英语测试

ENGLISH LANGUAGE TESTING

〔美〕C·里德教授 讲授
李庭芗 李守京 编译
李守明 刘来牛

人民教育出版社

英 语 测 试

[美] C·里德教授讲授

李庭芗 李守京
李守明 刘来牛 编译

人民教育出版社

1986·北京

英 语 测 试

[美]C·里德教授讲授

李庭芗　李守京
李守明　刘来牛　编译

*

人民教育出版社出版
新华书店北京发行所发行
北京房山印刷厂印装

*

开本 787×1092 1/32　印张3.875　字数78,000
1986 年 12月第 1 版　　1987年11 月第 1 次印刷
印数 00,001— 13,000
书号 7012·01183　　定价0.53元

译者的话

1982年9月到1983年7月，美国威斯康星州立大学麦迪逊分校英语系教授查理士·里德 (Charles Read) 博士来北京师范大学外语系任教。理德教授为外语系英语教学法研究生和本科高年级学生开设了应用语言学和普通语言学等课。《英语测试》这本书是根据里德教授讲授应用语言学中的英语测试部分的讲课录音稿整理、翻译而成的。

理德教授曾在北京师范大学附属中学听英语课，并同附中英语教师进行了座谈，还分析了附中学生英语期中考试的试卷。针对我国中学生英语考试的实际情况和需要，他深入浅出地讲述了英语测试的一些基本问题，诸如测试的种类、对试题的要求、中学英语试题的剖析、测试成绩的评定等等。另外，里德教授还介绍了两篇关于英语测试的资料，一篇列举了教师在英语测试中应避免的常见错误，一篇介绍了美国以英语为第二语言的测试出版机构，本书也作为附录附于书后。

测试和测试分析是英语教学的重要组成部分。正确而有效地利用各种形式的测试，科学地剖析测试的结果，对于评价教学效果，从而对教学工作进行必要的调整和改进都是有益的。里德教授在讲课中所用的材料是我国学生的英语试卷，有针对性地论述了英语测试的有关问题，因之，我国英语教师读起来会感到亲切而有所启发。

里德教授讲课的英语录音由李守京和李守明整理并译成汉语，附录一由刘来牛译成汉语。全部译文由李庭芗最后审阅并作了适当的订正。

译稿又承夏聿德教授赐阅，并对译稿中的统计图表和说明作了一些订正，谨在此表示谢忱。

编 译 人

1986 年 1 月 20 日

目　录

一、有关语言测试的一些基本概念

大家在上中学时，每个学期都要参加英语测试。大家也都参加过影响你们一生的高校入学考试。当然，绝大多数应试的考生没有被录取，这虽然对他们并不是一件幸运的事，但他们也会把它看作是生活中的一件大事。同学们上了大学以后，每门课程都要进行测试。这些测试也都很重要。从学生的角度看，如果测试题出得好，学生的水平就能得到如实的反映。可是我们将会看到，在现实生活中，没有一次测试是尽善尽美的。任何测试都存在着这样或那样的缺欠。当然，为了能对全体考生做出公正的判断，就应该尽最大努力把题目出得更好些，以提高试题的质量。

从国家利益来看，测试在很大程度上对发展教育事业，并从中得到最大的效益，起着至关重要的作用。这对中国的教育事业是这样，对美国的教育事业也是这样。测试的结果对研究工作也具有重要意义。我们看到，当前在世界各国，特别是在中国，将会有越来越多的人从事语言学习和语言测试的研究工作。中国现在非常重视发展教育，而学习外语对于教育的发展在一定程度上有其特殊的意义。只有把试题出好，对测试结果的研究才能卓有成效。反之，如果试题出得不好，测

试的结果不仅失去了意义，甚至会造成一些假象和误解。总而言之，无论从国家的利益出发，还是从个人增长知识的角度看问题，都必须尽量保证测试题的高质量。遗憾的是，多数试题，特别是各级学校自己出的试题，往往不很理想。这次讲座将一一列举和讨论影响试题质量的一些因素。下面先从测试的一些基本概念说起。

(一) 测试的种类

1. 测试的五种基本形式

测试基本上有五种：成绩测试、水平测试、语言禀赋测试、诊断性测试和分班测试。

1) 成绩测试(Achievement Tests)

成绩测试的目的是检查学生是否学会了教材的某一部分内容。这种测试包括学期考试、中学毕业考试等等，也包括用来检查学生掌握某种句子结构等知识的阶段性小测验。无论出于以上哪种用途，成绩测试的目的都是要检查学生对所学过的知识的掌握情况。学过什么内容就考什么内容。例如，期末考试的目的就是要检查学生是否已掌握了这个学期所学的内容。成绩测试通常由任课教师命题。

2) 水平测试(Proficiency Tests)

水平测试是根据一定的要求，检查学生的语言知识或综合运用语言的能力是否合格。例如，每个申请去美国留学的学生都要参加"英语作为外语的测试"，又称"托复(TOEFL)"。这种测试在世界范围内进行。它是由美国的一家测试机构主

办的。大多数招收外国留学生的美国大学都要求外国留学生参加“托复”测试，并以测试成绩作为录取的先决条件。“托复”不同于成绩测试。在美国出“托复”试题的人无从知道各国学生学习过哪些内容，他们也不考虑学生用过何种教材。“托复”的目的是全面考核学生的英语水平及综合运用英语的能力。

3）语言禀赋测试(Aptitude Tests)

语言禀赋测试是近年来在美国流行的一种测试，它引起了不少争议。这种测试的目的是判断学生学习语言的禀赋或潜在能力，而并不计较学生已经学会了多少东西。它的依据是，有些人比别人更善于学习语言。在一些学校中，当报考人数多于所要录取的人数时，就进行语言禀赋测试，看哪些学生学好这种语言有更大的可能性。考生在参加语言禀赋测试前，可能还不曾学过这种语言。换句话说，如果语言系不可能录取全部考生，就可以通过语言禀赋测试选拔其中的佼佼者。人们相信这些学生有可能获得较好的学习成绩。语言禀赋测试的目的在于预测学生学习语言的潜力，但它实际上是否真能取得这样的效果，目前社会上还是有争议的。

4）诊断性测试(Diagnostic Tests)

诊断性测试与成绩测试相仿，是针对某一部分具体的教学内容或语言知识进行的，目的是判断学生学习某一句型或某一语法结构是否有困难。诊断性测试常以两种语言的对比分析为基础。例如，比较汉语和英语的异同，我们发现英语中的定冠词“the”在汉语里是没有的。可以进行一次小测验，判断定冠词是否真是学生学习英语的一大障碍，以及哪些学生用定冠词有困难。这样一来，教师教学时就能有针对性地讲授，

以期达到最佳效果。而且还可以专门给有困难的学生多做些定冠词的练习，而已经掌握了这种用法的学生就可以少做练习。诊断性测试又以错误分析为基础。可以先分析学生作业中出现的错误类型，找出学生最感困难的句子结构，然后对全体学生进行诊断性测试，找出困难最大的学生，再为他们专门补习。诊断性测试的目的是了解学生在某一方面的学习困难，以便针对问题进行教学。

5）分班测试(Placement Tests)

分班测试的目的在于妥善地将学生按程度分班或编组，以利教学。例如，新生过去学过的课本不是一种，程度也千差万别，我们需要知道他们分别属于初级班、中级班或高级班。这就要进行一次分班测试。我们威斯康星大学有很多外国学生，我们为他们开设了多种课程。可是，并不是所有的学生的英语都已经过关，达到了无须额外补习就能顺利地听懂专业课的水平。针对这种情况，我们首先进行一次水平测试，然后再对英语水平不够高，需要额外补习的学生进行一次分班测试，以便确定他们各自需要哪方面的帮助。我们为这些学生开设了写作课、阅读课、语音课等等。在分班测试的基础上向学生宣布，哪些学生得在听力和理解能力上下功夫，才能听懂大学的课程，哪些学生的阅读能力尚待提高等等。分班测试是为了找出新生需要哪些帮助，以确定他们编在哪个班里学习为宜。

以上提到的五种测试形式并不总是孤立存在的。例如，根据一次水平测试的结果，可以确定某学生是否需要进一步提高英语水平，甚至还可判断出他最薄弱的环节是什么。这就是把水平测试与分班测试结合在一起了。学校的期中考试

是成绩测试，我们可以从中发现是否每个学生都学会了指定的课程。如果有意识地在期中考试中加进一些特殊的句型或语法结构，我们就可以查出哪些学生学习这部分内容有困难，从而根据每个学生对各个题目解答的情况，分别为他们准备一些有针对性的辅导课。这样，期中考试就有了诊断性测试的性质。有时很难断言某次测试属于哪一类型，也并不是所有的测试都能明确地贴上"水平测试"、"成绩测试"等标签。有的测试就叫做"英语作为第二语言的测试"，使人很难确定它究竟应归为哪一类。

总之，为了达到不同的目的，就要进行不同形式的测试。测试之前，教师应认真考虑要进行的是哪一种测试，这有助于确定考什么内容，以及怎样充分发挥测试的作用。

2. 常模参照测试和标准参照测试 (Norm-referenced Tests, Criterion-referenced Tests)

常模参照测试的目的是确定学生在掌握英语知识和运用英语能力上的差异，因此必须使学生的测试分数拉开，然后参照每个学生的成绩把他们区分出好、中、差诸等。"托复"就是常模参照测试的一例。每个人的"托复"成绩都要同别人的成绩进行比较，分数最高的人就是测试的优胜者。

标准参照测试则与常模参照测试不同。在标准参照测试中，所有学生的成绩都仅仅同应该达到的标准比较，看他们是否达到了既定的标准，或掌握了规定的技能，而不去比较学生之间的成绩差异。假设我现在进行一次测试，要了解大家是否领会了测试的可靠性和有效性这两个概念，这就是一次标准参照测试。我只想知道你们大家是否都能正确地答出这两

个概念，而无意将在座各位的成绩加以比较；只要把问题答对了就可以，而不计较你们之间的差异。由此可见，在这种标准参照测试中，如果每个学生都得了满分，那是十分可喜的。这证明所有的学生都学会了这部分内容。与此相反，如果在常模参照测试中人人都得一百分，那它就是一次失败的测试，因为它没有达到预期的目的。很可能由于试题太容易，没法区分出学生之间的差异。常模参照测试的理论是，确认人们之间存在着差异，测试的目的就是要显示这些差异，找出哪些学生学习得好，哪些学生学习中等或较差。假设我们被指定为中国的高等院校招生考试出英语试题，结果考生们都得了一百分，那么我们的工作就失败了。我们的试题应该能使考生的成绩分出优劣，使英语学得最好的学生得高分，因而最有希望被录取。标准参照测试却不在乎考生的差别，它要解决的是另一个问题，即：谁学会了某一课书，掌握了某一概念或言语现象，以及谁没有学会。

在教师计划测试时，弄清常模参照测试与标准参照测试的区别是很有必要的。甚至应试的考生弄清这一点也是有益的。如果是标准参照测试，就只检查学生是否掌握了某些知识；如果是常模参照测试，那就要比较学生成绩的优劣，那么，每个考生就都应尽力发挥出自己的最高水平，跟别人竞争。对这样两种不同类型的测试要区别对待。一般说来，水平测试是全面考核学生综合运用英语的能力，如听、说、读、写的能力，属于常模参照测试；成绩测试和诊断性测试一般属于标准参照测试。

3. **分列式测试和综合性测试** (Discrete-point Tests,

Integrative Tests)

近来，语言测试又出现了另一个重要的分类形式：分列式测试和综合性测试。分列式测试中的各个项目分别检查学生的一种技能或一类知识，如，某种句子结构或某个单词。中国的高等院校入学考试就属于分列式测试，包括要求学生用适当的时态填空、填冠词、选择适当的单词填空等。综合性测试则要全面检查学生综合运用语言的能力，例如，写作课的结业测试可以是一篇两小时的作文，检查学生全面运用语言的能力，要他们把语言运用于实际。写作时可以用过去时，也可以用现在时，在这方面学生不受限制，也不专门考某个句型。在美国，当国务院或外交部招聘驻外国使馆雇员时，他们往往挑选懂得派驻国语言的人。在对候选人进行测试时，不采用“托复”测试的形式，而是通过“外语面试”（Foreign Service Interview）选拔录用人员。凡是想到美国驻中国大使馆工作的人都必须通过面试。其内容只是让应试人坐下来同主试人用汉语交谈，就象美国大使馆的工作人员可能会同中国人进行的交谈那样。这就属于综合性测试。它所考核的是在实际生活的场合中，全面、自然而又有实际意义地运用语言进行交谈的能力。这正是雇员们将来在使馆工作时所必须具备的能力。

围绕分列式测试和综合性测试展开了一场争论。传统的课堂测试是分列式测试，“托复”也是分列式测试。但近来出现了这样一种议论：“这种测试方式不妥当。我们所要检查的应该是学生运用语言的实际能力。”大家知道，交际能力是指能在具体的场合中恰如其分地使用语言。有的人学了很多

语法知识，但一到具体场合，却往往张口结舌，这就是由于交际能力差的缘故。只有通过综合性测试，例如一次面试，才能考查出学生是否真正会运用语言。分列式测试常有这样的项目："把下列句子改写为被动语态。"大家都做过这样的试题吧。赞成综合性测试的人会说："谁管你会不会变被动语态？在实际生活中永远不会有人向你提出这样的要求，但却要求你能用英语交谈、写信、填写申请表，或者读懂一段英文说明书。这些才真正是要求你做的事。"具体来说，假设我们的工作是在中国民用航空总局教授机场技师学英语。民航有关于如何保养与维修飞机的小册子，其中有些可能是用英语写的。因此，技师们要正确地维修飞机，就得能够读懂这类说明书。那么，我们应该怎样考核这些技师的英语水平呢？是让他们变主动语态为被动语态，还是让他们按照英语说明书正确地拆卸发动机呢？显而易见，问题的关键不在于他们会不会把主动语态变为被动语态，而在于能不能够读懂说明书，并且按照说明去工作，哪怕他们需要借助字典查生词也是无妨的。因此，赞成综合性测试的人认为："我们需要的测试方式应该是让学生看一段英文说明书，然后照着操作，或是就说明书的内容回答一些问题，来证明他确实读懂了。"但赞成分列式测试的人反驳说："进行这样的综合性测试，教师无法限定学生必须使用某些特定的句子结构，因而也就无从了解学生会用什么，不会用什么。"不难看出，这种观点是有一定道理的。大家都知道在综合性测试中使用语言有迴避现象。当学生遇到他不认识的词，或者用起来有困难的句型时，就会兜个圈子绕过去，避免使用这些词语，而代之以较有把握的词语。学习语言

的人都知道这个窍门。在综合性测试中，教师无法限定必须使用的词句，因而只要学生感到某种句型用起来把握不大，那么，在谈话或者作文时他肯定不会用它，这是毫无疑问的。结果是，教师就无法发现他的这个困难。所以，要想诊断出学生的薄弱环节，用分列式测试可能更合适些。针对这一论点，赞成综合性测试的人再次反驳说："你们说得对，但是完全有可能精心设计一篇综合性测试题，使学生不得不用某些词语和结构。"

这场争论的另一个焦点是评分问题。赞成分列式测试的人认为分列式测试评分准确。象"托复"及高校入学考试这样的分列式测试，能很准确地算出每个学生应得的分数；但评阅作文就不然了。主试人的兴趣和爱好会在一定程度上影响评分的高低。这样一来，由于渗进了主观因素，评出的分数中能有多大比重是可靠的呢？这是评阅作文经常面临的问题。教学中常有这样的情况，把同一篇作文先后交给两位教师评阅，得到的往往是两个悬殊很大的分数。一位教师给这篇作文评"A"，另一位教师却给它评"C"。我们的目的是要考查学生的交际能力，综合性测试又被公认为能反映学生的水平，但是结果却令人失望，因为同一篇作文得到了两个不同的分数。从作文的主题到教师对其观点是否同意等等多种因素，都会对教师的评分产生影响。有人做过这方面的调查，证明与教师持相同观点的作文往往得高分。而在分列式测试中就不会出现这类问题。分列式测试的评分是很客观的，并且由于通常只有一个正确答案，所以评分也是机械的，不可能渗入主观因素。从这一点看，分列式测试更为公正。

现在大家了解了争论双方的意见。一方强调试题要能反映学生运用英语的真实水平，另一方坚持说评分要客观，双方各执己见，也各有道理。在实际应用时，要根据任务和具体情况作出选择，不可一概而论。

（二）测试的可靠性和有效性

这一节谈谈测试的两个重要原则问题：可靠性（或称信度 Reliability）和有效性（或称效度 Validity）。任何形式的测试都要面临这两个现实问题。事实上，无论用什么尺度衡量任何事物，都离不开这两项原则。先从可靠性谈起。

1. 测试的可靠性

测试的可靠性指测试结果的稳定性。假设我们考同一个学生五十道题，第一次测试，他答对了四十题；第二天再考，只答对了十题；一个星期后，同一篇试题又考了第三次，他答对了二十题。这样的测试就是不可靠的。测试的目的是要衡量某种相对稳定的东西，例如学生的英语知识，因而测试的结果不应该在一、两天内出现戏剧性的变化。如果几次重复测试的结果是稳定的，或基本上一致，这种稳定性就是测试的可靠性。

现在我们谈谈上文中提到过的作文和面试的评分问题。如果我要求学生写一篇作文来确定他写作课的结业成绩。我给其中一篇作文“B+”。为了检验一下自己的评分，我把这篇作文拿给另一位教师看。我先不告诉他我的评分，只向他说明我对学生的要求。我本以为那位教师对这篇作文的评价