DUI WAI HANYU JIAOXUE DE
YUYAN CESHI

对外汉语教学的语言测试

张林林 编著

广东高等教育出版社

DUI WAI HANYU JIAOXUE DE
YUYAN CESHI

对外汉语教学的语言测试

张林林 编著

广东高等教育出版社
广州

图书在版编目（CIP）数据

对外汉语教学的语言测试/张林林编著．—广州：广东高等教育出版社，2013.6

ISBN 978-7-5361-4873-4

Ⅰ.①对… Ⅱ.①张… Ⅲ.①汉语-对外汉语教学-水平考试-教学参考资料 Ⅳ.①H195

中国版本图书馆 CIP 数据核字（2013）第 074886 号

广东高等教育出版社出版发行

（地址：广州市天河区林和西横路）

邮政编码：510500　　电话：（020）87557232

广州市穗彩彩印厂印刷

890 毫米×1 240 毫米　32 开本　8.875 印张　230 千字

2013 年 6 月第 1 版　2013 年 6 月第 1 次印刷

印数：1～2 000 册

定价：20.00 元

目　录

第一章　中国传统的考试制度

第一节　中国传统考试的发生和发展

测试，通俗地说就是我们平常所说的考试。我国是考试的故乡。孙中山先生在《五权宪法》中写道："现在各国的考试制度，差不多都是学英国的。穷流溯源，英国的考试制度，原来还是从我们中国学过去的。"19 世纪时，有一位美国人在谈到考试制度时说过这样的话："中国现在的政治有一点使我们很感兴趣……就是文官必先经过考试及格取得学问上的资格，而后始能任职。在纠正恶习这一点上，中国人是走在我们前面了。同样，中国社会上都非常重视教育，也走在我们的前面。"①

运用精确的测量手段测试学习能力和学习成果，这不是我们的首创，但提出这种形式并把它制度化，这应该是我们中国人的发明。

现在，人们一提到科举考试总是首先想到隋唐时代的科举考试，其实，作为一种考核、选拔的手段，作为完整的教育中的一个环节，考试在西周时期就已经存在。

据《史记·五帝本纪》记载，尧考察继承人舜，把自己的两个女儿许配给舜，"以观其内"，指派九个男子与舜共事，"以

① 邓嗣禹：《中国考试制度史》，431 页，台北：台湾学生书局，1982。

观其外”，并且以五典考察其德，以百官考察其能。[①] 原始社会后期的部落首领是由民主选举产生的。因此，我们完全有理由说，考核、考试的产生是顺应了一种社会需要。

考核、考试产生的另一种社会需要是学校教育。我国西周时期就有比较完备的学校教育制度，对入学年龄、各阶段的学习内容以及相应的考核标准都有明确的规定。贵族儿童6～9岁在家学习简单的数字，学习方位、处所、顺序概念的表达；10岁入学，在学校寄宿，学习书写、记数、音乐、舞蹈等；从13岁开始，学习礼、乐、射、御；[②] 20岁行冠礼，开始学礼制、仪节和行为道德规范等。当时不仅有分门别类的学习内容，而且还有明确的考核要求。如射、御作为两种基本的军事技能训练，都分别有白矢、参连、剡注、襄尺、井仪以及鸣和鸾、逐水曲、过君表、舞交衢、逐禽左等具体考核指标。[③] 西周时期的考试无论是在考试功能上还是在考试类型上，都对后来的考试制度产生了深

---

① 《史记·五帝本纪》“乃以二女妻舜以观其内，使九男与处以观其外。”“试舜五典百官，皆治。”五典，又称五教，即父义、母慈、兄友、弟恭、子孝；百官，即众官职官位；试百官，意即考察其在各个岗位上的能力。

② 礼，是政治伦理，包括奴隶社会的宗法等级世袭制度、道德规范和礼节。乐，是综合艺术，主要指六代乐舞，包括黄帝时代的“云门”、尧时代的“大章”、舜时代的“大韶”、夏时代的“大夏”、商时代的“大濩，同音字”、西周时代的“大武”。“云门”、“大章”、“大韶”和“大夏”是文舞，“大濩，同音字”和“大武”是武舞。射，即射箭，御，指驾御战车，二者都是军事技能。

③ 白矢等五项是射箭技艺的考核指标。白矢，考核射箭者的臂力，要求箭穿透靶子；参连，考核射箭者的连发速度，要求射箭者能连发四箭；剡注，考核所射箭头是否锋利，箭靶上的箭头是否朝下；襄尺，考核射箭者能否礼让，君臣同射，为臣者应该后退一尺；井仪，考核射箭者箭法是否精准，所射四箭能否排成“井”字状。鸣和鸾等五项是驾御战车技艺的考核指标。鸣和鸾，要求御者能使和与鸾这两种装饰铃有节奏地共鸣；逐水曲，要求御者沿着曲折的河道奔驰而不颠簸；过君表，要求御者进辕门时不碰到石碏；舞交衢，要求御者能在十字路口轻盈地来往穿梭；逐禽左，要求御者驾车追赶禽兽时能使禽兽都往左边跑，以便君主射杀。周礼规定，君主田猎，自左方射之。

远的影响，打下了深深的烙印。

秦代的官学叫学室，在学校教育方面有两点比较突出：一是明习法令；二是识字写字。秦代以“以吏为师”、“以法为教”的“吏师”制度的目的在于选拔刀笔小吏，因此，所考法律、书写、诵书等都与刀笔小吏的工作息息相关。

汉代“独尊儒术”，兴办太学。汉代太学既是当时的最高学府，又是考试的最高主管机关。汉代太学考试的目的有二：一是通过考试选拔人才，充实官吏队伍；二是激发学生学习儒家经典的兴趣。

魏晋南北朝时期是学校教育走向衰落，而学校教育的考试制度，具体地说，就是太学考试制度逐步健全的时期。魏晋伊始，虽然曹操提出过“唯才是举，以备录用”的用人政策，不拘一格选任贤才，但曹丕即位后听从了吏部尚书陈群的建议，推行“九品中正官人法”。九品中正官人法也叫九品中正制，具体做法：①设置中正：郡置小中正官，州置大中正官；由司徒选择“贤有识鉴”的现任朝廷官员兼任其原籍的郡小中正或州大中正。②品第人物：中正官员负责察访与之同籍的士人，了解其家世源流，整理其德才表现材料，做出总的评价。“家世”谓之“品”，“德才”谓之“状”。中正官根据其品状定其等第。等第分为九品：上上、上中、上下、中上、中中、中下、下上、下中、下下。③按品授官：中正官将品第人士的材料定期造册上报司徒府，司徒核定上报尚书录用。一般来说，品第高者任高官，品第低者任卑职。

南北朝后期的察举制已经孕育着国家设科招考和自由报考的分科考试制度，科举考试制度已经呼之欲出了。

隋朝初年，虽然也实行过九品中正制，但隋文帝很快废除了这种制度，把选官任人的权力集中到朝廷的吏部。一般认为，隋炀帝创设进士科标志着科举考试制度的正式产生。

科举考试制度，无论是从政治上，还是从技术操作的层面上看，是古代选士制度的分水岭，也可以视为古代选士制度的一次重大改革。从政治层面上看，科举考试把录取、任用权完全由中央掌控，限制了门阀士族把持选士的局面，为庶族地主参政开辟了道路，扩大巩固了统治阶级的基础；从技术操作层面上看，科举考试把以察举为主改为以考试为主，所谓声名德望不再是主要依据，使轻门第、重才学、任人唯贤的选士工作有了相对客观的标准。

唐代的科举考试制度日趋完善，常设的科目有进士、秀才等几十种之多，还增设了不少考试科目。唐代的考试方法也有所发展，常用的方法有帖经、墨义、策问、诗赋和口试等。唐代的考试在程序上也有变化，考生先得经过省试，即尚书省礼部试，礼部考试通过了算是有了“出身”，但还不能得到官职，还得参加吏部试。

相对而言，唐是盛世，经济发达，政治开明，唐代的科举进步不小，相对较为合理，与前代相比有几点明显的不同：一是将选拔士子的权力牢牢控制在中央，中央集权的措施使庶族参政的欲望得以实现，这样的措施强化了统治集团的社会基础；二是把读书、应考、做官获禄三者结合得更为紧密，使得士子认为这是读书人的不二法门；三是继续改变只重品行、门第，忽视知识才能的选士标准。它使得考试本为一项与学校教育相联系的评价手段变为了一项巩固统治、驯服臣民的政治措施。

宋代的科举考试基本上是沿袭唐制，但也有一些变化。这首先表现在考试科目的设立和考试内容上。宋代沿用了前代的许多科目，也开了如制科、词科、绘画试等新的科目。制科是皇帝亲自策问的考试，考试内容由皇帝临时确定，当时人们称之为“大科”，为众科之首。北宋的兴学和科举考试改革几起几落，都有一个共同的特征，即坚持把学校教育与科举考试制度结合起

来，坚持育才是取才的前提条件这样一个观念。

元代的科举考试一开始就颇费周折。元朝统治者把人分为四等，科举考试的有关规定与考生所属等级有联系。如蒙古人属于第一等，色目人属于第二等，蒙古人和色目人参加科举考试只需考两场。汉人属于第三等，南人[①]属于第四等，汉人和南人参加科举考试必须考三场。蒙古人、色目人参加汉人、南人的考试，虽也考三场，若被录取，所授官级可比汉人高一等级。可见，元代的科举考试充满着民族歧视。

明清大抵实行荐举和科举两种方式，但实际上执行的是以科举为主的选拔人才的制度。明清时期的科举考试分为四个步骤：第一步是“童试”，由州、县长官主考，通过者为“生员”，俗称秀才。真正意义上的科举考试是从第二步乡试开始的。乡试是省一级考试，每三年举行一次。第三步是会试，会试是中央级的考试，乡试后的第二年二月在京城举行。皇帝钦点考官，由礼部主持。第四步是廷试，也叫殿试，殿试不是选拔考试，由皇帝亲自主持。殿试只考策问，考生必须当场作答。殿试考中称甲榜、甲科。

中国传统的科举考试作为一种制度，从隋唐算起至结束1 300 多年。鸦片战争之后，一些有先进思想的知识分子从中国处处落后、动辄挨打的现状中，看到了中国缺乏经世致用的人才，看出了科举考试取士的弊端，呼吁开办新式学堂，废除科举考试制度，提出了改革的主张。

清末科举考试制度的改革是分三步走的。第一步是改革科举考试的内容。康有为等人在《公车上书》中第一次向光绪皇帝提出废除八股文的请求，戊戌变法时清朝采纳这些意见，下诏废除八股文取士制度，规定童试、乡试、会试一律改试策论。戊戌

① 南人，指长江以南的汉人。

变法后，八股文一度复活，1901 年清朝廷第二次明令废除八股取士，改试策论。第二步，压缩科举取士的名额。光绪二十七年（1901）起张之洞等人先后提出递减科举取士名额，以学堂生员补充的建议。第三步，直接提出废除科举考试制度。科举考试不废除，对学生有很大影响。袁世凯、赵尔巽、张之洞等人奏请停止科举，兴办新式学堂。他们认为“科举不停，学校不广，士心既莫能坚定，民智复无由大开，求其进化日新也难矣”。迫于形势，清朝于光绪三十一年（1905）宣布停止科举考试，宣告了中国古代考试制度的终结。

从我国考试制度的发生、发展来看，往往是由一个良好的愿望发端，在实施的过程中，或是由于制度本身的先天不足，或是由于执行者的居心叵测，结果总不外乎是异化严重，流弊频出。西汉确定的太学考试制度和察举制度、魏晋南北朝时期的九品中正制，这些对政治和文化教育事业都起到过积极的作用，为古代考试制度积累了丰富的经验。但察举权多操纵于诸侯王、公卿之手，推荐也只重声名不重才行，所察举之人未必是真正的人才。

隋唐创立的分科考试取士的科举考试制度，一开始也起过进步作用，较好地解决了中央集权和调动地方、个人积极性的矛盾。但是，科举制度将读书、应考、做官三件有联系的事情以唯一的目的关系把它们串联起来，导致了科举考试控制教育，学校变成了科举考试的培训机构。

科举考试制度逐步建立起来的一套从内容到形式的范式，最初可以起到强化考试客观性和标准化的作用，但最终出现的是毫无生机的八股文和试帖诗。尤其是考试内容不能与时俱进，出现了重文轻理，所培养的人才可以坐而论道，但不能经世致用。我们不能简单地把中国近代社会与工业文明、科学技术、社会经济生活格格不入，全部归咎于科举考试制度，但那种局面的形成与长期实行科举考试不无关系。

第二节　语言测试与传统考试的区别

语言测试与传统的考试虽然有着千丝万缕的联系，但作为一种评价方式，语言测试与传统考试有以下几个方面的不同：

第一，从评价性质上看，传统考试是一种主观评价，即便是明清时期发展比较成熟的八股文、试帖诗在客观性和标准化方面已见端倪，但从根本上看，还是主观的。评判官认为好，皇帝喜欢，那就是好。

现代语言测试是运用数学手段的科学测量。运用数学手段是要走出经验科学，通过量化，使其走进经典科学的殿堂。科学上的量化有两个好处：一是使测试趋于稳定，二是用量化的形式使结果更准确。手段、工具的科学保证了测试的客观性。

第二，从目的来看，传统考试，尤其是古代的科举考试，其目的是单一的，就是取士。考试所呈现出来的所有信息仅仅把它作为选拔的依据之一。

现代语言测试的目的是多元的。我们可以把通过测试得到的信息作为选拔的某种依据，可以作为评价学习者能力的依据，也可以作为判断学习者对某种语言知识掌握程度的依据，还可以作为调整教学内容、教学目标、教学手段等的参考系数。简而言之，现代语言测试得到的信息可以用于与教学相关的研究、实验、反馈、诊断以及选拔等许多方面。

第三，从对这两种评价方式自身的价值判断上看，传统考试由于目的单一化，从某种意义上说，是为考试而考试，把评价视为一种孤立现象。

现代语言测试把它作为与学习相关的整个系统工程中的一个环节。如水平测试可以作为评价教学目标是否达到的依据，可以作为语言能力等级判定的依据。传统把它作为一种目的，现代语

言测试则把它看成是一种手段，一个与目的相关的环节。

第四，从所使用的题型来看，传统考试由于是一种主观评价，因此所使用的题目也大多是主观题，如科举考试的八股文、试帖诗以及后来的作文、古文翻译，开放性非常强。现代语言测试使用了大量的单项选择、多项选择等客观性题目。

第二章　语言测试的基本概念和基本原则

第一节　语言测试的基本概念

语言测试是语言教学的必要环节，也是对语言教学效果的一种评估手段。教授语言的教师，无论是教授母语的，还是教授第二语言的，作为教学过程的后续环节，在完成计划的教学任务后，往往要进行单元测验，或期中考试，或期末考试。这些测验或考试都是语言测试。

语文教学和研究、语言教学和研究，包括第二语言的教学和研究都与语言测试有着密切的联系，但是，语言测试与语文教学和研究、语言教学和研究并不是“孪生姐妹”。人们对语言测试的认识有一个逐步深化的过程，对这个概念以及这个概念所代表的方法有一个从不自觉到自觉，从被动使用到积极利用，从主观感知到科学认知的过程，这在对外汉语教学的语言测试的发展上表现得尤为充分。半个世纪来，人们在对外汉语教学中运用语言测试的实践中，使语言测试作为一种必要的教学环节在逐步发展和不断成熟起来，对语言测试价值的认识也在逐步加深，它的作用在逐步显现出来。

在语言测试学科中，测量（measurement）、测试（testing）、评价（evaluation），这几个概念经常用到，它们有相同之处，都是指对对象的评说，但又有所区别。

测量是用量化的方法描写事物本身具有的数量属性特征的过程；语言测量的目的是获得量化信息，通过量化信息来判断被测试者与语言相关的某一属性程度的手段或过程。

测量是一种观察事物数量属性特征的手段和方法，但在具体使用时有所不同。有些事物其数量特征是显性的，我们可以直接观测到，比如 A4 纸的数量特征是 210 mm × 297 mm，我们只要用标准的量具就可以测定。对这些具有显性数量特征的事物进行测量属于客观测量。而有些事物的数量特征是隐性的，一般不能一眼就看出。如果我们要了解其数量方面的属性特征，就得通过对与其数量特征密切相关的现象、表征的考察来推论其数量上的属性特征。我们之所以能够通过相关的现象或表征来揭示事物的某些属性特征，是因为事物的内在属性不是孤立的，它们总是相互联系，相互依存，相互制约的，人们语言能力方面的数量属性特征就属于这种类型。我们可以通过人们在交际过程中的具体表现，用一种量化的方法把人们在使用语言过程中能力上的程度差异，用量化的方式把它描写出来。通过不同的具体表现推断他们在语言运用上的程度差异，这样的测量相对前者而言多少带有一定的主观性。

测量作为描写事物数量属性特征的一种方法包含着三个要素：

首先，被测量的事物是可以通过特定的手段观察的，即便是隐性的，也可以通过相关的现象推论出来。人们的语言能力虽然是在动态交际中才呈现出来，但语言能力有程度上的差异是不可置疑的。这种差异会在言语交际中不自觉地流露出来。假如某一事物是不可观察的，那么，测量也将无法进行。

其次，被测量事物的属性特征可以从量化的角度进行描写，可以将它的属性特征转化为量化信息。这里所说的可以进行量化描写包含着两个层面的含义：一是指事物的属性特征本身存在着

转化为量化信息的可能性。比如某受试的测试作文语句通顺规范，层次清晰，老师根据评分标准给了 90 分；另一受试的测试作文频频发生语病，或是搭配不当，或是不合汉语的用语习惯，老师按照评分标准给了 65 分。90 分与 65 分这两个数据是上述两位受试在写作能力上的数量特征，是他们程度差异的量化形式。另一层含义是我们能够将其量化，并且对这些量化信息进行分析，解释造成这样的数量差异的原因。我们还能够对这些量化信息进行推测，对这些量化信息进行前瞻性的解读。

再次，测量的客观性。客观性是测量的“立身之本”，这与前面提到的客观测量和带有主观性的测量不是同一概念。无论是前面所说的客观测量，还是带有主观性的测量都必须具有客观性。测量的客观性是针对测量的方法和规则而言的，是指测量过程中所采用的方法、所依据的规则，都必须摒除个人的主观价值判断，或者说所有的价值判断都必须统一到一个标准上来，对所有的受试用同一个标准。客观性是测量的基础，是测量结论可靠的保证，它直接影响到结论的可靠性。客观性越高，其结论的可靠性就越高，反之就越低，偏差就越大。

客观性的高低来源于标准的统一和对标准的认可。统一标准就是统一尺度，只有在统一尺度下才有可能得到同一的结论。对语言能力的判断从本质上说是一种个人的价值判断，绝对的一致是没有的。比如对某受试口语能力的判断，甲老师听后把他定在某等级上；乙老师可能把他定在另一等级上，等级上的差异可能是细微的，也可能是巨大的。这是很可能发生的事，因为甲老师可能是从口语表达的流畅性角度加以考察，而乙老师则从语音的标准角度加以评判，角度的不同势必造成结果的误差。对标准的认可程度上的差异也会造成评判结果的误差。同样是从表达流畅性的角度加以评判，对标准掌握的宽松严紧也会造成结果的不同。从这一点来看，测量的标准尺度不仅仅是一个制定标准的问

题，同时也是一个执行标准的问题，二者不可偏废。

测试是对行为样本所做的客观的标准化的测量。测试作为一种对事物属性特征进行量化描写的方法或过程，应该具备以下三个要件：

首先是行为样本的典型性。测试作为一种描写事物属性特征的方法或过程，它不需要，也不可能对设计范围内的所有个体进行穷尽式的描写，它只是选取一些个体进行描写。所选取的个体必须具有典型性，个体所具有的特点就是这一类事物的特点。样本的典型性对测试的有效性有着直接的影响。

语言作为交际工具，它直接做功于言语交际活动，存在于言语交际活动之中，人们语言的不同能力体现在不同交际场合、谈及不同话题、对不同交际对象、运用不同语言形式、实现不同交际目的的言语交际活动中。简而言之，不同的交际因素会运用到不同的语言能力，所以，不同的言语交际样本会反映不同的语言能力。因此，我们所进行的语言测试必须有着明确目的，所抽取的样本应该与我们的测试目的相吻合，应该能充分反映出语言使用者特定的语言能力，这也就是说样本应该具有典型性。

样本的典型性是一个看似简单、操作起来却十分复杂的问题。这种复杂性表现在以下几个方面：一是言语交际是一种有多种因素参与的综合性活动，是诸多因素共同作用的结果，我们有时会一时难以判定哪些因素包含在其中；二是哪些因素在起着主要作用，哪些因素起着次要作用，我们也无法判定。从这个意义上说，选择行为样本的过程就是一个甄别、确定典型性的过程。

其次是测试的客观性。这里所说的测试的客观性总的来说是指测试要与客观实际相符合，具体有以下含义：一是测试的项目应该与受试的实际水平相符合，难度应该定在一个符合受试实际的水平上，不能脱离实际地过高或过低，只有把难度确定在一个符合受试的水平上，才能把受试实际水平的差异展现出来；二是