

# 中 外 语 言 测 试

## 基 本 理 论 与 实 践

修旭东 著

青岛海洋大学出版社

· 青 岛 ·

## 图书在版编目(CIP)数据

中外语言测试基本理论与实践/修旭东著. —青岛:  
青岛海洋大学出版社, 2002. 9

ISBN 7-81067-388-2

I. 中... II. 修... III. 语言能力—测试—研究  
IV. H09

中国版本图书馆 CIP 数据核字(2002)第 066206 号

青岛海洋大学出版社出版发行

(青岛市鱼山路 5 号 邮政编码:266003)

出版人:李学伦

日照日报社印刷厂印刷

新华书店经销

\*

开本:850mm×1168mm 1/32 印张:9.0 字数:220千字

2002年10月第1版 2002年10月第1次印刷

印数:1~3000册 定价:17.60元

## FOREWORD

This book on testing is valuable because there is a pressing need in all sections of the education community for a greater awareness of the test development process. This book aims to raise the level of professional competence in developing reliable and valid tests. We need to develop tests that actually test what we want to test and which provide us with results that we can depend on, i. e. , our tests must be both valid and reliable. The volume covers an impressive range of the areas we need to understand in language testing and will help readers develop their knowledge and skills in the area.

It has been said the language testing is the cutting edge of Applied Linguistics. It is through language testing that we will gain insight into the nature of language proficiency and also establish more clearly the development of language acquisition. Testing is not just an academic pursuit, however as its importance in the learning cycle shows.

Teachers need information on student progress during the life of a course as well as summative assessments

of ability at the end. Well constructed tests can provide this information. Teachers need to evaluate formatively to make appropriate decisions concerning modifications to teaching procedures and learning activities; to steer their way through the syllabus in action, i. e. , to shape and influence the process. Is new material being introduced too quickly or too slowly? If it is, the effectiveness of learning may well be diminished. Teachers need to decide when to move on in a unit or to the next unit. If the next units are dependant on what has gone before then the teacher needs to be sure the students have mastered the objectives of a particular unit before proceeding. This can lead to necessary modifications in a programme.

There is a need to monitor the performance ability of students; a need to establish as clearly as possible what it is they can do. This can lead to a descriptive profile of a learner's communicative performance or a record of achievement. Formal tests devised for monitoring achievement can be extremely helpful to individual students, can help identify areas of strength or weakness and provide focus for future learning objectives. They can also be motivational by providing an idea of progress.

Tests have a valuable formative and summative role in education. It is the aim of this book to help practitioners in China to play their part.

**C. J. Weir**

**Aug. 2002**

# 前 言

学好、用好语言是人类文明与进步的必要条件之一。从文化人类学的角度来看,由于种族和民族的不同,语言可以有母语、第二语言和外语之分。但是,从语言学习与应用的角度来看,无论是母语还是母语之外的其他语言,都存在着一个怎样才算是学好和用好语言的问题。那么,怎样才算学好、用好了语言呢?从事物的相对性和绝对性来看,“学好、用好”只能是相对的,学无止境才是绝对的。尽管如此,要看是否真正地“学好、用好”,总需要一定的标准来衡量。需要什么样的标准、怎样来测量,这些都需要相关的理论和依据作指导。改革开放以后,以英语语言测试研究为代表的理论与实践在我国产生了广泛而深远的影响,这对我国的英语教学与评估是一件非常有意义的事情,而在语言共性的基础上,它对汉语和其他语种的教学与评估,无疑也会产生有益的影响。事实上,在我国,不仅仅语言测试,甚至其他科目的测试乃至对教育质量的评估指标,均需进行深入研究,否则,素质教育将难以全面而科学地进行。实践中,尽管强调不要把分数当成“指挥棒”,但很少有人能够真正做到。在这种情况下,测试的质量就显得尤为重要。多数人会认为,得90分的考生比得85分的考生优秀,但大多数人所采用的这种判断

“标准”在多大程度上具有科学性呢？这就很难回答了。我们或许可以回答，到目前为止某某考试是比较科学或合乎标准的，但永远不可能回答某某考试已经是最科学、最合乎标准的。测试的质量如同商品的质量一样，“没有最好，只有更好”。如果“最好”了，为什么高考后会出现那么多“高分低能”的人呢？Hughes(1989)认为：事实上，没有最好的测试或测试技术。既然“没有最好、只有更好”，对语言测试同样需要永不间断地进行探索与研究。

本书所论述的“中外语言测试”是指国内外对以英语为代表的语言所进行的测试，作者在该书中主要对国内外以英语为实例的语言测试理论和实践进行阐述和研究。

从语言学角度看，语言测试属于应用语言学研究的范畴。同时，从测试学角度看，它要符合测试的基本要求。“测试”的定义有许多种，目前公认较为权威的是 Anastasi(1982)的定义(刘润清、韩宝成，2000)：测试是对行为样本所做的客观的标准化的测量。该定义包含三个基本因素：(1) 行为样本：在语言测试中，它是指对语言能力表现行为的有效的抽样。(2) 客观的测量：在语言测试中，它主要指测试项目的难易度和区分度，测试结果的可靠性和有效性。(3) 标准化的测量：在语言测试中，它主要指测试题目的编制、测试的实施、分数的解释等一整套系统程序所具备的严密性，从而使测试结果具有可比性。

Spolsky(1979)把语言测试的发展分为三个阶段：前  
此.2.为试读,需要完整PDF请访问: [www.ertong.com](http://www.ertong.com)

科学阶段(Pre-scientific trend)、心理测量与结构主义阶段(Psychometric-structuralist trend)和心理学与社会语言学阶段。

根据测试的定义和语言测试发展阶段的划分,本书主要介绍和研究了语言测试第三个阶段的理论与实践。在这一阶段中,我们可以看到:Bachman(1990)关于语言测试理论(信度、效度、项目分析等)的阐述成为语言测试界的里程碑;Weir(1993)关于四项基本技能的测试框架为语言测试的设计提供了蓝图;桂诗春教授对分数统计与解析的阐释开了我国英语标准化测试的先河。本书的特点是把国内外语言测试的主要理论与实践进行了比较全面、系统和由浅入深的介绍和探讨。

本书的主要内容包括:语言测试的主要类型、原则(信度、效度、项目分析等)、试卷的总体设计、四项基本技能的测试框架、分数的分析与解释等。对语言测试的前沿进行探讨的章节主要有:信度的综合归纳理论、项目反映理论、四项基本技能测试的真实性和交互性原则。

本书的适用对象包括英语本科高年级学生、语言学研究生、外语教学工作者等,并对从事汉语及其他语种教学的工作者也有一定的参考价值。

本书参考了大量国内外原著,在成书过程中,英国的Don Porter教授、Cyril J. Weir教授给予了热情的指导与帮助,Weir教授在百忙中为本书作序,曾与本人同在英国里丁大学语言科学系作过访问学者的朋友以及在国内

工作的亲友们对本书也提出过许多宝贵意见,烟台师范学院外语系的有关同事也给予了许多帮助,在此一并致谢!

由于笔者水平所限,书中错误、缺点难免,敬请指正。

修旭东

2002年6月

# 目 录

第 1 章 语言测试的意义	(1)
第 2 章 语言测试的类型	(4)
2.1 从用途的角度区分	(4)
2.1.1 水平测试	(4)
2.1.2 成绩测试	(5)
2.1.3 诊断测试	(6)
2.1.4 学能测试	(6)
2.1.5 分级测试	(6)
2.2 从分数参照框架的角度区分	(7)
2.2.1 常模参考性测试	(8)
2.2.2 尺度参考性测试	(9)
2.3 从评分方式的角度区分	(10)
2.3.1 主观性试题测试	(10)
2.3.2 客观性试题测试	(10)
2.4 从结构的角度区分	(11)
2.4.1 直接测试	(11)
2.4.2 间接测试	(11)
2.5 从综合与分离的角度区分	(12)
2.5.1 综合性测试	(12)
2.5.2 分离性测试	(12)
2.6 从规模的角度区分	(13)
2.6.1 大规模测试	(13)
2.6.2 课堂测试	(13)

第 3 章 信度 .....	(14)
3.1 CTS-理论 .....	(17)
3.1.1 真分数与误差分数 .....	(17)
3.1.2 平行试卷相关性及其误差方差计算 .....	(18)
3.1.3 内部一致性信度 .....	(19)
3.1.4 再测信度 .....	(25)
3.1.5 对等信度 .....	(25)
3.2 G-理论 .....	(26)
3.2.1 综合归纳范围和测量范围 .....	(27)
3.2.2 考生人数 .....	(29)
3.2.3 范围平均分数 .....	(29)
3.2.4 综合归纳信度系数 .....	(30)
3.2.5 CTS-理论和 G-理论中的标准误计算 .....	(33)
3.3 IR-理论 .....	(36)
3.3.1 单维假设 .....	(36)
3.3.2 项目特征曲线 .....	(36)
3.3.3 能力分数 .....	(38)
3.3.4 项目信息功能 .....	(38)
3.3.5 测试信息功能 .....	(39)
3.4 尺度参考性测试分数的信度 .....	(40)
3.5 影响信度的因素 .....	(41)
第 4 章 效度 .....	(43)
4.1 信度与效度 .....	(44)
4.2 效度的证据基础 .....	(46)
4.3 内容效度 .....	(46)
4.4 效标关联效度 .....	(48)
4.4.1 共时效度 .....	(48)
4.4.2 预测效度 .....	(50)

4.5	实验效度	(51)
4.6	效度的计算	(51)
4.6.1	效度相关系数的计算	(51)
4.6.2	线性回归	(53)
4.6.3	因素分析	(54)
4.7	影响效度的因素	(57)
<b>第5章</b>	<b>项目分析</b>	<b>(58)</b>
5.1	难度	(58)
5.1.1	客观性试题项目难度指数计算	(58)
5.1.2	主观性试题项目难度指数计算	(59)
5.1.3	项目难度的表示方法	(59)
5.1.4	项目难度的选择	(61)
5.2	区分度	(63)
5.3	难度与区分度的关系	(68)
5.4	灵敏度	(69)
<b>第6章</b>	<b>试卷的设计</b>	<b>(70)</b>
6.1	测试类型的决定	(71)
6.2	测试内容的决定	(72)
6.3	测试题型的决定	(75)
6.4	如何克服主观性试题中的主观因素	(77)
6.4.1	主观性试题的误差来源	(77)
6.4.2	改进办法	(78)
6.5	试题编写	(80)
6.5.1	多项选择题	(80)
6.5.2	完形填空	(84)
6.5.3	听写	(89)
6.6	试卷中项目数量和各部分内容比重的决定	(91)
6.6.1	选择题数量的决定	(91)

6.6.2	各部分内容所占的比重和项目数量的分配	(92)
6.6.3	测试细目表的编制	(93)
6.7	Alderson 关于试卷设计的理论说明	(94)
6.7.1	试卷设计的总体考虑	(94)
6.7.2	测试效度的说明	(98)
6.7.3	核对试卷设计说明的清单	(101)
<b>第7章</b>	<b>口语测试</b>	<b>(103)</b>
7.1	口语测试的目标	(103)
7.2	口语测试的实施条件(Conditions)	(105)
7.3	口语测试的形式	(112)
7.3.1	间接测试形式	(112)
7.3.2	考生间的交互形式	(118)
7.3.3	考生与考官的交互性	(120)
7.4	评分标准	(128)
<b>第8章</b>	<b>听力测试</b>	<b>(136)</b>
8.1	听力理解的模式	(136)
8.1.1	听说交互模式	(137)
8.1.2	辨认选择模式	(138)
8.1.3	Richards 模式	(139)
8.2	听力测试的目标	(141)
8.3	听力测试的实施条件	(146)
8.4	听力测试的方法	(150)
8.4.1	利用图画或图形测试听力	(150)
8.4.2	利用多项选择测试听力	(151)
8.4.3	利用听写形式测试听力	(152)
8.4.4	利用简答形式测试听力	(154)
<b>第9章</b>	<b>阅读理解测试</b>	<b>(159)</b>

9.1	阅读理解测试的目标 .....	(160)
9.2	阅读理解测试的实施条件 .....	(162)
9.3	阅读理解测试的方法 .....	(165)
9.3.1	选择填空 .....	(165)
9.3.2	C-测试 .....	(166)
9.3.3	简答题 .....	(167)
9.3.4	多项选择 .....	(183)
9.3.5	正误判断 .....	(184)
9.3.6	句子排列 .....	(184)
9.4	阅读理解测试中任务设计的步骤 .....	(185)
<b>第 10 章</b>	<b>写作测试 .....</b>	<b>(187)</b>
10.1	写作测试的目标 .....	(187)
10.2	写作测试的实施条件 .....	(190)
10.3	写作测试的方法 .....	(192)
10.4	写作测试的评分 .....	(203)
10.4.1	整体评分法 .....	(204)
10.4.2	分解评分法 .....	(209)
10.4.3	整体评分与分解评分相结合 .....	(212)
<b>第 11 章</b>	<b>测试成绩的分析与解释 .....</b>	<b>(214)</b>
11.1	分数集中趋势度量指标 .....	(214)
11.1.1	算术平均数 .....	(214)
11.1.2	中数 .....	(216)
11.1.3	众数 .....	(218)
11.1.4	平均数、中数、众数之间的关系 .....	(219)
11.1.5	几何平均数 .....	(219)
11.2	分数离散趋势度量指标 .....	(220)
11.2.1	方差与标准差 .....	(221)
11.2.2	相对标准差 .....	(222)

11.2.3	全距	(223)
11.2.4	四分位差	(223)
11.3	分数的分布	(224)
11.3.1	正态分布	(224)
11.3.2	偏态分布	(225)
11.3.3	偏态值与峰值	(226)
11.4	百分位表的编制	(227)
11.5	标准分	(230)
11.6	主观性试题分数的调整	(232)
附录		(234)
I	Transcript of lecture on issues in the Women's Liberation Movement	(234)
II	阅读的微技能(J. Munby, 1978)	(239)
III	正态曲线的面积和纵线表	(244)
参考文献		(256)

# Contents

<b>1</b>	<b>The Significance of Language Testing</b>	(1)
<b>2</b>	<b>Language Testing Types</b>	(4)
2.1	Types Divided on Purpose	(4)
2.1.1	Proficiency Test	(4)
2.1.2	Achievement Test	(5)
2.1.3	Diagnostic Test	(6)
2.1.4	Aptitude Test	(6)
2.1.5	Placement Test	(6)
2.2	Types Divided on Frame of Reference	(7)
2.2.1	Norm-referenced Test	(8)
2.2.2	Criterion-referenced Test	(9)
2.3	Types Divided on Scoring Procedure	(10)
2.3.1	Subjective Test	(10)
2.3.2	Objective Test	(10)
2.4	Types Divided on Construction	(11)
2.4.1	Direct Test	(11)
2.4.2	Indirect Test	(11)
2.5	Types Divided on Integrativeness and Discreteness	(12)
2.5.1	Integrative or Global Test	(12)
2.5.2	Discrete or Discrete-point Test	(12)
2.6	Types Divided on Dimension	(13)
2.6.1	Large-scale Test	(13)

2. 6. 2	Classroom Test .....	(13)
<b>3</b>	<b>Reliability</b> .....	(14)
3. 1	CTS-Theory (Classical True Score Measurement Theory) .....	(17)
3. 1. 1	True Score and Error Score .....	(17)
3. 1. 2	Parallel Test .....	(18)
3. 1. 3	Internal Consistence .....	(19)
3. 1. 4	Test-retest Reliability .....	(25)
3. 1. 5	Equivalence/Parallel Forms Reliability .....	(25)
3. 2	G-Theory (Generalizability Theory) .....	(26)
3. 2. 1	Universes of Generalization and Universes of Measures .....	(27)
3. 2. 2	Populations of Persons .....	(29)
3. 2. 3	Universe Score .....	(29)
3. 2. 4	Generalizability Coefficients .....	(30)
3. 2. 5	Standard Error of Measurement in CTS-theory and G-Theory .....	(33)
3. 3	IR-Theory (Item-Response Theory) .....	(36)
3. 3. 1	The Unidimensionality Assumption .....	(36)
3. 3. 2	Item Characteristic Curves .....	(36)
3. 3. 3	Ability Score .....	(38)
3. 3. 4	Item Information Function .....	(38)
3. 3. 5	Test Information Function .....	(39)
3. 4	Reliability of Criterion-referenced Test Score .....	(40)
3. 5	Factors that Affect Reliability Estimates .....	(41)
<b>4</b>	<b>Validity/Validation</b> .....	(43)
4. 1	Reliability and Validity .....	(44)