语言测试原理

（英）阿兰·戴维斯／著
任福昌 吴 平 任筱萌／译

经济科学出版社

语言测试原理

（英）阿兰·戴维斯 著

任福昌 吴 平 任筱萌 译

经济科学出版社

1997年·北京

图字：01—97—1483 号

语言测试原理
（英）阿里·戴维斯 著
任福昌 吴 平 任筱萌 译
*
经济科学出版社出版、发行 新华书店经销
三河永和印刷厂印刷
出版社电话：62541886 发行部电话：62568479
经济科学出版社暨发行部地址：北京海淀区万泉河路 66 号
邮编：100086
*
850×1168 毫米 32 开 6 印张 152000 字
1997 年 12 月第一版 1998 年 6 月第二次印刷
印数：1001—3000 册
IBNS 7-5058-1287-4/G·241 定价：9.20 元

图书在版编目（CIP）数据

语言测试原理/（英）戴维斯（Davys，A.）著；任福昌等译. -北京：经济科学出版社，1998. 2

ISBN 7-5058-1287-4

Ⅰ. 语… Ⅱ. ①戴…②任… Ⅲ. 语言-测验-研究 Ⅳ. H09

中国版本图书馆 CIP 数据核字（97）第 23403 号

责任编辑：齐伟娜

责任校对：段建英

封面设计：张卫红

版式设计：代小卫

技术编辑：刘 军

序

从一般意义上说，教和学的目的都不是，也不应该是为了测试（test）或考试（examination）[有关这两个术语，不少人是作区分的，戴维斯（AlanDavies）认为考试是测试中特殊的一种，其差别可以忽略不计，笔者同意这样的观点，在下面的论述中将它们视之为同质的]。事实是，测试在教学中总是处于主导地位。以高考为例，为了使尽量多的学生通过高考进入高等学校学习，几乎每所中学都会采取各种措施，包括文理分班，搜集各种高考辅导材料，由最富有经验的教师授课，组织课外辅导班，定期开家长会等等。因此，不管承认不承认，测试就是指挥棒。

既然测试是指挥棒，它的反馈作用可以影响教学的各个方面，考卷设计和命题就是一件非常严肃的[illegible]考试是教学工作的一部分，由此类推，研究测试理论，提高命题水平不只是专家的事，也是每个教师科研工作的一部分。

每个教师在开课之前都必须回答三个问题：为什么教？教什么？怎么教？测试也不例外，在着手命题之前，我们必须弄清测试的目的、内容和方法。当然，对于广大的教师来说，主要回答的是后两个问题，亦即测什么和怎么测的问题。语言跟数学等一些自然科学是很不一样的。乔姆斯基把语言分为语言能力和语言运用两个方面。显然，掌握语言知识是熟练地运用语言的必要条件，但不是唯一的条件。语言的特殊性意味着语言测试的异质性，这尤其要求语言教师具备必要的语言测试知识。但从我国当前外语教学领域的情况来看，尽管一些著名专家、学者对语言测试的研究已达到了相当高的水平，对测试的许多方面都作过精辟的论

述，但不能说每一个语言教师都能在理论指导下从事试卷的设计和命题工作。事实上，客观性测试系最先进的测试这种说法仍时有所闻，更令人不安的是，多项选择题这种题型似乎用得太滥了，撇开是考语言还是考解题技术这一问题不谈，掌握语言不只是个识别的问题。因此，普及测试的基本知识仍是一件刻不容缓的事。

《*Principles of Language Testing*》（《语言测试原理》）系戴维斯的重要论著之一，该书出版于1990年。戴维斯早年从事英语作为母语和第二语言的教学，有丰富的语言教学知识和经验。他60年代初开始致力于测试的研究，已有30余年，著作甚丰，在理论上颇有建树，曾在英国爱丁堡大学应用语言学系执教，现在澳大利亚墨尔本大学任澳国家语言文字研究所语言测试研究中心主任，在语言测试领域是一位很有影响的人物。该书提纲挈要地回顾了30年来语言测试的发展史，亦即语言测试何以从应用语言学的边缘成为核心，测试理论如何随着语言理论的发展而变更。在横向上，该书对语言测试的目的、内容和方法以及对语言测试的评价作了广泛[illegible]在探索过程中，作者对测试中的一系列关系都作了比较[illegible]的分析，其中包括语言测试与语言教学和语言学的[illegible]主观性测试与客观性测试的关系，常模参照测试与标准参照测试的关系，信度与效度的关系等等。

该书已由任福昌等人翻译成中文。该书中文本的出版无疑会大大提高我们对测试的认识并改善我们的试卷设计和命题的水平。

方　立

1997年7月31日于北京语言文化大学

目　　录

第一章　绪　　论

本书认为，语言测试在语言教学中处于中心地位。它为语言教学提供了目标，并对教师和学生成功地实现这些目标进行监控。它对教学的影响（即所谓的“反馈”作用）是强大的，而且人们通常感到这种影响是消极的。语言测试对语言的讲授和学习提供了试验和调查的方法学。其影响如此之大，作用如此之突出，值得给予比以往更多的重视和研究。如果把有关语言教学的著作进行比较，就会发现这些著作无论是有关方法学、教学大纲、语言资料的，还是有关语言理论的，都对语言测试的阐述极其简略。毫无疑问，造成这种现象的原因，在一定程度上是由于怀疑对语言的各种定量分析方法，同时也还因为过分相信充满人类智慧的文化知识不能简化为纯粹的数字的结果。

语言测试把教学引向注重评价，并时常冠以“考试”的名称，出现这种情况自然也是无可厚非的。所谓“反馈”作用流行甚广，以至于人们愿意接受它是可以理解的，而终止它则被认为是否定消极的，并试图把它搞得尽善尽美，使其影响达到最大的程度。其实不管怎样（不能或不可能发生），评价不会毫无作用。然而，持这种观点的人又非常广泛，并且认为：

1. 大多数语言考试缺乏结构方面的考虑，考核主要集中在打分上，而不是集中在考前准备上。

2. 忽视语言考试给语言教学带来的严重影响。这里，语言测试与考试在思想上确实存在着一种有趣的区别。一般地讲，测试的影响较大，并要求有详尽的计划和管理，而考试则无需如此精心准备。

我们认为，考试（examination）是测试（test）的特殊形式，后面的讨论一般将忽略这种区别。我们不但将讨论其研究价值，而且还将讨论与语言测试有关的实际和理论问题，以及它对有关语言和语言学习的思辩性描述和理论阐述的贡献。

我们的观点是，语言测试在语言教学中，的确处于中心地位。有鉴于此，它在相当大的程度上是从属于应用语言学这一学科的。本书的目的就是要研究这种中心性，并考察语言测试、语言教学和应用语言学之间的相互关系，进而提出测试在多大程度上处于从属地位，在多大程度上处于主导地位。

语言测试的研究具有典型的工具性和机制性，主要解决“是什么”和“怎么做”的问题（Oller1979，Henning1987）。现在，所需要的也是我们努力要做的是，更能从人文主义角度陈述语言测试的范围和作用。我们要研究的问题有：测试的目的是什么，与“是什么”和“怎么做”相对应的测试的原因问题，即为什么要进行测试。我们坚信，语言测试具有主导作用，反馈作用具有积极影响。我们还将强调语言测试对于应用语言学的贡献，在我们看来，语言测试已成为应用语言学的一个密不可分的部分。我们还将进一步讨论应用语言学对语言测试的发展产生影响的范围。我们认为，语言测试已经成为应用语言学中突出的优势，而应用语言学离开了语言测试是不可想象的。这一观点在后面还要深入讨论。

本书无意囊括语言测试的各个领域，也不想使本书成为入门式的教科书。其他著作有此类作用（如 Oller 1979，Henning 1978，Hughes 1989，Bechman 1990），但我们的目的是研究应用语言学学科范围内的语文测试的范围，进而支持有关语言测试发挥导向作用的观点。语言测试对应用语言学的贡献有：

1. 使应用语言学的理论框架转为实际运用；

2. 成为为教学进程和教学大纲确立的目标和标准；

3. 为应用语言学的经验研究提供方法学，无论这种研究是否

就是语言测试研究，对语言的掌握、判断、通俗性研究、理解和使用进行调查；对语言教学的方法和材料进行比较性试验。

语言测试所做的是让人们的注意力集中在语言学和应用语言学的主导思想的含义上。直到这两种学科已运用到实践中，其思想也得到了描述和解释，但还是模糊不清，而且含义多变。测试构成了种种选择，排除了模糊不清的东西，并揭示了到底回避了哪些东西：在测试者明确地提出他或她的主导思想的基础上，测试是描述语言的最为明确的形式。这种作法适用于各种测试，尤其适用于语言测试，因为语言的本质即是无限的，又无法一一说清楚。照此看来，一些富于创造性的技能（如音乐方面）和以掌握内容为主的科目却完全不一样，因此，语言只是一个特例而已。

语言和语言学习这种特例有三个基础：知识——控制关系；样本问题；以及对测试以外的而非其自身的标准进行界定的问题。语言能力似乎取决于知识和控制，这两者之间的关系最终也不能强行拆开。精通一门语言与了解一门语言，这两者之间的平衡一直是个难题。

语言测试的样本问题影响着各种语言行为，包括语言学家、教师和测试者的行为。因为活语言无法进行全面的描述、教学或测试，因此，测试样本是个两难的问题，既可说是完全不恰当的，又可说是十分恰当的。但是，从必要性上讲，测试者还是使用了各种样本。让我们对测试样本做一些解释，从这些解释中，也可以推论出对语言学家和教师的样本完全相同的结论。如果说使用测试样本是不恰当的，这是因为任何样本都无法展示语言的多变性和无限性的本质；同时，这些样本又是恰当的，因为语言的生命力是依靠其自身的创造力来体现的。

因此，对语言测试的要求是苛刻的。测试者则认为："测试代表了或本身就是语言。"对此，尽管我们已经表达了一种极端的观点，但事实是任何认真的语言测试都必须带有强烈的样本意识。除了其他的考虑之外，尽管样本浓缩了语法和教科书内容，覆盖了

语言的主干，并且占用了大量的时间去阅读、研究和教学等，与此相比较而言，测试所占用的时间还是太短了。的确，一项用来检查语言熟练程度的测试，如超过三个小时，往往就被认为时间太长了。而且在这三个小时之内，到底包括了多少语言内容呢？这一困境在全面检查语言熟练程度的测试中表现得更为突出。

语言标准的不精确性产生了测试的模糊性。这里的问题是，假如语言外在标准的有效性是可以找到的话，那么，这种标准既无法界定，也不可靠。下面，我们举两个例子来说明：一是中学毕业前夕举办的语言掌握程度的测试；二是用来决定进一步深造的语言熟练程度的测试。在这两种测试中，应该使用何种外在标准呢？在第一种测试中，标准是教师的判断，还是通过了升学考试？或者是在实际需要语言的环境中适宜程度的表现？在第二种测试中，标准是学生在学业结束前完成了学业（如获得学位、毕业证书或博士学位），还是学生掌握了主修课程的内容？这些标准都无法准确地界定：学业成功似乎是可以界定的，不像“基本掌握”的说法，这一说法客观地讲是无法达到的，但是，我们自然无法弄清什么是学业成功或者什么是不成功。在本书的作者看来，“不成功”（或“成功”）只有在一段较长时间临近结束时，才能显示出来。例如，一名连续10年未能毕业的博士生是否每年都可算作不成功呢？更进一步说，这种成功也很难归因于语言，因为它必然和许多其他因素相联系（如一般能力和专长，努力程度和实际应用水平，身体状况、动机和运气等等）。

当然，我们常常可以听到有人说，这种观点过于琐细，完全能够回答，这正如亚历山大对高迪安难结（The cordian Knot）采取快刀斩乱麻的态度，或是象约翰逊博士攻击伯克雷主教时那样的作法。这并不是说，当我们考察成绩、流利程度和水平时，我们已经了解它们了。但是，这种并非一派胡言的方式将无法奏效，因为观察者们的观点彼此不同。这样，我们又回过头来去控制我们所能控制的东西了，这就是测试，因为测试是最佳标准。

标准的不适当、缺乏信度以及常常表现出的无效性，这一切正是直接测试和间接测试要解决的永恒问题。假如真像我们曾认为的那样，根本不存在什么标准，而且测试本身的责任就是首先建立自己的标准的话，那么合乎逻辑的推论就是直接测试（如口头测试）相对于间接测试（如听写或语法测试）而言，则没有任何特殊的价值，因为任何可靠的行为都不足以作为标准。因此，间接测试可以按自身的标准进行，正象直接测试那样。幸运的是，从我们对语言测试的要求以及持有的明智判断出发，我们不能把测试划分得过于绝对化。无论是外在标准还是直接测试都并非毫无希望或毫无意义，而无法不屑一顾地将其抛弃。我们不久将会看到这一点。

我们需要考察语言测试涉及的三个领域：语言、评判标准和语言能力，并介绍本书的有关章节。

语　言

人们可以从不同的方面研究语言，譬如从语言学的角度，包括心理语言学和社会语言学；从应用语言学的角度，包括语言的教与学；也可以从非专业的角度进行研究。语言学的观点认为，语言是各种系统（如语调系统、时态系统等）的网络，这些系统之间存在着相互联系，同时又和非语言系统相互联系。系统的形成是结构性的，其联系也是结构性的。语言系统主要有音系学、语法学和语义学系统。这些系统和语音学、讲演学一起标明语言系统逐步蜕化的极限。

心理语言学认为，语言首先是生产过程的认知行为，是与发展相联系的学习和思维。对语言能力进行概念界定，属于心理语言学的边缘范畴，因为任何测试都是与形成概念的某种能力相联系的，也就是说是与思维过程相联系的。

社会语言学认为，语言在时间和空间上具有内在的多变性，而

且从系统的角度来讲也是如此。因此，社会语言学努力的目标是双重的：一是把社会环境同语言学的理论与阐释结合起来，这样，语言的规则就可以确切地描述社会环境对规则的影响；二是描述语言与社会间的系统性关系，也就是说把语言看作是有声的社会现象。从某种程度上讲，社会语言学的第二个目标把对社会学的关注同对语言学的关注看得同样重要，也许更重要。然而，它并未说明对语言变化的研究到何处结束，也未说明社会现实研究从何起步。例如一篇与性有关的演说，既可以从社会变量的角度认为是系统变化的一个方面，也可以认为是反映社会现实的一种形式，它传递了有关男性与女性的社会作用的信息。同样，双语种的交换，既可以从社会角度认为，是一种语言的联系，或者语言的丧失或保留，也可以从与克里奥耳化相似的进程角度认为，是社会发展进程作用于语言变化而产生的社会语言学的影响。然而，就语言测试而言，正是第一种观点，即语言学内部变化的结合，才是语言测试的着眼点，也才更有价值。

语言学家一直试图通过对居于中心的语言术语变异的阐释（它总是在方言学形态的外围进行）扩展为语用学和会话，进而在核心理论中界定语言的变化，但这一努力也提出创造可变规则或“软”规则（罗斯 1979）等方面的要求。因此，语言测试必须更具兼容性，内容更丰富。但出于规则的需要，不能把由实际操作的理想化而给语言学的灵活性造成的困难和给评判标准造成的困难等量齐观。当前或许更有希望的不是把社会语言学的发展融入测试中，而是在测试中，融入第二语言习得研究中，测试构想正在发展的程序和已被认可的方法。这种语言间互动的研究，没有对发展提供出各种要求，倒是更加关注语言个体的发展。

评 判 标 准

正如不久我们将会谈到的一样，由于我们认识到可信范围内，可能存在的不确定性，因此，评判标准的准确性取决于对信度和效度的评价。信度可以直接用各种方式来衡量，概括地说，就是具有可比性的两套测试类型之间的比较，从而可以估计评判标准的连续性。根据“错误规范评判法”，可以在测试分数中设计一定数量的模糊问题，也可以用这种方法来评估信度。从逻辑上说，分数围绕某一中间值向外扩展得越大，它被取代的可能性也就越强。这种方法的不利方面是，假如每个人的得分相同，那么，从信度角度而言，测试结果等于没有提供任何信息。

项目分析既涉及信度，又涉及效度。其根据在于测试类型具有较高的区分度，也就能被重复使用，而项目之间显而易见的区分度本身就预示着测试的设计是令人满意的，也就是说把各种类型的考题集中在一份试卷中。项目分析方法的目的是确定测试的同质性，两次测试的类型之间的相同性越多（不是完全一样），他们在相同范围内得以评判的可能性也就越大，因此可以看出，测试的结果是有效的（即效度），同时这种结果具有稳定不变的性质（即信度）。这两种方法可以使人们相信，把测试类型的分数相加起来是一种需要，而且还能获得测试的总分数。测试类型的同质性越强，把各自的得分加起来的效果也就越明显。其原理如同卷尺的原理一样，后面一个厘米或英寸的长度可以累加到前面，也就是说，他们具有累加性质。项目分析的其他方法（如项目反应理论）将在后面分析。

尽管信度着眼于如何确保测试是一种评判的尺度，而效度则着眼于这种尺度的本身，并试图提供一种理论分析框架，这种框架旨在给测试提供保险的机制。人们不必为存在着若干种效度的事实而感到惊奇。这些形式将在下面谈到，但应该指出的是，它

们都是关于测试手段与测试范围之间的关系问题。我们将会看到，测试范围可以界定得较狭义一些，如从卷面效度入手；也可界定得广泛一些，如从结构效度入手（详见第三章）。

测试评判问题也是一个实际问题。只有实际上可行，测试才会成为可能，记住这一点是非常重要的。当然，一项完善的测试（极其有效和可靠）实际上往往是不可能的，因为在其所需要的环境中，测试要占用大量的时间和训练有素的人力，甚至需要代价高昂或精心准备的媒介系统或评分程序等。那么（人们可能要问）怎样才能使测试既可靠又有效呢？这是一个很合理的问题。关于“有效性”的一种解释是，如果测试在实际上不可行，那不得不承认它也就缺乏效度，因为这种测试对其所要测试者来说则是无用的。然而在实验性样本的基础上建立信度和效度则是可能的，那么作为其运用的结果，承认在较为不利的条件下的不现实性也是可能的。

语 言 能 力

语言能力必须从测试用途和测试目的来体现。我们至少可以分为五种用途：成绩测试、水平测试、性向测试、诊断测试以及潜能测试。测试的目的可以分为两类：以常模为参照的测试（Norm－referenced）和以标准为参照的测试（Criterion－referenced）。这些将在以后各章中全面讨论，这里将简要提及一些。

在成绩测试中，目的是要衡量对所讲授的内容的掌握程度，或者是对教学大纲、教科书及其他资料的掌握程度。换句话说，这类测试是以提供给学生清晰和公开的教学内容为基础的。而水平测试则不是针对公开发布的教学内容，而是测试对语言的掌握和特定语言用途的关系，如考察被测试人是否掌握进行学术研究所应具备的语言能力，或者做医生、飞行员、汽车司机和滑冰教练所应具备的语言能力等。

性向测试考核的是学习语言的能力。进行这种测试的目的是要体现前面种测试没有覆盖的语言学习动力（过程）问题，因为前两种测试是对某一些时间点上的能力和知识的考核。从另一个角度看，性向测试是为了考核发展的潜力。诊断测试确实只是名义上的，但与水平测试有着严格的区别，这并非模糊不清。“真正的”诊断测试是为了检验学生到目前为止的情况（限于某些特定的领域），通过学生拥有的某一方面的知识，从而确定修正错误所需要的知识。潜能测试（Pre--achievement）实际上是成绩测试的一种特例，其目的不是考核已教过的知识的掌握情况，而是考核将要讲授的知识。许多入学考试都属这种类型，因为其内容与将要使用的教学大纲紧密联系。这种测试就好比筛子过滤一样，成绩优异的学生可以免修后面的课程。

常模参照考试的理论基础是，特性（如身高）与能力（如知识、语言）的正态分布概念。毫无疑问，这种概念可能是不真实的，测试有规则地分布仅仅是由于最初设计的原因，而不是因为确实反映了现实。常模参照考试一般利用等级排列顺序，它在一定范围内将个体按等级进行排序。近来正在讨论的标准参照考试似乎已表明自己是一种新方法。在这种测试中，标准或目标取代了排序，它们与考试任务相联系，也就是说，现在强调的是将要承担的各种任务而不是个体。但是，依据类推法，以标准为参照的测试一直作为以常模为参照的测试的特殊使用形式，因为在后一种测试中，成绩测试或水平测试是经常被使用的，而前一种测试无法客观地选择出一套比例。这两种测试从本质上说，是同一现象的两个方面：一方面考察人们能做什么，另一方面考察人们需要做什么。

接下来的各章都围绕着一个总的主题，即需要的基础性的语言测试不确定性与对测试确定性的要求之间创造性地寻求一种平衡。用来解释这些概念的更为常见的术语是信度和效度，我们将广泛地使用这些概念。但当我们想坚持这种平衡是一种普遍要求

时，我们也会使用不确定性和确定性这两个概念。

第二章提供了本书使用的术语和概念的定义。这一章是绪论已进行讨论的继续。在第三章，我们从定义开始深入讨论语言测试所能提供的信息。在定义和划定界限之后，第四章将努力为语言测试在应用语言学中定位，首先是通过连续体类型范畴的概念这样做的。在第五章中表达了这样的观点，即语言测试是已知知识和未知知识之间的平衡，正如我们已经看到的，这种平衡用效度、信度及其他相对称术语来表示。我们将为这些统一体建立一种模式，并在效度和信度之间提出一种调和的解决办法。

第五章主要讨论不确定性和确定性问题。在这里，我们将介绍目前研究的主要内容，即语言测试不能没有两个极端的平衡。这两极是由效度和信度之间巨大的分野来体现的。第六章论证了语言测试在应用语言学中的中心作用问题。本书提出了与此有关的各种观点，这些观点结合在一起对不确定性和确定性之间的互动提供了有力的支持。同时，这一章的另一个目的，就是对语言测试中存在的问题作出一些评论。

第七、八、九章进一步讨论了语言测试在各种不同的评价体系中的应用问题。第七章集中评价了不同方式和内容的测试的方法论问题，并提出评价体系可以集中在学习、成绩测试和学习方法上的问题。接着简要介绍了语言测试研究的六种类型。有些类型未作介绍，将更加详尽地研究。这些类型涉及从使用测试资料的研究到其他应用语言学领域的测试研究等极其宽阔的领域。第八、九章运用第七章提出的方法论，通过举例，对各种形式的测试评价指标进行更为深入的研究。语言测试的运行借助于不确定性和确定性之间形成的平衡，它主要在应用语言学研究中发挥作用。这一观点通过对例证的研究进一步得到强调和证实。