# 认知与语言测试

广东外语外贸大学　桂诗春

## 一、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，而项目反应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)；而测度本身还必须有效，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，它感兴趣的是每一个考生的行为，可称之为教育心理测量模型 (Educational Psychometric Measurement Models, EPM), 表示如图1：

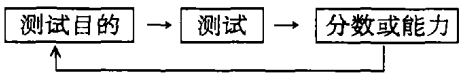


图1

这种EPM目前还是考试中的主流，但从认知科学的角度来看，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能站得住；一个模型的好坏主要是看它能否很好的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，每道题目的语境效应来源甚多，很难说题目是局部独立的，因此误差也不是全不相关的。考生的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作是不可企及的“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：怎样对测试的内部特征作试验，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的这些问题提出来的。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分的潜在倾向模型，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：

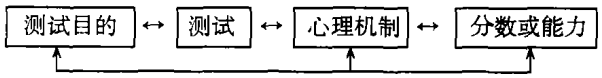


图2

从认识科学的角度看，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，它就更易被激活。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,它建立在陈述性知识的基础上,包括陈述性知识和它的使用条件。陈述性知识是静态的,而程序性知识是动态的;它强调区别与概括,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)认为技能自动性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,即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,对他独立掌握技能会起到促进作用;如果他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大脑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化的,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许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。

## 二、在认识科学指引下的语言测试

要建立崭新的CIP模型需要假以时日,必须由认知心理学家和认知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),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,但是这些孤立测量的隐含的语言知识和综合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不够全面的地方,好象语言知识和语言能力是可以分立的。语言知识不等于语言能力,但是语言能力包含语言知识,就像程序性知识包含陈述性知识一样。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,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,也有不那么完整和全面的。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,得到中间分数者介乎两者之间。不过如果是选择题,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) 能力的自动化程序。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,但怎样测量效率却是个问题。程序性知识执行得较快,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) 组合的程序。在组合中,几个产生式合而为一,以加快执行的速度。组合的程序有大小之分,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;反面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。认识测试必须考虑观察考生组合程序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中常会出现。

(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、为促进对事实性知识和记忆而生成的表象。元认知过程指的是人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)、理解监察策略(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。

## 三、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,使之提供更多的信息。应该指出的是,这一探索还比较浮浅,有待完善。

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同学生的语言水平高低之间的关系。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(1992)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,他只需按一个键。计算机将考生的答对题数,每题的第一次读文时间,答题时反复读文的时间,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。结果如表1。

| 考 生 | 答对率(%) | 标准差 | 阅读时间(秒) | 标准差 | 答对题平均用时(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 | 7.454 |
| 总计 46 人 | 74.71 | 11.85 | 1530.499 | 369.288 | 29.533 | 11.855 |

表1 考生答题情况

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.576(显著性水平为.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.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.901。答对率和时间所占的权重为.85和.15。

另一个实例与知识的完全和不完全有关,着重于摸索出一个反映这两种情况的计分方法。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特点。

张权(1992)用句子组合(jumbe 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予提示(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2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)。

| | 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 | 0.29 | 0.26 |
| 总数 | 4 | 3 | 3 | 2 | 3 | 2 | 1 | 18 | 0.514 | 0.65 | 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 | -50 | 1.56 | 0 | | | |
| 概率 | 0.81 | 0.59 | 0.59 | 0.38 | 0.38 | 0.38 | 0.17 | 0.5 | | | |
| Q值 | 0.19 | 0.41 | 0.41 | 0.62 | 0.41 | 0.83 | 0.83 | | | | |

表2 项目难度值和考生能力值的预处理(用PROX法)

把概率和百分比相比较，相差不大。由此看出第一个项目(11)最易，概率为.81。第七个项目(17)最难，概率为.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.5 (意味着有一半的人可能答对)，经提示答对者只能得.5(即1-p或Q)。如该项目很易，为.81，则经提示答对的，只能得.19；相反，如果项目很难，为.17，则可得.83。所以我们应把表2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3：

| | 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 | | | |

表3 项目难度值和考生能力值的预处理(用PROX)法

从上表可以看出，S1答对5题，但最后一题是经提示答对，而该题的难度为.59，稍易于平均值，故应给予.41，这个考生最后的能力值为.599，转换成概率为.65。如果用原始分计算，答对5题为全部题目(7题)的.71，答对4题，则为.57。因为最后一题是经提示后才答对，所以他的分数应在.57与.71之间，而.65是比较合理的。再看S2与S3，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.71，而S3答对的是第6题，难度没有那么大，为.38，故他们应分别给予.83和.62。经最后计算，他们虽然都答对4题，但他们的能力值略有不同，一为.213(概率为.55)，一为.077(概率为.52)。

【原载《外语教学与研究》1992年第3期】

## 参考文献

[1] Anderson, J. R.1983. *The architecture of cognition.* Cambridge, MA:Harvard University Press.

[2] Embretson, S.E.1985.Multicomponent latent trait models for test design.In S.E. Embretson (ed): *Test Design: Development in Psychology and psychomertics.*N.Y.: Academic Press.

[3] Johnson-Laird, P.N.1983.*Mental models:Towards a cognitive science of language,inference and consciousness.* Cambridge, MA:Harvard Univecsity Press.

[4] 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.

[5] Snow,R.& Lohman,D.1989.Implications of cognitive psychology for educational measurement.In Linn (ed): *Educational Measurement,* 3 Edition.N.Y: Colier-MacMillen.

[6] Weinsein, C. E. & Meyer, D. K. 1991. Implications of cognitive psychology for testing:contributione from work in learning strategies. In M.C.Wittock & E.L.Baker(eds.): *Testing and cognition.* London:Prentice-Hall International.

[7] 桂诗春(1991)，题库建设。载国家教委考试中心(主编)：《题库建设理论与实际》。北京：光明日报出版社。

[8] 季刚孟(1992)，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》。长沙：湖南教育出版社。

[9] 张权(1992)，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》。长沙：湖南教育出版社。