文章编号:1000-5862(2015)05-0441-08

多级计分题项目功能差异常用检测方法及比较

张 龙,涂冬波*
(江西师范大学心理学院,江西 南昌 330022)

摘要:项目功能差异是确保测验公平的统计技术手段.多级计分题目为教育测量和心理测量中不可或缺的题型,而目前还未见有公开发表的文章较为全面地将常用多级计分题 DIF 检测方法作一概括,该文从参数类与非参数类 2 个视角对多级计分 DIF 检验方法进行论述与比较,为实践应用者在方法选用上提供借鉴及支持,最后对多级计分 DIF 检验进行讨论.

关键词:项目功能差异;多级计分题;检测方法
中图分类号:B 842.1 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.05.01

0 引言

项目功能差异(Differential Item Functioning,DIF)指来自不同群体但具有相同能力水平或熟练水平的被试对某个题目正确回答的概率不同,则这个题目就存在项目功能差异[1],项目功能差异是衡量试题是否对某个群体有偏差现象、对试题的公平性进行综合评价的过程[2].测验的有效性、稳定性和公平性是测验质量的重要方面,前两者有效度、信度等相应的指标来衡量,而公平性则迟迟未受到人们的关注,直到 Coffman 使用统计显著性检验考察组间差异的思路来进行项目功能差异分析,DIF 检验才发展起来.自 20 世纪 90 年代至今,DIF 已成为测试研究领域的一个热点问题[3].1999 年,美国《教育与心理测试标准》对测试公平不光单独将其列出,还将其作为全书的 3 大板块之一给予前所未有的重视,同时,也提出高利害测验实施之前必须在不同性别、年龄、种族、文化背景、语言背景的考试群体间进行项目功能差异分析的要求[4].

在实际中,多级计分题目为教育测量和心理测量中不可或缺的题型,如心理测量中的 Likert 型量表多为 5 或 7 级计分,它是在我国广泛应用的态度、人格自陈量表的主要形式之一,而且通常在教育测量中多级计分题目(如论述题)的重要性更大,因此更有必要保证这些题目的公平性;此外,多级计分题目多为主观题,它更易受到文化和环境因素的影响从而导致 DIF.因此,针对多级计分题目进行 DIF 检测的必要性日益凸显,然而目前国内学界焦点主要集中在 2 级计分题 DIF 检测方法的理论研究和应用上,对多级记分题 DIF 研究涉及较少,仅有文献[5-6]进行过 2 项多级计分 DIF 实证研究;文献[7]在项目反应理论框架下对 Likert 型多级计分题的 DIF 检测方法进行探究,以及文献[8]将多级计分题 DIF 检测方法作为变通的题组 DIF 检测方法引入篇章阅读测验 DIF 检验.多级计分 DIF 检测方法研究已经发展比较成熟,其中既有由 2 级计分 DIF 检测方法改进后得到,也有针对多级计分题而提出,因此本文对常用多级计分题目 DIF 方法做一个较为全面的阐述及比较,为实践应用者在方法选用上提供借鉴及支持.

1 非参数类多级评分 DIF 检验方法

多级计分题目的 DIF 检测方法可以划分为 2 类:(i)非参数检测方法;(ii)参数检测方法.非参数类方法直接计算 DIF 指标,不涉及确定的回归模型或测量模型,也不涉及总体或模型的参数而关注总体的分布形态,并且具有对样本容量的要求小、方法简便直观、易于理解的优点,包括 GMH、the Mantel

收稿日期:2015-04-17
基金项目:国家自然科学基金(31100756,31300876,31160203,31360237),教育部人文社科项目(11YJC190002),高等院校博士点基金(20123604120001),江西省社会科学规划重点项目(13JY01),江西省教育科学规划项目(12YB088,13YB029)和江西师范大学青年英才培育计划资助项目.
通信作者:涂冬波(1978-),男,江西南昌人,副教授,博士生导师,主要从事心理统计与测量的研究.

test、Poly-SIBTEST、P-STND 等.

1.1 GMH 和 the Mantel test

GMH(generalized Mantel-Haenszel χ^2)[9] 和 the Mantel test[10] 是由 MH 方法推广而来,是 2 种简单且实用的非参数方法,它们适用于 2 种不同的情况,the Mantel test 适用于当反应类别是顺序变量时,它通过比较某个题目 2 个组群间的平均分来实现,而 GMH 则不需要反应类别是顺序变量的前提,它通过比较某个题目在 2 个组群间的得分分布来实现.

GMH 的指标为

$$\chi^2_{\mathrm{GMH}} = \left[\sum_{m=0}^{T} \boldsymbol{N}_m - \sum_{m=0}^{T} E(\boldsymbol{N}_m)\right]' \left[\sum_{m=0}^{T} V(\boldsymbol{N}_m)\right]^{-1} \cdot \left[\sum_{m=0}^{T} \boldsymbol{N}_m - \sum_{m=0}^{T} E(\boldsymbol{N}_m)\right],$$

其中 m 为匹配得分,一般使用测试总分($m = 0,1,2,\cdots,T$),$\boldsymbol{N}_m$ 为一个有 $c-1$ 个元素的向量,c 为计分等级数,里面的元素是参照组在 m 匹配情况下该得分等级的人数,$E(\boldsymbol{N}_m)$ 为在假设不存在 DIF 的情况下理想人数,$V(\boldsymbol{N}_m)$ 为该向量的方差向量;该指标服从自由度为 $c-1$ 的卡方分布.

The Mantel test 的指标为

$$\chi^2_{\mathrm{Mantel}} = \left(\sum_{m=0}^{T} F_m - \sum_{m=0}^{T} E(F_m)\right)^2 \Big/ \sum_{m=0}^{T} \sigma^2_{Fm},$$

该方法即为 MH 方法的一个顺延,它将 MH 方法的 $m \times 2 \times 2$ 的列联表拓展为 $m \times c \times 2$ 的列联表,m 为匹配得分,一般使用测试总分($m = 0,1,2,\cdots,T$),F_m 为目标组所有人在 m 匹配得分水平上所有分数之和,$E(F_m)$ 为在假设不存在 DIF 的情况下理想的分数之和,S^2 为该水平上所有分数的方差,该指标服从自由度为 1 的卡方分布.

为了更好地解释 the Mantel test 方法的效果,Penfield 等[11] 移植 Liu-Agresti 累计比值比指标来作为该方法的效果量,该指标为

$$\hat{\varphi}_{LA} = \left(\sum_{m=0}^{T} \frac{1}{N_m} \sum_{j=1}^{c-1} A_{mj} D_{mj}\right) \Big/ \left(\sum_{m=0}^{T} \frac{1}{N_m} \sum_{j=1}^{c-1} B_{mj} C_{mj}\right),$$

其中 m 为匹配得分,一般使用测试总分($m = 0,1,2,\cdots,T$),N_m 是一个有 $c-1$ 个元素的向量,c 为计分等级数,里面的元素是参照组在 m 匹配情况下这个得分等级的人数,A_{mj},B_{mj},C_{mj} 和 D_{mj} 是该题在这个等级的参照组与目标组 2×2 答题表的人数分布,该效果量倒数的对数为 0 时表示没有 DIF. 同时,Penfield 还开发了 DIFAS 软件来计算该对数值[12].

Jorge Carvajal 等[13] 通过操纵样本大小、DIF 形态、群体差异 3 个变量的模拟实验,结果显示即便在样本量小于 200 时,该指标的一类误差仍控制较好,且其作为效果量的表现不受样本大小影响.

Wang Wenchung 等[14] 模拟实验表明 The Mantel test 在题目区分度变化和群体间的能力不等时表现出膨胀的一类误差,而增加用来匹配被试的题目数量有助于减少一类误差. R. Zwick[15] 研究发现在题目是一致性 DIF 时,the Mantel test 的统计检验力比 GMH 要高,但当题目存在非一致性 DIF 时,GMH 表现更好,但两者的缺陷是都不能检测非一致性 DIF. 同时,Wang Wenchung 建议,为更好地控制一类误差,匹配分数应为一组无 DIF 的试题加上待研究试题的分数. Ángel M Fidalgo 开发 GMHDIF 软件[16],该软件可以实现 GMH、the Mantel test、MH 方法,也可对 2 级或多级计分题进行检测.

1.2 Poly-SIBTEST

该方法是 Chang Huahua 等[17] 对同时性项目偏差估计 SIBTEST(Simultaneous Item Bias Test)[18] 在方法针对多级计分题目的拓展和改进,它具有对原有方法改动小、简单易用、可以对一组题目同时进行检测的优点,但使用该方法的前提是需要有一组无 DIF 题目作为锚题、满足目标组和参照组独立、两群体之间的平均水平没有差距的条件. Poly-SIBTEST 和 SIBTEST 的基本逻辑一样,一个项目不存在 DIF 即在被试能力匹配水平 θ 上,参照组与目标组的被试在该项目上的平均得分一样,其效应值指标为

$$\hat{\beta} = \sum_{m=0}^{T} w_m (\bar{Y}_{RM} - \bar{Y}_{FM}),$$

其中 m 为匹配得分,一般使用测试总分($m = 0,1,2,\cdots,T$),F_m 为目标组所有人在 m 匹配得分水平上所有分数之和,$\bar{Y}_{RM}$,$\bar{Y}_{FM}$ 分别为参照组和目标组总分为 m 的被试在该题的平均分,w_m 是权重,为得分为 m 的人在总人数中的比例.

效应值可知 DIF 的大小,但还需要统计检验量来告诉它是否显著,有如下呈正态分布的统计检验量:$\hat{\beta}_{UNI} = \beta / \hat{\sigma}_\beta$,其中 $\sigma^2_\beta = \sum_{m=0}^{T} w^2_m \left(\frac{\hat{\sigma}^2_{Rm}}{N_{RM}} - \frac{\hat{\sigma}^2_{Fm}}{N_{FM}}\right)$,$\hat{\sigma}^2_{Rm}$ 为参照组中得分为 m 被试在该题得分方差,N_{RM} 为参照组中得 m 分的总人数,在 0.05 的显著性水平上,当 $|\hat{\sigma}^2_{Rm}| > 1.96$ 时,表明该题存在显著 DIF,目前可用软件 Poly-SIBTEST(DIFPACK 1.7) 来对 2 级计分和多级计分题进行检测.

D. Bolt[19] 模拟实验表明,Poly-SIBTEST 比 the Mantel test 具有统计检验力更高、更准确的特点,并且当作答与模型不拟合时,Poly-SIBTEST 不受影

响,而 GRM-LR、GRM-DFIT 则表现出膨胀的一类误差,Chang Huahua 等[17] 进行的模拟研究比较了 the Mantel test,P-STND,Poly-SIBTEST 法,研究结果表明当项目区分度变化时,较之 the Mantel test 和 P-STND 法,P-SIBTEST 法对一类错误控制更好. 但是 L. R. Gabriel [20] 发现 Poly-SIBTEST 对非一致性 DIF 不具统计检验力. 另外 R. Shealy [18] 基于模拟数据的实验表明,SIBTEST 至少需要 20 道锚题,而 Poly-SIBTEST 方法需要多少道锚题则尚无明确推荐的标准.

1.3　P-STND

STND(Standardization) 方法即标准化方法,由 N. J. Dorans [21] 提出. 该方法的基本逻辑是,若一个项目无 DIF,则有 $E_{FM}(Y \mid X) = E_{RM}(Y \mid X)$.

$E(Y \mid X)$ 为项目分数对测验分数水平 Z 的回归,即项目分数对测验分数的回归在理论上是不受群体的影响而应该完全相等的. 在计算时,STND 指标是计算各个匹配分数水平上 2 组人群的正确作答比例之差的加权平均数,然后求和. M. T. Potenza 等[22] 提出的 P-STND 沿用了这个思路,只是在计算时更为细致,其指标为

$$STND_{ED\text{-}DIF} = \sum_{m=0}^{T}\left[E_{FM}(Y \mid X) - E_{RM}(Y \mid X)\frac{N_{FM}}{N_F}\right], \quad (1)$$

其中 m 为匹配得分,一般使用测试总分($m = 0,1,2,\cdots,T$) ,此处 $E(Y \mid Z)$ 表示该组被试在 m 匹配得分水平上该题的平均分,N_{FM}/N_F 是权重,为得分为 m 的目标组人数占目标组总人数的比例.

P-STND 方法与其他非参数方法如 Poly-SIBTEST 有类似之处,也具有简单易用的特点,Fang Tian [23] 发现它具有不能检测非一致性 DIF,具有易受样本容量影响. P. Narayanan [24] 研究还表明它还易受到群体之间分布的差异性影响,当群体水平不一致时存在一类误差膨胀的缺点,因此在实际中较少使用.

2　参数类多级评分 DIF 检验方法

多级计分题目的 DIF 检测方法另一类是参数检测方法,它多涉及项目反应理论或其他模型,在使用过程中需要求解相关模型的参数,操作起来相对复杂,但当试题与模型拟合时 DIF 检验结果更精确,它包括 LDFA、GRM-LR、PHGLM、GRM-DFIT 等方法.

2.1　逻辑斯蒂克判别函数分析法(LDFA)

逻辑斯蒂克判别函数分析法 LDFA [25] (Logistic Discriminant Function Analysis) ,它在逻辑斯蒂回归方法的基础上发展而来. 通常在用逻辑斯蒂回归检验 2 级计分题目的 DIF 时,将被试作答反应 U 视为因变量,被试能力变量 X 和群组变量 G 作为自变量,即估计 $P(U \mid X,G)$. 但当处理多级计分题目的时候,反应变量 U 的多个取值,使得逻辑斯蒂回归方程不再适用,但逻辑斯蒂回归法检测 DIF 具有方便易用、准确率高、一类误差小、可检验非一致性 DIF 等优良特性,因此较多研究者想将该方法推广运用到多级记分题中,但大多朝 P-LR 方法(Polytomously Logistic Regression Procedure) 的方向努力,如此一来,每一个多级计分题的数据都要分解成若干个 2 值记分题的数据组,也就是每个题目均要用一组模型来表示,该方法不但难以理解,计算起来也相当复杂不便,实际应用中也不曾使用,因此在本文中不予介绍,由于上述理由,Miller 等进一步提出可以组群变量为因变量建立逻辑斯蒂克回归方程来估计 $P(G \mid X,U)$. 该方法的模型为

$$P(G \mid X,U) = \frac{e^{(1-G)(-a_0-a_1X-a_2U-a_3XU)}}{1+e^{(-a_0-a_1X-a_2U-a_3XU)}}, \quad (2)$$

$$P(G \mid X,U) = \frac{e^{(1-G)(-a_0-a_1X-a_2U)}}{1+e^{(-a_0-a_1X-a_2U)}}, \quad (3)$$

$$P(G \mid X,U) = \frac{e^{(1-G)(-a_0-a_1X)}}{1+e^{(-a_0-a_1X)}}.$$

在这一系列回归方程中,方程(3) 为紧缩模型,方程(1) 和方程(2) 为扩展模型,方程(3) 为被试属于哪个群体仅仅与他的能力有关,方程(2) 为被试属于哪个群体还与他的作答有关,这即符合一致性 DIF 的含义; 方程(3) 为被试属于哪个群体也与他的能力与作答的交互作用有关,这即符合非一致性 DIF 的含义; LDFA 方法用检验 α_2,α_3 系数显著性的方法来检测 DIF,系数 α_2 用来检测一致性 DIF,α_3 用来检测非一致性 DIF.

系数的显著性检验是通过比较 2 个模型似然函数的对数差进行的,2 个相邻等级模型的似然函数对数差近似地服从自由度为 1 的卡方分布; 分别计算出这 2 个模型似然比卡方拟合统计量 G^2,通过检验 G^2diff 是否达到统计显著水平,就可以得知该系数的显著性,从而得知项目是否存在非一致性 DIF 和一致性 DIF.

LDFA 作为逻辑斯蒂回归方法的一种,继承了 P-LR 可以检测一致性和非一致性 DIF 的优点,同时,该方法使得一个项目的 DIF 检验只需建立一个回归方程. 克服了 P-LR 需用一组回归方程来描述一个项目的 DIF 和计算复杂的缺点,并且使用起来高

效、简便.

该方法目前有 Juana Gómez-Benito 等[13] 提出了 2 个效果量: 拟合度(R^2) 和条件似然比(ΔLR),实验显示,显著性检验加上该效果量后,可以有效降低一类误差. J. Spray 等[25] 和 Fang Tian[23] 发现 LDFA 在检测一致性 DIF 时与其他方法表现的一样好,而 Wang Wenchung 等[14] 的模拟实验表明 LDFA 和 the Mantel 在检测非一致性 DIF 时比 GMH 方法更具检验力. 然而,当群体间的水平差异超过一个标准差时,LDFA 方法的一类误差会急剧失控; Fang Tian[23] 也有相似的发现. Elizabeth Kristjansson 等[26] 比较了 4 种检验方法(the Mantel test,GHM,LDFA,累积逻辑斯蒂回归) 发现这 4 种方法在检测一致性 DIF 时检验力均较高,但当试题的区分度 b 高时 LDFA 方法的检验力会比较低.

2.2 GRM-LR

S. Kim 等[27] 提出似然比法(likelihood ratio test,LR). 它基于等级反应模型(GRM),因为项目反应理论(IRT) 的假设是项目参数的估计不受考生能力的影响,因此 LR 法的基本思路是通过检测 2 个群体的项目参数是否有差异来侦探 DIF,具体通过比较一个待测题的参数在 2 个群体相等的模型(紧缩模型) 和一个待测题的参数在 2 个群体上变动的模型(扩张模型) 来实现,其假设是每个项目在 2 个群体上的各个项目参数相等.

H_0: $a_{jF} = a_{jR}, b_{ijF} = b_{ijR}(j = 1,2,\cdots,m)$;

H_1: 该题的参数至少有 1 个不相等;

该方法具体实现过程如下:

(i) 将试题分为 2 个部分锚题,即没有 DIF 的题和待检测是否存在 DIF 的题;

(ii) 分开估计目标组和参照组的项目参数和能力参数,得出锚题的线性关系;

(iii) 用锚题的线性关系作为联系 2 个群体的桥梁得出 2 个群体的能力的线性关系;

(iv) 将待检测的项目加入到锚题中,构建紧缩模型(compact model) 和扩张模型(augmented model),比较 2 个模型的似然函数值,其中紧缩模型通过等值限定 2 个群体的所有题目(包括待检测题) 的参数均相同,即假设没有 DIF 的情况,在同一参数尺度上得到对数似然函数值; 扩展模型即针对待检测项目,释放项目参数对 2 组被试相同的限制,容许该项目在 2 组被试的参数分开自由估计,其他项目仍限定参数相同,得到总的对数似然函数值;

(v) 计算 2 个模型下对数似然函数 -2 倍的差为 $G^2 = -2(\log l_{\text{compact}} - \log l_{\text{augmented}})$,其中 l 为 likehood,服从卡方分布,自由度等于该项目自由变动的参数个数. 如果大于卡方临界值,则认为该项目存在 DIF.

LR 方法使用的效果量是两群体的期望反应函数的平均绝对差(average unsigned difference, AUD),再以目标组的密度加权

$$AUD = \left(\sum_{q=1}^{Q} |ERF_R(\theta) - ERF_F(\theta)| g_F(\theta)\right) / Q,$$

其中 ERF 即该能力值在该题上的期望得分,它把每一个等级的得分乘以该等级的概率值再累加,Q 为积分权点,一般在 -4 到 4 的能力范围内以 0.1 为步长,共 81 个积分权点.

该方法可以通过 Multilog 程序间接实现,也可以直接通过 D. Thissen[28] 开发的 IRTLRDIF v2.0b 软件实现. 同时,C. M. Woods[29] 表明该方法的统计检验力会随样本量、题目区分度、锚题数、DIF 大小的增加而增加. C. M. Woods[30] 的实验表明,在群体间水平存在差异或锚题存在功能差异时,LR 方法会表现出膨胀的一类误差. C. M. Woods 还比较了 GRM-LR 和 Poly-SIBTEST、GMH、the Mantel tests,发现即便在违背被试群体成正态分布的假设下,LR 方法也比其他方法一类误差更小,表现更为稳健,Poly-SIBTEST 的一类误差最大,这表明尽管 Poly-SIBTEST,GMH、the Mantel test 是非参数方法,对被试和题目要求满足的前提更少,但却并不如 LR 方法.

2.3 PHGLM

由于 HGLM(hierarchical generalized linear logistic model) 广义多层线性逻辑斯蒂模型在解释具有结构性的数据如教育考试或成就测试时具有独特的优势,加上它相比 IRT 模型而言可以同时估计出项目和能力参数的优点,由此使得它被学界重视起来,Natasha J Williams 等[31] 开发了 PHGLM(polytomous hierarchical generalized linear logistic model) 多级计分广义线性逻辑斯蒂模型来检测多级计分题 DIF,以反应类型为 3 级的题目(如同意、一般、不同意为例),第 1 层是题目指示层,可表示为 $\eta_{1ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{(k-1)j}X_{(k-1)ij}$, $\eta_{2ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{(k-1)j}X_{(k-1)ij} + \delta_j$,其中 η_{1ij} 为被试 j 对题目 i 的反应类型为第 1 级的概率与反应类型为第 2、3 级的概率比值的对数值,η_{2ij} 为反应类型为第 1、2 级的概率与反应类型为第 3 级的概率比值的对数值,X_{qij} 为题目指示变量,一般 k 个题目有 $k-1$ 个变量,当下标 $q = i$ 时,即当要分析第 1 题时,$q = 1$,第

1 个指示变量就等于 - 1,而其他指示变量等于 0,δ_j 为题目等级间的区分度差异,在不同题目和被试间均保持恒定.

第 2 层是被试层,表示为 $\beta_{0j} = \gamma_{00} + \gamma_{01}(gender)_j + u_{0j}$,$\beta_{1j} = \gamma_{10} + \gamma_{11}(gender)_j$,…,$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(gender)_j$,$\beta_{1j} - \beta_{(k-1)j}$ 为题目效应值,它在不同题目间变动但在不同被试间恒定,β_{0j} 为截距,包含了所有被试的平均题目效应 γ_{00} 和在不同题目内保持恒定但在不同被试间变动的题目效应,以及在个人层面对被试作答产生影响的因素,如性别 g,γ_{q1} 表示性别的效应值,若显著则表示存在性别方面的 DIF. 结合这 2 层可得到等级 1 与等级 2、3 的概率对数比为 $\eta_{1ij} = -\left[(\gamma_{q0} - \gamma_{00}) + (\gamma_{q1} - \gamma_{01})g_j\right] + u_{0j}$. 等级 1、2 与等级 3 的概率对数比为 $\eta_{2ij} = -\left[\gamma_{q0} - \gamma_{00}) + (\gamma_{q1} - \gamma_{01})g_j\right] + \delta + u_{0j}$.

在层级 2 中用到的性别变量可以是 2 分变量,也可以是其他的连续变量,因此,PHGLM 方法的独特优势是除了可以对 2 个群体间 DIF 进行检测外,还可以对多个群体或某个连续的群体变量进行分析,该方法的另一独特优势是可以对 DIF 的成因进行分析,即如果把想考察的解释变量加入回归方程后,之前存在的 DIF 现在变得不显著,那么就可以认为该因素是影响 DIF 的原因之一. 该方法的缺点是操作起来会比较复杂,它的实现可用 HLM 软件.

Natasha J Williams 等[31] 对 PHGLM 和 GMH 进行了比较研究,发现 2 种方法统计检验力基本接近并都相当高,只有在 DIF 题目的比例高于 50% 时才会降低准确性. Cari H Ryan [32] 对上述研究进行了后续研究,发现 GMH 在大容量样本的情况下表现与 PHGLM 相当,2 种方法均能较好地控制一类误差.

2.4　GRM-DFIT

GRM-DFIT 是 C. P. Flowers 等 [33] 提出的适用于基于等级反应模型(GRM) 的多级计分项目和测验功能差异方法(Differential Functioning of Items and Tests,DFIT),它不但可以进行项目层面的功能差异检测,还可以进行测验层面的功能差异检测. DFIT 法的基本思路是: 具有相同潜在特质的 2 组被试,如果它们在某项目或测验上的真分数不一样,则认为该项目或测验存在功能差异.

DFIT 法的具体做法是:

(i) 先分开估计目标组和参照组的项目参数,从而获得了 2 套项目参数;

(ii) 然后用从两群体能力得来的线性转换关系将参照组的项目参数转换到和目标组的参数相同量尺上,此时就有 2 套在同一个尺度上的项目参数;

(iii) 再使用目标组的能力分布分别计算出每个项目上 2 套期望得分,然后计算 2 者之差.

也就是说对于目标组被试 s,可以得到 2 个期望得分,一个是使用目标组参数得到的,另一个是使用转换后参照组的项目参数得到的. 因此,目标组被试 s 在项目 i 上的真分数之差为

$$ES_i(\theta_s) = \sum_{k=1}^{m} P_{ik}(\theta_s) X_{ik},$$

$$d_i(\theta_s) = ES_{iF}(\theta_s) - ES_{iR}(\theta_s),$$

$P_{ik}(\theta_s)$ 为目标组中能力水平为 θ 的被试 s 在项目 i 的 k 等级上的答对概率,$ES_{iR}(\theta_s)$ 和 $E_{SiF}(\theta_s)$ 分别为多级计分模型下,参照组和目标组中被试在项目 i 上的期望得分.

NCDIF(noncompensatory DIF) 指标检测项目层面上的功能差异,它和大多项目水平 DIF 指标一样,假设除了研究项目外其他项目都是无 DIF 的,项目间不可以相互补偿. 题目 i 在目标组所有被试上的项目功能差异指标如下,E_F 表示目标组能力分布下的期望: $NCDIF_i = E_F\left[d_i(\theta_s)^2\right]$.

NCDIF 指标的显著性检验指标可以表示为

$$\chi^2_{NF} = \sum_{s=1}^{NF} (d_{is} - \mu_{di})^2 / \sigma^2_{di} = N_F(NCDIF_i) / \sigma^2_{di},$$

其中 N_F 为目标组人数,对于 NCDIF 的显著性检验,N. S. Raju 等 [34] 提出一般的做法是: 根据不同项目计分等级设置不同临界值,对于 2 级计分题临界值为 0.006,3 级为 0.016,4 级为 0.054,5 级为 0.096,6 级为 0.15. 实际使用中可以对多级计分题型使用 Multilog 软件对参照组和目标组作答数据分别估计参数,等值后使用 N. S. Raju [35] 的 DFIT8 软件对题目进行检测. D. Bolt [19] 研究发现,GRM-LR 法比 DFIT 法效果更好,并且即便在小样本情况下,IRT-LR 法与 DFIT 法这 2 种要求大样本的方法检验力也均优于 Poly-SIBTEST 法. 随着样本容量增加,DFIT 对轻度非一致性的 DIF 统计检验力有所提高,同时,当试题与模型不拟合时,相比 GRM-LR 方法,DFIT 方法也表现出较低的一类误差.

3　总结和讨论

综合以上的方法来看,非参数的方法虽然具有简单易操作的优点,但皆具有对非一致性 DIF 缺乏

检验力的缺陷,而在多级计分题目中更容易出现非一致性 DIF,即在一个得分等级上偏向目标组而在另一个得分等级上偏向参照组,由此使用参数方法在检测非一致性 DIF 时成为必须. 而目前的参数类方法又可以划分为 2 类:(i) 以线性回归为基础的方法,如 LDFA,P-HGLM,P-LRDIF,这类方法的结果比其他方法更为精确,但缺点为操作复杂且较费时;(ii) 以 GRM 为基础的方法,如 GRM-DFIT,GRM-LR,这类方法的结果也都比较稳健和精确,但由于是基于 IRT 构造,因此对样本容量有一定的要求.

表 1 多级评分 DIF 检测方法及其特点概览

类型	方法	特点		相应软件
非参数类多级评分 DIF 检验方法	GMH	比较题目在 2 群组间的得分分布;题目存在非一致 DIF 时,统计检验力比 the Mantel test 好	不能检验非一致性 DIF	GWHDIF
	the Mantel test	适用于顺序变量的反应类别,比较题目在 2 群组间平均得分;检验一致性 DIF 时统计检验力比 GMH 好		
	Poly-SIBTEST	可对 1 组题目同时进行检测;比 the Mantel test 有更准确更低的一类误差膨胀	需满足目标组和参照组独立,且 2 个群体之间平均水平无差距;对非一致性 DIF 缺乏检验力;	Poly-SIBTEST (DIFPACK 1.7)
	P-STND	方法简单易懂,容易操作	易受样本容量、群体之间差异性影响;不能检测非一致性 DIF,实际应用较少	
参数类多级评分 DIF 检验方法	LDFA	一类误差小、一致性 DIF 检验力较强;可检验非一致性 DIF	当群体间的水平差异过大时,LDFA 方法的一类误差会急剧失控;当试题的区分度高时检验力会比较低	HLM、R
	GRM-LR	比起非参数方法来更为稳健	在群体间水平存在差异或锚题存在功能差异时,LR 方法会表现出膨胀的一类误差,需要大样本	IRTLRDIF v2.0b
	PHGLM	适用解释具有结构性的群体数据;还可分析多个群体或某个连续的群体变量;一类误差控制较好;能对 DIF 成因进行分析	操作比其他方法更为复杂,花费更大	HLM
	GRM-DFIT	可以进行项目层面的功能差异检测和测验层面的功能差异检测	计算较为繁琐,需要大样本	Huhilog,DFIT8

多级计分方法的发展经过这么多年研究者的努力,已经有了长足的进步,但相对于 2 级计分的 DIF 检测方法仍显得落后许多,这不仅体现为方法在数量上远少于 2 级计分,也体现为目前存在方法操作的复杂性且较费时,还没有一个既操作简单又功能全面的方法,也即在多级计分 DIF 检测的方法中还没有一个在经济性和有效性上达到平衡,但总体来说还是以基于 IRT 的参数方法为最优选择,从发展的角度,本文对多级计分题 DIF 检测的未来研究方向做出以下几点展望:

1) 从理论的角度来看,开发出能检测一致性和非一致性 DIF,并且操作简单、易于理解的新方法是值得努力的方向,同时目前方法均针对 2 个群体间的 DIF 检测,这在实际使用中存在较大局限性,如同时针对多个民族、多个地区的被试群体进行 DIF 检测就存在问题. 因此从发展趋势来看,能同时在多群体间进行 DIF 检测的方法是一个可行的发展方向[31].

2) 当前的多级计分 DIF 研究在探讨方法优劣时主要使用模拟数据作为数据来源,但理想的模拟数据与真实数据之间存在较大的差距. 未来研究可以用实证研究来综合比较各方法,探讨在不同的使用条件下方法的优劣,为实际使用者提供借鉴.

3) 在认知诊断测试中也存在多级计分的题目,目前对于认知诊断中的 DIF 研究不多,仅有王卓然等[36],Zhang Wenmin[37],Li Feiming[38],Hou Likun 等[39] 对 DINA、修改的 HO-DINA 模型(modified higher order DINA model) 检测 2 级计分题 DIF,对于认知诊断中多级计分题目 DIF 检测方法尚无人探索. 此外,DIF 检测在认知诊断中也表现出一些不同的特点,如文献[37] 发掘了在认知诊断中独有被试的匹配模式,即将个体的属性掌握模式作为匹配标准,并比较了这 2 类匹配标准(试题总分和个体属性模式),发现在匹配个体属性掌握模式时,DIF 检测结果更为精确[37]. 另有文献[38] 用修改的 HO-DINA 模型(modified higher order DINA model) 检测 DIF 和 DAF(Differential Attribute Functioning)[38],DAF 则代表了群体在属性上的优势和缺点. 因此可

以从如何拓展这些方法、在认知诊断领域是否具有不同的特性和价值入手,这均具有较大的研究价值和前景.

4　参考文献

[1] Kim M. Detceting DIF across the different language groups in a speaking test [J]. Language Testing,2001,18(1):88-114.

[2] Kunnan. Fairness and validation in language assessment: selected papers from the 19th language testing research colloquium [C]. England: Cambridge University Press,2006:1-14.

[3] Juana Gómez-Benito,Ma Dolores Hidalgo,Bruno D Zumbo. Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items [J]. Educational and Psychological Measurement,2013,73(5):875-897.

[4] 美国教育研究协会,美国心理学协会,全美教育测量学会. 教育与心理测试标准 [M]. 燕妮琴,谢小庆,译. 沈阳: 沈阳出版社,1999.

[5] 李莉. 多等级试题项目功能差异(DIF)参数方法的检测研究 [D]. 南昌: 江西师范大学,2005.

[6] 宋丽红. LDFA 方法及其在项目功能差异分析中的应用研究——以高考英语试卷分析为例 [D]. 南昌: 江西师范大学,2008.

[7] 涂冬波,戴海琦. 项目反应理论下 Likert 型量表的 DIF 检测方法初探 [J]. 江西师范大学学报: 自然科学版,2007,31(3):311-315

[8] 郑蝉金,郭聪颖,边玉芳. 变通的题组项目功能差异检验方法在篇章阅读测验中的应用 [J]. 心理学报,2011,43(7):830-835

[9] Somes G W. The generalized Mantel-Haenszel statistic [J]. The American Statistician,1986,40(2):106-108.

[10] Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure [J]. Journal of the American Statistical Association,1963,58(303):690-700.

[11] Penfield R D,Algina J. Applying the Liu-Agresti estimator of the cumulative odds ratio to DIF detection in polytomous items [J]. Journal of Educational Measurement,2003,40(4):353-370.

[12] Penfield R D. DIFAS: differential item functioning analysis system [J]. Applied Psychological Measurement,2005,29(2):150-151.

[13] Jorge Carvajal,William P Skorupski. The effects of small sample size on identifying Polytomous DIF using the Liu-Agresti estimator of the cumulative common odds ratio [J]. Educational and Psychological Measurement,2010,70(6):914-925.

[14] Wang Wenchung,Yeh Yali. Effects of anchor item methods on differential item functioning detection with the likelihood ratio test [J]. Applied Psychological Measurement,2003,27(6):479-498.

[15] Zwick R,Donoghue J,Grima A. Assessment of differential item functioning for performance tasks [J]. Journal of Educational Measurement,1993,30(3):233-251.

[16] Ángel M Fidalgo. GMHDIF: a computer program for detecting DIF in dichotomous and polytomous items using generalized Mantel-Haenszel statistics [J]. Applied Psychological Measurement,2011,35(3):247-249.

[17] Chang Huahua,Mazzeo J,Roussos L. Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure [J]. Journal of Educational Measurement,1996,33(3):333-353.

[18] Shealy R,Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF [J]. Psychometrika,1993,58(3):159-194.

[19] Bolt D. A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods [J]. Applied Measurement in Education,2002,15(2):113-141.

[20] Gabriel L R,Stephen S,Oleksandr S C. The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test [J]. Applied Psychological Measurement June,2008,33:251-265.

[21] Dorans N J,Kulick E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test [J]. Journal of Educational Measurement,1986,23(4):355-368.

[22] Potenza M T,Dorans N J. DIF Assessment for polytomously scored item: A framework for classification and evaluation [J]. Applied Psychological Measurement,1995,19:23-27.

[23] Fang Tian. Detecting DIF in polytomous item responses [D]. Ottawa: University of Ottawa,1999.

[24] Narayanan P,Swaminathan H. Identification of items that show nonuniform DIF [J]. Applied Psychological Measurement,1996,20(3):257-274.

[25] Spray J,Miller T. Identifying nonuniformn DIF in polytomously scored test items [R]. ACT Research Report Series,1994.

[26] Elizabeth Kristjansson,Richard Aylesworth,Ian Mcdowell,et al. A comparision of four methods for detecting differen-

tial item functioning in ordered response items [J]. Educational and Psychological Measurement, 2005, 65(6): 935-955.

[27] Kim S, ohen A S. Detection of differential item functioning under the graded response model with the likelihood ratio test [J]. Applied Psychological Measurement, 1998, 22(4): 345-355.

[28] Thissen D. IRTLRDIF v2.0b: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [D]. Chapel Hill: University of North Carolina at Chapel Hill, 2001.

[29] Woods C M. Empirical selection of anchors for tests of differential item functioning [J]. Applied Psychological Measurement, 2009, 33(1): 42-57.

[30] Woods C M. Likelihood-ratio DIF testing: Effects of nonnormality [J]. Applied Psychological Measurement, 2008, 32: 511-526.

[31] Natasha J Williams, Natasha Beretvas S. DIF identification using HGLM for polytomous items [J]. Applied Psychological Measurement, 2006, 30(1): 22-42.

[32] Cari H Ryan. Using hierarchical generalized linear modeling for detection of differential item functioning in a polytomous item response theory framework: an evaluation and comparison with generalized Mantel-Haenszel [D]. Georgia: Georgia State University, 2008.

[33] Flowers C P, Oshima T C, Raju N S. A description and demonstration of the polytomous-DFIT framework [J]. Applied Psychological Measurement, 1999, 23: 309-326.

[34] Raju N S, vander Linden W, Fleer P. An IRT-based internal measure of test bias with applications for differential item functioning [J]. Applied Psychological Measurement, 1995, 19: 353-368.

[35] Raju N S, Oshima T C, Walach, A. H. DFIT8 [EB/OL]. [2009-11-19]. http: // coeweb. gsu. edu/coshima/EPRS 9360/DFIT8/DFIT8%20manual. pdf.

[36] 王卓然,郭磊,边玉芳. 认知诊断测验中的项目功能差异检测方法比较 [J]. 心理学报, 2014, 46(12): 1923-1932.

[37] Zhang Wenmin. Detecting differential item functioning using the DINA model [D]. Greensboro: The University of Nouth Carolina at Greensboro, 2006.

[38] Li Feiming. A modified Higher-Order DINA model for detecting differential item functioning and differential attribute functioning [D]. Georgia: The University of Georgia, 2008.

[39] Hou Likun, de la Torre J, Nandakumar R. Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model [J]. Journal of Educational Measurement, 2014, 51(1): 98-125.

The Common DIF Detection Methods Introduction and Comparison for Polytomous Items

ZHANG Long, TU Dongbo*

(School of Psychology, Jiangxi Normal University, Nanchang Jiangxi 33002, China)

Abstract: Differential item functioning (DIF) is a statistical technique to ensure a fair test. Multi-level scoring items are indispensable for educational measurement and psychometrics, but there is still no published articles completely described DIF detection method for multi-level scoring items in generally, this article class the non-parametric polytomous DIF detection methods and parameters polytomous DIF detection methods, these two categories of such methods were described and compared, and follow-up development were discussed.

Key words: differential item functioning; polytomous items; detection methods

(责任编辑: 冉小晓)