心理学报 2013, Vol. 45, No.8, 921−934
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2013.00921

IRT_Δb 法和修正 LR 法对矩阵取样 DIF 检验的有效性

张 勋[1] 李凌艳[1] 刘红云[2] 孙 研[1]

([1]北京师范大学认知神经科学与学习国家重点实验室, 北京 100875)
([2]北京师范大学心理学院, 北京 100875)

摘 要 矩阵取样测验包含多个题册, 单个题册的总分不能直接作为匹配变量用于 DIF 检测。本研究首先基于模拟数据, 同时采用 IRT_Δb 法, 以及用 IRT 模型估计的考生能力作为匹配变量修订后的 LR 法对矩阵取样测验进行 DIF 检测, 分析二者进行 DIF 检测的有效性及其相关影响因素; 并根据已有的 LR 法 DIF 判断标准划定出 IRT_Δb 法分类标准; 最后使用实证数据加以验证。结果显示：矩阵取样测验中, IRT_Δb 法和修正 LR 法均能较好地区分 DIF 量不同的题目; 样本量、题册中 DIF 题目的比例和考生群体间真实能力的差异对两种方法的检验力、犯 I 类错误的概率和分类结果都有较大影响。

关键词 矩阵取样测验; 项目功能差异; Rasch 模型; Logistic 回归
分类号 B841

1 问题提出

项目功能差异(Differential Item Functioning, DIF)是一种用于判断测验题目是否存在偏差的测量学指标。经过几十年的研究, 测量学家提出很多用于对测验进行 DIF 检测的统计方法(Clauser & Mazor, 1998)。目前, 大部分方法在实践应用中被广泛接受, 在保证测验的公平性方面发挥着重要的作用(Allen, Donoghue, & Schoeps, 2001; Organization for Economic Co-operation and Development, 2005, 2009)。

常规的 DIF 检验方法通常基于单题本测验, 在这种测验情境下, 测验总分能较好地反映考生群体的真实能力, 因而可以直接作为匹配变量用于对题目进行 DIF 检验(Dorans & Holland, 1993; Holland & Thayer, 1988; Swaminathan & Rogers, 1990; Thissen, Steinberg, & Wainer, 1993)。近年来, 随着教育测量与大规模评价项目的发展, 矩阵取样(Matrix Sampling)技术得到广泛使用。矩阵取样测验采用多套平行题本进行测试, 各题本包含的内容仅仅是整个测验的一部分(李凌艳, 辛涛, 董奇, 2007), 不能代表整个测验的构念(Construct)。而且, 不同的考生组作答不同的题本, 导致完成不同题本的考生的总分不能进行直接比较。如果将考生在单个题本上的总分作为衡量考生群体真实能力的匹配变量进行 DIF 检测, 所得结果将是不准确的(Allen & Donoghue, 1996)。因此, 与传统单题本测验相比, 如何选择并控制能较为准确衡量考生群体真实能力的匹配变量就成为了对矩阵取样测验进行 DIF 检测的首要问题。

目前, 在大规模教育测评项目中针对矩阵取样测验使用的 DIF 检测方法主要有两类, 一类是美国教育考试服务中心 (Educational Testing Service, ETS) 用于美国教育进展评估项目 (National Assessment of Educational Progress, NAEP)的方法, 其联合多个题本的总分对考生群体真实能力进行匹配, 然后采用 Mantel-Haenszel (MH) 法和 SIBTEST 法进行 DIF 检测(Allen et al., 2001); 另一类是国际学生评价项目 (Programme For International Student Assessment, PISA) 采用的 IRT_Δb 法, 即基于 Rasch 模型来判

收稿日期: 2011−08−02
通讯作者: 李凌艳, E-mail: lilingyan@bnu.edu.cn

断由不同群体考生估计得到的题目难度之间差异的方法(Organization for Economic Co-operation and Development, 2005, 2009)。NAEP 的方法保留了 MH 方法和 SIBTEST 方法的优点, 如可以对待检题目按照 ETS 设定的“ABC”标准进行分类处理(Dorans & Holland, 1993)。然而, NAEP 的方法忽略了“完成不同题本的考生所得的总分不能直接比较”这一事实; 同一分数段上, 作答不同题本的考生其真实能力存在差异, 这可能对检测结果产生影响。此外, 由于仅匹配了待检测题目所在题本的总分, 其所得测验总分不能完整地代表整个测验所要测验的内容, 因此不够准确。PISA 的方法由于使用 Rasch 模型对所有题本的题目进行同时性估计(Concurrent Calibration), 对考生群体能力的匹配更加准确。然而, 该方法所得的衡量 DIF 效应的 Δb 值尚缺乏一致的分类判断标准。

总体来说, 目前针对矩阵取样测验进行 DIF 检测的研究相对较少, 尚不能对下列问题加以回答: 在矩阵取样条件下, 不同 DIF 检验方法的有效性如何? DIF 效应的检测受何种因素影响? 不同方法检验出的 DIF 效应分类标准是否可以比较或参照? 而对上述问题的回答能帮助研究者基于自身测验设计的特点选择适合的 DIF 检测方法, 确保测验编制的质量与公平。

2 矩阵取样测验 DIF 检测方法

在矩阵取样测验中, 基于项目反应理论 IRT 模型得到的能力值比测验总分能更好地反映考生群体的真实能力, 而且对于完成不同题本的考生, 基于 IRT 模型得到的能力值通过等值技术可以被标定到同一个尺度上, 从而进行直接比较和进一步计算(Dings, Childs, & Kingston, 2002)。因此, 对于那些常规的、基于 IRT 模型类的 DIF 检测方法(Camilli, 2006; Thissen et al., 1993), 可以通过对多个题本进行同时估计, 然后直接用于矩阵取样测验进行 DIF 检测, 如 PISA 所使用的 IRT_Δb 法; 而对于那些基于观测分数(题本总分)的检测方法, 如逻辑斯蒂克回归方法 (Logistic Regression, LR)和 MH 法等(Holland & Thayer, 1988; Shealy & Stout, 1993; Swaminathan & Rogers, 1990), 则可用基于 IRT 模型得到的能力值替代题本总分作为匹配变量, 以修正其对考生群体能力匹配的准确性, 从而很好地用于对矩阵取样测验的 DIF 检测。

2.1 IRT_Δb 法

IRT_Δb 法的基本思想是, 在 Rasch 模型中, 如果某道题目的作答概率不能被考生的能力及该题目的参数完全解释时, 则说明这道题目存在 DIF (Wu, Adams, Wilson, & Haldane, 2007)。在使用 Rasch 模型标定题目的难度时, 如果分别使用两个亚群体 (比如男生和女生) 的作答数据对相同的题目集进行两次独立的 Rasch 分析, 而且两次分析估计得到的题目难度都能很好地拟合各自的数据, 那么对两次独立分析的题目参数进行校准后, 如果题目的难度值在两次分析中的差异 (Δb) 依然显著, 则说明该题目在这两个亚群体间存在着 DIF。难度值差异可以通过近似正态分布进行显著性判断, 公式[1]如下:

$$Z = \frac{\Delta b}{SE(\hat{b})} \tag{1}$$

在矩阵取样设计下, 由于每个学生只完成所有测试题目中的部分题目, 因此对于每个学生能力的估计误差较大。传统的对个体能力估计的方法, 如边际极大似然估计(Marginal Maximum Likelihood Estimation, MMLE) 和贝叶斯后验期望估计(Expected A Posteriori, EAP)等, 得到的个体能力的点估计汇总到组层面所得到的组能力估计是有偏的 (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992)。为了获得组水平能力的无偏估计, 可直接在极大似然函数中包含总体的参数和原始的反应(Mislevy, 1985), 采用多个值表示学生个体能力的可能分布, 即 PV (Plausible Value)值。PV 值基于学生对部分题目的反应作答以及其他相关信息, 可以看成是基于多重插补(multiple imputations)技术得到的一组特殊的值, 因此 PV 值并不是传统意义上的个体分数(Mislevy, 1993), 其详细原理可参考 Mislevy (1991) 与 Von Davier, Gonzalez 和 Mislevy (2009)的研究。

2.2 修正的 LR 法

在常规的基于观测分数的 DIF 检测方法中, LR 法对匹配变量的控制非常方便, 能够很好地与 IRT 所估能力值结合起来。并且, LR 方法具有能够同时检测一致性 DIF 和非一致性 DIF(Swaminathan & Rogers, 1990), 以及衡量 DIF 效应较为统一的标准(Jodoin & Gierl, 2001)等优点。因此, LR 方法可以作

[1] 在公式(1)中, $SE(\hat{b})$ 表示的是难度差异的标准误。

为一种较好的、被修正以用于对矩阵取样测验进行 DIF 检测的方法。

在 Logistic 方程中, 用 p 表示考生答对题目的概率, 用 θ 表示考生的能力值, G 表示组别, 用 LR 方法检测 DIF 时, 可以构建以下 3 个模型(Zumbo, 1999):

模型 1: $Logit(p)= b_0+b_1\theta$ (2)

模型 2: $Logit(p)= b_0+b_1\theta+b_2G$ (3)

模型 3: $Logit(p)= b_0+b_1\theta+b_2G+b_3\theta*G$ (4)

模型中, b_0 是截距, b_1 是能力值系数, b_2 是组别系数, b_3 是能力值与组别交互作用系数。当 b_2 不为 0 且 b_3 为 0 时, 则表示存在一致性 DIF; 当 b_3 不为 0 (不论 b_2 是否为 0)时表明存在非一致性 DIF。

模型 1 的假设是: 不存在 DIF, 只有考生的能力值对其在某道题目上面作答正确的概率有影响; 模型 2、模型 3 分别假设, 控制了考生能力后, 考生组别、组别与考生能力的交互作用对考生正确作答的概率有影响。DIF 检测的零假设是"不存在 DIF", 即: $b_2=b_3=0$, Swaminathan 和 Rogers (1990)指出其近似服从卡方分布, 可对其进行显著性检验。

LR 方法用于衡量 DIF 效应量较常用的指标是各模型间的 Nagelkerke 的 R^2 增量[2] (Nagelkerke, 1991), 记为 ΔR^2。其用于判断 DIF 效应的标准如下(Jodoin & Gierl, 2001):

• A 类题目—可以忽略的 DIF: $\Delta R^2< 0.035$;

• B 类题目—中等 DIF 效应: $0.035\leqslant\Delta R^2\leqslant 0.070$, 并且卡方检测显著;

• C 类题目—较大的 DIF 效应: $\Delta R^2> 0.070$, 并且卡方检测显著。

在常规的 LR 方法中, 考生的能力值通常是用测验总分进行衡量, 但是在矩阵取样设计测验中, 用基于 IRT 模型得到的考生能力值代替测验总分, 对考生能力的匹配更加准确。而且, 考虑到矩阵抽样设计下学生个体能力估计误差较大这一特点, 可将 PV 值以及 5 个 PV 值的均值作为能力匹配变量的结果。

基于上述对已有研究成果的分析和比较, 本研究拟对 LR 方法的匹配变量进行修正, 探查其在矩阵取样测验中进行 DIF 检测的有效性, 并对基于 Rasch 模型所使用的 IRT_Δb 法的有效性进行验证, 同时探讨对这些方法有效性可能产生影响的相关因素, 并尝试划定 IRT_Δb 法判断 DIF 效应时的分类标准。

3　模拟研究

3.1　实验设计

3.1.1　矩阵取样测验作答反应及 DIF 题目的生成

本研究基于 Rasch 模型, 采用 Matlab (R2009a)软件, 模拟生成符合部分平衡不完全组块设计(Partially Balanced Incomplete Block, pBIB)的"作答数据"。整个"测验"共有 120 道题目, 分为 8 个组块(A-H), 每个组块 15 道题目。然后, 按照 National Center for Education Statistics (NCES)官网提供的 NAEP 所使用的 pBIB 题册组合方式(National Center for Education Statistics, 2009), 由两个组块构成一个题本, 共形成 16 个题册(如图 1 所示)。这一矩阵取样设计方式很好地平衡了各组块出现的次数, 并能很好地把各题本链接起来。

通过改变目标组和参照组间的难度参数的差异产生测验中有 DIF 的题目, 其中, 所有具有 DIF 的题目均是目标组对应的难度值大于参照组, 即所有具有 DIF 的题目均有利于参照组。

3.1.2　实验条件

本研究共控制 4 个变量, 分别是 DIF 效应大小、题本中所含 DIF 题目的比率、考生群体间真实能力的差异和样本量, 4 个变量所设定的水平数及各水平对应的 DIF 题目数如表 1 所示。其中, DIF 效应的大小用于判断 DIF 检测方法是否能区分出不同 DIF 效应量的题目, 即判断 DIF 检测方法是否有效; 而题本中所含 DIF 题目的比率、考生群体间真实能力的差异和样本量三个因素则是以往研究中认为对 DIF 检测最有影响的因素(Finch, 2005; Roju, van der Linden, & Fleer, 1995)。

其中, DIF 题目的比率(高、低)、考生群体间真实能力差异(有、无)和样本量(小、中、大)相互结合共 12 种条件。在设计 DIF 效应大小时, 考虑在 DIF 题目比例低的情况下, DIF 效应的大小包含 3 个水平(0.3, 0.5 和 0.8), 而在 DIF 题目比例高的情况下, 为了更加深入考察高比例 DIF 题目下, 各因素对估计 DIF 效应大小的影响, 额外添加三种 DIF 效应水平 (0.4, 0.6 和 0.9)。最后, 每种条件均被重复 30 次。

[2] 公式为: $R^2=1-(\frac{L(0)}{L(\theta)})^{\frac{2}{n}}$, 其中 $L(0)$表示基线模型的似然值; $L(\theta)$表示估计模型的似然值; n 表示样本量。

组块 / 题册	A	B	C	D	E	F	G	H
题册 1	■			■				
题册 2	■	■						
题册 3		■	■					
题册 4			■	■				
题册 5		■		■				
题册 6	■		■					
题册 7					■			■
题册 8					■	■		
题册 9						■	■	
题册 10							■	■
题册 11					■		■	
题册 12						■		■
题册 13			■				■	
题册 14				■				■
题册 15		■			■			
题册 16	■					■		

图 1 pBIB 题册设计

表 1 研究中所控制 4 个变量各水平对应的 DIF 题目数

考生群体间真实能力差异 [b] (D)	样本量 [c] (S)	DIF 题目比例 (P) 及真实 DIF 值大小 [a]								
		低比例 (P1=7.5%)			高比例 (P2=15%)					
		0.3	0.5	0.8	0.3	0.4	0.5	0.6	0.8	0.9
D1: 0	S1: 250	3	3	3	3	3	3	3	3	3
	S2: 500	3	3	3	3	3	3	3	3	3
	S3: 1000	3	3	3	3	3	3	3	3	3
D2: -1	S1: 250	3	3	3	3	3	3	3	3	3
	S2: 500	3	3	3	3	3	3	3	3	3
	S3: 1000	3	3	3	3	3	3	3	3	3

注：a：DIF 效应为题目难度值在目标组和参照组间的差异;

b：考生群体间真实能力差异为 0 时, 目标组与参照组的能力分布均为 $N(0,1)$; 考生群体间真实能力差异为-1 时, 参照组的能力分布为 $N(0,1)$, 目标组为 $N(-1,1)$;

c：该样本量为每组每个题本对应的样本量, 对于所有题本总的样本量分别为 8000、16000 和 32000; 目标组和参照组样本量相同。

3.1.3 考生能力匹配变量的选取

在矩阵取样测验中, IRT 对考生能力的估计有多种不同的方法, 其中 EAP 方法和多重插补方法被最为广泛地接受(Von Davier et al., 2009; Wu, 2005)。在使用 LR 法时, 本研究将分别使用 EAP 和随机抽取 5 个 PV 值、5 个 PV 值均值 7 种变量作为匹配变量进行 DIF 检测。

3.1.4 DIF 效应的分析

本研究同时采用 IRT_Δb 法和修正 LR 法对以上条件中的题目进行 DIF 检测。其中, IRT_Δb 法采用 Conquest (v2.0)软件自带的 DIF 分析功能, 通过构建多面模型, 即把题目难度表示为题目、组别和题目与组别的交互作用这一模型, 这样所得的题目难度在组间的差异, 即为 DIF 值, 记为 Δb。对于修正 LR 法, 首先利用 Conquest (v2.0) 软件得到矩阵取样条件下考生能力的估计值, 然后通过 SPSS 17.0 构建 Logistic regression 回归模型进行检测。通过对回归系数进行显著性检验, 并考察 DIF 效应量增量 ΔR^2 两个指标对 DIF 检测的结果进行判断。

3.2　研究结果与分析

首先, 通过分别分析两种方法检测所得 DIF 效应量与真实 DIF 效应量之间的关系初步判断两种方法的有效性。

然后, 通过统计检验力(power)和犯第 I 类错误的概率(不存在 DIF 的题目被判为存在 DIF 的概率)两个指标对两种方法的有效性及相关影响因素进行分析。在进行检验力和 I 类错误分析时, IRT_Δb 法判断题目存在 DIF 的标准是 Conquest 所得题目难度值的差异显著不为 0, LR 法判断题目存在 DIF 的标准是组别系数显著不为 0, 显著性水平均取 0.01。

最后, 通过将 IRT_Δb 法和修正 LR 法所得 DIF 效应值之间建立函数关系, 将已有的 LR 法分类标准带入函数, 计算相应的 IRT_Δb 法分类标准, 并比较两者分类的结果的一致性。

3.2.1　有效性分析

(1) IRT_Δb 法的有效性验证

本研究发现, IRT_Δb 法所得 DIF 效应值的均值总是比真实的 DIF 效应值小, 如当两组难度值的真实差异为 0.3 时, IRT_Δb 法所得的难度值差异的均值为 0.238。但是, IRT_Δb 法所得 DIF 效应值的均值与真实的 DIF 效应值基本呈线性关系, 两者相关的 R^2 值达到 0.9949, 即, IRT_Δb 法估计所得的 DIF 效应值随着真实 DIF 效应值的增大而增大, 对于不同真实 DIF 效应值的题目, IRT_Δb 法估计所得的 DIF 效应值也存在差异, 说明 IRT_Δb 法能很好地区分 DIF 效应值大小不同的题目, 即检测结果有效。

各种条件下 IRT_Δb 法的统计检验力和犯 I 类错误的概率如表 2 和表 3 所示。

从表 2、表 3 可得, 真实 DIF 效应量大小、DIF 题目比率、考生群体间真实能力差异和样本量对 IRT_Δb 法的检验力有影响。随着真实 DIF 效应量的增大, 检验力增强, 对于中等及更高的 DIF 效应量(真实 DIF 值大于等于 0.5), 检验力在各种条件下均达到或接近 100%; 在较低的真实 DIF 水平下(真实 DIF 值小于 0.5), 高 DIF 题目比率时的检验力低于低 DIF 题目比率时的检验力; 考生群体间真实能力有差异时的检验力与没有差异时的检验力基本相当; 样本量越大, 检验力也越高。DIF 题目比率、考生群体间真实能力差异和样本量对 IRT_Δb 法检验力的影响存在交互作用。具体表现在：在低样本量情况下, DIF 题目比率对检验力有较大影响, 样本量较大时, DIF 题目比率和考生群体间真实能力对检验力的影响很弱; 在低样本量及高 DIF 题目比率时, 考生群体间的真实能力差异对检验力有较大影响, 低 DIF 题目比率时, 考生群体间真实能力差异对检验力的影响较小。

另外, 从表 2、表 3 可得, DIF 题目比率、考生

表 2　低 DIF 题目比率时 IRT_Δb 法各条件对应的检验力和 I 类错误

考生群体间真实能力差异(D)	样本量(S)	真实 DIF 值大小			I 类错误
		0.3	0.5	0.8	
D1: 0	S1: 250	0.73	1.00	1.00	0.09
	S2: 500	0.96	1.00	1.00	0.12
	S3: 1000	1.00	1.00	1.00	0.16
D2: -1	S1: 250	0.79	1.00	1.00	0.11
	S2: 500	0.94	1.00	1.00	0.13
	S3: 1000	0.99	1.00	1.00	0.18

表 3　高 DIF 题目比率时 IRT_Δb 法各条件对应的检验力和 I 类错误

考生群体间真实能力差异(D)	样本量(S)	真实 DIF 值大小						I 类错误
		0.3	0.4	0.5	0.6	0.8	0.9	
D1: 0	S1: 250	0.59	0.93	0.98	1.00	1.00	1.00	0.17
	S2: 500	0.84	0.98	1.00	1.00	1.00	1.00	0.25
	S3: 1000	0.99	1.00	1.00	1.00	1.00	1.00	0.43
D2: -1	S1: 250	0.72	0.84	1.00	1.00	1.00	1.00	0.18
	S2: 500	0.91	0.99	1.00	1.00	1.00	1.00	0.28
	S3: 1000	0.97	1.00	1.00	1.00	1.00	1.00	0.44

群体间真实能力差异和样本量对 IRT_Δb 法的 I 类错误有影响。高 DIF 题目比率时 I 类错误率高于低 DIF 题目比率时的 I 类错误率; 考生群体间真实能力有差异时 I 类错误率高于没有差异时的 I 类错误率; 样本量越大, I 类错误率越高。其中, DIF 题目比率对 IRT_Δb 法的 I 类错误的影响较大, 样本量次之, 考生群体间真实能力差异的影响较小。

(2) 修正 LR 法的有效性分析

如前所述, 在 LR 法中使用不同匹配变量策略应得到 7 个衡量 DIF 效应的 ΔR^2 值: EAP 的 ΔR^2 值、5 个 PV 值的 ΔR^2 值、5 个 PV 值均值的 ΔR^2, 最后还考察了 5 个 PV 值分别作为匹配变量所得的 ΔR^2 值的均值作为参照。

总体来说, 各种匹配方式下检测所得的 ΔR^2 均随着真实 DIF 效应值的增大而增大(见图 2), 这说明在匹配 IRT 所得的考生能力值后, LR 法所得衡量 DIF 的效应值能区分不同 DIF 效应的题目。其中对于有 DIF 的题目, 30 次模拟研究中, 在所有真实 DIF 效应值水平下, 匹配 EAP 所得能力值估计的 ΔR^2 的均值 (见图 2 的 R2theta 值) 和匹配 5 个 PV 值的均值得到的 ΔR^2 的均值(见图 2 的 R2pvm 值)比单独匹配 5 个 PV 值所得 ΔR^2 的均值(见图 2 的 R2pv1~R2pv5 及 R2mpv)高, 并且前两种匹配方式所得的 ΔR^2 值的标准差比分别匹配 5 个 PV 值所得的 ΔR^2 值的标准差小。对于没有 DIF 的题目, 匹配 EAP 所得能力值检测得到的 ΔR^2 的均值和匹配 5 个

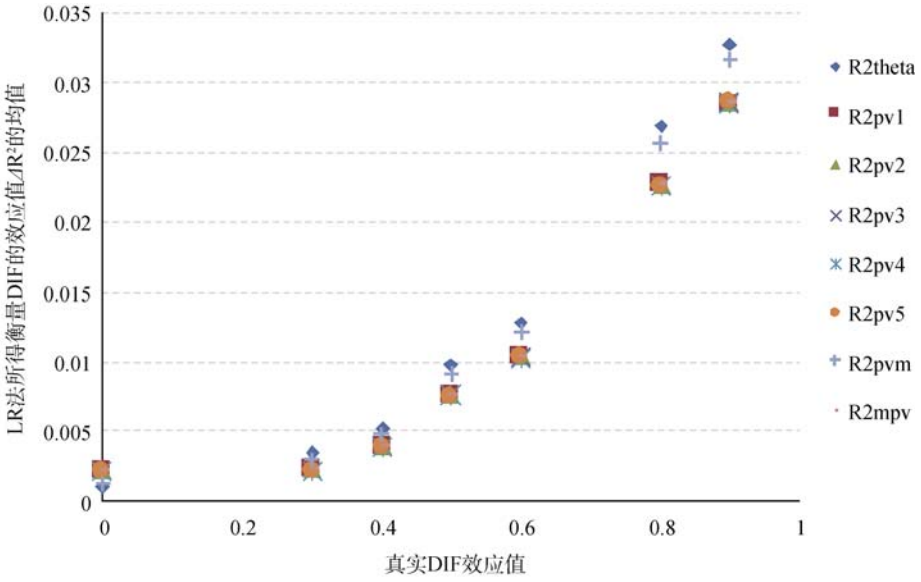

图 2 不同匹配变量下 LR 法所得衡量 DIF 的效应值 ΔR^2 的均值

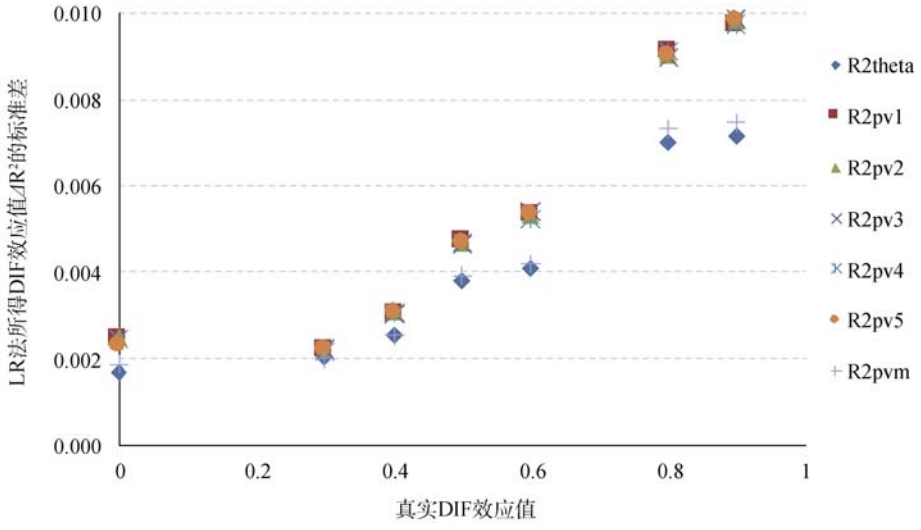

图 3 不同匹配变量下 LR 法所得衡量 DIF 的效应值 $\Delta R2$ 的标准差

PV 值的均值所得到的 ΔR^2 的均值更接近 0, 单独匹配 5 个 PV 值所得的 ΔR^2 的均值与 0 差异最大。

综合以上分析, 可以认为, LR 法在匹配 EAP 能力值和匹配 5 个 PV 值均值时, 其进行 DIF 检测更加有效和可靠。为了探索不同条件对 LR 法在匹配了 IRT 所得的能力值后进行 DIF 检测的影响, 本研究以匹配 EAP 所得能力值和匹配 5 个 PV 值均值进行 DIF 检测所得的 ΔR^2 值作为因变量进行进一步探索。

LR 法在匹配 EAP 所得能力值和匹配 5 个 PV 值均值(PVM)时各条件下的统计检验力和犯第 I 类错误的概率如表 4、表 5 所示。

从表 4、表 5 可得, 真实 DIF 效应量大小、DIF 题目比率、考生群体间真实能力差异和样本量对匹配 EAP 能力值后 LR 法的检验力有影响。随着真实 DIF 效应量的增大, 检验力有较大的增强, 对于中等及更高的 DIF 效应量(真实 DIF 值大于等于 0.5)时, 匹配 EAP 能力值后 LR 法的检验力在各种条件下均较高(除当考生群体间存在能力差异、样本量为 250 且真实 DIF 值为 0.5 时, 检验力为 0.78 外, 其他条件下均达到或接近 100%)。在较低的真实 DIF 水平下, 高 DIF 题目比率时的检验力低于低 DIF 题目比率时的检验力; 考生群体间真实能力有差异时的检验力总体略低于没有差异时的检验力; 样本量越大, 检验力越高。DIF 题目比率、考生群体间真实能力差异和样本量对匹配 EAP 能力值后 LR 法检验力的影响存在交互作用：在中等样本量情况下, DIF 题目比率对检验力有较大影响, 样本量较大时, DIF 题目比率和考生群体间真实能力对检验力的影响很弱; 在低样本量或中等样本量及高 DIF 题目比

表 4　低 DIF 题目比率时 LR 法各条件下的检验力和 I 类错误

能力匹配	考生群体间真实能力差异(D)	样本量(S)	真实 DIF 值大小			I 类错误
			0.3	0.5	0.8	
EAP	D1: 0	S1: 250	0.43	0.98	1.00	0.02
		S2: 500	0.83	1.00	1.00	0.03
		S3: 1000	0.99	1.00	1.00	0.04
	D2: -1	S1: 250	0.48	0.90	1.00	0.01
		S2: 500	0.82	1.00	1.00	0.02
		S3: 1000	0.99	1.00	1.00	0.04
PVM	D1: 0	S1: 250	0.43	0.99	1.00	0.02
		S2: 500	0.84	1.00	1.00	0.02
		S3: 1000	0.99	1.00	1.00	0.04
	D2: -1	S1: 250	0.34	0.88	1.00	0.03
		S2: 500	0.68	0.99	1.00	0.04
		S3: 1000	0.97	1.00	1.00	0.11

表 5　高 DIF 题目比率时 LR 法各条件下的检验力和 I 类错误

能力匹配	考生群体间能力差异(D)	样本量(S)	真实 DIF 值大小						I 类错误
			0.3	0.4	0.5	0.6	0.8	0.9	
EAP	D1: 0	S1: 250	0.43	0.78	0.91	0.97	1.00	1.00	0.04
		S2: 500	0.68	0.97	0.99	1.00	1.00	1.00	0.09
		S3: 1000	0.99	1.00	1.00	1.00	1.00	1.00	0.20
	D2: -1	S1: 250	0.33	0.61	0.78	0.99	1.00	1.00	0.04
		S2: 500	0.62	0.90	0.98	1.00	1.00	1.00	0.07
		S3: 1000	0.96	0.99	1.00	1.00	1.00	1.00	0.16
PVM	D1: 0	S1: 250	0.42	0.78	0.91	0.97	1.00	1.00	0.04
		S2: 500	0.71	0.97	0.99	1.00	1.00	1.00	0.08
		S3: 1000	0.99	1.00	1.00	1.00	1.00	1.00	0.18
	D2: -1	S1: 250	0.27	0.51	0.70	0.97	1.00	1.00	0.07
		S2: 500	0.44	0.83	0.93	1.00	1.00	1.00	0.15
		S3: 1000	0.91	0.98	1.00	1.00	1.00	1.00	0.32

率时, 考生群体间的真实能力差异对检验力有较大影响, 在低 DIF 题目比率时, 考生群体间真实能力差异对检验力的影响较小。同时, DIF 题目比率、样本量对匹配 EAP 能力值后 LR 法的 I 类错误率影响较大, 其中 DIF 题目比率对 I 类错误的影响最大, 样本量的影响次之; 考生群体间的真实能力差异对 I 类错误的影响很小。

从表 4、表 5 可得, 匹配 PVM 后 LR 法的检验力也受真实 DIF 效应量大小、DIF 题目比率、考生群体间真实能力差异和样本量的影响。随着真实 DIF 效应量的增大, 检验力有较大的增强, 对于中等及更高 DIF 效应量(真实 DIF 值大于等于 0.5)时, 匹配 PVM 后 LR 法的检验力在各种条件下均较高(除当考生群体间存在能力差异、样本量为 250 且真实 DIF 值为 0.5 时, 检验力为 0.70 外, 其他条件下均达到或接近 100%)。在较低的真实 DIF 水平下, 高 DIF 题目比率时的检验力低于低 DIF 题目比率时的检验力; 考生群体间真实能力有差异时的检验力明显低于没有差异时的检验力; 样本量越大, 检验力越高; 其中, 样本量的影响较大, 在样本量大的情况下, DIF 题目比率、考生群体间真实能力差异的影响作用很小。同时, DIF 题目比率、样本量对匹配 PVM 后 LR 法的 I 类错误影响较大, 其中 DIF 题目比率对 I 类错误的影响最大, 样本量、考生群体间的真实能力差异对 I 类错误的影响次之。

对比修正的 LR 法在匹配 EAP 能力值和 PVM 的结果可得, 匹配 EAP 能力值后 LR 法的 DIF 检测结果高于匹配 PVM 后所得的结果。在考生群体间存在能力差异时, 匹配 EAP 能力值后 LR 法的检验力略高于匹配 PVM 后的检验力, 其 I 类错误略低于匹配 PVM 后的 I 类错误; 在考生群体间不存在能力差异时, LR 法匹配 EAP 能力值和匹配 PVM 后的检验力和 I 类错误相当。对比 IRT_Δb 法与 LR 法所得的检测结果可得, IRT_Δb 法的检验力和 I 类错误总体均高于 LR 法, 考生群体间真实能力差异对 IRT_Δb 法和 LR 法的影响效果相反。

3.2.2 分类标准分析

在探讨两种方法所得的 DIF 效应值之间的关系时, 本研究选用不同真实 DIF 效应对应的所有不同条件下两种方法所得的衡量 DIF 效应值的均值进行探索。图 4 显示了匹配 EAP 能力值后 LR 法所得到的 ΔR^2 与 IRT_Δb 法所得的 Δb 之间的关系, 两者之间的关系可以近似拟合为一条幂函数, 拟合后其变异解释率达到 98.41%。根据这个拟合函数, 带入 LR 法对 DIF 题目进行分类的标准的临界值(Jodoin & Gierl, 2001), 经过计算, 得到 DIF 分类标准的两个分界点对应的 Δb 值分别为: 0.8532 和 1.2379。图 5 显示了匹配 PV 值均值后 LR 法所得到的 ΔR^2 与 IRT_Δb 法所得的 Δb 之间的关系, 两者之间的关系也近似拟合为一条幂函数, 其拟合后解释率达到 96.40%。根据这个拟合函数, 带入 LR 法对 DIF 题目进行分类的标准的临界值, 经过计算, 得到 DIF 分类标准的两个分界点对应的 Δb 值分别为: 0.8562 和 1.2156。由于在实际应用中, DIF 分类存在一定的模糊性。因此, 结合两种不同匹配方式所得的 DIF 分类标准分界点的 Δb 值, 可以把 IRT_Δb 法分类标准的两个分界点的 Δb 值确定为 0.85 和 1.23。

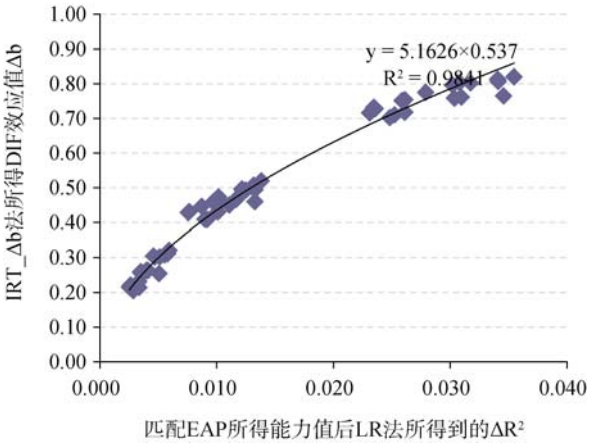

图 4 匹配 EAP 所得能力值后 LR 法所得到的 ΔR^2 与 IRT_Δb 法所得的 Δb 之间的关系

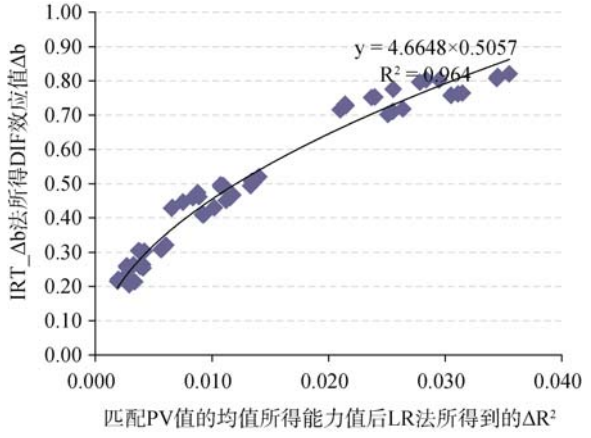

图 5 匹配 PV 值均值后 LR 法所得到的 ΔR^2 与 IRT_Δb 法所得的 Δb 之间的关系

根据 LR 法的 DIF 分类标准及本研究所确定的 IRT_Δb 法的分类标准, 可以对本研究所模拟的

DIF 题目的检测结果进行分类, 其分类结果如表 6、表 7 所示。在结果呈现时, 还选取了真实 DIF 效应值为 0 的题目作为参照。在具体挑选题目时, 选择每种条件下(共 12 种条件)估计的 DIF 效应值最大的前三道题。

从表 6、表 7 可得, 对于中等和低 DIF 效应量(真实 DIF 效应量小于 0.8)的题目, 三种方法均检测为 A 类 DIF; 对于高 DIF 效应量题目(真实 DIF 效应量大于等于 0.8)的题目, 大部分题目被检测为 A 类, 少部分 B 类, 没有 C 类。

随着样本量的增加, 三种方法检测出的 B 类 DIF 的比例均减少。

在低 DIF 题目比例条件下, 三种方法检测所得的 B 类 DIF 均比高 DIF 题目比例条件下所得 B 类 DIF 题目多。

考生群体间真实能力无差异时, LR 法检测出的 B 类题目的比例高于 IRT_Δb 法所得 B 类题目比例, 而考生群体真实能力有差异时则相反; 而且, 考生群体间真实能力无差异时, LR 法检测所得的 B 类题目的比例高于考生群体间真实能力有差异所得的检测结果。

LR 法在匹配 EAP 所得能力值和匹配 PVM 时所得的分类结果较为相近, 但是, 在考生群体间真实能力无差异时, 匹配 EAP 所得能力值的 LR 法所得的 B 类题目比例大体上略低于匹配 PVM 时所得结果, 而在考生群体间真实能力存在差异时, 则相反。

4　实证研究

在大型国际教育测量与评价项目中, 由于各国经济、文化和教育等的差别, 某些测验题目在不同国家间会出现难度差异, 说明这些题目在国家间可能存在项目功能差异。结合模拟研究中的已有分析, 进一步探讨在实证的矩阵取样测验中, IRT_Δb 法和修正 LR 法的检测结果是否具有较高的一致性。

表 6　低 DIF 题目比率时三种方法 DIF 检测分类结果

真实能力差异(D)	样本量(S)	真实 DIF 效应值	LR 法匹配 EAP 所得 θ 值的结果			LR 法匹配 PV 值的均值的结果			IRT_Δb 法所得的结果		
			A 类	B 类	C 类	A 类	B 类	C 类	A 类	B 类	C 类
D1: 0	S1: 250	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	50.00	50.00	0	70.00	30.00	0	80.00	20.00	0
	S2: 500	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	74.44	25.56	0	72.22	27.78	0	88.89	11.11	0
	S3: 1000	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	85.56	14.44	0	84.44	15.56	0	94.44	5.56	0
D2-1	S1: 250	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	80.00	20.00	0	84.44	15.56	0	78.89	21.11	0
	S2: 500	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	95.56	4.44	0	96.67	3.33	0	86.67	13.33	0
	S3: 1000	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.8	94.44	5.56	0	96.67	3.33	0	92.22	7.78	0

表 7 高 DIF 题目比率时三种方法 DIF 检测分类结果

真实能力差异(D)	样本量(S)	真实DIF效应值	LR 法匹配 EAP 所得 θ 值的结果			LR 法匹配 PV 值的均值的结果			IRT_Δb 法所得的结果		
			A 类	B 类	C 类	A 类	B 类	C 类	A 类	B 类	C 类
D1: 0	S1: 250	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	81.11	18.89	0	80.00	20.00	0	82.22	17.78	0
		0.9	51.11	48.89	0	51.11	48.89	0	56.67	43.33	0
	S2: 500	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	93.33	6.67	0	92.22	7.78	0	95.56	4.44	0
		0.9	56.67	43.33	0	55.56	44.44	0	67.78	32.22	0
	S3: 1000	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	98.89	1.11	0	98.89	1.11	0	100	0	0
		0.9	62.22	37.78	0	56.67	43.33	0	74.44	25.56	0
D2: -1	S1: 250	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	91.11	8.89	0	93.33	6.67	0	85.56	14.44	0
		0.9	66.67	33.33	0	77.78	22.22	0	70.00	30.00	0
	S2: 500	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	95.56	4.44	0	98.89	1.11	0	91.11	8.89	0
		0.9	81.11	18.89	0	90.00	10.00	0	74.44	25.56	0
	S3: 1000	0	100	0	0	100	0	0	100	0	0
		0.3	100	0	0	100	0	0	100	0	0
		0.4	100	0	0	100	0	0	100	0	0
		0.5	100	0	0	100	0	0	100	0	0
		0.6	100	0	0	100	0	0	100	0	0
		0.8	100	0	0	100	0	0	97.78	2.22	0
		0.9	85.56	14.44	0	93.33	6.67	0	80.00	20.00	0

4.1　数据与方法

本研究选用 PISA2003 年的测试数据, 对韩国和美国学生在数学测验上的作答数据进行 DIF 分析。PISA2003 年的学业测验以数学学业测验为主, 共 84 道题目, 被分配到 13 个题册中, 由于测验中还包含科学、阅读等学业相关题目, 所以, 各题册中包含的数学测验题目并不相等, 题册间题目采用的是不完全平衡组块的矩阵取样设计。由于有一道数学题目只在美国学生中施测, 因此, 实际分析的题目为 83 道。PISA2003 年参与施测的学生中, 韩国学生共 5444 名, 其中, 男生 3211 名, 女生 2233 名; 美国学生共 5455 名, 其中, 男生 2740 名, 女生 2715 名(Organization for Economic Co-operation and Development, 2005)。

数据分析基于模拟研究结论展开：采用 EAP 和 5 个 PV 值的均值作为两种考生能力的匹配变量使用 LR 法, 以及 IRT_Δb 法对真实测验数据进行 DIF 检测。按照前文分类标准对 DIF 效应进行划分、对比。

4.2　研究结果

根据 IRT_Δb 法的 Δb 大于 0.85 这一标准, 所检测的题目共有 11 道存在 DIF, 根据 LR 法的 ΔR^2 大于 0.035 这一标准, 修正 LR 法共检测到 8 道有 DIF 的题目, 结果如表 8 所示。

如表 8 所示, 匹配 EAP 所得能力值与匹配 PV 值的均值进行 DIF 检测所得的结果基本完全一致。IRT_Δb 法检测得到的 11 道有 DIF 题目包含了修正 LR 法所得的结果, 从所得的 DIF 效应量看, 除了 M800Q01 和 M421Q03 两道题目外, IRT_Δb 法和修正 LR 法所得的 DIF 效应量的大小关系基本一一对应, 即 Δb 值越大, 则 ΔR^2 值越大; 从 DIF 分类结果看, 两者有 6 道题目的分类结果一致, 对于 M800Q01 和 M421Q03 两道题目, 其按照 LR 法的分类标准被分为 B 类, 但是其 DIF 效应量较为接近 LR 法达到 C 类的临界值 0.07。M421Q02T、M144Q01T 和 M150Q01 三道题目按照两种方法所得的分类结果也存在差异, 但是, 从两者的 DIF 效应量来看, 这三道题目的 DIF 效应量均处于 A 类和 B 类分界的临界值附近。因此, 总体来说, 修正 LR 法与 IRT_Δb 法所得的结果具有较高的一致性。

5　研究讨论

本研究利用 Logistic 回归方法易控制影响变量的特点, 将 IRT 模型经过等值后所得的考生能力估计值作为衡量考生能力的匹配变量以进行 DIF 检测, 结果显示, 修正 LR 法能有效、便捷地对矩阵取样测验的题目进行 DIF 检验。这说明, 对于矩阵取样测验, 当选择基于 IRT 模型所得的能较准确地衡量考生能力的估计值作为匹配变量后, 对其进行 DIF 检测与对普通单题本测验进行 DIF 检测的过程一样。这提示我们, 可以通过这一思路对其它常规 DIF 检测方法加以修正。因此, 对于其它几种在实践中应用比较广泛的 DIF 检测方法, 如 MH 法和 SIBTEST 法, 在运用于矩阵取样测验时, 今后若能研发出能自由选择匹配变量的相应软件, 就可以提高 MH 法和 SIBTEST 法在矩阵取样测验中进行 DIF 检测的适用性。

虽然研究中 IRT_Δb 法和 LR 法得到的项目参数和能力参数均采用 Conquest 2.0 估计得到, 但是由于两种 DIF 检测方法原理上的差异(IRT_Δb 法的

表 8　PISA2003 数学测验在韩国与美国考生群体间 DIF 检测结果

题目	IRT_Δb 法		匹配 EAP 所得能力值		匹配 PV 值的均值		有利于的组别
	Δb	类别	一致性 DIFΔR^2	类别	一致性 DIFΔR^2	类别	
M420Q01T	2.608	C	0.225	C	0.221	C	美国
M505Q01	−2.300	C	0.196	C	0.199	C	韩国
M828Q03	1.814	C	0.131	C	0.128	C	美国
M800Q01	−1.702	C	0.063	B	0.064	B	韩国
M421Q03	−1.418	C	0.064	B	0.067	B	韩国
M124Q03T	1.412	C	0.122	C	0.120	C	美国
M408Q01T	1.186	B	0.050	B	0.048	B	美国
M179Q01T	1.114	B	0.059	B	0.058	B	美国
M421Q02T	0.918	B	0.034	A	0.032	A	美国
M144Q01T	−0.886	B	0.027	A	0.029	A	韩国
M150Q01	−0.876	B	0.029	A	0.030	A	韩国

结果更加依赖群体能力的估计, 而 LR 法需要在不同群组个体能力参数匹配后对项目 DIF 进行检测), 因而两种方法之间的差异也有可能是由于不同层面(群体和个体)能力估计精度的差异而导致的。矩阵取样设计一般用于大规模的抽样设计, 往往更加关注群体能力的估计而非个体能力估计的精度, 这也是为什么在类似 PISA 这样的大规模矩阵抽样设计下, 采用 IRT_Δb 法的原因(Organization for Economic Co-operation and Development, 2005, 2009)。同时, 本研究的结果表明, 合理匹配被试能力估计的 LR 方法也可以得到有效的 DIF 检验结果, 但是对实际应用者来讲, 采用 LR 方法应该注意其局限性：在矩阵抽样设计中, 由于每个考生只完成了测验中的部分题目, 因此传统估计方法对考生个体能力估计的精度较差, 这时, 矩阵取样设计具体的题目组合方式、题册生成方式对能力估计的精度会有不同程度影响(Mislevy, Beaton, et al., 1992; Mislevy, Johnson, et al., 1992)。本研究模拟数据时的矩阵抽样设计中每个学生完成的题目个数相对较多(30 题), 此时合理匹配被试能力估计后, LR 法同样能进行有效的 DIF 检验。但是, 对于每个学生完成测验题目更少的情况, LR 方法的检验效果有待进一步考察。

对于影响因素, IRT_Δb 法所得的 Δb 值和 LR 法的检验力、I 类错误和分类结果均明显受样本量的影响。对于检验力和 I 类错误, 显著性很显然会受到样本量的严重影响, 而分类结果随着样本量的增加倾向于判为更多的 A 类 DIF, 这说明随着样本量的增加, 所检测到的题目的效应量的值变小, 这一结果则与显著性判断的结果相反。显著性判断对低 DIF 效应量的题目有鉴别性, 而分类标准所得结果仅对高 DIF 题目有鉴别性, 这也说明在实践中需要将两者结合对 DIF 题目进行进一步判断。

题本中 DIF 题目比率和考生群体间真实能力的差异对两种方法的检验力、I 类错误和分类标准也存在一定影响, 并存在一定的交互作用。且考生群体间真实能力的差异对 LR 法和 IRT_Δb 法的影响效果正好相反, 如果为了获得更高的检测率, 在考生群体间真实能力存在差异时, IRT_Δb 法较佳。

在实际应用中, DIF 的分类标准具有很好的指导意义, 本研究通过 IRT_Δb 法所得的 Δb 值和 LR 法所得的 ΔR^2 值, 初步探索出 IRT_Δb 法所对应的 DIF 分类标准的临界值。从模拟和实证数据的 DIF 题目的分类结果来看, 根据本研究所得的 IRT_Δb 法的分类标准对测验题目进行 DIF 分类的结果与 LR 法的分类结果基本一致, 这在一定程度上说明这一分类标准的有效性。但是, Hidalgo 和 López-Pina (2004)对 LR 法和 MH 法进行 DIF 检测所得 DIF 分类结果的比较显示, LR 法得到的 A 类题目的数目明显多于 MH 法得到的 A 类题目, 而 LR 法所得的 C 类题目的数目明显少于 MH 法所得的 C 类题目。这说明相比于 MH 法, LR 法现有的分类标准依然太宽松, 导致很多 MH 法检测为严重或中等 DIF 的题目被 LR 法标记为中等或可以忽略。因此, 鉴于实践应用中通常把 MH 法的分类标准作为“黄金标准”, 当其它 DIF 检测方法在保证能得到稳定有效的衡量 DIF 大小的效应值时, 有必要对其所得的效应值与 MH 法的 ΔMH 值之间的关系进行进一步的研究, 从而将各种方法对 DIF 检测结果的分类标准加以统一。只有这样, 由不同种方法所得的 DIF 检测结果才具有可比性, 同时对实践应用者才有参考价值。同时, 在实际应用中还应该注意到由 LR 方法所得到的 ΔR^2 值本身的一些局限性, 在 DIF 检测中 ΔR^2 值的取值范围往往较小, 看起来差异不大的 ΔR^2 值在 DIF 效应大小的描述上却有可能表现出比较大的差异。在 LR 方法下如何选用更加科学合理的指标区分 DIF 效应的大小, 是使用该方法值得进一步探讨的问题。

参 考 文 献

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*(2), 231–251.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report (NCES, 2001–509)*. Washington, DC: National Center for Educational Statistics.

Camilli, G. (2006). Test Fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 221–252). New York: Praeger.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Dings, J., Childs, R., & Kingston, N. (2002). *The Effects of Matrix Sampling on Student Score Comparability in Constructed-Response and Multiple-Choice Assessments*. Washington, DC: Council of Chief State School Officers.

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278–295.

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item

functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903–915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

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

Li, L. Y., Xin, T., & Dong, Q. (2007). Application of matrix-sampling methods on large-scale assessment. *Journal of Beijing Normal University (Social Sciences)*, (6), 19–25.

[李凌艳, 辛涛, 董奇. (2007). 矩阵取样技术在大尺度教育测评中的运用. 北京师范大学学报(社会科学版), (6), 19–25.]

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58*(1), 79–85.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131–154.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692.

National Center for Education Statistics (NCES). (2009). Example of a Partially Balanced Incomplete Block (pBIB) Design. Retrieved March 31, 2013 from http://nces.ed.gov/nationsreportcard/tdw/instruments/booklet_design_pbib.asp

Organization for Economic Co-operation and Development (OECD). (2005). *PISA 2003 Technical report*. Paris: OECD.

Organization for Economic Co-operation and Development (OECD). (2009). *PISA 2006 Technical report*. Paris: OECD.

Roju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353–368.

Shealy, R., & Stout, W. (1993). 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. *Psychometrika, 58*(2), 159–194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–114). Hillsdale, NJ: Lawrence Erlbaum.

Von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114–128.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER ConQuest. Version 2.0. Generalised Item Response Modelling Software [Computer software]*. Camberwell: ACER Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Applying IRT_ΔB Procedure and Adapted LR Procedure to Detect DIF in Tests with Matrix Sampling

ZHANG Xun[1]; LI Lingyan[1]; LIU Hongyun[2]; SUN Yan[1]

([1] *National Key Laboratory of Cognitive Neuroscience and Learning*, *Beijing Normal University*, *Beijing* 100875, *China*)

([2] *School of Psychology*, *Beijing Normal University*, *Beijing* 100875, *China*)

Abstract

Matrix sampling is a useful technique widely used in large-scale educational assessments. In an assessment with matrix sampling design, each examinee takes one of the multiple booklets with partial items. A critical

problem of detecting differential item functioning (DIF) in such scenario has gained a lot of attention in recent years, which is, it is not appropriate to take the observed total score obtained from individual booklet as the matching variable in detecting the DIF. Therefore, the traditional detecting methods, such as Mantel-Haenszel (MH), SIBTEST, as well as Logistic Regression (LR) are not suitable. IRT_Δb might be an alternative due to its abilities to provide valid matching variable. However, the DIF classification criterion of IRT_Δb was not well established yet. Thus, the purpose of this study were: 1) to investigate the efficiency and robustness of using ability parameters obtained from Item Response Theory (IRT) model as the matching variable, comparing with the way using traditional observed raw total scores ; 2) to further identify what factors will influence the abilities in detecting DIF of two methods; 3) to propose a DIF classification criteria for IRT_Δb.

Simulated and empirical data were both employed in this study to explore the robustness and the efficiency of the two prevailing DIF detecting methods, which were the IRT_Δb method and the adapted LR method with the estimation of group-level ability based on IRT model as the matching variable. In the Monte Carlo study, a matrix sampling test was generated, and various experimental conditions were simulated as follows: 1) different proportions of DIF items; 2) different actual examinee ability distributions; 3) different sample sizes; 4) different size of DIF. Two DIF detection methods were then applied and results were compared. In addition, power functions were established in order to derive DIF classification rule for IRT_Δb based on current rules for LR. In the empirical study, through conducting a DIF analysis for American and Korean mathematics tests from Programme for International Student Assessment (PISA) 2003, the consistency of the classification rules between IRT_Δb and LR were further examined.

The results indicated that in the matrix sampling design, both IRT_Δb method and adjusted LR method were sensitive to the diverse DIF magnitude. It was also found that the power, type I error, and the final classification of both methods were also influenced by the sample size, percentage of items with DIF, and ability differences between the focused group and the reference group.

In conclusion, it was found that both the IRT_Δb method and adjusted LR method can be used to detect DIF in matrix sampling tests. A classification rule for IRT_Δb was proposed, which are: 0.85 between negligible DIF(A) and intermediate DIF(B), 1.23 between intermediate DIF(B) and large DIF(C). Meanwhile, it was suggested that researchers would take this rule as a tentative principle since the ΔR^2 was limited between a narrow interval and the classification rule of LR was very flexible compared to classification rule of MH. Further studies could be conducted to take MH, IRT_Δb as well as LR into consideration simultaneously to give more comparable and consistent classification rules for different methods.

Key words Differential Item Functioning; Matrix Sampling; Rasch model; Logistic regression