# 语言测试的信度与效度之间的关系

李翌豪
(江苏师范大学 外国语学院,江苏 徐州　221116)

[关键词]　测试信度;测试效度;语言测试

[摘　要]　测试的信度和效度是用来衡量一门考试是否有效和可靠的两个关键因素,任何测试的开发与评估都应当把二者纳入到重点考虑范围之内。然而,一些研究者对于测试信度的定义往往过于理论化,甚至把其成立的基础建立在某些理想化的客观条件之上;同时他们在测量信度时所采取的过于机械化的统计方法,也导致了其结果不能准确地反映出试题的客观稳定性。考虑到信度与效度之间不可避免的逆反关系,为了满足语言测试的首要目的,测试的开发者应当首先给予效度最大限度的重视。由于"套题"中包含了复杂的"题内相关性",所以它并不利于测试在数据上达到令人信服的"可靠性系数"值;但如果以此为依据就盲目地将其排除在外,我们就忽略了其在考查被测能力方面的作用,从而忽视了对测试整体效度的把握。

[中图分类号]　H08　　[文献标识码]　A　　[文章编号]　2095-5170(2016)05-0088-05

## 引言

众所周知,在一门测试的开发过程当中需要考虑诸多因素,尤其是对专门为第二语言或外语学习者所打造的语言能力测试而言,试题开发者对其信度和效度的把握无疑是试题开发当中最关键的两个环节。值得注意的是,测试的信度与效度并不总是以一种自然的、互补互促的状态存在;而且在很多情况下,二者之间还会呈现出一种让试题开发者难以调和的逆反关系,即对某一方的偏重势必会影响到另一方的作用。为了有效地测试出考生的真实语言水平,笔者认为任何形式的语言测试的开发都应当围绕着如何使试题更加有效度这一点来展开;如此一来,在有效地平衡信度与效度之间存在的反比关系的同时,如何保证测试本身的有效性,也就成为试题开发者们需要解决的首要问题。本文将重点讨论一些过往的针对测试信度的纯理论化定义,以及基于统计学方法论的信度测量值是如何使试题开发者们忽视了考试本身的考查效力,进而影响他们开发和选择有助于提高测试效度的题型。

## 一、测试信度的定义

根据美国教育研究协会 1999 年颁布的"教育与心理测试标准"[1](AERA et al. 1999),测试信度指的是测试结果是否稳定可靠,即同一套测试在对同一组测试对象进行的反复测试中,受试者是否可以保持稳定、一致的分数。基于此定义,Chalhoub-Deville 和 Turner (2000)提出了"测量误差"[2]的概念,其中涵盖了除受试者主观因素(即语言测试所考查的目标语言能力)之外的许多可以阻碍其考试发挥的客观因素;随即他们又从考生的实际测试成绩与其"真分数"[3]之间的差距这一角度出发,重新定义了测试信度——前者越接近后者,此测试的信度就越高。由于真分数只是代表了测试中不存在"测量误差"[4]时的真值或客观值(lord, Novick 1968),但是在任何实际的测试中,误差是不可能被完全避免的;所以 Hughes (2003)[5]进一步提出,如果一项测试可以在其实施过程当中将各种客观因素可能造成的测量误差减少到最小程度,并且使受试者发挥其最大潜能,那么此测试的结果就是可靠、可信的。基于笔者的分析,Hughes (2003)[6]对语言类测试的信度也进行了类似的定义:同一组语言试题被多次用

[收稿日期]2016-03-16
[作者简介]李翌豪,男,江苏徐州人,江苏师范大学外国语学院讲师,美国宾夕法尼亚大学教育学硕士。

于考查同一组考生，并且假设在此期间考生的被测语言能力没有明显变化，如果每次的成绩都相近，那么此语言测试便具有较高的信度。

综上所述，对于语言测试的信度来说至关重要的因素共有四点：

1. 同一测试被反复实施后受试者的成绩保持一致。

2. 同一组受试者在考试期间被测语言能力没有显著变化。

3. 测试实施过程当中“测量误差”不存在或已被减少到最小。

4. 测试成绩与“真分数”接近。

如果一个语言测试可以同时满足以上四个前提条件，就足以说明其测试结果的可信性。尽管如此，我们不难看出，想要在同一组考生（其被考查的语言能力在一定时间内还需维持不变）中重复地应用同一项语言测试，在实际操作当中可行性并不高。Hughes(2003)[7]就曾指出，在对同一项语言测试的重复操作当中，两次相邻的测试之间的间隔如果过短，学生对于部分试题的答案就会有较清晰的记忆，况且他们没有充足的时间去实现目标语言能力的提高，从而可能会导致两次测试的结果相差不大，这就意味着此测试的信度值虚高；如果两次测试之间的间隔过长，学生就有足够的时间去完善被测语言的能力，那么后一次的测试成绩可能会远远高于前一次，但如此一来此测试的信度就会大打折扣；因此在上述两种测试条件下，想要完全地抵消由时间安排不同所造成的对信度的影响是很困难的。即使在高新技术的支持下，某些研究可以人为地抹去受试者对于前一次测试的记忆，并且确保他们在参加第二次测试之前没有实现任何新的知识积累或能力提升，从而两次测试的结果十分吻合一致。试问，如此这般得到的测试信度意义何在？这种看似科学、准确的信度值又有什么实际参考价值？

笔者认为，一项语言测试的真正目的，在于其是否可以准确地、及时地为目标语言教师提供关于学生学习状态的反馈信息。如果同一组学生两次或多次参加同一测试的成绩相差巨大，那么这种看似“惨不忍睹”的低信度，至少可以表明他们在此期间突飞猛进或者一落千丈的实际学习效果，从而帮助教师在未来的实际教学当中作出相应的调整和安排。

对语言测试的实施者来说，诸如扰人的噪音、不合标准的测试设备或考试环境等可能会导致“测量误差”的客观不利因素，一定要设法去避免；但是像考试恐惧、紧张不适、难解压力之类的主观不利因素，则完全需要受试者自己在实际考试过程当中进行调节和克服。因此，对于一个被测语言能力突出但是心理素质偏弱的受试者而言，就算此测试在其设计、实施、监督各方面都做到无可挑剔，它也不一定能够真实地反映出此人现阶段的被测语言水平；换言之，抛开客观不利因素，“测量误差”完全可以由主观不利因素引起；而受试者在面对客观上完美无缺的标准化语言考试时所产生的紧张、焦虑等心理障碍，完全可以使其不能够百分之百发挥出应有的水平，甚至发挥失常。由于“真分数”强调的是一种受试者，在不受任何主观和客观不利因素的影响下，完全发挥出自己所有潜能所得到的成绩，那么“真分数”和实际成绩之间的对比可能会帮助我们判断一项测试的信度。尽管如此，在现实的考试环境中，我们每个人都会或多或少地经历各种不同程度的主、客观不利因素；虽然过硬的心理素质并不属于纯语言能力范畴，但是一个受试者在面对“测量误差”时的心态足可以影响其实际的语言应用和发挥。所以，受试者的实际测试成绩综合地反映了其语言能力和心理素质，而这二者对于未来想在任何环境中（考试、交谈、书信……），正确、自信地运用被测语言的受试者来说都是不可或缺的。如果“测量误差”被控制在一个可以接受的范围之内，那么对教师或者高校录取委员会而言，受试者的实际测试成绩可能会比其所谓的“真分数”更加具有参考价值。

虽然“测量误差”和“真分数”等概念可以帮助我们从纯理论角度去衡量一项测试的信度，但是如果试题开发者们过度地强调信度的重要性，并且盲目地把理想化的“真分数”，或者所谓的“零误差”测试环境作为理论依据和评判标准，那么，他们会不可避免地忽视了试题开发的重中之重——测试的效度。

## 二、测试信度的测量

作为心理或教育测验中最常用的信度评估工具，“可靠性系数”[8]这一概念首先由 Cronbach 提出（通常也被称之为“Cronbach 系数”），用于量化测试的信度（Kupermintz，2003）。“可靠性系数”的数值范围通常是最低值 0.00 至最高值 1.00 之间（Gliem and Gliem，2003）[9]，而且根据 Hughes (2003) [10]的进一步解释，除非同一组受试者不

管参加多少次同一项测试其成绩都完全相同，否则此测试不可能真正地得到 1.00 的“可靠性系数”；这就意味着“可靠性系数”能够达到 1.00 的测试在现实当中并不存在，那么一项测试的实际“可靠性系数”值越接近 1.00，其测试可信度就越高。

Hughes(2003，p.39－40)[11]在其研究中主要讨论了三种测量“可靠性系数”的方法：

1)重复测试法(the retesting method)。即用同一套试题在两个不同时间内来测试同一组受试者，这样便可以获得两组分数，然后计算出两组分数的相关系数。当然，Hughes(2003)也指出，想要同时避免考生残留性记忆或其自身能力进步对信度所造成的负面影响是极其困难的。

2)交替形式法(the alternative method)。对同一组受试者使用试题类型完全相同，难易程度相当，但具体题目不同的两套对等试题先后进行两次测试，然后计算出两次成绩的相关系数。

此做法的局限性在于其可操作性不高，并且试题开发的人力与物力成本较高。

3)对半法(the split－half method)。测试在同一时间内进行，只是将整套测试的试题分成对等的两份来分别计分(通常是按单、双数分成两份)，从而通过对两组成绩的对比得出相关的信度系数。由于有效地避开了前两种方法中存在的客观问题，对半法已经成为了目前最常用、最实用的测量“可靠性系数”的方法。

Wainer 和 Lukhele(1997)[12]就曾经在对托福考试的研究中使用了对半法：他们试图通过对具体“可靠性系数”值的分析，将托福考试的整体信度与其阅读和听力部分的信度做了深度对比。

这里有两个概念需要解释一下：

1)套题(testlet)。指“和一项测试的某个具体环节或内容紧密相关的一组试题”[13](Kiely，1987，p.190)；换句话说，由一门考试的某一部分所产生的、并且只跟这一部分考试内容有关的几个或多个试题就构成了一个套题。

2)“题内相关性”[14](local dependence)。指一个“套题”内所有试题之间相互关联、相互依赖的关系。(Yen，1993)

根据以上两点定义，由于听力理解或阅读分析中大部分的题目都源自于某一段音频或某一篇文章，所以托福考试的阅读和听力部分将会包含比其独立的语法或词汇试题更多的“套题”和“题内相关性”(Wainer and Lukhele，1997)[15]。鉴于托福考试中属于不同部分或不同“套题”的题目之间复杂的关系，在使用对半法衡量其信度时，尤其要注意的是：原本属于同一个“套题”的试题，必须被同时分配到两份试卷的某一份当中，否则整套测试的信度值就会虚高(Thorndike，1951)[16]；同时，由具有高度“题内相关性”的多个试题所组成的某个“套题”，也应当被作为一个不可分的独立题来对待(APA's 1966)[17]。因此，某一个“套题”在对半法中可以被随机分配到任何一份试卷里，但是其内部所有题目不能被分开而同时存在于两份试卷当中；否则考生对同一个“套题”内容的正确或错误理解，将会以相似程度的得分或失分呈现出来，如果同时很多个“套题”中的子题都被分散到两份试卷当中，那么，不论考生的分数高与低，其两份试卷的成绩可能会相当接近，从而导致了看似一致，实则虚高测试信度。

Wainer 和 Lukhele (1997)[18]在他们的研究中也发现了类似的问题：由于托福大多数题目之间存在着或多或少的“题内相关性”，所以把两份试卷中的所有题目都当成是独立题的做法，无异于自欺欺人，而且对半法无法真正做到将所有“套题”完整地分配，所以实验所得出的高达 0.95 的“可靠性系数”值，并不具有很高的说服力和参考价值。基于上述分析，虽然二人没有明确地赞成用更多的独立题来取代“套题”以提高信度系数值的可靠性这一做法，但是他们为语言测试的开发者们指出了“套题”的三个主要缺陷：

1)一项测试所包含的“套题”越多，其题与题之间的“题内相关性”就越复杂，那么这项测试的整体信度系数值就极容易虚高。

2)由于缺少相关的背景内容(如文章或音频)，就题目本身而言，每一个独立题(如词汇、语法)都比“套题”中的每一个子题带有更多的考试信息。

3)同等条件下，用较少的独立题可以实现用较多的“套题”才能够达到的信度系数标准，这也体现了前者在节约测试开发成本方面的优势。

不可否认，拥有多重“题内相关性”的“套题”，往往不能够帮助一项测试实现纯数据上的稳定，但是没有任何证据可以否认其在反映考生真实语言水平方面的作用。同样，拥有更多考试信息或答题线索的独立题，也不一定比“套题”中的子题更准确、更可靠、更具考查效力。

笔者认为，任何为了得到可靠的信度系数值而盲目地减少“套题”在一项语言测试中所占比例

的做法都是错误的：实际上作为一种复合类题型，“套题”及其相关的子题有时更能够准确地衡量考生综合运用被测语言的能力。例如，对某个单词词义的把握，通常就可以使考生在相关的词汇题上游刃有余，但这不足以确保其在阅读或听力部分依然可以信步闲庭：他还需具备在复杂的句式中判断出此单词句法成分的眼力；在抽象的上下文语境中领悟出其隐涵语义的功力；以及在听力对话中辨识出其口语特征的反应力等等。除此之外，因为其大多数答题线索或答案都已经被包含在阅读段落或听力音频里，所以“套题”的子题本身也不必带有过多的答题信息，而且每一个子题也只需和“套题”的部分背景内容有关。因此，独立的词汇题型并不能够全面、深入地为阅卷者提供关于考生综合语言素质的信息；它虽然有助于数据化的测试信度，但着实有碍于实践中的测试效度。

## 三、信度与效度一到底谁更重要？

鉴于之前的分析与讨论，我们不难看出，测试信度的定义与测量更多强调的是一种基于统计学方法论的量化研究。然而，这种对纯数据化信度的过分追求必将导致测试开发者们忽略了对效度的把握。依据 Hughes (2003)[19] 的阐述，对于语言测试而言，效度是指其是否可以有效、准确地考查和反映受试者的目标语言能力。很明显，单从定义上来看，效度比信度更直接地关乎于测试的本质，即其是否测试了它要测试的内容以及对应测试的内容所测试的程度。所以，测试效度应该被作为开发和评价某一项测试最根本、最重要的依据(AERA et al., 1999, p. 9)[20]。事实上，信度的维持和效度的优化，总会引起二者之间难以调和的矛盾：如果以保证数据上的信度稳定为目的，那么实际效度将不可避免地被削弱，反之亦然；“套题”被独立题所取代可以仅仅是为了一个稳定的信度系数值，而大范围使用“套题”虽有助于提高测试的综合考查力度，但如此一来其“可靠性系数”将不再有说服力。

透过二者之间此消彼长的矛盾，我们应当清楚地认识到，虽然“可靠性系数”可以量化信度，使其更直观地被人们所理解，但是试题的设计、开发、目的与功效是不可以简单地用数学公式来计算的。语言测试的目的永远只能是考查受试者目标语言的能力，为了实现这一目标，任何存在于信度与效度之间的分歧，都应当从维护后者的角度出发去解决。正如 Hughes(2003)[21] 所概括的那样，一个有效的测试永远都是可靠的，但一个可靠的测试可能仅仅实现了统计数据上的稳定而无法真正地体现实际的考查效果。因此，除了将客观“测量误差”减少到尽可能小的程度之外，其他任何跟信度有关的问题，都应当作为试题开发者的次要考虑，而他们的主要精力应该始终放在如何保证测试的效度上。

## 四、结论

综上所述，我们确实应当把信度和效度作为测试开发环节中重点考虑的因素。特别是对于考查第二语言或外语能力的语言测试来说，关键就在于如何以保证其对目标语言的考查效力为出发点，有的放矢地调解信度与效度之间的矛盾。一项测试可靠与否，必须以其是否可以如实地反映受试者的纯语言能力(听、说、读、写)为唯一标杆；任何试图重复或分解试题，从而机械地量化其可靠性的做法都是不可取的。

当然，我们也不能一味地贬低信度的价值；况且，我们可以通过对具体信度值的分析来更直观、更便捷地了解一项测试，是否可以使受试者稳定、自信、有效地发挥其被测能力。虽然某些有利于提高效度的题型会有碍于我们得出一个令人满意的信度系数值，但是一个纯数据上可靠的测试并不一定能够帮助老师真正地了解考生的实际学习情况；而像“套题”这一类需要考生运用综合语言能力的试题，由于其内部多变的题型和复杂的“题内相关性”，致使其无法顺应特定的统计规律，从而拉低或夸大了测试整体的“可靠性系数”。对于语言测试的开发者来说，二者之间某一方的加强势必导致另一方的衰减，此时应当把效度放在首位，尽可能地将测试打造成衡量被测语言能力的利器；任何偏离此方向的做法，都会使语言测试本身失去其应有的意义。

## [参 考 文 献]

[1][20] AERA, APA, NCME, 1999. Standards for Educational and Psychological Testing. , American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Washington, DC.

[2][3] Chalhoub－Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exam,

IELTS, and TOEFL. System,28, 523—539.

[4](美)罗德(Lord,Frederic M.)、(美)诺维克(Novick,Melvin R.)(1968):《心理测验分数的统计理论》,叶佩华译,1992 年版,第 624 页。

[5][6][7][9][10][11][19][21]Hughes, A. (2003). Testing for language teachers. Cambridge: Cambridge University Press.

[8]Kupermintz, H. (2003). Lee J. Cronbach's contributions to educational psychology. In B. J. Zimmerman and D. H. Schunk (Eds.). Educational psychology: A century of contributions, pp. 289—302. Mahwah, NJ, US: Erlbaum.

[9]Gliem, J. A., & Gliem, R. R. (2003). Calculating, Interpreting and Reporting Cronbach's Alpha Reliability Coefficient for Likert—Type Scales. Presented at the Midwest Research—to—Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8—10,2003.

[12][15][18]Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? Educational and Psychological Measurement, 57, 741—758.

[13]Wainer, H., & Kiely, G. (1987). Items clusters and computerized adaptive testing: A case for testlet. Journal of Educational Measurement, 24, 185—202.

[14]Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187—213.

[16]Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), Educational measurement. Washington, DC: American Council on Education.

[17]American Psychological Association (APA). (1966). Standards for educational and psychological tests and manuals. Washington, DC: Author.

# The Relationship between Reliability and Validity of Language Test

LI Yi-hao

(*School of Foreign Studies, Jiangsu Normal University, Xuzhou* 221116, *China*)

**Keywords**: Reliability; validity; Language Test

**Abstract**: There are lots of concerns involved in developing a test, especially a language test for second and foreign language learners, but the two most important ones that any test developer should take into consideration are reliability and validity. However, reliability and validity are not always mutually contributive to each other, and sometimes even an inverse relationship could exist between them. Aiming at assessing and reflecting the true language ability of test—takers, I think any test—developing activities should be organized around how to make the test more valid, in other words, how to resolve the potential tension between reliability and validity and ensure the latter one at the same time. My article will discuss how certain assumptions and suggestions underlying the theoretical definition and statistical measurement of reliability might distract test developers' attention away from validity and negatively affect their decision concerning the selection and construction of test items conducive to validity.

[责任编辑:江奎元]