现代语言测试与评估丛书

曾用强 总主编

形成性评估研究

李清华 王伟强 张 放 著

科学出版社

现代语言测试与评估丛书

曾用强　总主编

形成性评估研究

李清华　王伟强　张　放　著

科学出版社

北京

图书在版编目(CIP)数据

形成性评估研究 / 李清华，王伟强，张放著. —北京：科学出版社，2014.2
（现代语言测试与评估丛书 / 曾用强主编）
ISBN 978-7-03-039664-8

I. ①形… II. ①李… ②王… ③张… III. ①语言-评估-研究 IV. ①H09

中国版本图书馆 CIP 数据核字(2014)第 017486 号

责任编辑：刘彦慧　张翠霞 / 责任校对：鲁　素
责任印制：钱玉芬 / 封面设计：无极书装

联系电话：010-6401 9074　电子邮箱：liuyanhui@mail.sciencep.com

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

双青印刷厂印刷

科学出版社发行　各地新华书店经销
*
2014 年 2 月第　一　版　开本：A5 (890×1240)
2014 年 2 月第一次印刷　印张：7
字数：230 000

定价：58.00 元

(如有印装质量问题，我社负责调换)

形成性评估是教室里沉睡的巨人，是唤醒他的时候了！

（Brookhart，2007：57）

总　序

据 Spolsky 考证，正式的语言测试起源于中国东汉时期的科举考试。但是，现代意义上的语言测试却于 20 世纪中叶诞生在英美等发达国家。20 世纪 60 年代，外语测试作为一门新的学科从外语教学中独立出来。作为语言学、教育与心理测量学、计算机技术等的交叉学科，半个世纪以来，国外大批语言学家在语言测试领域取得了卓越的成就。两本专业期刊——《语言测试》(*Language Testing*)和《语言评估季刊》(*Language Assessment Quarterly*)相继诞生，在应用语言学界的影响越来越大，语言测试已经成为应用语言学领域的显学之一。在我国这样一个具有浓郁考试文化的大国，考试一直备受国人青睐，甚至受到顶礼膜拜。中国的高考和大学英语考试(CET)虽有数百万考生规模，但在研究方面仍落后于英美等发达国家。至今，以美国 ETS(Educational Testing Service)开发的 TOEFL 和英国剑桥大学考试委员会(CESOL)的 IELTS 为代表的西方国家的研究水平仍执全球之牛耳。但是，我们欣喜地看到，国内学者没有妄自菲薄，外语测试研究日益受到人们的重视，并已经取得一些成绩，一些学者的论文在国际期刊发表，国内专业期刊《外语测试与教学》也在 2011 年于上海外国语大学问世。虽然有若干著述出版，但它们大多关注于测试的开发实践，而对测试理论与实践的研究较少。在此背景下，我和我的博士们策划了“现代语言测试与评估丛书”。在国内学界享有盛誉的科学出版社高瞻远瞩，大力扶持外语研究，欣然同意出版这套丛书，可谓语言测试界的盛举。丛书为国内语言测试研究者提供一个平台，系统展示国内外语言测试领域的新成果，特别是国内学者的原创研究，供广大同行分享。

本丛书分为两个系列：①语言测试与评估：研究系列；②语言测试与评估：实践系列。前者以理论和研究为重点，主要面向应用语言学的研究生和语言测试研究者，后者以实践和应用为重点，主要读者是广大外语教师和教育行政管理人员。本丛书的第一系列围绕语言测试的热点和经典问题展开，主要涉及以下话题：语言测试的效度理论与实证研究、基于任务的语言测试研究、形成性评估研究、 ESP 测试研究、动态评估研究、语言测试的评分研究、语言测试计算机自动评分研究、基于语料库的语言测试研究、语言测试的后效研究等。每一本书在保持各自特色的前提下，主要内容包括：①系统介绍相关理论；②本领域的主要研究方法；③实证研究成果分析；④附录：本领域近十年的研究成果。

本丛书的编著者均为语言测试方向的博士和知名学者，相信本丛书的出版将进一步促进国内语言测试研究的发展。本丛书是一个开放的平台，欢迎广大同仁提供自己的新作，欢迎广大读者提出批评与建议。本丛书编著者愿与国内同行一起，为使我国从考试人数大国早日发展成为考试研究大国而努力。

曾用强

2013 年于广州

前　言

数百年以来，教育评估主要用来选拔和问责(accountability)。近几十年来，教育评估改革的浪潮席卷全球(Berry，2011)。从世界范围来看，教育评估正在从“对学习的评估”(assessment of learning)转向“为学习的评估”/“学习性评估”(assessment for learning)，从“测试文化”(testing culture)转向“评估文化”(assessment culture)。评估的促学作用越来越受到重视。形成性评估(formative assessment)具有巨大的促学潜力，也被称为“学习性评估”(Berry & Adamson，2011；Berry，2008；Stiggins，2005；Broadfoot & Black，2004；Stiggins，2002)或“促学评估”(Huhta，2008)。尽管斯克里芬(Scriven)早在1967年就明确提出将评估分成形成性的和总结性的两类评估的思想，但直到Black和Wiliam(1998a)关于形成性评估的综述的发表，形成性评估实践才在全世界的教育评估中迅速普及起来。

在外语教育领域，形成性评估的研究和实践也风靡全球，但没有得到像大规模高风险测试一样的重视。尽管面临诸多问题和困难，课堂形成性评估的巨大潜在作用吸引了越来越多的学者和教师。很多评估改革对课堂形成性评估寄予厚望。如何改变高考，以及大学英语四级、六级考试和英语专业四级、八级考试等总结性评估(summative assessment)一统天下的局面，使外语评估既达到选拔人才的目的又对教与学产生应有的促进作用？形成性评估如何成为标准化测试的有效补充？如何提高外语课堂评估的质量？如何设计开发适合我国实际的课堂形成性评估？使用哪些技术标准来评价形成性评估的效果？这是我国教育测量界和语言测试界亟待研究的课题。实际上，一些有识之士已着手这方面的工作，

例如，韩宝成(2009)、罗少茜(2003)、李清华(2012，2006)、李清华等(2013)、王华和富长洪(2006)、王红艳和解芳(2004)等对国外测试评估理念进行的讨论，曹荣平等(2004)、李清华(2008)、王华和甄凤超(2008)、文秋芳(2011)、周娉娣和秦秀白(2005)等进行了形成性评估的试探性研究。《英语课堂教学形成性评价研究》(罗少茜，2003)介绍了形成性评估在幼儿园和小学的英语教学中的应用，《新课程与教育评价改革译丛》(董奇，2005)系引进国外的原版著作的译文，这两种书的读者都为中小学教师和管理者，但在理论和研究层面多有欠缺。

本书分为四章和附录。第 1 章阐述形成性评估的理论基础，包括形成性评估的起源、形成性评估的界定，以及形成性评估与终结性评估、动态评估、课堂评估的关系，主要从学习理论、心理学理论、测量学理论的角度阐述形成性评估的理论基础，讨论形成性评估的效度和信度问题。第 2 章介绍国内外教育领域，特别是 ESL/EFL 领域的形成性评估的实践，分别从课堂形成性评估、大规模形成性评估、计算机辅助形成性评估三个维度以实例说明形成性评估的方法。第 3 章讨论形成性评估的研究方法，分别从量的研究方法和质的研究方法入手，并辅以实证研究成果分析。第 4 章从全球和中国两个视角讨论形成性评估的现状，指出亟待研究的问题，并展望未来发展的前景。附录部分介绍本领域近十年来的研究成果，供研究者参考之用。

本书的编写分工如下：王伟强负责第 3 章，张放负责第 2 章第 3 节及附录部分，其他章节和全书统稿由李清华负责。

本书被荣幸列于《现代语言测试与评估丛书》并获得出版资助，在此特别感谢作者的博士生导师曾用强总主编的支持和帮助！

科学出版社致力于扶持学术研究，刘彦慧女士为本书的编辑付出了巨大心血，作者深怀感激。

本书面向在读硕士、博士，语言测试研究者和广大外语教师，既有深入的理论探讨，又有丰富的实践方法和研究个案，还提供本领域的最

新研究成果，为读者展示形成性评估的学术思想，促进我国形成性评估的研究与发展。由于作者水平所限，书中难免存在疏漏之处，恳请广大同仁不吝批评指正。

作 者

2013年于广州

目　录

总序
前言
第 1 章　形成性评估的理论……1
1.1　形成性评估溯源……2
1.2　形成性评估的概念……10
1.3　形成性评估的理论基础……40
1.4　形成性评估的理论……77
第 2 章　形成性评估的实践……88
2.1　形成性评估的操作框架……88
2.2　课堂形成性评估实例……93
2.3　大规模形成性评估实践……126
2.4　计算机辅助形成性评估……133
第 3 章　形成性评估的实证研究……143
3.1　形成性评估实证研究的方法论……145
3.2　以定量研究方法为主的实证研究……149
3.3　以定性研究方法为主的实证研究……159
3.4　小结……165
第 4 章　形成性评估的现状与未来……166
4.1　国外形成性评估现状……166
4.2　中国形成性评估现状……169
4.3　形成性评估展望……171
参考文献……175

图 目 录

图 1.1 测试、测量、评估和评价的关系……14
图 1.2 自主学习、形成性评估和自我评估的关系……28
图 1.3 诊断性评估和形成性评估与其他评估的对比……36
图 1.4 形成性评估与其他评估的关系……37
图 1.5 评估的使用论证链……55
图 1.6 评估论证的流程……56
图 1.7 形成性推断的效度评价框架……65
图 1.8 信度块示意图……69
图 1.9 形成性评估的效度验证框架……72
图 1.10 形成性评估的社会文化模式……84
图 1.11 形成性评估理论框架……86
图 2.1 KLT 行动理论……89
图 2.2 形成性评估的循环框架……91
图 2.3 课堂评估的过程和策略……92
图 2.4 正式任务的形成性评估的过程……93
图 2.5 教学流程图……95
图 2.6 本课程中期的教学流程……96
图 2.7 形成性评估的操作框架……97
图 2.8 形成性评估流程……106
图 2.9 形成性评估流程修正……113
图 2.10 外语教学的形成性评估体系……114
图 2.11 portfolio 的基本特征……117

图 2.12　实验班和对照班的教学与评估安排……119
图 2.13　PBWA 操作模式……120
图 2.14　能力证据的构成……135
图 3.1　研究思路……149

表 目 录

表 1.1 形成性评估和总结性评估的对比……15
表 1.2 课堂形成性评估与总结性评估的特征……18
表 1.3 评估目的与评估类别的关系……19
表 1.4 语言行为表现评估和传统标准化语言测试的主要差异……25
表 1.5 语言行为表现评估的主要优点……26
表 1.6 形成性评估的特征的变体……39
表 1.7 语言测试效度观的对比……52
表 1.8 Lynch 和 Shaw 的效度理论框架……54
表 1.9 课堂测试效度的证据来源……63
表 1.10 大规模评估与课堂评估概念比较……70
表 1.11 形成性评估效度验证的主要问题及收集数据的方法……74
表 1.12 评估类型及其特点……78
表 1.13 实证主义范式与后实证主义范式的对比……79
表 1.14 心理计量范式、情景化范式和个人化范式的比较……80
表 1.15 社会建构主义的课堂教学理念及评估原则……82
表 2.1 形成性评估主要的教学过程……90
表 2.2 “应用语言学文献阅读与评价”课程安排……94
表 2.3 阅读每篇文章需要回答的问题……98
表 2.4 评价文章需要回答的问题……98
表 2.5 口头报告与小组合作的要求……100
表 2.6 参加者背景……118
表 2.7 教学活动安排……138
表 3.1 研究工具……156

第 1 章

形成性评估的理论

形成性评估尽管具有巨大的促学潜力，但没有得到像大规模测试一样的重视。大规模测试的统治地位影响了课堂评估重要性的发挥(Cizek，2009)。尽管面临诸多问题和困难，课堂形成性评估的巨大潜在作用吸引了越来越多的学者和教师。很多评估改革对课堂形成性评估寄予厚望。例如，我国香港地区的“为学习的评估”改革项目明确提出，从 2007~2008 学年起，基于学校的评估部分取代外部大规模测试的作用(Davison，2004)；英国威尔士地区从 2007 年起取消 11~14 岁学生的标准化测试(Leung & Lewkowicz，2006)。中国外语教育中出现的围绕高考、大学英语考试(College English Test，CET)等大规模测试的应试教育，已引起越来越多的关注与担忧。近几年来出台的《英语课程标准(实验稿)》(2001 年)、《普通高中英语课程标准(实验)》(2003 年)和《大学英语课程教学要求》(2004 年)则试图改变中国英语教学与评估的现状，突出了形成性评估的重要地位。形成性评估为何难登大雅之堂，为何不能成为标准化测试的有效补充？如何提高外语课堂评估的质量？如何设计开发适合我国实际的课堂形成性评估？使用哪些技术标准来评价形成性评估的效果？这是我国教育测量界和语言测试界亟待研究的课题。实际上，一些有识之士已着手这方面的工作，例如，国家基础教育课程改革“促进教师发展与学生成长的评估研究”项目组主持的“新课程与教育评估改革译丛”(2003~2005 年)，韩宝成(2009)、罗少茜(2003)、李清华(2006)、

王华和富长洪(2006)、王红艳和解芳(2004)等对国外测试评估理念的讨论，曹荣平等(2004)、李清华(2008)、王华和甄凤超(2008)、文秋芳(2011)、周娉娣和秦秀白(2005)等进行的形成性评估的试探性研究等。为了改变高考和大学英语四、六级考试等总结性评估一统天下的局面，使外语评估既达到选拔人才的目的又对教与学产生应有的促进作用，课堂形成性评估大有可为。

毋庸讳言，形成性评估面临巨大挑战。人们使用“形成性评估”时，其概念往往相去甚远；形成性评估的实践灵活多变，似乎无章可循；形成性评估的理论也鱼龙混杂。没有理论指导的实践是盲目的，没有理论基础的学术研究方法和结果是混乱的。本章讨论形成性评估的理论问题：追溯形成性评估的源头，理清形成性评估的概念，明确形成性评估的理论基础，构建形成性评估的理论。

1.1 形成性评估溯源

形成性评估是教育评估(educational assessment)的理论之一，教育评估是从教育评价(educational evaluation)和教育测量(educational measurement)发展而来的。教育评估的渊源可追溯至中国西周时期的科举考试制度，但现代意义上的教育评估却产生于美国。从世界范围来看，教育评估的发展历史可大致划分为以下四个阶段。

1.1.1 教育评估的萌芽时期(19 世纪中叶至 20 世纪 30 年代)

19 世纪下半叶，西方国家在实验心理学和数理统计学的基础上，用检查考核的手段对个体在智能上的差异进行定量研究，即探索对学生学力的客观化、标准化测量，从而为教育测量理论的建立奠定了基础。

1845 年，美国的梅恩(Mann)首先在马萨诸塞州波士顿文法学校引入统一试卷的书面考试。

1879 年，德国心理学家冯特(Wundt)在莱比锡大学建立了世界上第一个心理学实验室，摸索出一套心理测量的方法。1882 年，英国科学家高尔顿(Galton)在伦敦设立了人类学实验室，对人类心理遗传及个体差异进行研究，并设计了许多统计方法。冯特和高尔顿的研究对学生学力测量的研究产生了较大影响。1894 年，美国心理学家卡特尔(Cattell)在哥伦比亚大学使用各种测量考核本校大学生。这些研究为 20 世纪初教育测量运动奠定了基础。美国学者桑代克(Thorndike)和他的学生做出了突出的贡献。桑代克的《心理与社会测量导论》(1904 年)具有划时代的意义。他提出了“凡是存在的东西都有数量，凡有数量的都可测量”的著名信条。他与他的学生积极投入教育测量工作，陆续编制了各科标准测验[①](standard test)(包括学业测验、智力测验和人格测验)和标准测量表(standard scale)。此后，各种修订的量表层出不穷(如 1905 年问世的《比奈-西蒙智力量表》)，智力测验风靡世界。随着教育测量的发展，人们开始用学生学力测验结果来评估学校教育。与学力测验相关的升级率、退学率、教学效率等指标及其评估方法开始出现。教育测验运动取得了巨大成就。然而，人们逐渐认识到，教育测验尽管能使考试客观化、标准化，并能把人的能力换算成数字，甚至个别差异的程度也可以度量，但测验毕竟不能测量人的全部，即便是研究成果最多的学力测验，也不能测得学力的全部领域。例如，社会态度、创造力、兴趣、鉴赏力等重要的学力内容，因难以数量化，教育测验不能充分把握，往往被教育者冷落。因此，对教育测验的批评越来越多。

教育评估正是为弥补教育测验之缺陷而发展起来的。当然教育评估并非取代教育测验，而是在重视原有教育测验的同时，也注重测验以外的评估方法，从而把所有能够用以考查教育效果的方法综合起来，以评定教育是否实现全部教育目标。

---

① test 在教育和心理测量界多译为“测验”，在语言测试界统称为“测试”。

1.1.2 教育评估的形成时期(20 世纪 30 年代至 50 年代)

美国在 1933~1940 年开展了一次历时八年的课程改革，史称“八年研究”，该研究的目的是改革中学的核心课程(core curriculum)。为检验课程改革实验的结果，分析新课程与大学学习的关系，全面衡量学生的各项进步，成立了以泰勒(Tyler)为首的评估委员会，他们进行了卓有成效的评估改革实验研究，取得了一系列新成果。后人把以泰勒为首的评估委员会八年的研究成果《史密斯-泰勒报告》，称为“划时代的教育评估宣言”，泰勒也因此享有“教育评估之父”的美誉。

泰勒教育评估的基本思想包括：①教育是使人的行为方式发生变化和改进的过程；②这些形形色色的行为方式的变化就是教育目标；③教育评估就是确定教育目标实现程度的过程；④人的行为是复杂的，有的可以量化，有的难以量化。因此，除测验以外，还需要用其他各种评估手段(如观察、调查等)来检查教育效果。

泰勒教育评估的过程分为四个步骤：第一步，确定教育目标；第二步，设计评估情境；第三步，选择和编制评估工具；第四步，分析评估结果。这就是著名的泰勒模式(行为目标模式)，这一模式把确定教育目标作为评估过程的核心和关键，行为目标模式也因此而得名。从教育评估的历史来看，泰勒对教育评估发展的贡献突出表现在以下四个方面：

第一，泰勒在实验研究和评估实践的基础上首次提出了“教育评估”的概念，使评估与测验区分开来，并据此提出了富有创新意义的评估体系和模式。

第二，泰勒用具有学生行为对应物的具体教育目标作为评估标准，用预定的结果作为尺度来衡量学生的进步水平，从而避免了教育评估的任意性和主观性，在一定程度上提高了评估的客观性和科学性。

第三，泰勒的评估是一种目标参照评估，他注重的是绝对的教育目标实现的程度，而不是像以往测验那样只关心学生团体中的成绩差异和