

主观题评价中基于IRT的评分者效应模型

宫子恩 朱殷睿

浙江师范大学教师教育学院

DOI:10.12238/mef.v5i2.4838

[摘要] 主观题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、疲劳等主观因素的影响, 因此在评分的准确性、严厉度和集中度等方面容易导致评分者效应。为了控制评分者效应的大小, 本文介绍了基于项目反应理论的多面Rasch模型、评委束模型、阶层评分者模型及等级反应多水平侧面评分者模型, 并对其进行了评述。

[关键词] 主观题; 评分者效应; 评分者模型

中图分类号: G640

文献标识码: A

Rater Effects Model Based on IRT in Subjective Question Evaluation

GONG Zien, ZHU Yinrui

College of Teacher Education, Zhejiang Normal University

[Abstract] The score of the subjective test basically depends on the subjective impression of the rater, and is easily affected by the knowledge level, comprehensive ability, hobbies, emotions, fatigue and other subjective factors of the rater. Therefore, it is easy to lead to rater effects in terms of scoring accuracy, severity and concentration. In order to control the size of rater effects, this paper introduces and comments on the many-facet Rasch model, rater bundle model, hierarchical rater model and multi-level side rater model of hierarchical response based on project response theory.

[Key words] subjective question; rater effects; rater model

1 引言

主观题作为一种重要的题目形式, 由于形式多样、考查灵活以及能够探查考生深层次的能力, 备受测验编制者和命题者的青睐, 已成为众多大型测验和考试的重要组成部分。在大型测评项目中, 为控制评分偏差, 通常需要多个评分者重复对学生的试卷进行评定。与传统的单一评定相比, 考生得到某一特定分值的概率不仅取决于考生的能力与项目的难度, 而且受到评委的宽严程度、评定量表上特定分数等级的难度等因素的影响。这种评分者自身以及不同评分者之间在多次评定中, 评分准确性、宽严度和集中度的不一致性, 会直接导致评分者内信度 (intra-rater reliability) 和评分者间信度 (inter-rater reliability) 的降低, 从而降低评分结果的信度, 这种现象称为评分者效应 (Rater effects)。

从测量学的角度, 估计评分者效应的测量模型持续发展, 其估计的精度也逐步提高。例如Linacre提出的多面Rasch模型 (Many Faceted Rasch Model, MFRM), Wilson和Hoskens的评委束模型 (Rater Bundle Model, RBM), 以及Patz的阶层评分者模型 (Hierarchical Rater Model, HRM); 康春花和孙小坚等在等级反应模型 (GRM) 基础上提出多水平多侧面模型。

2 基于IRT的评分者模型

2.1 多面Rasch模型

多面Rasch模型是对传统Rasch模型的拓展, 可以独立估计评分者宽严程度、考生能力和项目难度等, 并校正主观评分中各因素对评分结果的影响, 提高评分结果的信度。多面Rasch模型Linacre的表达式如下:

对于0, 1计分情形:

$$\log odds(k: k-1) = \theta_n - \delta_i - \rho_{ir} \quad (1)$$

式中, θ_n 是被试的能力参数; 是项目的i的难度参数; 是评分者r在项目i上的宽严度参数;

对于多级情形:

$$\log odds(k: k-1) = \theta_n - \delta_i - \gamma_{ik} - \rho_{ir} \quad (2)$$

式中, 其他符号的意义与 (1) 式相同, γ_{ik} 是项目的参数分布 (多级评分时)。

如果定义为评 $\pi_r(x_{ni})$ 分者r给学生n在项目i上得 x_{ni} 分的概率, 则评分者给所有学生n打分的分数向量n的似然方程为:

$$L(x_{ni}) = \prod_i [\pi_r(x_{ni})] \quad (3)$$

如果假设有几个不同的评分者对学生的作业进行评定, 则此时学生n的

分数向量就有两套指标,即项目和评分者指标,

$$L(x_{ni}) = \prod_{i,r} [\pi_r(x_{ni})] \quad (4)$$

Patz认为该模型未充分评估多个评分者对同一作业评定之间的相依性。

2.2 评委束模型

多面Rasch模型没有考虑到评分者之间的局部相依性,测验中条件相依性或局部相依性最普遍的现象就是项目子集共用同一个刺激材料,如阅读理解测验。Rosenbaum提出了项目束独立的方法来解决这一问题,其思路并非是把似然方程表示为单个项目概率之积,取而代之的是把似然表示为项目集(束)概率之积,项目束(题组)内容各项目之间是条件相依的,而在项目束之间认为是条件独立的。Wilson和Adams把项目束独立性概念应用到分部评分模型(PCM)及其相关模型,称为项目束模型的模型,对其进行拓展,得到评委束模型。假定考虑两个评分者的情景,把评定了相同学生作业的评委定义为一束(评委束),在0-1计分情形下,如表1所示。

表1 两个评分者评定相同学生作业的评委束模型(学生n,项目i)

两名评分者所评定的分数		
1	2	$\log odd$
0	0	$-k_{i12}$
1	0	$\theta_n - \delta_i - \rho_{i1}$
0	1	$\theta_n - \delta_i - \rho_{i2}$
1	1	$2\theta_n - 2\delta_i - 2\rho_{i1} - \rho_{i2} - k_{i12}$

表中中和分别为评分者1和2的宽严度参数,为额外参数(Extra Parameter)表示评分者评定的交互作用参数,是表示条件相依性的参数,当不为0时,增加了评分者之间的不一致性,通常情况下,我们希望这种不一致性尽量少发生。总的来讲,项目束模型是一种用来解释项目束内部项目之间局部相依性的模型,而评委束模型是用来解释当评分者对这些项目提供了重复评定时评价局部相依性的模型。

2.3 阶层评分者模型

为了对评分者相依性进行校正,Patz提出了HRM模型。HRM模型通过组合

来自多个评分者的信息以了解被试表现、项目参数,它解释了对同一被试作业的不同评定者之间的边缘相依性,通过把数据产生过程分解为两阶段而校正了传统IRT侧面模型低估标准误差偏差的问题。

在第一阶段,假定存在一个理想不可观察的评定变量,描述被试i在项目j上的成绩。该变量遵循标准的分步计分模型(PCM)或其它IRT模型,HRM表达式如下:

$$\log it P[\xi_{ij} = \xi | \theta_i, X_{ijr} \in \{\xi, \xi - 1\}] = \theta_i - \beta_j - \gamma_{j\xi} \quad (5)$$

其中, $\theta_i \sim i.i.d. N(\mu, \sigma^2)$, $i = 1, \dots, N$; $\xi_{ij} \sim$ IRT模型,如PCM, $j = 1, \dots, J$; $\xi_{ij} \sim$ 信号检测模型, $r = 1, \dots, R$ 。 ξ_{ij} 可以看作在等级得分点

的所有可能的反应空间中,对被试i在项目j反应理想反应。

在第二个阶段,一个或多个评分者对被试i在项目j上的作答反应产生一个评分k,这个分数可能与理想评定类型一致。在HRM中,这一评分过程用离散信号检测问题来模型化,通过评分概率矩阵进行计算。Mariano证实,在侧面模型下,当评分者而非项目增加时,测验信息会无限制性提高。他通过严格证明,表明在HRM下,不管评分者数目如何,测验信息会因测验长度得到合理的限定,并且其能力估计的标准误差永远不会小于无评分者理想评定易时的IRT模型。

2.4 等级反应多水平侧面评分者模型

目前,常用估计评分者信度的是多面Rasch模型及多水平随机系数模型(MRCM)等。MRCM通过将评分者和受评者分离,在评分者层面加入预测变量来分析其对评分结果的影响,但MRCM无法估计宽严度参数,且是线性模型,而评分者和受评者的心理特质及其行为之间的关系常常是非线性的。在同时考虑评分者的系统偏差和评分者特质对评分结果影响的情况下,多面Rasch模型及多水

平随机系数模型并不适用。为此,Hung和Wang将多面Rasch和MRCM相结合发展出广义线性或非线性混合模型。Wang和Liu基于广义分部评分模型(GPCM)发展出广义分部评分多水平侧面模型(G-MLFM),以适合CR题的多级评分情境。等级反应模型由于不对每步加工过程本身做出反应,避开了因加工任务类型的不同而可能导致测验中各项目提供的信息量被低估的问题。

因此,在多面Rasch模型、MRCM和GRM的基础上构建等级反应多水平侧面模型(Grade Response Multilevel Facets Model, GR-MLFM),用于建构反应题如数学应用题问题解决,解法多样的评分情境,以考察评分者的宽严度及其相关的影响因素。

水平1公式:

$$\log it(p_{nij}^*) = \log \left[\frac{p(x \geq j)}{p(x < j)} \right] = a_i(\theta_n - B_{ij} - D_{kn}) \quad (6)$$

水平2公式:

$$\theta_n = \gamma_{00} + \gamma_{01}X_1 + \gamma_{02}X_2 + \mu_{0n}, \mu_{0n} \sim N(0, \sigma_{\theta}^2) \quad (7)$$

$$D_{kn} = \gamma_{k0} + \gamma_{k1}W_1 + \gamma_{k2}W_2 + \mu_{kn}, \mu_{kn} \sim N(0, \sigma_k^2) \quad (8)$$

$$\text{其中, } p_{nij}^* = \frac{\exp[a_i(\theta_n - B_{ij} - D_{kn})]}{1 + \exp[a_i(\theta_n - B_{ij} - D_{kn})]}$$

表示被试获得等级j及其以上等级概率。 θ_n 为第n个被试的能力值, γ_{00} 指被试总体的能力均值, μ_{0n} 为被试n能力值的随机效应,服从均值为0,方差为1的正态分布, X 为被试能力的预测变量。 a_i 为项目i的斜率, B_{ij} 为项目i从反应类别j-1到类别j的阈值。 D_{kn} 为评分者k评定被试n时的宽严度, γ_{k0} 为评分者K的平均宽严度, W 为影响评分者宽严度的预测变量, μ_{kn} 为评分者k评定被试n的随机效应,服从均值为0,方差为的正态分布。 $p(x \geq j)$ 为被试在项目i上反应类别为j及以上类别的概率, $p(x < j)$ 为被试在项目i上的反应类别低于类别j的概率。

上式呈现了一个两水平GR-MLFM的完整模型,水平2有两个随机成分:被试能力和评分者特质。该模型也可形成3水平模型,如在水平2的基础上,被试(评分者)嵌套于学校或组织等,依次可形成更多水平的模型。

3 总结与讨论

上述前三种模型都是随机系数多分逻辑特模型(RCMLM)的子模型,并且都是Rasch模型的拓广形式。其中HRM是一个二阶的潜结构模型,该模型本质上可以包含其他模型,其不同之处在于多了理想评定的中间层。如果将中间层理想评定变量移除,则在给定情况下,所有评定者观察评定变成局部独立了,此时评定数据的似然值是通过把公式(2)中所有观察到的被试、项目、评分者组合连乘而得到。当使用标准测验信息函数计算,Patz认为当每个项目的评分者数增加时,即使不增加项目,IRT侧面模型会产生被试能力其精确的测量值,同时IRT侧面模型会出现低估标准误差偏差的现象。模型拟合研究也同样表明,IRT侧面模型的形式没有遵循真实评定数据和明确考虑由于嵌套于评分者内或者嵌套于被试内的评分相依性的变异轨迹。

基于等级反应模型的评分者效应模型在多项模拟研究中表现出良好的性能,但是这一模型仍需继续探究。第一,模型方面。首先是包含预测变量(如评分者的自信心、责任心、情绪稳定性及评分经验)的全模型的模拟和实证研究。

其次是3水平及多水平模型的研究,如被试或评分者嵌套于学校或组织等。第二,难度参数随机效应模型,未来研究可把影响项目难度的刺激特征或认知属性加入模型,使模型更具普适性。

总之,评分者评分的模型估计参数的精度、准确度越来越高,可供选择的模型有很多,但在应用时仍需依据数据类型来选择适合的模型、方法。尽管多面Rasch模型没有考虑到评分者之间的相依性,但应用时间最长,及其简洁性与便利性,多面Rasch模型及它的变式是上述各模型中应用最广的模型。关于RBM以及HRM的研究相对出现较晚,较少的一些研究主要集中在模型的比较方面。GR-MLFM提出时间较短,其应用性还有待扩展。

【参考文献】

- [1] Wolfe, E.W. Identifying Rater Effects Using Latent Trait Models[J]. Psychology Science, 2004(45):35-51.
- [2] 李中权, 孙晓敏, 张厚粲, 等. 多面Rasch模型在主观题评分培训中的应用[J]. 中国考试(研究版), 2008(01):26-31.
- [3] Patz, R.J., Junker, B.W., Johnson, M.S., et al. The Hierarchical Rater Model For rated Test Items and its Application to Large-scale Educational Assessment Data[J]. Journal of Educational and Behavioral Statistics, 2002, 27(03):302-314.
- [4] Wilson, M. & Adams, R.J. Rasch

Models for Item Bundles[J]. Psychometrika, 1995, 60(02):181-198.

[5] Mariano, L.T. Information Accumulation, Model Selection and Rater Behavior in Constructed Response Student Assessments[J]. Carnegie Mellon University, 2002, 35(01):274-279.

[6] Hung L.F., Wang W.C. The Generalized Multilevel Facets Model for Longitudinal Data[J]. Journal of Educational and Behavioral Statistics, 2012, 37(02):231-255.

[7] Wang W.C., Liu C.Y. Formulation and Application of the Generalized Multilevel Facets Model[J]. Educational and Psychological Measurement, 2007, 67(04):583-605.

[8] 纪凌开. 分部评分模型与其它几种多级模型的比较[J]. 心理科学, 2004, 27(04):1000-1001.

[9] 康春花, 孙小坚, 曾平飞. 等级反应多水平侧面模型及其在主观题评分中的应用[J]. 心理科学, 2017, 40(06):1483-1490.

作者简介:

宫子恩(1996-), 男, 汉族, 山东威海人, 硕士研究生在读, 研究方向: 基础教育测量。

朱殷睿(1993-), 男, 汉族, 浙江金华人, 硕士研究生在读, 研究方向: 测验等值方法。