第26卷第3期 宁波大学学报(教育科学版) Vol.26 NO.3
2004年6月 JOURNAL OF NINGBO UNIVERSITY (EDUCATIONAL SCIENCE) June 2004

常模参照性测试与尺度参照性测试比较

许华琳

摘要:不同的教学目的运用的语言测试各不相同。常模参照性测试与尺度参照性测试是两种应用广泛、发展成熟的测试模式。在常模参照性测试与尺度参照性测试比较的基础上,从两种测试的概念、目的、考试内容的设计和对分数的解释等方面论述常模参照性测试和尺度参照性测试在英语教学中的作用。

关键词:语言测试;常模参照性测试;尺度参照性测试

中图分类号:G423 **文献标识码**:A **文章编号**:1008-0627(2004)03-0079-03

在过去的数十年里,随着应用语言学研究的不断发展,与之相应的语言测试学研究也越来越受到重视,并且逐渐成为一门独立的学科,有着自己的研究领域和研究方法。语言测试学包含了诸多元素:语言学、心理测量学、语言教学论等。[1](312) Upshur 曾经指出:语言习得、语言教学发展与语言测试的研究有着内在的、本质的相互联系。语言测试为语言习得和语言教学服务,同时后两者反过来对前者也起了促进作用。[2](54) 例如,语言测试常常在二语习得的研究中作为衡量语言能力的重要手段。语言测试也是检测教和学的有效性信息的有价值的来源,语言教学者运用测试来检查学生的学习效果,并对不同的教学法进行评估,从而实现测试的目的。不同的测试目的、不同的教学背景,运用的测试方法也各不相同。常模参照性测试与尺度参照性测试是语言测试学中两种发展较为成熟、应用广泛的测试模式。本文在常模参照性(norm-referenced)测试与尺度参照性(criterion-referenced)测试比较的基础上,从两种测试的概念、目的、考试内容的设计和对分数的解释等几个方面论述常模参照性测试和尺度参照性测试在英语教学中的作用。

1. 常模参照性测试与尺度参照性测试的概念

常模参照性测试是测试者通过考试结果对受测者个体的语言行为与某一个特定群体(所有参加该次考试的考生)做比较,从而确定这一特定群体中受测者个体之间的语言能力的差异。[2](72) 由于通过常模参照性测试可以使受测者的语言行为得以最大程度地区分,因此,它又被称为"心理测量学"的测试,因为大多数心理测量学的理论模式是建立在对个体受测者分数的一个常规分布的推测和偏差最大化的基础上的。[2](74) 心理测量学测试有四个主要特征:(1)用数值(分数)来评价人们的行为;(2)青睐高度结构化的试题;(3)关注行为的结果,而不是行为的过程;(4)主张用考生在各相关测试中的平均分来作为其某种行为的评价结果。[3](557) 在小规模的常模参照性测试中,常模指同一组学生的平均分数,在大规模的标准化考试中,常模代表不同地区、不同时间参加同一类型考试(不一定是同一份试卷)受测者的平均水平。

尺度参照性测试是测试者通过考试结果对某个考生的语言行为与预先设计的能力、目标或技能标准做比较,以此来衡量该考生的语言行为是否达到该标准。[2](74) 它的最大特点是直接描述受测者在测试中表现出来的语言行为。用一个事先决定的尺度去比较所有的考生。[4](18) 尺度参照性测试对受测者之间的差异、解答特定试题的表现没有兴趣(不做受测者个体之间的差异比较),它所注重的是将受测者考试时的行为扩展到更广泛的行为范围。

收稿日期:2003-10-08

作者简介:许华琳,浙江大学外国语学院应用语言系讲师(杭州 310058)。

2. 常模参照性测试和尺度参照性测试的不同目的

Bachman 认为就语言测试和对测试结果的解释两者来说,测试者的惟一最重要的考虑因素是某个特定的测试方法的目的是什么,即测试是为什么服务的。[2](54)根据常模参照性测试的概念我们可以知道,该测试的目的是甄别受测者的语言水平,即哪个受测者的语言水平好,哪个水平差。例如分级考试(placement tests),当教学计划开始实施之前,为了了解学生的语言程度,通过常模参照性测试,把语言程度相差悬殊的学生加以分类,按语言程度划出等级不同的班级,使教师能够有针对性地制定教学方案,因材施教。常模参照性测试还被用于各类选拔性考试中。

由于尺度参照性测试只考虑受测个人在测试中的语言行为,不考虑测试中别的受测者的表现,不做受测者个体之间的差异比较这一特点,它常常用于诸如学业考试(achievement tests)、资格考试以及各类会考和统考中。

3. 常模参照性测试和尺度参照性测试不同的设计特点

结构主义理论认为语言是可分解的一个系统,它是由语音、词汇、语法等元素构成的一个有限集合,这些有限集合的成分可以构成无限集合的句子。而所谓掌握一门语言就是掌握语言中这些元素并用来生成和理解无限数量的句子的能力。因此可以用离散的题目(discrete items),来逐项检测学生是否掌握了这些元素。[5](245)建立在这一理论基础上的常模参照性测试的设计具有以下三个特点:第一,由于一次测试不可能包含所有的语言成分和语言技能,常模参照性测试采用抽样的方法来抽取相应的语言成分和听、说、读、写活动。其选题标准往往取决于试题的统计数据。理想的常模参照性考试的选题是 50% 的受测者能够正确答出该题。如果一道选题,90% 的受测者都能答对,那么这就不是好的选题,因为它不能把好的学生和差的学生区别开来。同样 90% 的受测者都不能答对的选题则太难,也不能达到考试的目的。[1](315)因此在常模参照性测试中,太难和太易的选题都应排除。试题的难度值在 0.3 ~0.7 之间、区分数值高于 0.3,都被认为是好题。第二,考试内容具有形式的客观性和实质的一致性。形式的客观性是指常模参照性测试的设计者把抽取的语言成分和听、说、读、写活动转化成高度结构化的客观题,如最常见的多项选择题,一道多项选择题只测一个语言成分或一个语言技能。一致性指不同时间举行的同种考试内容难易程度是相同的,即使考试的形式有变化,考试内容的实质还是保持一致。第三,考试的管理和评分程序实行标准化。常模参照性测试的阅卷通常采用累计答对题目数的方法,有时也采用修正分数的公式来排除考生的盲目猜测对测试结果的影响。[3](558)常模参照性考试程序的标准在这一次考试和下一次考试中不会发生变化。考试的衡量尺度是公开的,考试的信度(reliability)和效度(validity)是经过审慎地检验和论证的。

由于测试的目的不同、方法不同,尺度参照性测试的设计与常模参照性测试的设计有很大的不同。尺度参照性测试是在分数的基础上,通过事先制订的行为准则来预测考生的实际的语言运用行为,因此,对这些预先制订的测试标准的定义必须是明确的、科学的。尺度参照性测试试题的舍取以是否违背预先制定内容细则和命题标准为原则。只有在科学的、合理的行为参照准则的前提下,才能保证测试结果(考分)的准确性。因此如何保证考试的信度是尺度参照性测试设计者的首要考虑问题。假设,我们制定了衡量考生语言行为的细则,并以此为依据从试题库中选择可以代表这些语言能力的选题。那么怎样才能确保这些选题是能够代表这些能力的呢? 如何确保这一次测量完全可靠呢? 也就是说,每一次用同样的"尺子"测量考生的语言行为,得到的都是同样的结果。桂诗春教授曾用 Livingston 提出的一个理论公式来测算尺度参考性测试的信度:

$$P_C^2(T_x,X)=\frac{\sigma_r^2+(\mu_x+C_x)^2}{\sigma_x^2+(\mu_x-C_x)^2}$$

$\sigma^2 r$ = 真正分数的方差,σ_x^2 = 观察分数的方差,μ_x = 平均分,C_x = 通过某一尺度的分数;这个公式主要是按照分数偏离尺度的分数来重新定义方差[4](136)。通常信度值 >.90 的测试是比较

理想的测试。我们在进行尺度参照性测试设计时,还应注意以下三个方面:(1)在不同的试卷中测试者要衡量的能力数量有所不同。(2)衡量每一种能力的试题数量和最低标准值随能力的种类的不同而有变化。(3)决定考生通过与否取决于考生的分数是否达到或者超过能力、目标和技能所要求的最低标准[6](173)。

4. 常模参照性测试和尺度参照性测试对分数不同的解释

解释常模参照性测试的某个具体的分数是以一个特定的群体的语言行为或称为常模(norm)为参照对象的。这一特定的群体是指参加该类考试的所有考生,他们的语言行为(norm)通常用平均数($\bar{x}$)(mean)来表示,而标准差s(standard deviation)则表示该群体的分数的分布状况。一个设计合理的常模参照性测试,考生的分数分布呈现典型的"钟型"正态曲线。[2](72)这个理想的测试模式是:50%的分数低于平均值,50%的分数高于平均值;34%的分数高于一个标准差(+1S),34%的分数低于一个标准差(-1S),27%的分数在一个或两个标准差之间(高于平均值13.5%或低于平均值13.5%),只有5%的分数远离两个或更多的标准差(远离平均值)。图1为Bachman的常模参照测试的典型正态分布图。[2](73)

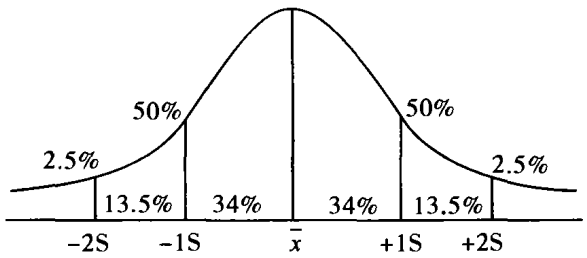

图1　Bachman的常模参照测试的典型正态分布图

我们运用Bachman的分析模式,可以准确地比较某个考生在某次常模参照性测试中与其他考生的语言行为差异。如:1998年6月的全国大学英语四级统考的平均成绩是56.71,标准差为13.91。[5](252)由此可以推断,如果某个考生的成绩为70.62分,那么其分数高于平均分一个标准差(56.71 + 13.91 = 70.62),这个分数说明该考生的语言能力高于84%的参加这次四级考试的考生。

尺度参照性测试是按照事先制定的评分标准来确定受试者的语言能力,因此在尺度参照性测试中,常常要先确定一个划线分(cut - off score),用它来表示事先制定的标准。例如,某一次学业考试共计测试题100题,每题1分,划线分为80分,即考生得分在80分以上就通过该考试,也就是说,他的语言能力达到了测试者事先制定的语言行为标准;而低于80分的考生则未能通过该考试,其语言能力也就未能达到该标准。我们知道就考试而言,由于各种原因,考生的考试分数与他实际的分数会有误差,这种情况肯定存在,那么如何确定考生的考试分数与实际分数之间的误差值呢?语言测试学家Berk的公式可以用来估计考生考试所得的分数可能偏离他的真正分数的界限。该公式是:

$$SE_{means}(x_i) = \sqrt{\frac{x_i(n - x_i)}{n - 1}}$$

x_i代表某一考生的分数,n等于该次考试的总题数。[2](213)例如:某次测试共30道试题,某个考生的考分为25分,根据Berk的公式得出该考生的考分与其实际分数的误差值为2.28,那么在95%的情况下,他的真实的分数不会高于$25 + 2SE^m$($25 + 2 \times 2.28$),也不会低于$25 - 2SE_m$($25 - 2 \times 2.28$)。

综上所述,常模参照性测试与尺度参照性测试作为语言测试学中的两种重要的测试模式,根据不同的测试目的、不同的教学方法的需要,正被广泛地应用于教学和教育管理的各个领域;同时又为广大的语言教学工作者在语言测试的实践中提供了理论依据,并在此基础上不断改进,使语言测试与语言教学能够得到完美的结合。

参考文献:

[1]R R van Oirsouw. Applied linguistics and the Learning and Teaching of Foreign Languages[M]. London: Edward Arnold, 1984.

[2]Bachman L. Fundamental Considerations in Language Testing [M]. 上海:上海外语教育出版社,1999.

[3]王振亚. 语言测试的目标与实现手段[A]. 语言学[C]. 北京:外语教学与研究出版社,2003.556-567.

[4]桂诗春. 标准化考试:理论、原则与方法[M]. 广州:广东高等教育出版社,1986.

[5]杨惠中. 语言测试与语言教学[A]. 中国的语言学研究与应用[C]. 上海:上海外语教育出版社,2001.239-258.

[6]徐强. 交际法英语教学和考试评估[M]. 上海:上海外语教学出版社,2000.

(责任编辑　徐鸿钧)