2005年6月 湖北经济学院学报(人文社会科学版) Jun 2005
第2卷 第6期 Journal of Hubei University of Economics(Humanities and Social Science) Vol. 2 No. 6

试论语言测试中信度与效度的关系

曹 辉

(中国地质大学 外语系,湖北 武汉 430074)

摘 要:信度与效度是评估语言测试中的及其重要的两个标准。本文从信度的概念、信度的评估,效度的概念、效度的评估,以及现代语言测试的偏向等方面阐述了语言测试中信度与效度的关系。

关键词:信度;效度;评估

测试是教学过程中必不可少的组成部分,它一方面检查学生对知识的掌握情况和教学中存在的问题,对今后的教学提供指导和帮助;另一方面,它也是衡量教学任务和教学大纲执行情况的一种手段,严密科学的试卷能客观地反映出考生的实际水平。怎样才能知道试卷质量呢?这就需要对试卷的质量进行评估和检验。评估的主要标准是什么?一般要看它的信度、效度、难度、区分度、实用性及后效作用,其中信度和效度是它的两个重要标准。信度与效度原是计量学中的两个重要概念,20世纪30年代被引入语言测试领域。20世纪60年代,以LadO等为代表的结构主义测试学家对这两个概念进厅了系统的阐述和论证,标志着语言测试已形成科学的体系,成为一门独立的学科。可以说,语言测试理论及实践它的发展和纷争都是以信度与效度为主线进行的。信度与效度是语言测试永恒的主题。

一、信度

信度也称可靠性。即当被测量对象本身发生变化,用同样的"尺子"去重复测量时,总是获得类似的结果。因此,信度也被称为一致性。如果说一个测试的信度高,便是指一个考生的成绩经反复测试都保持一致。对一组考生来说,便是指这组考生的成绩序列经反复测试都大致相同。如果说某次测试完全可靠,那么便指这次测试排除了一切误差,绝对地准确。即一个考生的成绩经反复测试后完全保持一致,一组考生经反复测试后,其成绩序列完全相同。其实,绝对的准确是不可能的,因为人的因素是不稳定的。信度的评估一般用考试结果的相关系数来表示。相关系数等于1,表示该试卷完全可靠;相关系数等于0,则说明该试卷不可靠。在实践中,人们往往根据具体情况对试卷的信度提出不同的要求。通常是客观题的信度高于主观题,客观题的信度系数一般定在0.90以上。如果一份试卷既有客观题又有主观题(如EPT)信度系数最好不低于0.80。英语测试中,一份好的试卷在词汇、结构和阅读部分的信度系数一般应在0.90-0.99之间;阅读理解部分的信度系数在0.8-0.89之间,口试的信度系数一般在0.70-0.79之间。 对信度系数的评估方法主要有3种:再测信度、平行试卷信度和内部一致信度。

二、效度

(一)效度的概念

考试的效度(有效性)指考试是否检测了它所要检测的东西,是否达到了它所预定的目的。效度是一个相对的概念。这主要是因为个人或群体的语言能力特征,只能通过其行为样本间接推测,而不可能直接测得。由此,推测的结果只能是相对有效,而不是绝对有效。从这个意义上讲,效度是一个程度上的概念,它反映了根据考试分数作出推论或预测的准确性程度。

(二)效度的评估

1. 内容效度

内容效度是指考试的内容是否具有代表性和综合性或者说是否考了应考的内容。一份试卷往往不可能包括所有要考的内容,所以选择内容的方法非常关键。例如:有30个题目是考英语语法的,而有20个题目是考主谓一致性方面的,这就很难完全反映出考生的语法能力。因为英语语法覆盖面很广且包括动词时态、动词短语、介词短语、不定式短语、分词短语等等。所以这样的考试,其内容效度必然很低。

内容效度的确定,一般是命题人员或审题人员对试卷的内容、题目的难易度、区分度等进行严格的分析。

2. 效标关联效度

预测效度是指考试的结果和预言是否有效。例如,要通过考试选拔学生进一步深造,该考试是否选拔了应该选拔的学生,有没有选错或选漏?这就要看考试是否起到了预测作用。一份具有很好的预测效度的试卷,应该能够正确地预测学生未来的行为。有些考试(如分班考试、水平考试)由于与将来的学习有关,所以应特别注意预测效度。预测效度的评估可用计算再测信度所用的Pearson的积距率公式来计算其相关系数。共时效度是用来将新的考试和已经公认的考试作比较,以便证明新的考试的效度。例如较短的时间内(一般不超过两周)让同一组学生参加EPT和TOFEL考试,如果考试结果的相关系数非常高说明两者的相关性很高。如果说TOFEL是一个公认的标准化考试,那么EPT也是一个标准化考试,共时效度的评估仍然可用Pearson的积距率法来计算相关系数。

共时效度与预测效度的关系是:两者都以某种独立的而且可靠的效标作为参照量,把所测试的分数与效标分数作比较,计算其相关程度。共时效度要求相比较的两次考试由同组考生在同一天或时间相隔很近(一般不超过两周)的情况下完成。预测效度则要求相比较的两次考试的间隔是半年、一年,甚至更长时间其测试对象仍然是同一批考生。

3. 结构效度

结构效度(也称实验效度)指一个考试所检测的能力是否符合语言、语言学习和语言行为理论中所假设的能力。这里说的结构是任何关于语言理论中所假设的能力或特征,这些理论可以是心理学方面关于语言学习或语言学习习得的理论,也可以是语言学方面关于阐述语言的本质或语言的功能的理论等。结构效度可以帮助建立心理学或语言学理论中的假设,也可以推翻这些理论中的假设。对于结构效度的评估,统计手段

有7种——检验、因素分析和其他的多元分析，鉴于手段复杂，一般都由计算机进行运算。

三、信度与效度的关系

语言测试以语言能力为测量目标，而语言能力是抽象的通过具体的语言行为体现出来的，任何测试都不必要更不可能测量所有的语言行为，因此语言测试的目的是通过对受试者语言行为样本的测量结果来推测受试者的语言能力。一个语言测试包含两个基本的过程：一是确定能有效说明受试者语言能力的语言或推测行为并在此范围内选取有效的样本，二是保证测试结果真实准确反映受试者的语言行为、语言测试的效度反映所测试的语言行为与语言能力的关系，语言测试的信度说明考试结果与语言行为的关系，没有信度意味着测试结果不是受试者语言行为的真实反映，没有效度只有信度的测试也毫无意义，因为它只是准确地测量了与语言能力不太相关或毫不相关的东西，在此情况下，我们同样无法从考试结果中推测受试者真正的语言能力，要实现一个语言测试的目的，信度与效度缺一不可，这是二者关系的统一性一面。另一方面矛盾和对立构成了二者关系的主要特征，效度要求使语言测试注重语言的整体性、艺术性，信度要求则使语言测试强调语言的科学性。因此，任何测试都难以兼有极高的信度和极高的效度。

四、现代语言测试的偏向

信度和效度的统一与对立的性质决定了同时具有高信度和高效度的语言测试是不存在的，任何语言测试都必须在二者之间进行平衡和折衷。然而这种平衡和折衷在现代语言测试中并没有得到很好的体现，现代语言测试的一个重要偏向是过于注重信度忽视效度，这种偏向的出现主要有以下原因：首先任何一门学科的产生和发展既反映了社会的要求同时也受相关学科以及社会技术进步的影响和制约。70年代以来，语言学、应用语言学、心理语言学、语篇分析及第二语言习得等相关学科的飞速发展，为语言测试注入了丰富的思想内容。在诸多有关语言和语言学习的理论中存在相互矛盾和对立之处，这又从另一方面阻碍了语言测试理论的发展和进步。至于语言测试理论情况更是如此，测试的研究者们往往对效度的论证得出不同甚至相反的结论，比如综合型测试（integrativetests）在效度方面优于离散型测试（Oiler，1979），但也有人经过研究认为在测量结果方面二者没有什么差别（Farhady 1979）。再比如有人质疑用多项选择题（MCQS）测量阅读理论能力的有效性，但也有人认为MCQS能有效地考察阅读理解能力（金艳，吴江，1998）。

现代语言测试重信度轻效度的倾向的主要表现，是测试内容和形式脱离语言运用的实际，重知识轻能力注重领会式技能的考查忽视复用式技能的考察。这种偏向在测试题型上的表现是测试以客观题为主，从而导致多项选择题的泛滥使用，以致于它在相当长的一个时期内似乎成了语言测试的唯一方式（Hughes，1989）。现代语言测试的这种偏向在应试教学（teach to the test）的作用下给外语教学带来了严重的负面影响，阻碍了外语教学培养交际能力这一目标的实现，外语学习者往往经过多年的学习尽管可能以较好的成绩通过考试但其语言实用能力却很低下。

五、语言测试的效度重于信度

语言测试应侧重考虑效度要求，在此基础上尽可能地追求信度。首先，从理论上讲效度是比信度更重要的一个属性，在语言测试中占有中心地位。信度和效度是两个相互排斥的属性，如果必须做出选择的话，效度毕竟更为重要，而信度并不是第一位的。有时为了提高效度而牺牲一定程度的信度是必要的，然而如果为提高信度而牺牲效度，我们的测试就变成了准确测量我们测量目标以外东西的工具（Weir，1990）。其次，从信度与效度的关系来看，语言测试如果首先保证了高信度则必然效度很低或没有效度；反之，如果首先考虑效度，信度虽然会受到一定的损害，但决不是不可获得，我们能够使一个高效度的测试增加一些信度，但我们难以使一个高信度的测试更加有效。

最后也是最重要的，语言测试以效度为主导有助于改善其对外语教学的影响。语言测试给外语教学带来的影响即人们常说的反拨效应，是衡量语言测试的重要标准之一，因此也被许多测试学家称为反拨效度。现代语言测试对效度的忽视在很大程度上也就是对反拨效应的忽视，使语言测试脱离了外语教学。语言测试以效度为主导无疑将给外语教学带来积极的影响，推动外语教学向培养学生实用语言能力方向发展。

要提高语言测试的效度必须设计出能有效反映受试者语言能力的题型并努力使主观题的评分尽量客观化。这方面近年来我国的学者们做了一些有益的探索和研究。如有人主张引入交际测试，有人论证了用听写代替多项选择题来测试听力理解的可行性。当然要提高语言测试尤其是大规模的语言测试的效度无论采用何种可行的测试题型与多项选择题相比都会加重阅卷方面的负担。那么应当如何看待这种负担呢？Hughes（1989）的一段话也许有助于我们的思考。他说采用能产生良好反拨效应的测试方法，采用不能产生良好反拨效应的测试方法使我们负担不起之前，我们必须问自己这样一个问题：语言测试不具有良好反拨效应的代价是什么？那种产生消极影响的测试使教学双方在与真正的教学目标并不相关的活动中浪费了大量的时间和精力，如果比较一下我们为此付出的代价我们就会得出这样的结论：我们真正负担不起的是不使用具有良好反拨效应的测试。

六、结束语

信度与效度是语言测试两大基本要求，信度与效度的关系问题是语言测试的根本问题。从信度与效度等测试标准出发来评价或取舍一种测试模式或测试题型是必要的但还远远不够，更重要的是要考虑它对教学的影响，看它是否有利于教学目标的实现。对此著名学者李筱菊（1997）有过精彩论述：一个语言测试的真正价值不在于它能用多少数据去说明什么，而在于它能否给人们带来好的快乐的结果，有助于使人变得更完美。这好的快乐的结果主要是指良好的反拨效应，如果为通过某一个重要考试，教学双方花费大量时间和精力而学生并没有获得相应的语言能力，这种结果是无论如何不能让人愉快的。注重语言测试的效度，改善其对外语教学的影响是90年代及未来语言测试发展的方向。

参考文献：

[1] Bachman LF. Fundamental Consideration in Language Testing [M]. London: Oxford University Press, 1990.

[2] Butler Christopher. Statistics in Linguistics [M]. London: Basil BlacKWell Ltd Oxford, 1985.

[3] Huges Arthur. Testing for Language Teachers[M]. London: Cambridge University Press, 1989.

[4]李筱菊 语言测试科学与艺术[M] 长沙：湖南教育出版社，1997.