现代语言测试模型

Modern Approaches to Language Testing

王振亚 著

河北大学出版社
Hebei University Press

现代语言测试模型

Modern Approaches to Language Testing

王振亚 著

河北大学出版社
Hebei University Press

图书在版编目(CIP)数据

现代语言测试模型=Modern Approaches to Language Testing/王振亚著.—保定：河北大学出版社，2009.5
ISBN 978-7-81097-360-1

Ⅰ.现… Ⅱ.王… Ⅲ.测试-语言模型 Ⅳ.H087

中国版本图书馆CIP数据核定(2009)第042384号

责任编辑：臧燕阳 Tel:0312-5921826 E-mail:zyyzmq@yahoo.com.cn
装帧设计：王占梅
责任印制：闻 利

出版：河北大学出版社（保定市五四东路180号）
经销：全国新华书店
印制：河北新华印刷一厂
规格：1/32(880mm×1230mm)
印张：9.75
字数：245千字
印数：0001～2000册
版次：2009年6月第1版
印次：2009年6月第1次
书号：ISBN 978-7-81097-360-1/H·71
定价：16.00元

前　言

本书评述了心理测量学－结构主义语言测试、综合语言测试、交际语言运用测试、Bachman 的交际语言能力和测试方法等现代语言测试模型。本书力求对这些语言测试模型的产生背景、理论基础、本质特征、主要采用的测试题型、优缺点等方面做详尽的评述。本书的要点，是心理测量学－结构主义语言测试（本书倾向于称之为语言知识－技能测试）采用的知识－技能测试目标模型存在理论缺陷，但在测试方法方面有灵活性、经济性等优势。综合语言测试的整体语言能力假设已经被推翻，但其倡导的完形和听写测试已被融入心理测量学－结构主义语言测试和交际语言运用测试，在继续发挥作用。作为交际语言运用测试理论基础的交际能力模型，比知识－技能模型先进，但也存在不足。交际语言运用测试对测试方法限制较多，故灵活性和经济性低于心理测量学－结构主义语言测试。心理测量学－结构主义语言测试和交际语言运用测试代表语言测试的发展方向，并有很高的互补性。Bachman 的交际语言能力概念和测试方法理论框架对心理测量学－结构主义语言测试和交际语言运用测试模型均有理论指导意义，但 Bachman 的交际语言能力概念本身也存在缺陷。

书中难免有错误和遗漏，望相关人士批评、指正。

王振亚

2008 年 10 月于北京语言大学

目 录

第一章　语言测试的一些基本概念

一、测试、测量、评估和评述

在教育测量学和语言测试文献中，有四个词义接近，使用频率很高的术语，分别是测试（testing）、测量（measurement）、评估（evaluation）和评述（assessment）。很多测试学者（如 Ebel and Frisbie, 1991; Davies, et al. 1999 等）都曾经讨论过它们的异同。

测试是一种特殊测量技术，以获取量化（由数字体现）的信息为目的，该信息反映应试者所掌握的某一方面的知识或能力的程度。在教育测量（除语言测试以外的其他教育测试）中，典型的测试由一套问题构成。测试中的每一个问题都有一个正确答案。这些问题由应试者口头或书面回答。测试中的问题和测量态度、动力、兴趣、偏好等性格因素以及学习策略、认知风格等认知因素的问题不同。后者由被测量人员根据自己的实际情况来回答，答案的正确与否不是由评分人确定。评分人无从知道这些答案是否反映了被测量人员的实际情况，只能假定被测量人员是诚实的，他们的答案反映了他们的实际情况。而对测试问题答案（包括问答题）的正确性，学科专家会取得一致意见，不会受到他们个人的价值观和好恶的影响。在语言测试中，很多问题也有正确答案。例如，多项选择题、正误判断题、配伍题等一般只有一个正确答案。而完形填空、其他形式的填空题、简短答案题则可以有

不只一个正确答案，但语言专家仍可以就其正确性取得一致意见。使语言测试比较复杂的是有些主观性测试，如口语面试和作文测试，没有正确答案。但在多数情况下，语言专家仍可以就应试者提供的答案是否达到测试要求、达到测试要求的程度或反映出的知识、能力水平达成一致意见，并以分数的高低来体现应试者掌握该项测试测量的知识或能力的程度。

测量指收集量化的信息来决定被测量之物（不限于知识或能力）存在的程度。在这一点上测量和测试是一致的。但在测量中可以使用那些不要求评分人对其答案做出正误判断的问题。例如，我们可以根据被测量人员对一组问题的答案判断出他们有内向或外向倾向。这些答案是否反映被测量人员的实际情况只有他们自己知道，评分人无法，通常也无需，对其做出正误判断。因此，测量包括测试。测试是测量的一种形式。测试由一套可以由评分人对其答案做出正误判断的题目构成。而测量可以由这样的题目构成，也可以由一套无法也无须评分人对其答案做出正误判断的题目构成。测试和测量的结果都必须是由分数体现的量化信息。

评估不局限于量化手段。在评估过程中，也经常使用面晤、问卷调查、观察等定性手段来系统地收集信息，目的是做出价值判断或决定。对一个语言教学计划进行评估，能够为教育管理者、教师，甚至学生家长提供有关语言教学质量的信息，也能够决定该语言教学计划的未来。在语言教学效果评估中，语言测试是常用的手段之一，如运用教学计划前测试和教学计划后测试来考察学习者的进步情况，运用学业测试来考察学习者掌握学习内容的情况等。

评述是使用最宽泛的术语，既可以和测试换用，也可以广义地指收集语言数据，包括测试数据，又可以狭义地指不包括测试的各种评估手段，如面晤、个案研究、问卷调查、观察等。总之评价收集定量和定性信息，供评述人了解被评述之物的现状。

从某种意义上说，评估是评述的一种形式，是需要做出价值判断或决定的评述。测量是评估的一种形式，仅采用定量手段收集信息。测试是测量的一种形式，仅由评分人可以对其答案做出正误判断或是否达到测试要求的判断的题目构成。

二、测试的种类

测试可以按不同的标准分成很多种类。很多教育测量或语言测试著作（如 Harris，1969；Harrison，1983；Heaton，1988；Allison，1999 等）都对测试的种类做过专门介绍。

（一）测试在教育中的作用类别

测试可以根据其在教育中的作用或功能分成若干种类，包括学业测试（achievement / attainment tests）、进展测试（progress tests）、水平测试（proficiency tests）、学能测试（aptitude tests）、诊断测试（diagnostic tests）、分级测试（placement tests）等。

1. 学业测试

学业测试考察学习者掌握教学大纲规定的学习内容的情况。学业测试通常在一门课程结束的时候实施，测试内容的选择应以该课程的教学大纲规定的教学目标和教学内容为依据，不受具体课程和教材的影响。很多学业测试采用标准化测试形式，由测试专家组命题。我国高中的各学科的会考和大学中的非英语专业英语四、六级考试和英语专业英语四、八级考试是典型的学业测试。

2. 进展测试

进展测试又称课程进展测试（class progress tests），和学业测试很类似。课程进展测试可以在一门课程的不同阶段或结束时实施，目标是考察学习者掌握课程或教材内容的情况。课程进展测试通常由任课教师命题，参加考试的学习者人数较少。测试的正式程度一般低于

学业测试。

3. 水平测试

水平测试考察学习者掌握学科知识的水平，不以教学大纲、课程计划或教材为命题依据。应试者可以有不同的学习背景。在这一点上水平考试与学业和进展测试有明显不同。有些水平考试以考察应试者的一般知识或能力水平为目的，例如，我国的公共英语等级考试(PETS = Public English Test System)考察的就是应试者的一般英语水平。有些水平考试则用来确定应试者是否具备接受某种教育或培训的学科知识水平。例如，托福(TOEFL = The Test of English as a Foreign Language)和雅思(IELTS = The International English Language Testing System)都是以考察应试者是否具备接受英语国家高等教育的英语水平为目标设计的。也有一些水平测试用来考察应试者是否具备其所从事的职业要求的语言水平。例如，我国的职称英语考试就属于这一类水平考试。

4. 学能测试

学能测试又称预测性测试(prognostic tests)，用来考察应试者学习某一学科的潜能。学能测试通常在应试者开始学习相关学科之前实施，以预测其将来学习该学科的结果。J. Carroll 和 S. Sapon (1958；1967)设计的 MLAT(Modern Language Aptitude Test)和 Pimsleur(1964；1966)设计的 PLAB(Pimsleur Language Aptitude Battery)是著名的语言学能测试。MLAT 和 PLAB 的测试内容并不完全相同。

MLAT 考察四种能力：

1)语音编码能力：识别语音，建立语音和体现语音的符号之间的联系，并能稍长时间地记忆这些联系的能力。

2)语法敏感性：识别句子中词语的语法功能的能力。

3)归纳式学习能力：在很少的指导下从新的语言材料中推断语言

形式、规则、格式的能力。

4)机械记忆能力:快速有效地学习和记忆音义之间联系的能力。

后来机械记忆能力被排除了,只有前三种能力被保留下来。MLAT 由五部分测试内容组成,分别是:数字学习、语音符号、拼写提示、句子中的词语、成对的相关词语。语音编码能力主要由"语音符号"部分测量。这部分实际测量的是建立语音和语音符号之间的联系和辨音能力。语法敏感性主要由"句子中的词语"部分测量。这部分要求考生在句子中挑出具有同样语法功能的词语。R. Gardner 和 W. Lambert(1965)的研究表明这部分测试和考生的一般学业有密切关系。MLAT 并没有明确地测量考生的归纳式学习的能力。

PLAB 考察三个方面的内容:

1)言语智能:对词语的熟悉程度和分析言语材料的能力。

2)学习动力。

3)听觉能力。

PLAB 包括六个部分:学生各科成绩的平均积分点、兴趣、词汇、语言分析、辩音、语音－符号。言语智能由"词汇"和"语言分析"两部分测量。"词汇"部分测量的是考生的母语词汇知识。学习动力由"兴趣"部分测量。听觉能力由"辩音"、"语音－符号"两部分测量。

在 MLAT 和 PLAB 测量的各种能力中只有语音能力才是语言学能的成分,其他能力很难和一般智能甚至情感因素区分开。但这两项测试都具有语言学习预测能力。

5. 诊断测试

诊断测试的目的是确定学习者学习中的困难和存在的问题或已经学过但尚未掌握的教学内容,以便教师在后面的教学活动中采取补救措施。尽管诊断测试一词在教学和测试文献中使用频率很高,但很少有测试完全为诊断目的而设计。学业测试和进展测试,甚至水平测试,都可以用于诊断目的。辨音测试、词汇测试、语法测试、某些有控

制的写作测试等都比较适合提供诊断信息。

6. 分级测试

分级测试考察应试者的学科知识或能力水平，并以此为依据把他们纳入适当的课程计划中。例如，北京普通高校的公共英语教学实行分级教学模式。入学新生都参加英语分级测试，根据测试结果，他们分别进入一、二、三或四级英语学习。在不实行分级教学的教学单位，应试者可以根据其英语分级考试成绩，分别进入慢班、普通班、快班学习。

(二)其他种类的测试

1. 速度测试和强度测试

速度测试(speed or speeded tests)和强度测试(power tests)是一对测试形式，测量应试者知识或能力的两个不同方面。

速度测试测量应试者解决问题的速度。一般速度测试的题目比较容易，如果不是在时间压力下，一般应试者都可以提供正确答案。但速度测试题目数量大，且有时间限制，因此很少有应试者能够完成。

强度测试测量应试者的知识或能力，题目数量不大，但有难度。应试者常常不能完成测试中的全部题目。原因不是时间不够，而是应试者不具备完成全部题目所要求的知识或能力。

大多数的语言测试都是强度测试。强度测试并不是没有时间要求。几乎没有测试允许应试者自行决定测试时间的长短。一般强度测试要求至少百分之七十五的应试者应能完成百分之九十五以上的题目。

2. 主观性测试和客观性测试

主观性测试(subjective tests)和客观性测试(objective tests)是根据评分的方式来区分的，因此又被称作主观性评分测试(subjectively marked / scored tests)和客观性评分测试(objectively marked / scored tests)。

在阅卷过程中，评分人需要对应试者提供的答案的正确性或满足测试要求的程度做出主观判断的测试即是主观测试。简短答案题(short-answer items)和开放性试题(open-ended items)是可以构成主观测试的试题。简短答案题(如填空题)，答案简短，而且正确答案的数量有限。开放性试题(如作文题)可以有很多答案，常常只有满足测试要求的程度上的差异，而无正误之分。因此，开放性试题阅卷的主观性要高于简短答案题。当测试的结果对应试者有重大影响时，可以通过由多个经过培训的评分人共同阅卷来提高主观测试的信度。

在阅卷过程中，评分人不需要对应试者提供的答案的正确性或满足测试要求的程度做出主观判断的测试即是客观测试。封闭性试题(closed-ended items)，如多项选择题(multiple-choice items)、正误判断题(true-false items)、配伍题(matching items)等，都是客观性试题，正确答案明确，评分人无须做出个人判断。

主观性测试和客观性测试仅相对评分而言。所有的测试都是由命题人主观设计的，都是由应试者根据主观判断来完成的。

3. 分立式测试和综合式测试

分立式测试(discrete-point tests)和综合式测试(integrative tests)反映测试设计者的语言观和语言能力观。分立式测试的倡导者认为语言是由小的语言成分(如语音、词汇、语法结构等)构成，人的语言能力也是可分解的，由具体的成分构成。在分立式语言测试中，一个题目只测量应试者掌握一个单独的语言成分的情况，也就是说测量应试者的一点知识或能力。分立式题目彼此独立。应试者的语言水平由其掌握的语言成分之和体现。在语言测试中，典型的分立式试题是多项选择题和正误判断题。20 世纪中叶，分立式测试在语言测试中占据了主导地位。但也就从那时起，分立式测试受到了越来越多的批评，人们对它的效度提出了质疑。

而在实际语言使用过程中，人们需要借助与语言有关的各种知识

与能力才能满足交际语言运用的需要。例如，写作要求语相、词汇、语法、语篇、思想组织、文化意识等多方面的知识与能力的参与才能完成。而交谈除要求语音、词汇、语法、语篇、思想组织、文化意识等多方面的知识与能力的参与外，本身还是涉及语言理解（听）和语言生成（说）两个语言运用层面的活动。传统的翻译、阅读、写作、口语面试等都是综合性测试。完形程序和听写也都是综合测试。他们有一个共同特点：要求应试者在做一个题目时融合多方面的知识和技能。有些学者认为综合式测试的效度高于分立式测试。但也正因为其综合性，综合性测试能提供的诊断性信息有限，或不能提供诊断信息。

4. 标准化测试和教师自主命题的测试

典型的标准化测试（standardized tests）具有下列特征（不是所有标准化测试都具备下列所有特征）：

1）有标准化的测试内容。水平测试的内容由反映某一语言能力的理论或应试者语言需要的观点的一套测试规范决定。学业测试和课程进展测试的内容通常由教学大纲或课程计划决定。不同试卷的内容要经过测试内容等值研究。

2）由命题小组的专家命题，有一套严格的命题规范。通常，出现在试卷中的题目要经过试测、项目分析和修正。

3）整份试卷要经过信度和效度分析，在这两个方面要达到标准化测试的要求。

4）测试的实施和阅卷有标准程序。测试原始分经常转换成百分位和标准分。

5）根据应试者分数的分布为应试者群体建立常模。

6）标准化测试使用的不同试卷经过使用统计方法所做的测试等值，因此测试的分数总能代表相同的能力或知识水平。

7）标准化测试通常要求达到很高的信度水平，以求在不同的测试环境中实施的测试结果具有可比性。为达到这一目的，标准化测试比

较多地采用客观试题。

8)测试的结果是判断应试者语言能力或水平的唯一依据。

国外的“托福”、“雅思”、“密执安英语语言水平测试”(the Michigan Test of English Language Proficiency),我国的“公共英语等级考试”,非英语专业大学英语四、六级考试都是标准化测试。

顾名思义,教师自主命题的测试(teacher-made test)就是语言教师为自己所教的学生准备的考试。这类测试一般有如下特征:

1)由教师本人命题、实施、评分。

2)测试内容通常直接由课程目标决定,从课程内容中选取。

3)应试者和教师彼此熟悉,因此应试者通常对测试内容、评分标准和测试结果的使用都有所了解。

4)由于只有一名评分人,评分标准比较一致。

5)对学生的最终评估还有其在其他测试和活动中的表现为依据,某次测试不是评估学生的唯一依据。

5.形成期评估和终结性评估

形成期评估(formative evaluation)和终结性评估(summative evaluation)也是测试文献中常出现的一对术语。形成期评估和和终结性评估是对教学进行评估的常用方法。

形成期评估亦称形成期评述(formative assessment),在教学过程中实施,对教学过程进行考察,以确定学生是否按教学计划顺利地学习。形成期评估的主要作用是为教师和学习者提供教学过程的反馈信息,为教师调整教学方式或教学材料提供依据。形成期评估采用各种方法搜集细节信息。教师对课堂学习活动的观察、课堂问答、学生作业完成情况、课堂小考和其他非正规的考查都可以为形成期评估提供信息。每一种方法都可以为形成期评估提供反映教学过程的某一个方面的微观信息。除了评估高度系统化的个别教学(individualized instruction:一种学生可以在学习目标、学习内容、学习方式、学习进度

方面自主选择的教学模式)过程以外,正规测试在形成期评估中较少使用。

终结性评估模式亦称作终结性评述(summative assessment),通常在一个教学计划或一个教学阶段终结时进行,以便确定学习者是否掌握了该教学计划规定的内容,或确定学习者是否掌握了该教学阶段应掌握的教学内容,从而可以进入下一阶段学习。终结性评估常采用正规测试来取得信息。正规测试不能反映学习者学习过程各方面的微观信息,但能反映学习者在某教学阶段或教学计划中的宏观学习情况。单元测试和期末测试是终结性评估的常用方法。

6. 常模参照性测试和准则参照性测试

常模参照性测试(norm-referenced tests)和准则参照性测试(criterion-referenced tests)是根据解释应试者所得分数的方法区分的两类测试。

常模参照性测试的作用是根据应试者在测试中的表现(分数的高低)将其排成等级序列,应试者所得的分数本身没有确切含义,不能说明这些应试者达到何种水平或标准。一组应试者会在同一测试中得到一组分数。这些分数的意义是其在同组分数中的地位或等级。也就是说,一个分数要经过和同组的分数比较才能明确其意义。我们可以为一个重复使用的大规模测试(如"托福"等)建立常模,即应试者在历届测试中取得的成绩,通常由平均分和标准差来表示。通过用标准差作为尺度来衡量一个分数和平均分的距离可以确定该分数在全部分数中的位置,由此确定该分数的意义。我们也可以用百分位来体现分数在全部分数中的位置。常模参照性测试把应试者按分数排成等级序列的作法可以用来建立分界线。比如,我们可以根据需要允许百分之六十、五十或其他任一数值的应试者通过某项测试。"托福"是典型的常模参照性测试。尽管 600 分是一条重要的分界线(很多北美的大学把旧"托福"600 分作为来自非英语教育背景的申请者的入学标准

之一），但这一分界线是根据经验建立起来的，并非事先确定的标准。通过追踪调查，人们发现旧“托福”600 分以上的外国留学生接受以英语为教学媒介的教育没有重大语言障碍。

准则参照性测试测量应试者应掌握的某一目标行为领域的知识或能力的水平。准则参照性测试以事先确定的标准作为测量依据。应试者的分数只与这些既定的标准比较，而不必彼此比较以确定相对位置。准则参照性测试的主要功能是确定应试者能做什么，不能做什么或达到目标的程度。“雅思”是典型的准则参照性英语水平测试，采用九分制，每一个分数对应一套行为描述。知道了某人的“雅思”成绩，就了解了其英语语言行为的特点。典型的准则参照性的学业测试的测试领域是某一课程的内容。从这个意义上说，准则参照性测试对教师明确教学目标和确定在教学中这些目标的实现程度有重要意义。学校中的期中考试、期末考试、毕业考试，公务员、医生、律师、驾驶员、厨师等职业证书考试，公共英语等级考试、原非英语专业大学英语四、六级考试（有分界线的）等也都是准则参照性考试。

因为在测试文献中准则（criterion）一词的多义性，有些教育测量和语言测试学者使用领域（domain）一词来取代它。因此，准则参照性测试又称领域参照性测试（domain referenced tests）。领域可以泛指 1）一套数量很大，紧密联系，又相互独立的技能或行为，2）一些联系松散，相互独立的技能或行为，3）若干套有某些联系的匀质的技能和行为，4）一单一技能和行为。在教育测量中，一个单一的教学目标、一教学单元或一课程计划的一组教学目标都可以看做一个领域。因此，测量教学目标实现与否或实现程度的测试又可以称作目标参照性测试（objective-referenced tests）。学校中的期中考试、期末考试、毕业考试，非英语专业大学英语四、六级考试等都可称作目标参照性考试。

7. 直接性测试和非直接性测试

直接性测试（direct tests）和非直接性测试（indirect tests）是根据