主观评分中多面 Rasch 模型的应用*

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,介绍了国外相关的典型应用,并且讨论了该模型的应用条件。

关键词:项目反应理论;多面 Rasch 模型;心理测量;主观评分;考试

中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2006)01 - 0070 - 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)等等三个方面,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,要考虑实际的工作负荷,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,将基本的双面 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]。

本文的后续部分将介绍多面 Rasch 模型的原理。之后,用几个国外的相关研究举例说明这个模型的应用。最后,对介绍进行总结。

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)。

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

B_n:考生 n 的能力值;

D_i:项目 i 的难度;

P_{ni}:考生 n 在项目 i 上获得分数的概率。

对于公式(1)进行数学转换,可以得到公式(2),

* 基金项目:北京语言大学资助(05YB01)。

这就是 *Rasch* 模型[4]。

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

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：

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

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：

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

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。举例说明如下：假设进行动物组织切片制作的考试，要求被试制作某种动物的 5 种组织的切片，每个切片从组织切取、组织处理和切片染色 3 个方面进行评分，那么，每个切片可以视为一个任务，3 个方面可以视为 3 个项目。显然，任务（某一种组织切片）的难度是它的组成项目（组织切取、组织处理和切片染色）的难度所决定的；然而，对于不同的任务（不同的组织），相同的项目（如组织处理）可能具有不同的难度。进一步拓展这个示例，如果这个考试要求对于 2 种不同动物提交上述组织切片，这个时候可以将不同的动物视为不同的任务，而每一种组织切片视为任务的组成项目，上述模型仍然适用；此时，如果要求同时研究切片评分中的组织切取、组织处理和切片染色 3 个方面，还可以将这 3 个方面视为项目中的子项目，将上述模型进一步拓展成为五面 Rasch 模型。

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参数（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。

3　主观考试分数等值的应用

应用多面 Rasch 模型估算得到各个面上的测量值，不仅是各面上测量结果的线性转换，而且已经将各个面之间的相互作用分离开来，因此，它所估算的被试能力测量值直接可以用作等值之后的分数。

Lunz 于 1990 的研究结果是一个主观考试分数等值应用的典型例子。这个研究是针对组织学切片临床考试而进行的，217 个被试按规定各自提交 15 个组织切片（即 15 个测试项目），由 18 位评分员评定分数。原始分数由三个部分组成：组织切取和组织处理，都取值 0/1，分别表示不可接受和可以接受；染色质量，取值从 0 到 3，代表很差到超出普通水平。三个部分相加为原始分数，因此，原始分数取值从 0 到 5。

评分过程安排两天时间，首先是 3 个小时的评分培训，之后进行评分。因为将所有被试的所有切片（$217 \times 15 = 3255$ 个）都交给每一个评分员进行评分从工作量安排上是不现实的，所以，该研究中进行了如下评分设计：将 15 个切片 5 个一组分成 3 组，将每一个被试的 3 组切片轮回发放到不同的评分员，轮回发放的设计中确保任何两个评分员都会拿到某一些相同被试的不同切片组。在这个评分设计中，使用被试做为连接手段（anchor）实现了不同评分员之间的连接（link），因此它不要求每一个评分员对

于每一个被试的每一个切片进行评分，每个评分员的工作量得到了降低。

应用多面 Rasch 模型，同时估算各面的测量值、标准误和符合度统计参数。

分析测量值数据，18 个评分员中，严厉度的最低值为 -1.19 logit(标准误 0.24)，最高为 1.08 logit(标准误 0.20)；15 个测试项目中，项目难度最低为 -0.62 logit(标准误 0.10)，最高为 0.85 logit(标准误 0.07)。被试的表现最低为 -1.0 logit，最高为 4.0 logit，平均为 1.1 logit1。在各面测量值的同时计算时，各面之间的相互影响已经隔离，因此，被试的能力值中评分员严厉度的差异已经得到了纠正，表 1 以两个被试的评分数据对这个纠正过程进行了说明。两个被试 C 和 D 的原始分数都是 64，但是由于被试 C 所对应的评分员更加严厉，经过多面 Rasch 模型估算后被试 C 的表现是 1.67 logit，而被试 D 只有 0.85 logit。

借助符合度统计参数的分析，可以发现各个面上的异常点。例如，通过分析评分员的符合度统计参数，Lunz 在其研究工作中发现某评分员的评分过于一致(Infit 0.6，Outfit 0.7)，这个数据提醒研究人员和评分管理人员进一步分析该评分员是否有对所有被试给出相近分数的倾向。

Engelhard 在 1992 年将多面 Rasch 模型应用于作文评分的实验研究。研究中选取了 4 个面 被试、评分员、作文题目(8 个不同作文题)、作文评分子项目(如文体、句子结构等 5 个子项)。该研究中，不仅分析了评分员的影响，而且发现不同的作文题目之间和不同的评分子项之间都存在着难度差异[6]。

表 1 经过评分员严厉度纠正之后的被试测量值与原始分数的比较

项目	评分员及其严厉度			平均严厉度	被试表现		原始分数	评分员一致性	
	1 组	2 组	3 组		测量值	标准误		Infit	Outfit
被试 C	7	8	6						
	-0.23	0.71	-0.02	0.15	1.67	0.34	64	0.8	0.9
被试 D	2	17	11						
	-0.48	-1.19	-0.31	-0.67	0.85	0.34	64	1.0	0.9
	相差			0.82	0.82	0.48			

4 评分员效应识别的应用

主观评分的研究中，一般主要关注评分员严厉度差异对于分数造成的影响。然而，评分员的不一致性不仅表现在严厉度的差异上，还表现在准确度和集中度的差异上。Wolfe 2004 的研究工作中，明确提出并且数学定义了评分员效应的三种分类，即在严厉度、准确度和集中度上的不一致性，同时提出了使用多面 Rasch 模型对于这三类评分员效应进行识别的方法。

Wolfe 提出了两个主观评分的统计指标 :残差方差和残差与期望值之间的相关系数，分别用 SDr 和 Rr-E 表示。残差的定义是评分员评分与分数期望值的差。利用多面 Rasch 模型估算得到的测量值，通过统计转换可以计算得到相关的分数期望值。通过分析研究，Wolfe 指出这两个指标与评分员效应之间存在如下关系：

(1)评分准确时，残差方差很小，残差与期望值的相关系数也很小；

(2)评分不准确时，残差方差大，但残差与期望值的相关系数很小；

(3)评分集中时，残差方差很小，残差与期望值之间为负相关；

(4) 评分分散时，残差方差大，残差与期望值直接为正相关；

(5)如果不存在上述评分员效应，只存在严厉度上的评分员效应，残差方差很小，残差与期望值的相关系数也很小。

显然，只使用上述指标无法区分严厉还是宽松，然而，多面 Rasch 模型直接估算出评分员的测量值(logit)就是评分员的严厉度，使用这个数据可以直接进行评分员严厉度的分析。

5 试题评审质量控制的应用

多面 Rasch 模型还可以应用于考试开发过程中专家审题的质量控制。Zhu，Ennis 和 Chen 在 1998 的研究工作中，就将多面 Rasch 模型应用到试卷开发中审题结果的处理。

该研究的实验工具是一个 150 道试题的价值取向测试(VOI－2)。128 个大学的教育工作者和 103 个体育学校的教育工作者做为审题专家参加实验对这些试题(或称为项目)进行评价。实验中设计 6 个面：性别、种族、任职状况、评分员严厉度、内容域和试题。为了减轻每个专家的工作量，实验进行了如下连接设计：将 150 道试题分为 5 组，其中一组做为共同题，与剩下 4 组分别组成 4 套 60 道试题的试卷。4 套试卷均匀寄送给 231 位专家，专家按照试卷中附带的书面要求进行 1 至 5 分的试题评审。虽然没有一位专家对于 150 道试题进行了完整的评审，但是预先设计的共同题连接方案使得所有试题的专家评分在经过多面 Rasch 模型处理之后都转换到同一个量表之上。

借助多面 Rasch 模型进行数据处理可以得到各个面上的测量值(logit)和符合度参数(fit statistics)，通过综合分析这些数据，该研究得到了一些有趣的发现，例如在评分员严厉度之上，女性要比男性宽松，大学教育工作者要比体育学校的宽松，白人要比其他人种宽松等等。该研究最后还指出，对于统计工具所给出的异质点，不应该简单的删除，而应该深入探讨这些异质点出现的原因，以提高审题工作的管理水平[7]。

6　结语

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。除了本文介绍的主观评分等值、评分员效应识别和试题评审质量控制等等典型应用以外，多面 Rasch 模型还可以有许多其它应用，例如将被试的各种分类做为数据处理的一个面，可以进行考试公平性的研究。

然而，这个模型的应用也具有它自身的限制。首先，与基于传统理论的方法相比，项目反应理论的分析方法具有更高的计算复杂性；其次，该模型具有严格的前提假设，在应用之前必须对其适用性进行考查；另外，对于多评分员多试卷的主观评分，必须合理设计连接方案；最后，应用该模型时一般都有复杂的评分管理设计，它要求阅卷评分的组织管理具有一定的信息化水平，完全依靠手工的阅卷评分管理难以满足它的要求。

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　Engelhard G, Jr. The measurement of writing ability with a many－faceted Rasch model. Applied Measurement in Education, 1992, 5：171－191.

7　Zhu W, Ennis C D, Chen A. Many－faceted Rasch modeling expert judgment in test development. Measurement in Physical Education and Exercise Science, 1998, 2：21－40.

Application of Many－faceted Rasch Modeling in Performance 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 an 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, as well as three existing research results are introduced as examples. The conditions of the application of this model is also discussed in this paper.

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