

教育学术文丛



# 基础教育学业成就评价 国际视野与本土实践

Jichu Jiaoyu Xueye Chengjiu Pingjia  
Guoji Shiye yu Bentu Shijian

张雨强 著



北京师范大学出版集团  
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP  
北京师范大学出版社

教育学术文丛



# 基础教育学业成就评价 国际视野与本土实践

Jichu Jiaoyu Xueye Chengjiu Pingjia  
Guoji Shiye yu Bentu Shijian

张雨强 著



北京师范大学出版集团  
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP  
北京师范大学出版社

---

图书在版编目 ( CIP ) 数据

基础教育学业成就评价: 国际视野与本土实践 / 张雨强著. — 北京: 北京师范大学出版社, 2017.3 (2018.8 重印)  
ISBN 978-7-303-22020-5

I . ①基… II . ①张… III . ①基础教育—学业评定—研究 IV . ① G632.47

中国版本图书馆 CIP 数据核字 (2017) 第 012600 号

---

营 销 中 心 电 话 0537-4459916 010-58808015  
北师大出版集团华东分社 <http://bnuphd.qfnu.edu.cn>  
电 子 信 箱 hdfs999@163.com

---

出版发行: 北京师范大学出版社 [www.bnup.com](http://www.bnup.com)  
北京市海淀区新街口外大街 19 号

邮政编码: 100875

印 刷: 济南荷森印务有限公司

经 销: 全国新华书店

开 本: 710 mm × 1000 mm 1/16

印 张: 16.25

字 数: 260 千字

版 次: 2017 年 3 月第 1 版

印 次: 2018 年 8 月第 2 次印刷

定 价: 36.00 元

---

策划编辑: 赵玉山 李 飞                      责任编辑: 刘萌萌 施文燕  
装帧设计: 耿中虎                              责任印制: 李 飞

版权所有 侵权必究

反盗版、侵权举报电话: 0537-4459907

读者服务部电话: 0537-4459903, 4459916

本书如有印装质量问题, 请与服务部联系调换。



## | 第一编 典型学业成就评价体系的比较研究 |

第一章	美国 NAEP2009 科学评估框架及启示 .....	2
	一、NAEP2009 科学评估简介 .....	2
	二、NAEP2009 科学评估框架分析 .....	5
	三、NAEP2009 科学评估框架的启示 .....	8
第二章	美国 NAEP2009 科学课程评价试题编制研究 .....	14
	一、NAEP2009 科学课程评价体系简介 .....	14
	二、NAEP2009 科学课程评价试题编制研究 .....	15
	三、NAEP2009 科学课程评价试题编制的启示 .....	22
第三章	苏格兰促进学习的评价模式述评 .....	25
	一、AiFL 模式是什么 .....	25
	二、AiFL 模式的发展轨迹 .....	26
	三、AiFL 模式的要素构成 .....	28
	四、AiFL 模式的有效运转 .....	31
	五、AiFL 模式的经验启示 .....	32
第四章	台湾学生学习成就评量资料库及其启示 .....	36
	一、TASA 的建构背景及目的 .....	36
	二、TASA 的建设构架 .....	37
	三、TASA 的研究方法 .....	39
	四、TASA 引发的思考与启示 .....	41



第五章	台湾评量人才养成工作坊及其启示	45
一、	台湾“评量人才养成工作坊”概况	45
二、	台湾“评量人才养成工作坊”培训规划	46
三、	台湾“评量人才养成工作坊”对大陆的启示	50
第六章	学业成就评价与课程标准匹配度的比较研究	55
一、	评价与标准匹配度的缘起与价值	55
二、	评价与标准匹配度的分析维度	56
三、	评价与标准匹配度的分析技术	58
四、	评价与标准匹配度的分析程序	60
五、	讨论与建议	61
第七章	课堂教学与学业成就评价的研究综述	65
一、	促进学生学习与发展是课堂评价的定位	65
二、	走向整合的评价：评价与学习的三维视角	65
三、	形成性评价的国际视野	67
四、	教师评价素养提升与评价质量	68
五、	大型评价项目与课堂评价和学生学业成就	69
第八章	走向整合的学习评价	70
一、	对促进学习的评价的理想三维模型的理解	70
二、	评价高水平认知能力的关键在于限定结构	73
三、	形成性评价是一个复杂的尝试	74
第二编 学业成就评价的技术与方法		
第九章	基于课程标准的学业成就评价之韦伯模式研究	78
一、	韦伯模式的研究缘起与意义	78
二、	韦伯模式的分析框架	80
三、	韦伯模式的应用研究	81
四、	韦伯模式的反思	84

第十章 判断评价与课程标准一致性的若干问题 .....	88
一、Webb 一致性程序概要 .....	89
二、本研究使用的一致性标准 .....	90
三、挑战与问题 .....	94
四、结论 .....	98
第十一章 PISA2006 科学试题设计和开发及其启示 .....	100
一、PISA2006 科学评估框架 .....	100
二、PISA2006 科学试题设计 .....	102
三、试题开发 .....	103
四、附加样题 (科学素养第 10 单元: 酸雨) .....	108
五、PISA2006 试题设计和开发的启示 .....	110
第十二章 PISA2006 科学评估对我国化学试题编制的启示 .....	115
一、PISA 简介 .....	115
二、PISA2006 科学评估框架分析 .....	115
三、PISA2006 科学评估样题分析 .....	119
四、PISA 对我国初中化学试题编制的启示 .....	121
第十三章 高中化学教科书课后习题难度预估方法探究 .....	124
一、影响课后习题难度的假设因素及其统计分析 .....	124
二、建立难度预估方程 .....	129
三、难度预估方程的检验 .....	132
<b>第三编 学业成就评价的本土探索</b>	
第十四章 初中毕业生学业考试的市域比较研究 .....	136
一、各省市初中毕业生学业考试方案的梳理与比较 .....	136
二、我国初中毕业生学业考试存在的问题 .....	143
三、我国初中毕业生学业考试的改革建议 .....	145



第十五章 基于区域方案比较的普通高中学业水平考试研究 .....	149
一、各省（市）普通高中学业水平考试方案的系统梳理 .....	149
二、我国普通高中学业水平考试存在的问题 .....	156
三、我国普通高中学业水平考试的改革建议 .....	158
第十六章 义务教育阶段科学学业成就评价框架的开发研究 .....	164
一、科学学业成就评价的国际经验 .....	164
二、义务教育科学学业成就评价开发的整体设想 .....	170
三、科学学业成就评价的试题编制 .....	174
第十七章 普通高中综合素质评价的实施路径 .....	182
一、省市如何推行综合素质评价工作 .....	183
二、高校如何进行综合素质评价 .....	185
三、普通高中学校如何落实综合素质评价 .....	187
四、如何将综合素质评价结果纳入高校招生录取 .....	189
五、如何将综合素质评价结果用于普通高中发展 .....	190
第十八章 英语开放性学业成就评价与课程标准的匹配策略初探 .....	194
一、开放性学业成就评价与课程标准的内涵 .....	194
二、基于标准的英语开放性学业成就评价 .....	196
三、总结 .....	202
第十九章 中考化学试题与课程标准的一致性初探 .....	207
一、学业成就评价与课程标准的一致性内涵 .....	207
二、利用韦伯一致性分析标准分析中考化学试题 .....	208
三、对《全日制义务教育化学课程标准（实验稿）》的反思 .....	211
第二十章 高考化学试题与课程标准的一致性初探 .....	214
一、韦伯模式的内涵 .....	214
二、高考化学试题与课程标准的一致性分析 .....	215
三、问题与反思 .....	218

第二十一章 高中化学学业水平考试与课程标准的一致性研究 .....	220
一、研究对象和研究工具 .....	220
二、研究过程 .....	221
三、内容主题划分方法与一致性讨论 .....	228
四、结论与反思 .....	229
第二十二章 初中化学教材习题与课程标准的一致性分析 .....	231
一、一致性及其研究工具 .....	231
二、教材习题与内容标准的认知水平编码 .....	232
三、习题与课程标准的一致性分析结果 .....	237
四、结论与建议 .....	239
第二十三章 高中化学教材习题与课程标准的一致性研究 .....	242
一、研究工具 .....	242
二、高中化学选修4教科书课后习题与课程标准的一致性分析 .....	243
三、结论与反思 .....	248
后 记 .....	251





## 第一编

# 典型学业成就评价体系的比较研究





## 第一章 美国 NAEP2009 科学评估框架及启示

美国教育进展评估 NAEP (National Assessment of Educational Progress) 是美国目前最权威的全国学生学业成就评估体系，它通过测量学业成就趋势来衡量全国教育目标的达成与否，从而对国家基础教育质量监控产生至关重要的作用。NAEP 正式建立于 1969 年，由美国国会委托美国教育部国家教育数据中心 (NCES) 负责运行，经费由联邦政府负责。主要评估数学、科学、阅读、写作、地理、美国历史、艺术、公民及其他学术性科目中学生应该掌握的知识和技能。评估年龄段依次是 9、13、17 岁，1983 年后改为 4、8、12 年级。NAEP 报告提供多学科的学生学业表现 (包括基本技能与高级技能) 的相关信息，以及基于不同种族或人种、性别、社区类型与地理区域的比较信息；同时，其数据也能显示出学业成就与背景变量 (如作业花费时间、家长教育水平等) 之间的关系。

### 一、NAEP2009 科学评估简介

2006 年 NAEP 推出了新的“NAEP2009 科学评估框架”，同期还推出了“NAEP2009 科学评估与试题规则”姊妹文件。NAEP2009 科学评估框架的开发集中了数以百计人的智慧，包括国家权威科学家、科学教育工作者、决策者以及评估专家。在与国家评估理事会 (the National Assessment Governing Board, NAGB) 的契约下，由 WestED (非盈利性教育研究与服务中介机构) 与重点州学校管理委员会 (the Council of Chief State School Offices, CCSO) 历经 18 个月开发了该框架，此间听取了各委员会、地区委员会与公众的意见与建议，并接受了 NAGB 的外部监督。

### （一）评价框架概况

NAEP2009 是根据广泛接受的国家科学教育标准和评估水平进行设计的，同时也考虑到了地方课程标准。为了打造国际竞争人才，使美国年轻人在科技方面保持国际竞争力，框架在设计过程中，充分参照了国际评估框架，如国际数学与科学趋势项目（The Trends in International Mathematics and Science Study, TIMSS）与国际学生评估项目（Program for International Student Assessment, PISA）。

科学内容和科学实践是 NAEP2009 科学评估的两大内容。科学内容涉及物理科学、生物科学、地球与空间科学三个领域，同 NAEP2005 相比，增加了空间科学评估；科学实践则分为识别科学原理、应用科学原理、实施科学探究、进行科学设计四个维度。

在具体的评价技术方面，NAEP 有两种方式：一种是选择反应，如多项选择；另一种是建构反应，如简答题。对于每一年级，试卷中选择反应与建构反应题目分量各占 50%。为了进一步考查学生的综合理解、科学研究技能，每个学生必须完成 30 分钟附加的交互式计算机任务或动手操作任务；各年级测验中，这两类题目不得少于 4 个，应该包含至少 1 道交互式计算机任务或动手操作任务，而且前者数量不能多于后者。<sup>[1]</sup>交互式计算机任务有四种类型：研究和分析信息，使学生主动寻找有关信息进行问题解决；以经验为基础进行调查，使学生为达到问题解决而研究；模拟和模仿，使学生巧妙地运用科学处理多变的问题，培养学生预测和解释的能力；概念图，探测学生组织知识进行逻辑推理的能力。

### （二）科学评估功能

近 40 年来，NAEP 提供了完整的信息，报告了美国 4、8、12 年级教育状况与进展情况，近年来也为部分州及一系列大的城市学区提供了教育数据。NAEP 的目的是及时提供公正准确的学生学业成就测量信息，并报告学生在阅读、数学及其他学科领域内的学业成就进展情况。鉴于 NAEP 严格的设计与执行方法，其报告正越来越多地为各州管理相应评估学科而服务，也作为设计其他大规模评估的模型等。具体而言，NAEP 的主要功能有：<sup>[2]</sup>



### 1. 管理工具

作为唯一实施中的国家学生教育进步研究项目，NAEP 已经越来越成为获得学生知识与技能等相关信息的重要信息源。决策者、教育工作者与公众都能通过各种途径获得相应评估信息。因此在国家、州与学区等各级水平上，NAEP 可以用作管理学生阅读、数学、科学及其他学科学业成就的工具之一。例如，NAEP 报告（国家报告卡）比较了全国不同地区之间某一学科的学业成就差异，或不同时期内的某一学科差异，或同一年级中不同被试群体的差异。NAEP 还报告 9、13、17 岁学生在阅读、数学等领域内的长期学业成就倾向。<sup>[3]</sup>此外，对个别州级标准在多大程度上反映了国家框架规定的核心知识与技能等，也得到了较为公正的反映。如果某州使用的是独立标准（不使用国家标准），比较研究就仅限于州级标准与 NAEP 标准之间的匹配度分析。即便如此，NAEP 依然是衡量科学课程中 4、8、12 年级学生知识与技能的权威指标。

### 2. 评估开发与评估方法的模型

NAEP 评估框架与评估规则本身就是国家、州级与地方课程与评估的资源库。那些被广泛采用的评估框架与评价规则开发程序，本身就意味着某个学科内学生知识与技能的当前研究成果。另外，NAEP 使用了设计极为严密的程序来开发相应的评估工具，而且试测、外部审查与内部审议等程序都保证了 NAEP 评估具有极高的信度与效度，这些复杂严密的方法完全可以作为其他评估工作者的工具。加上“不让一个孩子掉队法案”（No Children Left Behind NCLB）中已经规定的评估要求，各州可以把 NAEP 作为模型来自行指导其州级评估的开发工作。

### 3. 研究资源与决策依据

NAEP 数据包含了学科学业成就结果（以两种形式汇报：等级评价与学业成就水平），各种次级水平上的关于学校、教师、学生的背景信息（如 12 年级西班牙裔男生的选课模式），州级水平的结果，某些城市大区的信息，州与地区参加 NAEP 的历史纪录，公开的评估试题、学生答案与评分指南等。NAEP 网站提供了各种用户友好型分析软件，让决策者与研究者等可以全方位地研究 NAEP 数据，编制重要测验，或定制图表化的 NAEP 评估结果。这些数据与软件等工具可以用于决策、次级研究及其他目的。

## 二、NAEP2009 科学评估框架分析

就结构而言, NAEP2009 科学评估沿用了历史框架, 分为科学内容与科学实践两大部分。但经过比较, 无论是在科学内容与科学实践上, 还是在评估与试题规则上, 都可以看出 2009 科学评估明显的转变。这种转变彰显了美国科学教育评价研究的新进展, 也透视了美国国家科学教育政策的走向与人才培养的趋势。

### (一) NAEP 科学评估: 历史的视角

NAEP 科学评估每四年一次, 数学与阅读评估每两年一次。NAEP1996、2000、2005 三个科学评估框架在十几年的时间内并未得到充分发展。考虑到国际竞争的加剧和经济全球化的大背景, NAEP2009 在前三者的基础上做了较大革新, 给科学评估进一步发展提供了参照, 是新的 NAEP “科学动向”, 也为发展科学评价和提高研究科学水平提供了一把钥匙。我们拟通过二者间的比较, 管窥 NAEP 科学评估的新进展。

表 1-1 NAEP1996 ~ 2005 与 NAEP2009 科学评估项目比较<sup>[4]</sup>

	1996-2005 框架	2009 框架
科学内容	评估内容的相关科学标准几乎没有涉及	科学内容从既有的标准与评估框架(如, 国家标准, 基准, TIMSS, PISA 与州级标准)中生发而来
	内容领域包括: 物理科学、生命科学、地球科学	内容领域包括: 物理科学、生命科学、地球与空间科学
	不同学科领域与年级之间试题分配比例的建议: 4 年级与 12 年级几乎相同, 8 年级更强调生命科学	不同学科领域与年级之间试题分配比例的建议: 4 年级三者比重相同, 8 年级强调地球与空间科学, 12 年级强调物理科学与生命科学
	内容以较为独立的句子与短语形式呈现	科学内容以相应内容次级标题的图表形式呈现(如, 物理科学中的“力”与“影响运动的力”)
	框架使用三个抽象主题: 系统、模型、变化形式	
		框架使用了横跨三个学科的内容
	评估试题涉及“科学本质”	



续 表

	1996-2005 框架	2009 框架
科学实践	“知识与操作”维度被分为概念理解、具体调查、实践推理三个方面	“科学实践”维度分为识别科学原理、应用科学原理、实施科学探究、进行技术设计四个方面
		“科学本质”的考查放在了“科学实践”中，尤其是应用科学原理与科学探究上
	“科学实践”的评估很大程度上是“基于经验的”	“科学实践”的评估重视了科学研究以及那些能反映认知复杂性的试题
	评估包含了实践推理题目（如应用科学知识对日常生活问题提出有效对策）	
		评估包含了技术设计题目（如通过系统程序应用科学知识与技能以解决真实情境中的问题）
	45% 的评估试题集中于概念理解上	60% 的评估试题集中于概念理解上（如识别与应用科学原理）
		评估包含了学习进步情况（如不同学段中某个科学表现的连续序列之变化情况）
评估试题		试题包含了科学史、科学与技术的关系等问题
	评估试题包含了纸笔试题与操作表现型任务	评估试题包含了纸笔试题与操作表现型任务，以及交互性计算机任务
	“框架”中未呈现相应科学知识或科学实践的例题；只是在“评估规则”中出现了一些建议	在“框架”与“评估规则”中都呈现了相应科学与科学实践的例题
		“框架”与“评估规则”包含了针对残疾学生与英语语言学习者的评估指南
		“框架”包含了如何编制试题与解释试题的案例
		学生关于科学原理的原始概念也被详尽评估了

## (二) 科学评估框架分析

### 1. 2005 框架

评估筹划委员会规定了 NAEP 科学评估的五个特征,<sup>[5]</sup> 作为其下属计划委员会的工作指导方针:

- 评估框架要反映全体美国学生达到高水平科学理解力所必需的知识、技能与能力等相关科研成果。相应地,它必须涵盖知识、应用有组织的事实性信息、概念之间的联系、统整科学的核心思想、思维、实验室技能等;必须基于科学领域中对教学、学习与学生表现的最新研究成果。

- 框架与 NAEP 科学评估必须陈述清楚科学知识的本质与实践,并区分与其他知识途径的不同;反映科学的量化方面,以及生命科学、地球科学与物理学中的概念;直面科学与技术引发的社会问题与事件;包含设计、材料应用、利益权衡等方面的实践性问题的解决;考虑到学生发展的不同水平;保证不同背景的学生在评估中均可以客观、公正地有机会展示自己的知识与表现。

- 评估试题必须与所评估目标相一致。评估学生表现的策略应该包括:表现型任务,可以允许学生进行表现、展示对测验材料的科学理解;开放题,能让学生展示其理解水平与科学中的交际能力,及其对科学概念与概念间联系的创造能力(而不是简单确认);长期收集学生作品,能显示出学生在严格控制时间的标准化测验外还能做些什么;多项选择,展示学生对概念的理解,以及按照科学方式对多个科学概念的联系能力。

- 评估必须包含每一年级的、三种不同熟练水平的、广泛的试题类型。

- 必须收集学生的统计信息与背景信息。必须从学生、教师、管理者等处收集有关教学计划与课程实施系统的附加信息,以便确定它们与学生学业成就的关系,给教学计划与课程决策者提供信息。

### 2. 2009 框架

通过对 NAEP2009 科学评估框架的分析,结合框架本身的说明,可以看出其主要有如下特点:<sup>[6]</sup>

- 框架的开发基于国家科学标准与评估体系,以及州级课程标准。但框架目的是为开发评估系统提供信息,而非是为了提倡具体的教学方法,亦非为了表征科学内容与技能的全部范畴。



- 为了使美国青年在科学与技术上具有国际竞争力，框架开发程序对国际评估框架给予了特殊考虑，如 TIMSS 和 PISA。

- 原始材料中包含的大量科学原理，使评估框架必须聚焦于每一学科中的基础知识与拓展知识，因而相应地减少了对科学内容的评估。

- 框架基于科学知识，框架开发程序来自于既有的测验与评价研究成果，同时也有大量的经验性事实提供支撑。

- 那些需要依靠非科学观点才能做出的对自然现象的解释，并未反映在框架中。

- 科学内容的呈现很具体，根据不同年级制定相应图表，让读者可以清楚地看到学生的认知复杂性随着不同年级的渐变过程。

- 科学内容的描述力争避免出现错误，语言应用上力争准确明晰又应避免技术化，以便于广大社会人士能够读懂框架。

- 框架聚焦于学生的概念理解，即他们的知识，以及对科学事实、概念、原理、法则与理论的应用；也反映了学生参与科学探究及技术设计的能力。

- 推荐了一些新的试题形式，如交互式计算机任务等。

- 对科学课程中交际能力与量化推理能力的使用情况进行评价。尽管并未规定相应试题类型对应的具体考试时间，但框架规定所有的三个年级（4、8、12）都必须包含相关试题。

### 三、NAEP2009 科学评估框架的启示

NAEP 是美国全国性学业成就评价的整体框架，深入研究其建立与运转机制、评估程序与评价技术，对于建设我国全国性学业成就评价体系，并进而建立全国教育质量监控系统，具有相当的借鉴价值。尤其是 NAEP 关于标准匹配度的保证技术、学业成就评价数据库的建设与灵活运用策略等，更是具有较高的技术性与科学性。

（一）呼唤整体性国家评估体系的建立，提供可供参考与借鉴的国家评估范本

教育评估与评价应该有个国家水平的整体评估体系，并且能对地方水平与学校水平的课程评价和教学评估提供一个值得借鉴的范本。NAEP 评估体系基



本可以分为国家层面评估（含长期趋势评估）、州级评估、试验地区评估三个层面。国家层面评估涵盖所有学科，每两年一次（数学与阅读、长期趋势评估、科学、美国历史等每四年一次）；州级评估学科只有阅读、数学、科学三门，与国家层面评估同期进行；试验地区评估为不定期安排，主要考虑各种具体需要。长期趋势评估是为了追踪美国学生学业成就变化的趋势，从而为国家教育决策提高科学依据。NAEP 在这方面值得我们学习，它依据国家课程标准开发，同时又参考了地方课程标准，兼顾了国家标准的统一性与地方标准的特殊性二者之间的协调。而这一做法也可以为我国的三级课程管理体制提供一条新思路。

在日益强调国家课程标准与学业成就标准水平的今天，如何在“校本化进程”中保证“国家标准底线”与“区域评价灵活”二者间的和谐统一，更是评价研究工作者关注的重点之一。<sup>[7]</sup>在监控了基于国家课程标准的基础教育质量的前提下（即我们常说的国家水平上的“基于标准的评价”），NAEP 也考虑到了那些不实施国家课程标准、采用州级独立课程标准的个别州与地区。对这些特殊区域的教育评估，NAEP 则采用了比较地方与国家课程标准以及评估标准之间的匹配度的分析方法。即虽然这些州与地区看似“游离于国家标准之外”，但国家评估体系对这些“独立标准”的区域并未“放任自流”，而是以国家课程标准与评估标准为参照系，来比对区域教育进展。

## （二）重申国家评估的功能定位，立足教育评估的发展性与持续性

国家教育评估与评价的功能应该定位于发展性，即为国家、地方与学校的课程决策者与执行者提供改进学业成就的决策信息，而不是为了对不同地区与学校之间的学业成就进行优劣比较。NAEP 报告（国家报告卡）与 NAEP 州评估做法相同，不报告参与评估的个别学生或学校的信息，而是以人口群体（如 4、8 年级学生）和子群体（如男生、女生、亚裔学生、非裔学生等）来报告学生的学业成绩、教育体验、学校环境等信息。这种有效措施是发展性教育评价的实际体现，也从根本上杜绝了各地区与学校根据国家报告卡来“攀比”的可能。

国家评估体系不应是孤立的，其丰富信息应该能够为相关评估与评价研究（地方、学校）提供基本的参考数据。国家评估体系是建立在教育评价最新研究