经济法试题 DIF 的参数法检测研究*

李 力[1,2],戴海崎[2],董圣鸿[2],欧冬明[3]
(1. 南昌大学 教育学院,南昌 330029;2. 江西师范大学 心理学系,南昌 330027;3. 人事部考试中心,北京 100080)

摘 要:该研究基于项目反应理论的 Samejima 等级反应模型(GRM),在 MULTILOG 软件支持下,应用参数检测方法,对某年度全国性资格考试的某科目试卷中经济法部分的 21 个项目做了 DIF 检测分析。结果如下:存在性别 DIF 的项目一个,存在民族 DIF 的项目四个,存在工作性质 DIF 的项目一个。其中项目 68 在民族层面上表现为一致性 DIF,项目 64 既存在民族 DIF 又存在工作性质 DIF。通过对项目统计量、反应曲线的分析和专家的讨论,文章最后还分析了产生这些 DIF 的几个可能的原因。

关键词:项目功能差异;等级反应模型;项目偏差;项目特征曲线
中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2007)04 - 0088 - 05

1 问题的提出

如果具有相同潜在能力水平的两个群体在同一个项目上有不同的作答表现,就说明该项目存在项目功能差异[1](Differential Item Functioning,DIF)。测验中存在项目功能差异,就可能导致测验偏差,导致测验的不公平。人们已经越来越认识到项目功能差异检测和研究的重要性。测验专家们希望通过对项目功能差异的检测和分析,能编制出更公平、公正的测验。在国外,有关 DIF 的研究已有较大的发展,特别是在美国,不仅理论研究发展很快,而且应用上也非常广泛。我国对 DIF 的研究起步较晚,研究的数量不多,程度也不深,对象主要限于二级记分题,方法限于经典测量领域,即用经典理论的方法进行探测(如:MH、STND、P - MH、SIBTEST 等)。对于多级记分题采用项目反应理论的参数方法作 DIF 检测分析的还未见报告。DIF 检测的参数方法就是在一个特定的项目反应理论模型下,对所估项目参数进行比较从而实现对 DIF 的研究[2]。

该研究基于项目反应理论的 Samejima 等级反应模型,采用参数检测方法,对某年度全国性资格考试中某科目试卷的试题进行了项目功能差异的检测分析。

2 项目反应理论框架下多级记分试题 DIF 的检测方法

在项目反应理论框架下探测等级模型试题是否存在 DIF 的主要计算工具是 Thissen 的 MULTLOG(7.03)[3]。检验项目是否有 DIF 的统计原理是似然比大小的比较。似然比法 LR Test(GRM - LR test;Kim&Cohen,1998;Thissen et al,1986)就是通过检测两组等级项目的参数是否有差异来侦查测验项目是否存在 DIF[1]。通常检测步骤如下:

第一步:选取一组项目,以其中若干无 DIF 的项目作为对照项目,另一个项目作为研究项目(需检测是否存在 DIF 的项目)。

第二步:比较每个研究项目在两个模型下的相对适合度,即:部分等值模型和全等值模型间统计量的拟合度(McClelland,1989)。

在全等值模型中,确立一个被研究的项目,把其他项目暂时当作无 DIF 的项目,组成"铆题",然后进行极大似然参数估计,求出项目参数和似然函数值。之后对研究项目的参数进行限制,即设定该项目在目标和参照两组上的参数值相等,这就是部分等值模型的等值阶段。对该模型再进行极大似然估计,求出部分等值模型下的项目参数和似然函数值。

第三步:根据项目参数反应模型,算出该组项目每种反应类型的概率之积,即似然函数。然后计算两个模型下对数似然函数 - 2 倍和的差值:

$G^2 = -2[\log - likelihood\ function\ L\ from\ the\ compact\ model - log - likelihood\ function\ L\ from\ the\ augmented\ model]$

第四步:由于 G^2 近似服从 χ^2 分布,最后进行是否有 DIF 的假设检验。如果 G^2 值在 α 水平上超出了 χ^2 临界值,即拒绝不存在 DIF 的原假设[4]。

3 检测分析设计

3.1 检测分析材料

该研究采用某年度全国性资格考试中某科目试

* 基金项目:国家教育考试"十五"科研规划项目(2006023)。

卷作为研究材料。整份试卷一共有 105 个项目，均为选择题。该试卷从内容上分为经济法等六个不同的部分。每个部分占有不同的题量。该研究选取了 21 道经济法项目作为研究对象。具体项目分别为 57、58、59、60、61、62、63、64、65、66、67、68、69、70、99、100、101、102、103、104、105。其中前十四个项目为 0－1 记分题，每题 1 分，其中只有一个正确答案。后七个项目为多级记分题，每题 2 分，均有 2 个或 2 个以上选项符合题意，错选不得分，少选得 0.5 分。项目总分为 28 分。

3.2　检测分析样本抽取

该研究的数据资料由国家主管本项考试的考试机构提供。除了香港、澳门、台湾外，从全国 31 个省和直辖市 20 多万考生中随机抽取了 10000 人，其中包括总体中所有的 2107 位非汉族考生。从性别、民族、东－西部地区、企－事业机关单位四个方面，分别对该试卷做 DIF 分析。

3.3　测验的单维性检验

利用单维项目反应模型分析测量数据的前提是所测特质必须是单维的，这样才能满足 IRT 模型的局部独立的假设[5]。在该研究中，用因素分析的方法对 105 道试题进行因子分析，求得项目四项相关矩阵的最大特征值为 13.62，次大特征值为 2.347，两者之比远远大于 5，因此可以把该试卷所有项目归为同一维度范畴。

3.4　目标组和参照组的能力匹配

该研究分别把男性、汉族、东部地区、事业机关单位作为参照组，把女性、非汉族、西部地区、非事业机关单位作为目标组。项目功能差异表现为两个具有相同潜在能力水平的群体在项目上有不同的成绩，因此，还存在对所抽取的被试的一个能力匹配问题。做法是对经济法 21 个项目上的 10000 个被试进行了能力估计，把来自于总体的大样本容量目标组被试能力值视为正态分布。从能力分布（－3、3）正态分布区间，按能力高、低分数分为 12 个等级，以精确到小数点后三位有效数字，按能力值相等的原则选取与目标组被试相等的参照组被试数，匹配后人数均为 1000 人。

表 1　匹配分组被试在经济法 21 个项目的基本统计量

类　别	被试量	α 系数	平均分（答对率）	标准差	偏度（skew）
参照组（男性）	1000	0.763	15.902（0.567）	4.632	－0.039
目标组（女性）	1000	0.769	15.896（0.568）	4.653	－0.061
参照组（汉族）	1000	0.748	15.966（0.570）	4.328	0.056
目标组（非汉族）	1000	0.717	15.842（0.566）	4.275	0.080
参照组（东部地区）	1000	0.751	16.155（0.577）	4.288	－0.029
目标组（西部地区）	1000	0.766	15.962（0.542）	4.32	－0.08
参照组（事业机关）	1000	0.763	16.603（0.593）	4.397	－0.122
目标组（非事业机关）	1000	0.755	16.457（0.584）	4.081	－0.104

从表 1 中 21 个统计项目的基本统计量可以发现，目标组和参照组被试双方的各种统计量非常接近。其中男女性、东西部地区、事业和非事业六个组的偏态系数为负值，表明数据分布稍右偏，有一个小小的左尾，但偏态系数均远小于 1，结合被试的能力正态分组，说明该数据还是属于正态分布。同时八个组所检测的 α 系数值也均达到了 0.70 以上，说明所有项目间的一致性程度也较高。

3.5　检测分析步骤

在单维性和测量的不变性已经得到验证，相同能力的被试也已经匹配好之后，就可以进行 DIF 的检测分析了。整个的检测分析分三步完成。

第一步：DIF 的检测。分别从性别、民族、东－西部地区、企－事业机关单位四个匹配层面，编写应用 MULILOG 软件作参数估计的命令程序，首先分别估出 21 个项目在基线模型上的项目区分度和项目难度的参数值，而后在全项目等值模型和部分项目等值模型下，用 MULILOG 软件编写等值程序估出所有的参数值，比较 $G^2 = G_i^2 - G_{hl}^2$ 与 $\chi_a^2(df)$ 的大小。如 G^2 不超过规定值，便可接受项目 i 为等值模型的假设，得出关于项目 i 参数等值的结论，即不存在 DIF；反之则拒绝该假设，确定项目 i 存在 DIF。然后再在此基础上进行下一个项目的检验，直至全部项目检验结束。

第二步：DIF 特征曲线分析。对于存在 DIF 的项目，分别作目标组和参照组的项目特征曲线图。比较两曲线之间的差异，从而对 DIF 的具体表现做进一步的分析。

第三步：DIF 根源的探讨。会同经济专业专家针对项目内容、性质等从社会学、心理学角度进行分析，寻找产生 DIF 的具体原因。

4　DIF 的检测分析结果

4.1 DIF 的检测结果

关于“经济法”21 个项目按性别、民族、地区和单位性质匹配分组的 DIF 检测的结果，如表 2。

表 2 匹配分组进行的 DIF 检测结果

类别	项目	参数	基线模型		等值模型	部分等值模型		df 的变化	G^2 的变化
			参照组	目标组	参照(目标)组	参照组	目标组		
性别		a	0.73	0.90	0.81	0.72	0.92		
		b_1	0.25	-0.58	-0.35	-0.03	-0.57		
	102	b_2	0.29	-0.47	-0.27	0.02	-0.46	2	17.5
		b_3	1.15	0.41	0.61	0.89	0.41		
		b_4	3.08	1.97	2.34	2.85	1.93		
民族	64	a	0.87	0.80	0.83	0.88	0.75	2	14.7
		b	1.68	1.94	1.80	1.70	2.03		
	68	a	0.93	0.88	0.88	0.91	0.91	2	14.0
		b	-1.29	-1.52	-1.40	-1.27	-1.46		
		a	0.37	0.23	0.37	0.36	0.90		
	99	b_1	0.30	-0.72	-0.05	0.33	-0.30	2	20.3
		b_2	0.78	-0.57	0.27	0.82	-0.26		
		a	0.11	0.44	0.14	0.40	0.44		
	101	b_1	1.32	0.70	1.45	0.63	0.67	2	17.6
		b_2	2.97	1.10	2.70	1.08	1.07		
单位	64	a	0.86	0.68	0.76	0.85	0.71	2	37.5
		b	1.92	1.80	1.68	1.78	1.66		

注：$\chi^2_{0.001}(df=2)=13.815$

以上是运用 MULILOG 软件，在四个不同的层面上进行 DIF 检测后，存在 DIF 项目的结果汇总表。在性别分组检测中，项目 102 在两个模型上被检测的 G^2 值为 17.5，远大于 $\chi^2 0.001(df=2)=13.815$，差异性非常显著。男女双方在该项目上的区分度 a 和阈值 b 都不相等。该项目的 DIF 被检测为非一致性 DIF[6]，表现为男性区分度低于女性，但男性各个等级上的阈值都比女性高。

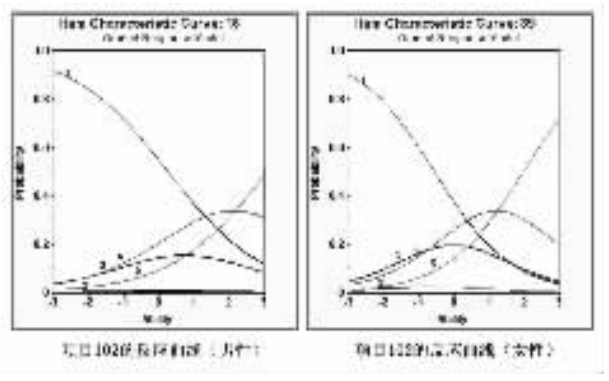
在民族这个层面上，共有 4 个项目的 G^2 值大于 13.815。其中项目 68 的区分度相等($a=0.91$)，阈值 b 不相等，为一致性 DIF[6]。其它 3 个项目区分度和阈值都不相等，为非一致性 DIF。汉族区分度在项目 64 上比非汉族要高，但阈值却比非汉族低；在项目 68、99、101 三个项目上汉族区分度和阈值都比非汉族要高。项目 99 和项目 101 都是多级记分题，但都只估出 0.5 和 1.0 两个等级上的阈值，其中项目 99 汉族区分度和阈值都比非汉族要高，而项目 101 区分度比非汉族低，阈值却比非汉族高。

地区层面表现为 21 个项目的 G^2 差值都比 13.815 要小，经济法所有的项目在地区这个层面上没有 DIF 出现。经济法 21 个项目在单位性质这个层面上，只有第 64 个项目被检测出有 DIF 并且 G^2 值非常大(37.5)，远远大于临界的卡方值 13.815。项目 64 是个二级记分题，事业机关区分度和阈值都比非事业机关要高，该项目 DIF 为非一致性 DIF。

4.2 DIF 的特征曲线分析

存在 DIF 的项目，其在项目特征曲线上表现为同一项目有不同的特征曲线。项目无 DIF，表示双方的项目特征曲线重叠、无差异，说明这些项目对于双方被试(不论特质高低)的刺激是等值的，不存在项目与性别、民族或工作性质之间的交互作用。为了能更好地对出现 DIF 的项目做进一步的探讨，下面来研究出现 DIF 项目的反应曲线图之间的差异。

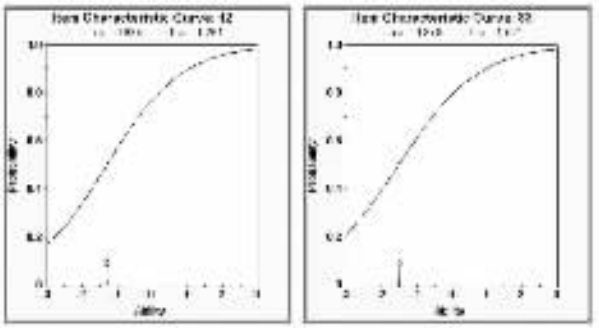
4.2.1 存在性别 DIF 的的项目反应曲线分析



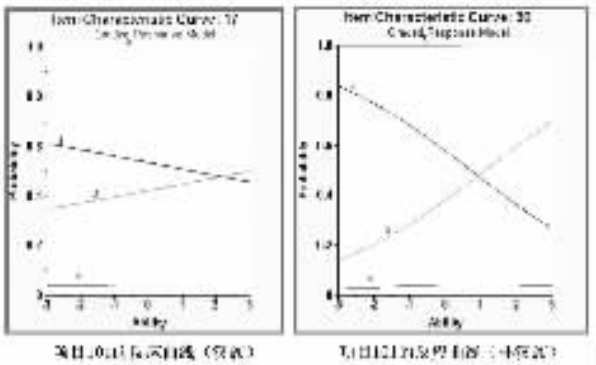
在项目 102 男女反应曲线图内各有 5 条曲线，表示特质 θ 的被试对每个等级反应的概率变化。该项目在目标组和参照组中都有着不等的区分度和难度，属于“非一致性 DIF”。由于男性区分度低于女

性,男性的 5 条曲线的起伏程度稍逊于女性。女性的阈值 b1 ~ b4 低于男性,相应的曲线位置略偏左。曲线 2 表示男女双方在 0.5 分值 P(k = 1)上的反应概率都很低,男性接近 0,曲线 3 和曲线 4 的主要差异在于:女性特质位于 - 1 ~ 1 处的被试在该等级 P(k = 2)和 P(k = 3)得分的概率高于男性,且最高概率达到 0.2;在得分值为 1.5 分 P(k = 3)的被试中,女性被试特质大多数集中于 0 ~ 3,而男性被试的分布广度大于女性。在分值为 2.0 即 P(k = 4)方面,曲线 5 表明:男性的 b4 阈值为 3.08,曲线幅度缓于女性,且在 - 3.0 ~ 3.0 特质范围内,最高概率低于 0.6,而女性接近 0.8。从两组项目特征曲线走向来看,中等能力水平被试的得分值大多数集中在 1.5 分值上。该项目的整体趋势是有利于目标组,这和均值显著性检验的结果是一致的。

4.2.2　存在民族 DIF 的项目反应曲线分析

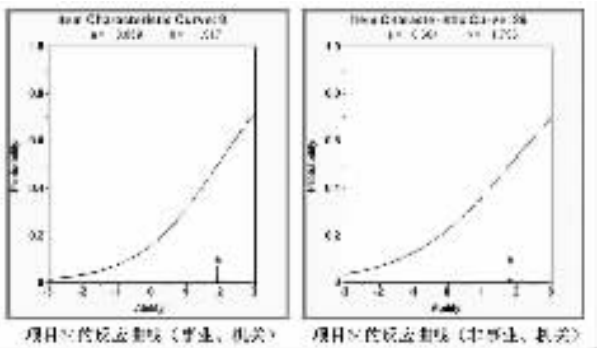


被检测存在 DIF 的项目 64 和项目 68 都是二级记分题,汉族区分度都要高于非汉族,曲线变化比较一致。这里以项目 68 为例,从该特征曲线可看到,曲线偏右,项目难度值较大,两组被试在该项目上得分概率都较低。非汉族曲线较汉族曲线走势平缓,在 - 3 ~ 3 特质领域内,汉族被试的通过人数比非汉族多,有利于汉族被试;而项目 68 区分度适中,阈值较低,两组被试在该项目上的通过率都比较高,有利于非汉族被试作答。



存在 DIF 的项目 99 和项目 101 都是多级反应试题。由于在 P(k = 3) 和 P(k = 4)等级上无正确作答分数,MULTILOG 软件没有估计出这两个等级上的项目阈值,因此只有三条运算特征(OCCs)曲线。项目在 0.5 分值等级上,曲线 2 退化明显,该等级上得分被试极少。以项目 101 为例,从它的反应曲线图来看,汉族区分度低于非汉族,汉族曲线的起伏程度稍逊于非汉族。特质水平较低的被试在项目 101 上得分的概率汉族被试比非汉族被试要高,同时特质高的非汉族被试得分概率要高的多,最高超过了 0.7,而特质高的汉族被试得分最高概率却低于 0.6。对于项目阈值,两个项目汉族的难度值都高于非汉族,两个项目的整体趋势有利于目标组。

4.2.3　存在工作性质 DIF 的项目反应曲线分析



项目 64 在民族和工作性质两个层面上都被检测出 DIF。对于事业机关被试该项目阈值为 1.971,对于非事业机关被试项目阈值为 1.796,难度值偏大。特质水平处于 0 的事业、机关被试的通过率不到 0.15,而非事业、机关被试的通过率超过了 0.2。双方被试在该项目上的得分都比较低。

4.3　DIF 根源的探讨

通过对存在 DIF 项目内容的分析,结合具体的项目统计量和反应曲线可以对出现 DIF 的项目进行原因探讨。但是对这些存在 DIF 的项目是否存在真正的项目偏差,还应从多方面进行验证。该研究对项目统计结果的分析和经济法的三位专业教师专家讨论,分析了产生 DIF 的几个可能原因:

1)性别 DIF 产生的原因:男女对经济法知识的理解和掌握差异;男女对经济法学科的兴趣和爱好差异;男女识记能力的差异性。如项目 102 的内容是问股份有限公司在向国务院证券监督管理机构提出股票上市交易申请时,应当提交哪些文件。该项目检测出来的结果表现为有利于女性,这可能一方面是女性对识记性项目的掌握能力比男性要强,另一方面女性一般从事秘书、文件整理等相关工作,这可能也和实际正确作答情况有关。(项目 102)

2)民族 DIF 产生的原因:少数民族区域和汉族

区域经济的发展状况存在差异;近年来东西部地区大力发展经济的意识有了很大的提高;国家对少数地区政策的倾斜和地方对经济、法学的重视。四个被检测有民族 DIF 的项目中,三个项目有利于非汉族被试作答,这和我国少数民族分布区域和企业分布区域有一定的关系。我国非汉族人口主要居住在西部地区,而西部地区也是我国工业的主要分属地,因此西部地区被试在这些项目上的作答表现出了一定的优势。(项目 64、68、99、101)

3)工作性质 DIF 产生的原因:非事业机关单位被试因工作方面的性质,对上市公司、企业发展相关的现代企业制度等法律规定在工作过程中有更实际性的理解,因此在经济法试题的测试中对实践性较强的项目的正确作答概率要高于事业、机关单位被试。(项目 64)

5 小结

该研究用基于项目反应理论 Samejima 等级反应模型(GRM)下的参数方法,在 MULTILOG 软件支持下,对某全国性的资格考试某科目试卷中经济法部分 21 个项目(既有二级题也有多级题)做 DIF 分析。检测结果如下:在 21 个项目中,不存在性别 DIF、民族 DIF、地区 DIF 和工作性质 DIF 中任何一种 DIF 的项目有 16 个。在地区被试的比较中,没有探测到 DIF;在性别和工作性质被试和民族被试的比较中也只有少数几个项目被检测到出现 DIF 现象,其中性别 DIF1 个(项目 102),民族 DIF4 个(项目 64、68、99、101),工作性质 DIF1 个(项目 64)。项目 68 在民族层面上表现为一致性 DIF,项目 64 既存在民族 DIF 又存在工作性质 DIF。通过对项目统计结果的分析和专家的讨论,文章分析了可能产生 DIF 的几个因素并概括如下:区域经济的发展差异;国家政策倾向方面的差异;工作经验和社会阅历的差异;不同群体的心理特征的差异;对考试内容的预备知识掌握程度和试题的理解程度等方面的差异。

参考文献

1 Steinberg L. Context and serial – order effects in personality measurement: Limits on the generality of measuring changes the measure. Journal of Personality and Social Psychology, 1994, 66: 341 – 349.

2 Smith L, Reise S. Gender differences on negative affectivity: an IRT study of Differential Item Functioning on the Multidimensional Personality Questionnaire Stress Reaction Stress Reaction Scale. Journal of Personality and Social Psyhology, 1998, 75: 1350 – 1362.

3 Thissen D. MULTILOG: A user's guide. Mooresville, IN: Scientific Software Inc, 1991.

4 Paul W. Holand, Howard Wainer. Differential Item Functioning. Published by Lawrence Erlbaum Associates, 1993.

5 漆书青,戴海崎,丁树良. 现代教育与心理测量学原理. 高等教育出版社,2002.8.

6 曹亦薇. 项目功能差异在跨文化人格问卷分析中的应用. 心理学报,2003,35(1):120 – 126.

The Exploration and Research of Parametric Method for DIF on the Items of The Laws Pertaining to the Economy

Li Li[1], Dai Haiqi[2], Dong Shenghong[2], Ou Dongming[3]

(1. Education College, Nanchang University, Nanchang 330029; 2. Psychology Department, Jiangxi Normal University, Nanchang 330027; 3. National Personnel Examinations Authority, Beijing 100080)

Abstract: In this study, based on the response the 21 optional item of the laws pertaining to the economy in some year intermediate economic test throughout country, a parametric procedure using the Graded Response Model and MULTILOG was used to detect DIF. The parametric procedures research found 4 items showed nation – related DIF; 1 item showed gender – related DIF; and only 1 item showed work classification – related DIF. Nation – related DIF (item 68) is a uniform DIF and item 64 appears both nation – related DIF and work classification – related DIF. According to the results of analysis among item statistics, characteristic curves and decisions of specialists, results show some possible causes of DIF.

key words: Differential Item Functioning; the Graded Response Model; Item Bias; Item Characteristic Curve