首届汉语考试国际学术讨论会论文选

《首届汉语考试国际学术讨论会论文选》编委会

北京语言学院出版社

首届汉语考试国际学术讨论会论文选

《首届汉语考试国际学术讨论会论文选》编委会

北京语言学院出版社

(京)新登字157号

图书在版编目(CIP)数据

首届汉语考试国际学术讨论会论文选/《首届汉语考试国际学术讨论会论文选》编委会编. —北京:北京语言学院出版社,1994

ISBN 7—5619—0432—0

Ⅰ. 首…

Ⅱ. 首…

Ⅲ. 汉语—考试—留学生教育—国际会议

Ⅳ. H195—53

出版发行:北京语言学院出版社

(北京海淀区学院路15号 邮政编码:100083)

印　　刷:北京语言学院出版社印刷厂印刷

版　　次:1995年4月第1版　1995年4月第1次印刷

开　　本:787×1092毫米　1/16　印张:22.75

字　　数:560千字　印数:1—1600册

定　　价:30.00元

出 版 说 明

首届汉语考试国际学术讨论会于1992年8月14日至18日在北京语言学院举行。来自澳大利亚、德国、菲律宾、韩国、加拿大、美国、日本、新加坡、意大利、中国和香港等11个国家和地区的98位专家和学者出席了讨论会。大会收到海内外人士提交的学术论文80余篇，经组织有关专家评选，确定61篇为参加会议的论文。这次讨论会论文内容重点突出，主要集中在四个方面：(1) 中国汉语水平考试（HSK）研究；(2) HSK用户调研报告；(3) 汉语水平考试与对外汉语教学；(4) HSK考试与大纲评估、其他汉语考试研究与介绍。

本届讨论会由中国国家对外汉语教学领导小组办公室和北京语言学院联合主办。讨论会结束后，成立了《首届汉语考试国际学术讨论会论文选》编辑委员会。编委会经过认真审编，选定51篇作为论文集的入选论文。在编辑过程中，对一些原稿在文字上或技术上做了必要的处理。

对于篇目的选定和稿件的处理，难免存在疏漏和舛误，敬请批评、指正。

编者

1994年2月

目　录

认知与语言测试 …… 桂诗春（1）
有关较高层次汉语考试的一些问题 …… 张清常（8）
汉语水平考试与语言应用能力 …… 龚千炎（12）
论汉语能力和汉语考试 …… 王德春（18）

汉语水平考试（HSK）研究

汉语水平考试（HSK）的理论基础探讨 …… 刘英林（26）
汉语水平考试的等值问题 …… 郭树军（36）
汉语水平考试的分数体系 …… 谢小庆（49）
汉语水平考试结构效度初探 …… 张　凯（59）
汉语水平考试（HSK）信度、效度分析报告 …… 何　芳（68）
试谈 HSK 所考察的组词造句能力——HSK 内容效度研究之一 …… 孙清顺（75）
关于高等汉语水平考试的设计 …… 刘镰力　宋绍周　姜德梧（83）
高等汉语水平考试试测结果的统计分析和对课程设置的评估 …… 刘镰力　李　明（93）
HSK 题库建设综述 …… 李　航（100）
开发计算机辅助自适应性汉语水平考试的设想 …… 谢小庆（111）

HSK 用户调研报告

北京大学留学生 HSK 入学考试分班入系分析——兼论分班入系 …… 杨德峰（117）
北京语言学院 HSK 入学测试编班分析 …… 贾永芬　方　玲（125）
HSK 对对外汉语教学的重要指导作用 …… 徐甲申（132）
建造以 HSK 为参照的教学链 …… 耿二岭（执笔）（139）
HSK 的效应与对外汉语教学 …… 吴勇毅　徐子亮（147）
汉语水平考试成绩报告及相关的几个问题 …… 吕效东（161）
南京大学留学生 HSK 考试情况分析 …… 陈雅静（167）
对“R/F”水平测试的反应、分析与建议 …… 许和平（176）
关于汉语水平考试的几点建议 …… 陈满华（179）
外国学生 HSK 结果报告及分析 …… 王　虹（188）
延边朝鲜族中小学学生汉语水平考试（HSK）成绩分析 …… 刘明章　姜永德（196）
维吾尔、哈萨克等族学员汉语水平考试刍议 …… 赵学会（204）
新疆少数民族学生使用《HSK 听力试题》调查报告 …… 孙　岿　邬婉荣（210）
汉语水平考试（初级与中级）1991 年 6 月 15 日在新加坡首次
　举行的成绩检讨与改进建议 …… 卢绍昌（215）

汉语水平考试与对外汉语教学

汉语水平考试（HSK）与对外汉语课堂教学 …… 赵贤州（219）
语言的社会功能与对外汉语水平测试 …… 周明朗（223）
汉语水平考试与“汉语的双轨制教学” …… 于淑华（执笔）（233）
如何改进对外国人的基础汉语教学——汉语水平考试的启迪 …… 何　平（237）
HSK（高级）口语测试模式的研究与设计 …… 崔　良（243）
简议汉语口语测试 …… 任筱萌（251）
从汉语水平考试谈日本学生汉语语言技能掌握的特点 …… 彭恒利（258）
中高级汉语课程测试问题 …… 李　杨（264）
试论中级汉语的测试 …… 陈田顺（271）
一般汉语考试试题库的建设 …… 罗守坤（278）
对外汉语教学 CAI 系统 …… 冯恭己等（282）
HSK 计算机辅助系统：功能分析与开发策略 …… 郑成义（289）

HSK 考试与大纲评估、其他汉语考试研究与介绍

对外汉语教学学科建设的新进展——评《汉语水平考试研究》 …… 贾甫田（294）
对《语法等级大纲》（试行）的几点意见 …… 吕文华（300）
从“文化测试”说到“文化大纲” …… 陈光磊（306）
关于中小学语文（汉语）水平考试总体构想的若干问题 …… 王桐生（311）
说词与写词的比较以小学五年级学生为例 …… 梁荣基（317）
从汉语水平考试看韩国学生的对策 …… 许　璧（324）
日汉语专业口译考试设计 …… 井出静（329）
北美汉语水平考试浅析 …… 刘　濂（336）
方言地区的普通话测试——香港的普通话水平测试的述评 …… 王培光（342）
美国“汉语能力试验”（CPT）的特点及可改进之处 …… 吴小燕（347）
英国外交人员汉语语言津贴考试评介 …… 李　瑜　金乃逵（351）

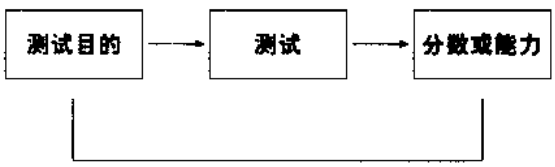
认 知 与 语 言 测 试

桂 诗 春

1. 认知与测试

迄今为止，测试的目的都是为了测量出一个考生在量表中的位置，以观察该生是否完成特定的教学要求或达到一定的专业水平。经典测试理论关心的是考生在分数量表（原始分或标准分）中的位置，而项目反映理论则关心考生在能力量表中的位置。为了保证不同的位置能充分反映出考生的不同的水平，量表必须准确、可靠、有区分性，符合分数分布的一般规律（正态分布）；而测试本身还必须有效，即考了要考的内容。这一类考试都是以取得考试成绩为目标，它感兴趣的每一个考生的行为，可以称之为教育心理测量模型（Educational Psychometric Measurement Models，EPM），表示如下图：

[图一]



这种 EPM 目前还是考试中的主流，但从认知科学的角度来看，也有其不足之处。Snow 等（1989）指出其主要问题在于：

（1）测试的项目不一定在心理上站得住：一个模型的好坏主要是看它能否很好地描述某种实验性数据，而不是看它是否符合心理的合理性。

（2）这些模型的一些假设，如项目的局部独立性、项目难度的单维性和认知心理学的实验结果不一致。在阅读理解的测试中，一篇文章后有几道题目，对每道题目的语境效应来源甚多，所以很难说题目是局部独立的，因此误差也不是全不相关的。考生一个分数反映了他的处理技能、策略、知识结构，也很难说是单维的。特别是在我国大量进行应试训练，分数更不可能是考生真正能力的反映。

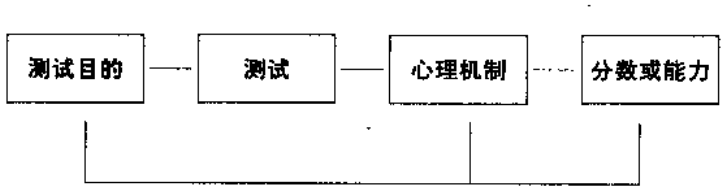
（3）这些模型把项目和分数看成是不可企及的“黑箱”，因此一个考试是否考了它所说要考的内容，即是否有效，成了它们要反复论证的问题。

认知心理学对测试提出的挑战实际上也是很多测试专家长期以来所面临的挑战：怎样对测试的内容特征作试验，怎样检验 EPM 的各种假设，怎样使这些模型能够更好地从心理学的角度去解释测试行为，怎样使关于测试的建立、评分、解释上升成为明白易懂的理论。当然认知心理学对教育也提出一个更大的挑战：对作为教育目标的学能和学业成就，以及有关的

教育测量提出更完善理论。

另外一类模型，认知信息处理模型（Cognitive Information—processing Models，CIP）正是针对 EPM 的这些问题提出来了。这些模型用以发展和检验实质性的心理理论，企图解释人类认知系统的内部机制，从而穿透我们所知之甚微的“黑箱”。所有的 CIP 都企图对输入的信息进行认知操作的具体过程和步骤作出假定。简单的模型和认知心理试验差不多，只有一、两个反映不同处理阶段的功能的参数，复杂的模型是一些用来分析更为复杂的过程的数学模型，实际上是一种计算机模拟。CIP 的提出并不是为了取代 EPM，而是为了提供更为丰富的信息，因此最近已有不少人谈到这两种模型的结合，例如，Carroll 等人主张使用以三参数模型为基础的个人特征函数来考察受到多种影响的项目难度下的测试单维性。Embretson（1985）则试图发展一种多成份的潜在倾向模型，把包括技能和知识的 CIP 模型和潜在倾向的 EPM 模型结合起来。Misley & Verhelst（1987）又把项目反应理论应用到评估那些考生使用了不同处理策略的项目。CIP 可简单地图示如下：

[图二]



从认知科学的角度看，人类的知识分为两大类：一是陈述性知识，一是程序性知识。

陈述性知识是关于事实本身的知识，以语义网络的形式保存在人的记忆里，信息的提取决定于它在网络中是怎样组织的，如果新的知识单位和已有的知识单位连接得比较好，它就能更容易被激活。EPM 感兴趣的是考生是否掌握某一些知识单位，而 CIP 感兴趣的是这些单位在网络中是怎样组织的。我们怎样知道知识在大脑中是怎样组织的呢？一个简单的办法是观察知识的提取过程，或者是看它的提取速度。提取速度快就意味着知识组织得比较合理。更进一步的考虑是观察考生提取时用了什么策略，如学科内的概念是怎样组织的？原有的知识或信念系统对新知识的习得有何影响？两者（如母语系统和外语系统）发生矛盾时，学生是怎样处理的？

程序性知识指的是怎样进行各种认知活动的知识，它在陈述性知识的基础上建立，包括陈述性知识和它的使用条件。陈述性知识是静态的，而程序性知识则是动态的：它强调的区别与概括，特别是自动化。我们通常所说的能力或技能其实就是程序性知识。Anderson（1983）认为技能自动化过程经历过三个阶段：（1）知识仍然为陈述性的，但必须使用普遍性的程序性知识进行有意识的处理。学生必须了解他们应该怎样做，只不过他们所做的尚未达到自动化的程度。（2）学生通过反馈进行练习，建立起某些产生式，就是在记忆中存储那些导致成功的条件的过程，故名程序化。由于人的工作记忆的容量有限，一些复杂的程序不可能一次就达到程序化。不正确的程序和正确的程序都同样有机会被学到，故需要有反馈。反馈可以由学生自己产生，这意味着他已经建立起一个评判自己行为是否合适的内部标准，对他们独立掌握技能就会起到促进作用；如果他们还未能建立起内部标准，就只能依赖教师或计算机从外部提供反馈。这些反馈不是随时随地都有的，所以学生就有可能学习到一些不正确的程序。经过反复的练习，那些分别学到的、连续执行的程序就有可能组合成为一个产生

式系统（如儿童要一笔一笔地写自己的名字，而成人则可以一笔签名）。(3) 程序的应用范围可以扩展，如开汽车可以从走公路发展到走山路；也可以专门化，如开跑车比赛。技能之所以能够达到自动化，主要是产生式可以归并和组合，成为产生式系统，而且这些系统的应用范围可以扩展或专门化。

程序性知识与陈述性知识的存储、提取和使用都是有组织的，图式就是它们的组织方式，可以说是知识的高级结构。Johnson—Laird (1983) 把这种高级结构称为心理模型 (mental model)，这是因为人类对客观世界的理解是通过在心中建立一个关于它的工作模型，然后再使用这些工作模型来控制、解释、预测事物的发生。心理模型可以是具体的，也可以是非常抽象的。生成这些模型来组织和归纳输入的信息，使之能够进行合理的推断和预测，是一种技能，与阅读理解、算术文字题的理解和其他的很多领域都有密切的关系。

学生在学习过程中建立了许许多多专门化的陈述性和程序性知识，包括词语和概念的网络；事物和方位的认知图表；物质世界和社会的图式；个人和他人的信念、价值、目标、计划；各种推理和求解的技能和策略。这些知识因人而异，往往是部分的、不完全的，甚至不正确的，而且与特定的场合有关。新知识的学习对已有的知识是一种冲击，可以对它进行补充、完善、修正，但已有的知识也会起反作用，歪曲新知识。CIP 对考生知识的评估应该考虑到新知识是如何习得的，它与已有知识是如何组合和变化的、它的应用范围是如何扩展的、它的执行是如何程序化的，等等。

2. 在认知科学指引下的语言测试

要建立崭新的 CIP 模型需假以时日，它必须由认知心理学家和学科的专家共同研究和开发。但是 CIP 对我们的许多启发，可以探讨和逐步试验，一个完整的模型不是一夜之间可以建成的。下面我们想结合语言测试，从过程的角度来谈几个值得考虑的问题：

(1) 语言知识的测试。就语言知识而言，有明示的、有意识的知识（如语法规则），但也有隐含的、直觉的知识（如感觉到那样说不对，但又说不出道理），我们的传统外语教学教的是明示的语言知识，但是考的是隐含的语言知识。这里虽然存在某些语言知识的转换，但是这些孤立地测量的隐含的语言知识和综合的运用语言知识的能力仍有一段距离。近十多年来，我国花了很大的气力来扭转这个方向，使语言测试的重点从语言知识转移到语言知识的运用能力。但是往往在提法上，有些不够全面的地方，好像语言知识和语言能力是可以分立的；其实语言知识不等于语言能力，但是语言能力却包含语言知识，就和程序性知识包含了陈述性知识一样。以过程作为目标的认知测试关心的不仅是能力的测量，而且是知识怎样上升为能力；如果考生缺乏某一方面的能力，这是因为他缺乏知识基础，还是因为他已具有知识，不过知识未能转化为能力？就我国的外语教学而言，已有的母语知识和新学的外语知识有一致的地方，也有矛盾的地方。两者怎样结合在一起，成为统一的、兼容的语言知识系统，更是值得深入考察。

(2) 输出的正确评估。从认知的角度看，考生的知识有比较完整和全面的，也有不那么完整和片面的，也有全无所知的。从一份考卷的整体看，高分者的知识是比较完整和全面的，低分者的知识是十分零碎的，所知无几的。得到中间分数者介乎两者之间。但是如果是选择

题的，则一道题目的分数不是0，就是1，这往往不能反映考生的知识结构。还有的能力往往不是单维的，而是多维的，几个维度之间的关系（权重）又怎样处理？这都是认知测试要解决的问题。

（3）能力的自动化程度。程序性知识都有自动化程度的问题，语言测试中的听、说、读、写、译的能力都有效率（流利）性的差异，但是怎样测量效率却是个问题。程序性知识的执行得较快，且注意资源使用得较少，评估必须考虑到时间的因素。目前有的考试靠增加题量，使题量超过一般考生所能完成的限度，通过题量的完成数来看效率；但是每一部分、每一个题目的效率却仍难以测量。有的考试靠增加面试，来直接观察考生的效率；但面试的评分标准主观性太强，评分员之间难以统一。

（4）组合的程度。在组合中，几个产生式合而为一，以加快执行的速度。组合的程度有大小之分，能力强的考生能够把较多的产生式合成一个产生式系统，能力差的考生却要逐个执行。但是组合会出现定向效应（set effect），即程序的定型化，正面的定向效应可以加速解决问题的能力；反而的定向效应就是僵化，反而会延缓解决问题的时间，而且往往是考生错误的来源。认知测试必须考虑观察考生组合程度的大小和定向效应的正负。

（5）概括能力。概括是知识和技能的延伸和转移，把阅读理解中的能力转移到听力理解，把口语能力转移到笔语，都是概括。概括能力的提高往往是语言能力提高的结果。但是在把部分的能力上升为整体的能力时，会出现所谓过度概括，即忽略了某些事例的特殊性，这在语言规则的概括时常会出现。

（6）学习策略与元认知。学习策略指的是人们用以提高他们习得和保存信息的认知过程，例如在阅读过程中所进行的摘要和推理，为了促进事实性知识的记忆而生成的表象。元认知过程指的是人们对自己思维过程的意识和控制，例如为了提高注意力、知识习得、知识保存而采取的一系列的学习策略。Weinstein & Meyer（1991）提出了几种人们常用的策略：重复性策略（rehearsal strategies）、增添性策略（elaboration strategies）、组织性策略（organization strategies）、理解监察（comprehension monitoring strategies）。他们认为这是从知识的测量转到学习的测量，对了解学生的学习过程，从而提高教学效果有很大的作用。

从上述的几个方面看来，认知测试和以往的测试在指导思想、用途、实现手段等方面都有不同的重点，我们不妨归纳为以下几点：

（1）它着重在了解过程和群体行为，而不仅是了解结果和个人行为。

（2）它是一种诊断性考试，旨在提供关于考生的知识结构、能力水平、思维特点、学习策略、元认知过程的诊断性的信息。

（3）它在多数情况下是一种机助测试，从利用到计算机显示试题、评估成绩、统计速度、计算能力到使用计算机模拟认知和使用语言过程。

（4）它是在心理测量的基础上发展起来的，因此必须使用计算能力的数学模型。

3. 两个实例

下面举两个实例，说明我们怎样尝试改善目前的语言测试，使它能提供更多的信息。应该指出的是，它们对过程的探索还比较浮浅，可以进一步完善。

一个实例是关于阅读理解的。大家都承认阅读理解是一种语言能力，阅读能力之高低，除了体现在理解的正确与否外，还体现在阅读速度的快慢。在阅读训练中，还有所谓快速阅读，目的也是在于提高阅读的效率。但我们关心的是理解正误和理解速度和学生的语言水平高低之间的关系，换句话说，增加速度的参数会不会更好地评估学生的阅读能力？因为阅读的过程牵涉到许多方面的知识的提取和不同策略的应用，水平高的学生提取迅速，策略使用得当，速度自然会快些。季刚孟（即将出版）让英语本科三年级学生在计算机屏幕前自行控制阅读材料的出现，然后再显示问题，让学生选择答案。如果学生想再看材料，他只需要按一个键。计算机将考生的答对题数，每题的第一次阅读文章的时间，答题时反复阅读文章的时间，以及用于答题的时间（除去阅读短文的时间）等数据整理收集起来。然后利用考生二年级期末的一次水平考试成绩及各科总平成绩作为参照点，对实验数据进行分析，找出时间与答对率之间的关系，得出加权的方法。结果如下表：

[表 1]

考生答题情况

考　生	答对率（%）	标准差	阅读时间（秒）	标准差	答对题平均时间（秒）	标准差
高 15 人	82.44	8.31	1342.924	87.452	26.222	5.779
中 15 人	76.00	8.93	1587.152	106.076	31.058	6.212
低 16 人	66.25	12.04	1653.236	143.647	31.208	7.464
总计 46 人	74.71	11.85	1530.499	369.288	29.533	11.855

从平均值来看，阅读能力强的学生答对率高，阅读材料的时间和答题的时间也短，而阅读能力差的学生则反之。相关分析表明，考生的答对率与被参照水平考试分数之间的相关系数只有 0.576（显著性水平为 0.001）。它只能反映考生真实水平的 33%。如果把时间因素考虑在内进行加权，则相关系数可以提高到 0.751。如果把学生中的异常个案（水平考试的分数和各科总成绩距离很大者）除外，则相关系数可达 0.901。答对率和时间所占的权重为 0.85 与 0.15。

另一个实例与知识的完全和不完全有关，着重于摸索出一种反映这两种情况的计分办法。目前的客观题目的评分不是 1，就是 0，不足以反映人类知识的一般的特点。张权（即将出版）采用句子组合（jumble sentences）的测试方式来了解学生的句子知识是否完全，如果学生答不出来则给与提示（给出句子的第一个实义词）。学生接受了提示后仍答不出，可以说是不具备这方面的知识，应给与 0；但如果学生经提示后答出，说明他不是全不懂，而仅是知识不完全，因此所得的分数应该高于 0，但低于 1。怎样才能给出一个合适分数？我们试图用 Rasch 的单参数模型来解决这个问题，首先是把包括未接受和接受提示的全部答对的项目作一次预处理，接受提示后才答对的项目，用星号标出。这样我们可以计算出项目的相对的难度值（相对的难度值指所有的难度值的平均为 0）。[表二]假定有 7 个项目，5 个考生，用 PROX 法（桂诗春，1991）计算出其项目难度值和考生的能力值，表示为对数单位，为了便于观察，我们把难度值和能力值均转换成概率（p 值）。

[表二]

项目难度值和考生能力值的预处理（用 PROX 法）

	I1	I2	I3	I4	I5	I6	I7	总数	%	能力	概率
S1	1	1	1	1	1*	0	0	5	0.71	1.03	0.74
S2	1	1	0	1	0	0	1*	4	0.57	0.32	0.58
S3	1	0	1	0	1	1*	0	4	0.57	0.32	0.58
S4	0	1	0	0	1*	1*	0	3	0.43	−0.32	0.42
S5	1	0	1	0	0	0	0	2	0.29	−1.03	0.26
总数	4	3	3	2	3	2	1	18	0.514	.065	0.52
%	0.8	0.6	0.6	0.4	0.6	0.4	0.2				
难度	−1.44	−0.38	−0.38	0.50	−0.38	0.50	1.56	0			
概率	0.81	0.59	0.59	0.38	0.59	0.38	0.17	0.5			
Q 值	0.19	0.41	0.41	0.62	0.41	0.62	0.83				

把概率和百分比相比较，相差不大。由此看出第一个项目（I1）最易，概率为 0.81。第七个项目（I7）最难，概率为 0.17。经过提示后而答对的项目，应该参考该项目的难度值来评分：如果该项目的难度为 0.5（意味这着有一半的人有可能答对），经提示答对者只能得 0.5（即 1−P 或 Q）。如该项目很易，为 0.81，则经提示答对的，只能得 0.19；相反，如果项目很难，为 0.17，则可得 0.83。所以我们应把［表二］中的 Q 值代入有星号的分数值，然后再用 PROX 法计算其最后的项目难度值和考生的能力值如［表三］：

[表三]

项目难度值和考生能力值的最后计算（用 PROX 法）

	I1	I2	I3	I4	I5	I6	I7	总数	%	能力	概率
S1	1	1	1	1	0.41	0	0	4.41	0.63	0.599	0.65
S2	1	1	0	1	0	0	0.83	3.83	0.55	0.213	0.55
S3	1	0	1	0	1	0.62	0	3.62	0.52	0.077	0.52
S4	0	1	0	0	0.41	0.62	0	2.03	0.29	−1.02	0.27
S5	1	0	1	0	0	0	0	2	0.29	−1.03	0.26
总数	4	3	3	2	1.82	1.24	0.83	15.88	0.45	−0.23	0.44
%	0.8	0.6	0.6	0.4	0.36	0.25	0.17				
难度	−1.73	−0.67	−0.67	0.21	0.37	0.98	1.51	0			
概率	0.85	0.66	0.66	0.45	0.41	0.27	0.18	0.5			

从上表可以看出，S1 答对 5 题，但最后一题是经提示后答对，而该题的难度为 0.59，稍

易于平均值，故应给予 0.41，这个考生最后的能力值为 0.599，转换为概率为 0.65。如果用原始分计算，答对 5 题为全部题目（7 题）的 0.71，答对 4 题，则为 0.57。因为最后一题是经提示后才答对，所以他的分数应在 0.57 与 0.71 之间，而 0.65 是比较合理的。再看 S2 与 S3，大家都答对 4 题，但是最后一题经提示后才答对，S2 答对的是第 7 题，难度最大，为 0.17，而 S3 答对的是第 6 题，难度没有那么大，为 0.38，故他们应分别给予 0.83 和 0.62，经最后计算他们虽然都答对 4 题，他们的能力值应略有不同，故一为 0.213（概率为 0.55），一为 0.077（概率为 0.52）。

参 考 文 献

1. Snow，R. & Lohman，D.（1989），Implications of Cognitive Psychology for Educational Measurement，in Linn（Ed.）：*Educational Measurement*，3 Edition. N. Y：Collier—MacMillan.
2. Embretson，S. E.（1985），Multicomponent latent trait models for test design. In S. E. Embretson（Ed.）：*Test Design：Development in Psychology and Psychometries.* N. Y.：Academic Press.
3. Misley，R. J.，& Verhelst，N.（1987），Modeling item responses when different subjects employ different solution strategies. Technical Report RR—87—47—ONR，Educational Testing Service，Princeton，NJ.
4. Anderson，J. R.（1983），*The architecture of cognition.* Cambridge，MA：Havard University Press.
5. Johnson—Laird，P. N.（1983），*Mental models：Towards a cognitive sciience of language，inference and consciousness.* Cambridge，MA：Harvard University Press.
6. Weinstein，C. E & Meyer，D. K.（1991），Implications of cognitive psychology for testing：contributions from work in learning strategies. In M. C.
7. Wittrock & E. L. Baker（Eds.），*Testing and cognition.* London：Prentice-Hall International.
8. 季刚孟：《阅读测验中的速度参数》，载桂诗春（主编）：《中国学生英语学习心理》，湖南教育出版社。
9. 张权：《提示在语言测试中的意义和作用》，载桂诗春（主编）：《中国学生英语学习心理》，湖南教育出版社。
10. 桂诗春：《题库建设》，载国家教委考试中心（主编）：《题库建设理论与实践》，光明日报出版社，1991年。

〔作者单位：中国广州外国语学院〕

有关较高层次汉语考试的一些问题

张 清 常

这里所谈的汉语考试是属于较高层次的，尤其是带有中国特色的。我勉强地杜撰一个词语，叫它人文的（liberal）语文考试。这人文的涵义接近于西方的 liberal arts，liberal education 的 liberal。它以语言学和语文学为基础，却辐射出文、史、哲、经济、艺术、科学等方面的光辉。它的表现形式是语文考试，使人接触到的是人类智慧的火花。这样语文考试的结果，以我的毕生经验与体会，认为这才是中国的语文考试的理想。

至于“略识‘之’‘无’”而已，例如传说中白居易（772年—846年）这位伟大诗人在婴儿时期，口不能言，考他哪个是“之”字他却能指出；又如许慎《说文解字·序》所引汉朝《尉律》所规定的“学僮”识字写字的考试，这类的要求层次不高的考试，不在本篇讨论范围之内。

本篇着重要谈的只是下述的第一个问题：中国历来对语文水平的观察品评是综合性的；如果以考试的形式来体现，那就是人文的语文考试。至于靠后面的五个问题，只是零星意见，就不必多罗嗦了。

一

万事万物以语言文字为载体，语文的学习必须直接联系万事万物。语文考试不是仅仅解决语言学范围之内所要求解决的语言结构问题，因为这不够，这样只是达到目标，尚未达到目的。目的仍在于考察这样的语文能力，在阐述解决实际问题的答案时，是否令人满意。否则，尽管念起来字正腔圆，声调铿锵，词汇丰富，语法正确，字体完美，其结果仍是废话一大堆。各种八股文章、假大空的演讲辞之类，可以凭借某种特殊权势使之横行无阻，但是在历史的无情的审查之下，这些东西仍是不能通过公正的语文水平标准的。当今在语文考试阅卷时有一种特殊的说法，认为一篇作文，尽管写得怎样文不对题，空话连篇，阅卷的教师总不能给打零分，人家写了一大篇了嘛。其实这种看法不对。对于语文试卷不敢打零分，这有种种原因，其中之一就是不理解语文考试标准的要求。

中国历来对语文水平的考察品评是综合性的。在形式上，要求不论是口语，或是文章，语文结构必须合理，这是最起码的条件。语文表达必须合乎中国习惯，因此旧日评语有“不通”、“欠通”、“通顺”、“清通”、“通畅”、“畅达”等层次，必须能够把自己的思想表达清楚。在语文畅达，思路条理化，达到基本合格的水平以后，评定其语文水平并不是以形式为主，语文只是载体，评议主要重在思想内容。这里需要解释的是，所谓思想内容是广义的而不仅局限于

要求符合当时的政治标准。用中国人的老话来说，所要求的很宽，但必须有头脑，有眼光，有见识，言必有物。

上古先秦诸子百家争鸣，尽管有一些著作和说辞的语言形式并非特别讲究，而他们著作的价值，游说舌辩的成效，首先在于思想内容。两汉从武帝起开始选拔秀才，虽然制度尚不严密，大概要通过策论来衡量才能见识，因为政府考试的目的在于选拔治国平天下的人才。至于网罗文学词章之士，那是另一回事，用另一种标准。

中古隋朝制定科举。唐朝白居易在元和元年（806年）为了应“举制”的考试，自撰策问的模拟题，自己作文回答，共习作七十五篇。真到了考场，所用不过百分之一二，便顺利通过，得到官职。由于自己费了一番心血，所以把这七十五篇编成《策林》四卷，收入《白氏长庆集》。这一批习作，比较全面地表达了他对当时的经济、政治、军事、外交、文化、教育等方面的观点，并提出了解决问题的方案。这件事反映出：一、白居易当时敢于说话，而且话说得有分寸，没惹祸，反而得官做。例如《策林》第二十一篇，主要是说人民之所以贫困，乃是由于皇帝贪得无厌，穷奢极欲，鱼肉百姓。如果不是白居易应考时把内容和口气写得比较和缓，那就是当时对于读书人思想的束缚还不是太严厉，也还希望听一点人民的呼声。二、这时这种语文考试，仍是注重思想内容的。

唐诗在中国文学史上居于显著的地位。唐朝的语文考试，诗赋是必考的一项。当然，官府科举考试仍以策论为主。至于日常生活中，作诗以显示才华，表达心意比较方便。唐诗的精品很多，这大概是由于三百年间全民练习写诗，当然会涌现出、筛选出大量的名家名篇。但，即使是一首小诗，品评其高下，思想内容仍是很重要的。

据说白居易初到长安，带着自己的诗文习作去拜见著作郎顾况。顾况看到他的名字是“居易”（取自《礼记·中庸》，安心于平凡和谐），便说：长安是争名夺利的首都所在，“米价方贵，‘居’亦不‘易’。”看到第一首诗，是白居易十五岁时所作《赋得古原草送别》：“离离原上草，一岁一枯荣。野火烧不尽，春风吹又生……”，大为赞赏，说：“道得个语，‘居’即‘易’矣。”从此白居易名声大振。十五岁的习作就能说出“野火烧不尽，春风吹又生”这样富于深刻哲理的千古名句，当然是很了不起的。

到了后代，传说有一首咏雪的打油诗：“一片两片三四片，五片六片七八片，九片十片十一片，飞入梅花（一作芦花）都不见。”从前小孩子描“红模子”学写字，也有四句：“一去二三里，烟村四五家，亭台六七座，八九十支花。”如果你抓的是语文结构完整就算100分，后者可得满分而前者只能得25分。如果你抓思想内容，前者可以及格而后者可以淘汰；前者是打油诗，后者只是韵语。

为了从幼儿时期便训练语文学习，要开拓思路扩展眼界，所以让儿童们从“红花”想到“绿叶”，从“山高”想到“月小”，从“水落”想到“石出”。并没有向幼儿讲语音语法词汇修辞，面实际上却是依此而设计的。更重要的是训练思路的活泼，切忌迟钝呆板；既要“举头望明月”，也要“低头思故乡”。

过去以“对对子”测试语文水平。上联出个“孙行者”，如果连孙行者就是《西游记》齐天大圣孙悟空那个泼猴都不知道，那就无从说起了。如果教师所要求的只是词义解释，语词结构分析，这三个字的标准读音和写法，那也罢了，至此为止，这还有什么可说的呢？可以得满分的“对千”是“祖冲之”，因为这表明应试者知道世界文化名人里面有个中国南北朝的伟大

科学家，祖冲之算出圆周率π的值、约率和密率值，要比欧洲早一千多年。祖冲之还是杰出的天文学家和物理学家。祖：孙，冲：行，之：者相对，非常工整。而出这个题的人认为“胡适之”可能更好。因为这比“祖冲之”又深了一层，猴子又叫加犬旁的猢狲．所以吴承恩写《西游记》给齐天大圣这个泼猴找个姓就是孙。能够在“对对子”的时候想到以胡：孙，这证明应试者的头脑更活跃。

当然，在今天，再提倡旧时代的“对对子”是不合适的，因为现代汉语的语音打破了古代平仄规律，单音字词为主线的时代已经过去，大量的双音词和多音词不断滋生……。可是，语文考试不宜停留于语文结构本身的分析为止，应该如何引导到重视思想内容，开拓眼界，搞活思路，成为人文的语文考试，这是值得我们一起来思考的问题。

中国古代的考试由人文的语文考试，退化为追求形式的躯壳，也由于科举。到了明清两朝，改用八股取士，八股就是极端形式主义的死格式。题目专从《四书》及朱熹注里面出，光怪陆离，无奇不有，早已走上绝路，成为文字游戏。可悲的是，虽然清朝末年便已取消八股，而八股的阴魂不散，借尸还魂，于是有七八股、洋八股、党八股等等，“假大空”就是新时代新形势下的伪劣商品式的语言文字游戏。

因此，语文教学和语文考试的改革，实在是很迫切需要而且关系重大的事。

二

其他有关语文考试的，只有一些零散而且不成熟的意见，所以只是三言两语。

（一）不论今天现代汉语如何发展与丰富，看起来单音词、单音字从几十万年前形成华语，五六千年前形成华文，就是华语华文的命根子。单音词、单音字一直是构成新词语及词汇发展的基础。国家对外汉语教学领导小组办公室汉语水平考试部《汉语水平词汇与汉字等级大纲》统计所得汉语常用字词基本数据是：词汇总计8822，汉字总计2905。这一统计数据是科学的，可信的。从学习汉语来说，掌握常用词汇8822个不算多，掌握常用汉字2905个不算少。我这里提供一件事例供考虑。据统计：名著所用过的不同汉字的字数，《韩非子》为2680个，《红楼梦》为4462个，《毛泽东选集》一至四卷为3002个。

因此我认为要想迅速提高汉语水平，如何善于利用2905个汉字去掌握构词方法，扩展词汇，使每个汉字都活跃起来，这可能是最直接而且迅速见效的办法之一。汉语考试可以在督促学生掌握汉字，活跃汉字这方面多设计些方法，既考试又诱导。如果行之有效，学生掌握一两万条词语不算难事。

（二）语文考试中发现，学生词汇贫乏是个严重问题，这是“学生腔”的表现之一。基本词汇并不真正熟练掌握，对其涵义、义项、用法等马马虎虎，这样，基本词就活跃不起来。而所谓“丰富”“扩展”词汇的途径就是追求一些花里胡哨的字眼儿。再加上某些考试命题者不在词汇的要害方面进行督促检查，喜欢出些如何解释成语之类的题，于是学生背诵成语成风。又一再闹“俗语”、“惯用语”[9]、“歇后语”……热，这些工具书重复出版。精粹的成语、俗语等都是汉语的宝贵财富，应该认真学习掌握，准确理解，恰当使用，非常必要。但如果生吞活剥，胡乱使用，则不但措词失当，而且舍本逐末，反而有助于“假大空”，有助于“耍贫嘴”。词汇贫乏的根本原因是：既生活贫乏，又精神世界空虚。