中国心理学前沿
2019 年 10 月第 1 卷第 8 期

基于项目反应理论的项目功能差异检测方法：现状与挑战

尹　昊*　郭嘉程　赵何钧　刘志炜

曲阜师范大学教育学院心理学系，曲阜

邮箱：qfnuyinhao@163.com

摘　要：在教育和心理测试中，检测和消除项目功能差异（Differential Item Functioning，DIF）对于测验的公平性和有效性具有重要意义。因此，DIF 检测方法的研究态势一直比较活跃，但是以往研究主要集中在单一情境下的模型开发和应用以及各种方法的比较上，不同测验类型和测验情境的 DIF 检测方法尚未完善，对 DIF 来源进行合理解释的研究仍然缺乏。本文在项目反应理论框架下，从参数和非参数两个方面具体介绍 DIF 检测方法的原理及适用条件，阐述了国内外 DIF 检测方法的研究现状和最新进展，对当前研究所面临的问题和未来研究的发展趋势提出建议和展望，以期对研究者后续研究有所助益。

关键词：项目反应理论；项目功能差异；DIF 检测方法

收稿日期：2019-09-10；录用日期：2019-09-29；发表日期：2019-10-11

文章引用：尹昊，郭嘉程，赵何钧，等. 基于项目反应理论的项目功能差异检测方法：现状与挑战 [J]. 中国心理学前沿，2019，1（8）：537-555.
https://doi.org/10.35534/pc.0108034

The Detection Methods of Differential Item Functioning Based on the IRT Framework: Research Status and Challenges

Yin Hao* Guo Jiacheng Zhao Hejun Liu Zhiwei

Department of Education, Qufu Normal University, Qufu

Abstract: Detecting and eliminating differential item functioning （DIF） is of great importance for test fairness and validity in educational and psychological measurements. Therefore, research of DIF detection methods has been relatively active. However, previous studies mainly focused on the models for development and application under a single situation as well as the comparison of DIF detection methods. DIF detection methods for different test types and educational situations have been uncompleted. Moreover, the research on the explanation of the source of DIF is still lacking. In this study, within the framework of item response theory, the principle of DIF detection methods is specifically introduced from two aspects of parameter and non-parameter, and the applicable conditions of various methods are carefully elaborated. Meanwhile, the status and latest progress of DIF detection methods are summarized. Finally, the limitations in the DIF detection methods are put forward, along with several primary perspectives, in order to assist researchers' further research.

Key words: Item response theory; Differential item functioning; DIF detection methods

Received: 2019-09-10; Accepted: 2019-09-29; Published: 2019-10-11

1 引言

在教育与心理测量中，公平性是一个重要的考量。《教育和心理测试标准》认为测验的公平性是指在测试过程中给予被试公平的待遇以及为所有被试提供平等的机会［1］。项目功能差异（Differential Item Functioning，DIF）是影响测验公平性的重要因素。DIF 指被试所属的群体不同（如不同性别、种族、文化、地域等）但其能力水平（或潜在特质）相同时，正确回答某些项目的概率存在差异。如果项目中存在 DIF，则说明被试在该项目上的得分不仅取决于被试的知识或能力水平，而且在很大程度上取决于被试所属的群体，即该项目对不同的群体存在不公平现象。

DIF 的检测方法始于 20 世纪 70 年代，在 20 世纪 80 年代形成一般的统计框架［1］。近年来，很多研究者致力于 DIF 检测方法的开发。综合而言，这些方法可以分为两大类：非参数化检测方法和参数化检测方法。非参数化检测方法包括：逻辑斯蒂克回归（Logistic Regression，LR）［2］、Mantel－Haenszel 方法（MH）［3］、标准化（Standardization）［4］、同时性项目偏差测验（the Simultaneous Test Bias，SIBTEST）［5］［6］［7］和 Breslow－Day（BD）［8］等。基于项目反应理论（Item Response Theory，IRT）的参数化检测方法包括：似然比检验法（Likelihood Ratio Test，LRT）［9］、Raju 面积测量法［10］［11］和 Lord 卡方检验法［12］等。在实际应用中还存在包含多个目标组的情况，因而研究者在两个组（一个对照组和一个目标组）基础上开发了适用于多组的 DIF 检测

https://doi.org/10.35534/pc.0108034

方法。例如，广义的 MH 方法（the Generalized Mantel–Haenszel Test，GMH）[13]、广义的 LR 方法（Generalized Logistic Regression，GLR）[14]、广义的 Lord 卡方测试（Generalized Lord’s Chi–square Test）[15]。

本文将首先介绍测验中常用的三参数 IRT 模型；然后以此为基础，详细阐述 IRT 框架下非参数和参数的 DIF 检测方法，以及各种方法的适用条件和影响因素；最后，指出当前研究的局限性以及未来的研究展望。

2 项目反应理论模型

IRT 描述的是被试在测验项目上的正确作答反应同潜在特质水平之间的定量关系[16]。IRT 模型有一参数、两参数和三参数逻辑斯蒂克模型，分别用 1PL、2PL 和 3PL 表示。3PL 是最常见 IRT 模型之一，它包括三个项目参数：区分度、难度和猜测参数[17][18]。3PL 项目反应函数（IRF）采用以下这种形式：

$$P(x_j=1|\theta_n,\ g_j,\ d_j,\ a_j)=g_j+\frac{1-g_j}{(1+exp[-(d_j+a_j\theta_n)])} \quad (1)$$

其中，θ_n 是第 n 个被试的能力水平，g_j 是猜测系数，d_j 是项目难度，a_j 是项目区分度。在 3PL 中，如果猜测参数值设置为 0，则得到的模型为 2PL 模型，如果令 $a_j=a$ 可以获得 1PL。基于 IRT 的参数化检测方法一般将被试潜在特质水平作为匹配变量，比较不同组被试之间项目参数的差异，如果项目参数估计值在不同组之间存在显著差异，则认为该项目存在 DIF。非参数方法一般将原始分数作为匹配变量，比较不同组被试在测验项目上作答的表现，如果存在显著性差异则认为存在 DIF。在同一项目上，如果一组被试的能力水平均高于另一组被试，说明项目中存在一致性 DIF，若被试的能力水平和组别之间存在交互作用，则说明项目中存在非一致性 DIF。在基于 IRT 的参数化方法中，特定模型的选择影响所假设的 DIF 效应类型[19]。

3 非参数检测方法

3.1 MH 方法

MH 方法［3］［20］是一种常用的检测 DIF 的方法，根据测验总分有条件地检测被试和项目作答反应之间是否存在关联。设 N_m 为目标组和对照组中总分为 m 的被试数，根据被试分组和项目 J 的作答情况，可以将这 N_m 个被试分为一个 2×2 列联表（如表 1）。

表 1　2×2 列联表

Table 1　The 2×2 Contingency Table

组别	项目反应		合计
	答对	答错	
对照组	A_m	B_m	N_{mR}
目标组	C_m	D_m	N_{mF}
合计	N_{m1}	N_{m0}	N_m

表 1 中 A_m 和 B_m 分别是对照组中正确和错误作答项目 J 的人数，C_m 和 D_m 分别为目标组中正确和错误作答项目 J 的人数。对照组和目标组的被试总数分别为 $N_{mR}=A_m+B_m$ 和 $N_{mF}=C_m+D_m$；正确和错误作答项目的数量分别为 $N_{m1}=A_m+C_m$ 和 $N_{m0}=B_m+D_m$。α_{MH} 的计算公式可以表达为：

$$\alpha_{MH}=\frac{N_{mR}}{N_{mS}}=\frac{\sum_{m=1}^{J-1} A_m D_m / N_m}{\sum_{m=1}^{J-1} B_m C_m / N_m} \tag{2}$$

当 $\alpha_{MH}=1$ 时，项目无 DIF 值；$\alpha_{MH}<1$ 时，目标组难度较低；$\alpha_{MH}>1$ 时，对照组难度较低。Mantel 和 Haenszel 为检测 α_{MH} 是否等于 1，提出了 MH_{χ^2} 统计量［3］（以下称 MH）：

$$MH_{\chi^2}=\frac{\left\{\left|\sum_{m=1}^{J-1}\left[A_m-E(A_m)\right]\right|-0.5\right\}^2}{\sum_{m=1}^{J-1} Var(A_m)} \tag{3}$$

其中，

$$E(A_m)=\frac{N_{mR} N_{m1}}{N_m} \tag{4}$$

https://doi.org/10.35534/pc.0108034

和

$$Var(A_m)=\frac{N_{mR}N_{mF}N_{m1}N_{m0}}{N_m^2(N_m-1)} \quad (5)$$

公式（3）的分子中，–0.5 是亚茨（Yates）连续性校正项，目的是使 MH 更好地近似连续型卡方分布。

MH 统计量的虚无假设是没有 DIF 存在，即项目反应和被试之间没有联系，若无 DIF 存在时，MH 统计量遵循自由度为 1 的卡方分布，如果 MH 的值大于临界值时，则断定项目存在 DIF。美国教育测验服务中心（ETS）为与其 Δ 量表相匹配［20］，对 α_{MH} 进行了变换，转换公式如下：

$$\Delta_{MH}=-2.35\ ln(\alpha_{MH}) \quad (6)$$

Mantel 和 Haenszel 提出的 MH 统计量在多个目标组情况下的 DIF 检测能力很差［3］。进而，Somes 提出了 GMH 统计量，并详细介绍了 GMH 统计量及应用［13］，但是 MH 和 GMH 还是更适用于检测一致性 DIF，对于非一致性的 DIF 检测能力很低。因此，Clauser、Mazor 和 Hambleton 提出了一种标准 MH 程序的变式——vaMH［21］，用于非一致性 DIF 检测，并发现其比 MH 和 LR 有更好的统计检验力，并且能更好控制 LR 的一类错误率［22］。

此外，Penfield 比较了在多个目标组情况下 MH 统计量，使用 Bonferroni 不等式对 α 水平进行校正后的 MH 统计量，以及 GMH 统计量三种 DIF 检测方法［23］。其结果表明，在大多数条件下 GMH 是最适合于检测 DIF 的程序，因为它的一类错误率保持在 0.05 显著性水平，并且有很好的统计检验力。在 Penfield 的研究中，影响因素有目标组数、样本量、能力水平分布差异和组内 DIF 数量，研究发现，目标组数量对 GMH 统计量的一类错误率没有影响，样本量的增加会使其统计检验力变好。Guilera，Gómez–Benito，Hidalgo 和 Sánchez–Meca 对用于检测 DIF 的 MH 统计量的相关研究进行元分析［24］，结果发现，测试长度对一类错误率和统计检验力的影响最小；样本量的增加会使统计检验力增大，综合各方面因素样本量在 1000 左右将是最有效的；目标组和对照组的样本量不同会使一类错误率增大，统计检验力降低；使用项目净化程序可以显著降低一类错误率，提高统计检验力；检测 DIF 所用模型与 Rasch 模型的差异越大，一类

错误率越大，统计检验力越小；项目的 DIF 比例越高，一类错误率越大，统计检验力越低，比例在 20% 以上尤为明显。

3.2 标准化与 SIBTEST 方法

标准化方法是一种类似于 MH 的更容易理解的描述和解释 DIF 的方法[25]。标准化方法中，通过比较每组被试正确作答概率和测验总分的比例检测 DIF 是否存在。标准化的 p 值差异（ST–p–DIF）可以看作是目标组和对照组的被试正确作答概率的差异的加权平均值，其表达式为：

$$ST\text{–}p\text{–}DIF=\frac{\sum_{m=1}^{J-1}\omega_m(P_{mF}-P_{mR})}{\sum_{m=1}^{J-1}\omega_m} \quad (7)$$

其中，$P_{mF}=\frac{C_m}{N_{mF}}$ 和 $P_{mR}=\frac{C_m}{N_{mR}}$ 是目标组与对照组内被试正确作答某一项目的概率，ω_m 是目标组内被试测试总分 m 所占比重。ST–p–DIF 统计量取值范围为 –1 到 1，统计量的值越接近 0，则 DIF 项目越少。Dorans 和 Holland 还提出了针对多项选择项目的标准化测试[26]，标准化方法的结果通常与 MH 的结果一致，两种方法之间关系密切，MH 方法中存在的问题也在标准化方法中有所体现。

SIBTEST 方法可以看作是标准化方法的一种推广[5]。SIBTEST 分为可疑分测验和匹配分测验项目两部分，是用于检测一致性 DIF 的方法，比较的是目标组和对照组被试的正确作答概率，可表示为：

$$\hat{\beta}_{uni}=\sum_{c=0}^{C_X}\hat{P}_c\,(\bar{Y}_{Rc}-\bar{Y}_{Fc}) \quad (8)$$

其中，Y 是可疑分测验项目未加权的总分，X 为匹配分测验项目未加权的总分，C_X 为 X 中最有可能取得的总分，$\hat{P}_c$ 为在两组中正确作答该项目的被试的比例，$\bar{Y}_{Rc}$ 和 $\bar{Y}_{Fc}$ 分别是对照组和目标组在匹配分测验上总分为 X=c 时，所有被试所得 Y 值的平均数。

进而，得到式（8）的渐近标准误差估计：

$$\hat{\sigma}\,(\,\hat{\beta}_{uni}\,)=\left\{\sum_{c=0}^{C_X}\hat{P}_c^2\left[\frac{\hat{\sigma}^2(Y_R|c)}{N_{Rc}}+\frac{\hat{\sigma}^2(Y_F|c)}{N_{Fc}}\right]\right\}^{1/2} \quad (9)$$

https://doi.org/10.35534/pc.0108034

其中，$\hat{\sigma}^2(Y_R|c)$ 和 $\hat{\sigma}^2(Y_F|c)$ 分别是匹配测验分数为 C 时，目标组和对照组被试在可疑分测验上总分的方差。N_{Rc} 和 N_{Fc} 是当 $X = c$ 时，对照组和目标组的被试数。计算得出 $\hat{\beta}_{uni}$ 和 $\hat{\sigma}(\hat{\beta}_{uni})$ 后，可得统计量 $\hat{B}$：

$$\hat{B} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\hat{\beta}_{uni})} \tag{10}$$

当项目不存在 DIF 时，SIBTEST 统计量近似服从标准正态分布。式（7）中，两组被试需要有相同的能力水平，但是在实际测验中两组的被试的能力水平可能不同。Li 和 Stout 开发一种新的程序，交叉的 SIBTEST（Crossing SIBTEST, CSIBTEST）[6]，CSIBTEST 能区分项目中存在 DIF 的类型。在实际测验中，目标组和对照组的被试有着不同的能力水平，可以表示为：

$$\hat{\beta}_{cro} = \sum_{c=0}^{k_c-1} \hat{p}_c (\bar{Y}^*_{Rc} - \bar{Y}^*_{Fc}) + \sum_{c=k_c}^{C_X} \hat{p}_c (\bar{Y}^*_{Fc} - \bar{Y}^*_{Rc}) \tag{11}$$

$\bar{Y}^*_{Rc}$ 和 $\bar{Y}^*_{Fc}$ 是去除 DIF 影响后校正的 Y 值平均分，k_c 表示对照组和目标组两条 IRF 曲线交叉点处的值。式（11）的渐近标准误差估计为：

$$\hat{\sigma}(\hat{\beta})_{cro} = \left\{ \left(\sum_{c=0}^{k_c-1} + \sum_{c=k_{c+1}}^{C_X} \right) \hat{p}_c^2 \left[\frac{\hat{\sigma}^2(Y_R|c)}{N_{Rc}} + \frac{\hat{\sigma}^2(Y_F|c)}{N_{Fc}} \right] \right\}^{1/2} \tag{12}$$

由此可得 CSIBTEST 统计量：

$$\hat{B} = \frac{\hat{\beta}_{cro}}{\hat{\sigma}(\hat{\beta}_{cro})} \tag{13}$$

Philip 对 Li 和 Stout 提出 CSIBTEST 的原理进行了改进，提出用大样本 χ^2 假设检验方法代替 CSIBTEST 统计量和随机化方法 [7]。研究结果表明，SIBTEST 和 CSIBTEST 统计量会受到样本量大小的影响，样本量增大会使统计检验力更高。特质分布差异会使 SIBTEST 的一类错误率增加 [27]。与 LRT、MH 方法相比，测试长度对 SIBTEST 的影响更大，增加测试长度会对 DIF 检测结果产生负面影响 [28]。Finch 和 French 研究表明，组间能力差异、DIF 百分比和生成数据的基本模型不会对 SIEBTEST 统计量的一类错误率控制产生影响 [29]。

3.3 LR 方法

LR 方法 [2] 是根据测验总分、被试以及两者之间的相互作用，发展出的一个测量项目正确作答概率的 Logistic 模型。通过对被试主效应的测试，可以检

测出一致性 DIF，通过对相互作用的测试，可以检测出非一致性 DIF。Logistic 回归模型为：

$$logit(\pi_n)=\tau_0+\tau_1 M_n+\tau_2 G_n+\tau_3 (M_n G_n) \quad (14)$$

其中 π_n 是被试 n 正确回答项目 j 的概率，M_n 是被试 n 的总分数，G_n 是不同组的被试，τ_0, τ_1, τ_2 和 τ_3 是回归系数。如果 τ_2=0 且 τ_3=0，项目 j 不存在 DIF；如果项目 j 存在一致性 DIF，则 $\tau_2 \neq 0$ 和 τ_3=0；如果 $\tau_3 \neq 0$，项目 j 有非一致性 DIF。

Millsap 和 Everson 提出可以将 LR 方法推广到两个以上目标组的条件下[25]，Van den Noortgate 和 De Boeck 提出了逻辑混合模型，可以用于检测多组的情况[30]。Kanjee 建议在使用 LR 统计量前，先将所有目标组合并为一个，这种方法避免了成对比较，并可以控制一类错误率膨胀现象，增强统计检验力，但会存在错误标识 DIF 项目的情况[31]。Magis 等人提出了 GLR 方法[14]，是基于 Swaminathan 和 Rogers 的 LR 方法改进得出：

$$logit(\pi_{ng}) = \alpha + \beta M_n + \alpha_g + \beta_g M_n \quad (15)$$

其中，π_{ng} 是 g 组的被试 n 正确作答概率，α 和 β 是斜率和截距，M_n 是被试 n 的总分。如果一个组的被试和项目作答反应存在交互，α_g 和 β_g 至少有一个不等于 0，则存在 DIF。GLR 的优点是即使对照组不明确时，也可以使用。根据研究，被试群体规模越大，模型参数的估计越好，因此 DIF 识别的准确率也越好。LR 的统计检验力不受组间能力差异的影响[29]，Rogers 和 Swaminathan 指出，在高难度和高区分度的项目中，LR 一类错误率过高，可能是由于猜测参数对高难度项目的影响较大[32]。Svetina 和 Rutkowski 研究显示，组数对 GLR 的表现没有影响[33]；在没有 DIF 的情况下，有膨胀的一类错误率，而在有 DIF 的情况下非常保守，在 DIF 值很大时统计检验力一般都很高。Lee 提出 LR 是渐进抽样分布并不适用于小样本条件，其提出基于渐进 LRT 分布的 LR 统计量可以适用于小样本的情况[34]。

3.4 BD 方法

Breslow–Day（BD）方法是通过确定项目作答反应与被试之间的联系是否是

https://doi.org/10.35534/pc.0108034

一致的[8]，如果一致则存在一致性 DIF；如果不是，则存在非一致性 DIF[35]。该统计量表示为：

$$BD = \sum_{m} \frac{[A_m - E(A_m)]^2}{Var(A_m)} \quad (16)$$

其中，A_m 为对照组中正确作答所研究项目得分 m 的被试人数，E 和 Var 分别表示该值的期望值和方差。

其中，$E(A_m)$ 是二次方程的正根，等于以下两个根中的正值：

$$E(A_m) = \frac{\hat{\alpha}\,(N_{mR}+N_{m1})+(N_{mF}-N_{m1}) \pm \sqrt{\rho}}{2(\hat{\alpha}-1)} \quad (17)$$

其中，$\hat{\alpha}$ 是优势比的估计值，ρ 的计算如下：

$$\rho = [\hat{\alpha}\,(N_{mR}+N_{m1})+(N_{mF}-N_{m1})]^2 - 4\hat{\alpha}\,(\hat{\alpha}-1)\,N_{mR}\,N_{m1} \quad (18)$$

$Var(A_m)$ 计算如下：

$$Var(A_m) = \left[\frac{1}{E(A_m)} + \frac{1}{N_{mR}-E(A_m)} + \frac{1}{N_{m1}-E(A_m)} + \frac{1}{N_{mF}-N_{m1}-E(A_m)} \right]^{-1} \quad (19)$$

BD 统计量也是渐进卡方分布。Hosmer 和 Lemeshow 在多层的列联表分析中提到了同质性优势比的 BD 方法[36]。Camilli 和 Shepard 指出 BD 方法可以用来检测非一致性 DIF[37]。Aguerri，Galibert，Attorresi 和 Maraño 使用 BD 与 LR 和 MH 进行了比较，发现 BD 更适合于检测非一致性 DIF，在短测验中 BD 在一类错误率表现优于 LR 和 MH，说明 BD 更适合于非一致性 DIF 的检测以及短测验条件[38]。

4 参数检测方法

4.1 似然比检验（Likelihood Ratio Test，LRT）

LRT 是一种基于参数和模型的程序，既可以检测一致性 DIF，亦可以检测非一致性 DIF。在使用 LRT 程序进行 DIF 检测时，假设对照组和目标组之间没有项目参数的差异[9][39][40]。它由两种模型组成：缩减模型和扩展模型。在缩减模型中，所研究项目的项目参数在两组中被约束为相等；在扩展模型中，

所研究项目的项目参数不受约束。

LRT 统计量 G^2 的计算公式为：

$$G^2=-2log\left(\frac{L_C}{L_A}\right) \quad (20)$$

其中，L_C 为拓展模型的极大似然估计值，L_A 为缩减模型的极大似然估计值，又可转换为：

$$G^2=-2LL_C-(-2LL_A)=-2LL_C+2LL_A \quad (21)$$

其中，LL_C 为给定的缩减模型参数的最大似然估计的对数似然；LL_A 为给定的扩展模型的最大似然估计的对数似然。G^2 统计量是卡方分布，自由度等于两个模型中项目的参数个数差值。如果统计量值显著大于临界值，则存在 DIF 值。

不同的 DIF 项目比例和模型类型是影响 LRT 的有效因素，DIF 项目的比例增加会导致一类错误率的增加；当目标组和对照组的样本大小改变时，LRT、SIBTEST 和 MH 的一类错误率减小，统计检验力增加。在 Kim 和 Cohen 的研究中，研究者使用 LRT 程序检测 DIF 在不同的样本大小和能力匹配情况下的有良好的一类错误率控制［41］。样本量的大小会影响 LRT 的统计检验力表现，当两组平均能力分布不同时，LRT 依然有良好的一类错误率控制［42］。

4.2 Lord 卡方法

Lord 卡方检验可以对各类项目反应模型进行拟合，但在拟合时需要将不同组中得到的项目参数转化到同一标准上。该方法使用的 Q_j 统计量有如下形式：

$$Q_j=(v_{jR}-v_{jF})'(\Sigma_{jR}-\Sigma_{jF})^{-1}(v_{jR}-v_{jF}) \quad (22)$$

其中，$v_{jR}=(a_{jR}, b_{jR}, c_{jR})$ 和 $v_{jF}=(a_{jF}, b_{jF}, c_{jF})$ 分别是对照组和目标组的项目区分度、项目难度和猜测系数的向量，Σ_{jR} 和 Σ_{jF} 是对照组和目标组作答反应的方差—协方差矩阵。Q_j 统计量的虚无假设是两组被试的项目参数相等，该统计量是渐进卡方分布，自由度等于模型中估计参数的个数。

Kim，Cohen 和 Park 提出了广义 Lord 卡方测试，将 Lord 卡方测试扩展到多个目标组的情况下［15］。广义的 Lord 卡方统计量根据式（22）改进：

$$Q_j=(Cv_j)'(C\Sigma_j C')^{-1}(Cv_j) \quad (23)$$

https://doi.org/10.35534/pc.0108034

其中 $v_j = (v_{jR}, v_{jF})'$ 是对照组和目标组中项目参数估计值的向量，Σ_j 为反应的组块对角矩阵，其中每个对角组块为每组被试项目参数的方差—协方差矩阵。C 矩阵是一个设计矩阵，用于项目参数在组之间进行比较［15］。

广义 Lord 统计量也是渐近卡方分布，自由度与设计矩阵 C 的等级有关。只有样本量大于 1000 时，Lord 卡方方法对项目参数估计才能更加稳定［43］。Woods，Cai 和 Wang 评估了 Lord 卡方 Wald 检验的改进版，在样本量和 DIF 百分比不同的情况下，将 Wald–1、Wald–2 算法和 LRT 方法同时进行两组和三组的评估［44］。结果表明，两组情况下 Wald–1 的表现较好，Wald–2 的一类错误率很大，在多组情况下 LRT 和 Wald–1 方法的表现都好于 Wald–2。

4.3 Raju 面积测量法

Raju 面积测量法［10］［11］将面积测度分为两种“无符号测度”和“有符号测度”，通过计算目标组和对照组的项目特征曲线间的面积大小来进行 DIF 检测。其公式可表达如下：

$$A_i = \sum_{j=1}^{k} |P_{i1}(\theta_j) - P_{i2}(\theta_j)| \Delta\theta_j \qquad (24)$$

若两曲线间的面积越大，说明项目存在 DIF 的可能性就越大；若两曲线间的面积接近于 0，则项目无 DIF［16］。Raju 曾推导出用于 1PL、2PL 和 3PL 模型中的面积计算公式［10］，当使用 3PL 模型时，其公式如下：

$$A_i = (1-c)\left| \frac{2(a_2-a_1)}{Da_1 a_2} \ln[1+e^{Da_1 a_2 (b_2-b_1)/(a_2-a_1)}] - (b_2-b_1) \right| \qquad (25)$$

其中，D 是一个比例常数，通常设置为 1.7，a 为项目区分度，b 是项目难度和 c 是猜测系数。使用 2PL 模型时，没有猜测参数；若是使用 1PL 模型时，上述公式简化为 $|b_2-b_1|$［45］。Raju 推导出式（25）的标准误，式（25）所得面积除以标准误，可以转化为 Z 分数，用以检测其显著性［11］。但是，该方法假定目标组和对照组的项目特征曲线的猜测参数相等，否则 A_i 的值是无穷大的，无法检测是否显著。

Cohen 和 Kim 的研究采用不同的测试长度、样本量、DIF 项目比例以及使用

2PL 模型进行项目参数估计，结果发现与 Raju 面积方法相比，使用 Lord 卡方方法的一类错误率更低，随着测试长度和 DIF 项目比例的增加，一类错误率也增加，并且随着样本量的增加或 DIF 统计量的显著性水平从 0.01 变为 0.05，一类错误率降低［46］。

5 总结

本文对用于 DIF 检验的参数和非参数方法进行了归纳梳理。综合而言，参数化 DIF 检测方法因其具有参数不变性的特点，并且由于 IRT 模型种类众多，可以适应不同情境下的 DIF 检测；参数化方法在检测一致性和非一致性 DIF 方面也具有较为明显的优势。但是，相较于非参数化方法，参数化方法计算过程更为繁琐，耗时较长，不同的 IRT 模型对于不同检测方法的统计检验力有很大的影响，同时，有一部分方法对于样本容量也有一定的要求。一般而言，需要大约 1000 人才可能得到较为可信的结果［47］，Clauser 和 Mazor 指出，对于 MH、SIBTEST 和 LR 等传统而有效的检测方法，每组被试量大约在 200 到 250 人之间是合适的［48］，而在实际测验中存在小样本的情况，样本量可能满足不了模型的要求［21］。简言之，无论是参数化方法还是非参数化方法，各种 DIF 检测方法都有各自独特的优势，进而存在各自相对优势的使用条件与范围。因此，研究者在实际中对 DIF 进行检测时，可以结合各种方法的优点和适用条件，且在必要时同时选用多种方法，以获得更为准确的检测结果。

DIF 检测方法经过近些年的发展，已有较为完善的框架，但是适用于多组情况和多级计分的 DIF 检测方法还不是十分完备，需要研究者进一步关注。在今后的研究中，除了进一步研究和完善适用于多级计分、题组和多组情况下的 DIF 检测方法外，还可以根据不同情境和不同测验类型等实际情况对 IRT 模型进行拓展。在通过模拟研究探究 DIF 产生的机制、检测 DIF 的方法之外，还需要进一步结合真实测验探究 DIF 的检测与处理方法。例如，陈冠宇和陈平将基于广义线性混合模型和非线性混合模型构建的 IRT 模型，定义为解释性项目反应理论模型，在 IRT 模型的基础上加入预测变量，在刻画被试和项目间关系的基础上

https://doi.org/10.35534/pc.0108034

进一步解释相关变量影响，进而拓展 IRT 模型的应用范围［49］。

参考文献

［1］Penfield R D，Camilli G. Differential item functioning and item bias［J］. Handbook of Statistics，2007，26（26）：125-167. https://doi.org/10.1016/S0169-7161(06)26005-X

［2］Swaminathan H，Rogers H J. Detecting Differential Item Functioning Using Logistic Regression Procedures［J］. Journal of Educational Measurement，1990，27（4）：361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

［3］Mantel N，Haenszel W. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease［J］. Journal of the National Cancer Institute，1959，22：719-748.

［4］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，1986，23：355-368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

［5］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（2）：159-194. https://doi.org/10.1007/BF02294572

［6］Li H H，Stout W. A new procedure for detection of crossing DIF［J］. Psychometrika，1996，61（4）：647-677. https://doi.org/10.1007/BF02294041

［7］Chalmers R P. Improving the Crossing-SIBTEST Statistic for Detecting Non-uniform DIF［J］. Psychometrika，2018，83（2）：376-386. https://doi.org/10.1007/s11336-017-9583-8

［8］Breslow N E，Day N E. Statistical methods in cancer research. Volume I – The

analysis of case-control studies[J]. IARC scientific publications, 1980, 32(32): 5-338.

[9] Thissen D, Steinberg L, Wainer H. Use of item response theory in the study of group differences in trace lines [M]. In H. Wainer & H. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum, 1988, 147-172.

[10] Raju N S. The area between two item characteristic curves [J]. Psychometrika, 1988, 53: 495-502. https://doi.org/10.1007/BF02294403

[11] Raju N S. Determining the significance of estimated signed and unsigned areas between two item response functions [J]. Applied Psychological Measurement, 1990, 14: 197-207. https://doi.org/10.1177/014662169001400208

[12] Lord F M. Applications of Item Response Theory to Practical Testing Problems [M]. Hillsdale, NJ: Erlbaum, 1980.

[13] Somes G W. The generalized Mantel-Haenszel statistic [J]. The American Statistician, 1986, 40: 106-108. https://doi.org/10.1080/00031305.1986.10475369

[14] Magis D, Raiche G, Béland S, et al. A Generalized Logistic Regression Procedure to Detect Differential Item Functioning Among Multiple Groups [J]. International Journal of Testing, 2011, 11(4): 365-386. https://doi.org/10.1080/15305058.2011.602810

[15] Kim S H, Cohen A S, Park T H. Detection of Differential Item Functioning in Multiple Groups [J]. Journal of Educational Measurement, 1995, 32(3): 261-276. https://doi.org/10.1111/j.1745-3984.1995.tb00466.x

[16] 漆书青. 现代测量理论在考试中的应用 [M]. 武汉：华中师范大学出版社，2003.

[17] Swaminathan H, Hambleton R K, Rogers H J. 21 Assessing the Fit of Item Response Theory Models [J]. Handbook of Statistics, 2006, 26(6): 683-718. https://doi.org/10.1016/S0169-7161(06)26021-8

[18] Lord F M，Novick M R. Statistical Theories of Mental Test Scores [J]. Journal of the American Statistical Association，1968，66（335）：651. https://doi.org/10.2307/2283550

[19] Hanson B A. Uniform DIF and DIF Defined by Differences in Item Response Functions [J]. Journal of Educational & Behavioral Statistics，1998，23（3）：244–253. https://doi.org/10.3102/10769986023003244

[20] Holland P W，Thayer D T. Differential Item Functioning and the Mantel–Haenszel Procedure [J]. ETS Research Report Series，1986，2：1–24. https://doi.org/10.1002/j.2330–8516.1986.tb00186.x

[21] Clauser B E，Mazor K M，Hambleton R K. The effects of score group width on the Mantel–Haenszel procedure [J]. Journal of Educational Measurement，1994，31（27）：67–78. https://doi.org/10.1111/j.1745–3984.1994.tb00435.x

[22] Fidalgo A M，Ferreres D，Muñiz J. Utility of the Mantel–Haenszel Procedure for Detecting Differential Item Functioning in Small Samples [J]. Educational & Psychological Measurement，2004，64（6）：925–936. https://doi.org/10.1177/0013164404267288

[23] Penfield R D. Assessing Differential Item Functioning Among Multiple Groups：A Comparison of Three Mantel–Haenszel Procedures [J]. Applied Measurement in Education，2001，14（3）：235–259. https://doi.org/10.1207/S15324818AME1403_3

[24] Guilera G，Gómez–Benito J，Hidalgo M D，et al. Type I error and statistical power of the Mantel–Haenszel procedure for detecting DIF：a meta–analysis [J]. Psychological Methods，2013，18（4）：553–571. https://doi.org/10.1037/a0034306

[25] Millsap R E，Everson H T. Methodology review：Statistical approaches for assessing measurement bias [J]. Applied Psychological Measurement，1993，17：297–334. https://doi.org/10.1177/014662169301700401

[26] Dorans N J，Holland P W. DIF Detection and Description Mantel–Haenszel

and Standization [M]. In P. W. Holland & H. Wainer (Eds), differential item functioning. Hillsdale, NJ: Erlbaum, 1993: 35–66.

[27] Roussos L, Stout W. A Multidimensionality–Based DIF Analysis Paradigm [J]. Applied Psychological Measurement, 1996, 20(4): 355–371. https://doi.org/10.1177/014662169602000404

[28] Atalay K K, Arsan N, Gök B, et al. Comparing Performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel–Haenszel Methods in the Determination of Differential Item Functioning [J]. Kuram Ve Uygulamada Egitim Bilimleri, 2014, 14(6): 2186–2193.

[29] Finch W H, French B F. Detection of Crossing Differential Item Functioning A Comparison of Four Methods [J]. Educational & Psychological Measurement, 2007, 67(4): 565–582. https://doi.org/10.1177/0013164406296975

[30] Noortgate W V D, Boeck P D. Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models [J]. Journal of Educational & Behavioral Statistics, 2005, 30(4): 443–464. https://doi.org/10.3102/10769986030004443

[31] Kanjee A. Using logistic regression to detect bias when multiple groups are tested [J]. South African Journal of Psychology, 2007, 37: 47–61. https://doi.org/10.1177/008124630703700104

[32] Rogers H J, Swaminathan H. A Comparison of Logistic Regression and Mantel–Haenszel Procedures for Detecting Differential Item Functioning [J]. Applied Psychological Measurement, 1993, 17(2): 105–116. https://doi.org/10.1177/014662169301700201

[33] Svetina D, Rutkowski L. Detecting differential item functioning using generalized logistic regression in the context of large–scale assessments [J]. Large–scale Assessments in Education, 2014, 2(1): 1–17. https://doi.org/10.1186/s40536–014–0004–5

https://doi.org/10.35534/pc.0108034

[34] Lee S. Detecting Differential Item Functioning Using the Logistic Regression Procedure in Small Samples [J]. Applied Psychological Measurement, 2017, 41 (1): 1-14. https://doi.org/10.1177/0146621616668015

[35] Penfield R D. Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF [J]. Alberta Journal of Educational Research, 2003, 49 (3): 231-243.

[36] Hosmer D W, Lemeshow J S. Best Subsets Logistic Regression [J]. Biometrics, 1989, 45 (4): 1265-1270. https://doi.org/10.2307/2531779

[37] Camilli G, Shepard L A. MMSS: Methods for Identifying Biased Test Items [J]. Bms Bulletin of Sociological Methodology, 1994, 4 (45): 145-146.

[38] María E A, María S G, Horacio F A, et al. Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test [J]. Quality & Quantity, 2009, 43 (1): 35-44. https://doi.org/10.1007/s11135-007-9130-2

[39] Thissen D, Steinberg L, Gerrard M. Beyond group-mean differences: The concept of item bias [J]. Psychological Bulletin, 1986, 99 (1): 118-128. https://doi.org/10.1037//0033-2909.99.1.118

[40] Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models [M]. Hillsdale, NJ: Erlbaum, 1993.

[41] Kim S H, Cohen 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. https://doi.org/10.14662169802200403

[42] Ankenmann R D, Witt E A, Dunbar S B. An Investigation of the Power of the Likelihood Ratio Goodness of Fit Statistic in Detecting Differential Item Functioning [J]. Journal of Educational Measurement, 2010, 36 (4):

277–300. https://doi.org/10.1111/j.1745–3984.1999.tb00558.x

[43] Hambleton R K, Swaminathan H, Rogers H J. Fundamentals of Item Response Theory [J]. Bms Bulletin of Sociological Methodology, 1992, 36: 83–83.

[44] Woods C M, Cai L, Wang M. The Langer–Improved Wald Test for DIF Testing with Multiple Groups: Evaluation and Comparison to Two–Group IRT [J]. Educational and Psychological Measurement, 2013, 73 (3): 532–547. https://doi.org/10.1177/0013164412464875

[45] 余民宁. 试题反应理论（IRT）及其应用 [M]. 新北：心理出版社, 2009.

[46] Cohen A S, Kim S H. A Comparison of Lord's Chi Square and Raju's Area Measures in Detection of DIF [J]. Applied Psychological Measurement, 1993, 17 (17): 39–52. https://doi.org/10.1177/014662169301700109

[47] 曾秀芹, 孟庆茂. 项目功能差异及其检测方法 [J]. 心理科学进展, 1999, 17 (2): 41–475.

[48] 李凌艳, 张勋. DIF 分析实际应用中的常见问题及其研究新进展 [J]. 考试研究, 2010, 2: 75–84.

[49]陈冠宇, 陈平. 解释性项目反应理论模型: 理论与应用[J]. 心理科学进展, 2019, 27 (5): 937–950. http://kns.cnki.net/kcms/detail/11.4766.R.20190320.1003.026.html.