文章编号:1000-5862(2014)06-0593-07

基于多级计分题目的分步功能差异检验

李美娟[1],刘红云[2*]
(1. 北京教育科学研究院,北京　100191;2. 北京师范大学心理学院,北京　100875)

摘要:对分步功能差异如何在项目功能差异的检测和解释中发挥作用进行阐述:(i)从国外分步功能差异的模型、方法原理、分类模式、应用和结果解释等方面对这一方法的进展和应用情况进行概括和综述,旨在对国内测验公平性的研究提供借鉴;(ii)通过实际测验的数据,采用 DSF 分析的方法对测验中题目及不同等级分数的 DIF 进行了检验,进而对产生 DIF 的原因进行更深入的分析,以对测验内容的审核和题目的修订提供更具体和具操作性的依据.

关键词:多级计分题目;项目功能差异;分步功能差异
中图分类号:B 841.7;TP 301.6　　**文献标志码**:A

0　引言

从 20 世纪 60 年代起,美国教育界就开始关注性别与种族在测验结果上的差异,即测验公平的问题. 测验的公平性是测验研发者、使用者,乃至整个社会所普遍关注的一个非常重要而又异常复杂的问题. 对于中国这个考试大国来讲,为了提高试题质量,对于一些高考、公务员等考试进行题目的公平性检验是十分必要的. 美国的教育研究学会(AERA)、心理学学会(APA)、教育测量年会(NCME)认为测验的公平必须满足 4 个条件:(i)项目没有偏差;(ii)所有的考生都有平等的机会证实自己对于测验内容掌握的熟练性程度;(iii)所有的考生都有平等的机会学习测验内容(除了就业、认证或者入学考试);(iv)不同类别考生的分数分布是相同的[1]. 中国教育学会教育测量与统计分会认为测验公平性是指如果一个测验对来自不同团体而具有相同能力或熟练程度的个体所测得的特性相同,则说明该测验具有公平性,如果测得的特性不同,则说明该测验不公平而具有偏差[2]. 即公平性检查的目的是找出是否存在测验范围之外引起组间差异的因素.

项目偏差这个概念是美国在 20 世纪 60 年代提出的,主要用于对跨文化团体、性别、种族差异的研究. 一直以来,对于项目偏差的研究,项目功能差异(DIF,differential item functioning)一直发挥着非常重要的作用,DIF 是项目偏差的充分而非必要条件. 相对于项目偏差,DIF 是一个有关统计分析的术语,表示不同团体相同能力水平的被试对于相同测验题目的通过率却不同,引起 DIF 的原因是 2 组被试在与测验所测的能力无关的知识或经验上存在差异[3-5]. 目前大多数检测 DIF 的方法都集中在 2 级计分题目上,其中包括(i)非参数方法:MH,SIBTEST,LRDIF,STND 等;(ii)参数方法:基于 IRT 的 Lord 卡方检验法,Raju 面积测量法和似然比率法(IRT 和 LRDIF);而对于多级计分题目 DIF 的检测方法多来源于 2 级计分题目检测方法的拓展,目前也有许多关于多级计分题目 DIF 检测的方法,其中包括标准化均值差异法、Mantel 的卡方检验、广义 Mantel-Haenszel 法、多级 SIBTEST 法、逻辑斯蒂克判别函数分析法、累积发生比方法等. 但是,这些传统的多级计分题目检测 DIF 的方法只能提供项目水平上的 DIF 指标,不能测量题目在哪个分数水平上存在 DIF,进而也不能进一步解释 DIF 的产生原因.

纵观国外对于 DIF 的研究,大多数研究者集中在其方法的探讨上,有少数研究涉及到 DIF 检测的影响因素,如样本量、维度,以及模型的参数等方面. 而国内对 DIF 的研究也比较早,主要是对 DIF 相关概念以及检测方法的研究. 之后也有不少研究者使用实际数据对 DIF 检测方法进行应用,并对几种方

收稿日期:2014-06-19
基金项目:国家自然科学基金(31100759),全国教育科学"十二五"规划教育部重点课题(GFA111001),北京市与中央在高校共建课题(019-105812)和北京市教育科学"十二五"规划 2012 年度青年专项课题(CHA12109)资助项目.
通信作者:刘红云(1972-),女,山西夏县人,教授,博士,主要从事心理统计、心理测量与评价的研究.

法进行比较,还有一些研究者将 DIF 的检测直接应用到心理测验中,对心理测验的公平性进行初步探讨. 但是很少有研究对 DIF 的解释进行深入分析,或者对产生 DIF 的原因进行挖掘,从而使测量在心理学的实际应用中变得更有意义. 近年来,对多级计分项目的 DIF 的研究有进一步细化和深入的趋势,本研究的目的在于回顾 DIF 研究方法这一领域的新进展及应用,介绍一种新的检测 DIF 的方法——分步功能差异(DSF)检验法,同时结合一个实际测验,简要介绍这一方法的具体应用. 本研究的目的在于为研究者进一步探讨产生 DIF 的原因提供更充分的依据和途径.

1 分步功能差异(DSF)的相关概念

1.1 分步函数的定义

分步功能差异(DSF)可以检测多级计分题目的不同分数水平上是否存在 DIF,即通过分步函数的特征(基本参数)得到特定能力的被试在各个分数水平上正确做答的概率[6]. 其分步函数根据 IRT 模型的不同具有不同的形式. 基于不同形式的分步功能特征的含义是不同的,最常用的是等级反应模型(GRM)下的累积形式和分部计分模型(PCM)下的连接形式的分步功能差异.

分步函数主要是在多级计分题目上,个体从低分数水平跨越到高分数水平的概率,对于一个有 r 个分数水平的多级计分题目,则有 $J=r-1$ 个分步水平. 例如,一个 4 级计分题目,分数水平定为 0,1,2,3,$r=4$,分步水平 $J=3$,结果用符号 Y 表示. 其累积形式的分步函数是:(i)被试从分数水平 0 到分数水平 1 或者高于 1 的概率,即 $Y\geqslant 1$ 概率;(ii)被试从分数水平 1 到分数水平 2 或者高于 2 的概率,即 $Y\geqslant 2$ 概率;(iii)被试从分数水平 2 到分数水平为 3 的概率,即 $Y=3$ 概率. 而其连接形式的分步函数是:(i)被试从分数水平 0 到分数水平 1 的概率,即 $Y=1$ 概率.(ii)被试从分数水平 1 到分数水平 2 的概率,即 $Y=2$ 概率.(iii)被试从分数水平 2 到分数水平 3 的概率,即 $Y=3$ 概率.

1.2 分步函数的参数

每个分步水平均使用 2 参数 Logistic 回归模型进行参数估计[7]:

$$P(Y\geqslant j\mid\theta)=\frac{\exp[D\alpha(\theta-b_j)]+G\omega_j}{1+\exp[D\alpha(\theta-b_j)]+G\omega_j},$$

其中 b_j 为 j 分步水平的难度系数,且每个分步水平的难度系数是不同的;a 为分步水平的区分度系数,且每个分步水平的区分度系数是相同的;θ 为被试的能力水平;D 为 1.7. $G=0$ 为目标组,$G=1$ 为参照组. a 描述了每个分步水平能够区分高低能力被试的程度,b_j 描述了通过该分步水平的概率为0.5的特定被试的能力水平. 在 GRM 模型中,假设 b_j 随着分步水平的提高而增加,而在 PCM 模型中,则没有这样的假设. $\omega_j=0$ 表示不存在 DSF,$\omega_j>0$ 表示参照组占优势,$\omega_j<0$ 表示目标组占优势.

1.3 一致性 DSF 和非一致性 DSF 的概念

在 DSF 的分析中,一致性 DSF 和非一致性 DSF 是基于 j 个分步水平的 DSF 分析. 一致性 DSF 指 j 个分步水平的 DSF 均相同,而非一致性 DSF 是指 j 个分步水平的 DSF 不完全相同[8]. 由此可见,虽然 DSF 和 2 级计分题目 DIF 的研究较相似,但是对于非一致性 DIF 和 DSF,组间 a 参数差异的不同是区分两者最重要的因素. 在 2 级计分题目中,a 参数的不同表示非一致性 DIF 的存在,而非一致性 DSF 表示在不同的分步水平上 2 组 DIF 方向不一致或 DIF 大小程度不一致,如 DSF 分析结果在第 1 个分步水平上有利于男生组,在第 2 个分步水平上有利于女生组,以上属于非一致性 DSF 的一种情况.

非一致性 DSF 的检测方法与 2 级计分的非一致性 DIF 检验方法是相同的,但是相关研究文献中还没有真正应用过,所以应用的价值还有待进一步证实.

2 DSF 的估计

已有研究关于 DSF 的估计方法主要有参数和非参数 2 类方法,其中参数法主要有 IRT 方法,而非参数法主要有 odds 比率法和 Logistic 回归法. 这些方法曾是检测 2 级计分 DIF 的方法,所以在应用时要注意:(i)必须将所研究题目的等级水平转化为 j 个分步水平,(ii)必须对每个分步水平独立分析.

2.1 分步水平的建构方法

虽然从理论上讲构建分步水平的方法有多种,但主要的是以广义分部计分模型[9](GPCM)为基础的连接方法(AC-LOR)和以等级计分模型[10](GRM)为基础的累积方法(CU-LOR),这 2 种方法对 DSF 的定义如前所述,但是 2 种概念下 DSF 的结果和解释是否相同也是 DIF 研究者们需要深入考察的一个内容. 对以这 2 种模型为基础的 DSF 发生比方法进行了统计特征的模拟研究比较,结果发现累

积方法下的 DSF 结果更稳定[7],精确性更高. 另外,将 2 种方法应用于实际数据时[11],当不存在 DSF 或者 DSF 很小时,两者结果一致. 但是第 1 种方法缺乏独立性,一个水平存在较大的 DSF,将伴随着高水平反方向的较大 DSF. 当存在较大的 DSF 时,使用第 2 种方法更容易获得显著的结果,而且这种方法标准误更小,稳定性和检验力更强. 研究还发现,当仅有一个分数水平上存在 DSF 时,第 1 种方法的精确性更强,解释更加合理.

2.2　参数估计方法

IRT 检测 DSF 的基础是比较参照组和目标组在多级计分题目的每个分步水平上的差异[8],表示为 $\Delta(b_j) = b_{jF} - b_{jR}$. 如果 $\Delta(b_j) = 0$,则不存在 DSF. 若 $\Delta(b_j) > 0$,则表示参照组占优势. 反之,目标组占优势. $\Delta(b_j)$ 为 j 分步水平上参照组和目标组的有符号面积测度[12],这与 Raju 对 2 级计分题目 DIF 的面积测量法是相同的. 因此,DSF 的效应大小的衡量标准与 Raju 的面积测量法的衡量标准是相同的.

常用的检验标准是: 若 $|\Delta(b_j)| < 0.25$,则表示较小的 DSF 值. $|\Delta(b_j)| < 0.50$,则表示中等的 DSF 值. 如果 $|\Delta(b_j)| > 0.50$,则表示较大的 DSF 值. 检验 IRT 模型下不存在 DSF 的虚无假设的方法有 2 种,其中一种是将 $\Delta(b_j)$ 除以标准误,并且假设其是标准正态分布的. 另外一种方法是似然比检验法,即将紧缩模型(2 组项目参数固定) 和扩展模型(自由估计 2 组分步函数参数) 的似然值进行比较.

2.3　非参数方法

与检验 DSF 的参数方法比较,在实际应用中非参数方法更受欢迎,因为其不受样本量、数据与模型拟合程度的影响,而且易操作.

2.3.1　发生比方法(odds ratio)　发生比方法(odds ratio) 主要是比较参照组和目标组成功通过 j 分步水平的发生比,该发生比的自然对数就是 λ 值,即不同能力水平被试的 λ 值[13]. λ 的算法为

$$\mathring{\lambda}_j = \ln\left[\sum_{k=1}^{m}\frac{A_{jk}D_{jk}}{T_k}\Big/\sum_{k=1}^{m}\frac{B_{jk}C_{jk}}{T_k}\right],$$

其中 A_{jk} 为 k 能力水平的参照组成功通过 j 分步水平的人数; B_{jk} 为 k 能力水平的参照组未成功通过 j 分步水平的人数; C_{jk} 为 k 能力水平的目标组成功通过 j 分步水平的人数; D_{jk} 为 k 能力水平的参照组未成功通过 j 分步水平的人数; 若 $\lambda_j = 0$,则表示在 j 分步水平上不存在 DSF; 若 $\lambda_j > 0$,则表示在 j 分步水平上,题目得分会有利于参照组; 若 $\lambda_j < 0$,则表示 j 分步水平上,题目得分会有利于目标组.

发生比(odds ratio) 方法可以检验 DSF 的显著性,检验方法为

$$z(\hat{\lambda}_j) = \hat{\lambda}_j / SE(\hat{\lambda}_j),$$

其中 $SE(\hat{\lambda}_j)$ 的算法如下:

$$SE(\hat{\lambda}_j) = \Big(\sum_{k=1}^{m} T_K^{-2}\big(A_{jk}D_{jk} + \mathring{\alpha}_j B_{jk}C_{jk}\big)\big(A_{jk} + D_{jk} + \mathring{\alpha}_j B_{jk} + \mathring{\alpha}_j C_{jk}\big)\Big) \Big/ \Big(2\Big(\sum_{K=1}^{m} A_{jk}D_{jk}/T_k\Big)^2\Big)\Big)^{1/2}.$$

另外,上述方程所检验的统计量服从标准正态分布的[14].

ETS 常用的判断标准为: 当 $|\lambda_j| < 0.43$ 时,则表示存在较小的 DSF 值; 当 $0.43 \leqslant |\lambda_j| \leqslant 0.63$ 时,则表示存在中等的 DSF 值; 当 $|\lambda_j| > 0.63$ 时,则表示较大的 DSF 值.

2.3.2　Logistic 回归　估计 DSF 的另一种非参数方法是 Logistic 回归[8],模型表述为

$$P(Y \geqslant j \mid X) = \frac{\exp(\beta_{j0} + \beta_{j1}X + \beta_{j2}G)}{1 + \exp(\beta_{j0} + \beta_{j1}X + \beta_{j2}G)},$$

其中 Y 为被试在某个项目上第 j 步的得分,X 为测验总分. G 为一个关于组别变量的虚无变量,并且是 $G = 0$ 代表目标组,$G = 1$ 代表参照组. β_{j2} 为 j 分步水平的 DSF 效应. 其中 $\beta_{j2} = 0$ 为 j 分步水平不存在 DSF,$\beta_{j2} > 0$ 则表示 j 分步水平上存在 DSF,题目得分有利于参照组,$\beta_{j2} < 0$ 则表示 j 分步水平上存在 DSF,题目得分有利于目标组. 这个方法也可以通过在模型中加入测验分数 X 和分组变量 G 的交互作用来考察是否存在非一致性 DSF.

显著性检验方法: 似然比方法,即将紧缩模型(无 $\beta_{j2}G$ 项) 和扩展模型(有 $\beta_{j2}G$ 项) 的似然值进行比较. 统计软件提供 β 的估计值,显著性水平以及模型的($-2 \times$ 似然值),以便进行适当的似然比检验. 该方法划定 DSF 范围的标准是 ΔR^2,若 $\Delta R^2 < 0.10$,则表示较小的 DSF 值,若 $0.10 \leqslant \Delta R^2 \leqslant 0.20$,则表示中等的 DSF 值. 若 $\Delta R^2 > 0.20$,则表示较大的 DSF 值[15].

2.4　3 种估计方法之间的区别和联系

IRT 参数估计要求样本量大,数据需与相关分步函数拟合,并且该方法比较耗时,建议使用 BILOGMG3、IRTLRDIF[16] 和 MULTILOG7. DIFAS 程序,均可计算 λj 和 $z(\lambda j)$[17]. 如果在观测分数与 IRT 模型拟合的情况下,并且将测验总分认为是能力水平的近似估计时,3 种估计方法的结果具有一定的关系,即 Logistic 回归(迭代法) 和 odds ratio(非迭代法) 方法估计的 β 值和 λ_j 是等值的[18],另外,这 2 个数值与 2 组难度系数的差异是成比例

的,其中比例系数就是区分度值[16].

3 使用 DSF 的结果检测 DIF

3.1 利用 DSF 效应模式识别 DIF 产生的原因

R. D. Penfield 等[19]根据 DSF 产生的轨迹将 DSF 分为普遍性 DSF 和非普遍性 DSF,普遍性 DSF 是指所有的分步水平都有 DSF 效应,说明导致 DIF 的因素在题目水平上造成影响. 而非普遍性 DSF 是指一些分步水平上存在 DSF,说明导致 DIF 的因素仅仅在一个或者少数几个分步水平上造成影响. 根据 DSF 产生的一致性将分为一致性 DSF、会聚性 DSF、发散性 DSF 3 种. 一致性 DSF 是指分步水平 DSF 值的大小和符号都相同,会聚性 DSF 是指分步水平的 DSF 值符号相同,大小却不同,发散性 DSF 是指分步水平的 DSF 值符号不同,详见表 1.

表 1 DSF 效应模式

一致性(DIF 产生的一致性)	普遍性(DIF 产生的轨迹)	
	普遍性	非普遍性
一致	所有分步水平都有 DSF 效应大小相同,符号相同	只有一步或少步有 DSF 效应大小相同,符号相同
会聚	所有分步水平都有 DSF 效应大小不同,符号相同	只有一步或少步有 DSF 效应大小不同,符号相同
发散	所有分布水平都有 DSF 效应符号不同	只有一步或少步有 DSF 效应符号不同

一致普遍性 DSF 对 DIF 的产生源于题目水平的特征提供了充足的证据,而一致非普遍性 DSF 说明 DIF 的产生不一定源于题目水平的特征,而是源于存在 DSF 效应的分步水平的特征. 会聚性 DSF 说明 DIF 可能源于题目水平的特征,也可能源于不同分数水平的不同特征. 会聚性 DSF 的解释很有挑战性,尤其在分步水平较多的情况下. 发散性 DSF 给 DIF 源于不同分步水平的特征提供了充足的证据,而且不同的分步特征使得有利的组别不同. 所以 DIF 研究者的任务就是检测定义分步水平的分数等级的特征,从而识别是一个特征对不同分步水平有影响还是多个特征分别对不同分步水平有影响.

3.2 基于 DSF 结果检验项目 DIF

每个分步水平不存在 DSF 是题目不存在 DIF 的充分必要条件. 这种方法也就是 R. D. Penfield 提出的 DIF 同时性分步水平检测方法(SSL),其源于发生比的 DSF 估计法[6]. SSL 基于分步水平,并且在 DSF 的符号和大小随着分步水平的变化而变化时,具有比其它 DIF 方法更强的检验力.

上述方法也就是 DIF 的 global 检验方法的一种,DIF 的 global 检验则关注无符号 DSF,因此它对发散性 DSF 是敏感的. 当分步水平的 DSF 符号不同、大小不同时,global 检验法对 DIF 的检测是比较敏感的,其中现有的 global 检验法包括,IRT 的似然比方法,多级逻辑斯蒂克回归方法,广义的 MH 卡方检验法,还有 SSL 法[7],但是在分步水平的 DSF 一致时,net 检验法的敏感性更强. DIF 的 net 检验基于所有分步水平有符号 DSF 的集合,它对发散性 DSF 是不敏感的. 其中 DIF 的 net 检验包括 Mantel 的卡方检验法、多级计分 SIBTEST 检测法、标准均值差异和其相关方法,以及 Liu-Aresti 的累积 common odds ratio 估计法. 因此,DIF 的 net 检验对发散性 DSF 是不敏感的,而 DIF 的 global 检验对发散性 DSF 是敏感的[21].

3.3 DSF 和 DIF 的联合使用

对于如何使 DIF 和 DSF 的检测最有效地发挥作用,最重要的是弄清楚两者在多级计分模型中评价测量不变性的优缺点. 在关注造成 DIF 的分数水平时,DIF 的检测并没有提供任何的信息. 相反,DSF 的检测却能提供给项目水平的 DIF 提供分数水平上的信息. 虽然 DIF 存在这样的缺点,但是有时 DIF 的检验力可能更强. 因为其分析综合了 j 个分步水平的结果. 总之,DIF 在非等同测量中可能更敏感,而 DSF 可以给非等同测量的形式提供更多的信息.

基于 DIF 和 DSF 检测的优缺点,建议在虚无假设为不存在 DIF 的多级计分题目中,测量等同的开始阶段则同时使用 DIF 的 net 检验和 global 检验. 前有研究发现:(i)当 DSF 效应不一致时(除了普遍性 DSF),global 检验法的检验力更强.(ii)当 DSF 效应一致时(普遍一致性 DSF),net 检验法的检验力更强. 如果结果接受虚无假设,则说明测量的等同性存在,如果拒绝虚无假设,则说明需要进一步的 DSF 分析[20].

因此,DSF 和 DIF 检验的联合可以提高敏感性,并且可以给题目提供更多的信息. DSF 的检测可以对 DIF 产生原因和轨迹提供更多的信息. 在实际应用中,建议同时进行 DIF 的 net 检验和 global 检验,如果两者中的一种检验结果显著,那么需要继续进行 DSF 的检测[19],所以建议同时使用 3 种方法对 DIF 进行检验,检验力会更强.

4　应用举例

以下是一个对 DSF 使用以及解释的实证研究. 本研究的研究材料是 Ralf Schwarzer 等编制的一般自我效能感量表[21],其中有 10 个题目,均为 4 点计分. 被试为美国人和香港人,其中美国被试 1 167 人,约占 48%,香港被试 1 152 人,约占 52%. 另外,在此研究中,美国为参照组,香港为目标组.

分别使用发生比方法,Logistic 回归法,IRT 方法对自我效能感量表的 10 个题目 DSF 分析. 结果如表 2 所示.

在表 2 中,使用 DIFAS 程序[17] 计算各分步水平上的 common log-ratio (λ_j),λ_j 值的标准误. 为了验证如何将 DSF 的分析与 DIF 结合在一起,每个题目也均进行了 global 和 net 检验,其中 DIF 的 global 检验对每个分步水平的 DSF 进行显著性水平为 Bonfereoni-adjusted Typed I error rate(0.05/J) 的显著性检验,而 DIF 的 net 检验使用 Liu-Agresti 累积 common Log-odds ration(LA) ,LA 值服从正态分布,可通过 Z 值对其进行显著性检验[22].

表 2　发生比的 DSF 数据分析

Item	step	λ_i	$SE(\lambda_i)$	$Z(\lambda_i)$	DSF SIZE	DSF FORM	DIF EFFECT
Item2	1	0.677	0.225 45	3.003*	大	普遍一致	LA = 0.873 $Z(LA)$ = 8.009*
	2	0.833	0.124 83	6.673*	大		
	3	1.173	0.213 58	5.492*	大		
Item3	1	−1.116	0.160 77	−6.942*	大	非普遍会聚	LA = −0.665 $Z(LA)$ = −6.065*
	2	−0.500	0.124 03	−4.031*	中等		
	3	−0.053	0.263 02	−0.0202	小		
Item4	1	−0.599	0.227 96	−2.628*	中等	非普遍会聚	LA = −0.496 $Z(LA)$ = −4.55*
	2	−0.415	0.132 98	−3.121*	小		
	3	−0.614	0.234 26	−2.621*	中等		
Item5	1	−0.686	0.219 43	−3.126*	大	非普遍会聚	LA = −0.430 $Z(LA)$ = −3.839*
	2	−0.442	0.135 86	−3.253*	中等		
	3	−0.070	0.237 58	−0.295	小		
Item7	1	1.268	0.233 24	5.436*	大	普遍一致	LA = 0.984 $Z(LA)$ = 8.708*
	2	0.956	0.137 04	6.976*	大		
	3	0.875	0.180 61	4.845*	大		
Item8	1	0.207	0.249 54	0.830	小	非普遍会聚	LA = 0.304 $Z(LA)$ = 2.951*
	2	0.269	0.123 48	2.178*	小		
	3	0.430	0.179 36	2.397*	中等		
Item9	1	0.644	0.339 59	1.896	大	普遍会聚	LA = 0.522 $Z(LA)$ = 4.619*
	2	0.489	0.139 56	3.504*	中等		
	3	0.547	0.176 65	3.097*	中等		
Item10	1	−1.852	0.283 52	−6.532*	大	普遍一致	LA = −1.614 $Z(LA)$ = −12.709*
	2	−1.649	0.142 25	−11.592*	大		
	3	−1.319	0.248 34	−5.311*	大		

在表 2 中第 1 列为题目,第 2 列为分步水平,第 3 列为 λ_i,第 4 列为 λ_i 的标准误,第 5 列为显著性水平为 Bonfereoni-adjusted typed I error rate(0.05/J) 的显著性检验,即 DSF 的 global 检验,第 6 列是根据判断标准判别的 DSF 效应大小,第 7 列为 DSF 模式,第 8 列为 DSF 的 net 检验.

表 3 中,使用 Logistic 回归法(SPSS) 和 IRT 方法(Multilog 软件) 对上述 10 个题目进行 DSF 分析,结果发现,Logistic 回归法和 IRT 方法计算的结果与发生比方法的计算结果基本相似,符合上文中的理论假设,另外,也说明该数据和 IRT 的分步函数是拟合的.

综上所述,本研究将使用发生比方法对研究结果进行解释,在本结果中,8 个题目的 global 检验结果显著,net 检验结果也显著. DSF 模式完全决定于 DSF 的大小,而不是 DSF 效应的显著性水平. 研究结果发现,2,7,10 题的 net DIF 检验显著,且 DSF 属于普遍一致型,由此可以说明造成 DIF 的原因在于题目本身; 2 题和 7 题的 λ 值为正,表明对于第 2 和 7 题讲,相同自我效能感的香港人和美国人,美国人在此题目上会得分更高,而 10 题相反,香港人得分会更高. 9 题属于普遍会聚型 DSF,说明造成 DIF 的

原因在于不仅在于题目本身,而且在于题目选项的设置,λ_j值越大,说明选项 j 的设置出现问题的程度越大,并且 λ 值为正,则说明在每个选项的设置上美国人得分都比较高,只是差异程度不同. 3,4,5,8 题的 DSF 属于非普遍会聚型,与前面一致的是,λ_j 值越大,说明选项 j 的设置出现问题的程度越大,并且具有中等或者较大程度 λ 值的选项 j 的设置标准比较容易出现问题. 总之,使用该问卷对美国人和香港人的自我效能感进行测量和比较是很不公平的.

表 3 Logistic 方法和 IRT 方法的 DSF 分析

Logistic 回归方法					IRT 方法			
Item	step	β_j	sig	ΔR^2	b_jF(SE)	b_jR(SE)	Δb_j(SE)	Z
Item1	1	-0.023	0.964	0	-4.40(0.23)	-4.30(0.44)	-0.10(0.35)	-0.29
	2	0.109	0.441	0.01	-1.50(0.07)	-1.57(0.10)	0.07(0.09)	0.81
	3	0.078	0.556	0.01	0.52(0.09)	0.50(0.07)	-0.02(0.08)	0.25
Item2	1	0.721	0.001	0.52	3.36(0.14)	-2.88(0.19)	6.24(0.17)	37.43*
	2	0.835	0	0.70	0.46(0.09)	0.10(0.08)	0.36(0.09)	4.23*
	3	1.260	0	1.59	1.64(0.15)	2.67(0.13)	-1.03(0.14)	-7.34*
Item3	1	-1.070	0	1.14	-1.34(0.08)	-2.19(0.13)	0.85(0.11)	7.89*
	2	-0.513	0	0.26	0.51(0.10)	0.06(0.07)	0.45(0.09)	5.21*
	3	0.080	0.756	0.01	2.34(0.22)	2.17(0.11)	0.17(0.17)	0.98
Item4	1	-0.548	0.016	0.30	-1.85(0.05)	-1.90(0.10)	0.05(0.08)	0.63
	2	-0.441	0.001	0.19	-0.40(0.05)	-0.46(0.05)	0.06(0.05)	1.20
	3	-0.499	0.023	0.25	0.94(0.10)	1.86(0.04)	-0.92(0.08)	-12.05*
Item5	1	-0.658	0.002	0.43	-1.76(0.05)	-1.86(0.10)	0.10(0.08)	1.27
	2	-0.427	0.002	0.18	-0.31(0.05)	0.37(0.04)	-0.68(0.05)	-15.01*
	3	-0.025	0.914	0	0.99(0.10)	1.08(0.05)	-0.09(0.08)	-1.14
Item6	1	0.120	0.794	0.01	-3.86(0.17)	-3.83(0.38)	-0.03(0.29)	-0.10
	2	0.046	0.726	0	-1.58(0.07)	-1.67(0.10)	0.09(0.09)	1.04
	3	0.418	0.002	0.17	0.20(0.08)	0.40(0.37)	-0.20(0.27)	-0.75
Item7	1	1.392	0	1.94	-2.61(0.08)	1.87(0.11)	-4.48(0.10)	-46.63*
	2	0.981	0	0.96	-0.81(0.06)	0.28(0.05)	-1.09(0.06)	-19.73*
	3	0.915	0	0.84	0.62(0.08)	1.11(0.06)	-0.49(0.07)	-6.92*
Item8	1	0.179	0.467	0.03	-2.70(0.09)	-2.68(0.18)	-0.02(0.14)	-0.14
	2	0.276	0.025	0.08	-0.66(0.07)	-0.52(0.07)	-0.14(0.07)	-2.00*
	3	0.471	0.007	0.22	-1.08(0.11)	1.34(0.07)	-2.42(0.09)	-26.21*
Item9	1	0.647	0.058	0.42	-3.13(0.10)	-2.65(0.19)	-0.48(0.15)	-3.17*
	2	0.516	0	0.27	-0.99(0.05)	-0.75(0.06)	-0.24(0.06)	-4.35*
	3	0.636	0	0.40	0.64(0.08)	0.95(0.06)	-0.31(0.07)	-4.38*
Item10	1	-1.851	0	3.43	-1.63(0.05)	-2.41(0.15)	0.78(0.11)	6.99*
	2	-1.656	0	2.74	0.04(0.06)	-0.67(0.05)	0.71(0.06)	12.85*
	3	-1.318	0	1.74	1.40(0.12)	0.84(0.05)	0.56(0.09)	6.08*

5 本方法的未来研究趋势以及局限

分步功能差异(DSF)检验法的优点是:(i)测量不变性水平高于 DIF 的整体测量方法. (ii) DSF 方法可以分数水平上(分步水平)分析产生 DIF 的原因. 即如果一个多级计分题目标记有 DIF,那么 DSF 可以分离题目的成分来确定导致 DIF 的原因给题目内容的审核以及修订提供依据. 造成 DIF 的影响因素是修订或者删除题目的关键[18]. (iii)越来越多的研究者对题目认知策略感兴趣[23],这就强调了研究者应在有关认知策略的测量特征上理解组别差异,而 DSF 可以对多级计分题目检测其认知策略的组别差异. 但是,面对一个显著的分步水平 DSF 值,研究者的任务就是将分步水平的 DSF 转为特定分数水平的 DSF. 2 种概念下 DSF 的解释是不同的,由于累积方法的 DSF 稳定性强,所以其是研究者们常用的一种方法. 例如,4 级计分题目的第 2 个分步水平上存在 DSF 表示 2 个最低分数水平到 2 个最高分数水平的过渡对于其中一个组来说要更难. 但是,仅

DSF 是不足以说明哪个高分数水平造成 DSF,有可能是第 3 个分数水平,也有可能是第 4 个分数水平,也有可能两者都有.

一些研究者提出的策略是,如果一个分步水平上存在 DSF(如,第 j 分步水平) 表示在第 j 个分数水平上存在着组间差异,说明 DIF 的产生是由于第 j 个分数水平的特征因素造成的; 如果第 j 和 $j+1$ 个分步水平均存在着组间差异,说明 DIF 的产生是由于第 j 和 $j+1$ 个分数水平的特征因素造成的. 但是通过这种方法计算的结果是有偏的,所以寄予在未来研究中能够发现一种能够对分步水平到分数水平进行准确转化的方法,也希望未来的研究能够更深刻得理解非一致性 DSF,并且进一步对检测非一致性 DSF 的方法进行研究和实践应用. 另外,DSF 是 DIF 研究领域的一种新方法,其可以在分数水平上检测 DSF,从而对 DIF 产生的原因深入探讨,但是无论从方法上来讲,还是从实践上来讲,这种方法还不是很成熟,所以期望未来大量的将其应用于心理测验的实证研究,进而为测验公平性提供充足的证据.

6　参考文献

[1] American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing [M]. Washington D C: American Psychological Association, 1999.

[2] 中国教育学会教育测量与统计分会. 测量术语测验公平性 [J]. 中国考试, 2003, 12(上半月刊) : 19.

[3] Holland P W, Thayer D T. Differential item performance and the Mantel-Haenszel procedure [C]. NJ: Erlbaum, 1998: 129-145.

[4] Penfield R D, Camilli G. Differential item functioning and item bias [C]. New York: Elsevier, 2007: 125-167.

[5] Zumbo B D. Three generations of DIF analyses: considering where it has been, where it is now, and where it is going [J]. Language Assessment Quarterly, 2007, 4(2) : 223-233.

[6] Penfield R D. Assessing differential step functioning in polytomous items using a common odds ratio estimator [J]. Journal of Educational Measurement, 2007, 44(3) : 187-210.

[7] Penfield R D. Three classes of nonparametric differential step functioning effect estimators [J]. Applied Psychological Measurement, 2008, 32(6) : 480-501.

[8] Penfield R D, Gattamorta K, Childs R A. An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items [J]. Educational Measurement: Issues and Practice, 2009, 28(1) : 38-49.

[9] Muraki E. A generalized partial credit model: application of an EM algorithm [J]. Applied Psychological Measurement, 1992, 16(2) : 159-176.

[10] Wim J van der Linden, Ronald K Hambleton. Handbook of modern item response theory [M]. New York: Springer-Verlag New York Inc, 1997: 85-100.

[11] Gattamorta K A. A comparison of adjacent categories and cumulative DSF effect estimators [D]. Miami: University of Miami, 2009.

[12] Cohen A S, Kim S H, Baker F B. Detection of differential item functioning in the graded response model [J]. Applied Psychological Measurement, 1993, 17(4) : 335-350.

[13] Penfield R D. A nonparametric method for assessing differential step functioning in polytomous items [C]. San Francisco: CA, 2006.

[14] Hauck W W. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio [J]. Biometrics, 1979, 35(4) : 817-819.

[15] Jodoin M G, Gierl M J. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection [J]. Applied Measurement in Education, 2001, 14(4) : 329-349.

[16] Thissen D. IRTLRDIF v. 2. 0 b: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. 2001, Unpublished ms.

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

[18] Alvarez K, Penfield R D. Using differential step functioning(DSF) to refine the analysis of DIF in polytomous items: an illustration [C]. Washington D C, 2007.

[19] Penfield R D, Alvarez K, Lee O. Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: an illustration [J]. Applied Measurement in Education, 2009, 22(1) : 61-78.

[20] Penfield R D. Distinguishing between net and global DIF in polytomous items [J]. Journal of Educational Measurement, 2010, 47(2) : 129-149.

[21] Schwarzer R, Jerusalem M. Generalized self-efficacy scale [EB/OL]. [2014-05-16]. www. thefindingsgroup. com.

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

(下转第 609 页)

[12] Cui Yang,Gang Wei,Chen Fangjiong. An estimation-range extended autocorrelation-based frequency estimator [J]. EURASIP Journal on Advances in Signal Processing,2009(10):961938.
[13] 杨德钊,宋凝芳,林志立,等. 基于自相关及相位差法的高精度频率估计算法 [J]. 北京航空航天大学学报,2011,37(8):1030-1033.
[14] Fu H,Kam P Y. Sample-autocorrelation-function-based frequency estimation of a single sinusoid in AWGN [C]// IEEE 75th Vehicular Technology Conference,Yokohama,2012:1-5.
[15] 曹燕. 含噪实信号频率估计算法研究 [D]. 广州: 华南理工大学,2012.
[16] 邹昕,叶志清. 基于量子双向传态的多方量子通信网络的构建方案 [J]. 江西师范大学学报: 自然科学版,2013,37(5):492-496.
[17] Lank G W,Reed I S,Pollon G E. A semicoherent detection and Doppler estimation statistic [J]. Aerospace and Electronic Systems,1973,9(2):151-165.
[18] 吴柳雯,叶志清. 用 4 粒子 Ω 纠缠态实现多粒子隐形传态 [J]. 江西师范大学学报: 自然科学版,2013,37(6):561-564.

An Improved Fitz Frequency Estimation Algorithm with Fast Speed and High Accuracy

WANG Fang,CHEN Yong,YE Zhi-qing
(College of Physics and Communication Electronics,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Fitz frequency estimation algorithm frequency estimation variance in the high SNR is higher and there is a big gap between the CRB. An improved Fitz frequency estimation algorithm,which first defines the modified autocorrelation function weighted by generalized Kay window has been proposed,and then calculates the sum of the weighted phases of the modified autocorrelation function,finally gets the frequency estimation of the complex sinusoidal signal in AWGN. Computer simulation and analysis shows that: the frequency estimation variance of improved algorithm decreases about 2 dB when the data length is 24 and the signal to noise ratio is 20 dB,while the calculated amount of improved algorithm and original algorithm is about the same. In the other words,the proposed algorithm to meet the real-time requirement,achieves a higher frequency estimation precision.

Key words: frequency estimation; autocorrelation; real-time; Cramer-Rao bound

(责任编辑: 冉小晓)

(上接第 599 页)

[23] Leighton J P,Gierl M J. Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes [J]. Educational Measurement: Issues and Practice,2007,26(2):3-16.

The Differential Step Functioning in Polytomous Items

LI Mei-juan[1],LIU Hong-yun[2*]
(1. Beijing Academy of Educational Sciences,Beijing 100191; 2. School of Psychology,Beijing Normal University,Beijing 100875)

Abstract: The research mainly introducesDifferential stepfunctioning (DSF) how to play a role in the examination and interpretation of differential item functioning(DIF) effect. There are two parts in the research. The first part summarizes and reviews models,methods,patterns,applications,result interpretation about DSF abroad,which aims to provide some reference for domestic test fairness. Using DSF analysis methodology by testing actual data,the second part verifies the DIF in test items and differentlevel of steps,and takes further analysis to the reason for DIF production. Therefore,it provides more specific and operational basis for the item review and revision.

Key words: polytomous items; differential item functioning(DIF); differential step functioning (DSF)

(责任编辑: 冉小晓)