中国境内外 HSK 考生成绩公平性的分析

任杰
北京语言文化大学汉语水平考试中心　北京 100083
E-mail:renjie@blcu.edu.cn

摘要：本研究利用项目功能差异（Differential Item Functioning，简称 DIF）理论，对 HSK 考生中不同两个群体——中国境内考生和境外考生，进行题目的反应分析，考查 HSK 的题目是否存在不利于某一群体的。我们采用 MH 方法检测 DIF，利用标准化的离散分析方法鉴别 DIF 的真伪，并寻找造成 DIF 的原因。由数据分析的结果可知，中国境内考生听力占优势，而境外考生语法和阅读成绩好于听力；有利于中国境内考生的题目多于境外考生，不过，所占比例很小。因此，HSK（初、中等）A 试卷的题目对于中国境内与境外考生基本公平。

关键词：项目功能差异（DIF），汉语水平考试（HSK），MH 方法，标准化的离散分析方法。

DIF Detection and Analysis in HSK scores of examinees taken in China and overseas

Ren Jie
HSK center of Beijing Language and Culture University ,Beijing 100083
E-mail:renjie@blcu.edu.cn

ABSTRACT: The research uses Differential Item Functioning(DIF) theory to detect whether there is different item performance in HSK testing between examinees taken in China and overseas. We use MH Method to detect the DIF, and do Standardized Distractor Analysis Method to distinguish which item having DIF is true and search the reasons causing this DIF. The result shows that the examinees in China perform better than the matched overseas ones on the Listening items, and less on the Sentence Structure and Reading .The overall item's number favoring the examinees in China is just a little more . So, the items in the test A of HSK Elementary-Intermediate are fair with examinees taken in China and overseas basically .

Keyword: DIF, HSK, MH method, Standardized Distractor Analysis method.

1. 问题提出

汉语水平考试（HSK）是测试母语非汉语者的汉语水平而设立的一种国家级标准化考试。为了保证HSK题目中不出现或少出现对某一群体有利或不利的情况，必须对施测后的数据进行题目公平性分析。造成不同群体在题目反应上的差异的原因主要有两种：1. 题目有偏差。由于题目所测知识与技能不是考生必须掌握的，并且只有某一群体的考生才具有，那么这类题目就对另一群体的考生不公平，这类题目是我们应该避免的；2. 题目测试的内容是考生应该掌握的知识，只是由于学习环境等各种因素的影响，掌握的程度不同。此外，造成这种反应上的差异还可能由于一些尚未清楚的原因。我们对HSK试卷进行了亚裔与非亚裔的公平性分析，找寻过在HSK试卷中是否存在性别歧视等。

本文试图从项目功能差异的角度来分析中国境内与境外考生对HSK题目的反应，寻找差异，分析差异，避免可以避免的差异，以此保证HSK成绩的公平与合理性。

2. 研究材料

HSK（初、中等）某试卷（以下简称A卷），数据是1999年参加该试卷考试的国内日本考生（男生457人，女生592人）和国外日本考生（男生525人，女生789人）的成绩。

3. 研究方法

项目功能差异（Differential Item Functioning，简称DIF）指的是某题目在不同群体间表现出与测验目的无关的功能性差异。方法是当两组被试中具有相同能力的人，在某题上的正确回答概率相同，则该题没有项目功能差异。

MH分析方法（Mantel和Haenszel于1959提出）和标准化分析方法（standardization）是目前较为广泛采用的两种方法。其中标准化的离散分析方法是经过改造的标准化分析方法，本文利用该方法解释DIF出现的原因，而利用MH分析方法寻找DIF。

(1) MH分析方法

MH分析方法首先需要确定参照组、目标组和匹配变量，并且根据匹配变量的不同能力水平将数据分组（匹配小组）；其次分别计算在相应的匹配小组中参照组R和目标组F在某题目上的答对、答错人数，计算某题目的固定偏移比αMH和固定偏移比的标准化值MH D-DIF。

表1 某匹配小组j在某题目i上人数分布情况

组	正确人数	错误人数	总人数
参照组R	A_{ij}	B_{ij}	NR_{ij}
目标组F	C_{ij}	D_{ij}	NF_{ij}
总计	$M1_{ij}$	MO_{ij}	T_{ij}

利用固定偏移比(constant odds ratio) âMH 来表示某题目的功能差异程度：

$\hat{a}MH=(\Sigma A_{ij}D_{ij} / T_{ij}) / (\Sigma B_{ij}C_{ij} / T_{ij})$

âMH>1，则表示 R 比 F 表现好；âMH=1 表示无差异项目

为了便于理解，ETS（Education Testing Service）将固定偏移比标准化为：

MH D-DIF = $-2.35\log(\hat{a}MH)$

该值的正值表示对目标组有利，负值表示对参照组有利。

它的标准误是：

SE (MH D-DIF)

$= [2.35 / C]\times \mathrm{sqrt}\{ \Sigma [(A_{ij}D_{ij}+\hat{a}MH\ B_{ij}C_{ij})\times[A_{ij}+D_{ij}+\hat{a}MH(B_{ij}+C_{ij})] / (2\ T_{ij}* T_{ij})]\}$

其中 $C = \Sigma (A_{ij}D_{ij} / T_{ij})$

DIF 的三种水平（ETS 的分类标准）：

A 级：可忽略：若 MH D-DIF 与零无显著差异，或绝对值小于 1；

B 级：中间的：若 MH D-DIF 与零有显著差异，绝对值至少为 1，并且（1）小于 1.5，（2）不显著大于 1.0；

C 级：显著的：若 MH D-DIF 显著大于 1.0，且绝对值大于或等于 1.5。

(2) 标准化的离散分析方法

此方法通过计算两团体在某题目上对选项 A、B、C、D 的反映率的标准化差异 STD P-DIF(A)、STD P-DIF(B)、STD P-DIF(C)、STD P-DIF(D)来解释 DIF 出现的原因。

首先，计算在 i 题目 j 总分水平上目标组和参照组的考生选 A 的比率：

$PF_{ij}(A)= AF_{ij}/NF_{ij}$　　$PR_{ij}(A)=AR_{ij}/NR_{ij}$

其中，AF_{ij}和 AR_{ij}是目标组和参照组在 i 题目 j 总分水平上选 A 的人数。

其次，计算比率的差异：　$D_{ij}(A)= PF_{ij}(A)- PR_{ij}(A)$

最后，应用这些差异的标准化权重函数把不同总分水平上个体被试的分数水平差异结合起来得到对选择 A 的标准化差异 STD P-DIF(A)：

STD P-DIF(A)= $\Sigma W_{ij} D_{ij}(A)/ \Sigma W_{ij}$　　其中，$\Sigma W_{ij} = NF_{ij}$

STD P-DIF(B)、STD P-DIF(C)、STD P-DIF(D)的计算方法同上。

STD P-DIF(A)的正值表示这个题目目标组选 A 的比率高于参照组；负值表示这个题目目标组选 A 的比率低于参照组。

(3) 自编的 MH 和标准化分析方法的通用软件

本研究需要对大量的数据进行抽样、评分、计算和比较，为此，我们使用 VISUAL FOXPRO 语言，编制了 MH 分析方法和标准化分析方法的通用软件。

4. 研究过程

(1) 样本抽样。

为了比较想比较的环境差异，抽样时尽量避免其它因素的干扰，所以相互比较的中国境内外两个团体要求人数相等，性别相同，并且具有相同的文化背景（均为日本考生）。为此我们做了以下四次抽样：

样本①：从中国境内考生中提取全部日本男生 457 人；

样本②：从中国境内考生中随机抽取日本女生 592 人；

样本③：从中国境外考生中随机抽取日本男生 457 人；

样本④：从中国境外考生中随机抽取日本女生 592 人。

(2) 计算考生成绩。

(3) 分别以全部题目的总分和分测验的总分为匹配变量，以 5 分分组（原因见参考文献[1]），对**中国境内外日本男生**样本①和③进行计算，求其固定偏移比 MH D-DIF、标准误 SE (MH D-DIF)、选项 A、B、C、D 的反映率的标准化差异值 STD P-DIF (A)、STD P-DIF (B)、STD P-DIF (C) 和 STD P-DIF (D)。

(4) 分别以全部题目的总分和分测验的总分为匹配变量，以 5 分分组，对**中国境内外日本女生**样本②和④进行计算，求其固定偏移比 MH D-DIF、标准误 SE (MH D-DIF)、选项 A、B、C、D 的反映率的标准化差异值 STD P-DIF (A)、STD P-DIF (B)、STD P-DIF (C) 和 STD P-DIF (D)。

(5) 利用标准化的离散分析方法对 DIF 为 C 级和接近 C 级的题目进行分析。

5. 研究结果与分析

听力与阅读显然是两种不同的能力。HSK 总分相同的考生，完全可能在听力上差别很大，因此，我们主要以分测验的总分来考查 DIF。同时，我们也以全部题目的总分为匹配变量，考查了 DIF，虽然这时的 DIF 结论并不可靠，但是我们可以从中发现问题。

(1) 以**全部题目的总分为匹配变量**的 DIF 计算结果如下：

表 2　有利于境内、外考生的题目数量

分测验	日本男生 DIF 为 B/C			日本女生 DIF 为 B/C		
	总数	利于境内	利于境外	总数	利于境内	利于境外
听力理解	23	10/13		26	7 /19	
语法结构	3	1 /	2 /	9	/1	5 /3
阅读理解	19		4 /15	23		6 /17
综合填空	8	4 /1	2 /1	16	3 /3	6 /4
总计	53	15/14	8 /16	74	10 /23	17 /24

注：以全部题目的总分作为匹配变量、5 分分组、采用相等的样本量

由表 2 中日本男生对比组和女生对比组的数据均显示：

① 听力部分（共 50题）DIF 显著的题目**全部有利于中国境内**的考生；

② 阅读部分（共 50 题）DIF 显著的题目**全部有利于中国境外**的考生。

以全部题目的总分作为匹配变量时，DIF 显著的题目很多，根据经验，很可能匹配组有问题，为此，我们改变了分组大小，结果没有很大改观；改变匹配变量，出现 DIF 的题目明显减少。另外，计算各分测验的平均分及与总分平均分的比率，可知，以总分（全部题目的）相同的考生衡量他们的汉语综合能力比较合适，而衡量他们在听力和阅读测验的反应，显然是不合适的。所以，以各分测验的总分作为匹配变量来比较各分测验，结果更可靠。

表 3 日本男生境内外考生各 457 人

	听力		语法		阅读		综合		全题总分	
	境内	境外	境内	境外	境内	境外	境内	境外	境内	境外
平均分	29. 6	22. 6	19. 7	18. 0	36. 2	34. 5	27. 8	24. 6	113. 4	99. 9
占全题	26%	22%	17%	18%	31%	34%	24%	24%		

(2) 以**各分测验的总分为匹配变量**的 DIF 计算结果如下：

表 4 有利于境内、外考生的题目数量

分测验	日本**男**生 DIF 为 B 和 **C**			日本**女**生 DIF 为 B 和 **C**		
	总数	利于境内	利于境外	总数	利于境内	利于境外
听力理解	3	2/**1**		9	5/**2**	2/
语法结构	2	1/	1/	3	2/	/**1**
阅读理解	5	1/**1**	3/	8	3/**2**	1/**2**
综合填空	8	4/**3**	1/	8	2/**3**	1/**2**
总计	18	8/**5**	5/	27	12/**7**	4/**5**

注：以各分测验的总分作为匹配变量、5 分分组、采用相等的样本量

(3) 利用标准化的离散分析方法分析出现 DIF 的原因：

例 1：某一有关家庭生活的题目，请考生从给出的 A、B、C、D 四个选项中选择一个正确答案。境内、外男生对比表现出此题目对境内男生有利，DIF 为 B 级；境内、外女生对比同样表现出此题目对境内女生有利，DIF 为 C 级。分析原因如下：

表 5

对比组	标准化的固定偏移比	标准误	DIF 等级	题目有利于	差异			
					选 A	选 B	选 C	选 D
境内外男生对比	-1.4	0.36	B	境内	11	-1	-12	2
境内外女生对比	-1.7	0.34	C	境内	7	4	-14	3

由表中可以看出，标准化方法在答案中识别出 DIF，<C>是正确答案。日本男生境内外对比组的STD P-DIF(C)为-12%，目标组（境外考生）选 C 的比率远低于参照组（境内考生）。STD P-DIF(A)

为 11%，目标组（境外考生）选 A 的比率远高于参照组（境内考生）；日本女生境内外对比组的 STD P-DIF(D)为-14%，STD P-DIF(A)为 7%。两个对比结果均表明，境内考生选 C 的多，境外选 A 的多。此题目有利于境内考生。可能由于题目中使用了一个对男朋友的称呼语，并且涉及当代中国青年女子的恋爱观。境外考生由于没有语言学习的社会大环境，不太了解这部分内容。

例 2：某一语法题，请考生从给出的 A、B、C、D 四个近义词中选择一个正确答案。两个对比组同时表现出此题目对境外考生有利，DIF 分别为 B、C 级，分析原因如下：

表 6

对比组	标准化的固定偏移比	标准误	DIF 等级	题目有利于	差异			
					选 A	选 B	选 C	选 D
境内外男生对比	1.07	0.45	B	境外	-7	6	2	-1
境内外女生对比	1.57	0.39	C	境外	-7	9	0	-1

由表中可以看出，<B>是正确答案。两个对比组境外考生选 B 的多。此题目表现出有利于境外考生。但仔细分析题目，找不到出现这种结果的原因，唯一可以解释的原因是境外的日本考生对这部分的语法知识掌握的更好。

例 3：某一题目反应中国社会生活，请考生从给出的 A、B、C、D 四个词语中选择一个正确答案。两个对比组表现出此题目有利于境内考生，DIF 均为 C 级，分析原因如下：

表 7

对比组	标准化的固定偏移比	标准误	DIF 等级	题目有利于	差异			
					选 A	选 B	选 C	选 D
境内外男生对比	-1.5	0.39	C	境内	6	3	-13	3
境内外女生对比	-1.7	0.41	C	境内	5	2	-12	4

由表中可以看出，<C>是正确答案，中国境内考生答对的多。究其原因此题目反映中国社会生活，考查考生在真实语境中准确运用这四个词语的能力。涉及到语境时，在中国学习汉语的考生自然占优势。

例 4：某一题目，请考生在空格处填写合适的汉字。此题目有利于境内考生，DIF 均为 C 级，分析原因如下：

表 8

对比组	标准化的固定偏移比	标准误	DIF 等级	题目有利于	差异			
					选 A	选 B	选 C	选 D
境内外男生对比	-2.1	0.39	C	境内	-15	0	0	0
境内外女生对比	-2.1	0.34	C	境内	-16	0	0	0

由表中可以看出，DIF 显著。此题目有利于中国境内考生。原因在于，题目涉及的内容在中国的医院里有，而在日本表现形式却不一样。

以上只是从统计学的角度对题目进行分析，有其局限性。为此，在处理 DIF 为 C 级的题目时，要视具体情况而定，一般应该遵循两个原则：

① 有利于境外和有利于境内的题目数量应该尽量保持平衡。

② 如果判定某一题目所测的是测验希望测验的内容，即使 DIF 值较大，也应将此题目保留下来。如要删除，也需要经过训练有素的测试编制者和有关问题专家联合决定。

6. 结论

由表 2 可知，以全部题目的总分为匹配变量，即汉语的综合能力相当的考生相互比较，由于所考查的是考生的多种不同的能力，DIF 结论不可靠，但却可以看到，中国境内考生听力占优势，而境外考生语法和阅读成绩好于听力。在中国境内的绝大部分考生都在中国学习过汉语，正是由于具备运用汉语的社会大环境，他们的听说和交际能力更强。

以各分测验的总分为匹配变量，也就是对听力部分总分相同的考生进行比较，分析听力题目的差异，对语法部分总分相同的考生进行比较，分析语法题目的差异，以此类推。由表 4 可知，虽然女生对比组比男生表现出更多的 DIF 题目，但是趋势是一致的。有利于中国境内考生的题目多于境外考生，不过，所占比例很小。

因此，HSK（初、中等）A 试卷的题目对于中国境内与境外考生基本公平。

注：(1)考虑到种种原因，以上例子中的题目内容不便公开。

参考文献

[1] HSK 成绩中关于女性考生公平性的分析　任杰　李航　对外汉语教学论文集（2000）

[2]许雪立(1999) 关于非亚裔团体 HSK 初中等成绩的公平性分析

[3] DIF detection and description: Mantel-Haenszel and Standardization. Neil J. Dorans & Paul W. Holland (1993).In **Differential Item Functioning**, Paul W.Holland &Howard Wainer (ETS), 1993, pp35-66. Lawrence Erlbaum Associates, Hillsdale, New Jersy.

[4] Item and Test Charateristics That are Associated with Differential Item Functioning. Kathleen A.O'Neill and W.Miles Mcpeek(1993). In **Differential Item Functioning**, Paul W.Holland & Howard Wainer (ETS), 1993, pp255-276. Lawrence Erlbaum Associates, Hillsdale, New Jersy.

[5] Practical Questions in the Use of DIF Statistics in Test Development. Michael Zieky (1993). In **Differential Item Functioning**, Paul W. Holland & Howard Wainer (ETS), 1993, pp337-347. Lawrence Erlbaum Associates, Hillsdale, New Jersy.