

基础教育质量监测学科工具研发的思考

王燕春

(浙江省教育科学研究院浙江省中小学心理健康教育指导中心办公室,浙江杭州 310012)

摘要:基础教育质量监测学科工具的研发是一项系统而艰巨的任务。目前,关于基础教育质量监测学科工具的研发较少,对监测工具的研发程序、监测工具的质量指标研究的更少。本文借鉴PISA、TIMSS和NAEP三项国际测评经验,阐述了监测学科工具的特点,初步探讨了监测目标的确定、监测工具的设计、监测内容的确定、监测题目开发、监测工具的技术分析五大研发环节,明确了包括效度、信度、难度、区分度、猜测系数以及能力参数在内的六大监测工具的质量指标,并对监测工具的应用价值做了说明。

关键词:基础教育;质量监测;工具研发;双向细目表

中图分类号: G629.2, G40-058.1 **文献标识码:** A **文章编号:** 2095-3380(2017)02-0015-05

DOI: 10.13794/j.cnki.shjee.2017.0021

Reflections on the Research and Development of Tools Used in Elemental Education Quality Assessment

Wang Yanchun

(Zhejiang Primary and Middle School Mental Health Education Guidance Center Office,
Zhejiang Research Institute of Educational Science, Hangzhou, Zhejiang 310012)

Abstract: The research and development of tools used in the elemental education quality assessment is a systematic and arduous task. There is little research on development of these tools, and the research on the research and development process and the quality indicators are special lack. Therefore, by a summary of three international evaluation, PISA, TIMSS and NAEP, the paper expounds the characteristics of the evaluational subject, discusses the development of assessing tools. It includes five aspects: assessing objectives, assessing content, assessing subject development, assessing tool design, assessing tool technical analysis; and six quality indicators: validity, reliability, difficulty, the guess coefficients and the ability parameters. The paper also describes the assessment tool application value.

Keywords: Elemental education; Quality assessment; Tool development; Two-way specification table

学科工具的研发是实施基础教育质量监测的核心环节,其研发技术在很大程度上决定了监测工具的质量和监测目标的达成。学科工具的研发是一个多领

域专家共同参与、具有很高技术含量的工作,专业性极强,涉及心理与教育测量学、教育评价学、教育统计学、教育心理学等多个领域,然而目前我国在这些领

收稿日期 2017-01-05,修回日期 2017-03-29

域中的研究人员数量还相对较少,研究水平也有待提高。^[1]单独研发严谨、科学的监测工具的省份或地市比较少,可作借鉴的实践经验尤为缺乏。鉴于此,本文对学科工具研发模式、学科工具的特点、研发流程及工具的应用价值进行介绍,以求教于方家。

一、基础教育质量监测学科工具研发模式

基础教育质量监测工具研发借鉴了国际学生评价项目(The Programme for International Student Assessment,PISA)、国际数学与科学成就趋势研究(The Trends in International Mathematics and Science Study,TIMSS)、美国教育进展评议中心(National Assessment of Educational Progress,NAEP)测试等三大学生国际测评模式。^[2]

PISA 是世界经济合作与发展组织(OECD)的一项国际学生评价项目。其宗旨并不是筛选学生,而是为了改进教育政策,使教育政策的焦点从教育投入转向教育产出,评价 15 岁的学生在数字化的文化社会、成人生活的情境中所具备的知识和能力,是否能为终身学习奠定良好的基础。

TIMSS 是由国际教育成就评价协会(IEA)组织的对学生成绩进行跨国比较研究,其宗旨是评价各国或地区 4 年级和 8 年级的学生在数学和科学课程上的掌握情况,以促进全球数学和科学教学水平的提高。

NAEP 是美国目前最权威的学生学业成就评估体系,其评价宗旨是对学生总体学业成绩的普查,而不是对个人的诊断性评价,评价对象是随机抽样,评价科目根据不同年份抽取决定。NAEP 评价框架反映出当前教育发展的要求,同时又具有预测性和前瞻性,实现教学实践与研究成果之间的平衡。

综上,PISA 和 TIMSS 在评价形式与题目设计上有所相似之处,但侧重点有所区别。PISA 和 TIMSS 采用了相同的测试工具,并利用相似的心理测量学方法进行数据处理。同时它们还采用了趋势题目,在设计试题时保持不同轮次测评中部分题目相同,也就是设计一些锚题将各国学生在不同测试中的表现标定在同一量尺上,从而实现对其发展变化趋势的研究。它们之间不同之处在于 PISA 侧重于教育的产出,了解 15 岁年龄段的学生所具备的知识和技能,着重考查学生在真实生活情境中解决问题的能力,将能力作为测试的重要指标,而 TIMSS 侧重于各国课程中的共同部分,

表 1 三大国际评价项目比较

评估项目	目标	对象	内容	方式	周期
PISA	了解即将完成义务教育的各国初中学生是否具备未来生活所需的知识与技能的基本素养,并为终身学习奠定良好的基础	各参与国 15 岁学生	阅读素养、数学素养和科学素养	纸笔测试、背景问卷(4 类:学生、家长、学校和信息与通讯技术问卷)	每三年一次
TIMSS	了解学生对数学和科学课程的掌握情况,对各国数学教育和科学教育的情况进行比较研究	4 年级和 8 年级学生	数学与科学课程掌握情况	纸笔测试、背景问卷(学生、教师、学校和课程问卷)	每四年一次
NAEP	针对美国学生总体的学业成绩普查	4、8 和 12 年级的学生	各个学科的掌握情况	测试	每两年一次

从课程内容这个角度出发设计试题,考查学生的课程成绩。NAEP 和 TIMSS 整体上比较接近,都是基于对学生在校学习的课程效果的评估,但 NAEP 更侧重美国内部不同地区和学校间差异的比较(表 1)。^[3]

基础教育质量监测工具研发结合 PISA 的能力理念,借鉴 TIMSS 和 NAEP 对课程内容的重视,以“纸笔学科测试+背景问卷(学生、教师、班主任、管理干部、校长以及家庭问卷)”形式,使基础教育质量的现状及其影响因素得到客观、科学的反映,从而为教育教学质量的提升服务。

二、基础教育质量监测学科工具的特点

基础教育质量监测学科工具具有以学生能力为导向、锚题链接多个题本的特点。

(1) 以学生能力为导向。监测学科工具以考查达到监测年级的学生应该具备的能力为目标,全面监测学生的综合素质,着眼于学生群体的表现或整体的基础教育状况,从而避免了学生或教师想通过临时复习来提高个人成绩。

(2) 锚题链接各题本。监测目的是了解监测年级学生群体应该掌握的所有能力,可见监测包括了海量的内容。为了规避监测内容与参测学生有限的矛盾,采取锚题技术。选择多次测试比较稳定、难度区分度适当的高质量题目作为锚题,链接多个题本。让相邻的学生做不同的题本,由此既确保了监测内容的覆盖面,又不导致学生疲倦。

三、基础教育质量监测学科工具的研发程序

学科工具研发除了要组建一支热爱教育事业、经验丰富、具有一定教育教学理论基础且相对稳定的研发队伍外,最基础的工作就是要制定并遵循科学、规范的研发程序,这是确保研发工作按照既定工作目标顺利开展的前提。^[4]监测工具研发程序主要包括以下五大环节。

(1) 监测目标的确定。教育目标是教育工作的出发点,也是监测教育质量的基本依据,而课程标准是教育目标的细化。监测目标就是试图通过检测课程内容来反映学生所具有的能力。可见,监测目标来源于但不限于课程标准,它是在测试中可以检测到的那部分课程目标。学科工具包括内容与能力两个维度。

(2) 监测工具的设计。根据课程标准确定量化的行为指标体系,从而构建标准题库;依据监测框架编制题本细目表;根据题本细目表从标准题库中选题组卷。组卷时,需要明确监测实施形式、监测题目形式及具体题型;综合考虑监测工具题目量与监测时间,使学生的能力得到充分发挥而又不致于作答疲劳。确定监测题目的安排,一般根据题目形式与内容分类,从易到难排列。同时,要求锚题是测验长度的 $1/4$ 到 $1/3$ 之间。^[5]但对于超长测验,锚题比例依然采用这个比例的话,锚题的数量很大,会增加测验长度,对于题库建设,则会使锚题使用率太高,造成过度曝光,影响题库安全。Kolen和Brennan(2004)曾建议,经验法则是锚题量至少为测验总体量的20%。^[6]

(3) 监测内容的确定。研发团队集体研究监测学科的课程标准;分析监测学科所有版本的教材,找到监测年级各个版本教材的交集,作为研发题目的基础参考;借鉴PISA、TIMSS、NAEP等题目研发框架,从而确定该监测学科的内容维度、能力维度以及题目的比例。通过双向细目表的确定,充分体现监测的各项指标,包括题目的知识结构、能力结构、难度结构、题型结构、题量和分值分配结构等。双向细目表制订完成后,应严格按照细目表的要求进行题目研发;在研发题目过程中,也可依据实际需要对接目表进行微调和完善,但不宜随意进行大的改动。

(4) 监测题目的开发。题目研发之初,研发人员需要考虑题目的独立性、内容的代表性、表述的简明性、格式的规范性、创意的公平性、答案的确定性。研发人员提交的每一个题目均需要附题目清单(包括题型、

学科知识类别、认知能力、应用范围、背景、试题难度级别、预估答对率、评分标准、试题说明等信息),并集体逐一进行讨论,经过多次讨论修改后,进入预试阶段。

(5) 监测工具的技术分析。通过分层抽样,将监测工具在小范围内预试,分析工具的目的、内容覆盖率与权重、题目编制的规范性、与双向细目表的吻合度;根据学生的作答情况,对整套题本及每个题目的难度、区分度、猜测系数进行分析,并逐题修订题目、选项、答案及评分标准等;将修订后的工具再次预试,如有必要,需再修订、再预试,直到题目达到测量学质量指标。

四、基础教育质量监测学科工具的质量指标

监测工具的质量包含两个层面,一个是宏观层面,即监测题目的选配、组合是否符合监测目标,其指标主要为效度和信度;另一方面则是微观层面,即监测的题目质量,主要依靠题目分析来进行评价,其质量指标主要是题目的难度、区分度、猜测系数以及能力参数等。

1. 监测题本质量

监测题本的质量涉及效度和信度两个方面。效度是衡量监测有效性的质量指标,直接影响到决策的质量。效度是监测目标的达成程度,将效度作为监测工具的第一质量指标是合理的。效度以证据为基础,效度检验的过程就是积累和收集各种证据的过程。效度的证据主要包括监测内容、学生填答过程、监测题本的内部结构以及监测结果与其他变量之间的关系。信度是衡量多次监测结果一致性或可靠性程度的量化指标。由于监测本身的局限性,其结果除了受学生能力的影响之外,还受许多其他因素的影响,如测试的环境、学生个人特质、身体不适等。若监测的结果受到许多其他因素的影响,就不可能作为学生能力高低的标志,监测结果就不能做出正确、有效的解释,监测就没有很高的效度,从某种意义上说,信度是效度的前提。因此,将信度作为监测工具的指标是必要的。

2. 监测题目质量

由于监测工具研发是以项目反应理论为理论基础,在题目质量分析方面的优势非常明显。不但可以计算出试题的难度、区分度,还可以计算出学生的能力参数、题目信息函数,而且这些参数具有不变性。即试题的参数(难度、区分度及猜测系数)与学生的样本无关,学生的能力参数与所选用的题目无关。

(1) 难度。难度是指题目的难易程度,是体现题目与学生知识和能力水平相适合程度的量化指标。项目反应理论是以项目特征曲线的拐点处学生能力值作为试题的难度值,它的取值范一般为 $[-3,3]$,但在实际应用中取值范围在 $[0.2,0.8]$ 之间较好。题目难度与题型有关,不同类型题目的恰当难度如表2所示,^[7] 这些数据可供研发题目时参考。

表2 不同类型题目的恰当难度

题目类型	恰当难度
填空和短答	0.50
五选一选择题	0.70
四选一选择题	0.74
三选一选择题	0.75
是非题	0.85

经预试结果而得到的、基于数据的、客观的、实际的难度,为研发人员提供了反馈信息,是优化题目及题本重要参考依据。难度作为评价题目质量最重要、最基本的量化指标,不仅对学生的达标率有很大的影响,同时也直接影响题目的区分度、信度和效度,对基础教育教学也有很强的导向作用。

(2) 区分度。区分度是指题目对不同能力的学生的区分程度或鉴别能力。在项目反应理论中,题目的区分度与题目反应曲线拐点处的斜率有关。曲线越陡峭,区分度越高,当学生能力稍微偏离题目难度值时,学生正确做答题目的概率取值迅速增大或变小,表现出很强的区分作用。区分度的取值范围一般在 $(0,2]$ 之间,但在实际应用中,区分度的取值范围在 $[0.3,1.5]$ 之间较好。美国教育与心理测量学家艾伯(L. Ebel)提出了一个评价题目优劣的标准,^[8]如下表所示:

表3 评价题目优劣的标准

区分度指数	试题评价
0.4 以上	优
0.30-0.39	良好,修改会更佳
0.20-0.29	尚可,仍需修改
0.19 以下	劣,必须淘汰

区分度作为鉴别题目质量的重要指标之一,反映了题目对不同能力水平的学生的区分情况,作为修改或删除题目,调整题本的主要参考依据。在删除题目时,须结合双向细目表的整体构架,慎重确定是否删除。

(3) 猜测系数。猜测系数表示能力值很低的学生也能答对某题目的概率。猜测系数越大表示学生就越容易猜对该题,此时答对情况与学生能力水平高低无

关。在实际题目研发中,猜测系数越低越好,这样学生很难对该题进行猜测。一般来说,高质量的题目应有比较小的猜测系数,在理论上它的取值范围为 $[0,1]$,实际上多数小于理论上纯随机猜对概率,如四择一的选择題,猜测系数大于 0.25 的情况极少。

(4) 能力参数。在经典测量理论中,将学生的最后总得分作为学生能力的衡量指标,没有专门的能力参数。而在项目反应理论中,针对学生对题目的应答情况,利用项目反应模型来估计学生的能力参数值 θ ,估计的能力参数值是不随题目的不同而变化的。理论上 θ 的取值范围是 $(-\infty, +\infty)$,它是 $P(\theta)$ 与 θ 之间关系函数的定义域,但在实际应用中, θ 的取值范围是 $[-3, 3]$ 。一般来说,学生的总得分与能力参数总体趋势是一致的,但不完全相同。总分高的学生,其能力参数不一定高。^[9]

五、基础教育质量监测学科工具的应用价值

基础教育质量监测学科工具的应用价值在于客观反映全省(市)教育质量现状,科学诊断基础教育存在的问题和原因,为教育决策提供科学的依据。

1. 客观反映教育现状,科学诊断教育问题

监测学科工具以能力为导向,由锚题链接多个题本,覆盖了被监测学生应具备的知识和能力,客观反映该年级的学生群体的能力水平,反映了当前教育教学质量。结合配套的调查问卷提供的信息,可以科学诊断基础教育教学存在的问题及原因,为教师教学、学生学习、学校质量、教育发展指明方向。

2. 提供教育决策依据,促进教育质量提升

基于学科工具得到的教育大数据,将学科工具所获的数据与各调查问卷经过链接、清理、分析,将得到更加客观的结论,其中有些结论将彻底颠覆常理,此时不能武断地进行定论,还需结合省(市)的具体情况进行分析,从而得到更为合理、科学的参考信息,进而为制定更加贴近现实的发展规划提供数据支撑。科学有效的教育决策对教育质量的提升具有直接的效用。

3. 优化教育资源配置,注重教育公平

学科工具具有较强的明辨力,将不同水平的学校进行区分,客观反映不同地区学校水平的差异,找出此差异的原因,使教育资源得到合理、有效的配置,使教育资源的边际效益最大化,从而促进教育的均衡发展,对教育公平的落实无疑具有实际的推动作用。

4. 指引教育改革,推动学生全面发展

学科工具研发宗旨就是考查学生能力水平,尤其注重学生的创新能力和高层次思维能力,与传统的唯分数的教育观点截然不同,学科工具监测注重在当今数据时代下适应社会、运用知识的能力,关注学生个性发展、尊重学生人格完善,从而促进学生的全面发展。

参考文献

- [1] 董奇.构建具有中国特色的基础教育质量监测体系[J].人民教育,2007(13):2-3.
- [2] 姚霞.国家科学素养测评对我国科学学科测评的启示[J].考试研究,2013(2):53-63.
- [3] 占盛丽,文剑冰,朱小虎.全球化背景下PISA在美国基础教育质量评估体系中的贡献——基于美国的PISA与NEAP的比较[J].外国中小学教育,2010(5):1-6.
- [4] 王燕春.省级基础教育质量监测学科工具研发初探[J].时代教育,2016(12):33.
- [5] 熊建华,叶新蓉,丁树良,等.等值设计中锚题比例研究[A].The Third International Conference on Education Technology and Training[C].2010(8):197-200.
- [6] Kolen, M.J., Brennan, R.L. Test Equating, Scaling, and Linking: Methods and Practices [M]. New York: Springer Verlag,2004.
- [7] 刘新平,刘存侠.教育统计与测评导论[M].北京:科学出版社,2003:167;288.
- [8] 丁秀峰.心理测量学[M].开封:河南大学出版社,2001:115.
- [9] 王鹏.基于项目反应理论的自适应测试选题策略研究[D].沈阳:沈阳师范大学,2013:12.

(上接第14页)

辅人员等)对学生在校期间学习、生活等方面的需求给予关怀的指标等。针对学生作为产品属性而言,在服务质量体系中可以增加1个适用性维度,其指标主要为用人单位人才素质和能力的需求等指标。但基于此评价体系构建的高校教育服务质量评估体系是否适用尚需要进一步开展实证研究证实。

通过对当前的各种度量模型比较,结果显示不同的模型在应用中各有优缺点,其中模型(1)、模型(2)和模型(8)具有较好的应用前景。模型(1)的结果可用于服务质量问题的诊断^[8];模型(2)的结果可用于指导服务质量改进的优先行动方案制定^[13];而采用模型(8)的结果对总体服务质量进行评价具有更高的准确性。

参考文献

- [1] Parasuraman A, Zeithaml V A, Berry L L. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality [J]. Journal of Retailing, 1988(1):12-40.
- [2] ShankMD, Walker M, Hayes T. Understanding Professional Service Expectation: Do We Know What Our Students Expect in a Quality Education [J]. Journal of Professional Services Marketing, 1995(1):71-89.
- [3] 马万民.高等教育服务质量管理研究[M].上海:上海交通大学出版社,2005:54.
- [4] 余天佐,韩映雄. SERVQUAL 在高等教育服务质量评价中的应用研究述评[J].现代大学教育,2010(6):59-63.
- [5] Cronin J J, Taylor S A. Measuring Service Quality: A Reexamination and Extension [J]. Journal of Marketing, 1992(3):55-68.
- [6] Jain S K, Gupta G. Measuring Service Quality: SERVQUAL vs. SERVPERF Scales [J]. Vikalpa, 2004(2):25-37.
- [7] 吴学梅,张少文.基于SERVQUAL的高校教育服务质量评价研究——以泰州某本科高校为例[J].重庆高教研究,2013(5):19-23.
- [8] 朱国锋,兰杏芳,方旭华.差距理论在高职教育服务质量测评中的应用[J].浙江交通职业技术学院学报,2003(3):46-50.
- [9] 王齐(女).高校教育服务质量感知的定量研究[J].高等工程教育研究,2003(6):20-23.
- [10] 宋彦军.高职教育服务感知质量的测评[J].统计与决策,2009(10):53-54.
- [11] 王雪亘.基于顾客期望与感知的技师学院教育服务质量评价研究[J].机械职业教育,2013(12):3-5.
- [12] Pena, D., A methodology for building service quality indice [J]. Annual Quality Congress Proceedings, 1999(53):550-558.
- [13] 唐宇宏,唐旭琴,潘鸿.基于SERVQUAL模型的专业教学服务质量评价探讨[J].重庆医学,2016(34):4869-4871.
- [14] 孙美荣,汪雅霜.独立学院教育服务质量实证研究[J].上海教育评估研究,2016(5):14-18.