当代国外语言学与应用语言学文库

Language Test Construction and Evaluation

语言测试的设计与评估

J. Charles Alderson
Caroline Clapham
Dianne Wall

外语教学与研究出版社
Foreign Language Teaching and Research Press
剑桥大学出版社
Cambridge University Press

当代国外语言学与应用语言学文库

Language Test Construction and Evaluation

语言测试的设计与评估

J. Charles Alderson, Caroline Clapham and Dianne Wall 著

杨惠中　导读

外语教学与研究出版社

剑桥大学出版社

京权图字：01－2000－0161

图书在版编目(CIP)数据

语言测试的设计与评估/(英)奥尔德森(Alderson, J. C.)等著;杨惠中导读.—北京：外语教学与研究出版社，2000.7
ISBN 7－5600－1923－4

Ⅰ.语… Ⅱ.①奥… ②杨… Ⅲ.外语教学—测试研究 Ⅳ.H09

中国版本图书馆 CIP 数据核字（2000）第 66545 号

出 版 人：李朋义
责任编辑：晏小平
出版发行：外语教学与研究出版社
社　　址：北京市西三环北路 19 号（100089）
网　　址：http://www.fltrp.com
印　　刷：北京京科印刷有限公司
开　　本：650×980　1/16
印　　张：21.25
版　　次：2000 年 8 月第 1 版　2006 年 10 月第 3 次印刷
书　　号：ISBN 7－5600－1923－4
定　　价：27.90 元

*　　*　　*

当代国外语言学与应用语言学文库

Preface by Halliday

Foreign Language Teaching & Research Press is to be congratulated on its initiative in making these publications in linguistics available to foreign language teachers and postgraduate students of linguistics in China.

The books are a representative selection of up-to-date writings on the most important branches of linguistic studies, by scholars who are recognized as leading authorities in their fields.

The availability of such a broad range of materials in linguistics will greatly help individual teachers and students to build up their own knowledge and understanding of the subject. At the same time, it will also contribute to the development of linguistics as a discipline in Chinese universities and colleges, helping to overcome the divisions into "English linguistics", "Chinese linguistics" and so on which hinder the progress of linguistics as a unified science.

The series is to be highly commended for what it offers to all those wanting to gain insight into the nature of language, whether from a theoretical point of view or in application to their professional activities as language teachers. It is being launched at a time when there are increasing opportunities in China for pursuing linguistic studies, and I am confident that it will succeed in meeting these new requirements.

M. A. K. Halliday
Emeritus Professor
University of Sydney

王宗炎序

近年来，国际交往日益频繁，国际贸易急速发展，出现了一种前所未有的现象：学外语、教外语、用外语的人多了；研究语言学和应用语言学的人多了；开设这方面专业的高校也多了，语言学硕士生和博士生也多了。就是不以此为专业，学习语言学和应用语言学的也不乏其人。为了给从事这个专业的师生提供便利，同时又帮助一般外语教师、涉外工作者以及汉语研究者开阔思路，扩大视野，提高效率，我们献上这套内容崭新而丰富的丛书——英文版《当代国外语言学与应用语言学文库》。

文库首批推出54部外国英文原著，它覆盖了语言学与应用语言学26个分支学科。这批书是我们与各地有关专家教授反复研究之后精选出来的。出版这样大规模的语言学与应用语言学丛书，这在我国语言学界和外语教学界是破天荒第一次。

我们这样做，抱着什么希望呢？总的说来，是遵循教育部关于加强一级学科教育的指示，在世纪之交，推出一套书来给中国的外语教育领航，同时也给一般外语工作者和汉语研究者提供信息，拓宽思路。

我们希望这个文库能成为进一步带动外语教学改革和科研的发动机；我们希望它能成为运载当代外国语言学理论、语言研究方法和语言教学方法来到中国的特快列车；我们希望，有了这套书，语言学与应用语言学专业师生就能顺利地进行工作；我们希望，通过读这套书，青年外语教师和外语、汉语研究者能迅速把能力提高，把队伍不断扩大。

以上是我们的愿望，可是从广大读者看来，这个文库是否真的有出台的必要呢？我们想，只要大家看一下今天的客观情况，就知道这套书有填空补缺的作用，是让大家更上一

层楼的扶梯。

我们跟许多人一样，认为国内的外语教学和语言学与应用语言学研究是成绩斐然的，但是某些不足之处也无庸讳言。

在语言研究方面，有大量工作还等着大家去做。汉语语法研究，过去由于结构主义的启示，已经成绩卓著，可是现在虽则引进了功能主义，还看不出什么出色的成果。语料语言学是新兴学科，在我国刚刚起步，机器翻译从50年代就有人搞，然而其进展至今不能令人满意。

在语言理论方面，我们不时听到一些片面的、所见不全的论调。有人说，1957年前西方根本没有什么理论语言学，其创始者是Chomsky；也有人说，语言纯属社会文化范畴；还有人说，搞语言研究只有量化方法才是科学方法，定性方法不值得一提。

谈到外语教学，某些看法做法是分明不值得赞许的。有人以为交际教学只管听说，不管读写，也有人以为教精读课就是教阅读，不管口语。在分析课文时老师满堂灌，学生开口不得，是常见的；教听力课时老师只管放录音，对学生不给半点提示点拨，也并非罕有现象。

上述这些缺点，我们早有所知，现在我们更加明白，必须力图改进，再也不能安于现状了。为了改进，我们就得参考国外的先进理论，借鉴国外的有效措施。眼前这个文库，就是我们上下求索的结果。

在编辑这个文库时，我们在两方面下了功夫。

一方面，在选书时，我们求全，求新，求有代表性和前瞻性。我们不偏爱一家之言，也不只收一家外国出版社之书。语言学与应用语言学的主干学科固然受到了应有的重视，分支学科可也不忽视。语料语言学、语言统计学是新兴学科，我们收入了专著；句法学、语义学久已有人研究，我们也找到了有关的最新著作。

另一方面，我们邀请了国内知名的博士生导师、硕士生

导师为各书撰文导读，为读者铺平道路。语言学和应用语言学专著包罗宏富，初学者读起来可能觉得茫无头绪。为了助他们一臂之力，本文库中每一种书我们都请专家写了一万字左右的导读材料。哪怕书中内容比较陌生，谁只要在读书前看一下导读材料，读书后把材料再看一遍，一定能弄清脉络，掌握要点。

在结束本文时，我们想向爱好泛读的人们提个建议。语言和社会生活息息相关；我们靠语言与他人协作；通过语言继承传统文化，接受外国先进思想和科学知识；利用语言来教育下一代，帮助他们创造美好的未来；语言又反过来表达着我们的个性和我们充当的各种角色。学一点语言学和应用语言学，有助于增强我们的语言意识，对我们的工作和生活都是有利的。我们不妨把此事作为一个项目，列入自己的日程。持之以恒，必有所获。

王宗炎

中山大学教授
博士生导师

导　读

语言测试作为一门独立的学科，有自己的研究领域和研究方法，这一点今天已经为人们普遍接受。语言测试是一门跨学科的综合性科学，从语言学、语言教学法和学习理论取得科学内容，从心理测量学获得科学手段。语言测试是伴随着语言教学出现的，没有语言教学也就无所谓语言测试。语言教学是第一性的，语言测试为语言教学服务。语言教学的任务是培养学习者实际运用所学语言的能力，而语言测试的目的则是提供一种科学的测试工具，通过对学生语言运用的抽样，对学生的语言能力进行客观的、准确的、公正的评价。语言测试的规模可大可小，小如用于一个班级、一个年级、一个学校的各种考试，大如各种全国性的甚至是跨国界的考试。考试的规模愈大，考试的社会性愈强。

大规模考试对考生、对用户往往有重大影响。因此，语言测试工作者必须明瞭所负的社会责任，应尽最大努力，保证考试的科学性、客观性、公正性。本书全面介绍语言测试从设计、开发到实际运行各个环节的原理原则，也介绍了对语言测试项目的质量的评价标准。本书主要针对大规模考试进行讨论，但所涉及的原理原则当然也适用于中、小规模的语言测试。

语言测试是一项专业性极强的工作。一个大规模考试项目，从开发到成熟，至少要经过三个环节：设计阶段、实施阶段和考后阶段。

设计阶段包括制定考试内容规范、公布考试大纲、规定考试内容和试卷构成及试题形式、以及确定记分体制。

设计阶段的首要问题是定义所测量的语言能力，这是语言测试的理论基础，涉及一项语言测试的结构效度，也是使分数具有可解释性的根本依据。

不同的语言观决定语言测试采用不同的方法、不同的内容和不同的题型。用作者的话来说，任何语言测试项目都是一定的语言观的具体操作。

七十年代以前，在语言测试界占主导地位的是分析法，这是语言测试的心理测量－结构主义时期。其语言学理论基础是结构主义语言

学。该理论认为，语言是由语音、词汇、语法构成的一个系统，这一系统是可以分解的，因此，可以设计出离散的题目（discrete items），以逐项测验学生是否掌握了这些分解的元素。在这一时期，语言测试中使用得最为广泛的题型就是多项选择题。从心理测量的角度来说，由于采用离散题，每题只考核一个语言点，因此，在有限的答题时间内，考生有可能做大量题目，从而增加了采样的覆盖面。采样量大，信度就高。另外，采用多项选择题评分具有客观性，可以提高语言测试的信度。在这一发展阶段，语言测试开发了一系列统计分析方法——包括对试题难易度、区分度、整卷信度等等的定量分析方法，使语言测试成为一门既有坚实的理论基础又有成熟的统计方法的学科。

但是，这种离散的语言测试存在其自身固有的缺点，因为把语言的构成要素进行分解并孤立地逐项地进行教学和测试，必然忽视这些构成元素在更大的交际语境中的相互作用。

自七十年代中期起，语言测试界开始重视所谓总体综合法的研究，这一时期称为语言测试的心理语言学和社会语言学时期（psycholinguistic－sociolinguistic era）。新的理论认为，语言不但是一个可以分解的体系，更是一种动态的、具有创造性的功能体系。因为语言使用过程中冗余度很大，不能说缺少了哪一个语言点语言交际就无法进行。特别是社会语言学家提出了语言交际能力（communicative competence）的概念，认为使用语言不但要能够按照语法规则构造出合格的句子，而且还必须具有在不同语境中合理地使用这些句子的能力。这就对语言测试的效度提出了新的要求，而这一点是孤立地测试语言点的单纯离散题做不到的。这一时期采用得比较多的是综合题，如完形填空（Cloze）、综合改错、听写、口试、作文等等。

自八十年代中期以来，随着交际教学法的发展，交际法语言测试（communicative testing）受到了愈来愈多的重视。从交际法的角度来看，所谓掌握一门语言是指在一定的语境中能够使用所学的语言进行有效的交际，交流思想感情，达到相互沟通的目的。在这里作者着重介绍了 Bachman 提出的交际语言能力（communicative language ability）模型，这一模型被国际语言测试界广泛接受。从语言作为交际工具的角度来看，进行有效的语言交际，仅仅掌握语言形式是不够的，因为语言交际过程涉及交际的目的、语境、彼此的角色地位等等；同样的语言形式，由不同的人在不同的场合以不同的方式讲出来，其含义可能完全不同。因此，语言交际过程实际上是一种解释过

程（interpretation），是交际双方的协同过程（negotiation）。既然如此，语言测试就必须在真实的（authentic）语境中采用真实材料来进行，观察学生在真实语境中运用语言达到交际目的的能力，并以此来判断学生的语言水平。

从上面的简单回顾可以看出，对语言测试来说，不同的语言观不但决定考什么，而且决定怎么考，而这两者直接关系到语言测试的信度和效度。这是开发一个语言测试项目在设计阶段首先要解决的问题。这一点确定以后，还要制定语言测试的内容规范（test specification），包括考试的目的、用途、性质、内容、方法、试卷构成、时间长度、考生对象等等。语言测试所涉及的语言变量、交际功能、情景、交际活动等等都不是凭空制定的，应当通过需求分析（needs analysis）来确定，只有这样才能保证一项考试的科学性。最后，考试内容规范还要规定记分体制，使分数获得可解释性。如果作为考试结果的分数任意性很大而且又不可解释，那么这样的考试就没有什么意义。大规模考试可以采用正态分制，也可以采用等级分制。正态分制通过把考生按成绩排序，用均值和标准差来表示考生在考生群体中的相对位置，这是一种间接可解释的记分体制；等级分制则采用等级描述语来描写考生能力，这是一种直接可解释的记分体制。前者常用于常模参照考试，后者常用于尺度参照考试。当然也可以把两者结合起来，称为尺度相关—常模参照考试。

考试内容规范中的大部分内容应当公开，这就是考试大纲。考试大纲应相对稳定。完成了考试内容规范，还要设计相应的试卷，在预定的考生群体中作抽样试测。经过分析，如果达到了预定的要求，一个语言测试项目的设计阶段就结束了。

在实施阶段，要有一系列的质量保证措施，才能确保一个语言测试项目的科学性。

在大规模考试中，为了保证考试的信度和效度，对考试有很高的质量要求。首先要保证评分的客观性和一致性。根据评分是否受阅卷员主观判断的影响，试题可分为客观性试题和主观性试题。客观性试题的常见形式有多项选择题、简答题、完形填空等。因为客观性试题的答案是唯一的或有限的，所以常用机器阅卷。大规模标准化考试通常采用较多的客观性试题。

在大规模标准化考试中，通常普遍使用多项选择题。其命题工作是一项专业性极强的工作，命题难度大、周期长。一项大规模考试不

但对所考核的内容要有明确的规定，而且对每一部分都要有详细而严格的命题要求。要建立专门的命题员队伍；命题员要经过严格培训。试卷中的每一道题目，在能够用于实际考试前，都要经过命题、审题、预测、计算机试题项目分析、复审、构卷等漫长而复杂的过程。只有确保试卷在难易度、区分度等方面都达到了规定的要求，才能实际施考。这个周期往往长达一年之久。只有这样，才能保证一项大规模考试的质量。

主观性试题的常见形式有命题作文、问题解答等，答案常常是开放性的。主观题的阅卷需要采用一系列质量控制措施，包括阅卷员培训、阅卷过程质量监控等，目的是保证阅卷员本人和阅卷员之间评分的一致性。

为了使不同考次的分数具有可比性，每次考试以后在成绩发布之前，大规模考试还要经过漫长的数据处理过程，包括数据录入、加权处理、等值处理、分数的正态化处理等等。

在考后阶段，作为科学的考试，每次考试结束还有大量工作要做。首先，每次考试后要发布成绩公报。如果是常模参照考试，还应当公布建立常模的依据，要提供用户手册。对教师来说，考后报告提供大量反馈信息，让教师了解教学中的长处和短处，以便改进教学；对用户来说，考后报告应向他们提供关于考试目的、分数解释、信度效度等各种信息，以便在使用分数时能够进行正确决策；对考试机构本身来说，还要有详尽的试卷分析，包括各种描述统计、试题项目分析、试卷各部分相关、客观题信度、主观题阅卷信度等等，并找出可能存在的问题和改进方法等。以上这些都是每次考试以后都必须进行的考后分析。

对于一项大规模考试来说，还有一项必须做的工作，这就是效度研究（validation）。如果一项考试信度很高，但效度不高，也就是说没有考到应该考的内容，那么这样的考试有什么用呢？现代语言测试研究的重点是提高测试的效度。效度是一个实证问题。它不是靠测试项目的设计者怎么声称而获得，而必须靠实验来验证。效度要通过外部的、独立的标准进行评价，例如把学生成绩与教师的评价进行比较、把一项考试与某一公认的大规模标准化考试作相关分析、通过长期观察考生语言能力的变化来推断考试的效度等等。一项考试的效度愈高，对分数的解释愈准确。

语言测试在完成设计开发并付诸实施以后，还有一个不断改革和

完善的问题。这不但因为测试项目本身要保持其信度和效度、要始终保持其测量精度、防止某些题型的老化，而且语言测试要跟上语言学和语言教学新思想的发展，要采用各种新技术新方法，以便不断提高测量的信度和效度。

大规模考试的社会性必然带来如何正确对待考试的问题，也就是对考试的结果是正确使用还是误用的问题。如果不能正确对待考试，就有可能导致应试教育。但应试教育不是考试的直接后果。作为对学生成绩的鉴定，科学的考试始终是需要的。考试的任务是对考生能力进行科学的、客观的、公正的鉴定。那么，考试本身如何鉴定呢？有没有对语言测试项目进行鉴定的评价标准呢？由于英语在国际交往中的重要性，目前各种英语测试项目多如牛毛，这就难免良莠不齐。就作者在最后一章中专门讨论了对考试本身的评价标准问题。这里所说的标准是指考试机构的行为准则，考试机构应该遵循的专业质量标准。这无疑是十分重要而且有着现实意义的。为了撰写本书，作者在1990年前后曾对英国一些主要的语言测试机构进行过问卷调查，结果发现实际情况与理想状态相去甚远，由此更说明了建立语言测试专业标准的重要性。作者对调查结果进行了详细分析，并在每章后面加以讨论。

本书对语言测试的各个环节，从设计、开发、到具体实施都进行了深入浅出的讨论，并且理论联系实际，值得所有从事语言测试工作的人阅读。

第一章　源起与概述

在这一章里，作者首先谈到了为什么要写这样一本书，并且对语言测试所涉及的各个方面作了概述。

有教学就要有考试，考试是对学生能力的鉴定。作者特别指出，“测试”与“考试”这两个术语的界定十分模糊，因此，在本书中两个术语混用，不作区别。本书是为那些负责语言测试的设计开发的人写的，也是为那些虽然并不直接介入语言测试的设计开发、但要使用语言测试的结果的人写的。

语言测试的设计开发是一个很长的过程。从最初起草考试大纲，直到最后作为考试结果的成绩报道，都是考试设计开发全过程中的一个个不可缺少的环节。本书力求用深入浅出的语言，全面讨论考试的

设计、开发、施考等原则，目的是确保语言测试的专业质量。全书共分十一章。除第一章概述外，分别讨论考试内容规范设计、命题与审题、预测与试题分析、考官培训、阅卷信度控制、记分体制与成绩报道、效度研究、考后分析报告、语言测试的改革与完善。最后一章讨论语言测试的专业规范与标准。

语言测试是一门交叉学科，跨越语言学、应用语言学和心理测量学等学科。一名称职的语言测试工作者，除了要有丰富的教学经验外，还应当善于从应用语言学与其他相关学科吸收新知识、运用新技术。本书描述的一些涉及语言测试开发设计的原则，不但适用于英语，也同样适用于其他语种。

语言测试是一门内容丰富的学科，并不象一般人想象的那样简单，似乎出了题目打个分数就是语言测试。作为科学的语言测试，最关键的是测试的信度和效度，没有信度就谈不上效度；而一个考试如果没有效度，那么这个考试就没有什么存在的必要。要保证一个考试的信度和效度，就要有一系列的质量保证措施。一个考试的规模越大，社会后果就越大，其信度和效度也就越重要。一个考试的信度和效度是需要有实验的数据来证明的；检验一项考试需要有专业规范和标准。为了深入浅出地解释语言测试的原理原则，本书每章首先提出若干问题，引起读者思考。每章结尾还附有一张列表，供从事语言测试的人备用检查。

本书还有一个特点。英语是目前使用最为广泛的国际交际语言，因此，英语作为外语的语言教学和语言测试应运而生。在英国，英语作为外语的测试项目数量众多。这些测试项目都达到了语言测试理想的质量标准了吗？为了撰写本书，作者在 1990 年左右对英国一些主要的语言测试机构进行了问卷调查。结果发现，实际情况与理想的状态相去甚远。作者对调查结果作了详细分析，在每章后面进行讨论，这对于从事语言测试工作的人是很有启发和参考价值的。

此外，各章及书后附有丰富的参考书目，书后还提供了常用的统计公式和应用软件目录，这些都是本书非常实用之处。

第二章　考试内容规范

设计和开发一个语言测试项目，首先要解决的问题是考什么和怎么考。考试内容规范（test specifications）就是关于某项考试考什么和

怎么考的正式文件，也是建立某项考试结构效度（construct validity）的基础。

考试内容规范应当涉及一项考试的各个方面，包括考试的目的、用途、性质、内容、方法、试卷构成、时间长度、考生对象、分数解释等等。毫无疑问，这样一个文件对涉及一项考试的每一个人都十分重要：对考试设计者，这是保证考试质量的“蓝图”；对命题者，这是命题的依据；对教师，可以知道如何利用考试提供的反馈信息提高教学质量；对考生，了解了考试的题型、难度、时间长度等，有利于发挥自己的真实水平；对用户，了解了考试的性质和分数意义，有利于正确进行决策，防止考试的误用。考试内容规范中的大部份内容加上样本试卷，应当公开，这就是考试大纲；一部分内容应当保密，如命题须知等。

制订考试内容规范时，最根本的问题是定义所测量的是什么语言能力（language proficiency），并且说明为什么这一设计中的考试能够测量这种语言能力。这是语言测试的理论基础，涉及到一项语言测试的结构效度。只有明确了这一点，才能明确试卷各部分所测能力的相互关系，也才能在阅卷员和考官之间统一思想，提高阅卷信度。说到底，任何语言测试都是一定的语言观的具体体现和操作（operationalization），这是某项语言测试的理论框架。作者在这里着重介绍了目前国际语言测试界流行最广的 Bachman 等人提出的交际语言能力（communicative language ability）模型，主要分为两个方面，即组织能力（organizational competence）和语用能力（pragmatic competence），前者包括语法能力（grammatical competence）和语篇能力（textual competence），后者包括言语能力（illocutionary competence）和社会语言学能力（sociolinguistic competence）。

对考试有影响的因素还有题型（客观题、主观题）、文体（是否学术英语；文章的类型；阅读目的）、试卷各部分的时间分配、考试指令是否清楚等等，这些统称为考试的方法因素，都会影响到考生的能力发挥，最后影响到考生成绩。

当然，语言测试中涉及的语言变量、交际功能、情景、交际活动（activities）等等，都不是凭空制定的，而应当通过需求分析（needs analysis）来确定，只有这样，才能保证一项考试的科学性，保证该项考试考了应当考的内容并具有高的效度。

第三章　命题与审题

语言测试是环绕着试卷进行的，因此，命题是保证考试质量最重要的一个环节。称职的命题人员不仅应具备丰富的教学经验，丰富的学科知识，了解学生学习中的困难所在，还要有丰富的创造力和想象力。试题和练习在形式上有颇多相似之处，但目的不同。试题是以一定的方式方法，让考生通过一定的语言活动来表现其语言能力，以此作为对考生语言能力评价的依据。因此，试题不但须要求明确、没有歧义，而且要有信度和效度。

命题人员须经过一定的培训，不但应熟悉某项考试的内容规范，而且掌握必要的命题技巧。许多语言测试是围绕一篇篇文章来命题的。因此，命题的第一步是选择合适的文章。所谓合适，不但是指题材、体裁、难易度、长度，而且要求文章有一定的信息量和足够的题眼，即能够根据考试内容规范提出各种类型的问题，以考核内容规范所规定的各种语言技能。因此，作者提出应先定文章后命题，甚至提出不妨在平常阅读时就注意收集合适的文章，建立文章库。这些都是经验之谈，很实用。

紧接下来，作者分别就客观题和主观题对各种常见的题型如何命题提出了许多实际的建议。这些来自实践的经验，对具体从事命题的人很有参考价值。

审题是命题以后必须进行的一个环节。比较好的一个办法是请非命题人员从考生的角度先做一遍，并回答每道题考的究竟是什么，再与命题人员当初的设计意图比较、与考试内容规范比较。这样一来，必能发现许多值得改进的地方，提高考试的效度。

第四章　预测与分析

命题结束以后，试题必须经过预测，才能知道题目的质量。对于某道试题考什么、难易度如何，即使请若干有经验的教师进行判断，其结论也可能相去甚远。可见不预测不可能知道试题的质量。而题目未经预测，直接应用于考试，就好比一把尺子本身没有经过检验和校正就拿来丈量，如果在大规模考试中这样做，其风险之大就可想而知了。

预测必须保证三点，即采样要有代表性、保密性、样本要有一定

的量。

预测以后要对题目进行项目分析（item analysis）。经典的项目分析主要是求试题的难易度和区分度。求出题目的难易度和区分度以后，还可以进一步求出整卷的平均难易度和试卷的内部相关系数。经典算法的题目难易度依赖于考生的水平。也就是说，对一批考生来说很难的题目，对另一批考生来说可能很容易。要排除考生因素就要采用试题响应理论（Item Response Theory），试题响应理论的数学基础是概率论。该理论认为，考生答题是一种概率现象：考生能力越强，答对某题的概率越接近于1；能力越低，答对某题的概率越接近于0。这可以用试题特性曲线（ICC）来表示。首先，把考生能力和试题难易度放在同一座标上，再通过锚题作等值处理，这样就可以排除考生因素的影响。试题响应理论有单参数模型（只考虑题目难易度）、两参数模型（考虑题目难易度和区分度）和三参数模型（考虑题目难易度、区分度、和猜测因素）。采用试题响应理论为数学模型建立的题库，如果题目都经过标定，就有可能用于建立机助自适应考试系统。

最后，作者还深入浅出地介绍了描写统计学的常用基本概念。

第五章　考官培训

考官是指考试过程中负责对考生语言能力作出判断的人。这里不谈采用机器阅卷的客观题，就主观题部分而言，所谓考官，至少包括作文阅卷员和口语考试中负责评分的考官。

考试是一种教育测量。所谓测量，首先要有一种标准化的测量工具；其次，使用这种测量工具的人应能始终保持标准一致。对语言测试来说，要做到这一点，考官必须对所测量的语言能力有一致的看法，对评分标准有一致的理解，并且对评分标准的掌握始终保持一致。这里就涉及到考官培训。

所谓主观题就是需要考官在评分时作出判断的题型。考官培训首先必须要有明确的、可操作的评分标准，每个分数段要有明确的描述语（descriptor）；其次要选定标准样卷，标准样卷要能够正确反映各个分数段，而且要覆盖分数全距；此外，也可以选出一定数量的“问题卷”，即较难判分的卷子。最后用标准样卷和“问题卷”对考官进行培训，目的是统一标准，统一思想。

这些措施对于保证考试的信度非常重要。如果考官随便打分，用