北京语言大学汉语水平考试中心学术文库

语言测试与信息研究

柴省三 著

HSK考试管理信息化及基于统计的考试质量控制
从『试』偶拾
现状与对策——汉语作为第二语言的教学研究
汉语水平考试建设和计算机辅助教育
汉语测试初学集
汉语作为第二语言的测试研究
语言测试与信息研究
问学录——汉语测试与话语分析探究
张凯自选集
汉语测试探微
HSK与语言问题
汉语测试与句法研究探微

北京语言大学汉语水平考试中心学术文库

语言测试与信息研究

Language Testing and Information Research

柴省三 著

北京语言大学出版社
BEIJING LANGUAGE AND CULTURE
UNIVERSITY PRESS

图书在版编目(CIP)数据

语言测试与信息研究：汉、英／柴省三著．—北京：北京语言大学出版社，2011.4
（北京语言大学汉语水平考试中心学术文库）
ISBN 978-7-5619-3003-8

Ⅰ.①语… Ⅱ.①柴… Ⅲ.①语言－测试－文集－汉、英 ②信息语言－文集－汉、英 Ⅳ.①H09-53②H087-53

中国版本图书馆 CIP 数据核字（2011）第 061186 号

书　　名：语言测试与信息研究
责任印制：汪学发

出版发行：北京语言大学出版社
社　　址：北京市海淀区学院路 15 号　　邮政编码：100083
网　　址：www.blcup.com
电　　话：发行部　82303648/3591/3651
　　编辑部　82303647
　　读者服务部　82303653/3908
　　网上订购电话　82303668
　　客户服务信箱　service@blcup.net
印　　刷：北京联兴盛业印刷股份有限公司
经　　销：全国新华书店

版　　次：2011 年 4 月第 1 版　2011 年 4 月第 1 次印刷
开　　本：880 毫米×1230 毫米　1/32　印张：4.625
字　　数：138 千字
书　　号：ISBN 978-7-5619-3003-8/H·11042
定　　价：18.00 元

目　录

上编　语言测试

下编　信息研究

上编　语言测试

关于 HSK（初、中等）平行信度的实证研究[①]

一　引言

汉语水平考试（HSK）的测试目的是对母语非汉语者（包括外国人、华侨和国内少数民族考生）的汉语水平和能力进行评估，测试内容和测试方法以能够有效地反映被试实际具有的汉语水平为原则。与学业成就测验（achievement tests）、诊断性测验（diagnostic tests）等一般课堂教学测验和选拔性测验不同，汉语水平考试（HSK）不以特定的教材、教学大纲和某一具体教学方法为依据，也不考虑被试先前接受过何种汉语训练，而是以准确、有效地鉴定被试是否具备完成某项特定任务应该达到的汉语水平为原则。HSK 已经成为国内不少普通高等院校录取外国留学生的依据之一。因此，HSK 的信度如何，将直接决定对被试所作的各种评价和决策是否可靠、公平。本文拟对 HSK（初、中等）的平行信度（parallel reliability）进行理论探讨和实证分析。

二　平行信度的制约因素

根据对测验结果进行解释所依据的参照体系的不同，语言测试可

① 原载于《汉语学习》，2002 年第 2 期。

分为常模参照性测验（norm-referenced）和标准参照性测验（criterion-referenced）两种。常模参照性语言测验侧重于对被试团体中的个体之间进行比较，适用于选拔功能；而标准参照性语言测验则更注重将被试的测验结果与事先确定的参照标准进行对比，判断被试是否达到某种标准或达到标准的具体程度如何。HSK 是兼具常模参照性质的标准参照性水平测验（刘英林，1994）。尽管 HSK（初、中等）的考试结果也可以用于对被试的汉语水平进行对比、区分，但是 HSK（初、中等）的标准参照性更强，它有独立的考试大纲，用统一的汉语水平参照点对 1 至 8 级汉语水平进行了明确的界定，因此这种参照体系更适合于通过将被试的测验结果与相应的汉语水平等级进行比较而作出相关的评价结论。

由于 HSK（初、中等）属于标准参照性测验，因此它必须要有较高的平行信度，既要证明达到同样分数标准的不同被试具有相同的汉语水平，又要确保汉语水平相对稳定的同一个被试参加不同复本的测验时，其结果要有很好的稳定性、一致性。考查 HSK 平行信度高低的标准有两个，一是被试在测验复本上的考试总分的一致性如何；二是被试在测验复本上的分数结构性质（configuration）是否相同，证书的等级是否一样。在不考虑记忆和练习效应（practice effects）的前提下，如果一个被试正常参加测验复本 A 卷和 B 卷的考试结果基本一致，那么测验的平行信度就高，反之，如果被试在两个测验复本上的考试结果相差太大，那么测验的平行信度就比较低，由此根据某一次测验的结果对被试语言水平所作的任何评价和社会决策就有失公允。

1995 年，前国家教委规定（教外来［1995］668 号）：凡申请进入中国普通高等院校接受本科教育的外国人，均须参加 HSK 考试，并获得相应的最低合格等级的汉语水平证书。显然，HSK 不仅被国内高校作为录取外国留学生的统一标准，而且录取的参照标准是以证书等级而不是以 HSK 的总分为依据，因此在评估 HSK 的平行信度时，不仅要重视被试在复本测验上的考试分数信度，更要侧重对被试在复本测验上所获得的汉语水平证书等级一致性的考查。影响 HSK 平行信度的因素包括系统因素和偶然因素两种，其中最主要的系统因素有以下几个方面。

2.1 命题质量

命题质量的高低不仅影响测验的效度（validity），而且影响测验的信度。高质量的命题必须保证不同测验复本之间具有相同的被测因子（measured factors），复本试卷之间的总体难度（power）、区分度（discrimination）等指标基本相同，考试的构念（construct）科学合理，只有这样才能确保测验具有较高的同质性信度（homogeneity reliability）和复本信度。不过，在实践中要命制表面形式相同、试题难度和区分度完全相同的平行试卷显然是不现实的，因为无论命题的质量有多高，试卷之间的难度和区分度等指标总是有差别的，因此还要借助心理测量学手段对试卷进行等值处理（郑日昌等，1990）。

2.2 等值效果

所谓测验等值（equating）就是对测验之间进行单位系统的转换（郑日昌等，1990）。由于命题时很难保证平行试卷之间在项目难度、区分度、信度和分数分布之间完全相同，因此，我们对 HSK 的平行试卷进行了“铆题”式等值处理，以尽可能弥补因试卷复本之间质量指标差异而造成的信度损失，确保语言水平相同的两个被试在参加两个不同测验复本的考试时获得的 HSK 总分和证书等级基本相同。由于等值处理的方法很多（谢小庆，1998），而且在等值处理过程中也存在误差，所以，等值处理的方法和等值效果的好坏也必然影响测验的平行信度。

2.3 分数体系的界定

分数体系界定的核心是考试结果的表征（representation）是否合理、科学，即：不同的考试分数与相应的汉语水平等级之间的对应关

系是否有可靠的实证基础和理论基础作保证。HSK（初、中等）的分数体系是由8级构成的多等级体系，不同等级的证书和分数及相应的汉语水平之间有科学的界定，这种分数体系的可靠性和稳定性如何，无疑会制约HSK的平行信度。

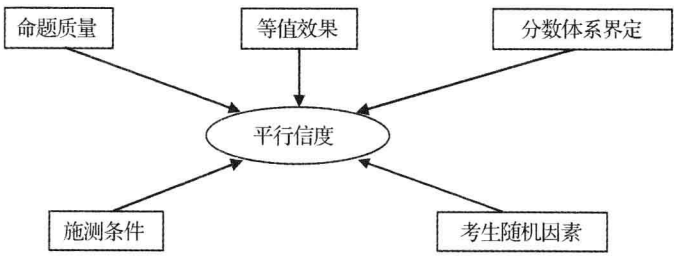

图1　平行信度制约因素示意图

除命题质量、等值方法和分数体系等系统因素以外，影响平行信度的因素还有被试自身和施测条件等偶然因素（见图1）。为了评估标准参照性测验的平行信度，国外许多心理测量专家在经典测验理论（CTT）体系下，提出了许多统计量来评价标准参照性语言测验的平行信度，这些方法主要包括：（1）测验分数平行信度；（2）分类一致性信度；（3）领域分数估计值信度。本文将采用前两种方法对HSK（初、中等）的平行信度进行实证分析。

三　HSK（初、中等）平行信度分析

3.1　研究样本

由于平行信度反映的是同一组被试在两个测验复本之间考试结果

的一致性程度，因此，在用复本方法估计信度时，两个测验必须连续施测，而且两次测验的时间间隔要适当（Bachman，1990）。对于HSK（初、中等）而言，由于测验复本之间采用了“铆题”进行等值处理，所以施测时间间隔的选择非常重要，如果间隔太短，被试对“铆题”的记忆效应就会干扰信度估计值，如果施测间隔太长，则被试的练习效应同样会影响考试的结果。国外学者（Nunnally，1964）建议两次施测的间隔一般以两周为宜，这样的时间间隔既可以有效排除记忆效应又可以将练习效应降低到最低限度。因此，本文以152名考生（样本构成见表1）为研究样本，对J325卷和J323卷的平行信度进行研究。两个测验复本在题型、长度、时限和指令等方面完全相同，两次考试的时间分别为2001年6月24日和2001年7月8日（间隔正好两周），两次考试均属于正规的证书考试，因此考生的动机比较明确，基本可以忽略考生主观因素对考试结果的影响。

表1　被试样本组成

国籍 / 性别	韩国	日本	印尼	越南	瑞士	合计
男	24	10	1	0	1	36
女	94	17	4	1	0	116
合计	118	27	5	1	1	152

3.2　同质性信度

测验的同质性信度（homogeneity reliability）是指同一个测验内部所有项目（items）之间的一致性程度，它反映了所有项目的测试目标是否明确、一致，被测的因子是否稳定统一。较高的同质性信度是语言水平考试的基本要求之一。对于高水平的测验而言，只有每个测验复本都具有较高的同质性信度，才有可能保证复本之间具有较高的平行信度。如果测验的同质性信度较差，那么测验的平行信度肯定也不会太高。由于HSK（初、中等）采用的是一题一分制记分方法，所以

我们采用库德—理查逊（Kuder-Richardson）提出的 K－R21 公式来分别估计 J323 卷和 J325 卷的听力理解、语法结构、阅读理解、综合填空及总分的同质性信度系数（具体结果见表 2）。

表 2　J323 卷和 J325 卷的 K－R21 系数

类别	听力理解	语法结构	阅读理解	综合填空	总分
J323 卷（7 月 8 日）	0.896	0.823	0.883	0.848	0.959
J325 卷（6 月 24 日）	0.847	0.806	0.845	0.831	0.948

从表 2 中的 K－R21 值可以发现，无论是 J323 卷还是 J325 卷，其听力理解、语法结构、阅读理解和综合填空的同质性信度系数都比较高，K－R21 系数均在 0.800 以上，而整卷的 K－R21 系数则分别达到了 0.948（J325 卷）和 0.959（J323 卷）。一般认为，客观性语言测试的同质性信度系数超过 0.90 就比较理想（Lado，1961），因此，J323 卷和 J325 卷的同质性信度比较高，说明 HSK（初、中等）的测试目标比较统一，测验项目之间具有很高的一致性。

3.3　测验分数平行信度

测验分数平行信度通常用被试在两个复本测验上所得分数之间的相关系数来表示，同一组被试在两个复本测验分数之间的相关系数越高，平行信度就越高。表 3 是 152 名被试在 J323 卷和 J325 卷上各部分原始分（raw scores）之间的相关矩阵，表 4 是被试在两卷上各部分标准分（standardized scores）之间的相关矩阵。通过对表 3 和表 4 的对比可以发现，无论是原始分之间还是标准分之间，被试在 J323 卷和 J325 卷上各部分的考试分数之间均显著相关（＊表示 0.01 水平显著），说明 HSK 在命题过程中对难度、区分度等指标的控制比较好，测验分数的平行信度较高；其次，被试在 J323 卷和 J325 卷各部分标准分数之间

的相关系数普遍高于原始分数之间的相关系数，尤其是总分之间的相关性差别更加明显，原始总分之间的相关系数为 0.881，而标准总分之间的相关系数则达到了 0.903，证明对 HSK（初、中等）平行试卷之间进行等值处理的结果有效地提高了测验分数的平行信度。

表 3　原始分相关矩阵

类别		J323 卷（7 月 8 日）				
		听力理解	语法结构	阅读理解	综合填空	总分
J325 卷（6 月 24 日）	听力理解	0.695 *	0.616 *	0.582 *	0.566 *	0.713 *
	语法结构	0.664 *	0.700 *	0.651 *	0.688 *	0.776 *
	阅读理解	0.585 *	0.614 *	0.754 *	0.664 *	0.763 *
	综合填空	0.537 *	0.583 *	0.712 *	0.764 *	0.749 *
	总分	0.733 *	0.735 *	0.794 *	0.780 *	0.881 *

表 4　标准分相关矩阵

类别		J323 卷（7 月 8 日）				
		听力理解	语法结构	阅读理解	综合填空	总分
J325 卷（6 月 24 日）	听力理解	0.721 *	0.619 *	0.579 *	0.570 *	0.731 *
	语法结构	0.669 *	0.718 *	0.676 *	0.697 *	0.801 *
	阅读理解	0.574 *	0.626 *	0.775 *	0.669 *	0.780 *
	综合填空	0.529 *	0.598 *	0.742 *	0.773 *	0.770 *
	总分	0.734 *	0.746 *	0.815 *	0.788 *	0.903 *

3.4　分类一致性信度

分类一致性信度是指用两个不同的测验复本对同一组被试进行施

测时，根据测验结果和相应的参照标准对被试语言水平进行分类的一致性程度（Hambleton，1973）。对于多等级制的标准参照性测验而言，分类一致性信度更具有说服力。分类一致性信度与测验分数信度不同，它对被试的具体分数并不敏感，而更注重分类结果的稳定性。评价分类一致性信度的指标有 P_0 指标和 K 指标两个，在计算 P_0 指标和 K 指标时不涉及被试在复本测验中的具体分数，而只涉及被试在两个复本测验中所获得的证书等级是否相同。表 5 是被试在 J323 卷和 J325 卷上获证等级的基本情况。

表 5　两次考试获证情况

J325 卷 \ J323 卷		7 月 8 日（J323 卷）								
		未获证	初等证书			中等证书			合计	等级相同人数
		1－2 级	初 C	初 B	初 A	中 C	中 B	中 A		
未获证	1－2 级	0	2	1	0	0	0	0	3	0
初等证书	初 C	1	3	4	0	0	0	0	8	3
	初 B	0	2	11	5	4	0	0	22	11
	初 A	0	0	3	8	12	1	0	25	8
中等证书	中 C	0	0	0	8	28	19	1	56	28
	中 B	0	0	0	0	6	20	12	38	20
	中 A	0	0	0	0	0	0	1	1	1
合计		1	7	19	21	50	40	14	152	71

P_0 指标和 K 指标是由 Swaminathan 和 Hambleton 等人提出的。这两个指标是用来衡量标准参照性测验分类一致性信度的指标，P_0 指标的基本含义是两个测验复本对被试一致区分的比例，而 K 指标则是指除去偶然因素外被一致区分的受试者比例。根据表 5 中的数据，我们可以求出 J323 卷和 J325 卷的 P_0 指标为 0.47，而 K 指标为 0.32。[①]

由于 K 指标排除了 P_0 指标中因偶然因素所造成的一致性区分，因而 K 指标的值与 P_0 指标相比更小，K 指标反映了在总的一致性区分中

① P_0 指标和 K 指标的具体计算方法见张厚粲、刘昕（1992）和 Hambleton & Novice（1973）。

真正由测验所作的区分一致性。尽管研究样本在 J323 卷和 J325 卷上的分数平行信度较高，但从 P_0 指标和 K 指标来看，测验的分类一致性信度偏低，说明被试在测验复本上所获证书等级的稳定性还不尽理想，具体原因还有待于进一步的实证性分析。

四　结论和建议

测验分数平行信度和分类一致性信度是从两个不同角度评价多等级标准参照性测验信度的两个基本指标，前者反映了被试在测验复本上得分的一致性；后者则反映了被试在平行测验上的考试结果与相应的等级标准相比较而所作决策的一致性。因此，一个高质量的标准参照性测验应该既要有较高的分数平行信度，又要追求较高的分类一致性信度。本文实证分析的结果表明：

（1）HSK（初、中等）的同质性信度、复本测验分数的平行信度均比较高，J323 卷和 J325 卷的同质性信度系数（K－R21 系数）分别达到了 0.959 和 0.948，被试在 J323 卷和 J325 卷上总分之间的相关系数在 0.900 以上，说明 HSK（初、中等）的命题质量和等值效果均比较好。

（2）HSK（初、中等）的分类一致性信度指标偏低，P_0 指标和 K 指标分别只有 0.47 和 0.32，对于一个大规模的标准化考试而言，HSK（初、中等）的分类一致性信度还有较大的提高余地。

（3）建议将 HSK（初、中等）的 8 级标准进行相应的合并或组合，或者在进一步深入研究的基础上将 HSK（初、中等）拆分为“初等汉语水平考试”和“中等汉语水平考试”，分别对不同水平的被试进行施测，以减少因等级划分过密而造成的分数等级之间临界“敏感区域”过多及其对证书平行信度的干扰。

参考文献

桂诗春（1986）《标准化考试——理论与操作》，广州：广东教育出版社。

刘润清、韩宝成（2000）《语言测试和它的方法》，北京：外语教学与研究出版社。

刘英林（1994）《汉语水平考试研究（续集）》，北京：现代出版社。

谢小庆（1998）关于HSK等值的试验研究，《世界汉语教学》第3期。

张厚粲、刘　昕（1992）《考试改革与标准参照测验》，沈阳：辽宁教育出版社。

郑日昌、漆书清、马世晔（1990）《考试的教育测量学基础》，北京：高等教育出版社。

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. & A. Palmer (1990) *Language Testing in Practice*. Oxford: Oxford University Press.

Carmines, E. G. & R. A. Zeller (1979) *Reliability and Validity Assessment*. Beverly Hills, CA: Sage Publications.

Davies, A. (1990) *Principles of Language Testing*. Oxford: Oxford University Press.

Hambleton, R. K. & M. R. Novice (1973) Toward an Integration of Theory and Method for Criterion-Referenced Tests. *Journal of Educational Measurement*, 10: 159-170.

Lado, R. (1961) *Language Testing, the Construction and Use of Language Tests*. London: Longman.

McNamara, T. (1996) *Measuring Second Language Performance*. London and New York: Addison Wesley Longman Limited.

Nunnally, J. C. (1964) *Educational Measurement and Evaluation*. New York: McGraw-Hill.

Spolsky, B. (1995) *Measured Words*. Oxford: Oxford University Press.

HSK 的获证信度及猜测概率分析[①]

一 引言

信度（reliability）是反映测试稳定性（stability）、一致性（consistency）和可靠性（dependability）的重要指标之一。所谓考试的信度是指同一个测验（或测验形式相等的两个或多个平行测验）对同一组考生施测两次或多次，其结果的稳定性和一致性程度。也就是说，同一个考生如果在“适当短”的时段内多次参加某种测验，其考试结果的波动性不大，每次得到近乎一样的分数，那么，就可以认为该测验的信度是高的；反之，如果考试结果忽高忽低，起伏不定，则说明该测验的可靠性不够、信度不高。信度和效度（validity）是语言测试中衡量考试质量高低的重要指标。任何考试，只有可靠、可信，方能有效。信度高是效度高的必要前提。

在语言测试中，影响考试信度的因素是多方面的，而对考试信度高低的评价往往是通过对考试分数的评估进行的。Bachman（1990）将语言测试中影响考试分数的各因素表示为下面的“路径图”（path diagram）。语言测试中的考试分数是由考试本身要测量的“交际语言能力”（communicative language ability）、考试所要测量的交际语言能力以外的“个人特质”（personal attributes）、“测试方法层面”因素（test method facets）和考试中的“随机因素”（random factors）协同

① 原载于《考试研究》，2001 年第 4 期。