基 于 标 准 的 评 价 研 究 丛 书

总主编◎崔允漷

A Study on the Theory and Technology of Development for Open-ended Test Items

开放题编制的理论与技术研究

著 张雨强 冯翠典

华东师范大学出版社

基于标准的评价研究丛书

总主编◎崔允漷

A Study on the Theory and Technology of Development for Open-ended Test Items

开放题编制的理论与技术研究

著　张雨强　冯翠典

华东师范大学出版社

图书在版编目（CIP）数据

开放题编制的理论与技术研究／崔允漷主编，张雨强，冯翠典著. 上海：华东师范大学出版社，2009
（基于标准的评价研究丛书）
ISBN 978－7－5617－6932－4

Ⅰ. 开… Ⅱ. ①张…②冯… Ⅲ. 试题—编制—研究
Ⅳ. G424.79

中国版本图书馆 CIP 数据核字（2009）第 018839 号

基于标准的评价研究丛书

开放题编制的理论与技术研究

主　　编　崔允漷
著　　者　张雨强　冯翠典
责任编辑　周志凤
审读编辑　方　强
责任校对　邱红穗
装帧设计　卢晓红

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
电话总机　021－62450163 转各部门　行政传真 021－62572105
客服电话　021－62865537（兼传真）
门市（邮购）电话　021－62869887
门市地址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　址　www. ecnupress. com. cn

印 刷 者　江苏通州市印刷有限公司
开　　本　787 ×1092　16 开
印　　张　19.5
字　　数　329 千字
版　　次　2009 年 4 月第 1 版
印　　次　2009 年 4 月第 1 次
印　　数　3 100
书　　号　ISBN 978－7－5617－6932－4/G · 3872
定　　价　39.00 元

出 版 人　朱杰人

教育部人文社会科学研究青年基金项目
《开放性学业成就评价与课程的匹配度研究》成果之一
教育部普通高等学校人文社会科学重点研究基地
华东师范大学课程与教学研究所研究成果

华东师范大学“985 工程”二期哲学社会科学
“教师教育理论与实践”创新基地建设成果之一
上海文化发展基金图书出版专项基金

迎接教育评价的新范式——代总序

崔允漷

半个世纪以来，教育评价的理论领域也许没有出现诸如布卢姆的教育目标分类学之类的重大成果，但教育评价的实践领域却发生了巨大的变革。这种变革源于知识观、学习观的变化，也与社会发展和教育发展目标的变化有关。从评价实践的变革中，人们似乎可以看到，教育评价实践领域正在发生一个范式转换，而在学生学业成就评价领域中，这种范式转换似乎更为明显。

一

教育评价，特别是学生学业成就评价领域在近几十年中正在发生着巨大的变革，这种巨大的变革是教育评价历史上从未有过的，具体表现在以下几个方面。

（一）“对学习的评价”依然受关注，但“为学习的评价”逐渐成为主流

为什么要进行学生学业成就评价？美国学者洛克希德（Lockheed，M. E.）认为当前学生学业成就评价有六个最普遍的目标：为高一级的教育选拔学生；认证学生的成就；监测成就变化的趋向；评价特定的教育项目和政策；促使学校、学区对

学生成就负责；诊断个体的学习需要。① 在这六个方面的目标中，有些是评价最原始的目标，如选拔，有些则是近几十年来才出现的目标，如监测、政策评价和问责。但除了诊断学生个体的学习需要外，为其余五个目标实施的评价基本上都属于“对学习的评价”（assessment of learning）。

在当前的学生学业成就评价实践中，认证性和选拔性评价依然具有非常重要的地位，因为个体进入社会生活需要获得相应的学业成就水平的证明，高一级教育资源和社会资源相对有限，需要相对公平的分配机制；监测性评价则源于国家对教育质量的责任和对有关教育质量的信息需求，近年来得到很大发展，这不仅表现在许多国家层面的教育质量监测体系的建立，也表现在众多国家参与国际性学生学业成就评价项目的热情；用于政策或项目评价的学生学业成就评价因为科学决策的要求而发展；学生学业成就评价同样被广泛地作为对地方、学校、教师和学生个体进行问责，并促使其对自己的职责承担责任的工具。

这里，我们可以看到，“对学习的评价”不仅没有削弱，反而正在得到加强。但是，当前教育评价领域中关注最多的还是“为学习的评价”（assessment for learning）。实际上，对“为学习的评价”的关注在 20 世纪 60 年代就已经开始。布卢姆的教育目标分类学表明教育者开始清楚地表达对这种评价的需求——专门设计教育目标并将该目标用于计划、教学、学习和评价这一循环圈中。② 这种教育评价是一种支持学习的评价模型，它关注相对于自己而非他人的个体成就；检测能力而非智力；发生于相对不受控制的情境中，因此不能产生“人人都可通用的”（well-behaved）数据；寻求最好的而不是典型的表现；如果不考虑作为标准化测验的特征的规则和限制，它是最有效的；体现着一种建设性的评价观，目的在于帮助而非惩罚学生。③ 而格拉瑟（Glaser）将标准参照测验和常模参照测验区分开来，就是一种将教育评价从经典的心理测量学中分离出来的重要尝试。

当前，对学习的关注已经成为教育评价改革的一个大观念（big idea）。教学、学习和评价三位一体的关系得以建立，评价被看成镶嵌于教—学过程之中的一个

---

① Lockheed, M. E.. Assessment and management: World Bank support for educational testing[C]//Little, A. and Wolf, A.. *Assessment in transition: learning, monitoring and selection in international perspective*. Pergamon Press, 1996: 29-30.

② Gipps, C.. Assessment for learning [C]//Little, A. &A. Wolf(eds.). *Assessment in Transition: learning, monitoring and selection in international perspective*. Pergamon Press, 1996: 254.

③ Gipps, C.. Assessment for learning[C]//Little, A. &A. Wolf(eds.). *Assessment in Transition: learning, monitoring and selection in international perspective*. Pergamon Press, 1996: 255.

成分。对各种新型评价方式的倡导,内部评价尤其是课堂层面的评价得到高度关注,对多元评价尤其是学生参与评价的倡导,对评价结果的适当运用的规范,等等,无不反映着“为学习的评价”的理念。即使在为监测、问责等目的而实施的学生学业成就评价中,促进学生的学习同样是一个重要的关注点。

(二)评价管理体制变革明显,平衡的学生学业成就评价体系正在形成

就传统而言,世界各国的教育管理体制几乎都可以归入两大阵营:集权和分权。这种传统的力量非常强大,经常在教育政策的制定过程中扮演着极为重要的角色。近几十年来,世界各国的教育管理体制正在发生急剧的变革。从表面上看,两种不同传统的教育管理体制似乎正在走向其对立面,但不同的路径指向的却是同一目标,即集权和分权之间适当的平衡。

在整个教育管理体制的变革中,评价管理权起着一个核心的作用——考试控制权的回收或下放被当作促成教育管理体制变革的强有力的杠杆。

在美国,《国家在危急之中》发布后,几乎每个州都采取了自上而下的旨在提高标准的努力,如提高对学术课程的要求,强化对教材的控制,州课程指南的运用等,但最为普遍的还是针对所有年级的全州考试,试图以此来重塑学校教育实践。联邦政府甚至每年出版排名表(Wall Chart),按照标准化测验的成绩对各州进行排名,而各州也将这种年度活动复杂化,将学区的排名公开化,而学区的管理层又进一步将学区内学校的排名公开化。在英国,1988 年教育法案中一个基本思路就是削减地方教育当局的教育权力和课程上的自由裁量权,将教育控制权由地方收归中央。政府特别相信考试控制权的回收具有将教育管理权从地方收归中央的能力,因此,全国性的考试系统建立起来了,学生必须定期参加考试,考试结果要公开发布。

而一些传统上高度集权的国家,考试系统的控制却在走向相反的方向。在法国,考试权力传统上高度集中于中央政府,从 19 世纪初开始中央政府就是全国考试所测量的标准的守卫人,考试系统就是中央政府小心保护的主要教育特权之一。但从 1980 年代早期开始,教育管理中实施了一种温和的权力下放政策,地方获得了更多的权力。

事实上,上述两种表面上看来完全相反的趋势本质上却是一致的。教育管理

上的分权趋向更多是从教育输出方面考虑的。基于对传统的输入或过程评估模式的反思,西方开始更多地从结果方面考虑教育的质量;而要使学校教育获得良好的结果,学校就必须获得相当程度的自主权,传统的自上而下管制的方法只能解决输入的问题,对于学校教育的真正改善效果有限。学校要改善,就必须获得相关的信息,因此分权发生了。但国家将教育权力赋予地方、学校层面并不意味着国家不需要相关的信息,实际上,所有层次的管理者都需要信息来改善其决策的效能。因此,保证分权和监控之间的适当张力——决策上的去中心化(过程规制与专业主义控制)与基于结果的问责——就成为几乎所有国家的共同选择,以实现国家规制和地方控制上的平衡。①在评价事务上,就是国家评价与地方、学校层面评价的平衡。

从一个国家内部看,评价的权力似乎也正在向两头延伸。一方面,国家层面的教育质量监测成为许多国家的现实做法和追求,②试图建立国家教育质量监测体系来监测预设的标准的达成状况,对地方、学校进行问责,进而保证教育目标的实现。而另一方面,人们又普遍相信,学习最终发生于课堂之中,当现实中教师将其三分之一到一半的教学时间用于评价及与评价相关的事务上时,如果评价不能在日常的课堂实践中有效地运行,那么其他层面(学区、州、国家和国际)上的评价完全是浪费时间和金钱。③ 因此,内部评价,特别是课堂层面的评价,也就成为当前评价改革的一个核心关注点。有些国家中教师自己实施的评价甚至被当作学生学业成就评价的核心成分,作为选拔、认证乃至问责的依据。如在英国的全国评价项目中,教师评价是最终评价的一个正式的成分。在澳大利亚的昆士兰,20 年来没有基于学科的公开考试,大部分学生评价来自于改善后的教师评价。

(三)传统的考试一统天下的地位被颠覆,诸多新型的评价方式得到广泛运用

在传统的学生学业成就评价中,其实并不缺少口试、对实际表现的观察之类的评价方式。可是,受到 20 世纪初心理测量在军事领域中的成功的诱惑和刺激,美

---

① Olmedilla, Juan M.. Tradition and change in national examination systems: A comparison of Mediterranean and Anglo-Saxon countries [C]//Ecksteino, M. A. & Noah, H. J.. *Examinations: Comparative and International Studies*. Pergamon Press,1992: 142.

② 参见《基于标准的学生学业成就评价》(崔允漷等主编,华东师范大学出版社2008 年版)的第十四章.

③ 转引自: Roschewski, P., Gallaher, C., Isernhagen, J.. Nebraskans research for the STARS[J]. *Phi Delta Kappan*, 2001(8): 611-615.

国常春藤联合会的入学考试机构——大学入学考试委员会(College Entrance Examinations Board, CEEB),在1926年采用难度更高的"军队阿尔法",编制了学术性向测验(Scholastic Aptitude Test, SAT)。这种标准化测验因为基于心理测量学而披上了"科学"的外衣,这使得它在科学主义甚嚣尘上的年代极具诱惑力。很快,这种基于心理测量学的考试取代了原本运用论文式考试的选拔性评价,而且发展成为学校教育系统中各种评价的主要形式。尽管后来有许多测验宣称测量"成就"而不是"性向",但依然采用SAT的形式,或者就是心理测量的形式——这种考试获得了在教育评价领域中一统天下的地位。

然而,在这样一种考试模式获得其统治地位的同时,所引发的批评和质疑也不绝于耳。SAT的倡导者布里格汉姆(Brigham, C.)就已经预见到标准化测验可能带来的消极后果。近年来,当知识观、学习观发生变化,尤其是建构主义的学习理论兴起之时,在美国,这种批评"多年来局限于进步的批评家和学术传统主义者,如今却经常出现在《华盛顿邮报》、《纽约时报》、《华尔街杂志》、《新闻周刊》上,甚至出现在黄金时段的电视节目上"①。而在我国,这种批评同样不再局限于学术媒体,而频繁地出现在《羊城晚报》、《北京文艺》、《中国青年》、《南方周末》之类的大众媒体上。批评针对的是这种考试模式更多关注结果的可比较性和公平性,很少考虑对学习的加强和支持。它鼓励学生对事实性知识的掌握,鼓励再生他人的观点,激励"肤浅的学习",不能导致对"高等级的思考技能"的学习;当考试具有高利害关系时,教师常被鼓励去追求更高的分数,"为不适当的考而教",而不是去更好地理解学生学习上的困难;一些消极的甚至不合伦理的实践就成为学生学习中的常态。在这种情况下,最完美、最有效的考试却导致最糟糕的学习。②

在这种背景下,标准化测验的统治地位受到了猛烈的冲击,尽管其市场依然巨大,但诸多冠以"表现性评价"(Performance Assessment)、"真实性评价"(Authentic Assessment)或者"另类评价"(Alternative Assessment)之名的新型评价方式正在成为众多评价项目的重要方法,甚至在诸如监测、问责、升学之类传统上由大规模的标准化测验所控制的领域发挥作用。这些新型评价关注高层次学习所要求的批判性思考和知识整合,要求评价任务本身是技能或学习目标的真实例子,而不是替代物,期望学生通过思考生成答案而不是在多个选项中选出正确答案。

---

① Berlak, H.. The need for a new science of assessment[C]//Berlak, H. & Newmann, F. M.. *Toward a new science of educational testing and assessment*. State University of New York Press, 1992: 8.

② 王少非. 校内考试监控研究[D]. 上海: 华东师范大学课程与教学研究所博士学位论文, 2007: 13.

比如表现性评价，目的在于测量学习者运用先前所获得的知识解决新异的问题或完成具体任务的能力，常常运用真实的生活或模拟的评价练习来引发最初的反应。① 表现性评价不仅评价学生“知道什么”，更重要的是评价学生“能做什么”；不仅评价学生行为表现的结果，更评价学生行为表现的过程；不仅评价学生在某个学习领域、某方面的能力，而且评价学生综合运用已有知识进行实践与表现的能力。在这些方面，表现性评价所能做的正是选择式、记忆式的纸笔考试所无法企及的。正因如此，表现性评价开始成为诸多学生学业成就评价项目中的重要评价方式。如英国各学科的 A 级考试中，借助于表现性评价的“中心评审课程作业”普遍占 20% —35% 的分数比重。② 又如，我国香港地区，名为“教师评审制”的表现性评价的分数在高考中占到了 20% —30% 的比例。③ 另外，国际教育成就评价协会（International Association for the Evaluation of Educational Achievement，IEA）在第三次国际数学和科学教育研究（Third International Mathematics and Science Study，TIMSS）中对表现性评价的运用也为大规模进行表现性评价提供了范例。④

（四）教育评价的心理测量学基础被动摇，新的教育评价文化正在兴起

相对于以往随意化的评价，心理测量学成为教育评价的基础无疑是教育评价发展史上的一个里程碑。然而，当评价的功能发生变化，转向对学习的促进时，教育评价的心理测量学基础就不可避免地受到质疑。

从根本上讲，心理测量学的诸多假定都来源于关于测验目的的假定。借用现代科学主义的话语，心理测量学将其所发明的测验称为“工具”，而且是一种外在于历史与文化的，不受感情或价值观影响的、公正无偏的科学的工具，这种工具的根本功能被假定为“选拔”，进而“安置”，即对个体或群体进行区分，然后将之归到被认为适当的位置上。区分的根据就是个体身上那种稳定的不变的东西，因此，基于心理测量学的测验就只能测量人类的少数特性，通常是那些不受教育影响的特

① Stiggins，Richard J.. Design and Development of Performance Assessments[J]. *Educational Measurement: Issues and Practice*. 1987(6).

② 冯生尧，谢瑶妮. 英国高考中的表现性评价：中心评审课程作业[J]. 比较教育研究，2006(8).

③ 冯生尧. 香港高考教师评审制的特点和启示[J]. 课程·教材·教法，2004(10).

④ 周文叶. 论表现性评价在综合素质评价中的运用[J]. 全球教育展望，2007(10).

性，也就是智力或自然倾向。测验就是要测出个体到底有“多少”这样的特性，而不关注个体在这些方面的表现有“多好”。当信度和常模成为教育评价的核心关注点时，教育评价就不再关注个体，而是关注个体与他人（常模）的比较，这导致学生在教育评价中的被动地位和无力感，因为他们能决定自己的成绩，但不能影响他人的成绩；同样，这种关注使得对统计分析的适合性成为教育评价（包括考试）设计的重要关怀，而对于评价在课堂中的意义，对于评价在促进学生学习和提高学业成就方面的意义，则基本上没有得到关注。而在教育评价中，所应当评价的东西与心理测量学期望测量的东西有本质的不同，学生学业成就显然是教育的结果，而不是不受教育影响的固有的不变的特质——相对于智力和自然倾向，换言之，作为教学的直接结果的成绩是“脏”的，它直接受到教学和教师的影响。就此而言，那种基于心理测量学的、看起来非常成熟的技术标准不能完全适合指向于不同目的、需要不同方法的教育评价。

而在柏拉克（Berlak，H.）看来，心理测量学范式中测验不只是不适合教育评价的问题，“植根于一个不合时宜的范式之中的标准化和标准参照测验阻碍了学校的更新和重构。当我们进入20世纪的最后十年时，至少对于那些外在于测验编制的人而言，标准化和大部分标准参照测验所基于的假定是明显站不住脚的。在这一范式的废墟之外，一种新的范式正从许多并不完美的解决教育成就评价的实践问题的努力中缓慢地出现……”①

有人将正在出现的新的评价范式描述为“评价文化”。② 它强调真实的情境化的测验，强调运用多元评价，强调对高层次技能而不是知识的再生产的评价；不仅关注对认知的评价，而且包括对元认知、情感和社会维度以及心理动力技能的评价；关注将评价整合到学习之中；并使学生越来越多地承担评价过程中的责任；“对学习的评价”与“为学习的评价”的整合。在这种评价文化中，传统的智慧正被超越，新的智慧正在出现：

教学智慧——关注学习；

学习智慧——反思性的、主动的知识建构；

---

① Berlak，H.. The need for a new science of assessment[C]//Berlak，H.，et al.. *Toward a New Science of Educational Testing and Assessment*. State University of New York Press，1992：12.

② Birenbaum，M.. New insight into learning and teaching and their implications for assessment[C]// Segers，M.，Dochy，F. &Cascallar，E.（eds.）. *Optimising new modes of assessment：In search of qualities and standards*. Kluwer Academic Publishers，2003：15.

评价智慧——情境化的、解释性的、基于表现的。

二

柏拉克将教育评价中心理测量学基础的动摇看成教育评价“范式转换”的一个环节。的确，基于心理测量学的教育评价不能解决现实的教育评价中的一些“例外”，因而竞争性理论或实践模式出现并尝试排挤它的情况就不可避免。就此而言，柏拉克的结论完全正确。

可是，柏拉克的视野也许狭窄了一些。当前教育评价中的范式转换也许不仅仅是因为以心理测量学为教育评价基础的观念被动摇，更是因为教育评价领域诸多信念、原理、实践方式的变化——在上一部分的粗略描述中，我们已经能够获得关于这种变革的一个图景。正是这一些变革使得教育评价真正发生了更大的范式转换。

范式（Paradigm）和范式转换（Paradigm shift）是美国科学史家托马斯·库恩（Tomas Kuhn）在 1962 年出版的《科学革命的结构》（*The Structure of Scientific Revolutions*）一书中提出的概念。如今，尽管这两个远非清晰的概念（库恩本人也承认他对“范式”一词的使用出现了“弹性”），却被广泛地套用到创新的乃至传统的领域。库恩没有对范式下过定义。他在《科学革命的结构》中给出了“范式”的 21 种用法，大致上可分三个层面：（1）哲学层面的范式，如：信念、有效的思维方式、标准、公认的看法、条理化规则等。（2）社会学层面的范式，如：公认的科学成就、具体科学成就、一套科学习惯、一套政治制度、司法裁决等。（3）人工科学层面的范式，如：教科书或经典著作、工具、仪器、类比、格式塔图像等。但总体来说，一个科学的范式就是一套关于现实的假设，以及说明它所面对的事实的一套规则。具体来讲，范式的基本含义应包括以下几个方面：共同遵守的“科学共同体信念”；公认的范例；与共同信念和模型相适应的方法。换言之，范式就是共同信念、模型、方法三者的有机统一。

所谓范式转换，实质上是一个旧范式灭亡和新范式发生的过程，在库恩的话语体系中它等同于科学革命。按照库恩的观点，当已有的范式不能说明与解释新出现的事实与社会现象，直接导致了反常和危机——也即“例外”——的出现，于是会导致一些争论，这些争论可能导致基础范畴、理论体系的创新，新的分析问题和

解决问题方法的产生，信念、理论、方法等重新组合，形成能够更好地解释现实和解决现实问题的新范式。这一范式转换的过程大致上可以简单地描述如下：范式1——常规科学（在范式1指引下积累的知识）——异例（即范式1不能全面解决的新现象）——危机（即范式1从根本上受到怀疑）——革命（即范式1全面崩溃）——范式2。

从第一部分所描述的评价变革看，原有的教育评价范式因为不能有效解决教育评价领域所出现的现实问题而受到根本的怀疑，一些基础的范畴，如评价的目的、评价的方式，甚至评价方法的基础都从根本上受到质疑。而且，当人们开始相信更重要的是“为学习的评价”而不是“对学习的评价”时，原本为“对学习的评价”而建立的那个范式从根本上崩溃也许就不可避免了。因为在库恩的范式结构中，处于最高层次的是世界观和价值。范式之所以得到承认，是因为科学共同体成员有共同的价值观念和标准。这些价值观念和标准决定了范式的第二层次，范式的思想内容，也就是一个特定时代和特定领域中的基本定律和基本理论。柏拉克所称的“废墟”之外正在生长的东西就是在“为学习的评价”的引导、规范之下生长起来的，这不仅表现在哲学层面，也表现在社会学和人工科学层面。这些新的东西正在得到教育共同体越来越多的成员的“选票”，成为指导教育评价的基本原则。

然而，一种新范式的产生也必然会留下种种有待解决的问题。按照库恩的观点，能被称为范式的应当是常规科学（Normal Science），而且这种常规科学还必须具备两个特征：足以空前地把一批坚定的拥护者吸引过来，使他们不再去进行科学活动中各种形式的竞争；同时又足以毫无限制地为一批重新组合起来的科学工作者留下各种有待解决的问题。①

如何让评价有效地促进学生的学习？这也许是新的教育评价范式亟待我们去解决的问题。特别是，当新一轮课程改革以课程标准的方式对学生的学习结果作了清晰、明确的规范之时，我们如何运用评价让学生更好地达成课程标准对他们所知和能做的期望，这就是我们迫切需要解决的问题。

教育有了课程标准之后，会是什么？国际经验已经告诉我们：随着我国课程标准的出台，随之而来的就是“基于标准”运动，如基于标准的课程设计，基于标准的教学，基于标准的评价，基于标准的问责，基于标准的资源开发，等等。《基础教育课程改革纲要（试行）》明确规定：国家课程标准是教材编写、教学、评估和考试

① ［美］托马斯·库恩.科学革命的结构[M]. 金吾伦等译. 北京：北京大学出版社，2003：8.

命题的依据，是国家管理和评价课程的基础。然而，新课程进入实验区已经 8 年了，我们的课程标准却依然更多是一个文本，未能对教学、评价产生明显的影响。原因在于，我们缺少这方面的知识基础，评价方面尤甚，而评价又恰恰被视为当前课程改革的瓶颈。

三

三年前，我们的团队由钟启泉教授领衔申报了教育部哲学社会科学研究重大课题攻关项目《素质教育课程评价体系研究》（课题编号：04JZD00025），我本人也申报了教育部哲学人文社会科学研究重点基地重大项目《基于标准的学生学业成就评价研究》（课题编号：05JJD880010），期望通过我们的工作获得关于评价的一些新的知识基础。尽管这一领域的研究对我们而言极具挑战性，但我们的研究还是取得了丰硕的成果。本套丛书即为我们的成果之一。

本套丛书的编写者都是具有博士学位的高校中青年教师，在教育理论和教育实践方面具有较为丰富的经验，而且对我国的教育现实及中国教育的发展前景有一个比较清醒的认识，也都乐于在评价方面贡献自己的学识和智慧。他们在新的评价范式之下探究了基于标准的评价的诸多领域，既包括基于标准的评价体系、模式等宏观领域，也包括评分规则及学科领域的评价等微观领域；既包括外部考试、校内考试等传统的纸笔考试，也包括表现性评价等新型评价方式；既涉及基于标准的问责等政策问题，也涉及学校内部的发展性评价等制度问题；既涉及开放题编制、评分规则开发等技术问题，也涉及教师整体的评价素养问题。所有这些研究领域都源于我国当前教育评价的现实问题，也是我国课程改革乃至素质教育推进中必须解决的关键问题。我们相信，关于这些领域的研究必将能为评价发展乃至教育发展提供重要的知识基础——至少提出了一些新范式关注的“有待解决的问题”。

在本套丛书出版之际，我感谢我们的研究团队成员为之作出的巨大努力，没有他们各司其职及团结协作，我们的研究任务不可能顺利完成；感谢教育部社会科学司对本项目提供的资助和华东师范大学“985”工程哲学社会科学“教师教育理论与实践”创新基地提供的平台；感谢华东师范大学出版社为本套丛书提供了出版机会，使我们的研究成果与更多的同仁分享。

前言

以传统封闭式学习活动与题目为主要测试工具的标准化纸笔测验，走过了漫长的发展阶段。作为其反叛与补充形式的开放性活动（open-ended activity），源于20世纪六七十年代医学领域的PBL，其主旨是让学生在真实性任务情境下，在解决有一定挑战性的建构性实践问题的过程中，启发创造性与开放性思维，培养合作与交流能力，锻炼问题分析与问题解决能力。它是一种“实践指向性”极强的任务型学习活动。而开放题（open-ended problem/open-ended test item）则是开放性活动在学业成就测验中的转换与变通形式，二者具有共同本质，是同一事物的两种形式。

关于NAEP、PISA、TIMSS三项国际评价项目的横向比较研究表明，三者中开放性试题（此处为粗略划分，除多项选择题以外）的比例分别是50%、40%、27%。① NAEP比后二者更注重开放题的应用，特别是具有多种答案的开放性题目的比例高达21%。② 由对NAEP科学评估项目的纵向比较研究可知，随着时间的推演，NAEP决策者越来越重视那些能检验学生高级学习成果的试题形式与学习活动，NAEP2000、2005、2009中建构反应型题目的比例占到40%。③ 而且，NAEP2005与2009中还分别增加了30分钟的计算机交互性作业，用于考查学生的信息素养与全

① 黄慧娟等. 关于三项著名国际学生评价项目的比较[J]. 福建师范大学学报，2004，(4). 141－146.

② Comparing Science Content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments. Technical Report. NCES 2006－026.

③ 此表由笔者根据NAEP2000、2005、2009科学评估框架，评估与试题规则及相关资料制作，从中可见NAEP对开放题等高层次评价工具的态度与政策变迁。

面科学素养。

学业成就评价是近几年来基础教育领域的热点问题之一。究其原因无非有三：一是从新课程改革的进程来看，课程评价成为当下的重要环节；二是考试（评价的主要部分之一）一直是个高利害性话题，也成为历次课程改革的重要关注点之一；三是近年来教育测量与评价领域虽无重大的理论突破，但在具体评价技术与工具的探索上有了很多值得关注的进展。

新一轮基础教育课程改革顺应时代和社会要求，强调以学生发展为本，以提高学生的科学素养为目的，以培养学生的创新精神和实践能力为主旨，对传统的教育理念进行全面反思。不断深化的中高考改革也对学生的创新精神和实践能力提出要求，体现了与课程改革的匹配。教育部《关于2000年初中毕业、升学考试改革的指导意见》明确要求："理科考试要结合具体问题考查学生对基本概念和原理的理解，以及运用这些概念和原理分析和解决简单实际问题的能力……在试卷中适当增加开放性题目。"《国家基础教育课程改革实验区2004年初中毕业考试与普通高中招生制度改革的指导意见》中进一步指出："命题要加强试题的开放性、灵活性，注重考查学生学习潜能和创造性思维能力，引导培养学生的创新意识。"特别是近几年来，开放题也成为高考改革中的一个重要方面。

在实践层面上，习题是学科教育实现课程目标的重要工具。传统习题以封闭题为主，强调标准答案，往往使学生不是在做"学问"而是在做"学答"。学生只需重复训练和简单模仿便可顺利破题，学生的质疑能力和问题解决能力无从培养。在这种习题训练中，学生的思维能力得不到很好的发展，批判性和独立性受到压制，求知欲也被消磨在机械枯燥的学习活动中，不能满足现代教育对人才培养的要求，也不能承载新课程改革理念，更不能把新课程改革理念切实转化为课堂教学文化。因此，寻求新的途径和手段势在必行。

开放题承载了新的教育理念，是在对封闭题的深刻反思中应运而生并不断发展的。开放题是学科习题的新题型，能鼓励学生积极探索、深入研究，能引导学生思考问题更具全面性、综合性。开放题是学生思维领域的开放，是学生个体意识的表现舞台，是激发学生发散思维和创新精神的良好媒介，是评价学生探究能力和科学素养的有效工具。

本书共七章，主要从理论基础、编制与评分等方面对开放题进行了相关论述。

学业成就评价的"去标准化"与"真实性"为开放题的兴起提供了滋生的土壤。第一章在澄清开放题的基本范畴的基础上，分析了开放题的研究现状与存在的问