

中国当代学术探索文库·语言文化卷

汉语测试、习得 与认知新探索



HANYU CESHIXIDE YU RENZHIXIN TANSUO

张旺熹 邢红兵 主编



中国书籍出版社

本书出版得到北京语言大学特色重点学科项目经费支持

汉语测试、习得 与认知新探索

张旺熹、邢红兵 主编

中国书籍出版社

图书在版编目 (C I P) 数据

汉语测试、习得与认知新探索 / 张旺熹, 邢红兵主编.

— 北京: 中国书籍出版社, 2012. 6

ISBN 978-7-5068-2804-8

I. ①汉… II. ①张… ②邢… III. ①汉语—语言教学—文集
IV. ①H19-53

中国版本图书馆CIP数据核字(2012)第087992号

策划编辑 / 安玉霞

责任编辑 / 安玉霞

责任印制 / 孙马飞 张智勇

封面设计 / 展 华

出版发行 / 中国书籍出版社

地址: 北京市丰台区三路居路97号(邮编: 100073)

电话: (010)52257153(发行部) (010)52257143(总编室)

电子邮箱: chinabp@vip.sina.com

经 销 / 全国新华书店

印 刷 / 三河富华印刷包装有限公司

开 本 / 710毫米×1000毫米 1/16

印 张 / 17

字 数 / 260千字

版 次 / 2012年07月第1版 2012年07月第1次印刷

书 号 / ISBN 978-7-5068-2804-8

定 价 / 38.00元

★ 版权所有 侵权必究 ★

汉语测试、习得与认知新探索

主 编 张旺熹 邢红兵

编 委 (按音序排列)

柴省三 陈 宏 郭树军 黄理兵

李桂梅 聂 丹 田清源 王佶旻

邢红兵 张宝林 张 凯 张旺熹

前 言

为激发研究生的科研热情和科研创新能力,加强研究生之间的学术交流,北京语言大学汉语水平考试中心2008年11月举办了“首届语言测试与习得研究生学术论坛”,会后精选了16篇论文集结出版,论文集《汉语测试、习得与认知探索》已由北京语言大学出版社于2010年3月出版。2010年11月又举办了“第二届语言测试与习得研究生学术论坛”,会后精选了21篇论文集结出版,论文集《汉语测试、习得与认知探索(续一)》由世界图书出版公司于2011年9月出版。2011年12月23日,我们又举办了“2011年研究生科研创新论坛”,中心的所有研究生导师以及70多名硕士、博士研究生参加了论坛,老师和同学们共同就语言测试、语言习得、语言认知及语言教学等领域的相关问题进行了交流和探讨。

本次论坛主要作为“汉语水平考试中心2010年研究生科研项目”结项总结的汇报展示,同时吸收部分未立项的研究论文。论坛共收到28篇研究生论文,研究方向涉及语言测试、语言习得和语法研究等方向。这些论文大部分得到汉语水平考试中心研究生科研项目的立项资助。经过一年的研究,同学们的论文都表现了较高的科研水平,主要表现在论文选题有价值,侧重于实证研究等方面。论坛结束后,我们组织了论文评审委员会,对提交论坛的论文进行了匿名评审,最后有16篇论文入选论文集。为了保持论文集的延续性,本届论坛的论文集定名为《汉语测试、习得与认知新探索》。

论文集编委会成员均为北京语言大学汉语水平考试中心的研究生导师,导师们在论坛组织、论文指导及论文评审等各个环节都全力投入,在对论文进行评审的同时,也提出了很多具体的修改意见建议,而且还在论坛举办时亲临现场,给学生以具体的指导。导师们的辛勤付出,对提高研究生的科研能力和论文写作水平都有重要的指导意义,也是论文集的整体水平得以保证

的前提。

汉语水平考试中心研究生管理办公室的李郁和金樱两位老师在论坛的组织和论文集的编选过程中做了大量的工作，论文集的初步编排工作由硕士研究生王专、李姝雯同学承担，王正刚先生为本书作了细致的编校和体例统一工作。感谢中国书籍出版社为我们提供了论文集出版的宝贵机会，感谢本书编辑安玉霞女士所做的细致工作。

北京语言大学汉语水平考试中心

2012年5月30日

目 录

题组项目功能差异检验方法在 HSK(高等) 阅读理解中的应用·····	周莉 耿岳 (001)
C.test 口语考试的对答结构分析 ·····	王保利 (024)
推断类试题在语言测试中的作用·····	邓泽茜 张巍 李建媛 (034)
韩国留学生汉语口语语篇逻辑连接手段使用考察·····	赵燕朋 (052)
语言测验的真实语料观·····	孙鹏阁 (068)
C.TEST(A-D) 阅读理解测试答题策略研究初探·····	黄 敬 (077)
留学生语篇表达中的偏误类型及分布·····	李姝雯 孟莹 张琳琳 赵奕 (093)
留学生汉语称谓语使用情况调查研究·····	郭肖芳 (123)
2010 年度新词语的语言学考察 ·····	谭 静 (141)
基于 HSK 动态作文语料库的韩国留学生“比”字句偏误研究 ·····	李 桑 (152)
汉语中介语语料库文本句标注与标注员语感·····	徐枫洁 (163)
外国学生“使”字句回避偏误的成因分析·····	郑云霞 (178)
汉语书面语中的话语标记研究·····	王 专 (192)
重动构式的语篇功能考察·····	郭雅静 (204)
“又”、“还”、“再”语篇衔接功能分析 ·····	靳 玮 (216)
也谈“大+X+的”结构 ·····	周也明 (232)
“A 的 A, B 的 B” 对举格式考察 ·····	刘雪莹 (247)

题组项目功能差异检验方法在 HSK(高等) 阅读理解中的应用

周莉 耿岳

提 要 本研究引入能够处理题组效应的项目功能差异检验方法,为篇章阅读测验提供更科学的项目功能差异检验方法。研究采用 GMH 法、P-SIBTEST 法和 P-LR 法对中国汉语水平考试(HSK)(高等)阅读理解试题部分尝试进行了 DIF 检验。研究表明,这三种方法的检验结果具有较高的一致性,该部分试题在性别与国别变量上不存在显著的 DIF 效应。本研究还将传统的 DIF 检验方法的结果与变通的题组 DIF 检验方法的结果进行了比较,结果表明后者具有明显的优越性。

关键词 项目功能差异 题组 多级计分 HSK 阅读理解

一 问题的提出

心理测量或者考试的目的是尽可能准确、公平地测出被试各种心理特质上的差异。研究者们通过项目功能差异(DIF)检验检测出测验中可能对不同群体的个体产生不公平待遇的项目。近些年,不同的测验条件也给 DIF 检验方法带来了很多挑战,比如说如何处理有题组效应的篇章形式的阅读理解试题的 DIF 检验。

题组(testlet)指的是具有共同刺激或者其他特征的一组项目(Wainer & Kiely,1987),也有研究者称之为项目束(item bundle)(Rosenbaum,1988)。阅读测试中一个篇章与有关的几个项目就构成了典型的题组。同一题组中的项目之间很有可能存在局部依赖性(local

dependence), 而局部独立性是使用项目反应理论的基本假设之一 (Lee, Brennan, & Frisbie, 2005)。Bolt (2002) 认为, 局部依赖性造成的模型拟合问题也会影响 DIF 检验的结果。因此, 我们有必要处理好题组项目间的局部依赖性问题。传统的项目功能差异检验方法忽略了篇章阅读测验项目的题组效应, 存在一定的不稳定性。

目前解决题组项目间的局部依赖性问题主要有两种方法。第一类方法是把同一个篇章的多个项目作为一个独立的整体, 把题组内每一个二级计分的项目看作是这个整体反应的一部分, 然后用针对多级计分的 IRT 模型进行拟合, 然后进行 DIF 检验 (Wainer & Kiey, 1987)。这样对多个二级计分项目进行的 IRT DIF 检验变成了对一个多级计分项目的 IRT DIF 检验。后来, 遵循同样的思路, 研究者把适用于多级计分的非 IRT DIF 检验方式引入到题组测验的 DIF 检验中来, 例如 P-SIBTEST 法, P-LR 法, P-Mantel 法等。

第二类方法是近几年提出来的题组 IRT 模型 (testlet response theory model, 简称 TRTM)。这类模型在传统 IRT 模型中加入一个被试层面的随机参数, 表示题组项目的依赖程度 (Wang et al., 2002)。虽然这种方法对题组效应的处理更加科学, 是真正的题组 DIF 检验方法。但是由于这种模型待估参数较多, 采用一般的算法非常复杂, 在实践中有一定的困难, 也就限制了该方法的应用。因此, 本研究采用的是第一种题组 DIF 检验方法, 即把适用于多级计分的非 IRT DIF 检验方式引入题组测验的 DIF 检验中来。

篇章形式的阅读理解在中国汉语水平考试 (HSK) 中占有很重要的地位。这种形式的题目倾向在篇章中考查被试的阅读理解能力、阅读速度以及词汇、语法等语言能力。本研究主要解决两个问题:

(1) 尝试用变通的题组项目功能差异检验方法对 HSK (高等) 阅读理解题进行 DIF 检验;

(2) 将传统 DIF 检验方法与可以处理题组效应的 DIF 检验方法进行比较, 探讨题组 DIF 检验方法的必要性和优越性。

二 研究设计

2.1 研究方法

我们选择 G-MH 方法、P-SIBTEST 方法和 P-LR 方法检测 HSK（高等）阅读理解试题。其中，G-MH 方法采用 Angel M. Fidalgo 教授编制的 GMH 软件完成，P-SIBTEST 方法采用 P-SIBTEST 软件完成，P-LR 方法在 SPSS17.0 里完成。

2.2 研究材料

本文所使用的研究材料是 2010 年 10 月 16 日在中国国内举办的中国汉语水平考试 HSK(高等) 的阅读理解试题。该试题包括两个部分，第一部分是快速阅读后用最简单的汉字回答问题；第二部分是传统意义上的阅读理解题，也就是阅读理解最常见的题型，即每段文字后都有若干个问题，每个问题都有四个备选答案，要求考生选择唯一恰当的答案。题目总数量为 40，分为 8 个题组。

这次考试共有 2733 名考生，男生 1183 人，女生 1550 人，中国少数民族考生 739 人，外国考生 1994 人。具体描述性统计结果如下表：

表 1 阅读理解试题总体描述性统计

题数	样本数	平均数	标准差	偏态值
40	2733	22.2774	7.4074	0.2140
峰态值	ALPHA 系数	标准误	平均难度	平均点双列
-0.6284	0.8628	2.7415	0.5569	0.4007

综合上述指标，我们可以看出这部分试题具有很高的可靠性，区分度高，难度适中，分数误差小。

2.3 研究假设

A：与男性考生相比，HSK（高等）阅读理解某些试题对女性考生在某种程度上具有不公平性。

B：与中国少数民族考生相比，HSK（高等）阅读理解某些试题对外国考生在某种程度上具有不公平性。

2.4 研究步骤

2.4.1 G-MH 方法

研究 A：男女考生的公平性问题

由于我们假设女生组可能受到不公平待遇，所以我们把女生组设为目标组，男生组设为参照组。具体步骤如下：

- (1) 整理数据，将女生组、男生组考生的题组成绩合并为一个数据库；
- (2) 将数据导入 GMH 软件中进行分析。软件分为三步完成数据分析：
 - ① 将 8 个题组的总分作为匹配变量，所有的项目都列入怀疑有 DIF 的范围；
 - ② 把筛选出来的项目剔除，将剩下的项目总分作为匹配变量，再次检验 DIF；
 - ③ 对检测出来的 DIF 项目进行显著性检验。

研究 B：中外考生的公平性问题

由于我们假设外国考生可能受到不公平待遇，所以我们把外国考生组设为目标组，中国少数民族考生组设为参照组。

数据分析步骤和研究 A 几乎完全一致，只是目标组和参照组分别换成外国考生组和中国少数民族考生组。

2.4.2 P-SIBTEST 方法

研究 A：男女考生的公平性问题

- (1) 整理数据，从男生组中随机抽取 1100 人，从女生组中随机抽取 1100 人。由于每一组的人数并不相同，我们要按照人数少的组进行抽样。
- (2) 分别将男生组和女生组的题组答题情况分开存放，这样有利于下一

步操作。

(3) 运用 P-SIBTEST 程序进行分析, 第一次将所有项目都列入怀疑有 DIF 的范围。

(4) 把筛选出来的项目剔除, 对剩下的项目再一次操作, 直到没有 DIF 项目出现, 得到最佳匹配变量。

(5) 利用上一步得到的匹配变量, 再运行程序, 对筛选出来的怀疑有 DIF 的项目再次检验, 得到最确定的 DIF 项目。

研究 B : 中外考生的公平性问题

(1) 整理数据, 从中国少数民族考生组中随机抽取 700 人, 从外国考生组中随机抽取 700 人。同样, 由于每一组的人数并不相同, 我们要按照人数少的组进行抽样。

(2) 以下步骤与研究 A 步骤完全一致, 不再赘述。

2.4.3 P-LR 方法

研究 A : 男女考生的公平性问题

(1) 整理数据, 将考生的题组答题情况存入 SPSS 中作为因变量, 以阅读测验部分的总分为匹配变量, 分组变量为组别 (男生组和女生组)。

(2) 运行写好的程序。

研究 B : 中外考生的公平性问题

数据分析和研究 A (男女考生的公平性问题) 几乎完全一致, 分组变量调整为国别 (中国考生组和外国考生组)。

三 研究结果与分析

3.1 所抽样本描述性统计结果

为了保证结果的准确性, 我们在抽样的时候考虑到了目标组和参照组能力本身的差异, 排除了一些总分较低的考生。描述性统计结果具体如下:

表 2 样本描述性统计

	样本数		平均分	标准差
性别组	男	1100	23.61	6.41
	女	1100	24.39	6.37
国别组	中	700	25.97	6.14
	外	700	25.14	6.08

3.2 三种题组 DIF 检验方法的结果

3.2.1 G-MH 方法

表 3 男女组显著性检验结果

题组 1	Stage 1:	QMH = 1.0992	df = 1	p = 0.2945
	Stage 2:	QMH = 0.2236	df = 1	p = 0.6363
题组 2	Stage 1:	QMH = 1.2982	df = 1	p = 1.2982
	Stage 2:	QMH = 0.1955	df = 1	p = 0.6584
题组 3	Stage 1:	QMH = 1.1187	df = 1	p = 0.2902
	Stage 2:	QMH = 0.2085	df = 1	p = 0.6479
题组 4	Stage 1:	QMH = 13.9062	df = 1	p = 0.0002*
	Stage 2:	QMH = 15.9874	df = 1	p = 0.0001*
题组 5	Stage 1:	QMH = 15.7773	df = 1	p = 0.0001*
	Stage 2:	QMH = 18.5004	df = 1	p = 0.0000*
题组 6	Stage 1:	QMH = 2.6187	df = 1	p = 0.1056
	Stage 2:	QMH = 0.3646	df = 1	p = 0.5460
题组 7	Stage 1:	QMH = 0.3684	df = 1	p = 0.5439
	Stage 2:	QMH = 0.3453	df = 1	p = 0.5568
题组 8	Stage 1:	QMH = 1.2983	df = 1	p = 0.2545
	Stage 2:	QMH = 0.1985	df = 1	p = 0.6560

*p<0.05

在男女考生组的研究中，共检测出两个有 DIF 的题组，分别是题组 4、5。

表 4 中外考生组显著性检验结果

题组 1	Stage 1:	QMH =19.8972	df = 1	p = 0.0000*
	Stage 2:	QMH = 16.7124	df = 1	p = 0.0000*
题组 2	Stage 1:	QMH =0.1390	df = 1	p =0.7093
	Stage 2:	QMH = 0.3859	df = 1	p = 0.5344
题组 3	Stage 1:	QMH = 24.0032	df = 1	p = 0.0000*
	Stage 2:	QMH =13.7536	df = 1	p = 0.0002*
题组 4	Stage 1:	QMH = 1.4209	df = 1	p = 0.2333
	Stage 2:	QMH = 1.2803	df = 1	p = 0.2578
题组 5	Stage 1:	QMH = 27.4219	df = 1	p = 0.0000*
	Stage 2:	QMH = 20.6532	df = 1	p = 0.0000*
题组 6	Stage 1:	QMH = 2.7070	df = 1	p = 0.0999
	Stage 2:	QMH = 3.1321	df = 1	p = 0.0768
题组 7	Stage 1:	QMH = 17.6367	df = 1	p = 0.0000*
	Stage 2:	QMH = 10.6657	df = 1	p = 0.0011*
题组 8	Stage 1:	QMH = 2.5996	df = 1	p = 0.1069
	Stage 2:	QMH = 0.0001	df = 1	p = 0.9935

*p<0.05

在中外考生组的研究中，共检测出 4 个有 DIF 的题组，分别是题组 1、3、5、7。

3.2.2 P-SIBTEST 方法

表 5 男女考生组显著性检验结果

题 组	Beta estimate	standard error	p-value	p-elim R F	MS SSD
1	-0.092	0.052	0.078	.01 .01	-0.11
2	-0.049	0.049	0.315	.00 .01	-0.11
3	-0.090	0.068	0.185	.00 .01	-0.10

续表

题 组	Beta estimate	standard error	p-value	p-elim R F	MS SSD
4	0.132	0.042	0.002*	.00 .01	-0.14
5	0.148	0.043	0.001*	.00 .01	-0.15
6	-0.071	0.048	0.144	.00 .01	-0.11
7	-0.017	0.050	0.731	.00 .01	-0.12
8	-0.083	0.052	0.112	.00 .01	-0.11

*p<0.05

在男女考生组的研究中，共检测出两个有 DIF 的题组，分别是题组 4、5，都是对参照组（男生）有利的。

表 6 中外考生组显著性检验结果

题 组	Beta estimate	standard error	p-value	p-elim R F	MS SSD
1	0.221	0.060	0.000*	.02 .01	0.09
2	-0.020	0.056	0.725	.01 .01	0.13
3	0.373	0.085	0.000*	.01 .01	0.06
4	-0.006	0.052	0.913	.02 .01	0.13
5	-0.286	0.052	0.000*	.02 .01	0.18
6	-0.165	0.057	0.004*	.02 .01	0.15
7	-0.345	0.061	0.000*	.01 .01	0.18
8	-0.089	0.064	0.165	.01 .01	0.14

*p<0.05

在中外考生组的研究中，共检测出 5 个有 DIF 的题组，分别是题组 1、3、5、6、7。题组 1、3 对参照组（中国少数民族考生）有利，题组 5、6、7 对目标组（外国考生）有利。

3.2.3 P-LR 方法

表 7 男女考生组显著性检验结果

题组	模型 A	模型 B	模型 C	df	卡方 (B-A)	卡方 (C-B)	df	卡方 (C-A)
1	906.990	909.441	913.022	1	2.451	3.581	2	6.032
2	1028.740	1029.591	1030.470	1	0.851	0.879	2	1.730
3	1634.564	1636.653	1637.111	1	2.089	0.458	2	2.547
4	525.932	539.295	540.933	1	13.363*	1.638	2	15.001*
5	591.239	606.701	606.702	1	15.462*	0.001	2	15.463*
6	744.634	747.405	750.000	1	2.771	2.595	2	5.366
7	1099.279	1099.373	1109.241	1	0.094	9.868*	2	9.962*
8	1039.849	1040.712	1040.743	1	0.863	0.031	2	0.894

在男女考生组的研究中,共检测出 3 个有 DIF 的题组,分别是题组 4、5、7,题组 4、5 为一致性 DIF,题组 7 为非一致性 DIF。

表 8 中外考生组显著性检验结果

题组	模型 A	模型 B	模型 C	df	卡方 (B-A)	卡方 (C-B)	df	卡方 (C-A)
1	550.464	565.781	567.475	1	15.317*	1.694	2	17.011*
2	578.482	578.482	578.495	1	0.000	0.013	2	0.013
3	1045.928	1067.535	1068.017	1	21.607*	0.482	2	22.089*
4	376.208	376.384	377.933	1	0.176	1.549	2	1.725
5	343.785	384.163	384.264	1	40.378*	0.101	2	40.479*
6	554.390	558.364	561.220	1	3.974	2.856	2	6.830*
7	823.570	844.875	845.314	1	21.305*	0.439	2	21.744*
8	734.130	734.707	734.866	1	0.577	0.159	2	0.736

在中外考生组的研究中,共检测出 5 个有 DIF 的题组,分别是题组 1、3、5、6、7,且均为一致性 DIF。

综合上述结果,我们将三种方法的显著性检验结果进行整理,结果如下(标注“√”表示某个题组在不同题组 DIF 检验方法的显著性检验中拒绝虚无假设):

表 9 性别 DIF 检验结果

题 组	G-MH	P-SIBTEST	P-LR
1			
2			
3			
4	√	√	√
5	√	√	√
6			
7			√
8			

表 10 国别 DIF 检验结果

题 组	G-MH	P-SIBTEST	P-LR
1	√	√	√
2			
3	√	√	√
4			
5	√	√	√
6		√	√
7	√	√	√
8			

从表 11 和表 12 可以看出,这三种方法的结果基本一致: 题组 4、5 在性别上存在 DIF 的可能性较大, 题组 1、3、5、6、7 在国别上存在 DIF 的可能性较大。但是我们需要进一步计算题组 DIF 的效应值指标,判断 DIF 效应(DIF effect size)的大小。为了提高题组 DIF 检验的准确性,避免不必要的人力与测试资源的浪费,我们利用 P-LR 方法提供的 DIF 效应大小的指标,对上述检验中存在 DIF 的项目进行分析。

目前该方法公认的 DIF 效应值指标是 Jodoin & Geirl(2001) 提出的:

A (可忽略 DIF): $\Delta R^2 < 0.035$

B (中等 DIF 效应): $0.035 \leq \Delta R^2 \leq 0.07$