汉语测试、习得与认知探索

张旺熹　邢红兵
主编

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

汉语测试、习得与认知探索

『首届语言测试与习得研究生学术论坛』论文选

张旺熹　邢红兵
主编

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

图书在版编目(CIP)数据

汉语测试、习得与认知探索 / 张旺熹，邢红兵主编．—北京：北京语言大学出版社，2010.2
ISBN 978-7-5619-2648-2

Ⅰ.①汉… Ⅱ.①张…②邢… Ⅲ.①汉语—语言学—文集 Ⅳ.①H1-53

中国版本图书馆 CIP 数据核字（2010）第 015508 号

书　　名：汉语测试、习得与认知探索
中文编辑：徐　雁
英文编辑：望　震
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　82303223
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：北京中科印刷有限公司
经　　销：全国新华书店

版　　次：2010 年 3 月第 1 版　2010 年 3 月第 1 次印刷
开　　本：850 毫米×1168 毫米　1/32　**印张**：9.625
字　　数：276 千字
书　　号：ISBN 978-7-5619-2648-2/H·10002
定　　价：32.00 元

凡有印装质量问题，本社负责调换。电话：82303590

前　言

2008 年 11 月 30 日，由北京语言大学汉语水平考试中心和北京语言大学研究生部联合举办的“首届语言测试与习得研究生学术论坛”在北京语言大学召开。本次论坛共收到北京语言大学汉语水平考试中心、北京语言大学对外汉语研究中心等单位以及北京师范大学等院校研究生提交的学术论文 34 篇，研究领域涉及汉语测试研究、汉语习得研究、汉语认知研究三个方面。

本次论坛提交的论文具有开放性的特点，涉及的研究领域广，论文起点高。为了让更多的研究者了解该领域研究生的研究状况，鼓励研究生发表自己的科研成果，提高研究生学术论文的写作水平，论坛结束后，我们组织了论文评审委员会，对提交会议的全部论文进行了匿名评审。最后筛选出 18 篇论文结集出版。入选论文包括语言测试研究论文 6 篇，语言习得研究论文 8 篇，语言认知研究论文 4 篇。论文集定名为《汉语测试、习得与认知探索》。

本论文集编委会的各位老师都是汉语水平考试中心的研究生导师，参与了论坛的组织、论文的指导及论文评审等各个环节的工作，在入选论文的修改方面也提出了很多具体的建议。这些辛勤的劳动，大大提高了研究生的科研能力和论文写作水平，对今后的研究生科研工作也有着重要的指导意义。

汉语水平考试中心研究生管理办公室李郁老师在论坛的组织和论文的收集过程中做了大量的工作。汉语水平考试中心硕士研究生刘慧芳为论文初稿的编辑和修改做了大量的基础性工作。王正刚先生为本书做了细致的编辑加工工作。感谢北京语言大学出版社为我们提供了论文集出版的宝贵机会。

谨对上述各方面人员为本论文集的出版所作的贡献表示衷心的感谢。

北京语言大学研究生部
北京语言大学汉语水平考试中心
2009 年 8 月

目 录

HSK（高等）主观性考试评分员间一致性检验 …… 刘婷雁（1）
实用汉语水平认定考试（C. TEST）［A-D］级题目
内部结构效度检验 …… 郭兴燕（27）
C. TEST（A-D）级造句题试题难度和应试者应答错误分析
…… 张淑男（44）
对 Bachman 语言交际能力模型中“graphology”一词的修正
…… 周潇潇（65）
IRT 参数估计方法综述 …… 卢恩玲（78）
Lyle F. Bachman 真实性研究述评 …… 周 聪（89）

外国学生汉语“使”字句习得考察 …… 梁婷婷（109）
基于“HSK 动态作文语料库”的韩国学习者“有”
字句习得偏误分析…… 张 颖（121）
日本学生汉语动宾式离合词习得偏误及原因分析
…… 朱旻文（134）
“可见”等十个篇章连接成分在 HSK（高等）作文
考试中的使用情况考察…… 陈煌煌（160）

华裔和非华裔高级汉语学习者阅读中伴随性词汇
习得策略使用的实验研究…… 李　昱（175）
不同教学方式对汉字书写习得的短时个案动态研究
…… 张海威　黄小雨（190）
对语言测试与二语习得在语言能力结构问题上的
研究接口可能性的一些思考…… 刘思维（204）
第二语言习得交际策略研究述评…… 付一鸣（217）

位移事件与“去 VP”结构中“去”的隐喻拓展
…… 李慧敏（236）
“A 把 B-VP”（B∈A）类“把”字句位移图式简析
…… 乌云赛娜（248）
说“相当” …… 李佳琳（258）
量词“口、把、眼、头”认知理据考释 …… 郝瑜鑫（278）

HSK（高等）主观性考试评分员间一致性检验

北京语言大学汉语水平考试中心　刘婷雁

提　要　本研究采用相关分析、频次分析等统计方法，对 HSK（高等）主观性考试的评分员间一致性作了一次较为全面的检验。检验结果显示：1. HSK（高等）主观性考试具有较高的评分员间一致性。2. HSK（高等）主观性考试评分员对评分等级的把握总体良好，但对中间评分等级的把握差异较大，且评分呈现出较高趋中性。3. 评分员间严厉度的一致性是影响主观性考试评分员间一致性的因素之一。

关键词　HSK（高等）主观性考试　评分员间一致性　相关系数

一　引言

1.1 理论基础

主观性考试是指评分员必须根据自己对评分标准的主观解释来对考生的反应作出主观判断的考试（Bachman，1990）。在主观性考试的评分过程中存在一定的主观性，不同评分员评出的分数往往会出现差异，导致评分误差。“评分员误差是主观性考试的一个主要误差来源。”（Cooper，1984；Wigglesworth，1993）因此，如何提高评分员间一致性（即评分员间信度）是保证主观性考试质量的一个重要方面。

所谓“信度”（reliability），也叫可靠性，是指测验分数的稳定性和一致性程度（张凯，2002）。根据测验分数误差的来源，信度可分为内在一致性信度（internal consistency reliability）、再测信度（test-retest reliability）、复本信度（parallel-forms reliability）和

评分员信度（rater reliability）（Bachman，1990；王孝玲，2004）。评分员信度是指评分员两次或多次评估同样试卷所得结果相对稳定的程度（周胜，1997）。评分员信度又可进一步分为评分员内信度（intra-rater reliability）和评分员间信度（inter-rater reliability）（Henning，1987；Bachman，1990；王孝玲，2004）。评分员内信度所考察的是同一评分员内部的一致性，因此，也称评分员内一致性；而评分员间信度所考察的是不同评分员间的一致性，因此，也称评分员间一致性。本研究主要关注的是 HSK（高等）主观性考试的评分员间一致性问题。

1.2 关于评分员间一致性的研究

主观性考试评分员间一致性问题，很早就引起了语言测试研究者的重视，在国外许多研究者的调查研究中都有所反映。例如：Diederich、French 和 Carlton（1961）请 53 名评分员按 9 个等级对 300 份大一新生的作文试卷进行评分，94% 的试卷得到了至少 7 个等级的成绩，没有一份试卷得到的成绩少于 5 个等级（转引自 Huot，1990）。D. Strach 和 E. C. Elliott 请 142 名评分员对两名被试的作文试卷进行评分，一名被试的得分从 62 分到 99 分不等，另一名被试的得分从 50 分到 98 分不等。H. Pieron 请 76 名评分员对同一篇作文评分，最低分与最高分相差了 13 分（转引自章熊，1994）。

国内类似的调查研究也有不少。1983 年，郑日昌对高考作文评分员的抽样调查发现，不同阅卷组对同一篇作文的评分误差高达满分的 60%。漆书清和曾桂兴分别在江西和广东每年选取三百至四百名评分员共同评阅 4 篇高考作文，都发现评分摆动的平均幅度在 40 分左右（转引自章熊，1994）。另外，聂建中和王正仁（1997）通过检验某省 1995 年高考英语 14 个口试评分员组的评分，同样发现不同评分员所评分数存在着较为显著的不一致性。

从以上国内外早期的调查研究来看，评分员间一致性问题是主观性考试评分中存在的一个突出问题，也是研究者一直关注的

焦点。但需要指出的是，以上调查研究仅对评分员间评分不一致的情况作了简单的统计和描述。

而另一些研究者则从影响评分员间一致性的因素方面进行了更为深入的探讨。例如，Alister Cumming（1990）发现新老评分员在对母语非英语者的作文进行评分时，评分结果存在较大的不一致性。Micheline Chalhoub-Devillie（1995）发现有教学背景的评分员与无教学背景的评分员在对母语非英语者的口语考试进行评分时，评分员对评分标准的不同方面的关注度不同，从而导致评分出现较大的不一致性。杨丽颖（2006）发现在国内大学生英语作文评分中，老评分员间的一致性明显高于新评分员。薄丽（2005）发现语言学背景和非语言学背景的两类评分员在 HSK（高等）作文考试评分中存在差异，且语言学背景的专业评分员对被试的评价更接近被试的真实语言水平。罗丹（2006）发现有阅卷经验和无阅卷经验的评分员在 HSK（高等）作文考试的评分中也存在显著差异。

另外，还有一些研究者从如何提高主观性考试的评分员间一致性方面进行了一些探索。例如，郭茜、邢如和沈明波（2003）对参加“清华大学英语水平考试”口语评分的两组评分员在参加培训前后的评分结果进行了相关分析，发现评分员在培训后具有很高的评分员间一致性。通过比较离散指数，他们发现培训对评分的影响很明显。李庆本和许雪立（1999）对 HSK（高等）口试评分过程的研究则表明：在同一时间段评分，评分员间一致性较为理想；在评分前让评分员听标杆样本并进行讨论，有助于增加不同时间段评分的一致性。

从以上研究来看，目前对 HSK（高等）主观性考试的评分员间一致性的研究并不全面，因此，本研究希望通过实证性研究的方法对 HSK（高等）主观性考试的评分员间一致性作一次较为全面的检验，并为相关研究提供可参考数据。首先我们要明确一个问题：主观性考试需要多高的评分员间一致性才能达到标准化测验对信度的要求呢？美国教育测验服务中心（Educational Testing

Service：ETS）对 TOEFL 口语考试评分员间一致性的检验显示：长面试评分员间一致性系数为 0.735；短面试评分员间一致性系数为 0.758（John L. D. Clark & Spencer S. Swinton，1979）。而北京语言大学汉语水平考试中心的检验则显示，HSK（高等）作文考试的评分员间一致性系数基本在 0.70 以上（罗丹，2006）。因此，我们认为，评分员间一致性系数达到 0.70 以上即符合大规模标准化测验信度的要求。

二　实证研究

2.1 研究问题

（1）HSK（高等）主观性考试具有多高的评分员间一致性？

（2）评分员对评分标准把握的一致性是否会影响 HSK（高等）主观性考试的评分员间一致性？

（3）评分员严厉度的一致性是否会影响 HSK（高等）主观性考试的评分员间一致性？

2.2 研究样本

本研究样本为 2006 年 10 月 15 日 HSK（高等）作文和口语考试评分员对考生表现的评分结果。HSK（高等）作文和口语考试均采用以整体评分法（Holistic Approach）为基础的 5 级分制评分（实际为 5 个基准级和 7 个辅助级，共 12 小级，以 3 级为合格线）。

在检验之前，要将评分员所评的评分等级转换为 12 个等级分数后，才能进行数据分析。具体转换方法如表 1 所示：

表 1. HSK 秩次等级转换表

评分等级	1	2 -	2	2 +	3 -	3	3 +	4 -	4	4 +	5 -	5
等级分数	1	2	3	4	5	6	7	8	9	10	11	12

需要指出的是：本研究仅对作文和口语考试评分员评出的原始成绩进行分析，复评成绩不在本研究范围内。

2.2.1 被试样本

本研究被试为2006年10月15日参加HSK（高等）作文和口语考试的5300名考生。排除成绩经过复评的考生后，实际被试样本为5300名作文考试考生和4930名口语考试考生。

2.2.2 评分员样本

本研究涉及的评分员为北京语言大学对外汉语教师和对外汉语专业研究生。作文和口语考试评分员各84名，参与复评的评分员不在本研究范围内。其中，作文考试评分员42组（每组2人），口语考试评分员28组（每组3人）。为保证各评分员的内部一致性，本研究涉及的所有评分员在评分之前均接受了统一的培训。

2.3 研究方法

本研究采用基于经典测量理论的相关系数法，对HSK（高等）作文和口语两项主观性考试的评分员间一致性进行检验。

HSK（高等）作文考试由两名评分员对同一被试进行评分，因此，我们采用适用于计算两列等级（顺序）变量之间相关的斯皮尔曼等级相关法（Spearman Correlation Coefficient）进行检验。

基本公式为：$r_R = \frac{3}{N-1} \cdot \left[\frac{4\sum R_x R_y}{N(N+1)} - (N+1)\right]$ （公式一）

其中，R_x 与 R_y 为两列变量各自排列的等级序数

N 为被试数目

HSK（高等）口语考试由三名评分员对同一被试进行评分，因此，我们采用适用于计算两列以上等级（顺序）变量之间相关的肯德尔和谐系数法（Kendall Coefficient of Concordance），又称W系数法，进行检验。

基本公式为：$W=\frac{S}{\frac{1}{12}K^2\ (N^3-N)}$ （公式二）

其中，$S=\sum R_i^2-\frac{(\sum R_i)^2}{N}$

R_i 为评价对象获得的 K 个等级之和

N 为被试数目

K 为评分员数目

由于在评分过程中，同一评分员会对不同被试的表现给出同一评分等级，所以需要使用校正公式：

$$W=\frac{S}{\frac{1}{12}K^2\ (N^3-N)\ -K\sum T} \quad \text{（公式三）}$$

其中，$S=\sum R_i^2-\frac{(\sum R_i)^2}{N}$，（同公式二）

$$\sum T=\sum\frac{n^3-n}{12}$$

n 为相同等级的数目

同时，本研究还将分别统计和分析 HSK（高等）作文和口语考试不同评分员对评分等级使用的相对频次，进一步考察和分析评分员对评分标准的把握和评分员的严厉度。

三　研究结果与分析

3.1 评分员间一致性检验结果与分析

3.1.1 HSK（高等）作文考试评分员间一致性检验结果与分析

根据公式一，应用 SPSS 16.0 计算斯皮尔曼等级相关系数，结果如表 2：

表 2. HSK（高等）作文考试评分员间斯皮尔曼等级相关系数表

评分员	评卷份数	斯皮尔曼等级相关系数（r_R）	评分员	评卷份数	斯皮尔曼等级相关系数（r_R）
组 1	127	0.598**	组 22	126	0.639**
组 2	127	0.756**	组 23	126	0.704**
组 3	127	0.638**	组 24	126	0.633**
组 4	127	0.566**	组 25	126	0.761**
组 5	127	0.717**	组 26	126	0.591**
组 6	127	0.627**	组 27	126	0.707**
组 7	127	0.683**	组 28	126	0.809**
组 8	127	0.715**	组 29	126	0.760**
组 9	126	0.720**	组 30	126	0.781**
组 10	126	0.833**	组 31	126	0.719**
组 11	126	0.626**	组 32	126	0.598**
组 12	126	0.520**	组 33	126	0.743**
组 13	126	0.681**	组 34	126	0.756**
组 14	126	0.744**	组 35	126	0.600**
组 15	126	0.680**	组 36	126	0.766**
组 16	126	0.734**	组 37	126	0.501**
组 17	126	0.652**	组 38	126	0.694**
组 18	126	0.597**	组 39	126	0.620**
组 19	126	0.719**	组 40	126	0.830**
组 20	126	0.717**	组 41	126	0.552**
组 21	126	0.748**	组 42	126	0.719**

注：*表示相关系数在 0.05 水平上显著，**表示相关系数在 0.01 水平上显著。

从表 2 可以看出，在本次 HSK（高等）作文考试评分中，每组两名评分员所评分数间的相关系数都达到了 0.01 显著水平。其中，最高的相关系数达 0.833（组 10），最低的为 0.501（组 37）。

为了使统计结果更为清晰，我们对表 2 中的斯皮尔曼等级相关系数进行了分段统计：

表 3. HSK（高等）作文考试评分员间斯皮尔曼等级相关系数分段统计表

斯皮尔曼等级相关系数	组数（组）	比例（%）
0.500 ~ 0.600	8	19.05
0.600 ~ 0.700	12	28.57
0.700 ~ 0.800	19	45.24
0.800 ~ 0.900	3	7.14

表 3 显示：本次 HSK（高等）作文考试共有 31 组评分员的评分员间信度系数在 0.600 ~ 0.800 之间，占总组数的 73.81%。信度系数在 0.800 以上的有 3 组，占总组数的 7.14%。而信度系数在 0.600 以下的有 8 组，占总组数的 19.05%。总体来看，本次 HSK（高等）作文考试具有较高的评分员间一致性。

3.1.2 HSK（高等）口语考试评分员间一致性检验结果与分析

由于本研究中大多数口语考试评分员组所评被试数量都超过了 166 名，无法使用 SPSS 软件对肯德尔和谐系数进行计算，所以本研究应用 Visual FoxPro 6.0 编写了相关运算程序进行计算，结果如表 4：

表 4. HSK（高等）口语考试评分员间肯德尔和谐系数（W 系数）表

评分员	评卷份数	肯德尔和谐系数（W）	评分员	评卷份数	肯德尔和谐系数（W）
组 1	159	0.681**	组 7	190	0.732**
组 2	150	0.702**	组 8	189	0.688**
组 3	189	0.758**	组 9	138	0.689**
组 4	189	0.701**	组 10	171	0.747**
组 5	170	0.581**	组 11	186	0.671**
组 6	190	0.782**	组 12	184	0.655**

续表

评分员	评卷份数	肯德尔和谐系数（W）	评分员	评卷份数	肯德尔和谐系数（W）
组 13	188	0.763**	组 21	169	0.752**
组 14	146	0.743**	组 22	172	0.705**
组 15	176	0.708**	组 23	189	0.667**
组 16	173	0.723**	组 24	179	0.752**
组 17	181	0.772**	组 25	189	0.738**
组 18	163	0.736**	组 26	189	0.691**
组 19	186	0.795**	组 27	170	0.806**
组 20	182	0.787**	组 28	173	0.749**

注：*表示相关系数在0.05水平上显著，**表示相关系数在0.01水平上显著。

从表4中可以看出，本次HSK（高等）口语考试中，每组三名评分员所评分数间的相关系数也都达到了0.01的显著水平。其中，最高的相关系数达0.806（组27），最低的为0.581（组5）。

与作文考试的检验相同，我们也对表4中的肯德尔和谐系数进行了分段统计：

表5. HSK（高等）口语考试评分员间肯德尔和谐系数分段统计表

肯德尔和谐系数	组数（组）	比例（%）
0.500～0.600	1	3.57
0.600～0.700	7	25.00
0.700～0.800	19	67.86
0.800～0.900	1	3.57

表5显示：本次HSK（高等）口语考试评分员信度系数也较多地集中在0.600～0.800之间，共26组，占总组数的92.86%。其中，信度系数在0.700～0.800之间的，共19组，占总组数的67.86%。总体来看，本次HSK（高等）口语考试也具有较高的评分员间一致性。

3.1.3 HSK（高等）作文和口语考试评分员间一致性对比分析

我们还对本次 HSK（高等）作文和口语考试评分员间信度系数的统计结果作了比较：

表 6. HSK（高等）主观性考试评分员间信度系数分段统计表

信度系数	作文考试		口语考试	
	比例（%）	累加比例[①]（%）	比例（%）	累加比例（%）
0.800～0.900	7.14	7.14	3.57	3.57
0.700～0.800	45.24	52.38	67.86	71.43
0.600～0.700	28.57	80.95	25.00	96.43
0.500～0.600	19.05	100.00	3.57	100.00

由表 6 可知，本次 HSK（高等）作文和口语考试的评分员间信度系数基本都达到了 0.700 以上。其中，作文考试评分员间信度系数达 0.700 以上的评分员组占 52.38%，口语考试评分员间信度达 0.700 以上的评分员组占 71.43%。这说明此次 HSK（高等）主观性考试的评分员间一致性程度总体较高。

同时，通过比较我们发现，本次 HSK（高等）作文考试评分员间信度系数低于 0.600 的评分员组所占百分比远高于口语考试，而信度系数高于 0.700 的评分员组所占百分比却远低于口语考试。因此，从总体上看，作文考试的评分员间一致性在整体上低于口语考试。造成这一结果的原因有待考察。

3.2 评分员评分等级使用相对频次分析

当评分员间一致性较高时，评分员所使用的评分等级也应在一定程度上保持相对一致。但实际上，评分员在使用评分等级时会出现一定的差异。我们可通过对比评分员评分等级使用的相对频次来考察各评分员间的评分差异，以探求影响评分员间一致性的因素。

评分员评分等级使用相对频次是指：针对所有评分员在整个