北京外国语大学语言学研究丛书

刘润清　胡壮麟　主编

语言测试和它的方法（修订版）

LANGUAGE TESTING AND ITS METHODS

刘润清　韩宝成　编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

北京外国语大学语言学研究丛书
刘润清　胡壮麟　主编

语言测试和它的方法

（修订版）

LANGUAGE TESTING AND ITS METHODS

刘润清　韩宝成　编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

(京)新登字 155 号

图书在版编目(CIP)数据

语言测试和它的方法/刘润清,韩宝成编著. －2 版(修订版), －北京:外语教学与研究出版社, 1999
ISBN 7－5600－1765－7

Ⅰ.语… Ⅱ.①刘… ②韩… Ⅲ.语言能力－测验 Ⅳ.H09

中国版本图书馆 CIP 数据核字(1999)第 72470 号

语言测试和它的方法(修订版)
编著: 刘润清 韩宝成
* * *
责任编辑: 孙 蓓
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. cn
印 刷: 北京外国语大学印刷厂
开 本: 850×1168 1/32
印 张: 8
字 数: 202 千字
版 次: 1991 年 11 月第 1 版
2000 年 5 月第 2 版 2000 年 5 月第 3 次印刷
印 数: 16001—18000 册
书 号: ISBN 7－5600－1765－7/H·1020
定 价: 10.90 元
* * *

第一章 语言测试的性质、目的及类别

作为语言教师，我们几乎天天和测试打交道。比如说，每次讲授新课之前，可能抽出几分钟的时间复习一下上一课学过的知识，或做单词拼写，或做短文听写等。每教完一课书，可能要进行一次测验，检查一下学生对本课掌握的情况。到学期中间，一般要进行期中考试，期末还要进行期末考试等等。在这样的一个教学过程中，我们不仅可以看到学生的学习及进步情况，同时也了解到教师的教学效果。因此，教学离不开测试。现代教育理论的发展，尤其注重人的素质的教育，强调发挥学生的主观能动性，强调因材施教，要检验教学的效果，离不开对学生的评价，通俗地讲，就是对学生进行测试。

一提到测试，有的老师会认为，“那还不容易?”他之所以认为容易是因为他经常给学生考试，出过多次试题，有些现成材料；或改卷多次，积累了一些试题；或可以估计学生可能在哪些方面弱一些。甚至有的老师认为市场上有各种各样的试题集，从中选些题目，拼成一份试卷，拿去考考学生不就行了吗？等等。这些看法实际上是对测试的一种误解，或者说对测试的性质以及测试题目设计的复杂性估计不足。

为了正确地、更好地、更有效地运用测试手段来检查学生的成绩及评价老师的教学效果，我们认为有必要对测试以及与测试有关的一些基本概念进行必要的说明。

1.1 几个基本概念:测量,测试,评价

除测试之外,工作中我们经常用到“测量”、“考试”、“测验”、“评价”等术语或说法。它们之间既有联系,又有区别,不能混为一谈。

1.1.1 测量

什么是测量? Stevens(1951)认为,“广义而言,测量(measurement)就是根据法则赋予事物数量。”也就是说,按照一定的规则给事物的属性指派数字或符号的过程就是测量。这是迄今为止公认的测量定义。

举例来讲,要测量一下桌子的高度,我们可以拿尺子来量一量,看看它有多高。尺子是人们根据一定的法则制定的量具,利用它就可以把事物的属性,即桌子的高度用数字表示出来,如,0.75米。这种测量属于客观测量,因为它基本上不受观察者的主观判断的影响。

由此可以看出,测量这一定义包含三个要素:

1. 事物及其属性

这是测量的对象或目标。上面提到的对桌子的高度进行测量,属于对物体进行测量,其属性——高度,是可以观察到的,可以进行客观测量的。在外语教学领域,我们感兴趣的是学生的语言能力,而学生的语言能力属于人的心理特征,是无法直接测量的,但是人的心理活动会在人的具体活动和行为中体现出来,所以只能通过测量其外显行为或外在表现特征来推论一个学生语言能力的高低。

2. 指派数字或符号

所谓指派数字或符号，就是用数字或符号来代表某一事物或事物的某一属性的量。如张三在本次阅读考试中得了87分，李四得了92分，我们说李四比张三多考了5分。数字本身没有意义，只是一种符号。我们用它来代表考生的阅读成绩，这时它就变成了量化的数，可以对其进行解释和分析。在一定的条件下，还可以对数据进行运算从而对事物的属性进行推测。

3. 法则

法则是指测量所依据的规则和方法，是测量的关键。法则不好或不可靠，得到的测量结果就会出偏差，失去测量的意义。简单来说，尺子不准，测量的结果就无法使人信服。对客观世界的物体进行测量时，由于有公认的测量法则或尺度，如测量物体的高度、重量等，一般不会出现大的偏差。而对人的某些特性（心理特征）进行测量时，则往往会出现较大的偏差。举例来讲，有几个评委对某学生的英语口语进行评定。评委A认为一个人的口语要好，必须发音准确，而该学生的发音好，所以他给打了个5分。评委B认为流利性最能体现一个人的口语水平，该同学尽管发音不错，但流利性差一些，所以她给他3分。同一名学生，让不同的评委去打分，成绩出现了偏差。这也很自然，原因是他们没有按照一个评定口语成绩的统一法则（rules）去给这名学生打分，结果造成了偏差。这个例子提醒我们，在对人的某些心理特征，如口语表达能力、阅读理解能力等等进行测量时，首先要制定一个便于操作的，稳定的法则或标准。这样得到的测量结果才可靠，才具有可比性。

1.1.2 测试

测试（test）又称测验。不同的心理学家对此下的定义不同。Anastasi（1982）认为，“测试实质上是对行为样本所做的客观的标准化的测量。”这个定义是人们公认的最权威的定义，它包含以下三

个基本要素：

1．行为样本

语言测试的目的是要测量受试者的语言能力。上面提到，语言能力是无形的，如何去测量？只能测量它的有形表现，这里所说的有形表现，是指语言表现，如说出来的话，写出来的句子，对测试题目所做的各种反应等等。这些行为，都是无形的语言能力的有形表现，用心理学术语叫“表征”(manifestation)。所谓行为样本，是指对语言能力表现行为的有效的抽样。我们知道，一个人的语言能力的表现行为会有各种各样的形式，测试时不可能也没有必要把它的全部表现行为都测到，只能选取一部分有代表性的抽样进行测量，然后据此对受试者的语言能力作出推测。

2．客观的测量

所谓客观的测量是指测量的标准是否符合实际。对于一项测试的客观性程度可以从这么几个方面去评价：(1)测试题目的难易度和区分度如何；(2)测试结果的可靠性程度如何？(3)测试结果的有效性如何？这几项指标是衡量一项测试质量的重要指标。

3．标准化的测量

标准化的测量是指在测试题目的编制、测试的实施、记分以及对分数的解释等方面有一套严密的系统的程序。只有这样，测试才有统一的标准，对不同人的测量结果才有可比性。凡是不标准化的测量，都没有可比性。

1.1.3 评价

Weiss(1972)认为，“评价(evaluation)是指为作出某种决策而收集资料，并对资料进行分析，作出解释的系统过程。”与测量、测试相比，评价的含义更广、综合性更强。Bachman(1990)指出，决策的正确与否，一方面取决与决策者本身的能力，另一方面则取决于收集到的信息的质量。在其它条件等同的情况下，如果收集到的信息越可靠，

相关性越强，那么，作出正确决策的可能性就越大。所以说，评价的一个很重要的方面就是要获得可靠的、相关的信息。

在谈到评价与测量及测试的关系时，Bachman 指出，在对个体（学生）作出评价时，我们可以从质量和数量两个方面进行描述，或只描述其中一个方面。所谓质量方面的描述是指对学生的行为作出定性的描述，如某某学生的口头表达能力优秀，书面表达能力优良等；数量方面的描述则是指某次测验的分数等。对于测试、测量及评价三者之间的关系，他用下面的图来表示（见下页）。

从图 1.1 可以看出，我们在对某教育目标（或学生的行为）作出评价时不一定用到测试或测量（如面积 1 所示），这种评价属于质量评价，或叫定性评价，如指出学生在学习方面存在的问题。有时在作出评价时只需测量，而无需测试（如面积 2 所示），对学生的口头表达能力定出级别就属于这种性质的评价。如果要检查学生学习的进步情况，通常就要对学生实施测试，这又是另一种性质的评价，即只通过测试对学生的成绩作出评价（如面积 3 所示）。许多情况下，测试只是作为一种科研的工具或手段，而不是用来作出评价（如面积 4 所示），在外语教学、第二语言习得研究领域，我们经常拿水平测试作为研究的工具。不用测试便可进行测量的情况（如面积5所示）在外语教学研究领域也经常碰到，在研究学生的

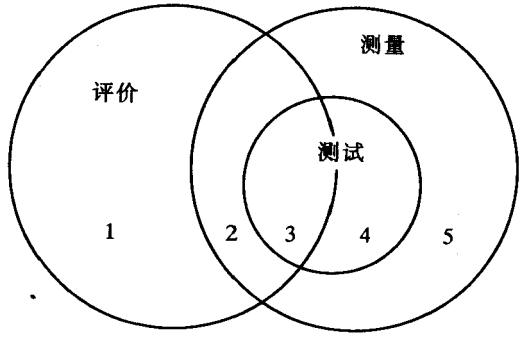

图 1.1 测试、测量及评价三者之间的关系

第二语言习得时，如果研究对象为来自不同国家的学生，人们一般按其母语情况编号。总而言之，并非所有的测量都是测试，并非所有的测试都属于评价，而且并非所有的评价活动都涉及到测试或测量。

1.2 语言测试的目的

通过上面的分析可以看出，测试，包括各种形式的语言测试，实际上是测量的一种形式，或者说是测量的一种工具（量具）。比如我们对学生的语言能力作出测量，就需要设计一份试卷对学生进行测试，然后根据测试的分数来评定学生的语言能力。通俗来讲，就是拿语言测试这把尺子，去量一量每个人的语言能力“有多高”。要想保证测量的准确性，首先必须保证这把尺子是准的，只有这样，才能达到语言测试的目的。那么，我们进行语言测试，通常要达到哪些目的，或者说，语言测试有哪些用途呢？

1.2.1 语言测试用于诊断及反馈

可以这样讲，语言测试首先是用来对学生的学习情况作出诊断，即通过语言测试来检查学生在哪些方面取得了进步，在哪些方面还存在着弱点。比如说，经过一段时间的教学之后对学生进行了一次测试，发现学生甲的阅读理解能力很弱，学生乙的书面表达能力较以前有了进步等，学生丙的发音方面还存在问题等。教师可以把这些信息反馈给学生、学校或者是学生的家长等。另外，根据这些反馈信息，针对不同的学生，教师可以制定相应的补救措施（如改进自己的教学手段等）来帮助学生克服自己的弱点，弥补自己的不足。

1.2.2 语言测试用于筛选或选拔

举例来讲，想到美国或北美其他国家大学读书的学生并非想学哪个专业，就一定能申请得到。校方除了要看申请者以前的学业成绩、有关人员的推荐信之外，还要求学生提交自己的托福成绩（TOEFL，Test of English as a Foreign Language），来决定申请者是否有资格进入某个专业学习。如果想学语言专业，而申请者的托福成绩不太好，对方就不会录取他到语言专业学习。

拿国内的情况来说，每年高考之后，有许多考生要进入外语专业学习，校方在招生时，除了要看考生的笔试成绩之外，还要对考生进行口试，以确定他是否有资格进入外语专业学习。这些都是语言测试用于筛选或选拔的例子。

1.2.3 语言测试用于编班

新生入学后，校方一般要对学生进行一次全面考查，有的称“摸底考试”。“摸底考试”的目的就是为了了解一下学生的入学水平以便把他们编入不同的教学组内，从而使各个教学组的学生在水平较近的起点上开始学习，免得大家参差不齐，互相牵制，从而影响教学效果和学习效果。

1.2.4 语言测试用科研或调查

作为教师，除了正常的教学工作之外，还要搞一些科研活动。如调查为什么在相同的学习环境下，有的学生进步快一些，而有的学生进步慢一些，影响学生外语学习有哪些因素？语言学习的过程是什么？语言习得和语言学习有什么区别？不同的教学方法会产生哪些不同的效果？不同的教材对学生会产生哪些影响等，要

进行这样的调查和研究，就需要根据不同的情况设计不同的试卷，对学生进行测试，然后根据测试的结果来比较、分析、判断、验证所提出的假设等。因此我们说，搞科研离不开测试。语言测试是科研的必要手段和工具。

语言测试的目的和用途还有很多，我们举出以上这些，无非是想说明语言测试的重要性。就目前而言，亟需解决的问题是普及语言测试的基础知识，让广大语言教师知道语言测试的基本原则，设计测试题目的方法及要求，设计出一套好的试题要从哪几个方面去考虑，如何对考试的分数作出解释，如何评估考试题目等。这项工作是外语教学改革中的重要任务之一。

1.3 语言测试的类别

Henning(1987)指出，语言测试有多少个目的，就会有多少种测试类别。此话的确有一定道理。我们设计任何测试，并非总是按照一个模式来进行，测试的目的不同，试卷的内容和要求自然不会一样。下面我们根据不同的划分标准，看看语言测试有哪些类别。

1.3.1 按学习阶段来分

如果按学习阶段来分，一个学期内可能有以下几种测试。

1. 编班测试(placement test)

编班测试往往是在新生入学后对学生进行的全面检查。目的是为了把学生按照程度不同进行分班或分组。现代教育理论强调因材施教，对不同类型、不同水平的学生要分别采取不同的教学方法、不同的教学内容。编班测试关心的是受试者目前的知识水平及能力，它考查的是学生的整体能力。这种测试把学生分出几大

组就可以，不用区分得十分细致。

2. 随堂测试(classroom test)

这是指每教完一课书之后进行的小型测验。这种测试份量小，时间短，不超出一周的教学内容，形式可以多样：拼写、听写、填空、释义、翻译等。最重要的是，题目不宜过难，大部分项目是复习本课知识，同时复习前面的知识。但是，设计这种测试也不是信手捻来的。负责的教师应该从长计划、分课安排，保证这一系统的测试有目的性、连续性、系统性。比如像冠词、介词的用法，首先确定一个学期的目标，然后分散到几次随堂测试中去，使一些重点项目不时地得到体现。这样做有助于帮助学生明确学习重点，帮助教师掌握教学情况。

3. 期中测试(mid-term test)

学期中间，一般要停课一周，进行复习，然后进行一次比较系统的测试。这种测验，不仅让学生在心理上有阶段感、轻松感，而且使学生有机会独立思考，对知识进行系统化。语言的规则不是互不相干的，而是密切联系的。孤立地学习一条规则时，学得再好也有局限性：由于学生所接触的素材有限，对规则的理解仍很肤浅。到了一定阶段，有机会把几条规则联系起来使用，是很有益处的。期中测试有利于这种认识过程。为了达到这个目的，期中测试的设计更应该加以研究。它不仅要体现教学大纲，突出重点项目，而且在随堂测试的基础上，要具有一定的综合性和系统性。使试题具有综合性和系统性是很不容易的，决不像某些教师认为的那样，只要考最后两三课书(自然是比较难的)就可以了。殊不知，最后两三课书不一定有代表性；不能说最后的语言事实学好了，前面的也一定都会了。应该说，某种语言事实或规则以不同形式多次出现过。设计的题目要能引导学生去对这种事实或规则进行综合分析，从而在更高的水平上认识它，掌握它。

4. 期末测试(end-of-term test)

与以上三种相比，期末测试应范围更广，分量更重，时间也更

长些。期末测试有三个目的:促使学生巩固所学知识,评价一学期的教学效果,调整下学期的教学安排。设计期末测试的原则是,以教学大纲为依据,全面反映出该学期学生应该掌握的教学内容,但是不再严格地参考教科书的具体内容,而是变化语言材料来考查学生对所授知识的掌握,同时测定学生解决问题的能力。

1.3.2 按用途来分

按照测试的用途来分类,可以分出四类语言测试:水平测试(proficiency test)、成绩测试(achievement test)、潜能(或素质)测试(aptitude test)及诊断测试(diagnostic test)。

1. 水平测试

水平测试用来测量学生的语言能力,即看看考生是否达到某一水平,从而决定其是否能胜任某一任务。这种测试与过去的教学内容和学习方式没有直接联系。它不考虑考生从前学过没有,也不考虑是如何学的。像美国的托福考试,英国的剑桥英语水平证书测试(University of Cambridge Certificate of Proficiency in English),我国自行设计的英语水平测试(English Proficiency Test,简称EPT)等,都属于水平测试。

2. 成绩测试

成绩测试考查学生对所学知识的掌握,它一般要参考某种教学大纲,甚至考虑到教学方法。上面讲的随堂测试、期中测试和期末测试,以及各学校的毕业考试,都属于成绩测试,因为它们都是针对以前所学的内容而设计的。我国的高考外语试题严格来讲属于成绩测试,因为它必须参考中学的外语教学大纲;但它又是水平测试,因为它常常包括一些考查学生解决问题和分析问题的能力的项目。

3. 潜能测试

与水平测试不同,潜能测试用来预示学生学习某种语言的潜

力和天赋。它不基于某种教学大纲，也不关心考生目前学会了多少东西。有时考生可能从未学过或从未接触过这种语言。现代心理学研究发现，有的人天生可能就有学习语言的天赋，利用潜能测试就可以发现和鉴别这些人材，以便发挥他们的特长。

潜能测试主要是测试受试者是否具备将来学习语言的天赋，无论与水平测试还是与成绩测试相比，潜能测试设计的题目往往多一些，目的是通过考查受试者模仿、记忆等方面的能力，判断他学习语言的潜力。

4. 诊断测试

诊断测试的应用目的与成绩测试恰恰相反，因为成绩测试所关注的是学习成功的程度，而诊断测试关注的则是失败的程度，即学习者在哪些方面犯了错误并借此找出补救的办法。诊断测试有时也用来发现教学方面存在的问题。诊断测试可以用来考查单个语言项目(如时态)，也可以是综合性的，其目的是为了改进教学，调整教学计划，进行个别指导。

成绩测试是回过以前，水平测试主要是展望未来，同时也注意过去，而潜能测试只是预见将来。诊断测试检查以往以图补救今后。

1.3.3 按考试方式来分

按照考试的方式，语言测试可分为分离式测试(discrete-point test)和综合性测试(integrative test)。

1. 分离式测试

所谓分离式测试，是指考题把知识和能力分解为若干小的单位，逐个地进行测量。例如，我们可以把语言分解为语音、语法、词汇等，然后在设计相应的测试题目。分离式测试一般集中考查语言的某一方面，或考查学生单方面的技能。其考试形式主要为多项选择题。

分离式测试的理论基础是:语言是由许多成份组成的,掌握一种语言就是要掌握这些构成成分;测试一个人的语言水平也就是考查他对这些成份的了解和使用。但有的人反对分离式测试,理由是:各种单项知识的总和不一定等于对语言全面的掌握。

2. 综合性测试

综合性测试是指一次同时考查语言的多方面的知识和技能的测试。现在常用的听写、完形填空、翻译、作文等都属于综合性测试。看下面这个例子:

One day, the wife of a Chinese king sat watching a worm as it ate some mulberry leaves. Soon it stopped _____. Then as it slowly turned its head from side _____ side, a very fine thread came out of its _____. It wrapped the thread around and around itself until it was shut _____ a little cocoon.

这里说的是蚕的发现。所填的空格分别是动名词、名词和介词。然而,要正确到填入所需的词,必须懂得上下文的词义和句法关系否则只是瞎填,不可能答对。这种题目表面上来看是让填单词,但也包含了词义、词类、句法及对上下文的理解,所以这种题目属于综合性测试题目。

虽然人们认为综合性测试可以较全面地考查学生的外语能力,但它也有自身的弱点,如进行写作、翻译、口试时,评分标准往往不好掌握,大规模测试中的评分工作需要大量的人力和时间。

1.3.4 按对考试分数的解释来分

不论举行什么考试,为了使考试的结果有意义,必须确定分数解释的参照标准。依据参照的标准不同,语言测试分为常模参照性测试(norm-referenced test)和标准参照性测试(criterion-refer-

enced test)。

1. 常模参照性测试

先解释一下什么是常模。常模是指一群类型相同的人在一类考试中的成绩,这个常模一般用该考试的平均分与标准差来表示。常模参照性测试是指参照某一个常模来对某考生的分数作出解释。假设某次 TOEFL 成绩的平均分为 512 分,标准差(以后要讲到这个概念)为 66,某考生在这次考试中得了 578 分,正好比平均分多出一个单位的值,即一个标准差的分数(512 + 66 = 578)。按照正态分布的原理,84.13% 的考生成绩低于得 578 分的考生。由此可以看出,常模参照性考试实际上是结合其他考生的得分情况来反映一个考生的分数,说明他在这个人群中的位置。这种方法特别有利于选拔学生。

2. 标准参照性测试

与常模参照性测试相反,标准参照性测试指在对考生的成绩作出评判时,参照一个事先规定好的尺度或叫标准,与这个尺度或标准相比,看看他是否达到了既定的要求。标准参照性测试所给的分数不是相对的,即不考虑其他考生的得分情况。如单词听写测试,如果听写 50 个单词,考生能写出 40 个就算通过,那么凡是能写对 40 个的考生都算通过。社会上有很多测试属于标准参照性测试。如驾驶员领取驾照,律师领取营业执照的考试都是标准参照性考试。

1.3.5 按试卷的评阅方式来分

Pilliner(1968)指出,按照试卷的评阅方式,语言测试可分为主观性测试(subjective test)和客观性测试(objective test)。

1. 主观性测试

主观性测试指试题的答案比较灵活,需要阅卷人对考生的作答情况作出主观判断的测试。语言测试中,简述题、翻译题、作文、

口试等都属于主观性题目。有的人认为,主观性测试命题容易,考生靠猜测得分比较困难,题目一般要求考生自由地表达思想,所以容易测出考生实际使用语言的能力。但主观性测试也有明显的缺点。第一,主观题考查的语言现象有局限性。第二,评分较麻烦。比如考生的一篇作文,让不同的评阅人去打分,因每个人的观点、看法、印象不同,最后的打分情况可能相差很多,尽管有时有严格的评分标准。我们认为,在较大规模的语言测试中,考虑到人力、物力、财力等方面的因素,主观性测试题目的量要适当。而在较小规模的测试中,尤其是以班级为单位的小型测试中,主观性测试题目的量可以多一些。

2. 客观性测试

与主观性测试相反,客观性测试答案惟一,不受评阅人的影响。多项选择题属于典型的客观性测试题目。客观性测试的主要优点是,答案固定,评分简单,多数情况下可以使用机器来阅卷,因此能节省大量人力、物力和时间。另外,客观性测试覆盖面一般较大,针对性较强,特别适合分离式测试。

然而,多数客观性测试只要求受试者打钩、填图字母、画圈等,再加上不少的猜测因素,无法测量考生实际使用语言的能力,因此遭到越来越多的语言教师的反对。评分人认为各类考试中滥用客观性测试题目对教学和学习带来的很坏的影响。

1.3.6 语言测试的其他分类

近年来,由于语言教学法中的交际法(communicative approach)的发展,又产生了交际性测试(communicative testing)。这种测试的基本思想是,语言能力不仅包括词汇、语法等知识,而且包括交际能力,即用得体的语言完成交际任务。这种测试题目的特点是,在选择项中,正确答案是语法正确并符合社会规范的句子。语法正确而不得体的句子,或者得体而又有语法错误的句子,