PEARSON

语言测评：原理与课堂实践（第二版）

H. Douglas Brown
Priyanvada Abeywickrama
编著

Language Assessment: Principles and Classroom Practices (2nd Edition)

英语教师职业发展前沿论丛

清华大学出版社

PEARSON

语言测评：
原理与课堂实践（第二版）

H. Douglas Brown
Priyanvada Abeywickrama 编著

Language Assessment:
Principles and Classroom Practices
(2nd Edition)

英语教师职业发展前沿论丛

清华大学出版社
北京

北京市版权局著作权合同登记号图字：01-2012-7348

图书在版编目（CIP）数据

语言测评：原理与课堂实践：第2版 = Language Assessment: Principles and Classroom Practices, 2nd Edition：英文/（美）布朗（Brown, H. D.），（美）阿贝维克拉玛（Abeywickrama, P.）编著. —北京：清华大学出版社，2013

（英语教师职业发展前沿论丛）

ISBN 978-7-302-30630-6

Ⅰ. ①语…　Ⅱ. ①布…②阿…　Ⅲ. ①语言教学—教学评估—研究—英文　Ⅳ. ①H09

中国版本图书馆CIP数据核字(2012)第273431号

责任编辑：刘细珍
封面设计：常雪影
责任校对：王荣静
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：187mm × 235mm　　印　张：27.25　　字　数：550千字
版　　次：2013年1月第1版　　印　次：2013年1月第1次印刷
印　　数：1 ~ 4000
定　　价：65.00元

产品编号：047711-01

丛书总序

改革开放30多年来，随着我国与世界各国交流和来往的广度和深度的不断发展，国民英语水平得到了普遍与大幅的提升。在我国发展的各个不同历史时期，国家也会对各个层次的英语教学适时做出新的调整，提出新的要求。进入21世纪以来最近的一次大学英语教学改革，作为我国高等教育教学质量工程的一项重要内容，在教育部的领导下，整体规划，分步实施，措施得当，取得显著效果。经过近十年的改革，我国大学英语教学的状况发生了巨大改变，基于计算机和课堂的新型教学模式在全国各高校基本全面建立，“以学生为主体，以教师为主导”的教学理念基本被广泛认同，各高校都已基本建立与本校办学特色相适应的大学英语课程体系，且注重加强课程内涵建设，学生的英语综合运用能力和自主学习能力普遍得到提高。

改革走到今天，经历了阵痛，也看到了成效，但依然方兴未艾。广大的高校英语教师面临学生英语水平的提高，面临高校师资队伍建设的新形势，面临职称晋升不断抬高的门槛，在亲历了大学英语教学改革浪潮的洗礼之后，尤其感觉到了从事高校英语教师这份职业的不易、挑战与压力。从教育部到高校各级教学单位的管理层，也越来越意识到，高等学校大学英语教学质量是关系到提高我国高等教育质量、办人民满意的教育的大事，而要提高英语教学质量，除了要改革教学大纲、教材系统、考试体系、教学模式和教学手段，更重要、也是更内核的是要转变广大英语教师的教学理念，不断提升他们的专业水平和教学能力。

我国的大学英语教师，普遍来说都是从高校取得英语语言文学及相关专业学位之后，即直接开始从事教学工作，不少年轻教师并没有接受过有关教育学和教学法的系统培训。而一个显而易见的道理是：一个好的英语教师仅仅具备扎实的英语语言技能是远远不够的，并不是自身英语水平高的教师就一定能教出英语好的学生。要搞好英语教学，咱们的英语教师还须不断学习现代教育理论、外语教学理论和外语学科理论，优化和完善自身的知识结构，掌握现代教育技术，提升文化素养，拓展国际视野，并具备将理论知识真正融会贯通到具体教学当中去的能力，如制定教学大纲、设计教学方案、驾驭课堂、充分利用教学资源、有效管理学生、科学测评学生能力等各方面的能力。更为重要的是，英语教师还应具备在本领域中可持续发展的能力。这就需要广大英语教师具备自主的终身学习意识和动力，具备自我发展的动力和能力，教师职业的专业化发展能力成为新时期对教师提出的新的和更高的发展目标。

20世纪80年代以来至今，我国陆续出现了一些旨在帮助广大英语教师夯实理论基础、完善知识结构、更新教学理念、掌握新兴教学方法的著作。其中，既有从国外引进的，也有国内学者执笔的；既有偏综合性和理论性的，也有重实践和应用的。这些著作的出版，对于英语教师自我提升教学水平和科研能力，起到了非常重要的推动作用。此类著作目前在我国不是太多，而是太少。清华大学出版社外语分社历来就有重视教学研究的优良传统，此次经过精心策划和遴选，全新推出的“英语教师职业发展前沿论丛”是一套开放性丛书，今年先行推出第一批，今后还将根据我国广大英语教学工作者的需要不断进行补充和丰富。我有幸被邀请参与该套丛书的编委工作，看到这样一批优秀的国外前沿理论著作即将能在国内被引进出版，感到十分高兴。该套丛书特色鲜明，优势突出，其最大的特色与优势主要体现在以下几个方面：

一、出版社与作者并重，内容权威。该系列丛书中的每一本都是从美国Pearson出版集团和McGraw Hill出版集团等世界知名出版公司引进版权。作者均为当代国际著名语言教学专家，如David Nunan现任加州Anaheim大学副校长，并于2008年创建了David Nunan语言教育学院，曾荣膺2002年美国国会颁发的在英语教育领域中做出杰出贡献奖；H. Douglas Brown是美国旧金山州立大学教授，曾任该校美国英语研究所所长和《语言学习》杂志主编。他们都曾任国际TESOL组织主席，在全球语言教学与研究领域的影响力广泛而深远，也为我国广大语言学习者和教学研究工作者所熟知。这套“英语教师职业发展前沿论丛”选择的第一原则就是：出自名出版社的名家代表性力作。

二、经典与前沿并行，更关注前沿。该套丛书中有一些属于教学法方面的经典著作，如子系列“实用英语语言教学法”所包含的6本，分综述篇、听力篇、口语篇、阅读篇、语法篇、少儿英语篇，另外还有两部语言测试与评估领域的经典之作，都是从事英语教学与研究的工作者奠定基本知识框架和掌握基本教学技能所需要的得力助手。同时，清华大学出版社此次在遴选入选书目时，更为关注的是国际上语言教学领域的发展动态与前沿方向。如《根据原理教学：交互式语言教学》与《语言测评：原理与课堂实践》，引进的都是近两年新改版的最新版次，在权威、经典、全面的基础上又增加了新热点问题的论述，包括后教学法条件、多元智力、自主性与交流意愿二原则、评价的再组织原则、教师发展与反思性教学、社会责任、批评教育学、标准化考试领域的最新研究成果等。另外，计算机辅助语言教学（CALL）、语音教学和跨文化交际教学等这些近年来的热门领域，在该系列中也都能找到国际上目前最前沿的论著。

三、理论与实践结合，更重实践。这套丛书最突出的一个特点就是理论与实践的统一，每一本书都是以一套完备的理论体系作为支撑，最终服务于实践指导，具有很

强的实用性和操作性。子系列“教学点津”（Tips for Teaching）的每一本都着眼于非常具体的教学技巧，理论研究与教师教学实践相辅相成、有效融合，同时还在书中提供了丰富而具体的课堂活动设计及可复制的课堂活动材料，展现活动设计范例和具体操作指导，让教师能快速学以致用。如《教学点津：计算机辅助语言教学（CALL）实用方法》一书就展示了100多个与教学内容配套的CALL相关软件和网页的彩色截图，随书附带的光盘还针对各章内容提供了“演示”和“模拟”功能，既形象生动，又易于上手进行实际体验和操练；《教学点津：语音教学实用方法》也是图文并茂，讲解清晰具体，配套的音频CD光盘还提供了所有可供选择的课堂活动的听力材料。其他的所有著作无一例外也都是一部部真正能为教师提升教学效果指点迷津的实用指南，其实用性价值在同类学术著作中无可比拟。

《国家中长期教育改革和发展规划纲要（2010－2020年）》中提到：教育大计，教师为本。教育部也从今年开始，在全国高校范围选派骨干英语教师定期举办“高等学校大学英语骨干教师高级研修班”，大学英语教师专业水平和教学能力的提升和培训进入常态化。“英语教师职业发展前沿论丛”的出版对于我国广大英语教师及英语教学法研究者来说，犹如一场及时雨，必将为他们的职业发展助一臂之力，为打造一支业务精湛、结构合理、具有较强英语运用能力、熟悉外语教学理论、掌握现代教育技术的高素质专业化英语教师队伍起到积极的推动作用。

王守仁

2012年11月于南京大学

中文导读

1. 引言

语言测试是一门理论和实践并重的学科，伴随语言教学而出现。作为语言教师，多年来经历过无数次的测验和考试，编写测验也是常有的事，有颇多的经验感受。但这并不等于每位教师都能科学、恰当地使用测试。要用它公正、有效地评价学生的学业表现，检查教学目标的实现程度，有关测试的概念和基本原则教师必须熟悉，有些基本的测评技术和方法教师也应当掌握。广大教师在实施恰当有效的教学的同时，还应重视测试题目与评价任务的开发和设计。H. Douglas Brown 与 Priyanvada Abeywickrama 合著的《语言测评：原理与课堂实践》（以下简称《语言测评》）一书正是基于这样的思考。

H. Douglas Brown 教授长期任教于美国旧金山州立大学，曾任该校美国语言研究所所长、国际 TESOL 组织主席、《语言学习》杂志主编等，主要从事二语习得和二语教学方面的研究，研究领域包括以策略为基础的教学指导、课堂语言评估，以及如何将第二语言习得研究工作与课堂教学法相联系等，撰写过多部著作。本书另一作者 Priyanvada Abeywickrama 也曾执教于美国旧金山州立大学，主要从事教师培训和语言评估方面的研究工作。

2. 全书结构与总体评价

《语言评估》是培生教育集团出版的系列丛书中的一本，初版于 2004 年，2010 年修订再版。全书共 12 章。每章包括以下内容：学习目标、章节正文、练习题、建议阅读书目、附录、术语表、参考书目、人名索引、主题词索引。全书大致可分为四部分：第一部分（第 1-2 章）介绍了语言测试的基本概念和原则，对语言测试的发展以及课堂评估的热点问题进行了讨论。第二部分（第 3-6 章）分别讲述了课堂语言测试的设计、基于标准的评价、标准化测试，并在第 6 章专门探讨了测试以外的评价方式，如行为测试、学习档案袋、日志、访谈、观察、自我评价和同伴互评等。第三部分（第 7-11 章）从能力构念的界定、测试类型、试题设计以及评分等方面，分别探讨听力、口语、阅读、写作、语法和词汇等语言能力各方面评价

的实践问题。第四部分（第 12 章）重点讲述了成绩评定与学生评价的原则和指南，探讨了有关学生评价的常见问题，对教师实施评价提出了具体的建议。

本书是评估课堂语言能力方面的一本经典著作，主要特点有：

（1）面向实践。本书是一本语言测试教科书，是语言测试工作者、测试命题者、研究者、一线语言教师不可或缺的指导手册。本书兼顾大规模标准化测试和课堂语言测试的需要，有理论指导，更注重实践。本书追溯了语言测试的发展历史，从“相关概念”到“基本原则”，再到“测试实践”，讲得详尽透彻，通俗易懂，并辅以大量实例加以阐释，着力解决课堂语言评估中的实际问题，非常实用。在每章“练习题”部分，作者为读者设计了不同的任务和参考案例，力图使读者结合自己的课堂教学和学习，对所学内容加以巩固。此外还针对各章所探讨的话题，提供了相关的书目，供读者进一步研究。

（2）推陈出新。本书立足学科前沿，全面探讨了语言测试的热点问题，反映了语言测试的发展趋势，尤其突出计算机化语言测试、基于标准的评估、语言测试的伦理问题等热点话题，并结合自身长期教学和研究的实践经验，对旧版内容进行了充分的修订。改进的内容包括：1）作者在每章开头部分增加了本章的提要和学习目标，并在第一版的基础上对参考文献进行扩充；2）增加一章新内容（第 11 章　语法与词汇评价），并在个别章节增加了一些内容，如在第 6 章增加了“量规评价”，在第 12 章提供了成绩评定和分数计算方面的网络资源，方便读者查阅；3）对前三章的内容结构进行调整，并根据研究现状对第 4 章进行重新撰写；4）作者对一些关键术语在文内以粗体标出，在书后术语表中给出详细解释，还对常见的主题词在索引部分列出，供读者查阅；5）在附录部分，作者针对目前广泛使用的测试（如 IELTS、MELAB、OPI、TOEFL、TOEIC、TSE、Versant Test、TWE 等），从试题特点、测试规范以及相关网站等方面作了介绍。

（3）内容全面。作者对目前语言测试领域流行的各种理论进行了全面梳理和介绍，可以说，与语言测试相关的内容基本都讲到了。本书所探讨的用以评价学生语言能力的方法，既包括针对传统意义上的听、说、读、写以及语法与词汇等各项语言技能的测试，又包括学生自评、同伴互评、学习档案袋、学习日志、叙事评价、核查表、课堂观察、访谈等另类评价手段（alternative assessment）。所引材料翔实，涉及范围广泛，实例丰富，针对性强。在每章后面还附有相关推荐书目和简介，供读者进一步研究。读完本书，我们不难发现新的语言评价范式的转变：强调真实情境化的测试；强调运用多元评价方式，将评价融合到学习之中，促使从“对学习的评价（assessment of learning）”向“为学习的评价（assessment for learning）”转变，使学生参与评价过程，从而形成新的教学、学习、评价观，即教学侧重关注学生发展，学习重在反思性、主动的知识建构，而评价主要通过情境化、叙事性、基于真实表现的方式呈现出来。本书注重测试、评价与教学的衔接，重视评价对教学的导向作用。既关注终结性评价，又重视形成性评价，使评价成为促进学生语言能力全面发展、

改进教师课堂教学、帮助教师专业成长的重要手段。

当然，本书也有值得商榷之处：1）作者在谈语言测试的有用性时，并未谈到交互性（interactiveness）。实际上交互性也是评估测试质量的一个重要方面。此外，本书作者除了论述课堂的传统测试，也探讨了形成性评价和真实性评价等评价方式，但未提及动态评价（dynamic assessment）。动态评价是对传统测验的挑战和超越，是一种能力测评的新视角，它通过教学和干预等手段把学生的学习过程和学习结果结合起来，测查学生的未来发展水平或学习潜能；2）对有些概念未作澄清。例如，作者多次提到行为测试与基于任务的测试，但对两者的关系和区别谈得较少。又如，对于后果效度（consequential validity）、影响（impact）和反拨效应（washback）是否为一回事，有何异同，作者没有仔细讨论。作者在第 4 章专门论述基于标准的评价，但未明确说明标准（standard）和基准（benchmark）的区别。3）在讨论听力、口语以及写作测试时，作者先介绍相关能力构念的界定以及测试类型，再讲述测试任务的设计。但在第 9 章论述阅读测试时，作者却先讲阅读的类型，再说阅读能力和阅读策略，然后介绍阅读测试的设计，这一章的论述和结构安排与前面三章明显不一致。

当然，上述所指之处不影响本书的科学性和系统性。读者若能熟读此书，并辅读相关著作，足以对语言测试有个比较全面的了解。建议读者能将所学内容应用到测试实践中去，学会编写课堂测验和考试题目，灵活使用各种评价工具，对外能满足公众和社会问责（accountability）要求，对内也能满足促进学生发展的需求。

3. 各章节内容梗概

第一章　语言测评的概念和热点问题

考试是教育评估的一种手段，从课堂测验到大规模标准化考试，世界各国都把考试看作学业评价、水平诊断以及人才选拔的重要手段。作者认为，考试纵然关乎学生的前途和教师的切身利益，但不应该歧视学生或给其造成心理负担，而应该有利于学生学习，关注学生的成长与发展，帮其重塑自信，让学生体验进步与成功。这也是作者撰写本书的初衷所在。

测试与评价是成功教学的基础，也是诸多教育决策的依据。社会各界特别是教育领域十分关注语言测试和教育评价的学科发展。那么，什么是测试？什么是评价？二者与教学、学习、测验以及评估等概念有何联系与区别？语言测试的目的和类别又有哪些？语言测试的发展现状和趋势又是怎样？本章重点探讨这些问题。

对多数人来讲，提到“评价”和“评估”，就会想到测试，认为评价、评估、测试甚至考试都是可以互换的术语。实则不然。“Assessment”虽通常译作测试，但它包含“评价”之意，有别于“testing”。在教育领域，评价的含义更广、综合性更强，强调主观估计

与客观测验统一。作者指出，评价是指对个体的特质水平进行价值判断的过程，而测试是评价的一种手段。科学地讲，测试是衡量一个人能力、知识水平或在某领域的表现的测量工具（measurement）和方法（method）。它也是根据一定的程序和规则，对考生的行为进行量化和解释的过程（Bachmann，1990）。同时，作者指出，测试测量的是考生的行为表现（performance），而结果反映的是考生的能力（competence）。作者还对语言评价与测试中的相关概念与关系作了详细对比分析，主要包括评价与测试、测量与评估、评价与学习、非正式评价与正式评价、形成性评价与终结性评价、常模参照测试与标准参照测试等概念。

一旦选择测试作为评价手段，我们就应该明确测试的目的，并据此选择恰当的形式。依据主要目的和用途，作者把课堂内的语言测试分为以下五类：成绩测试（achievements tests）、诊断测试（diagnostic tests）、分班测试（placements tests）、水平测试（proficiency tests）以及学能测试（aptitude tests），并逐一作了介绍。

纵观历史，随着外语教学理论的发展，语言测试经历了几个不同的发展阶段。在20世纪40至50年代，受行为主义和对比分析法的影响，语言测试偏重对语音、语法、词汇等语言形式的考查。20世纪70至80年代，应用语言学界对语言的本质和教学理论有了新的认识，交际理论受到重视，在测试界也相应出现了交际语言测试，测试的真实性和有效性成为学界关心的热点问题。作者以语言测试的发展脉络为主线，重点阐述了分离式测试、综合性测试、交际语言测试和行为测试等。

20世纪中期，语言教学和测试深受行为主义心理学和结构主义语言学的影响，产生了心理测量学-结构主义语言测试。它注重对词汇、语法以及双语翻译等语言形式的考查，而忽视真实语境中的语言运用。同时，语言测试出现了分离式测试法。其主要特点是：语言可分解为语音、形态、词素、句法以及语篇等各种成分和听、说、读、写等语言技能，强调分别测试不同的语言成分和语言技能，常用的题型是多项选择题。

由于分离式测试脱离语境，缺乏真实性，在交际法盛行的今天，其缺陷日益显现。与此相反，综合测试法（the integrative approach）侧重测试考生综合运用语言的能力，主张将语音、词汇、语法以及语篇等语言成分和听、说、读、写等技能综合起来，从整体上对学生的语言能力进行测试，它考虑到了语境的重要性。作为综合测试法的支持者，Oller更是将其与分离式测试对立起来。他提出“一元能力假设”（UCH），认为语言能力不是分离的，语言知识中的语音、语法、词汇等项目和听、说、读、写等方面的技能不能逐个地测试，而是在一定语境中综合地得以运用。综合性测试的形式主要有完形填空、写作、听写、口试等。

近几十年，语言教学方法逐步向交际式语言教学转变。交际法理论的核心是关于交际能力（communicative competence）的概念。这一概念由Hymes在1972年首先提出，在20世纪80年代经Canale与Swain加以补充，后由Bachman（1990）作了进一步完善，从而

丰富和发展了交际语言测试。交际语言测试最大的特点是测试内容的真实性(authenticity),它重在测试考生在真实语言环境中完成真实语言任务的交际能力。Bachman（1990）认为，语言能力由语言组织能力和语言使用能力组成。Bachmanh 和 Palmer（1996）同时强调策略能力在语言交际中的重要性。作者指出，在语言测试的设计过程中，我们需要考虑 Bachman 语言能力模型中的各个要素，尤其是语用和策略能力，应注重对真实测试任务的考查。

随着交际教学法的流行，以学生为中心（student-centered）的教学范式备受推崇。语言测试也由注重语言形式转向重视技能，再转向重视语言能力的实际运用，这种测试被称为行为测试（performance-based assessment）。它不同于传统的以多项选择题为主的客观语言测试，而是根据学生在现实任务（如口试、写作、开放式问答题、综合技能表现、小组表现以及其他互动任务等）中的具体表现，来直接测定被试者的语言行为表现和完成任务的能力。这种测试耗时耗力，但它是基于真实任务的语言测试，内容效度较高，而传统的纸笔测验则无法检测这样的交际能力。

受多元智力理论和建构主义学习理论的影响，传统的标准化测试饱受非议。测试专家和教师开始反思其弊病，倡导用非传统的、真实的行为测试取代标准化测试，以适应新的教学需要。同时，语言测试界也出现了一些新的研究视角。作者主要探讨了以下三个方面的热点问题：多元智力理论、非传统测试和计算机化语言测试。

传统智力理论认为，智力是以语言能力和数理逻辑能力为核心的。传统的标准化 IQ（Intelligence Quotient）测验脱离现实情境，其结果只能反映出学生在某个测验上的一次表现，无法对学生的多元智力给出完整描述。有鉴于此，Howard Gardner（1983，1999）将传统的智力观拓展为八个方面，奠定了多元智力的理论基础，分别是：逻辑 - 数学、语言、空间、音乐、身体 - 运动、自然认知、人际和自我认知。此外，Robert Sternberg（1988,1997）还将创造性思维与操纵策略看作智力的一部分，Daniel Goleman（1995）更是强调情感在认知加工中的重要性，提出情商（Emotional quotient，简称 EQ）的概念。不过，要将它们作为独立的智力提出来，还缺少足够的证据。目前，多元智力理论已经被用在教育、语言教学以及语言测试领域。交际性语言课堂日益重视学生学习能力的多样性，强调测量学生的人际沟通、全语言技能、创造性思维、学习过程以及意义协商的多维能力。

近年来，传统测试的统治地位受到了猛烈的冲击，诸多冠以“行为测试”（performance assessment）或者“另类评价”（alternative assessment）之名的新型非传统评价方式得到广泛应用。作者首先归纳了传统测试与非传统测试方法之间的区别。但同时指出，二者实则很难划清界限，很多评估形式介于两者之间，或二者兼而有之。我们不能认为传统的测试就一无是处。

随着计算机技术的发展，计算机被广泛运用于第二语言教学与测试。基于计算机的语

料库在语言测试中的应用备受关注，人们逐步将语料库用于语言测试的开发、选材、命题、评分等各个阶段。同时，计算机适应性测试（computer-adaptive test，简称 CAT）也得到迅速发展和应用，成为语言测试界的研究热点之一。作者认为，CAT 突破了传统纸笔测试的局限，具有许多优点：1）能用于各种课堂测试；2）可根据应试者的能力，自我调适测试项目的难度；3）可用于高风险标准化考试；4）提高了测试的个性化程度；5）可同时在不同场所用于大规模标准化考试，评分快捷；6）可用于开发自动作文评阅系统和语音辨识测试。尽管计算机化语言测试有上述诸多优点，但它也存在许多局限性，具体表现在：1）由于监管缺失，试题的保密性不高；2）试题质量缺乏可靠性；3）缺乏实用性或可行性；4）缺乏真人之间的交互性；5）测试任务缺乏真实性。

除了上面谈到的话题，作者还对直接性口语与写作测试、语料库语言学的进展、基于标准的评价、后果效度等热点问题作了探讨。更多的研究热点将在后面的章节中详细介绍。作者希望通过此书，让广大读者更深刻地体会测试在评价中的地位以及评价与教学的相互关系。最后，作者认为：1）定期评价，包括正式和非正式评价，能衡量学生的进步程度，可以增强学生学习的动机；2）合理的评价有助于强化和储存重要信息；3）评价能够发现学生的强项和不足；4）评价可以定期检测教学内容；5）教师应鼓励学生参与自我评价，增强其学习的自主性；6）评价能激发学习者为自己设定目标；7）评价有助于评估教学效果。

第二章　语言测评的原则

本章作者主要讲述语言测试的原则以及如何将这些原则用于评价一项考试和其他评价方式。如何知道一项测试是有效、合理、有用或者是高质量的呢？它符不符合语言测试的要求呢？作者认为评价一项测试，我们需要综合考查它的可行性（practicality）、信度（reliability）、效度（validity）、真实性（authenticity）和反拨效应（washback）。这些原则归纳起来体现了测试的有用性。

作为教师或命题者，我们纵使有许多美好的愿望和想法，但不一定都行得通。我们的工作总是受到财力、时间、人力等因素的制约，不得不调整或作某种妥协。因此，要搞好语言测试，从命题到施测，再到评分，每个阶段都需要考虑测试的可行性（practicality）问题。语言测试的可行性主要是指：从物力或财力、时间、人力资源、评分、施考以及成绩报告等环节上，测试能否得以实施，是否可行。作者认为，可行性主要体现在以下方面：测试开发和维护成本保持在预算范围之内；考生可在规定时间范围内完成测试任务；施考指南清楚、明晰；人力资源得到合理利用；施考所需的资源配置比较科学；试题设计和评分所需的时间和物资分配均衡。

语言测试质量评估要考虑的第二个问题，是测试结果的可信程度，即信度（reliability）。谈测试的信度，实际上谈的是测试结果的一致性和稳定性。作者认为，一项信度可靠的测

试应具备如下原则：受试者在不同场合测试结果应前后一致；评分指南应清楚详细；应有统一的评分标准；评分者使用评分标准时应保持稳定和一致；测试题目和任务清晰明了。作者同时指出，这种一致性和稳定性可能是关于受试者的，关于评分者的，也可能是施考或测试本身的，或者是关于以上几种因素不同组合情况下的。作者主要对受试者信度、评分者信度、施考信度以及测试信度作了介绍。

效度是评价测试质量的另一重要指标。效度，顾名思义，就是一次测试的有效程度，或者说，测试是否考了它所要考的。效度表明一种相关性，即测试结果与测试目标的关联程度。作者从多个角度概括了效度的基本特征，即：即测试应该测量它所要测量的内容；不测量不相关的变量；依赖经验证据进行效度验证；提供反映考生能力的有用信息；以理论依据为支撑。在具体讨论测试的效度时，作者主要从内容效度、效标关联效度、构念效度、后果效度、表面效度等几个方面进行考虑。实际上，效度是一个复杂的概念，反应效度或应答效度（response validity）也影响考试效果，它反映了考生答题时的策略和风格，即是否按试题设计的要求去做出应答。作者在这里并未提及。

真实性是语言测试又一重要原则。这一原则对于我们开发、评价某项考试是极为有用的。什么是语言测试的真实性，测试界对此看法不一，难以界定，因为人们的判断往往是主观的。Bchmann 和 Palmer（1996）认为，语言测试的真实性指目标语言使用任务特征与测试任务特征的一致性。一致性越高，测试的真实性就越强。实际上，完全的真实性是无法达到的，测试材料只能尽可能反映真实。作者认为，语言测试的真实性主要体现在以下方面：试题由自然语言构成、有语境提示、包含意义相关且有趣的话题、提供真实情景的任务。

反拨效应是语言测试的一种后果效度（consequential validity），有学者也将其称为超考试效度（beyond-the-test validity）。考试对教学有着无可否认的反拨效应，可以是正面的，也可以是负面的。如果考试给教学带来一种良好的导向作用，这就达到超过考试本身更重要的目的，就算具有好的反拨效应，反之则具有负面的反拨效应，不利于教学和学习。作者认为，测试的正面反拨效应体现在：1）对教学有积极影响；2）对学习有正面导向作用；3）利于学生备考；4）促进学生的语言发展；5）属于形成性评价而非终结性评价；6）为学生的最佳表现创造条件。

作者最后指出，语言测试涉及多个学科和相关领域，要设计有效的测试远不止这五条原则。在本章最后部分，作者重点探讨了如何利用语言测试的五个原则去评价课堂语言测试。作者建议，在评价一项考试时应考虑以下 8 个问题，分别是：1）测试程序是否可行（practical）；2）测试本身是否可信（reliable）；3）评分者信度是否有保证；4）测试是否具有较高的内容效度；5）测试的影响（impact）是否预先告知考生；6）测试程序是否公平；7）测试任务是否真实（authentic）；8）测试对学习者有无良好的（beneficial）反拨作用。

第三章 课堂语言测评的设计

前面介绍了语言测试的基本概念和原则，本章重点讲述如何利用这些概念和原则去设计课堂测试任务或修改现行的试题。命题是测试的中心环节，考什么和怎样考，对教与学直接起着“指挥棒”的作用。开发语言测试，一般要遵循一定的步骤。每一步需要做什么，每一阶段要拿出什么成果，这些都是测试成功的保证。在设计测试任务之前，作者建议命题人员对测试目的、考查目标、测试规范、题目筛选与设计要求、施考、以及评分步骤和反馈程序等问题要有清楚的认识。作者主要以阅读测验（Reading Quiz）、语法单元考试（Grammar Unit Test）、中期写作（Midterm Essay）、听说期末考试（Listening/Speaking Final Exam）等四个测试情景（Senario）为例，讲述了课堂测试设计的六个步骤，希望能够为命题人员在课堂测试设计和施测方面提供指导。

第一，确定测试目的。这是测试命题最重要的一个环节。测试目的决定着测试内容和具体目标，进而影响测试说明的设计、试题的编写以及评分标准的确定。测试的目的，Bachman 和 Palmer（1996）亦称之为测试的有用性（test usefulness），即用测试做什么。作者指出，要真正了解测试的目的和有用性，我们首先应弄清以下问题：1）是否有必要实施考试？若必要，测试的目的何在；2）测试对该课程有何意义；3）测试是否是衡量教学质量和学习效果的理想方式；4）与其他学生表现相比，测试的重要性有多大；5）是否需要用考试结果来判定学生的学习效果；6）测试能否对学生产生正面反拨作用；7）能否将测试结果作为后期教育资源的分配手段；8）测试对学生和教师有何影响。

第二，设计测试目标。为了保证课堂测试能够考查与教学任务相关的代表性样本，并能够测得学生所掌握的知识、技能和理解程度，很重要的一点是设计具体的测试目标或操作细目表，以此来指导测试题目的选择。作者以“语法单元测试”为例，谈了设计具体测试目标的要求和注意事项。

第三，编写测试规范。在正规考试中，除规定测试目的和目标外，一般都有测试规范来指导出题者编写试题。测试规范是编写试题的纲领性文件，它对测试内容、试卷各部分题型（如多项选择题、完型填空等）、测试任务类型（如写作、阅读等）、技能要求、评分方式以及成绩报告形式等都有明确规定。测试规范通常比较正式，对内容是严格保密的。

第四，编写测试题目。题目编写是测试开发的重要部分，但这一过程并不总是一帆风顺，总会遇到各种意想不到的问题。作者以中期写作和期末听说考试为例，从课程目标要求、考生特征、题目类型、题目指示语、题目的数量、时间分配、评分方法等方面阐述了题目编写过程中应注意的事项。作者强调，命题人员应综合考虑试题的可行性、信度、效度、真实性以及反拨作用等因素，全方位评价试题的质量和有用性。接着，作者重点讲述了多项选择题的特点和设计要求。多项选择题是最为通用的一种测试题型。与其他类型的

测试题型相比，它有若干优点。由于每道题只有一个正确答案，评分相对简单，而且分数分析也相对容易。除此之外，多项选择题可循环使用，其可行性和信度相对较高，因此被大量用于各类测试中。然而，多项选择题也饱受质疑。Hughes（2003）曾列举了多项选择题的局限性：1）只测试考生的识别能力；2）猜测（guessing）因素对测试结果影响较大；3）要编写可靠、高质量的题目并非易事；4）对测试内容有严格的限定；5）具有不良的反拨效应；6）考生容易靠旁窥或打暗号作弊。多项选择题看似容易设计，实际上是特别不容易命好的题目，通常比较耗时且需要专门技能。作者认为，在设计多项选择题目时，应遵循下面四个原则：1）一个题目只测一个语言要素或测试目标；2）题干和选项尽可能简洁、清楚；3）确保一个题目只有一个正确或最佳答案；4）用测试项目的评估指标来选择、剔除或修改测试题目。在大规模常模参照性测试中，测试项目分析很有必要。常用的项目分析指标有：项目难度（item facility or difficulty）、项目区分度（item discrimination）和项目干扰项的分析（distractor analysis）。至此，测试题目设计好以后，我们就要全面核查各测试题目，必要时可进行修改。关于测试题目的修改，作者提出了 10 条建议。作者指出，如果条件允许，可在小范围内进行试测，或让同行从头复核一遍，包括考试所需时间、试卷的结构等都要进行仔细的检查，再考虑定稿。

第五，施考。为了确保施考顺利进行，保证考生最大程度发挥其语言水平，作者认为考试工作人员应注意以下事项：1）告知考生考试要求（如考试时间、考试工具、考试题型、答题策略以及评分标准等）；2）宣读考场规则；3）回答考生提出的与考试相关的问题。同时，作者建议监考人员做到以下方面：1）提前到考场检查考场的布置是否到位，如灯光、温度、座位安排等；2）如需特殊技术设备，应预先调试；3）备用一些答题纸和常用的书写工具；4）准时分发试卷；5）坐在指定的位置监考；6）如果是限时考试，应提醒考生考试结束时间，督促其按时完成答卷。最后作者指出，这里所列举的只是施考过程常见的注意事项，也许在特定情况下还会有其他情形并未列在其中。

第六，评分与成绩反馈。要设计课堂测试，课堂教师一定要考虑如何评分。教师应综合考虑试题难度是否合适、时间分配是否合理、考生的答题反应如何、试题是否全面考查了学生的水平等因素，根据需求修改评分计划，以适应将来教学和测试的需要。实际上，要把大量评价数据概括为只有一个字母等级（例如 A、B、C、D、F 等）或一个数值，是一件很困难的事情。作者认为，如何给学生评定成绩，取决于一个国家的教育体制、文化、英语课堂实际、对等级的界定与理解以及考生对考试的期望程度等诸多因素。关于具体的等级评定方法，作者在第 12 章会作详细介绍。

学生的成绩可服务于多种功能，比如，向学生和家长反馈学生的进步程度、改善教学、便于教学管理和指导等。作者指出，反馈和报告学生进步程度的方法有很多，常用的有：字母等级评定法、总分数评定法、分项计分法、诊断评分系统、目标核查表、课堂讨论、

同伴评议、学生自我评定以及教师的口头评价等。

第四章 基于标准的测评

我国的考试有着悠久的历史，早在2000多年前的汉代，就用正式的考试来选拔官员。如今，我们仍然深受考试的影响，尤其是大规模标准化考试。多年来，无论是在学校、企业、政府，还是其他部门，标准化测试在成绩考核、选拔人才等方面都发挥着重要作用。“标准”是标准化考试的基本指标。既然是标准化考试，就应该在建立测量标准上下功夫，如命题标准、质量标准、评分标准等。可是，测试的“标准”（standards）又是什么？标准从何而来？为什么需要标准？需要什么样的标准？标准的效度如何？如何科学使用标准？诸如此类的问题，自考试产生以来就始终存在，探讨者代不绝人，概念也不无出新。这也是本章重点要探讨的问题。

作者首先讲述了标准化测试中标准的作用。标准化测试对测试目标或能力表现水平有明确的规定，并将标准应用于课程设计、教学实践、年度计划和学业评估之中。具体有关标准化测试的问题，将在第5章进行讨论。在基于标准的评价中，标准（standards）的含义既非标准参照测验中的“criteria”，也非直接作为评价结果基准的“benchmark”，而是国家对中小学生接受一定阶段的学校教育后，应该知道什么（what students should know）和能够做什么（be able to do）所做出的统一规定和表述，反映了国家对教师业绩和学生学习成就的期望，关于此，我国通常称之为课程标准，主要包括内容标准和表现标准。前者划定了学习的范围，回答学生应“学习什么”的问题，后者则规定学生在学完这些内容后应达到的水平，回答学生“能做什么”的问题。同时，标准也是教材编写、教学、评估和考试命题的依据。

如今，许多国家都已实施基于标准的教育（standards-based education），并制定出相应的学习标准（standards for learning）。欧洲国家为了尝试不同语言之间的学分认证，共同制定了《欧洲语言教学共同参考框架》（*Common European Framework of Reference*，简称CEFR）。同时，欧洲语言测试协会也制定了相应的测试与评估标准。标准运动对美国的教育公平性同样影响巨大。为了能给来自不同语言和文化背景的二语（ESL）或英语学习者（ELL）提供学习和能力发展机会，TESOL组织制定了ESL教育标准。作者主要介绍了Short（2000）有关ESL的三个教育目标，其中每个目标分别有三个内容标准。这三个目标是：1）能在社交环境中用英语进行交流；2）用英语能完成所有的学习内容目标；3）能恰当地进行社会和文化交流。2001年，美国联邦政府开始实施《不让一个孩子掉队》（*No Child Left Behind*，简称NCLB）的教育法案，旨在缩小所有学生之间学习成绩的差距，促进人人享受优质教育。

基于标准的教育改革，其工作重心是开发全国性的标准。只有制定标准的程序合理，

才能保证结果科学、有效、可信。这就需要课程专家、测试专家、教师和研究人员共同合作完成。如何评价英语学习者的语言能力已经成为一个国际性的课题，也是各国教育改革的重要部分，各国都在根据国情制定本国的英语语言能力评价标准。但对大多数学校、地区和国家来说，这仍是一个巨大的挑战。

在制定和开发评价标准的基础上，教育工作者需要设计与此标准相匹配的评价工具。基于标准的评价需要标准本身的进一步完善。如何基于标准进行评价有多种途径和方式，而课程标准、教学与评价之间的一致性（alignment）是基于标准的评价中很重要的方面，也是确保评价能否真正反映课程标准和教学过程的关键环节，需要对此进行深入研究。作者主要介绍了加州英语语言发展测试（California English Language Development Test，简称CELDT）。它由一系列测试题目组成，旨在测试跨年级英语语言发展水平的成就。具体内容作者在附录部分有专门介绍。

在欧美基于标准的改革运动之后，“Standards-Based Assessment（SBA）”成为大家默认的专有名词。如今，基于标准的评价在许多国家和地区非常普遍，它在高水平教育阶段（大学、社区大学、成人学校、语言学校）有着巨大的影响。成人学生综合评价系统（Comprehensive Adult Student Assessment System，简称CASAS）是针对全美国非母语英语课程（English as a Second Language，简称ESL）设计的全面评价系统。该系统涵盖80多个标准化测试工具，用以给学生分班、诊断学习者的需求、追踪并掌握进度、认证基本技能掌握程度等。除了检测读、写、听、说技能外，还包括高层次思维能力。与CASAS的评价体系相似，美国劳工部制定的SCANS报告（Secretary's Commission in Achieving Necessary Skills，简称SCANS），提出了每一个进入劳动市场的人所必备的关键技能，大体可分为五种基本技能：学习技能、思考技能、交流技能、技术技能、人际交往技能，主要涉及资源分配、人际交往、信息处理、制度理解以及技术应用五个方面。

学生的考试表现取决于教学质量，而教学质量又是教师专业化发展的结果。在制定学习标准的同时，也有必要对教学标准（standards for teaching）做相应的规定。教师专业标准已成为许多国际英语教学协会（TESOL）关注的焦点。Kuhlman（2001）认为教师标准在语言专业技能发展、语言文化交流以及教学规划与监管等方面发挥着重要的作用。不过，如何评估教师是否达到相应的标准，是一个复杂的问题。TESOL教师标准委员会提倡用行为测试来测试教师的语言专业知识、语言文化交流水平等。基于标准的教学和评估虽面临许多挑战，但其社会影响不可忽视，尤其在学生评价方面。

基于标准的评价有许多明显的优点，标准的实施固然能促进和完善教育体制，但它也会产生许多意想不到的负面后果。首先，缺乏针对基于标准的评价的问责机制；其次，基于标准的教育往往与标准化测试（如SAT、GRE、TOEFL等高风险标准化测试）关联密切，其后果效度或反拨作用值得我们思考。