语言测试术语及其运用
——有关语言测试术语的几点介绍

刘 冰
(淮南师范学院 外语系,安徽 淮南 232001)

摘 要:由于考试结果通常用分数来表示,因此考试的分数必须具有可解释性。而分数解释又涉及记分体制,为了使分数带有大量的信息,考试成绩在最终报道前,总要经过加权处理,等值处理,正态转换等一系列转换分析过程。作者对随之产生的有关英语测试的专业术语进行了详细的介绍,对其使用过程和作用作了详尽的分析。

关键词:集中趋势;离中趋势;效度和信度;后效作用
中图分类号:H310.4 文献标识码:A 文章编号:1009-2463(2002)06-0112-03

On Terms of English Testing

LIU Bing
(Huainan Teachers College, Huainan, 232001, Anhui)

Abstract: As test results are usually reported in the form of test scores, they must be interpretable, which involves the use of a certain scoring system. In large scale standardized tests, the test scores, before release, have to undergo a series of transformation processes, including score weighing, score equating and score normalization etc. The author of this paper and the terms concerned in English testing explains in detail how these scores are interpreted and discussed.

Key words: central tendency; dispersion; validity & reliability; wash-back effect

人们对传统的语言测试术语如水平测试,考试难度,平均分等并不陌生,但这些概念的隶属关系,实际含义及如何运用是值得研究的。现代语言测试的一大特点是语言测试与统计学的相结合。[1](P239)这样一来,不可避免地要产生一些新的测试学术语,本文将着重介绍有关知识及用途。

一、集中趋势(central tendency)

集中趋势指分数分布中有代表性的中间点,它主要由平均分(mean)、众数(mode)和中位数(median)来测量。平均分所有分数的算术平均数(arithmetic average),即分数之和与考生总数之商,通常由分式 $X = \sum x / N$ 表示,其中 X 表示平均分,$\sum x$ 表示个体分数之和,N 表示考生总数。众数指多数考生所获得的那个分数,即考生分数分布中出现次数最多的分数。如一组分数为 68、69、70、70、70、71、74,则众数为 70。中位数指位于分数分布中间的那个分数,如果分数组的个数为奇数则是中间的那个分数,如一组分数为 70、73、75、80、87、中位数即为 75。如果是偶位数,中位数则是中间两个分数的平均值,如一组分数为 70、73、75、79、83、86 则中位数为 75 和 79 的中间值 77。中位数试者的人数没有多大关系,平均分众数和中位数均用于测量分数分布的集中趋势,测试是否能够用不同的分数来区分不同水平的受测者,这是研究测试质量的极为重要的方面。根据教育统计学的原理,通常用曲线来表示分数的分布情况,如果分数呈正态分布(normal distribution),那么表示分数分布的曲线就叫正态分布曲线,这时算术平均数,中位数和众数是重合的。否则,分数呈偏态分布(skewed distribution)。这时,算术平均数,中位数和众数三者是不重合的。

要使测试的分数态到正态分布是很不容易的。水平测试,特别是选拔性的水平测试,分数一般应达到或接近正态分布。

二、离中趋势(dispersion)

集中趋势与分数分布的集中性有关,而离中趋势则与分数分布的离散性有关。全距(即最高分减去最低分)是描述离中趋势的最好方法。由于全距只涉及分数分布中的最大值和最小值,我们无法得知其它分数分布的情况,所以我们又要引入标准差的概念(即一组分数总体如何以平均分为中心,向两边离散分布)。标准差涉及到一组分数的每一个个体。

标准差可以帮助我们科学地进行测试评定,标准差大,意味着分数分布的离散趋势大,说明考生水平相差较大,分

收稿日期:2002-08-12
作者简介:刘 冰(1966-),女,汉族,江苏南京人,淮南师范学院外语系讲师,在读研究生。

数全距也相应较大。标准差小，则意味着分数围绕平均分紧凑分布，考生水平接近。

标准差的大小与测试目的有关。尺度参照性考试的目的在于测试考生是否达到某一要求，如果平均分较高，我们则希望标准差较小，这样分数都紧密分布于平均分周围，说明考生总体水平较好，考生同质性（homogeneity）强，达标理想。如果我们的测试目的在于选拔人才，较大的标准差能将各个不同的档次的考生分开，以达到选拔的目的。同时也表明考生的异质性（heterogeneity）较强。

三、试题易度和区分度（item facility & discrimination）

试题易度即难易度，是测试后通过题项分析（item analysis）计算出来表示题项难易程度的量数。为什么不叫难度呢？是因为这个数值与"容易的程度"成正比。题项越容易，易度值（facility value）就越高，最高可达到 1，表示所有受测者都做对了。题项越难，易度值就越低，最低可达到 0，表示没有人做对。易度值由分式 $P = \sum Cr / N$ 求得。其中 $\sum Cr$ 表示答对人数。N 表示总考生人数。例如当 $P = 0.80$ 时，即 80% 的考生都能答对。

试题的易度值是出题者的一个指导思想。由于语言测试的目的一般在于了解各个档次考生的语言能力情况，不正常的试题易度值都不能达到测试目的，除特殊情况外，试题的易度值一般应住于 0.4 – 0.6 或 0.3 – 0.7 之间。[1](P248)一般而言，在整张考卷中，中等难度试题占大部分（60%），两极（难或易）试题占小部分（各 15%），并且所有试题应按难度递增顺序排列。即 15% 易题 + 70% 中等题 + 15% 难题。以此确保各个档次的考生不至于对考试失去兴趣。[2]

试题的区分度指试题区分好、中、差考生的程度。由区分度指数（index of discrimination）表示。区分度指数越高，说明题目的区分力越强，最高可达到 1，表明试题能完全区分不同能力的考生，即作对试题的是水平好的考生，做错试题的是水平差的考生；反之，区分度指数越低，说明题项的区分力越弱，最低可达 – 1，表示试题完全反向区分。

在计算区分度指数时，把受试者按照所得分数的高低分成人数相等的两组，当高分组答对某项试题的人数比低分组多时，区分度指数为正数。当两组人数相等时，区分度指数为 0，当高分组人数比低分组人数少时，区分度指数为负数。计算试题区分度的方法很多，我们也可用下列比较简单的公式来计算实际的区分度。

$$D = \frac{H - L}{N}$$

其中 D = 区分度指数

H = 高分组答对某题的人数

L = 低分组答对某题的人数

N = 高分组（或低分组）人数

区分度指数旨在统计整张试卷总成绩好的考生每一项试题的成绩合格率。如果一项试题好的考生都能作对，而差的考生都做不出，说明试题的区分度高，试题质量好，区分了不同能力档次的考生。相反如果水平差的考生作对了，而水平好的考生反而做错了，则试题的区分度较差，应予以修改。因为它正好反向区分考生，与测试目的背道而驰。

一般而言，区分度指数低于 0.30 的试题需要谨慎对待。在检验试题质量的预测后，每题的易度和区分度指数通常被列于其后，试题只有经过这两方面的检测后才能投入使用，尤其是大规模国家级的考试试题。

四、效度与信度（validity & reliability）

评价一次考试质量高低的最重要的标准是效度和信度。[3]

考试的效度指的是考试在多大程度上测出预期要测量的东西或者说考试在多大程度上完成了预期的测量任务。达到了预期测试的目标。效度概念包含了两层含义，一是考试究竟测的是什么（what）；二是测试的程度有多大（how well）。

效度是一个相对的概念。效度的有效性总是相对于一定的目的、功能和范围而言的。对于某一目的是有效的考试，用于另一目的就未必有效。效度相对性的另一层含义是程度的相对性。即效度不是或有或无的关系，而只是高低程度上的不同。

考试的效度可分成三大类：内容效度，构想效度和效标关联效度[4]（舒运祥，2001：52）。

1. 内容效度（content validity）

内容效度是指考试是否真实地体现了它所要测量的内容，或考试的题目多大程度上代表了它所要测量的目标，考试实际上是一种抽样，事实上在每次有限的考试时间不可能考所有学过的东西，而只能挑一部分，这就叫抽样，我们通过考一部题目，来推断学生总的学习情况，因此，出题有没有代表性，就意味着考试的结果能不能说明学生的实际水平。[5]

内容效度对于成绩考试和标准参照考试都具有特别重要的意义，因为成绩考试的用途是测定学生对过去学过的知识和技能实际掌握的程度，因此大纲是成绩考试的命题依据，标准参照考试以原先制订的标准作为评价考试成绩优劣的依据，所以这两类考试都必须重视内容效度。

为了提高内容效度，应该事先拟定考试内容细目表，根据大纲和其他的教学目标和要求，列出所有待测的内容细目。然后按照细目表编写具体的试题。这样才能保证试题的代表性，避免只出容易编写的试题或收集手头现成的试题，偏离了考试目标。

2. 构想效度（construct validity）

考试的构想效度指考试实际测得的东西与理论所假设的能力要素或心理特征相吻合的程度。由于我们需要借助理论构想来判断考试成绩是否代表一个人的真实能力，所以构想效度也称理论效度。例如，交际能力包括四个因素：形式正确，合理可行，内容得体，实际运用。如果考试的结果证实它确实考核了这四个方面，那么可以认定这次考试的构想效度好，因此成绩高的学生，他的交际能力强，反之则不能说分数好就一定交际能力强，要想取得较高的构想效度，在试卷编写的过程中至少应包括如下步骤。

（1）对所要测试的语言能力、交际能力和其它心理语言特征，根据理论提出测试的构想，如，测试学生的阅读能力时，先要根据某种理论，列出影响阅读理解能力的要素，然后围绕这些要素编题。

（2）根据假定的测试构想编出试题。

（3）检验考试结果与测试构想的吻合程度。

以上三个步骤中，第 1、2 步是取得构想效度的关键。因为在编写具体试题前，对测试有一个总体的科学的构想是十分重要的。

3. 效标关联效度（criterion – related validity）

效标关联效度就是以考试分数与其效标分数之间的相关，来表示的效度，也称统计效度。效标关系效度表示的是考试与效标之间彼此拟合的程度，根据测试分数和效标分数获得的时间关系，可以将效标关联效度分为预测效度和共时效工，预测效度的考试分数获得在前，效标分数获得在后，这两个分数获得的时间有间隔。而共时效度两个分数几乎可以同时获得。

（1）预测效度（predictive validity）

预测效度指测验分数对考生未来行为和作业作出预测的准确程度。预测效度越高，区分学生学术研究能力的准确性也越强。检定预测效度的最重要，最难的一个问题就是选择一个最有效的效标。我国高考英语笔试的预测效度一般以大学的四、六级考试为效标。国外大学入学考试的预测效标一般以被录取新生在大学一年级各科平均成绩作为入学考试的效标。拿学能测试举例来说，如果测试得高分的考生在今后的外语学习中成绩优异，最低分的考生则成绩平平，那么该学能的测试便具有预测的有效性。再比如在选拔出国留学人员的外语测试中，如果通过外语测试的人，不能胜任在国外的学习和深造，则说明这种测试没有预测的有效性。

（2）共时效度（concurrent validity）

共时效度指考试分数与同时获得的某种令人满意的行为或另外一次有效考试之间的相关程度。由于需要检验的一次考试成绩和用作效标的考试成绩几乎是同时获得的。所以这种效度称为共时效度。作为效标的即可以是一个公认的效度很高的考试，也可以是教师在长期教学中形成的对学生的评定。

求证考试的共时效度至少有两个作用，第一，可以利用现成有效的考试来检验新编制测验的效度。第二，可以利用共时效度寻找一个简便的测试手段以替代一个已被证明的有效的，复杂的考试。例如 TOEFL 考试经过多年的实践与研究被公认为是有效的考试。曾经流行于我国的"英语水平考试"EPT（English Proficiency Test）在试用时，为了了解考试的有效性将 TOEFL 考试作为效标，让考生同时做 EPT 和 TOEFL 考试，求得相关系数在 0.85－0.87，从而说明 EPT 有关很高的共时效度。质量评估要考虑的第二个问题是测试结果的可信程度（或可靠程度）即信度（reliability）。谈信度通常从以下三个方面来研究并进行比较。

1. 再测信度（test－retest reliability）

再测信度就是用同一套试题在尽可能相似的环境中对同一批受试者测试两次。通过计算、比较两次成绩，求出相关系数。求相关系数用得方法比较多，主要有两种 Pearson 积矩率法和 Spearman 排列次序相关法。用前一种方法，只要知道人数，第一次和第二次测试的分数就能算出来；用第二种方法，是与研究两次测试的次序相关。根据成绩把受试者两次测试的名次排出来，比较两次测试名次上的差异，求出测试的稳定性。

2. 复本信度（equivalence）

复本信度常用等值法（the equivalent－forms method），是指对同一批考生进行测试的两份或复份试卷，在考试性质、内容、题型、题量、难易度等方面都一致或相等。用复本测试对同一组考生在同一时间内施考，所求得信度数称为等值系数。这种方法可以有效地防止学生因记忆，练习效果而产生的复测成绩的变化。同一组被试者在复本测试上所得的结果的相关系数就是复本系数；若复本个数在两个以上时，复本信度可用每两个复本测验结果的相关系数的平均数来表示。

复本信度反映了测验过程中的误差、形式不同的试题间测试结果的一致性以及学生的答题在一段时间内的稳定性，与其它测试信度相比，这类信度是最有用的。因为它能让我们估计测验结果在多种条件下的普遍意义。用此法求得的信度系数高，说明考生的测验成绩不仅代表现在的测验表现，而且，也代表今后一段时间内可能的表现。

3. 内部一致性信度（internal consistency reliability）

用分平法求信度对学生只需要进行一次测验，即可求得内部一致性信度。（也叫折半信度）。方法是将一次测验的奇数题目和偶数题目分开计分。把这两部分试题看作是两套独立的试卷，然后分别比较，得出两部分得分的相关系数。

在语言测试中用得较为广泛的是 Richardson 的 KR21 公式

$$KR21 = \frac{K}{K-1}\left(1 - \frac{M(K-M)}{KS2}\right)$$

其中 K = 试题总数

M = 试题总分的算术平均分

S = 总分的标准差

信度系数的判断是一项比较复杂的工作，它既要考虑到是主观型考试还是客观型考试，又要顾及用什么方法计算信度系数。一般而言，客观型考试和大规模考试的信度系数高一点才符合要求，而主观型考试和课堂测验的信度要求就低一些。

效度和信度是评估考试质量最重要的指标。它们之间的关系是单向的。测试的信度是指测试结果的稳定性和非偶然性，它要求同一个考试多次测验的结果基本不变。测试的效度指测试的内容要真正符合考前即定的目标，可靠的测试（信度高）可能是有效的（效度高），也可能是无效的（效度低），但不可靠的测试（信度低）必须是无效的（效度低）。

五、后效作用（washback effect）

语言测试的后效作用一般被定义为考试对教学的积极或消极的反馈作用，科学的考试会推动教学，不科学的考试只会起到反面作用。从宏观上而言，外语教学着重培养学生听、说、读、写译等基本技能，如果考试中缺乏考察某一能力的试题，结果就可能会导致教师和学生对这一能力培养的松懈。因此，科学的考试应采取多种题型，多种方法，相结合的方式，以取得正面的后效作用。

总的说来，科学的语言测试对语言教学的各个方面都应起着正面的、积极的、引导作用。

参考文献：

[1]邹申．英语语言测试[C]．上海：上海外语教育出版社，1999.
[2]王孝玲．教育测量[M]．上海：华东师范大学出版社，1989.
[3]Alderson. J. C "Language Test Construction and Evaluation" [J]. 剑桥：Cambridge：Cambridge University Press. 1995.
[4]舒运祥．外语测试的理论与方法[M]．上海：世界图书出版，2001.
[5]黄和斌．外语教学理论与实践[M]．南京：译林出版社，2001.

责任编辑：许有江