HSK 主观考试评分的 Rasch 实验分析*

田清源
(北京语言大学 汉语水平考试中心,北京 100083)

摘 要:主观评分中存在的不一致性导致主观评分的信度降低。多面 Rasch 模型基于项目反应理论,可以应用于评分员效应的识别和消除,从而提高主观评分的信度。该文介绍多面 Rasch 模型的理论和应用框架,设计了基于该模型的 HSK 主观考试评分质量控制应用框架,利用 HSK 作文评分数据进行了实验验证。

关键词:项目反应理论;多面 Rasch 模型;心理测量;主观评分;汉语水平考试
中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2007)01 - 0065 - 04

1 引言

主观考试一般要求被试者按照规定完成一定的综合性任务,评分员对于被试者完成任务的具体表现进行综合评定,给出一个综合分数。这种考试方式的固有特点使得它无法完全被客观考试所取代,许多知名的考试都采用或者部分采用主观考试。

主观考试的评分基本依赖于评分员的主观印象,容易受到评卷员的知识水平、综合能力、爱好、情绪、疲劳等等主观因素的影响。因此,主观考试的不同评分员之间存在着主观差异,同一个评分员在不同的时间也具有主观不稳定性。在评分的准确性(Accuracy/Inaccuracy)、严厉度(Harshness/Leniency)和集中度(Centrality/Extremism)三个方面,评分员自身在多次评分时难以保持一致,不同评分员对于相同被试的评分也难以相同,这些不一致的存在,直接导致评分员自身信度(intra - judge reliability)和评分员之间信度(inter - judge reliability)的降低,从而降低评分结果的信度,国外文献将这种现象称为评分员效应(rater effects)[1]。

为了消除评分员效应,提高主观评分的信度,人们引入了许多方法。从评分体系的管理上,最为常见的方法有两个:一是对评分员进行提前培训,力争让评分员达到统一的评分标准;二是对于相同的被试进行多人评分,使用原始评分的平均数做为评分结果。有研究表明,无论如何进行事前培训,评分员也无法在严厉度上保持一致[2];实际的主观评分中必须考虑工作的负荷,一般不可能让所有的评分员对于所有的被试进行评分,因此,原始分平均数也不能保证公平。从数学的方法上,人们引入了方差分析模型和结构方程模型,但是因为数据的不完整性(并非每一个评分员对于每一个被试都做出评分)以及原始分数的非线性特征(原始分数为等级分数,不是被试特质的线性表示),这些方法的适用都受到了限制。

Rasch 模型是项目反应理论(item response theory)的模型之一,将基本 Rasch 模型拓展为多面 Rasch 模型(many - faceted Rasch model)之后,它提供的统计框架可以消除主观评分中各个方面的因素对于评分结果的影响,提高评分结果的信度[3]。

文章的后续部分将介绍多面 Rasch 模型的原理。之后,基于多面 Rasch 模型,设计 HSK 主观考试评分质量控制应用框架。然后,利用 HSK 作文评分的数据进行模拟实验。

2 Rasch 模型原理及其拓展

2.1 Rasch 模型

项目反应理论之理论基础是:被试的能力是被试的潜在特质(latent trait),它与被试参加的考试以及具体的项目无关。Rasch 模型是项目反应理论的单参数模型,对于项目它只考虑难度参数。如果进行 0/1 评分,被试在某个项目上获得分数的概率可以表示为公式(1)。

$$P_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \quad (1)$$

B_n:被试 n 的能力值;

D_i:项目 i 的难度;

P_{ni}:被试 n 在项目 i 上获得分数的概率。

* 基金项目:北京语言大学资助项目(05YB01)。

对于公式(1)进行数学转换,可以得到公式(2),这就是 Rasch 模型[4]。

$$log\ \frac{P_{ni}}{1-P_{ni}} = B_n - D_i \qquad (2)$$

使用这个模型,可以同时估算项目的难度和被试的能力值,因此,它是一个双面模型。

对于多级评分,Rasch 模型可以做如下拓展:

$$log\ \frac{P_{nki}}{P_{ni(k-1)}} = B_n - D_i - F_{ik} \qquad (3)$$

F_{ik} :对于项目 i,评分等级 k 相对于等级 k - 1 的难度;

P_{nik} :被试 n 在项目 i 评定为 k 的概率;

$P_{ni(k-1)}$:被试 n 在项目 i 评定为 k - 1 的概率。

进一步,假设在考试中有多个任务,而每一个任务又是由若干项目组成,同时,再考虑不同评分员具有不同的评分严厉度,上述模型又可以拓展如下:

$$log\ \frac{P_{nmijk}}{P_{nmij(k-1)}} = B_n - A_m - D_i - C_j - F_{mik} \qquad (4)$$

A_m :表示任务 m 的难度;

C_j :第 j 个评分员的严厉度;

F_{mik} :对于任务 m 项目 i,评分等级 k 相对于等级 k - 1 的难度;

P_{nmijk} :被试 n 在任务 m 项目 i 由评分员 j 评定为 k 的概率;

$P_{nmij(k-1)}$:被试 n 在任务 m 项目 i 由评分员 j 评定为 k - 1 的概率。

使用这个模型,被试能力值、任务难度、项目难度和评分员严厉度能够同时得到估算,它是一个四面 Rasch 模型,是多面 Rasch 模型的一个典型示例。

这个模型中,任务和项目是相对的概念:任务是由若干项目组成,因此,任务难度是组成该任务的所有项目的难度的函数:在把任务做为一个面进行处理的同时,把项目也做为一个面来处理,能够对于不同任务中相同项目之间的难度差异进行估算和比较。

2.2 多面 Rasch 模型的应用框架

基于多面 Rasch 模型开发的统计工具 FACETS,可以同时估算各个面的测量值(logit),还可以估算这些测量值的标准误和符合度统计指标(fit statistics)。测量值中,各面之间的相互作用已经分离,被试的能力值不受其它面的影响。通过符合度统计参数,可以发现异常的原始分数,也可以发现其它各个面上的异质点。比较各面的测量值,深入分析异常原始分数和异质点的原因,可以对于主观评分有一个更加深入和准确的把握[5]。

因为多面 Rasch 模型在估算各个面上的测量值时,已经将各个面之间的相互作用进行了区分和隔离,所以,应用这个框架可以提高测量的区分信度。

国外,多面 Rasch 模型广泛应用于主观考试分数等值、评分员效应识别、试题评审质量控制和考试公平性等等领域[6]。

3 HSK 主观考试评分质量控制应用框架设计

3.1 HSK 主观考试评分质量控制应用框架

中国汉语水平考试(HSK)是由北京语言大学汉语水平考试中心设计研制,为测试母语非汉语者的汉语水平而设立的国家级标准化考试。在高等汉语水平考试中,包括作文和口试两个主观分测验:作文考试要求完成一篇 400 到 600 字的作文,口语考试要求朗读一段文章,另外口头回答指定的两个问题。两个主观分测验在考试后由评分员人工评分。

为了解决主观评分的信度问题,HSK 主观考试采取多评分员独立对一个被试评分,之后计算平均分数的评分办法。为了进一步控制评分质量,成立了统计专家、语言学家和对外汉语教学专家组成的质量监控小组,对于多个评分员之间分数差异较大的评分进行复评,对评分员的表现进行控制。这种主观评分质量控制应用框架效果显著,但也存在一些难以解决的问题,主要包括评分员的表现无法在整个评分员群体中衡量,不同主观试卷之间的难易差别无法衡量,以及评分工作总体负荷沉重等等。

以原有框架为基础,引入多面 Rasch 模型进行数据分析,可以进一步解决这些问题。具体设计如下:

1)对于一次考试,设计一定比例的被试为锚(anchor),所有评分员都对这些被试评分;对于其他被试,每一个被试只由一个评分员评分,评分员不知道锚的存在,或者知道锚的存在,也不知道哪一个是锚。评分数据由多面 Rasch 模型统计工具处理,在锚的连接作用下,每个评分员都能在整个评分员群体中得到衡量。被试的评分中消除评分员严厉度的影响,评分信度得到提高。

2)通过上述统计的 fit 指标,以及残差方差和残差与期望值之间的相关系数等统计指标,可以发现评分不准确以及评分过于集中或者分散等等异常的评分员,及时进行控制。

3)需要比较两次主观考试的难易差别时,多面 Rasch 模型中增加一个代表不同考试的面,在两次评分中使用一定数量的相同评分员,以这些评分员

为锚，实现两次考试的连接。这时，不同考试的难度差异得以发现和剔除，实现了不同主观试卷之间的等值。

4）从工作负荷来看，因为不需要对于每一个被试都进行多次评分，总体工作负荷能够降低。

3.2　应用框架适用性的理论分析

多面 Rasch 模型是基于项目反应理论的模型，它的适用具有一定前提条件。项目反应理论具有两个等价的前提假设：单维假设和独立假设。考察 HSK 主观考试评分中的被试和评分员两个面：对于被试来说，统计上，不同被试的能力值是可以视为相互独立的；对于评分员来说，不同评分员对于相同被试进行评分时，不知道其他评分员对于该被试的评分，因此，他们的评分之间没有相互影响，评分具有独立性。基于这些分析，理论上看，将被试和评分员各视为一个面的双面 Rasch 模型是适用的。

3.3　应用框架适用性的实验验证设计

为了进一步验证应用框架的适用性，利用 HSK 作文考试的实际数据，设计验证性的实验如下：从一次 HSK 作文评分中选取 406 名被试，编号从 1 到 406。这些被试由 8 名评分员评分，编号从 1 到 8。评分等级为 0 到 5 分，辅助以 2 + 、3 - 这样的辅助等级，实际为 12 级计分。为了方便处理，将上述评分转换为 0 到 12 分的计分，称为原始评分。每一个被试由两个评分员评分，被试与评分员的关系如表 1。利用多面 Rasch 模型进行处理的数据中，被试和评分员各为一个面，只在每两个相邻评分员之间保留 10 个共同被试做为锚，以模拟锚只占全体被试部分比例的情形，整理后的数据结构如表 2。利用 Rasch 分析得到的结果，进行两个方面的验证：评分员严厉度的识别以及被试能力值中评分员严厉度差异的消除。

表 1　实验采集的评分员与被试

评分员	被试
1	1-33
2	4-36
3	34-136
4	87-190
5	139-239
6	192-336
7	253-436
8	337-436

表 2　Rasch 模型处理的评分员与被试

评分员	被试
1	1-33
2	24-36
3	79-136
4	139-190
5	192-252
6	241-336
7	337-436
8	337-436

4　实验数据分析

按照实验设计，使用多面 Rasch 模型统计工具处理实验数据，得到被试和评分员测量值。图 1 概要显示了两个面的情况。表 3 为评分员的 Rasch 分析数据，从上向下按照评分员严厉度由高到低排列。数据表明，评分员 2 在所有评分员中最为严厉。评分员 5 最为宽松，评分员 1 次之。表中 InfitMS 是指信息加权均方离差符合度统计指标（information - weighted mean - square fit statistic），OutfitMS 是指常规均方离差符合度统计指标（conventional mean - square fit statistic）。根据这两个指标，评分员 5 趋向于给所有被试相同的分数，而评分员 2 的评分最为分散。

对于 Rasch 模型分析得到的评分员严厉度进行验证：每两个相邻的评分员都有一组共同被试。当相邻评分员严厉度测量值之差大于一个标准差时，计算他们所有共同被试（而非仅仅是锚）的原始分均值，以验证多面 Rasch 模型对于评分员严厉度区分的有效性。参照表 3 数据，对于评分员 2 和 1，评分员 6 和 5，评分员 4 和 5 以及评分员 8 和 7 进行上述比较。这四个评分员组中，都是前面一个评分员的 Rasch 分析严厉度高。表 4 中计算了四个评分员组所有相同被试原始评分均值，为了便于比较，也摘列了 Rasch 分析得到的评分员严厉度测量值。原始分均值计算使用了各组评分员所有的共同被试，数量最少为 33 个，最多为 70 个，而 Rasch 分析时相邻评分员只选用 10 个共同被试作为锚。表 4 中，Rasch 分析严厉度高的评分员，所有共同被试的原始评分的均值都明显偏低，这验证了 Rasch 分析在区分评分员严厉度差异上的有效性。

对于被试能力值中评分员严厉度差异的消除进行验证：表 5 中列出由评分员 1 和评分员 2 各自给出原始评分为 8 的两个被试，数据经过多面 Rasch

模型的处理,评分员严厉度的差异得以消除,虽然原始得分相同,被试 10 由宽松的评分员 1 评分,他的能力测量值为 - 0.19 logit,而被试 40 由最严厉的评分员 2 评分,他的能力测量值为 3.39 logit。

表 3 评分员 Rasch 分析数据

评分员	严厉度(logit)	标准误	InfitMS	OutfitMS
2	2.05	0.39	0.75	0.74
8	1.50	0.49	0.49	0.44
3	1.28	0.35	0.72	0.70
4	0.22	0.35	0.50	0.47
6	- 0.40	0.36	0.61	0.57
7	- 0.83	0.33	0.61	0.70
1	- 1.59	0.49	0.42	0.41
5	- 2.23	0.34	0.26	0.28

表 4 评分员严厉度 Rasch 分析的验证

评分员	被试人数	被试原始分数均值	原始分标准差	Rasch 分析评分员严厉度(logit)
2	33	5.45	1.371	2.05
1	33	6.42	1.173	- 1.59
6	61	6.43	1.408	- 0.40
5	61	6.84	1.344	- 2.23
4	55	6.31	1.763	0.22
5	55	7.16	1.686	- 2.23
8	70	6.00	1.274	1.50
7	70	7.10	1.625	- 0.83

表 5 被试 Rasch 分析数据摘录

被试编号	评分员编号	原始分数	调整后分数*	被试能力值(logit)
10	1	8.00	7.02	- 0.19
40	2	8.00	9.12	3.39

* 从被试能力值(logit)转换得到的量表分数。

5 结语

5.1 主观评分往往存在各种不一致性,它们表现为评分员效应,导致主观评分的信度降低。对于项目反应理论模型之一的 Rasch 模型进行的多面拓展所得到的多面 Rasch 模型在识别和消除评分员效应上有其优势。

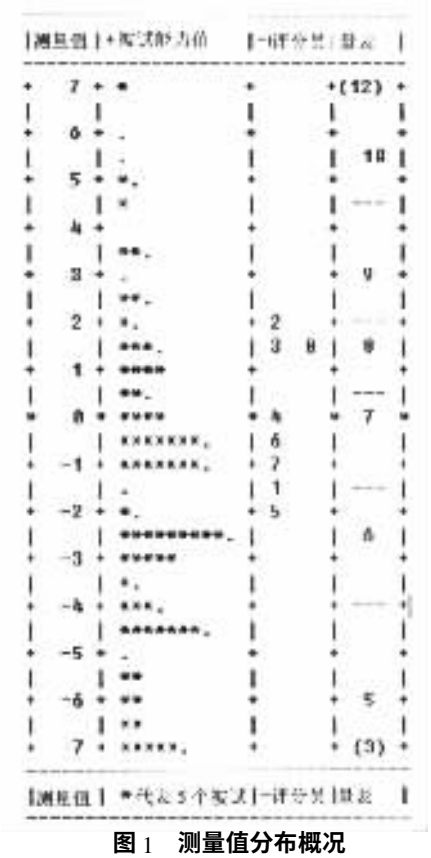

图 1 测量值分布概况

(其中被试能力值(logit) 均值 = - 2.18,标准差 = 3.29;评分员严厉度(logit) 均值 = 0,标准差 = 1.44)

5.2 基于多面 Rasch 模型,设计了 HSK 主观考试评分质量控制应用框架,并利用 HSK 作文评分数据进行了实验,实验数据的分析表明,使用多面 Rasch 模型能够有效区分评分员严厉度的差异,提高主观评分的信度。

5.3 评分员准确性判断方法,以及主观试卷的等值,今后将另外著文研究。

参考文献

1 Wolfe E W. Identifying rater effects using latent trait models. Psychology Science,2004,46(1):35 - 51.

2 Lunz M E,Wright B D, Linacre J M. Measuring the impact of judge severity on examination scores. Applied Measurement in Education,1990,3:331 - 345.

3 Linacre J M. Many - Facet Rasch Measurement. Chicago,IL: MESA,1994.

4 Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press,1960/1980.

5 Linacre J M. Facets - Rasch measurement computer program. Chicago, IL: MESA Press, 2003.

6 田清源. 主观评分中多面 Rasch 模型的应用. 心理学探新,2006,1:70 - 73.

Rasch Experimental Analysis of HSK Performance Test Rating

Tian Qingyuan
(HSK Centre, Beijing Language and Culture University, Beijing 100083)

Abstract: In the performance rating, reliability is bated by its subjective inconsistency. Based on the item response theory, many – faceted Rasch model provides a framework to improve the rating reliability by identifying and reducing the rater's effects. In this paper, the theory and application framework of this model is introduced. Based on this model, a quality control framework is designed for the HSK performance test rating, and an experiment is taken using the rating data of the HSK composition test.

Key words: item response theory; many – faceted Rasch model; psychometrics; performance rating; HSK test

---

(上接第 33 页)

selection task. Cognition, 1995, 57(1): 31 – 95.

15 Almor A, Sloman S A. Reasoning versus text processing in the Wason selection task – a non – deontic perspective on perspective effects. Memory & Cognition, 2000, 28(6): 1060 – 1070.

16 Beller S, Spada H. The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. Thinking and Reasoning, 2003, 9(4): 335 – 378.

17 Fiddick L, Cosmides L, Tooby J. No interpretation without representation: The role of domain – specific representations and inferences in the Wason selection task. Cognition, 2000, 77(1): 1 – 79.

18 Stone V E, Cosmides L, Tooby J, et al.. Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. Proceedings of the National Academy of Sciences, 2002, 99(17): 11531 – 11536.

19 Fiddick L, Spampinato M V, Grafman J. Social contracts and precautions activate different neurological systems: An fMRI investigation of deontic reasoning. NeuroImage, 2005, 28(4): 778 – 786.

20 Goel V. Evidence for dual neural pathways for syllogistic reasoning. Psychologia, 2003, 32: 301 – 309.

Perspective Effects in the Wason Four – card Selection Task

Yang Qun　Qiu Jiang　Zhang Qinglin
(Department of Psychology, Southwest University, Chongqing 400715)

Abstract: Perspective effects in the Wason four – card selection task occur when people choose mutually exclusive sets of cards (P, – Q and – P, Q) depending on the perspective they adopt when making their choices. It has been a robust phenomenon in both deontic and non – deontic reasoning and explained by different theories. A debate focuses on the domain specific or domain general procedure people use in the reasoning. In the present research, previous studies were reviewed and debates on this issue were analysed, which was expected to contribute to the future studies to further reveal the mystery of human reasoning.

Key words: Wason four – card selection task; perspective effects; domain specific; domain general