

汤姆森英语教师丛书

# LEARNING ABOUT LANGUAGE ASSESSMENT

Dilemmas,  
Decisions and Directions

语言评价研究  
困境、抉择与方向

Kathleen M. Bailey (美) 著

外语教学与研究出版社

FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

汤姆森英语教师丛书

# 语言评价研究：困境、抉择与方向

**Learning About Language Assessment: Dilemmas,  
Decisions and Directions**

**Kathleen M. Bailey (美) 著**

任 真 导读

**外语教学与研究出版社**

**FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS**

北京 BEIJING

## (京)新登字 155 号

京权图字: 01 - 2002 - 1456

### 图书在版编目(CIP)数据

语言评价研究:困境、抉择与方向/(美)贝利(Bailey, K. M.)著. —北京:外语教学与研究出版社, 2003.9

ISBN 7 - 5600 - 3689 - 9

I. 语… II. 贝… III. 英语—语言教学—教学法—英文 IV. H319

中国版本图书馆 CIP 数据核字(2003)第 089757 号

Kathleen M. Bailey

Learning About Language Assessment: Dilemmas, Decisions and Directions

Original Book ISBN: 0-8384-6688-5

Copyright © 1998 by Heinle & Heinle, a division of Thomson Learning

Original language published by Thomson Learning (a division of Thomson Learning Asia Pte Ltd). All Rights reserved.  
本书原版由汤姆森学习出版集团出版。版权所有,盗印必究。

Foreign Language Teaching and Research Press is authorized by Thomson Learning to publish and distribute exclusively this edition. This edition is authorized for sale in the People's Republic of China only (excluding Hong Kong, Macao SARs and Taiwan). Unauthorized export of this edition is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

本书由汤姆森学习出版集团授权外语教学与研究出版社独家出版发行。此版本仅限在中华人民共和国境内(不包括中国香港、澳门特别行政区及中国台湾)销售。未经授权的本书出口将被视为违反版权法的行为。未经出版者预先书面许可,不得以任何方式复制或发行本书的任何部分。

## 语言评价研究:困境、抉择与方向

Bailey, K. M. (美) 著

\* \* \*

项目管理: 孙平华

责任编辑: 关静瑞

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: 北京外国语大学印刷厂

开 本: 787×1092 1/16

印 张: 18

版 次: 2004 年 3 月第 1 版 2004 年 3 月第 1 次印刷

书 号: ISBN 7 - 5600 - 3689 - 9/H·1855

定 价: 25.90 元

\* \* \*

如有印刷、装订质量问题出版社负责调换

制售盗版必究 举报查实奖励

版权保护办公室举报电话: (010)68917519

# 汤姆森英语教师丛书

## 专家委员会

主 任 张连仲

副 主 任 王 蓓 李 力 程晓堂

委 员 (按姓氏笔划为序)

|     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|
| 马寅初 | 王 蓓 | 田贵森 | 任 真 | 刘兆义 | 孙平华 | 许时昇 |
| 何向明 | 何安平 | 张连仲 | 邱 峰 | 范东生 | 夏谷鸣 | 徐 庆 |
| 程晓堂 | 鲁子问 | 潘炳信 | 穆林华 |     |     |     |

本书导读 任 真

项目管理 孙平华

责任编辑 关静瑞

## 出版说明

当前,我国外语教育领域正在进行着全面而深入的改革。改革的内容涉及到外语教育的方针政策、基本理念、基本原则、教育文化、教育实施、教育评价等各个方面,也关系到新世纪亿万学生的成长和可持续发展。随着课改新理念、语言新方针和教育新策略的确定及课改的逐步推开,教师培训工作的重要性和关键作用变得越来越为突出。从某种意义上说,课改能否最终获得成功,取决于广大教师对课改的认识、取决于他们的语言理论水平和实际工作能力。

对英语教师来说,新一轮的课改提出了许多新的理念和原则:强调学习者的个性发展的“以人为本”的教育思想;强调语言学习与人的可持续发展的关系;强调语言目标的新内涵:知识、技能、策略、文化和情感等;强调语境教学与综合全面评价理论的引进。所有这些具有丰富内涵的理念,都需要广大教师学习和了解。

同时,国家英语课程改革对英语语言教学理论和实践提出了新的、更高的要求,研究学生如何学、教师如何教、考试如何考和教师如何在教学实践中提高自我素质已成为新世纪英语语言教学研究的主要焦点。针对国家英语课程改革的实际,外研社精选并引进国际著名出版机构——汤姆森学习出版集团设计的多部论著和专业教程(均为当代语言教育领域的力作)。由当今国际语言教育界的权威人士 David Nunan, Larsen-Freeman, Neil Anderson, Andrew Cohen, Tom Scovel 等编著并由国内英语教育界专家学者撰写导读。这些专家学者在认真阅读原著的基础上,结合中国外语教师的需求和现状,有针对性地用深入浅出的语言对原书作了概括、总结与评价,有助于英语教师和语言工作者、研究者对原著的学习和理解。本丛书 19 个选题都是目前英语教师关注的深层和热点专题。

在这 19 本书中,有几本是关于外语教学理论的,如 *Learning New Languages*, *Teaching ESL K-12*, *Understanding Language Teaching*, *Teaching English as a Second or Foreign Language*;有几本是关于语言教学技能和实践的,如 *Exploring Second Language Reading*, *Teaching and Learning Vocabulary*, *Teaching Second Language Writing*;有关于学法指导的: *How to Be a More Successful Language Learner*;有关于课程设计的: *Designing Language Courses*;有两本是关于语言测试与评估的: *Learning About Language Assessment*, *Assessing Language Ability in the Classroom*;有的是关于外语教学中科研方法的: *Doing Teacher Research*;有的是关于师资发展的: *Pursuing Professional Development*;有三本书是关于语境和文化教学的: *Teaching Culture*, *Teaching Language in Context*, *Teacher's Handbook*;还有三本是关于教学法的: *Working with Teaching Methods*, *Teaching*

*Languages, Methodology in TESOL.*

本丛书作为英语教师专业化教育/发展和英语专业硕士研究生教育的教材,从英语语言教学的理论和实践两个层面,为英语教师的教学实践提供理论指导,并为英语教师的在职教育和终生职业发展提供了丰富的资源,供在职大、中学英语教师和英语语言教学研究者、尤其是英语专业硕士研究生学习使用。

外语教学与研究出版社

# 语

## 言评价研究：困境、抉择与方向

---

### 导读

本书的写作目的是帮助关注教育测试与评价理论的教师了解这一领域的基本理念，并帮助他们分析课堂教学评价的案例。

近几年，大量的有关语言测试理论书籍的先后出版，说明越来越多的教师和教育管理者认识到测评在教学中的重要作用，并且关心这一领域的研究发展与动向。但是，大部分中小学教师由于语言及专业知识有限，对于理论性很强的专著存在着理解上的困难，本书会使他们消除畏惧心理。

本书与实际紧密结合，语言简朴，非常适合中小学教师和一切关注语言教学发展和教学测评的人士使用。也可用作教师培训的教材，每章作为一个独立的话题供大家讨论。

# 1

## 全书结构与总体评价



### 全书结构

全书共分 14 章，每章包含三部分：**教师之声** (Teachers' Voices)、**理论框架** (Frameworks) 和 **研究与反思** (Investigations)。“教师之声”以讲故事的形式把教师在实践中遇到的困惑告诉给读者，作者记录下与教师的对话，以这种方法引发读者思考。在“理论框架”部分中，每一章围绕一个主题介绍不同的评价手法及相关领域的研究，从理论上指导读者解决问题。在“研究与反思”部分，作者为读者设定了不同的任务及参考案例，力图使读者将相关知识应用到实践中去，结合自己的课堂教学与学习，尝试一些新想法。这部分不仅包括了教师可试用的任务 (Tasks)，还提供了相关书目，感兴趣者可以进一步研究。



### 总体评价

此书适合于什么样的读者？作者在一开始采用教学中常用的 **Twenty Questions** 的方式解答了老师可能提出的问题和疑惑。例如：很多老师会问他们可能不亲自参与测试编写工作，还有必要读这本书吗？另外有的老师认为 **Test** 仅仅是四个字母组成的一个单词，对学生更为重要。难道真是这样吗？其实不然。第一，你也许不是试卷的编写人员或教研员，但只要你参与

语言教学，就应学会编写自己的课堂练习、小测验，随时掌握学生对教学的反馈。特别是当前我国正逐步加强形成性评价的力度，教学评价不应仅仅局限在一两次大、小考试上，而应有系统地、随时地对学生的课堂表现与进步进行测评。第二，你也许还没成为老师或者刚从事教学事业，只要有心想成为一名合格的教师，就应该对语言评价有所了解。无论是进修、写论文、听报告，我们都会时常遇到有关评价方面的术语。评价是教学进程中不可缺少的环节。如果缺乏教学经验，在阅读理论性较强的专业书籍时，就会有似懂非懂的感觉，而阅读本书如同听一个老教师讲述自己的经历，使你受到启发，引发联想，从而深刻理解理论的内涵。第三，你也许没有上过研究生课程或尚未完成大学本科教育，担心对测试理论知之甚少，但只要你懂英语，就会看懂书中的专有名词。它们都用黑体标出，并在第一次出现时有例句或简单解释，书后的 Glossary 中也有注解。这不仅可以扩大专业词汇量，还可避免与同事谈论专业话题时由于不懂专业术语所产生的尴尬。第四，你也许认为在教学或辅导中，你经常使用已编好的卷子或标准化试卷，很少有机会甚至尚未有能力自己编写试卷，但只要充分认识到测试评价的重要性，就会有强烈的欲望想知道：什么样的试题是有效的？好的试题标准是什么？你还可以发现现有的试卷中所存在的问题与缺憾。试卷并不是检测学生学习表现与能力的惟一方法。教师的每一次提问，每一项作业，每一段评语，都是评价的一部分。第五，你也许会遇到家长、学生、亲朋好友向你咨询有关考试方面的问题。只有你懂得这方面的专业知识，才会给他们正确的指导。例如：2001 年高考时，北京有很多考生认为只要总分考过 600 分，就能上北大、清华，但当得知并非如此时，有些老师和家长认为分数“毛了”。这主要是对属于常模参照性考试的高考性质不了解，有关常模参照性考试将在后面的章节中详述。

# 2

## 各章节主要内容与导读

### 第一章 作为信息收集过程的评价

本书在开篇中讲了一个众所周知的故事。有一个农夫运送一只鹅、一筐谷物和一条狗过河，由于渡船载重有限，每次只能运送一人一物。如果运了粮食，狗就会吃掉鹅；如果先运狗，鹅就会吃掉粮食。农夫必须想出解决的办法。

作者通过这一故事引出教师在教学中因需要做出不同选择时的困惑。她的朋友在数年前还是一个初出茅庐的教师，当时被分到学校教书，年级组长让她使用第二册教材，并提出只有保证足够的生源才能开班。这位老师的困惑是，她无从得到任何有关这个班学生的成绩、教学进度甚至教学大纲的文字记载。通过这一案例，作者指出，在开展教学前有必要对学生进行语言能力和知识层面的测评，因为测评是收集信息的过程，它可以让我们因材施教。一个好的评价计划应该有效（valid）、可靠（reliable）、可行（practical），并对教学起到良好的反拨作用（washback）（在 Hughes 书中为 backwash）。测试的有效程度称为效度（validity）。如果语言评价能够提供给我们所需要的信息，这个测试就具有有效性。在制定评价计划时，首先要明确我们为什么要进行评价。如果我们想了解学生的口头语言表达能力但实际却测试了学生的语法知识，评价结果则无效。评价还要保持一致性，即评价结果的可信度（reliability）。如果教师在口试中由于劳累在给后面学生评分时要求过严或过松，就会造成评价结果的不一致，这不仅影响学生的成绩，也会

影响老师对学生能力的准确判断。评价需要人力、物力的消耗，例如人员、时间、费用等，所以评价的手段一定要可行，如果英语老师想采用面试（interview）方式来检验一个班 50 个学生的口语能力，显然会耗费很多时间，不太现实。另一个值得关注的问题是，评价对语言教学是否起到良好的反拨作用。如果我们的教学是以提高学生交际能力为目标，而测评内容是让学生花时间记忆一些不常用单词的拼写与定义，其结果对教学不会起到积极的促进作用，相反还会打击学生的积极性，产生挫败感。因此在测评过程中不仅要明确测评目的，还要明确要考的语言技能和层面，做到有的放矢。几年后，作者再次访谈当年的朋友，证实了适当的评价手段可以帮助教师了解学生状况，使教师对学生充满信心。作者以农夫的故事开始，是想告诉读者在各自的教学实践中，也许会遇到与书中相似或截然不同的经历，我们必须学会面对不同问题，找出解决问题的方法。因此在每章的“研究与反思”部分作者都设计了学习任务，希望读者能够运用理论知识，结合实例，具体探讨解决新问题的方法。这一部分对师资培训具有现实意义。可通过与同事们共同讨论，逐步提高独立承担教学任务的能力。该书在提供参考书目时不仅列出书名、作者和出版社，还对相关文章做了简单的介绍。

## 第二章 测试分析：听写

作者在本章分析了测评中常用的标准听写（standard dictation）题型。

做标准听写练习时，老师朗读听写内容，要求学生听懂其意并将原文全部写出。在使用听写中要考虑以下四方面：（1）选材。教师选听力段落时应明确考试目的，如果所选材料中含有过多的生词、人名或地名，不仅不利于老师朗读，也阻碍学生理解。（2）任务及完成情况。如果教师布置的任务设定为培养学生自我纠错的能力，就应在默写后给一分钟时间让学生自检。如果教师想考查学生是否能在听懂材料的基础上重新组句完成大意，就可以在朗读时减少停顿，拉长间隔。但这一切必须充分考虑到学生的能力，使他们充分发挥自己的水平。（3）评分标准。评分标准基于检测目的。小学教师在给学生听写单词时可能会重视拼写，因为要检测小学生识别单词和字母的能力。除此之外，现在越来越多的语言测试已不太看重拼写错误。在不影响句意、不引起误解的前提下，有些拼写错误是可以容忍的。例如：学生把过去时 learned 拼成 learnd 或 lernd，这表明学生抓住了语音，听懂了词义，但写错了字形。有些拼写错误是不能饶恕的，例如：在剑桥少儿英语考试中如果考生把 bed 误拼成 bad，虽然表面上可能是因为发音时嘴型偏大造成发音有

语言评价研究：困境、抉择与方向 · F9

误, 但 bad 在英语里与 bed 的词义截然不同, 因此不能接受。

根据考生的水平和考试目的, 听写也可采取另外两种形式。第一种称为部分听写 (partial dictation)。教师将给出听写材料的部分内容, 学生在听写过程中只需填写空缺部分, 这在一定程度上比标准听写的难度低, 以此可以突出考试重点。另一种称为渐进听写 (graduated dictation)。教师在朗读听写内容时每一停顿后字数逐渐增加, 句子逐渐加长。开始可能二三字一停, 逐渐到五六字一停, 十字一停, 直到学生感到稍许吃力才停止。这种方法可以引导学生从短时记忆向深层理解过渡。

### 第三章 评价目的的矛盾冲突

本章讲述了常模参照性 (norm-referenced) 考试与标准参照性 (criterion-referenced) 考试的区别。

这两种考试侧重于对分数的不同解释。常模参照性考试是将考生成绩与其他同考的学生成绩相比, 假如一个考生在百分制试卷中得了 95 分, 这个成绩是否算高分要看其他考生的成绩。如果一个学校录取前 10 名考生, 而 96~100 分之间的考生已有 10 名, 那么得 95 分的考生就可能落选。目前, 我国的高考就属这一类考试。有经验的教师知道没有固定的分数能确定学生是否能考入某所大学, 关键要看招生名额比例和报考人数的比例。当报考人数多, 需要从中筛选出有限名额时, 人们通常采用常模参照性考试。标准参照性考试是将考生成绩与事先设定好的标准进行对比, 不受其他考生成绩的影响。这如同考机动车驾驶证执照, 考试通过的标准是经过车管部门依据专业技术知识和要求严格制定的, 任何一个人只要达到这一标准就可以合法驾驶, 否则均不通过。它不存在通过率之说, 很多语言等级考试就是这样制定的。学生通过某种等级考试, 就意味着他们具备相应等级所描述的能力。一般来讲, 如果参加考试人数众多 (如 TOEFL), 考分又需分层拉开, 这时我们就需要采用常模参照性考试。作者认为常模参照性考试不适用于平时课堂测验, 因为我们考试的目的不是想把学生分成三六九等。

作者还提到潜能考试 (aptitude test)。这种考试不是针对学生某种语言能力的测试, 而是测评学生是否具备学习语言的潜在能力。

语言教学自始至终贯穿不同的考试, 因此, 明确考试目的、正确编写或选用不同类型的考试、准确解释考试分数以充分利用考试结果十分重要。通常一个学生要先通过入学考试 (admissions test) 获取资格, 然后参加分班考试 (placement test), 进入适合的班级。同一水平班的学生也可能在语言技能

的发展上有差异,教师可以用诊断考试(diagnostic test)来进一步判断学生的强项和弱项。如果学生总分很高,但听力很弱,教师将就此调整教学重点,对症下药。在教学过程中,教师需要随时检测学生对教材的接受程度,因此,小测验、单元测验便组成了阶段测试(progress test),用于检测学生是否不断取得进步。期中、期末或毕业考试属于成就考试(achievement test),用于考核学生是否在一定阶段内取得学习成果。诊断考试、阶段考试和成就考试都与教学大纲紧密相连。

## 第四章 背景知识在接受型技能评价中的作用

本章着重谈了文化背景知识对语言习得及测试所起的作用。作者首先讲了一个发生在波兰的笑话。两个陌生人站在一列正在行驶的火车的窗前,闲谈之中说起买汽车。甲问乙:如果世界上所有的汽车都供你挑选,你想买什么牌子的车?乙犹豫了一会儿说:也许是 Warszawa(波兰国产普通车)吧。甲说:你怎么才买这个车呀,如果你有很多钱,并且可以随心所欲地选呢?乙答:如果是这样的话,我要买 Wasrzawa。甲又强调说:如果你有大量的美钞、英镑呢?乙回答说:那我绝对买 Warszawa 了。最后甲说:那你肯定对汽车了解得太少了。乙说:不是我对汽车了解太少,而是我对你了解太少了。

如果读者对波兰上个世纪 80 年代政治与社会生活不了解,则很难理解这段笑话的意思,可见背景知识与文化的匮乏会阻碍语言交流。作者借这个笑话引出图示理论,即你所读到、听到、看到的语言文字、图形、符号和声音是否能激发起大脑某一区域的记忆,形成联系,从而帮助你去推断和理解信息。由于第二语言学习者用第一语言的文化背景去理解某些语言文字时常会遇到障碍,因此,在语言测试中设计者要充分考虑这一点。有些试题单从语言水平上讲并非很难,但如果考生缺乏对第二语言文化背景的了解,或受本身社会环境的局限,这些考题就答不出来。错误的联想也会引发错误的理解。举个简单例子:cereal(麦片粥)在英语里是常用词,而在中国对许多没有出过国的人来说就不知为何物,因为中国人很少在早餐时食用它。

## 第五章 完型填空题家族

本章与上一章紧密关联。如果我们在夜晚看到一家娱乐城屋顶上霓虹灯

的字缺胳膊少腿，我们仍能猜出字意，这是因为我们对母语非常精通，又有情景提示。作者通过这个例子再次说明，我们对语言的识别不仅依赖文字符号，还要依靠上下文、语言结构、背景知识等多方面信息的综合。完型填空系列的题型充分体现了人们具有整合利用残缺信息的能力。这些题型包括填空题（fill-in item）、完型填空（cloze test）和C-测试（C-test）。传统的填空题是一句一空，学生根据句意猜测所缺单词。虽然单词填空可覆盖多项考点，但存在有悖语言真实性的局限。在实际交流中，语言往往不是割裂使用，而是发生在情景之中，因此编写试题时应尽量提供完整语境供学生判断。与单词填空相比，完型填空则包含完整的语境，已被广泛用于语言测评中，以检验学生阅读理解和语言整体水平。C-测试是完型填空的变体，它与完型填空不同之处在于其所需填写的单词的第一个字母或头几个字母已给出，考生根据首字母线索和上下文意思填出适当单词。这在一定程度上降低了完型填空的难度。在本章中作者还介绍了编写完型填空时应注意的问题。

## 第六章 又一困境：不同的观念

在语言测试中我们应该注重语言知识还是语言技能？什么样的方法和题型更有效？作者在本章回答了这些问题。

她介绍了直接与间接考试（direct and indirect tests）、分项与综合考试（discrete-point and integrative testing）以及客观与主观评分的区别。直接考试是指口语、写作类考试，它直接运用与考试形式相关的语言技能。如果我们用写作文的形式来考查学生的写作水平，就是直接考试。但如果我们以多项选择方式让学生识别文章中的错误的方式来体现他们的写作能力就是间接考试。托福中，四选一查错的题型就属间接考试。分项考试是指每一题只考查一个语言点，例如只考辨音能力或语法单项，这曾经被认为是最公平、最客观的题型，因为它只能有一个答案。但事实上真正做到一题只考一点有很多局限。综合性测试侧重于考查学生综合运用语言的能力，涉及语言多层次、多技能，即语音、词形、词义、句法、语境及听、说、读、写等诸项。前文提到的听写和完型填空即属这类题型。客观评分试卷通常是选择题，不受个人主观意识影响，可使用机器阅卷。主观评分的试题（如口试或作文）在评分时往往受到评判人的主观影响。如果使用主观评分方法，不但要严格细致地制定评分标准，而且要严格培训评分者，以达到评分标准的一致性。

## 第七章 有用的统计学工具

## 第八章 相关分析家族

在这两章里，作者集中讲述了测试数据分析和试题分析。通过数据分析，我们可以了解测评的可靠性和有效性。数据分析可以告诉我们如何改进试卷和测评方法。作者并未深究统计学或测量学，只是讲解了一些基本的、常用的数据运算的过程，适用于小范围测评统计，便于教师应用。

常见的用来描述数据集中趋势（central tendency）的统计量是平均数（mean）。其计算方法是，把所有考生分数相加，然后除以考生总人数，公式为  $\bar{X} = \frac{\sum x}{n}$ ，其中  $\bar{X}$  表示平均数。X 上面没有“-”表示个体分数，n 为考生人数， $\sum x$  表示每个考生的分数相加。集中趋势反映数据向某一点集中的情况。但这还不够，我们还应考虑分数的分散情况。描述数据离散趋势的统计量是离散量数（measure of dispersion），它反映了一组数值离中的趋势。常用的统计指数有全距（range）和标准差（standard deviation）。全距指考生最高分与最低分之间的差值。如果将几个教学班的成绩相比较，他们的平均数有可能相近，但全距可能不一样。标准差可以更精确地反映一组数据的离散情况，也称均方差。要想了解分数的差异，要看它们之间的离中趋势。我们把一组分数里的每个分数与这组的平均值相比较，看它们平均离开平均值有多远。标准差值越大就说明这组分数离开平均值的平均距离越大，平均值越不具代表性。

如果考生考了两份不同试卷，我们可以计算出两份试卷之间的相关系数（correlation coefficient）。在研究相关系数时，一般使用散点图（scatter plot）。如果两份试卷之间不存在相关之处，散点分布杂乱无章，作者用不同的图表和例子进行说明。

而所谓相关分析，即“返回在某一集合上对两个系列求得的相关性”。后一章关于有关数据的种类、相关分析的基本概念、皮尔逊系数，重叠方差、秩相关系数（又称斯皮尔曼系数，服从由自由度和显著性水平决定的卡方分布），点二系列相关系数等的介绍，可以进一步帮助我们调查各种测试之间的相关性、一致性和评分的公正性。

## 第九章 多项选择和题项分析

除了对试卷和分数分析以外，我们还应该对题项进行分析。

本章即对多选题 (multiple choice) 进行了题项分析。虽然多选题具备诸多优点，如速度快、经济实用、可机读、评分客观、可靠性强、涵盖考点广等等，但在此作者并非想要赞许它的优点，而是提醒读者应充分认识对它的批评。作者指出：真正好的多选题很难编写，耗费人力，因此十分有必要对题项进行分析 (item analysis)。本章从难度、区分度、干扰项等方面分析多选题。难易度 (item facility) 是指一道题对考生的难易程度指数，通常用小数 0.0~1.0 表示。例如：有 40 人考试，20 人做对，该题难度指数为 0.5。难度指数为 0 表示没有一个人做对，难度指数为 1.0 则表示全部做对。区分度 (item discrimination) 比难易度更能提供进一步分析，它反映高分考生与低分考生在某一题项上的差别。难度值并没有区分好学生与差学生，特别是在常模考试中，这一差别更为重要。如果所有的高分学生都答错了某题，而低分学生却做对了，这就该引起我们思考：是题出得太难太偏，还是答案模棱两可？这就是为什么要计算区分值。

## 第十章 测评意义：听写作文和拼句组文

本章主要讲通过运用听写作文 (dictocomp) 和拼句组文 (strip story) 两种技巧测评语篇理解和文章结构的连贯性。

听写作文是由传统听写发展而来，但两者在评分上有区别。听写作文要求教师事先明确学生所写作文应包含的基本要点，这在我国高考作文中已有所体现。只要学生的作文涵盖了所给的要点并达意，就已完成基本测试任务，应给予及格成绩。

有些教师反映，一些考试在设计上没有突出对语用技能的考查，偏重知识性，缺乏真实性，这不适合当今强调语言交流的趋势。对此，作者描述了加拿大安大略教育学院 (OISE) 的测试项目以及他们在设计语言交流测试方面的原则。

拼句组文就要求学生既注重语境关联，又注重语言知识结构。在学校里，学生需完成许多写作任务，拼句组文被认为是有效的课堂活动。做练习时，每个学生必须记住自己的句子，然后通过集体的积极参与最后将故事或

段落拼凑出来。这样既训练了学生的口语和听力，又练习了写作，使学生在掌握了单词和句型的同时加深对句子之间关系的理解，对全面提高语言技能起到积极的作用。

## 第十一章 在情景对话中评价口试

本章重点介绍了口试中常用的情景对话技巧 (role play)。

出题者应使考题准确测出考生水平，虽然翻译 (translation)、问答 (direct question)、讲故事 (storytelling) 都能有效激发学生的表达能力，但在考高层次学生时，情景对话更能检验学生社交语言能力。

在编写情景对话的剧情 (scenario) 时，应考虑对话的情景是否符合学生的经历，是否可被学生理解。一旦超越他们的想像和阅历，结果就会不真实。

另外在口试中，如果对话是在考生和考官两者之间进行，考官的话语需有所节制，以便给考生更充分的机会展示自己的语言表达能力。因此在设计考试情景时双方对话的长短应平等。在保证对话自然流畅的前提下，考官要让考生主导谈话内容。本章结尾还谈到内部与外部评分一致性问题。

## 第十二章 写作评分的三种方法

作者在本章谈论了如何评判写作并探讨了主观评判过程中的问题。

传统的写作评分有三种方法：整体评分 (holistic scoring)、分解评分 (analytic scoring) 和客观评分 (objective scoring)。

首先作者阐述了何为评分标准 (scoring criteria)。评分标准不仅要明确规定正确答案，还应确认考生应完成的任务，并能鉴别完成任务的程度。整体评分是从整体上看学生的写作篇章，给一个总体分数评价。在评分过程中，考官不打分项分。通常采用 1 至 4、1 至 6 或 1 至 10 计分制。分解评分是按语言的不同组成部分——内容、文章结构、词汇、语用、标点等五方面分别打分，并根据考试要求决定各部分不同的权重。这种方法可使考生很容易知道自己文章到底哪方面不符合标准。但比起整体评分，分解评分较费时。不受任何主观干扰的评分方法应为客观评分。作者在本章中列举了几种不同的客观评分方法，很值得教师们参考，同时也提出问题供教师们思考。例如：客观评分只记错题是否恰当？如何评判不同程度的语法错误？如何区