Mc Graw Hill Education

语言项目中的测试：英语语言测评综合指南

James Dean Brown 编著

Testing in Language Programs:

A Comprehensive Guide to English Language Assessment

英语教师职业发展前沿论丛

Mc Graw Hill

清华大学出版社

Mc Graw Hill Education

语言项目中的测试：
英语语言测评综合指南

James Dean Brown 编著

Testing in Language Programs:
A Comprehensive Guide to English Language Assessment

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：01-2012-7857

James Dean Brown
Testing in Language Programs: A Comprehensive Guide to English Language Assessment
ISBN: 0072948361

图书在版编目（CIP）数据

语言项目中的测试：英语语言测评综合指南 = Testing in Langusge Programs: A Comprehesive Guide to English Language Assessment: 英文/（美）布朗（Brown, J. D.）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 978-7-302-30889-8

Ⅰ. ①语… Ⅱ. ①布… Ⅲ. ①英语—语言教学—研究—英文 Ⅳ. ①H319.3

中国版本图书馆CIP数据核字（2012）第291706号

责任编辑：蔡心奕
封面设计：常雪影
责任校对：王凤芝
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：**20.5　　**字　　数：**498千字
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：48.00元

产品编号：047709-01

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

语言测试是语言教学的重要组成部分，也是教育决策者和语言教师所关注的热点问题。目前，语言测试相关的书籍或教材主要侧重测试理论研究（如 Bachman, 1990; Bachman & Palmer, 2010; Alderson et al, 2000）、具体命题技巧（Heaton, 2000）或试题分析（Henning, 1987），主要面向的是一般性的考试，特别是大规模考试。但在实际外语教学中，教育决策者、语言教师和学生对语言测试的关注点和需求都不尽相同。Fulcher 和 Davidson（2007）系统地介绍了大规模考试（large-scale testing）和课堂测试（classroom-based assessment）的区别。随着语言教学的发展，课堂语言测试受到了越来越多的关注（Hill & McNamara, 2011），因此也需要兼顾不同读者群和针对不同测试类型的相关书籍。本书作者 Brown 教授基于多年的教学和研究经验，充分考虑到了语言测试的使用目的，从决策角度和教学角度将语言项目中的测试分为常模参照（norm-referenced）和标准参照（criterion-referenced）两种主要类型，旨在为语言项目决策者和语言教师提供实用、有效的工具和方法，以便他们各司其职，根据需求运用不同的测试手段，使更多的学生受益。

本书的主要框架分为前言、章节（1—11 章）、答案、术语表和索引、参考文献五个部分。每章内容的基本结构：先是相关理论和概念的介绍，进而阐述和讨论其在不同情境中的实践和运用，接着运用 Excel 工作表针对本章节的数据分析进行操作指导，然后以问题方式对本章节进行总结和回顾，最后是应用练习。

从内容结构上，本书的十一个章节可以分为以下三部分：

第一部分为 1—3 章，主要介绍了语言测试中的基本概念，以及如何选用和设计测试类型和题型。第一章首先介绍和讨论了两种不同语言测试类型（norm-referenced testing，criterion-referenced testing）的特点和区别，以及从决策角度和教学角度这两种测试类型所具有的四种主要功能，作者还解释了为何不存在可以同时满足四种功能的“万能考试”，本章的最后简单介绍了微软 Excel 工作表的入门知识；第二章着重探讨了在选用和设计语言测试时应考虑的理论和实际因素：前者包括语言教学法、语言能力和行为 / 表现之间的区别、孤立语言点考试和综合语言技能考试之间的差异等因素；后者包括考试公平性，考试成本和后勤保证等因素。本章还以表格的形式为考试的选用和设计提供了详细的参考指南；第三章主要讨论如何命制高质量试题，首先阐述了高质量考题的定义，提出了命题总则，接着介绍了三类不同的题型（receptive response item, productive response item, personal response

item），并给出了具体的命题指导。

第二部分为 4—7 章，主要介绍测试结果的量化分析和解读。第四章主要解释了两类测试（norm-referenced testing, criterion-referenced testing）中通常使用的试题分析方法。常模测试中主要分析题目的难度系数和区分度，作者通过 Excel 表格实例向读者演示了如何进行数据录入和相关运算；标准参照测试首先通过考察测试内容和测试目的的一致性对题目质量进行分析，然后根据区分指数（difference index）和 B 指数来反映考生在同一题目上的表现或成绩差异；第五章首先介绍如何描述和呈现考试结果，呈现形式主要为点状图、柱状图和折线图。然后讨论了三种不同的计量量表（类别量表、序位量表、连续量表）的区别及应用。接着作者介绍了描述统计学中的两大基本概念：趋中度（central tendency）和离散度（dispersion），并结合测试数据展示了相关公式和方法。趋中度主要包括均数（mean）、众数（mode）、中数（median）、中值（midpoint）；离散度包括全距（range）、标准差（standard deviation）、方差（variance）；第六章主要介绍如何对常模参照考试和标准参照考试中考生的测试结果进行分析和解读。在第五章的基础上，作者介绍了分数统计分析中的三个主要概念：概率分布（probability distributions）、正态分布（normal distributions）和标准分（standardized scores）；第七章的作用是承前启后，主要对语言测试中相关度这一概念及其计算方法进行介绍和说明，为后一部分信度和效度的量化分析提供相关统计分析准备。

第三部分为 8—11 章，主要介绍和讨论语言测试中的两大基本要素：测试信度和效度。第八章主要探讨在常模参照测试中的信度问题，作者首先讨论了导致测量误差的可能性因素，如测试环境、测试流程、评分流程、试卷质量和考生状态等。作者主要从试卷信度和阅卷员信度两方面来探讨语言测试的信度问题和测量方法；第九章主要介绍了在标准参照测试中的信度（用一致性和可靠度表示）及其检验方法。作者具体介绍了四种可靠度计算方法：临界缺损一致性方法（threshold loss agreement approaches）、平方误差缺损一致性方法（squared-error agreement approaches）、域分数可靠度方法（domain score dependability）以及置信区间方法（confidence intervals）。本章节涉及的统计概念和计算方法较为复杂，读者可根据需求进行选择性阅读；第十章探讨了语言测试中的另一重要概念，即效度问题。作者首先介绍了适用常模参照和标准参照两类测试的效度检验策略：内容效度和结构效度。然后介绍了主要适用于常模参照测试的标准相关效度检验策略。接着作者探讨了测试标准与信度和 / 或效度之间的关系。最后本章介绍了其他与效度相关的因素，特别是测试对教学的反拨效应，并就如何提高测试的正面反拨效应提出了具体建议；第十一章主要阐述了如何将语言测试融入语言教学体系中，并探讨了语言测试在课程设计和实施中的作用。

2012 年夏，笔者有幸结识了本书作者 Brown 教授的博士生 Soo Jung Youn 女士，她在夏威夷大学为本科生和研究生讲授语言测试课程，并使用本书作为这门课程的主要教材。

她从语言测试研究者和课程教师的角度总结了本书的如下优势：

1. Brown (2005) provides very thorough and comprehensive explanations on basic concepts in language testing with a balanced perspective of both theoretical and practical aspects.

2. It easily explains statistical concepts especially for language teachers. At the same time, the book is very accessible for researchers who are interested in quantitative language testing research.

3. The new edition focuses on "doing" aspects of language testing using Excel.

4. The book particularly emphasizes how quantitative analyses can be done for both NRT and CRT contexts, while many other testing textbooks emphasize NRT-related contexts.

5. During the last two years, I've consistently received positive feedback from students on Brown (2005), such as "will keep the book for professional career", "very easily written and accessible", or "a must-read book for all language teachers".

以下是 Soo Jung Youn 女士通过自己两年的课堂教学经验得出的针对使用本书作为语言测试教材时的相关建议：

1. For less experienced students: Every chapter needs to be thoroughly covered. However, depending on students' needs and knowledge in statistics, chapters that deal with quite technical aspects, such as Chapter 9, can be skipped or emphasized less.

2. For more experienced students: Chapters 1, 2, and 3 can be briefly covered and the instructor can quickly move onto later chapters depending on their current knowledge. Each chapter can be covered sequentially, but supplementary readings from other textbooks (e.g., Fulcher & Davidson (2007)) would also be useful to include, especially on validity, performance assessment, and inferential statistics.

3. Regardless of students' levels, review questions presented in each chapter can be used as part of quizzes or a midterm.

4. Additional quantitative data is useful: The instructor can include quantitative test data from their institutions (e.g., placement test results) for Chapters 4, 5, 6, 7, and 8. In addition to data presented in the book, actual data can be very useful for additional hands-on Excel exercises and to apply quantitative analyses (e.g., item analysis, descriptive statistics, correlation, reliability) to real data. By doing so, it will actually help students discuss how to revise the existing tests.

总之，本书将理论和实际相结合，紧扣课堂实际和教师需求，针对性强。每章的数据分析都有手把手的实际操作指导，每章后的总结和思考题便于读者梳理和消化所学知识，设计的练习和所附答案有助于巩固加强对已学内容的掌握和应用，有助于自学。本书适用

面广，读者群体可包括教育决策者、教学主管、语言教师、测试命题人员、测试研究人员以及相关学科研究生等。

参考文献

1. Alderson, J. C., Clapham, C. and Wall, D. 2000. *Language Test Construction and Evaluation*. Beijing: Foreign Language Teaching and Research Press.
2. Bachman, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
3. Bachman, L. F., & Palmer, A. 2010. *Language Assessment* in Practice: *Developing Language Assessment and Justifying their Use in the Real World*. Oxford: Oxford University Press.
4. Fulcher, G., & Davidson, F. 2007. *Language Testing and Assessment*: *An Advanced Resource Book*. New York: Routledge.
5. Heaton, J. B. 2000. *Writing English Language Tests*. Beijing: Foreign Language Teaching and Research Press.
6. Henning, G. 1987. *A Guide to Language Testing*. Boston: Heinle & Heinle Publishers.
7. Hill, K. & McNamara, T. 2011. Developing a comprehensive, empirically based research framework for classroom-based assessment, *Language Testing*, 29(3): 395-420.
8. Youn, S. J. 2012. Personal Communication.

张文霞

2012 年 11 月于清华园

This book is dedicated with love to my mother, Jeanne Yvonne Brown.

Among many other things, she taught me to love books.

Preface

As is often true in the language teaching field, this volume had its roots in a class that I teach quite regularly—in this case, a graduate-level course in language testing. While many books exist on language testing, none seemed to offer the types of information that I wanted to present in my class. I felt that some books were too technical and complex to be thoroughly covered in one semester, while others were too practical—offering many ideas for different types of language test questions, but very little on test construction, analysis, and improvement. As a result, this language testing book is designed to cover the middle ground. I have tried to provide a balance between the technical and practical aspects of language testing that is neither too complex nor too simplistic.

My overall goal was to provide information about language testing that would not only be immediately useful for making program-level decisions (e.g., admissions and placement decisions), but also information about testing for classroom-level decisions (i.e., assessing what the students have learned through diagnostic or achievement testing). These two categories of decisions and the types of tests that are typically used to make them are quite different.

The category of tests most useful for program-level decisions consists of tests specifically designed to compare the performances of students to each other. These are called norm-referenced tests because interpretation of the scores from this category of tests is linked closely to the notion of the normal curve (also known as the "bell" curve). Such tests are most commonly used to spread students out along a continuum of scores based on some general knowledge or skill area so that the students can be placed, or grouped, into ability levels. The administrator's goal in using this type of test is usually to group students of similar abilities together in order to make the teacher's job easier. In other situations, the administrator may be interested in making comparisons between the average proficiency levels of students in different levels, between different language institutions or among students across the nation. Norm-referenced tests are also appropriate for language proficiency testing. Notice that the purpose of the tests in the norm-referenced family is to make comparisons in performance either between students within an institution (for placement purposes) or between students across courses or institutions (for proficiency assessment purposes). In short, sound norm-referenced tests can help administrators and teachers do their jobs better.

In contrast, the criterion-referenced family of tests is most useful to teachers in the classroom (though administrators should be interested in these tests as well). Criterion-referenced tests are specifically designed to assess how much of the material or set of skills taught in a course is being learned by the students. With criterion-referenced tests, the purpose is not to compare the performances of students to each other, but rather to look at the performance of each individual student vis-à-vis the material or curriculum at hand. They are called criterion-referenced tests because interpretation of the scores is intimately linked to assessing well-defined criteria for what is being taught. Such tests are often used to diagnose the strengths and weaknesses of students with regard to the goals and objectives of a course or program. At other times, criterion-referenced tests may be used to assess achievement, in the sense of "how much has each student learned." Such information may be useful for grading student performance in the course, or for deciding whether to promote the students to the next level of study, as well as for improving the materials, presentation, and sequencing of teaching points. In short, sound criterion-referenced tests can help the teacher do a better job.

My primary motivation in writing this book was to provide practical and useful testing tools that will help language program administrators and teachers do their respective jobs better. The distinction between the norm-referenced and criterion-referenced tests will help administrators and teachers focus on the respective types of tests most appropriate for the kinds of decisions that they make in their work. Hence the topic of each chapter will be approached from both norm-referenced and criterion-referenced perspectives. After all, the decisions made by administrators and teachers affect students' lives, sometimes in dramatic ways, involving a great deal of time and money, other times in more subtle ways, including psychological and attitudinal factors.

I assume that teachers, though most interested in classroom tests, will also take an interest in program-level decisions. Similarly, I assume that administrators, though primarily interested in program-level decisions, will also take an interest in classroom-level tests. Each group is inevitably involved in the other's decision making—perhaps in the form of teachers proctoring and scoring the placement test, or perhaps in the form of an administrator evaluating the effectiveness of teachers' classroom tests. The types of decisions discussed in this book may interact in innumerable ways, and I think that any cooperation between administrators and teachers in making decisions will be healthy for the curriculum in general and test development in particular.

Regardless of whether the reader is a teacher, an administrator, or both, the goal of reading this book should be to learn how to do all types of testing well. Inferior or mediocre testing is common, yet most language professionals recognize that such practices are irresponsible and eventually lead to inferior or mediocre decisions being made about their students' lives. The tools necessary to do high quality testing are provided in this book. Where statistics are involved, they are explained in a straightforward "recipe book" style so that readers can immediately understand and apply what they learn to their teaching or administrative situations. If this book makes a difference in the quality of decision making in even one language program, the time and effort that went into writing it will all have been worthwhile.

This is the second edition of this book. Brown (1996a) was the first edition, and Brown (translated by Wada 1999) provided a Japanese translation. This edition differs in several ways from the first edition. Most prominently, this edition has been updated throughout to reflect the present state of knowledge on all the topics covered, including many new sections and new references. But also of importance, based on the feedback and suggestions of professors using the first edition of the book, the conceptual and computational explanations of the various statistical techniques in the first edition have been expanded to include clear directions for doing the various statistics in a spreadsheet computer program. Judging by feedback from readers, the first edition of this book was found to be useful by many. I hope this new expanded edition will prove even more useful in real language teaching situations like yours.

I would like to thank Kathleen Bailey, John Nelson, and Betsy Parrish for their helpful comments during the reviewing process. Also, I would like to thank Mark Nelson and Sophia Wisener for their help in the editing process.

Finally, I would like to thank Microsoft for permission to use their *Excel*™ program.

Contents