汪文义,何韵玲,宋丽红,等. 匹配变量纯化的测验偏差检验方法[J]. 江西师范大学学报(自然科学版),2022,46(5):447-452.
WANG Wenyi,HE Yunling,SONG Lihong,et al. The matching score purification for differential item functioning method[J]. Journal of Jiangxi Normal University(Natural Science),2022,46(5):447-452.

文章编号:1000-5862(2022)05-0447-06

匹配变量纯化的测验偏差检验方法

汪文义[1],何韵玲[1],宋丽红[2*],黄 涛[1]
(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 江西师范大学教育学院,江西 南昌 330022)

摘要: CSIBTEST方法是基于参照组和目标组2个测验信度对真分数进行估计,再按交叉位置分数将匹配分数划分为2类子样本,并分别计算其卡方统计量,然后将这2个独立的卡方统计量相加得到自由度为2的检验统计量. 鉴于测验信度具有群体依赖性,即不同群体的测验信度可能不尽相同,而CSIBTEST方法将参照组和目标组分别划分为2类子样本,有必要对子样本上的测验信度也进行估计,由此拓展了CSIBTEST. 新方法先使用CSIBTEST获得交叉位置参数,相当于进行DIF预分析,再使用子样本上的信度估计用于真分数估计,以在对匹配变量进行纯化后获得检测统计量. 模拟研究结果显示:相比SIBTEST和CSIBTEST,匹配变量纯化的测验偏差检验方法对存在DIF试题有着更高的统计检验力.

关键词: 测验偏差;项目功能差异;CSIBTEST;信度;考试公平
中图分类号: B 841 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2022.05.02

0 引言

从2018年到2022年的全国两会,“公平”与“质量”2个关键词始终贯穿政府工作报告的教育部分,如2022年的“促进教育公平与质量提升”. 2020年10月中共中央、国务院印发的《深化新时代教育评价改革总体方案》是教育评价领域的一份纲领性文件,凸显了教育评价在新时代教育事业发展中的重要地位[1]. 教育评价改革是新时代深化教育改革创新的关键突破口,是落实立德树人教育根本任务的重要方向标,是推动教育高质量发展、实现创新人才培养的重要保证. 教育学者需要结合新时代教育改革创新的国家战略,针对教育评价改革、新时代教育公平等一系列热点和重点问题展开系统地研究[2]. 有研究显示增值性教育评价本身的效果与想象的效果差别很大,并没有想象的那么公平和准确[3-4]. 当智能技术大规模应用于评价时,尤其要警惕不可解释、不透明的算法所做出的评价判断. 如用东南沿海地区初中生的英语口语语料训练的自动评分模型,全国使用或者用于西部地区就可能存在公平性问题[5].

测验公平是测验效度的基本问题. 测量偏差(test bias)是指与测验所测构念无关因素导致相同构念能力被试组在测验分数上的系统差异,通常采用项目功能差异(differential item functioning,DIF)来分析测验偏差. 常见的DIF方法主要有MH方法和GMH方法、STD和SMD方法、LR和LDFA方法、SIBTEST和P-SIBTEST方法等[6]. 下面主要关注SIBTEST及其推广方法CSIBTEST. SIBTEST是由R. Shealy等[7]于1993年提出的均匀或单向DIF检验方法,张华华等[8]在此基础上提出了多级评分的SIBTEST方法. 在SIBTEST方法基础之上,H. H. Li等[9]于1996年提出非均匀或双向DIF检验方法CSIBTEST,并采用模拟方法得到其统计量的抽样分布进行假设检验. CSIBTEST根据参照组和目标组的项目反应函数的交叉位置,对2个分段的SIBTEST统计量求和,以避免SIBTEST统计量正负相加而抵消. R. P. Chalmers等[10]对原CSIBTEST的统计量进行修改并使用卡方分布作为新统计量的渐近抽样分布. 本文主要使用R. P. Chalmers所改进的SIBTEST和CSIBTEST.

收稿日期:2022-04-25
基金项目:江西省社会科学基金(17JY10)和国家自然科学基金(62267004,62067005,61967009)资助项目.
通信作者:宋丽红(1981—),女,江西新干人,副教授,博士,主要从事教育测量研究. E-mail:viviansong1981@163.com

测验偏差检验方法 SIBTEST 和 CSIBTEST 主要基于克隆巴赫 α 信度系数[11]对真分数估计. 而近年来对克隆巴赫 α 信度系数受到了诸多批判和争论[12],本文将其他信度估计应用于 SIBTEST 和 CSIBTEST 方法中,并检验了信度估计对测验偏差检验的影响. 测验信度往往基于测验作答反应进行估计,不同群体的测验作答反应或题目间协方差矩阵不尽相同. 因此,测验信度具有群体依赖性,即不同群体的测验信度可能不尽相同[13-15]. CSIBTEST 方法基于参照组和目标组 2 个测验信度对真分数进行估计,再按照交叉位置分数将匹配分数划分为 2 类子样本,并分别计算其卡方统计量,然后将 2 个独立的卡方统计量相加得到自由度为 2 的检验统计量. 既然 CSIBTEST 方法将参照组和目标组按匹配分数划分为 2 类子样本,有必要对子样本上的测验信度也进行估计,将其应用于真分数估计,这样可校准能力分布差异,从而在参照组和目标组存在能力分布差异时偏差检验可能更有价值. 新方法先使用 CSIBTEST 获得交叉位置参数,相当于进行 DIF 预分析,再使用子样本上信度估计用于真分数估计,对匹配变量进行在一定程度上纯化,期望新方法可更高效检验存在 DIF 试题. 因为新方法源于匹配变量纯化的 DIF 检验思想[16-17],故被称为匹配变量纯化的测验偏差检验方法.

1 研究方法

1.1 信度系数

测验信度定义为测验真分数与测验总分方差之比 $\rho_{XX'}=\sigma^2(T)/\sigma^2(X)$. 克隆巴赫 α 系数是在真分数理论下提出来的信度估计[11]. 在真分数理论下,观测分数可以表示为真分数和误差分数之和,即被试 i 在第 j 题上的观测得分 X_{ij}、真分数 T_{ij} 和误差分数 E_{ij} 之间的关系可以表示为

$$X_{ij}=T_{ij}+E_{ij}=T_i+v_j+E_{ij}, \quad (1)$$

式(1)也被称为基本真分数等价(essential tau equivalency),将不同题目的真分数差异量限制为常数 v_j[12]. 若将式(1)看成因子分析模型特例,则 T_i、v_j 和 E_{ij} 分别代表(全局)因子得分、局部因子得分(题目的截距)和特殊因子分数. 含有 J 个试题的测验观察总分的方差可以表示为测验真分数方差和误差方差之和:

$$\sigma^2\left(\sum_{j=1}^{J}X_{ij}\right)=\sigma^2\left(\sum_{j=1}^{J}T_{ij}\right)+\sigma^2\left(\sum_{j=1}^{J}E_{ij}\right), \quad (2)$$

式(2)可简记为 $\sigma^2(X)=\sigma^2(T)+\sigma^2(E)$. 若将测验的观测得分矩阵的题目协方差矩阵记为 $\mathrm{Var}(X)=(\sigma_{jj'})$,其中 $\sigma_{jj'}$ 表示题目 j 和题目 j' 的协方差. 在误差分数与真分数不相关、题目之间的误差分数不相关并且误差均值为 0 的假设下[18],克隆巴赫 α 系数可将 $\mathrm{Var}(X)$ 中非对角线的均值 $\overline{\sigma}_1$ 作为题目真分数方差估计量,由此得到其信度估计公式[11]为

$$\alpha=J^2\overline{\sigma}_1/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right)=\frac{J}{J-1}\left(1-\sum_{j=1}^{J}\sigma_{jj}/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right)\right),$$

其中 $\overline{\sigma}_1=\sum_{j=1}^{J}\sum_{j'=1,j'\neq j}^{J}\sigma_{jj'}/(J(J-1))$.

若将 $\mathrm{Var}(X)$ 中非对角线元素的平方和 C_2 的函数用于估计真分数方差,可得到下面的信度估计公式[19-20]:

$$\lambda_2=\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}-\sum_{j=1}^{J}\sigma_{jj}+\sqrt{JC_2/(J-1)}\right)/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right).$$

在不同题目上因子负荷会有所不同,从而有如下单因子模型或同质模型(congeneric model)[12]:

$$X_{ij}=T_{ij}+E_{ij}=\lambda_jT_i+v_j+E_{ij},$$

若公共因子方差 $\mathrm{Var}(T_i)=1$,测验真分数方差 $\sigma^2\left(\sum_{j=1}^{J}T_{ij}\right)=\left(\sum_{j=1}^{J}\lambda_j\right)^2$,则信度估计公式[21-24]为

$$\omega_h=\left(\sum_{j=1}^{J}\lambda_j\right)^2/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right).$$

ω_h 也被称为测验的同质性信度[25].

GLB 信度系数采用以下多因子模型估计信度的上限,多因子模型为

$$X_{ij}=T_{ij}+E_{ij}=\sum_{k=1}^{K}\lambda_{jk}T_{ik}+v_j+E_{ij},$$

其中因子个数 K 为协方差矩阵 $\mathrm{Var}(X)$ 正特征值的个数. 每个题目的误差可通过共同度 $h_j^2=\sum_{k=1}^{K}\lambda_{jk}^2$ 来计算[20,26]. g_{lb} 的计算方法为

$$g_{lb}=1-\sum_{j=1}^{J}\sigma^2(e_j)/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right)=1-\sum_{j=1}^{J}(1-h_j^2)/\left(\sum_{j=1}^{J}\sum_{j'=1}^{J}\sigma_{jj'}\right).$$

1.2 匹配变量纯化的测验偏差检验方法

在新方法中,匹配变量纯化主要采用信度系数分组计算真分数回归,再基于分组的真分数估计并结合 CSIBTEST 进行测验偏差检验. CSIBTEST 是基于 SIBTEST 而提出的交叉或双向偏差检验方法. SIBTEST 是一种基于显著性检验的测验偏差检验的

非参数方法. 它是由 R. Shealy 等[7] 提出的，主要被用于测量工具单向偏差的检验方法中.

记参照组 $g_1 = R$ 和目标组 $g_2 = F$，样本量分别为 n_R 和 n_F. 测验项目数为 J，得分矩阵为 $\boldsymbol{U}_g = (\boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_m)$，其中 U_{gij} 表示在被试组 g 中被试 i 在项目 j 上的观测得分. 记匹配或有效项目集 S_1，项目数 $J_1 = K$. 剩下项目组成探查项目集 S_2，项目数 $J_2 = J - K$. 记参照组或目标组在有效项目集 S_1 上总分为 k 分的被试集合 A_{gk} 和对应的被试数 n_{gk}，合并组的总分为 k 分的人数比率为 $\hat{p}_k = (n_{Rk} + n_{Fk}) / (n_R + n_F)$，其中 $k = 0, 1, \cdots, K$. 基于参照组和目标组的测验信度，使用 CSIBTEST 方法由加权最小二乘法所得到的交叉位置分数记为 k_c. 匹配变量纯化的测验偏差侦查方法的主要步骤如下.

1) 基于 2 组的 2 个子样本的信度系数估计真分数. 考虑信度系数的估计与样本量有关，直接以匹配项目集上总分的 1/2(即 $K/2$) 为划界分数，将参照组和目标组的被试都分成高分组和低分组 2 个子样本 P_{g1} 和 P_{g2}，再计算对应的平均总分 $\bar{X}_{g1}$ 和 $\bar{X}_{g2}$，并使用 1.1 节信度估计方法得到各样本的信度估计值 $\hat{\rho}_{gc}(c = 1, 2)$，然后将其用于估计 $k(k = 0, 1, \cdots, K)$ 所对应的真分数:

$$\hat{V}_g(k) = \begin{cases} (\bar{X}_{g1} + \hat{\rho}_{g1}(k - \bar{X}_{g1}))/K, k < k_c, \\ (\bar{X}_{g2} + \hat{\rho}_{g2}(k - \bar{X}_{g2}))/K, k > k_c. \end{cases}$$

2) 采用线性插值方法估计 $\bar{Y}^*_{gk}$. 因为在参照组或目标组上能力分布差异会影响在匹配分数 k 下 2 组待探查项目集上被试总分的条件均值 $\bar{Y}_{gk}$，所以需要采用线性插值方法估计 $\bar{Y}^*_{gk}$:

$$\bar{Y}^*_{gk} = \bar{Y}_{gk} + \hat{M}_{gk}(\hat{V}(k) - \hat{V}_g(k)),$$

其中 $\bar{Y}_{gk} = \frac{1}{n_{gk}} \sum_{i \in A_{gk}} \sum_{j \in S_2} U_{gij}$，$\hat{M}_{gk} = (\bar{Y}_{g(k+1)} - \bar{Y}_{g(k-1)}) / (\hat{V}_g(k+1) - \hat{V}_g(k-1))$，$\hat{V}(k) = (\hat{V}_R(k) + \hat{V}_F(k))/2$.

3) 计算偏差估计量及其标准差. 基于交叉位置分数 k_c 和校正估计 $\bar{Y}^*_{gk}$，计算探查项目集上偏差估计量及其标准差:

$$\hat{\beta}^{(b)}_{\text{uni}} = \sum_{k=0}^{k_c-1} \hat{p}_k(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}), \hat{\beta}^{(u)}_{\text{uni}} = \sum_{k=k_c+1}^{K} \hat{p}_k(\bar{Y}^*_{Fk} - \bar{Y}^*_{Rk}),$$

$$\hat{\sigma}^2(\hat{\beta}^{(b)}_{\text{uni}}) = \sum_{k=0}^{k_c-1} \hat{p}_k^2(\hat{\sigma}^2(Y_R \mid k)/n_{Rk} + \hat{\sigma}^2(Y_F \mid k)/n_{Fk}),$$

$$\hat{\sigma}^2(\hat{\beta}^{(u)}_{\text{uni}}) = \sum_{k=k_c+1}^{K} \hat{p}_k^2(\hat{\sigma}^2(Y_R \mid k)/n_{Rk} + \hat{\sigma}^2(Y_F \mid k)/n_{Fk}),$$

其中 $\hat{\sigma}^2(Y_g \mid k) = S^2(Y_R \mid k) = \sum_{i \in A_{gk}} (\sum_{j \in S_2} U_{gij} - \bar{Y}_{gk})^2 / (n_{gk} - 1)$.

4) 构建检验统计量. 在原假设 $H_0: \beta_{cro} = 0$ 成立下，服从自由度为 2 的卡方检验统计量为

$$B^2 = (\hat{\beta}^{(b)}_{\text{uni}})^2 / \hat{\sigma}^2(\hat{\beta}^{(b)}_{\text{uni}}) + (\hat{\beta}^{(u)}_{\text{uni}})^2 / \hat{\sigma}^2(\hat{\beta}^{(u)}_{\text{uni}}) \sim \chi^2(2). \quad (3)$$

当没有交叉位置参数时，式(3) 退化为

$$\left(\sum_{k=0}^{K} \hat{p}_k(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk})\right)^2 / \left(\sum_{k=1}^{K} \hat{p}_k^2(\hat{\sigma}^2(Y_R \mid k)/n_{Rk} + \hat{\sigma}^2(Y_F \mid k)/n_{Fk})\right) \sim \chi^2(1).$$

5) 给出检验结论. 若 $B^2 > \chi^2_\alpha(m)$ (m 为自由度)，则拒绝原假设，否则不能拒绝原假设，其中显著性水平设置为 $\alpha = 0.05$.

2　模拟研究

2.1　研究设计

采用蒙特卡罗模拟验证新提出的匹配变量纯化的测验偏差侦查方法，检验其在各种条件下的表现. 参照已有相关研究的实验设计[10]，得分矩阵模拟采用 2 参数 Logistic 模型[27-28]:

$$P(U = 1 \mid a, d, \theta) = 1/(1 + \exp(-a\theta - d)),$$

其中截距参数服从标准正态分布(即 $d \sim N(0, 1)$)，区分度参数 a 服从对数正态分布(其中 $\mu = 0.25$，$\sigma = 0.5$)，$f(a) = e^{-(\ln a - \mu)^2/(2\sigma^2)} / (\sqrt{2\pi}\sigma)$.

参照组的能力水平服从 $N(0, 1)$，而目标组的能力分为 2 个水平，分别服从 $N(0, 1)$ 和 $N(0.25, 0.75)$. 参照组和目标组的样本量分为 2 个水平，均为 1 000 或 3 000. 测试长度固定为 25，第 25 个项目为存在偏差项目. 项目的偏差主要分为 3 种类型. 第 1 种为区分度不同类型(DIF-A)，即在参照组和目标组上的区分度分别为 1.0 和 1.5，而截距均为 0; 第 2 种为难度不同类型(DIF-B)，即在参照组和目标组上的截距分别为 0 和 −0.2，而区分度均为 1.0; 第 3 种为混合类型(DIF-AB)，在参照组和目标组上的区分度分别为 1.0 和 1.2、难度分别为 0 和 −0.2.

在 mirt 包中可以找到 SIBTEST 和修改的 CSIB-TEST 统计量的开源实现，以及推荐的渐近抽样分布[29]. 因此，对比方法使用在 mirt 包中 SIBTEST 和 CSIBTEST 的实现[10]. 同时，对 SIBTEST 函数代码进

行适当修改以得到本文所提出的新方法,简记为GCSIB. 每种实验条件重复 500 次,以得到方法的第Ⅰ类错误和统计检验力. 本文还考虑了 4 种信度估计对 3 种方法表现的影响.

2.2 实验结果

1) 第Ⅰ类错误率. 表 1 给出了在各条件下测验偏差侦查方法的第Ⅰ类错误率. 模拟结果表明:SIBTEST 和 CSIBTEST 能够达到与显著性水平相近的第Ⅰ类错误率,即第Ⅰ类错误率接近显著性水平 0.05. 而新方法在目标组为标准正态分布时的第Ⅰ类错误率略高于显著性水平,而当目标组存在能力分布差异时,新方法第Ⅰ类错误率控制较差. 而项目偏差类型、信度估计、样本量对各种方法的第Ⅰ类错误率影响较小.

表 1 在各种条件下测验偏差侦查方法的第Ⅰ类错误率

目标组	偏差类型	信度	$N=1\,000$			$N=3\,000$		
			SIB	CSIB	GCSIB	SIB	CSIB	GCSIB
$N(0,1)$	DIF-A	α	0.058	0.068	0.110	0.042	0.068	0.110
		λ_2	0.058	0.068	0.114	0.042	0.066	0.102
		ω_h	0.058	0.068	0.090	0.044	0.070	0.108
		g_{lb}	0.058	0.068	0.102	0.042	0.068	0.102
	DIF-B	α	0.060	0.076	0.090	0.036	0.044	0.080
		λ_2	0.060	0.076	0.088	0.034	0.044	0.078
		ω_h	0.060	0.076	0.080	0.040	0.050	0.070
		g_{lb}	0.060	0.076	0.086	0.036	0.044	0.074
	DIF-AB	α	0.056	0.058	0.090	0.058	0.070	0.102
		λ_2	0.056	0.058	0.094	0.058	0.070	0.114
		ω_h	0.060	0.058	0.064	0.058	0.064	0.086
		g_{lb}	0.056	0.058	0.080	0.058	0.070	0.102
$N(0.25,0.75)$	DIF-A	α	0.066	0.082	0.288	0.044	0.066	0.378
		λ_2	0.066	0.082	0.282	0.044	0.062	0.372
		ω_h	0.058	0.086	0.128	0.054	0.076	0.256
		g_{lb}	0.066	0.082	0.254	0.044	0.066	0.378
	DIF-B	α	0.052	0.058	0.286	0.058	0.068	0.382
		λ_2	0.052	0.058	0.274	0.056	0.068	0.370
		ω_h	0.054	0.058	0.128	0.054	0.080	0.266
		g_{lb}	0.052	0.060	0.260	0.054	0.072	0.372
	DIF-AB	α	0.044	0.050	0.262	0.062	0.074	0.440
		λ_2	0.044	0.048	0.260	0.062	0.074	0.420
		ω_h	0.050	0.056	0.120	0.068	0.084	0.276
		g_{lb}	0.044	0.050	0.250	0.062	0.074	0.424
		M	0.056	0.066	0.162	0.050	0.066	0.228

注:SIB,CSIB 和 GCSIB 分别表示 SIBTEST、CSIBTEST 和新方法,下表同.

2) 统计检验力. 表 2 给出了在各种条件下测验偏差检验方法的正确拒绝虚无假设的比率(即正确拒绝错误假设的比率). 从总体来看,SIBTEST 和 CSIBTEST 正确拒绝虚无假设的比率均低于新方法的,这表示新方法在识别存在偏差的题目上存在一定的优势. 各方法的表现受样本量影响较大,当样本量从1 000增加到 3 000 时,SIBTEST、CSIBTEST 和新方法正确拒绝虚无假设的比率增幅分别为 30%、44%和 34%.

当参照组和目标组的能力分布存在差异时,SIBTEST 和 CSIBTEST 的检验力有下降趋势;反过来,新方法在能力分布存在差异时表现相当好. 而项目偏差类型、信度估计、样本量对各种方法的检验力影响较小. 当样本量为 1 000 时,信度估计 w_h 与其他信度估计在各方法正确拒绝虚无假设的比率上有所差异,并且当样本量增大为 3 000 时,这种差异仍存在.

SIBTEST 对于 DIF-A 类型的偏差侦查表现较差,这主要是因为该方法主要针对单向偏差检验而设计,SIBTEST 对 DIF-B 和 DIF-AB 偏差类型均有较好的识别能力. CSIBTEST 正确识别 DIF-A 偏差类型比 DIF-B 和 DIF-AB 的检验力更高,当样本量为 3 000时,新方法对 3 种偏差类型的正确拒绝虚无假设的比率均值高达 95%.

表 2　在各种条件下测验偏差侦查方法的统计检验力

目标组	偏差类型	信度	$N=1\ 000$			$N=3\ 000$		
			SIB	CSIB	GCSIB	SIB	CSIB	GCSIB
$N(0,1)$	DIF-A	α	0.066	0.596	0.604	0.056	0.978	0.980
		λ_2	0.066	0.600	0.606	0.056	0.978	0.980
		ω_h	0.064	0.598	0.596	0.056	0.978	0.982
		g_{lb}	0.066	0.598	0.602	0.056	0.978	0.980
	DIF-B	α	0.510	0.488	0.500	0.948	0.932	0.920
		λ_2	0.510	0.490	0.504	0.948	0.932	0.916
		ω_h	0.508	0.496	0.502	0.948	0.934	0.906
		g_{lb}	0.510	0.488	0.504	0.948	0.932	0.920
	DIF-AB	α	0.426	0.440	0.432	0.884	0.912	0.898
		λ_2	0.426	0.434	0.436	0.884	0.910	0.894
		ω_h	0.426	0.438	0.424	0.884	0.906	0.904
		g_{lb}	0.426	0.440	0.436	0.884	0.912	0.896
$N(0.25,0.75)$	DIF-A	α	0.074	0.560	0.744	0.122	0.962	0.994
		λ_2	0.074	0.554	0.734	0.130	0.964	0.992
		ω_h	0.084	0.532	0.608	0.148	0.966	0.986
		g_{lb}	0.074	0.560	0.736	0.124	0.964	0.992
	DIF-B	α	0.500	0.498	0.678	0.946	0.936	0.948
		λ_2	0.496	0.494	0.674	0.946	0.936	0.952
		ω_h	0.468	0.466	0.582	0.930	0.918	0.914
		g_{lb}	0.498	0.494	0.672	0.946	0.936	0.942
	DIF-AB	α	0.374	0.356	0.804	0.822	0.836	0.992
		λ_2	0.370	0.354	0.802	0.820	0.828	0.992
		ω_h	0.330	0.338	0.660	0.792	0.792	0.978
		g_{lb}	0.374	0.356	0.800	0.820	0.830	0.992
		M	0.322	0.486	0.610	0.629	0.923	0.952

3　结论与讨论

考虑在测验偏差检验方法 SIBTEST 和 CSIBTEST 方法中主要使用克隆巴赫 α 信度系数,本文将其他信度估计应用于 SIBTEST 和 CSIBTEST 方法中,并检验了信度估计对测验偏差检验的影响. 研究结果显示,不同信度的估计对测验偏差检验的影响较小. 同时注意到测验信度具有群体依赖性,对 CSIBTEST 方法将参照组和目标组按匹配分数划分为 2 类子样本,并分别进行测验信度估计,将其应用于真分数估计,从而提出了改进的 CSIBTEST. 研究结果表明: 在参照组和目标组能力分布存在差异时,改进的 CSIBTEST 表现出相当优良的统计检验力.

虽然改进的 CSIBTEST 可以校正参照组和目标组能力分布的差异,并且表现出相当优良的统计检验力,但是新方法会以较高概率将没有偏差试题判为存在偏差. 从犯 2 类错误的风险来看,人们对高风险考试的公平性关注度较高. 因此,误将有偏差的试题视为无偏差,显然比误将无偏差的试题视为有偏差的风险更高. 从实际应用来看,可以充分发挥多种测验偏差检验方法的优势,新方法或者多种方法均将测验中某题判断为存在偏差试题,应该高度怀疑此题存在偏差. 对于统计检验方法怀疑的存在偏差试题及其偏差量可反馈给专家以做进一步判断.

4　参考文献

[1] 司林波,裴索亚,王伟伟. 新中国教育评价制度变迁的影响因素、基本规律与实践启示: 基于教育评价相关政策文本的扎根理论研究[J]. 大学教育科学,2021(6):69-77.

[2] 宣小红,檀慧玲,曹宇新. 教育学研究的热点与未来展望: 基于 2021 年度人大复印报刊资料《教育学》转载论文的分析[J]. 教育研究,2022,43(2):70-82.

[3] 赵勇. 教育评价的几大问题及发展方向[J]. 华东师范大学学报(教育科学版),2021,39(4):1-14.

[4] BITLER M, CORCORAN S, DOMINA T, et al. Teacher effects on student achievement and height: a cautionary tale[J]. Journal of Research on Educational Effectiveness,2021,14(4):900-924.

[5] 张志祯,齐文鑫. 教育评价中的信息技术应用: 赋能、挑战与对策[J]. 中国远程教育,2021(3):1-11,76.

[6] 汪文义,张华华. 统计测量视角下考试公平推动教育公平的对策[J]. 江西师范大学学报(自然科学版),2017,41(4):383-393.

[7] 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.
[8] 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.
[9] LI H H,STOUT W. A new procedure for detection of crossing DIF [J]. Psychometrika,1996,61(4):647-677.
[10] CHALMERS R P. Improving the crossing-SIBTEST statistic for detecting non-uniform DIF [J]. Psychometrika,2018,83(2):376-386.
[11] CRONBACH L J. Coefficient alpha and the internal structure of tests [J]. Psychometrika,1951,16(3):297-334.
[12] GREEN S B,YANG Yanyun. Commentary on coefficient alpha: a cautionary tale [J]. Psychometrika,2009,74(1):121-135.
[13] ANDERSSON B,LUO Hao,MARCQ K. Reliability coefficients for multiple group item response theory models [J]. British Journal of Mathematical and Statistical Psychology,2022,75(2):395-410.
[14] RAYKOV T. Examining group differences in reliability of multiple-component instruments [J]. British Journal of Mathematical and Statistical Psychology,2002,55(1):145-158.
[15] BENTLER P M. Covariate-free and covariate-dependent reliability [J]. Psychometrika,2016,81(4):907-920.
[16] LEE H S,GEISINGER K F. The matching criterion purification for differential item functioning analyses in a large-scale assessment [J]. Educational and Psychological Measurement,2016,76(1):141-163.
[17] CHEN Chengte,HWU B S. Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: the PISA example [J]. Applied Psychological Measurement,2018,42(3):206-220.
[18] 漆书青,戴海崎,丁树良. 现代教育与心理测量学原理 [M]. 北京: 高等教育出版社,2002.
[19] GUTTMAN L. A basis for analyzing test-retest reliability [J]. Psychometrika,1945,10(4):255-282.
[20] REVELLE W,YOVEL I. Psych: procedures for psychological, psychometric, and personality research [EB/OL]. [2022-01-06]. https://cran.r-project.org/web/packages/psych/psych.pdf.
[21] MCDONALD R P. Test theory: a unified treatment [M]. New Jersey: Erlbaum,1999.
[22] ZINBARG R E,REVELLE W,YOVEL I,et al. Cronbach's α, Revelle's β, and McDonald's ω_H: their relations with each other and two alternative conceptualizations of reliability [J]. Psychometrika,2005,70(1):123-133.
[23] ZINBARG R E,REVELLE W,YOVEL I. Estimating ω_h for structures containing two group factors: perils and prospects [J]. Applied Psychological Measurement,2007,31(2):135-157.
[24] JORGENSEN T D,PORNPRASERTMANIT S,SCHOEMANN A M,et al. semTools: useful tools for structural equation modeling [EB/OL]. [2022-01-06]. https://cran.r-project.org/web/packages/semTools/semTools.pdf.
[25] 温忠麟,叶宝娟. 测验信度估计: 从 α 系数到内部一致性信度 [J]. 心理学报,2011,43(7):821-829.
[26] REVELLE W,ZINBARG R E. Coefficients alpha,beta,omega and the glb: comments on Sijtsma [J]. Psychometrika,2009,74(1):145-154.
[27] BENTLER P M. Alpha,FACTT,and beyond [J]. Psychometrika,2021,86(4):861-868.
[28] LORD F M,NOVICK M R. Statistical theory of mental test scores [M]. Massachusetts: Addison-Wesley,1968.
[29] CHALMERS R P. mirt: a multidimensional item response theory package for the R environment [J]. Journal of Statistical Software,2012,48(6):1-29.

The Matching Score Purification for Differential Item Functioning Method

WANG Wenyi[1],HE Yunling[1],SONG Lihong[2*],HUANG Tao[1]
(1. School of Computer and Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. School of Education,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: The CSIBTEST method estimates the true scores based on the test reliability of the reference group and the focus group,then separates the matching scores into two kinds of sub-samples according to the crossing location,and computes the Chi-squared statistics respectively,and then adds the two independent statistics to obtain the test statistics with a degree of freedom of 2. From the view of the group dependence of test reliability,that is,the test reliability of different groups may be different,and the CSIBTEST method separates the reference group and the focus group into two sub-samples respectively,it is necessary to estimate the test reliability on the sub-samples for the extension of the CSIBTEST. The new method first uses the CSIBTEST to obtain the cross location,and then applies the reliability estimation on the sub-samples for the true score estimation as matching score purification to obtain the test statistic. The simulation study shows that the new method with matching score purification has higher statistical test power for the bias item than the SIBTEST and CSIBTEST.

Key words: test bias; differential item functioning; CSIBTEST; reliability; the fairness of testing

(责任编辑: 冉小晓)