能一方面可以用来筛选调查试题, 另一方面也可以为中国公民科学素质调查的后期分析提供便利。用来筛选调查试题很容易理解, 比如常见的将被试群体按性别分组。若分析结果显示在某些试题的作答上, 男性比女性有显著优势。这时就在提示我们仔细考虑, 产生这样的结果, 到底是因为男性在这一方面的确有优势, 还是因为这几个试题的问题情境偏向于男性被试。这对筛选修正试题, 提高测量信度有很大帮助。

进行项目功能差异分析对后期结论建议的得出也有重要帮助。比如将总体样本按每周的阅读时间分类, 分析科学素质的测量结果对不

同阅读时间的被试有何偏好。甚至还可以更具体一点, 考察对某一类群体(如 25 岁~35 岁, 月收入在 5 000 元)的科学素质水平, 哪些影响因素最为关键。当确认了这些因素与科学素质水平的相关关系, 这样就可以为科普工作的决策提供具体建议。还可以据此印发相关资料, 向公众普及科学素质的测量结果。这些资料不仅能让更多人关注了解科普工作, 还能让关注的人知道, 他是该去买几本书, 还是该抽空带着老人孩子一起逛逛科技馆。

3 启示和建议

数理统计理论在社会科学研究中的应用带来了更精细的研究方法和新广阔的分析视角。以 NAEP 等为代表的美国大规模教育测评项目已经率先迈出了关键的一步, 并取得了良好的社会效益和丰富的研究成果。参考借鉴他们的成功经验, 建构具有我国特色和国际领先的科学素质指标模型, 应成为我国公民科学素质调查的研究工作的重点之一。当然, 为达成这一目标, 还必须完成许多前期基础工作。

首先是相关专业人才的培养。美国在相关研究上的快速进展是以人才为基础的, 一个考试公司就有几百人上千人。大的考试机构, 如 ETS (美国教育考试服务中心), 有几千名员工, 其核心成员都是有着多年的心理测量、教育测量的研究背景, 拥有博士学位或是直接从高校引进而来。其中还不乏统计测量应用界的领军人物。中国是一个考试大国, 但实际上对考试进行专门研究的人员还不多。所以在物质投入增加的同时, 我们还需要重视智力资源储备。其中培养专业人才是重点, 特别是研究机构和高校对高水平人才的联合培养能极大地提高人才培养的效率和针对性。此外, 还可以引进国外的相关人才, 因为当前在美、澳等国有不少中国留学生修习统计相关的专业, 他们数理基础扎实且视野开阔, 而且相当比例的留学生愿意归国工作。培养和引进结合能较快地提升人才储备的数量和质量, 为进行高水平研究打下智力基础。

此外, 还要多进行各类小规模实验, 检

验新方案、新手段。现代的教育、心理测量研究是基于实证的，没有测量实验、没有实测数据就很难开展系统研究、取得理论突破。这方面中国科普研究所已经在进行创新和尝试。不仅在指标体系设定、内容维度划分上都提出了具有前瞻性的观点，而且已在利用多种数理统计方法进行后期分析和试题开发的尝试。

还需加强的一点是科普机构与高校、科研院所的交流合作。整体来看，高校在人才储备和理论基础上有优势，而科普机构在实践经验、

配套设施和资金上有优势。两者进行像如“馆校合作·科学教育”这样的合作研究，能实现优势互补，构成全方位、立体化的协同创新体系。

致谢 本文的写作思路，深受美国教育学会（AERA）院士、美国伊利诺伊大学香槟校区张华华（Hua-Hua Chang）教授所授的“教育与心理测量”课程内容的启发。文中部分关于NAEP的资料也参考了其讲稿。在此向其表示感谢。

参考文献

- [1] 何薇，张超，高宏斌. 中国公民科学素质调查现状思考[R]. 中国科普理论与实践探索——2009《全民科学素质行动计划纲要》论坛暨第十六届全国科普理论研讨会文集，2009.
- [2] 徐善衍. 关于我国公民科学素质调查工作的思考与建议[J]. 科普研究，2012(1): 19-22.
- [3] Institute of Education Sciences National Center for Education Statistics. NAEP History and Background[EB/OL]. [2013-06-01]. <http://nces.ed.gov/nationsreportcard/about/naephistory.aspx>.
- [4] 罗照盛. 项目反应理论基础[M]. 北京：北京师范大学出版社，2012.
- [5] Van Der Linden, Wim J., and Ronald K. Hambleton, eds. Handbook of Modern Item Response Theory [M]. New York: Springer, 1997.
- [6] 陈发俊，史玉民，徐飞. 美国米勒公民科学素养测评指标体系的形成与演变[J]. 科普研究，2009(2): 41-45.
- [7] 丁树良，罗芬，涂冬波，等. 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.

(编辑 张南茜)