高教社外语教师教育与发展丛书
体验英语教学系列

语言项目中的测试与评价

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

James Dean Brown

高等教育出版社
HIGHER EDUCATION PRESS
Mc Graw Hill

语言项目中的测试与评价

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

James Dean Brown

高等教育出版社
HIGHER EDUCATION PRESS

图字：01-2006-7303
James Dean Brown
Testing in Language Programs
ISBN: 0-07-294836-1

图书在版编目(CIP)数据

语言项目中的测试与评价 = Testing in Language Programs, a Comprehensive Guide to English Language Assessment/(美)布朗(Brown, J, D) —影印本.—北京：高等教育出版社，2006.12
(高教社外语教师教育与发展丛书.体验英语教学系列)
ISBN 7-04-020124-0

Ⅰ.语… Ⅱ.布… Ⅲ.语言—测试—英文 Ⅳ.H09

中国版本图书馆 CIP 数据核字(2006)第 146890 号

策划编辑 贾 巍 **责任编辑** 张歆秋 **封面设计** 刘晓翔
责任校对 张歆秋 **责任印制** 韩 刚

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100011
总 机 010-58581000
经 销 蓝色畅想图书发行有限公司
印 刷 北京七色印务有限公司

开 本 787×1092 1/16
印 张 21
字 数 480 000

购书热线 010-58581118
免费咨询 800-810-0598
网 址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
畅想教育 http://www.widedu.com
版 次 2006 年 12 月第 1 版
印 次 2006 年 12 月第 1 次印刷
定 价 32.00 元

物料号 20124-00

出 版 前 言

根据教育部大学英语教学改革的精神，《大学英语课程要求》提出要培养“学生的英语综合应用能力，特别是听说能力”，这其中包含了一些教学理念和教学模式的创新。要达到大学英语教学改革的预期效果，教师是十分关键的因素。大学英语教学改革的实践者是在教学第一线的广大英语教师，因此，针对目前我国大学英语教学师资短缺等问题，加强大学英语师资培训是摆在我们面前的一项刻不容缓的任务。为此，高等教育出版社策划引进出版了《高教社外语教师教育与发展丛书——体验英语教学系列》。

这是一套开放性的大型系列丛书，收入多位世界级语言教学专家的作品，具有权威性；内容涉及到外语教学方法、测试、评估等多个方面。丛书不仅系统介绍外语教学相关理论，更结合作者多年的教学经验，提供了大量实践案例，希望能够开拓我国外语教师教学及科研视野，培养教师在教学问题上独立思考、研究和创新的能力，成为我国外语教师教育与发展的助力器。

《高教社外语教师教育与发展丛书——体验英语教学系列》充分体现了体验式的教育理念，配合教育部大学英语教学改革推荐教材《大学体验英语》立体化系列教材及学习系统而出版，目的在于推荐新的教学理念，完成教学理念向教学实践的转化。

近期我社还将出版一系列为师范类学生、英语教师和英语研究者使用的英语语言教学丛书。我们由衷地希望这些教材的出版，对各高等院校的英语教学有所促进和帮助。

高等教育出版社

2006 年 11 月

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

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

导　读

梁茂成

语言测试及其研究简述

如同我们用尺子这种测量工具来测量物体的长度一样，语言测试是用来测量人们的语言能力的一种测量工具。但相比之下，测量语言能力比测量物体长度更为困难、更为复杂。

首先，我们对语言能力(即测量对象)需要一个准确的定义，搞清楚到底要测量什么。这方面的研究由来已久。语言能力是一种复杂而抽象的心理能力。早期的语言测试对语言能力的认识较为肤浅。随着人们对语言能力认识的加深，语言测试理论不断更新，对语言能力的认识也越来越科学、完整，并依此设计出形式多样的语言测试。对语言能力的完整认识是语言测试科学的理论方面，也是开发各种类型语言测试的理论基础。

其次，如同我们在测量长度时需要有一把顺手的尺子一样，工欲善其事必先利其器。语言测试中的又一重要研究领域是对测试本身的研究，其中主要包括对测试类型的研究和对测试中的题项的分析。语言测试的目的多样，有时为了选拔人才，有时为了对学生进行分班以便于因材施教，有时是为了检查学生对所学内容的掌握情况以便进一步改进教学。为了更好地达到测试的目的，人们不得不设计题目类型不同、内容各异、难易度不等的测试。与此同时，研究者还通过多种统计手段对试题的难易度、区分度、选项的分布等因素进行设定和调整，以确保测试工具具有一定的可靠性，能够诱导出参试者的真正语言水平。由于任何一种测试都无法满足所有语言测试的需要，对测试类型和测试中题项的分析成为语言测试研究领域的重要研究内容。

再次，对测试结果的描述和解释是语言测试学研究中的又一重要内容。如同我们在测量了物体的长度后需要读数并对结果进行报告一样，语言测试结果的报告是语言测试中不可缺少的一个环节，是语言测试的目的得以实现的重要保证。语言测试结果的报告及其解释主要以一些统计学知识为基础，其目的在于使考试组织者了解考试的总体情况，使应试者和语言教师了解自己或学生在测试中的成绩和存在问题，同时为人才录用、教学改革、分班教学等提供决策依据。

最后，测试的信度和效度是测试学中关注的另一重要问题。在测量长度时，如果我们用同一把尺子在短时间内测量同一物体所获得的结果应该基本相同；在语言测试中也一样，为了确保测得的结果可靠，在其他条件相同的情况下，同一批受试在两次测试中的结果应该基

本相同。这就牵涉到语言测试中的信度问题。此外，测试学中常常还需要关注所测得的数据是否真正反映了考生的语言能力的高低，如同我们关注用尺子所测得的数据是否真的反映了物体的长度一样。这与测试学中的效度有关。统计学方法是信度检验和效度检验的主要方法。

总而言之，语言测试研究主要涉及四个方面：

1) 语言测试要测量什么?即，语言能力包括哪些重要方面?

2) 语言测试中如何进行测量?即，什么样的测量工具才是最可靠的测量工具?

3) 如何报告和解释语言测试的结果?

4) 如何检验所测结果的准确性?

在以上四个问题中，第一个问题是一个理论问题，其余的三个问题主要是实践问题。由此可见，语言测试学是一门紧密结合理论与实践的综合学科。在理论层面上，它与应用语言学的其他领域联系紧密，而在实践层面上，它以统计学作为主要的技术基础。

《语言项目中的测试与评价》一书在兼顾测试学各方面的同时更注重实践，涉及以上四个问题但着重回答后三个问题，对语言测试全过程所涉及的众多方面进行了详细的描述和分析。

本书特点及作者简介

本书是语言测试方面的一本基础读物，由作者在自己多年研究生课程授课讲稿的基础上修改而成，是了解语言测试研究的一部综合指南。它是作者长期从事语言测试教学和研究的经验总结，适合于语言测试组织者、语言教师、研究生等，其清晰的脉络、合理的取舍和深入浅出的语言使其非常适合作高校语言测试学的教材。因本书具有较强的实用性，其第一版(Brown, 1996)已经被翻译成日语出版(Wada, 1999)。

本书的最大特点是通俗易懂、简略得当、实用性强。本书对测试的分类、测试命题方法甚至语言测试的历史等理论问题有浅显易懂的介绍和论述，同时突出实用性，对各种计算方法和统计分析的操作步骤进行细致的描述，甚至对如何利用电子表格处理软件分析和报告测试中的有关数据也逐步进行指导。此外，每章后还配有一些针对性很强的复习题和练习题，书后还附有一个十分有用的术语表以方便使用者查询。

本书的另一特点是一书多用。它兼顾了大规模语言测试和课堂语言测试的需要，同时也兼顾了考试组织和管理者、测试命题者、测试分析人员、语言教师、语言测试研究人员等与

语言测试相关的各种人员的需要，是快速了解语言测试学的极好著作，是语言测试组织者、语言教师、研究生等人员的案头必备。

本书作者James Dean Brown毕业于加利福尼亚大学洛杉矶分校(UCLA)，获得英语作为第二语言教学(TESL)硕士学位和应用语言学博士学位，曾在中山大学担任为期两年的高级学者，并在佛罗里达州立大学担任为期三年的助理教授，现为夏威夷大学马诺分校(University of Hawaii at Manoa)第二语言研究系(Department of Second Language Studies)的应用语言学教授，主要从事语言测试、课程设置、项目评价和研究方法等方面的教学和研究。

James Dean Brown是国际应用语言学界，特别是语言测试学界的著名学者，多年以来一直担任*TESOL Quarterly*和JALT期刊(*the JALT Journal*)的编委、TOEFL研究委员会(the TOEFL Research Committee)成员、TESOL研究顾问委员会(the TESOL Advisory Committee on Research)成员和TESOL执行委员会(the Executive Board of TESOL)成员。James Dean Brown曾在*TESOL Quarterly*, *TESOL Newsletter*, *Language Learning*, *Language Testing*, *Modern Language Journal*, *System*, *JALT Journal*, *The Language Teacher*和*RELC Journal*等多种学术期刊发表大量有影响的学术论文，并为多本学术专著撰写过章节。他编著出版过*New Ways of Classroom Assessment* (TESOL, 1998)一书。此外，James Dean Brown出版的专著主要有：*Understanding Research in Second Language Learning: a teacher's guide to statistics and research design* (Cambridge University Press, 1988)，*The Elements of Language Curriculum: A systematic approach to program development* (Heinle & Heinle, 1995), *Using Surveys in Language Programs* (Cambridge University Press, 2001)和*Testing in Language Programs* (Prentice-Hall, 1996)，并与Yamashita合作编写了*Language Testing in Japan* (JALT, 1995)、与Hudson合作编写了*Criterion-referenced Language Testing* (Cambridge University Press, 2002)、与Rodgers合作编写了*Doing Applied Linguistics Research* (Oxford University Press, 2002)、与Hudson等人合著了*Investigating Second Language Performance Assessments* (University of Hawaii Press, 2002)等。

本书内容简介

《语言项目中的测试与评价》全书共十一章，另附练习题答案、术语汇编、索引和参考文献等附录。

全书大致可以分为五个部分：第一部分(第1章)为基础理论和基本概念部分；第二部

分(第2—4章)讨论了开发、选用测试的宏观因素和微观因素，主要简述语言测试类型的选用、修改、开发和试题的设计，题项分析等；第三部分(第5—6章)讲述了如何报告测试结果，以及如何对测试结果进行解释，其中着重讲述描述统计方法及其含义，以便为选择基于常模的语言测试或基于标准的语言测试提供依据；第四部分(第7—10章)主要讲述了相关性分析、语言测试的信度检验和语言测试的效度检验，对相关性分析及其应用、信度和效度的检验方法等进行了详细的描述；第五部分(第11章)将语言测试与语言教学联系起来，主要阐述语言测试与课程规划、课程实施等之间的关系。

*** 第一部分(第1章)**

第一章：语言测试的类型及用途

本章主要介绍了语言测试的两种基本类型，即标准参照性测试和常模参照性测试。简言之，标准参照性测试常常目标明确，只用来测量考生对某些特定内容的掌握情况，这些内容常常与某一课程或某一阶段的所学内容有关。考生在考试前对考试中将涉及的内容比较清楚，而考试结束后，对考生成绩的解释不依赖于其他考生的成绩。换言之，考生所得的考分是一个绝对得分。如，某考生在某一门课的期末考试中只得了60分(满分为100分)，表明该考生对这一学期所学内容掌握得不好，可能只掌握了应该掌握内容的60%左右。相反，常模参照性测试用来测量整体语言能力(如英语水平、学术听力、阅读理解等)，测试中每一位考生的成绩只是一个相对成绩，该成绩只有与测试中所有其他考生的成绩相对比才有意义。考试之前，考生除了对测试的形式有所了解之外，对测试的内容常常一无所知。

为了便于测试组织者和语言教师选择合适的测试类型，作者从测试结果的解释、测量的对象、测试的目的、得分的分布、试题结构和试题所涉及的知识内容等六个主要方面分析了标准参照性测试和常模参照性测试之间的区别，并对这两类测试的用途(适用的决策目的)加以分析，同时指出，任何一种测试都无法满足所有的决策目的。

在本章的结尾，作者对电子表格处理软件Microsoft Excel作了简单介绍，并就本章内容准备了部分问题供读者复习时使用。

*** 第二部分(第2—4章)**

第二章：语言测试的选用、修改和开发

选用何种类型的测试需要考虑许多因素，其中既有理论方面的考虑，也有实践方面的考虑。这些考虑对测试组织者和语言教师十分重要。作者在第二章中对语言测试类型选择方面

需要考虑的一系列问题进行了逐个分析和讨论，所讨论的问题涉及理论和实践两个方面。

作者认为，由于各种语言教学方法侧重不同，在信念上存在分歧，语言测试开发者应该在充分讨论的基础上，结合本地或本国的教育政策，选择合适的语言测试类型。为了更好地说明这一点，作者对语言教学方法、特别是语言测试的历史进行了简短的回顾，提出在当今的应用语言学领域，多种教学理念和教学方法并存，在这种情况下，语言测试开发者应该把教学方法和语言测试结合起来。

本章中讨论的理论问题首先是语言能力与语言行为的区别。作者认为，语言能力是语言学习者的内在特征，而语言行为只能在有限程度上反映语言能力。由于语言测试中的成绩大多只能反映考生的语言行为的不同侧面，因此，对测试成绩的解释应该十分慎重。

本章中讨论的第二个理论问题是分项测试(如测试学生对英语冠词的掌握情况)和综合测试(如综合听写测试)的区别。作者认为，分项测试和综合测试各自有着自身的优缺点，且两者之间的区别有时十分模糊。

本章讨论所涉及的实践问题包括公平问题、开支问题、命题难易问题、施测难易问题、阅卷难易问题、理论问题的相互作用等。作者一再强调，在考虑选用、修改或开发某种测试类型时，对以上问题应该同时予以考虑。

在进行以上讨论之后，作者分别对选用、修改和开发测试时可能遇到的问题进行了分析，并提供了包含所有应该注意环节的核查表，供试题开发者、语言教师等参考。

最后，作者对Excel的基本操作进行了说明，这些基本操作包括几个基本概念、当前单元格的选择、创建工作表等。在结束本章内容之前，作者提供了一些复习题和实战练习题。

第三章：开发高质量的语言测试题项

本章主要讨论高质量语言测试所必备的基本部件。所有的语言测试中最基本的部件是题项，因此，作者在本章中首先对题项这一概念进行解释，进而对题项格式的基本准则进行分析。语言测试中的常见题项大致可以分为接受型应答题项、产出型应答题项和个人应答题项。接受型应答题项指只要求考生从所提供的选项中选择答案的题项(如多项选择题、正误判断题等)；产出型应答题项指要求考生使用所测试的语言，以口语或书面语形式进行产出性应答的题项(如填空题、简短回答题、作文等)；个人应答题项指要求考生根据个人兴趣以自己的方式应答的题项(如自我评估、面试、建立个人学习档案等)。

对题项格式分析应该遵循哪些准则，作者主要围绕四个方面讨论，它们分别是1)分析所

有类型的题项格式时应遵循的准则；2)分析接受型应答题项格式时应遵循的准则；3)分析产出型应答题项格式时应遵循的准则；4)分析个人应答题项格式时应遵循的准则。在对以上四个问题进行讨论的过程中，作者针对每一种题项格式分析中应该注意的环节，提供了一个检查清单，以便读者在需要时可以逐一核对，确保测试中的所有题项格式准确、规范、合理，能够充分检测考生对所测内容的掌握情况，且对所有考生公平、有效。

和其他各章一样，在本章的结尾，作者也提供了一些复习题和实战练习。

第四章：语言测试中的题项分析

本章主要介绍题项分析技术，全章分为两大部分。

第一部分讨论常模参照性测试中的题项分析技术，主要涉及难易度分析(item facility analysis)和区分度分析(item discrimination analysis)。难易度指某特定题项的答对率，即答对的人数与总人数之比，其取值范围为0—1；区分度指题项对考生的区分能力。对于一个区分度高的题项，考生水平越高，答对的可能性就越大，而考生水平越低，答对的可能性就越小。区分度通过计算高分组的难易度与低分组的难易度之差而求得，其取值范围在-1和1之间。

本章第二部分介绍标准参照性测试中的题项分析技术，主要涉及题项质量分析(item quality analysis)、区分指数(difference index)和B指数。在进行题项质量分析时，一方面应该考虑题项所涉及的内容，另一方面应该考虑题项的形式是否能够充分地测量期望测量的内容。区分指数的目的在于考察题项是否能够把掌握所测内容的考生与未掌握所测内容的考生有效地区分开来。区分指数通过计算后测与前测的难易度之差而得到，取值范围为-1到1之间。B指数指及格考生的难易度与不及格考生的难易度之差，所反映的是及格考生与不及格考生在某题项上的成绩差异。

在本章中，作者还就如何根据题项分析的结果选择合适的题项进行了必要的说明。在讨论每一种题项分析技术时，作者就如何在电子表格软件中进行计算操作进行了逐步讲解，并通过复习题和实战练习对所学内容加以巩固。

*** 第三部分(第5—6章)**

第五章：语言测试结果的描述

对语言测试的结果进行描述，其目的在于为考试组织者、命题者、语言教师乃至考生本人提供信息，告知他们考生们在考试中的实际表现，以便他们在以后的工作中作出相应的对

策。一般说来，测试结果的呈现方式有数据直接呈现和图形呈现两种，前者更为具体、准确，而后者更为直观、易懂。

在本章中，作者首先描述如何利用图形方式呈现测试中每一不同得分的出现频次。在作者看来，这种图形呈现方式十分方便测试组织者、考生和教师理解测试结果。

此后，作者讨论了测试中常用的三种量表(scales of measurement)，即称名量表(nominal scales)、顺序量表(ordinal scales)和连续量表(continuous scales)，其中的称名量表是最为初级的量表，只对考生进行简单的类别划分(如性别、母语背景等)；顺序量表对考生的某种能力按照特定的顺序进行排序(如按照成绩的高低进行排序)；连续量表是最为高级的量表，它不光对考生进行排序，并表明各考生之间的相对差距(如百分制得分)。这三种量表是统计学中的常见概念，对于理解测试学领域的许多统计问题十分重要。

本章讨论的第三个问题是描述统计学。与普通统计学教本所不同的是，作者在讲述描述统计学的基本知识时，把这些知识与语言测试紧密地结合了起来。在这一部分中，作者着重介绍了集中趋势和离散趋势两大问题，其中涉及到均数、众数、中数、中点、全距、标准差、方差等概念。在此基础上，作者详细地描述了在电子表格软件中如何进行描述统计学操作，图文结合，十分方便读者的理解。

本章中的另一的重点是测试结果的报告。作者不仅讨论了报表中应该包含哪些内容，还对成绩报表的呈现格式进行了必要的描述。最后的复习题和实战练习更有利于读者回顾和消化整章的内容。

第六章：语言测试得分的解释

语言测试的最终目的常常是便于决策者基于考生的分数作出相应的决策，因此对考生得分的解释尤为重要。在第六章中，作者主要讨论如何解释常模参照性测试和标准参照性测试中的考生成绩。作者结合第五章中学习的描述统计学原理，对考生的成绩围绕集中趋势和离散趋势两大方面，辅以各种图形进行多维度解释。

作者特别介绍了概率分布、正态分布和标准化分数三个主要概念。了解这些概念有助于语言教师把学生的成绩与参加考试的所有考生的成绩进行对比，从而进行更为准确的解释。

*** 第四部分(第 7 — 10 章)**

第七章：语言测试中的相关性

尽管描述统计学可以为我们选定、修改和开发测试提供重要依据，还可以帮助我们报告

测试结果，然而，仅靠描述统计学数据我们还无法断定测试是否可靠，我们还需要辅以其他统计手段。

本章主要介绍相关性分析。相关性分析是推断统计学(inferential statistics)的重要组成部分，可以帮助我们确定两组数据(如出勤率与考试成绩)之间是否存在关系。有时，我们认为两组数据间存在逻辑关系，但却并不知道这种关系的显著程度。相关性分析可以帮助我们确定我们认定的相关性是否存在并达到统计学要求的显著性，是否具有统计学意义。

作者首先介绍了相关性、相关系数等统计学概念，并通过例证和散点图等手段解释正相关、负相关、线形关系、确定系数、显著性、相关矩阵等概念。对于连续量表上的数据，我们进行相关性分析时一般采用皮尔逊积矩相关性(Pearson product-moment correlation)分析法。除此之外，作者还介绍了皮尔逊积矩相关系数的运算原理和计算方法，以及在电子表格软件中如何得以实现。与此同时，作者提醒我们，相关性分析存在一些潜在的问题，且皮尔逊积矩相关性的计算需要一定的前提，即 1)所分析的数据是连续数据；2)两组数据相互独立；3)数据呈正态分布；4)两组数据间存在线性关系。

鉴于皮尔逊积矩相关性分析只适用于连续量表上的得分，作者在本章中还向我们介绍了另一种相关性分析，并描述如何在电子表格软件中进行这种相关性分析。这种相关性分析是点二系列相关性(point-biserial correlation)分析，适用于分析一组称名变量与一组连续变量(如性别与考试得分)之间是否存在相关性。

如其他各章一样，本章结尾也附有一些复习题和实战练习。

第八章：语言测试中的信度

信度即一致性和可靠程度。语言测试中的信度包括两个方面。第一个方面是测试本身的信度。作为一种测量工具，测试与其他测量工具一样，应该具有一定的可靠性和稳定性，且能够测量它本该测量的东西。语言测试与其他测量工具一样，时常会产生一些误差。如果我们使用试卷复本(即等效测试题)对同一组学生重复进行测试，所得到的结果如果基本相同，我们则认为这一测试具有较高的信度。语言测试中信度的第二个方面是评分(主要是主观题的评分)的信度。同一个阅卷人在不同时间对评分标准的把握应该具有一致性和稳定性，而不同阅卷人对评分标准的把握也应该具有一致性。

在本章中，对于第一种信度，作者主要介绍常模参照性测试中的信度检测方法，并对信度系数的解释方法加以说明。作者首先分析语言测试中测量误差的可能来源，指出测试环

境、施测步骤、评分程序、受试考生、测试题和题项的设计等因素都可能导致测试中的测量误差。进而，作者对常模参照性语言测试中常用的三种信度检测方法，即再测信度检验法、复本信度检测法和内部一致性信度检测法进行了逐一介绍，并对其中的内部一致性信度的不同检测方法进行了更为细致的介绍，涉及分半信度(split-half reliability)分析法、Cronbach信度分析法、Kuder-Richardson公式等。作者还对内部一致性信度分析的各种方法进行了比较。

计算信度系数只是考察常模参照性测试的一致性的一种方法，我们也可以通过计算测量的标准误差(standard error of measurement)来确定测试得分的一致性。在本章中作者对测量的标准误差的含义和计算方法也进行了讨论。

对于第二种信度，作者分别讨论了评分员内部的信度和评分员间的信度，并介绍了两类信度的计算方法。

对于各种信度的计算方法，作者在本章中还细致地描述了如何利用电子表格软件进行操作，并通过复习题和实战练习的形式加以巩固。

第九章：语言测试中的可靠度

常模参照性测试中的考生成绩一般呈正态分布，且考生成绩的标准差相对较大(因考生水平差异较大)，因此我们可以利用相关性分析的方法对考生成绩进行信度检验。这一点第八章已经讨论过。然而，标准参照性测试则不同，测试的目的常常只是检测学生对指定内容的掌握情况，这样，如果对最近所学内容进行测试，考生成绩常常会呈偏态分布。比如，某教师在学期末对学生进行一次成绩测试，他所期望的可能是学生已经掌握本学期内所学的大部分内容，因此考试成绩会普遍偏高，标准差会较小，全距也会比常模参照性测试中的考生成绩的全距小得多。这种情况下，我们在分析测试的信度时，就不宜使用上一章所学的相关性分析的方法。

在本章中，作者主要介绍了适用于标准参照性测试的信度检验方法。为了与上一章中的常模参照性测试中的信度(reliability)相区别，作者分别使用一致性(agreement)和可靠度(dependability)等术语来专门指标准参照性测试中的信度。本章中涉及的可靠度检验方法主要包含三类，即临界缺损一致性(threshold loss agreement)分析法，平方误差缺损一致性(squared-error loss agreement)分析法和领域分数可靠度(domain score dependability)分析法。

首先，作者分别介绍了临界缺损一致性分析的两种常见计算方法。第一种是一致性系数(agreement coefficient)，即两次测试中成绩类别(按成绩高低将学生分为高水平和低水平)相

同的考生数与测试的总人数之比，取值范围在0.5和1之间。第二种是Kappa系数，其取值范围在0和1之间。由于这两种计算方法都要求对同一批学生进行两次测试，操作起来相对麻烦，作者又介绍了如何根据一次测试的成绩计算临界缺损一致性，并强调，临界缺损一致性的多种计算方法各有优势，实际操作中我们可以根据需要反复尝试多种方法。

与临界缺损一致性的计算方法不同的是，在平方误差缺损一致性的计算过程中，我们不是把考生按照成绩的高低简单地分为高水平者和低水平者，而是通过考察考生成绩与指定分数线之间差距的大小来计算一致性。

由于临界缺损一致性和平方误差缺损一致性系数都是以指定分数线为基点来计算的，有其固有的缺点。为了给读者提供更多的选择，作者还介绍了另外一种方法，即phi可靠度指数(phi dependability index，又称绝对误差的概化系数，即generalizability coefficient for absolute error)。该方法通过领域分数来计算测试的可靠度。

除了临界缺损一致性和平方误差缺损一致性两类可靠度检验之外，作者还向我们简要介绍了另外一类，即置信区间。这种可靠度计算方法的基本理念与常模参照性测试信度分析中的测量的标准误差(standard error of measurement)方法比较接近。

最后，就如何在电子表格软件中进行各种可靠度系数的计算，作者又逐一详细地加以说明，并配以复习题和实战练习加以巩固。

第十章：语言测试中的效度

评价测试的好坏，仅检验其信度是不够的，我们还需要对它的效度进行考察。为了说明这一问题，作者在第十章的引言部分给了这样的一个例子：众所周知，TOEFL具有较高的信度，然而如果我们用TOEFL考试对学生的数学能力进行测试，其结果是，信度很高，但我们其实并没有测试出学生的数学能力，只测试了学生的语言能力。可见，信度与效度是测试的截然不同的两个质量方面。信度是效度必要前提。任何测试若离开了信度就没有效度可言。

效度即有效性，指的是测试在何种程度上测出了它宣称或本该测量的东西。测试中的效度一般可以分为三种类型，即内容效度(content validity)、结构效度(construct validity)和标准关联效度(criterion-related validity)。对于标准参照性测试，我们只能分析其内容效度和结构效度，因为标准关联效度以相关性分析为基础，而相关性分析对标准参照性测试不适用。对于常模参照性测试，我们可以同时分析它的三类效度。

作者在本章中首先讨论了对常模参照性测试和标准参照性测试都适用的内容效度和结构效度。在考察内容效度时，人们关注的是测试是否能够代表它需要测试的内容。为了保证测试的内容效度，试题设计者在命题时可以将需要测试的内容分解，并根据题型，尽可能把这些内容逐一贯彻到题项之中。在命题最终确定之前，需广泛征求专家意见，并对试题进行预测，对其信度等指标进行统计。为了说明试题规划的全过程中所需要做的每一项工作，作者给出了一个很好的例子。

结构效度是效度的另一个重要方面。结构(construct)是一个心理学术语，用来指未经证实的概念，而结构效度实质上就是理论上的效度。为了使测试具有结构效度，作者向我们介绍了一种称作为分组区分研究(differential-groups studies)的方法。比如，我们并不知道什么是“语言能力”，但我们知道一些人具有较好的语言能力，而另一些人的语言能力则相对较差。根据分组区分研究的方法，我们选择两组其他条件相同但语言能力不同的人，对他们进行语言测试，如果语言能力较好的一组人测试成绩较高，而语言能力较差的一组人测试成绩较差，则我们可以得出结论，测试题能够区分语言能力的高低，因此具有结构效度。对此，作者在本章中也举例进行了说明。另外一种确立结构效度的方法称为介入性研究方法(intervention studies)。这种方法要求我们对同一组学生进行前测和后测，而两次测试之间进行介入性训练，使得学生在需要测试的方面的能力得到提高。如果两次测试成绩之间存在显著差异，我们则可以认为测试具有结构效度，能够有效地测量它应该测量的东西。

除了内容效度和结构效度之外，作者对仅适用于常模参照性测试的标准相关效度验证法也进行了讨论。标准相关效度检验方法要求学生测试得分与他们在其他被人们普遍接受的测试中的成绩高度相关。比方说，我们要验证一个英语水平测试的效度，可以让学生先参加该测试，然后再用TOEFL试卷测试他们，最后把他们在两场测试中的成绩进行对照，并分析两者间的相关性。如果两者高度相关，我们则有理由认为该测试具有效度。

除了以上内容之外，作者在本章中还讨论了标准的确定(standard setting)问题。测试常常被作为依据，用来决定学生是否可以就业、进入某课程的更高阶段、获得某课程的结业证书等，这些决定可能会给考生带来很大的影响，这就要求我们把握好标准，合理地确定分数线。这一确定分数线的过程称为标准的确定。标准的确定与测试的信度和效度关系密切。作者提出了一套确定标准时应该遵循的程序，借助测量的标准误差(standard error of measurement)和置信区间(confidence intervals)提高决策的合理性，值得决策者们借鉴。

在本章中，作者还讨论了反拨作用(包括如何减小负面的反拨效应、提高正面的反拨效

应等)以及测试中的不公正。与其他各章一样，本章的最后也附有复习题和实战练习。

*** 第五部分(第 11 章)**

第十一章：语言测试的实际应用

语言测试并非一个纯理论的学科，它与语言教学的众多环节之间存在着重要的联系。在本章中，作者引用 Brown(1995a)的课程设置中的主要环节的模型、教育技术中广泛应用的课程设置系统方法理论(systems approach)，并结合自己的亲身经历说明了语言测试与语言教学的其他环节之间的联系，详细描述了语言课程中如何结合测试进行各种决策。

参考文献：

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bailey, K. M. (1988). *Learning about language assessment: dilemmas, decisions and directions*. Heinle & Heinle.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

杨惠中，C.Weir. (1998). 大学英语四、六级考试效度研究 . 上海：上海外语教育出版社 .

刘润清，韩宝成 . (2000). 语言测试和它的方法，修订版 . 北京：外语教学与研究出版社 .

邹申 . 1999 英语语言测试——理论与操作 . 上海：上海外语教育出版社 .