语言学及应用语言学
研究生阅读大系

汉语测试与评估

方绪军◎著

復旦大學出版社

语言学及应用语言学
研究生阅读大系

汉语测试
与评估

方绪军◎著

复旦大学出版社

图书在版编目(CIP)数据

汉语测试与评估/方绪军著. —上海:复旦大学出版社,2013.9
(语言学及应用语言学研究生阅读大系)
ISBN 978-7-309-09764-1

Ⅰ. 汉… Ⅱ. 方… Ⅲ. ①汉语-测试-研究②汉语-评估-研究 Ⅳ. H195

中国版本图书馆 CIP 数据核字(2013)第 122685 号

汉语测试与评估
方绪军 著
责任编辑/韩结根 李婉茹

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@ fudanpress. com http://www. fudanpress. com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
江苏省句容市排印厂

开本 787×960 1/16 印张 15.75 字数 222 千
2013 年 9 月第 1 版第 1 次印刷
印数 1—4 100

ISBN 978-7-309-09764-1/H·2111
定价:28.00 元

序

随着我国国力的迅速增强，学习汉语的外国人愈来愈多，近年来对外汉语教学得到迅速发展，已经成为我国语言教学领域的一个重要分支。有教学就有测试，我国也已经开发了一些对外汉语教学方面的语言测试项目。

在教育领域，测试和考试很难区分，在学术讨论中一般也不严格区别。其实中国是考试的故乡，科举考试已经有一千四百多年的悠久历史，因此考试并不是一个新的概念。这样长久的历史传统，产生了两个重要影响：一个是考试具有权威性，“分数面前人人平等”，人们毫不怀疑地普遍接受；另一个是把考试看做行政行为，而不是学术行为，只要行政部门一纸红头文件就可以开始或者取消一个大规模考试项目。但是，考试是有社会权重的，尤其是大规模考试，考试结果往往决定学生一生的命运。对于这样重要的活动，人们自然要关心考试本身的科学性，只有保证测量的结果是准确的、公正的、精确的，“分数面前人人平等”才有意义，才能保证教育公平与公正。事实上教育考试（包括语言测试）属于心理测量与教育测量范畴，

是一个重要的学术领域。方绪军的《汉语测试与评估》结合汉语教学的特点对语言测试进行深入的分析讨论，对于提高我国语言测试的实践和理论研究水平有重要意义。

语言测试是一门内容丰富的独立学科，有着自己的研究领域和研究方法。语言测试是一门跨语言学、语言教学、心理测量学和教育测量学等领域的综合性学科，从语言学、语言教学法和学习论取得科学内容，从心理测量学和教育测量学获得科学手段。

语言测试，尤其是大规模标准化考试远比想象的复杂，且不说语言能力如何定义各有各的说法，即使有了一致公认的看法，心理测量又不同于物理测量，怎样把看不见的心理量准确地、公正地、稳定地测量出来，绝不是一件简单的事情，这些都对语言测试专业队伍的素质提出了很高的要求。

设计开发一个语言测试项目，要解决的是考什么和怎么考的问题。首先要定义所测量的语言能力，然后通过需求分析来确定语言变量、交际功能、交际情景、交际活动等。不同的语言观决定不同的测试方法、测试内容与题型。设计开发语言测试项目是一个漫长的过程，从最初起草考试大纲一直到最后作为考试结果的成绩或分数报告，涉及考试内容规范设计、命题与审题的质量控制、预测与试题分析、考官培训、阅卷信度控制、记分体制与成绩报告、效度研究、考后分析等，这些方面都达到了心理测量与教育测量的规范与标准，才是一项科学的考试。设计、开发并实施一项科学的考试是专业性极强的学术性工作。方绪军的《汉语测试与评估》对语言测试的各个环节进行了深入浅出的讨论。

语言测试与语言教学两者之间，语言教学是第一性的，语言测试为语言教学服务，把科学的语言测试的结果反馈给教学，使教和学都更具针对性，这是语言测试应有的功能。作者在书名中特别提到评估，对诊断测试进行了深入讨论，意在引导人们正确处理语言测试与语言教学的关系。

语言测试具有社会性，正像 Bachman 教授所说，“语言测试不是在试管中进行的。”语言测试的任务是对考生的语言能力水平进行测量，但测试结果往往作为用人部门在升学、就业、晋升等方面的决策依据，这种决策必然对考生、对社会产生重大影响，这就是考试的社会性。考试的社会性问题已经成为 21 世纪以来国际上语言测试学界的研究重点。考试（包括语言测试）其实只是一种测量工具，一个测试项目的社会权重越大，这项测试就成为高风险测试，受到测试各相关方（包括考生、教师、考生家长、学校、教育管理部门、采用测试结果的部门等）的关注程度就越高。因此，一方面，从事语言测试工作的专业人员必须尽一切努力提高语言测试的质量，另一方面，又必须普及语言测试知识，让语言测试各方面的相关者都能正确对待考试、正确使用考试结果。

语言测试学科的发展任重道远，作为心理测量学的一种实践形式的教育考试，包括语言测试，是从西方学的。毋庸讳言，我们需要继续学习和借鉴国外好的理论和方法。我们认为教育考试涉及一个国家的教育主权，中国学者有能力独立自主地开发各种达到心理测量学专业标准的语言测试项目，准确地评价学生的语言能力，更好地为语言教学服务。

杨惠中

2013 年 3 月 18 日

目　录

第一章　语言测试与评估的性质

第一节　语言测试与评估的含义

简单地说，语言测试与评估就是对人的语言能力进行测试与评估的手段。作为测试与评估对象的人，可以是外语学习者，也可以是母语使用者，可以是初学者，也可以是熟练使用者。

语言能力是人们运用语言参与语言交际活动、完成交际任务的能力。人类具有许多种能力，如观察能力、学习能力、分析能力、综合能力、推理能力、管理能力、创造能力、沟通能力、语言能力等，种种能力之间关系复杂，语言能力是这许多种行为能力中的一种。长期以来，人们对语言能力是什么，如何描述和评价人的语言能力之类问题，一直难有一致的意见。在一般场合，人们对语言能力的认识往往比较模糊。比如，在提到某人使用某种外语的水平时，人们经常用“一般”、“熟练”、“流利”、“精通”之类的词语来评价或描述，但对于什么是“一般”、什么是“熟练”、什么是“流利”、什么是“精通”，却没有十分一致的意见。

关注并研究语言能力的学科和领域也有很多，而且不同学科和领域对语言能力的研究在内容和方法上有交叉。

语言学研究语言符号系统及该系统如何运作，自然要研究人的语言能力。生成语言学派区分语言能力(competence)和语言表现(performance)，

并将语言能力作为研究的重点。语言是一种社会现象，不同的人群在地区、种族、文化背景、社会地位、职业等方面存在差异，他们在使用语言哪怕是同一种语言，在发音、用词、语法结构、语体风格等方面也有许多不同，即人们的语言能力各异，关注并研究这种现象，是社会语言学的研究内容。

随着人们社会分工的不同，不同的工作、不同的职位，人们对语言能力的要求也往往不同。比如，导游、翻译、记者、教师等在工作中，语言是最重要的工具，他们需要具有较高的语言能力水平，而驾驶员、厨师、建筑施工人员等的工作对语言能力则没有很高的要求。所以，人力资源部门很重视不同的岗位对语言能力的要求，以便根据这些要求适当挑选和安排工作人员。

人的语言能力又是不断发展的，它是人类所具有的一种特殊的心理属性。儿童随着年龄的增长、心理的成熟，语言能力得到不断的发展，以至达到一般成年人所具有的语言能力水平。人们学习并掌握外语的过程需要大脑进行大量的记忆、提取、理解、分析等活动，对于许多人来说，只有经过艰苦的努力，才能不断提高外语水平，逐渐形成使用外语的能力。因此，语言能力又是心理学研究的重要领域。

人具有语言能力当然也离不开生理的基础。一个人如果发音器官有问题，或者脑部语言区受损的失语症者或言语功能障碍者，他的语言能力就会受到影响。因此，生理学研究人的言语活动的生理机能，研究人如何发声，研究语音器官的工作原理，研究如何帮助言语功能障碍者恢复语言能力。

在众多关注并研究语言能力的学科或领域里，全部工作都围绕语言能力的培养、形成与提高而展开的，是语言教学。在语言教学活动中，设计教学大纲、安排教学内容、编写教材、组织教学等活动，目标都是为了培养和提高学生的语言能力。

语言测试是采用一定的手段对人的语言能力进行测量的工作。比较常见的是在语言教学活动中，教师为了检查学生的学习情况，编写试题并让学生作答，教师根据学生作答情况，给出成绩。这种活动人们习惯称为考试。考试的规模可大可小。小到一班一组甚至一人，大到成千上万人同时考（比

如，汉语水平考试 HSK)。随着外语交流在国际交往活动中日益频繁，范围不断扩大，交流手段与方式不断增多，语言教学规模也不断扩大，参加语言测试或需要进行语言测试的人数也日益增多，人们对语言测试工作科学性的关注程度也自然越来越高，随之形成了一套比较系统的概念、理论框架和工作规范。比如，编写什么样的试题，是多项选择题(multiple choice)还是构答题(constructive)；如何评价一道试题的质量，判断一道试题的优劣；一次测试安排多少试题合适，三十题、五十题还是八十题；怎样才能保证考生的成绩能够真实反映考生的实际语言能力水平；怎样才能保证所报告的考试成绩的可靠性。围绕这些问题，语言测试作为一门学科，已经作了大量的研究，提出了一些基本的原理、原则，形成了一些比较科学的操作规范。

由于教育和人才选拔需要有高质量的语言测试工具频繁地测量大规模人群的语言能力水平，于是形成了许多生产语言测试产品的专门机构，如我国的汉语水平考试委员会、大学英语考试委员会、英国的 UCLES(University of Cambridge Local Examinations Syndicate)、美国的 ETS(Educational Testing Service)等。一些语言测试产品，像汉语水平考试(HSK)、大学英语考试(CET)、雅思(IELTS)、托福(TOEFL)、托业(TOEIC)等，已经成了世界知名的语言能力水平测试工具，每年有成千上万甚至上千万人参加这些测试，这些测试产品对语言教学和外语人才选拔产生了极大的影响。

语言评估是利用各种与语言能力有关的资料和信息对人的语言能力状况或水平进行评价的活动。与人的语言能力有关的资料和信息可以是测试的结果，也可以是平时使用或学习语言的表现，也可以是教师对学生的评价，还可以是同学或其他相关人员的评价，甚至是学生对自己语言能力水平的自我评价，这些资料和信息都可以作为评价一个人的语言能力水平的依据。语言测试所报告的成绩通常作为语言评估的重要依据，但语言评估不一定要进行语言测试，不一定要采用语言测试的结果，它还可以利用其他的信息。因此，语言评估对人的语言能力进行评价时所采用的信息和方式比语言测试的范围更广。

本书主要讨论与汉语测试相关的概念、理论和方法，兼及汉语评估的手段及作用。

第二节　语言测试与语言教学

一般而言，有教学（各种教学活动）就有测试。语言教学过程通常包括教学目标设定、教学大纲设计、教材编写或选用、教学活动组织与实施、教学效果检测与评估等必需的环节，语言测试是对教学效果进行检测与评估的重要手段。

作为检测与评估教学效果的重要手段，语言测试与教学过程中的其他环节之间的关系密切，它们之间的相互关系大致如图 1－1 所示：

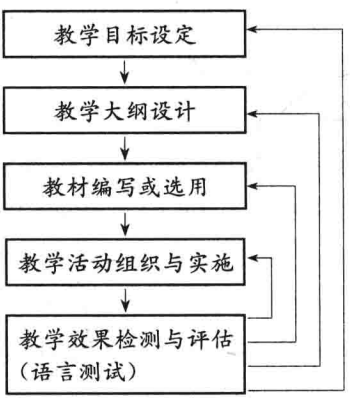

图 1－1　语言测试与教学过程中其他环节的关系

从线性的教学过程来看，语言测试是教学过程的最后一个环节。但在实际的语言教学过程中，语言测试会作用于教学过程的每一个环节，即教学过程中的每一个环节都可能与语言测试有关。

语言测试的结果可以反映语言教学目标的设定是否适当。学生轻易就通过了体现教学目标的测试，则说明所设定的目标可能过低；反之，则说明目标可能过高。

通过语言测试，也可以看出语言教学大纲的设计是否合理。教学大纲往往包括具体的教学内容或教学项目、各项内容所需的教学时间等，通过测试

结果，可以发现大纲规定的教学内容或项目的多少或详略、教学时间的要求等是否合理。

语言测试也可以帮助我们分析所编写或选用的教材与教学目标和教学大纲是否一致。教材是教学目标和教学大纲的具体体现，基于一定的教材编制语言测试，学生参加这种测试所得的结果可与教学目标和教学大纲相对照，由此可以看出使用该教材能否实现教学目标或在多大程度上能够实现目标。

语言测试对教学活动组织与实施的影响更加直接。语言测试可以在一个教学时间段结束时进行（如期末考试），也可以在一个教学时间段的中间进行（如单元测验），测试结果直接反映教学活动的组织与实施是否有效。学期中的测验不仅反映前一阶段的教学是否有效，而且为判断或决定下一阶段是否需要及如何调整教学内容和方法提供依据。

语言教学涉及教和学两个方面。语言测试不仅对教的方面有极大的影响，对语言学习也有很大的影响。诊断测试的结果不仅对改进教学有重要的参考意义，对于“学”意义更加重大：对学生的学习提出诊断意见，告诉学生在哪些方面已经取得了成绩，哪些方面有不足，还需要加强和提高。这些有助于学生明确努力的方向，对学习有提示和鼓励的作用。

大规模的、高风险的语言测试对语言教学的影响更大，这种影响有正面的，也有负面的。一项语言测试重视对语言运用技能水平的考查，就可能引导语言教学单位重视培养学生的语言运用技能，这是语言测试对教学的正面影响。社会上，许多人往往根据学生参加某些语言测试所取得的成绩来评价教学单位进行语言教学的质量，判断学生语言能力水平的高低。一些教学单位为了使参加这些测试的学生取得好成绩，甚至专门开设应试课程，做大量的模拟试题，传授所谓应试技巧，忽视对语言技能的培养与训练，这些是语言测试对语言教学的负面影响。

从事语言教学相关工作的人员（包括教师和教学管理人员等）要进行一些与语言教学相关的研究，语言测试是一种十分常用的工具。比如，研究语言教与学的效果，就经常需要编制测试或问卷，借以收集数据并作实证分析。

第三节 语言测试的心理测量学、统计学基础

与语言测试相关的学科和领域很多，除语言学（包括理论语言学、应用语言学）之外，心理学、教育学、统计学、社会学、计算机科学等都对语言测试工作的某方面起着支撑的作用。其中，心理测量学和统计学的一些基本原理和方法更是为语言测试所借用。

从心理学角度看，语言能力属于一种心理现象，语言测试就是对这种心理现象进行测量，因此，语言测试从理论上说是属于心理测量。事实上，语言测试也确实跟心理测量学息息相关，它在理论和方法上大量采用了心理测量学的理论与方法。

对心理现象（包括语言能力）进行测量，就像测量距离的长短和重量的轻重一样，必须使用一定的测量工具。尺子是测量长短的，秤是测量轻重的，语言测试就像尺子和秤一样，是一种测量语言能力状况的工具。

对心理现象进行量化或测量，通常要使用量表（scale），具体有称名量表、顺序量表、连续量表等。

（一）称名量表

所谓称名量表（nominal scale）就是根据某种标准把测量对象分成不同的类别。如，用1、2、3分别将一个班上的韩国学生、日本学生、印尼学生加以区分和归类。用什么名称去标记不同的组别是任意的。这三组学生可以分别用1、2、3去称名，也可以用2、3、1去称名，或用甲、乙、丙称名。所以称名量表也叫分类量表。

（二）顺序量表

在进行测量活动时，不仅把测量对象分成不同的类别，而且将不同的对象进行排序，建立起大小、高低的顺序关系，这就形成了顺序量表（ordinal

scale)。如,有些语言测试的结果用"级"表示,有一级、二级、三级等。可以规定一级最高,二级次之,三级最低,也可以规定三级最高,二级次之,一级最低,但在一共分三级的情况下不能规定二级最高,一级或三级最低,因为由"一"到"二"再到"三",或者是由"三"到"二"再到"一",都是有顺序的。要保持三者之间的相对顺序,这和称名量表只需称名不同。

(三) 连续量表

连续量表(continuous scale)不仅把测量对象分成不同的类别进行排序,而且表示出不同对象之间距离的大小。编制连续量表须要确定计量的起点和单位。计量的起点也叫"零点",但这"零点"并不等于"没有"或"无",就像温度计上的零度并不等于没有温度一样。比如,一位考生的作文得零分,我们就很难说他完全没有写的能力。这个"零点"只不过是为了测量时计量的需要而设定的一个起点。如果没有起点,就好像一条直线无始无终,是无法测量长短的。计量的单位跟测量对象的性质有关。测量对象的性质不同,测量结果往往用不同的单位来表示。如,测量长度可以用"米"、"公分"等做单位,测量重量则用"千克"、"克"等做单位,商品的价格用"元"、"角"等做单位。在语言测试中经常采用"分"、"级"、"等"为计量单位。像平时的学期考试满分 100 分、汉语水平考试(初、中等)满分 400 分,都是连续量表。

语言测试通常采用顺序量表和连续量表。用这两种量表测得的结果都要进行排序。如,采用顺序量表将测试结果分为六个等级,定义"一级"为最低,那么,这六个等级水平就可以排序为:六级>五级>四级>三级>二级>一级(">"表示"高于")。若采用连续量表(比如,用百分制),分数越高表示水平越高,那么,90 分>80 分>70 分>60 分>50 分>40 分……顺序量表的等级之间的距离不一定相等,而连续量表的每个分数单位之间的距离通常认为是相同的[①]。图 1-2 显示了顺序量表和连续量表的不同:

---

① 认定连续量表的每个分数单位之间的距离相同是主观的。比如,在一份含有词汇测试和语法测试的试卷上,考生在词汇测试部分得到的 1 分就很难说等同于他在语法测试部分得到的 1 分,甚至也很难说考生在一道语法试题上得到的 1 分等同于他在另一道语法试题上得到的 1 分。

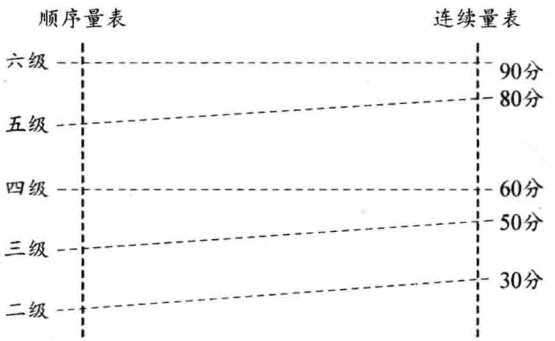

图 1-2 顺序量表与连续量表对照

从图示可以看出，与连续量表上的分数相比，顺序量表上的“三级”和“四级”之间相差 10 分，而“四级”和“五级”之间则相差 20 分。

以上三种量表的测量结果在具有区别性、显示顺序和显示差距大小方面存在差异，如表 1-1 所示：

表 1-1 三种量表的基本特征

量表 / 特征	称名量表	顺序量表	连续量表
具有区别性	+	+	+
显示顺序	−	+	+
显示差距大小	−	−	+

（说明：上表中，“+”、“−”号分别表示具有和不具有某项特征）

在语言测试领域广泛使用的试题难度与区分度、测试的效度与信度、等值等概念，都跟心理测量学原理有着十分密切的关系。心理测量学里的经典测量理论、项目反应理论等都直接影响着语言测试理论、方法与技术的发展与进步。

语言测试往往涉及考生群体方方面面的信息，考生群体人数越多，差异越大，涉及的各方面信息就越复杂。处理大量的多方面的数据信息自然离不开数据统计分析技术。在分析试题难度与区分度、报告并解释考生成绩、进行效度验证、分析测试信度等方面，都经常要利用多种统计技术。