曾光,张玉玲,谢晓尧,等. 一种改进的 4 参数等级反应模型和应用 [J]. 江西师范大学学报(自然科学版),2023,47(2):124-132.

ZENG Guang,ZHANG Yuling,XIE Xiaoyao,et al. The improved four-parameter GRM and its application [J]. Journal of Jiangxi Normal University(Natural Science),2023,47(2):124-132.

文章编号:1000-5862(2023)02-0124-09

一种改进的 4 参数等级反应模型和应用

曾　光[1],张玉玲[2],谢晓尧[1],黎瑞源[1*]

(1. 贵州师范大学贵州省信息与计算科学重点实验室,贵州 贵阳　550001;2. 贵阳市教育科学研究所,贵州 贵阳　550001)

摘要:针对在实际测验中各等级的猜测参数、失误参数可能存在不一致性问题,该文提出等级反应模型的改进模型. 以 2 等级项目为例,通过模拟数据检验模型发现:使用 4 参数 GRM 模型估计参数的误差随着猜测参数和失误参数的各个不一致性而增大,而改进后的模型具备更好的稳定性. 在实际的地理测验中,发现等级反应项目中的猜测度较低,但失误现象明显,并且参数之间的差异性较大,不可以忽略.

关键词:项目反应理论;等级反应模型;4 参数 Logistic 模型

中图分类号:B 841　　**文献标志码**:A　　**DOI**:10.16357/j.cnki.issn1000-5862.2023.02.02

0　引言

随着学科融合发展,IRT 开始被应用在临床医学[1-4]、经济管理[5-7]、体育[8-9]等,不再局限于教育与心理测量中. 并且现代蓬勃发展的统计学、计算机科学和数据科学进一步为 IRT 的发展注入活力,庞大的数据量与计算量不再是阻碍,估计精度有了明显提高. 这些需求与条件促使研究者要根据实际情况选择适当的模型,或发展新的模型. 目前的项目反应理论模型已有 20 多种,需考虑什么样的模型更能拟合好实测数据.

近些年,国内研究者越来越关注心理测验中的 4PLM,关于 4PLM 的理论与应用研究相继取得一些研究成果. 如刘玥等[10]选取了来自心理测验和成就测验的实际数据,分别采用传统模型和 4PLM 进行拟合,这说明 4PLM 能够显著提高模型对心理测验和成就测验数据的拟合性. 金英姿等[11]同样在语言测验中发现一些项目确实存在睡眠现象,加入失误参数进行数据拟合非常必要.

4PLM 主要应用在 0-1 评分分项目中,然而在许多测评被试潜在特质的案例中,情况不仅限于被试样本回答了一组由一定数量的 0-1 评分项目组成的问卷,得到的结果也不只是一系列代表“正确”或“错误”的反应. 如果需衡量某种仅出现在深度的思考过程中的潜在特质,则需提前准备具有复杂推理过程的项目,根据被试在评分标准中所体现的目标的达成程度对项目进行评分,这类题目被称作多级评分题. Samejima 提出的等级反应模型(grade response model,GRM)在多级评分形式的测验中已被广泛使用.

传统 GRM 实质是在单、双参数 Logistic 模型框架下建立的,但 4PLM 的优势促使某些研究者对等级反应项目的猜测现象和失误现象进行研究. 陈青等[13-14]基于 GRM,在保持 GRM 的特性(项目等级难度递增)的条件下,将猜测参数融合到项目的整体参数中,即认为被试在完成多级评分试题的整个过程中,各个等级的猜测程度应该是不变的. 之后,简小珠等[15]同样将失误参数作为项目参数融合到 GRM 中,而猜测参数的概率均匀分配到各个项目特征函数中,用它们反映多级记分试题上的猜测现象和失误现象,从而使得被试能力高估现象和低估现

收稿日期:2022-09-12

基金项目:贵州省教育厅自然科学基金(黔科教 2009(0034)号)资助项目.

通信作者:黎瑞源(1980—),男,广西桂林人,副教授,博士,主要从事教育测评技术及应用研究. E-mail:ruiyuan.li@126.com

象得到了有效的纠正.

GRM 作为一个减法模型,在参数估计时难以保证相减所导出的差为非负值. 上述模型的每个项目只包含一个猜测参数或失误参数的约束,虽然保证了 GRM 概率必须非负的要求,也造成了当多级计分项目的等级参数不一致时难以拟合的困境. 为了反映多级计分项目各等级的相对独立性和差异性,区别于各等级猜测参数、失误参数一致的模型(本文记为 4PL-GRM),探究等级参数的差异性,本文提出一种改进的 4 参数等级反应模型(本文记为 4NPL-GRM). 在坚持假设合理与继承等级模型的特性的原则下,保证概率必须非负的规定,为各等级赋予合理的等级参数,提高了模型的普适性和估计结果的准确性.

1　模型介绍

本文在 GRM 的基础上加入不一致的猜测参数 c 和失误参数 γ,提出了等级反应模型的改进模型. 原理如下:

设存在一个有 N 个等级的项目,被试 α 的能力为 θ_α,各个得分等级中的项目特征函数为 4 参数 Logistic 函数,则被试在项目上得分不低于 t 分的概率 $P^*_{\alpha t} = c_t + (\gamma_t - c_t)/(1 + \exp(-1.7a(\theta_\alpha - b_t)))$,其中 a 为项目的区分度,b_t 是项目第 t 个等级的难度值,且 $b_1 < b_2 < \cdots < b_N$,c_t 是项目第 t 个等级的猜测参数,且 $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_N \geqslant 0$,$\gamma_t$ 是项目第 t 个等级的失误参数,且 $1 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_N$.

令 $P^*_{\alpha 0} = 1$,$P^*_{\alpha(N+1)} = 0$,可推出被试恰好得某分的概率 $P_{\alpha t} = P^*_{\alpha t} - P^*_{\alpha(t+1)}$,$t = 0,1,2,\cdots,N$.

1.1　模型的非负性证明

虽然 4NPL-GRM 给各等级赋予不同的猜测参数和失误参数,但在上述参数的约束下,依然可以证明相减所导出的差为非负值,没有违背概率必须非负的规定,以下是证明过程.

由于

$$P^*_{\alpha t} = c_t + (\gamma_t - c_t)/(1 + \exp(-1.7a(\theta_\alpha - b_t))), \tag{1}$$

对式(1) 的 c_t 求偏导可得

$$\partial P^*_{\alpha t}/\partial c_t = 1 - (1 + \exp(-1.7a(\theta_\alpha - b_t)))^{-1}, \tag{2}$$

对式(1) 的 γ_t 求导可得

$$\partial P^*_{\alpha t}/\partial \gamma_t = (1 + \exp(-1.7a(\theta_\alpha - b_t)))^{-1}. \tag{3}$$

由于 $\exp(x)$ 恒大于 0,故

$$0 < (1 + \exp(-1.7a(\theta_\alpha - b_t)))^{-1} < 1. \tag{4}$$

结合式(2) ~ 式(4) 可得

$$\partial P^*_{\alpha t}/\partial c_t > 0, \tag{5}$$

$$\partial P^*_{\alpha t}/\partial \gamma_t > 0. \tag{6}$$

又由于 $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_N \geqslant 0$,$1 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_N$,结合合式(5) 和式(6) 可得 $P^*_{\alpha t} \geqslant P^*_{\alpha(t+1)}$,即 $P_{\alpha t} \geqslant P^*_{\alpha t} - P^*_{\alpha(t+1)}$,证毕.

通过图形描述,可以更进一步了解 4NPL-GRM 的特点. 图 1 给出了一个 3 等级的项目运算特征曲线和项目等级反应曲线,并给出该项目去除猜测参数和失误参数后的相应 GRM 曲线,以供比较研究. 参数 $a = 1.00$,$b = (-1.00, 0.00, 1.50)$,$c = (0.20, 0.10, 0.05)$,$\gamma = (0.95, 0.90, 0.80)$.

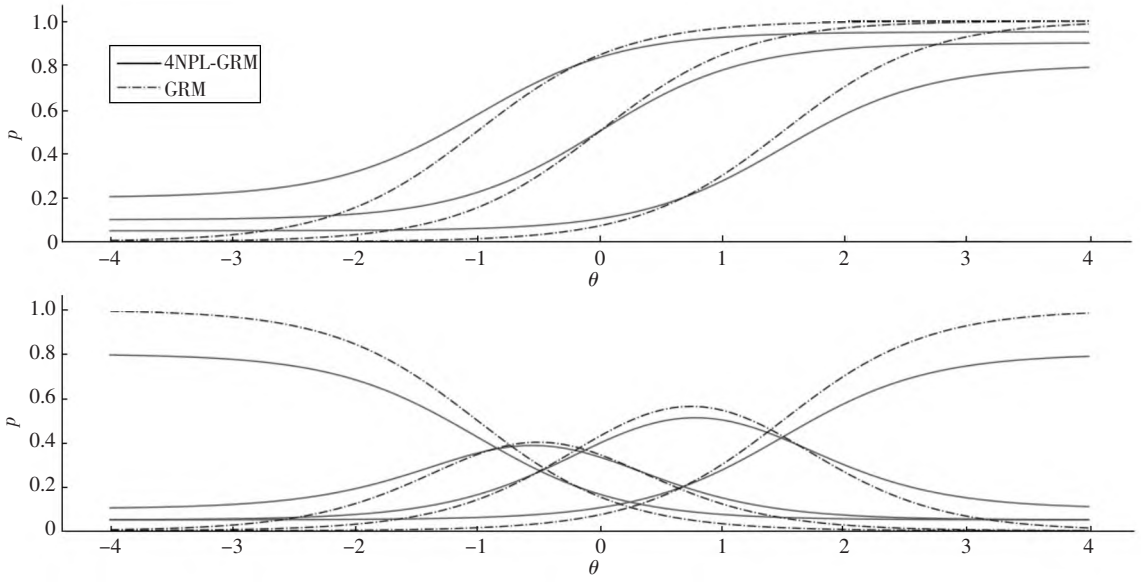

图 1　3 等级项目的项目运算特征曲线和项目等级反应曲线

传统 GRM 的项目运算特征曲线是由 2PLM 的项目反应曲线组合而成,而 4NPL-GRM 采用 4PLM,其上下渐进值不再固定为 1 和 0,而是同时逐级递减. 因此,观察图 1 的特征曲线,随着等级数增加,传统 GRM 的特征曲线表现为简单向右平移,4NPL-GRM 则可以被近似看作向右下方平移. 因为 4NPL-GRM 较 GRM 增加了逐级向下平移的趋势,在相邻2级相减构建项目等级反应曲线中,4NPL-GRM 的各级曲线左右渐进值必定远离 0 值,图 1 中 3 等级项目的等级反应曲线清晰地展现出这种特点.

1.2 模型的比较

为探究等级猜测参数与失误参数的存在必要性与差异性的影响,表 1 展示了 4 种模型的特征.

表 1 各个模型的特征比较

模型	等级数	区分度	难度	猜测参数	失误参数
4PLM	1	存在且唯一	存在且唯一	$c \geqslant 0$	$1 \geqslant \gamma$
GRM	> 1	存在且唯一	$b_1 < b_2 < \cdots < b_N$	$c_1 = c_2 = \cdots = c_N = 0$	$1 = \gamma_1 = \gamma_2 = \cdots = \gamma_N$
4PL-GRM	> 1	存在且唯一	$b_1 < b_2 < \cdots < b_N$	$c_1 = c_2 = \cdots = c_N \geqslant 0$	$1 \geqslant \gamma_1 = \gamma_2 = \cdots = \gamma_N$
4NPL-GRM	> 1	存在且唯一	$b_1 < b_2 < \cdots < b_N$	$c_1 \geqslant c_2 \geqslant \cdots \geqslant c_N \geqslant 0$	$1 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_N$

4PL-GRM 从传统的 GRM 发展而来,同样是一个减法模型,用减法模型来描述这个多级评分项目. 其参数分成 2 类,一类是描述等级的,如难度参数;另一类是描述整个项目的,如区分度参数,猜测参数和失误参数. 对于 4PL-GRM,当等级数为 1 时,模型简化成 4PLM,而当 $c_1 = c_2 = \cdots = c_N = 0$ 且 $\gamma_1 = \gamma_2 = \cdots = \gamma_N = 1$ 时,模型又简化为 GRM. 因此,该模型相较传统的 GRM 拟合能力和普适性更强. 而本文提出的4NPL-GRM 继承了4PL-GRM 引入猜测参数和难度参数的优点,在面对实际测验中多级反应项目(如当项目各选项考察的内容不一样)时,4NPL-GRM 突破了猜测参数和失误参数必须一致的限制,将它们考虑为等级参数,N 个等级就会有 N 个猜测和失误参数,当 4NPL-GRM 的各等级猜测参数和失误参数相同时,可以简化为 4PL-GRM. 因此,4NPL-GRM 进一步拓广了适用范围.

2 项目参数和潜在能力的条件估计

2.1 项目参数条件估计

为了检验各模型的拟合性能,使用固定项目参数的设计方法. 假定被试能力值 $\theta \sim N(0,1)$,被试个数设为40 000,设计 3 个由 100 道 2 级试题组成的测验. 其中,1 号测验 $a \sim U(0.7,3.5)$,$b_t \sim N(0,1)$,$b_1 < b_2$;2 号测验在 1 号测验的基础上增加项目猜测参数 $c \sim \mathrm{Beta}(3,17)$ 和失误参数 $\gamma \sim \mathrm{Beta}(17,3)$;3 号测验在 1 号测验的基础上增加等级猜测参数 $c_t \sim \mathrm{Beta}(3,17)$、$c_1 \geqslant c_2$ 和等级失误参数 $\gamma_t \sim \mathrm{Beta}(17,3)$、$\gamma_1 \geqslant \gamma_2$.

模拟被试作答:满分为 2 分的多级记分试题,被试得 0 分及 0 分以上(即 0 分,1 分,2 分)的概率为 $P^*_{\alpha 0} = 1$,被试得 1 分及 1 分以上(即 1 分和 2 分)的概率为 $P^*_{\alpha 1} = c_1 + (\gamma_1 - c_1)/(1 + \exp(-1.7a(\theta_\alpha - b_1)))$,被试得满分 2 分的概率为 $P^*_{\alpha 2} = c_2 + (\gamma_2 - c_2)/(1 + \exp(-1.7a(\theta_\alpha - b_2)))$. 由此进一步得出,被试恰好得 1 分的概率为 $P_{\alpha 1} = P^*_{\alpha 1} - P^*_{\alpha 2}$,被试恰好得0分的概率为 $P_{\alpha 0} = 1 - P^*_{\alpha 1}$,依据被试在试题上的作答概率,通过蒙特卡洛模拟方法产生被试得分.

估计方法采用适用条件广泛、原理简单的三点法[16],并结合潜在能力真值进行极大似然估计,得到项目参数. 最后,为了比较 3 种模型的偏差和返真性能,需要比较估计参数与真值,计算以下 3 种指标:平均偏差(mean error,M_E),平均绝对误差(mean absolute error,M_{AE})和均方根误差(root mean squared error,R_{MSE}).

$$M_E = \sum_{j=1}^{m} (x_j^* - x_j)/m,$$

$$M_{AE} = \sum_{j=1}^{m} |x_j^* - x_j|/m,$$

$$R_{MSE} = \sqrt{\sum_{j=1}^{m} (x_j^* - x_j)^2/m},$$

其中 x_j^* 和 x_j 分别表示模拟数据第 j 个项目参数的估计值和真值,评价结果见表 2.

考察 1 号测验估计结果的返真性,使用各个模型的 M_{AE} 与 R_{MSE} 指标进行比较,得到最大平均绝对误差 $M_{AE}(b_1) = 0.031\,7$,最大均方根误 $R_{MSE}(b_1) = 0.150\,7$,这些评价指标明显均在正常范围内,返真效果较好,这说明 1 号测验数据适宜被 GRM、4PL-GRM 和4NPL-GRM 同时拟合. 仅对比区分度和难度参数的返真效果,GRM 的 M_{AE} 与 R_{MSE} 整体小于

4PL-GRM,4PL-GRM 的 M_{AE} 与 R_{MSE} 又整体小于 4NPL-GRM,这说明在适用的基础上,模型拟真效果是不同的,对 1 号测验数据返真性排序为 GRM > 4PL-GRM > 4NPL-GRM.

表 2　各个测验的评价结果

数据	模型	评价指标	a	b_1	b_2	c_1	c_2	γ_1	γ_2
1 号	GRM	M_E	0.000 3	−0.020 4	0.001 6				
		M_{AE}	0.020 0	0.025 9	0.005 9				
		R_{MSE}	0.045 4	0.139 0	0.009 1				
	4PL-GRM	M_E	0.006 4	−0.021 1	0.001 8	0.000 5	0.000 5	−0.000 4	−0.000 4
		M_{AE}	0.020 5	0.026 7	0.006 5	0.000 5	0.000 5	0.000 4	0.000 4
		R_{MSE}	0.045 8	0.149 6	0.009 6	0.002 0	0.002 0	0.001 4	0.001 4
	4NPL-GRM	M_E	0.012 3	−0.015 6	−0.004 4	0.004 7	0.000 7	−0.000 6	−0.005 5
		M_{AE}	0.024 1	0.031 7	0.009 1	0.004 7	0.000 7	0.000 6	0.005 5
		R_{MSE}	0.047 3	0.150 7	0.018 2	0.012 1	0.002 1	0.002 9	0.019 6
2 号	GRM	M_E	−1.530 0	−0.149 5	−0.000 1				
		M_{AE}	1.530 0	0.314 5	0.271 6				
		R_{MSE}	1.700 2	0.543 2	0.3792				
	4PL-GRM	M_E	0.008 6	−0.002 5	−0.001 0	0.000 3	0.000 3	−0.000 6	−0.000 6
		M_{AE}	0.049 2	0.013 0	0.011 5	0.003 5	0.003 5	0.003 3	0.003 3
		R_{MSE}	0.070 2	0.020 7	0.016 8	0.005 2	0.005 2	0.005 7	0.005 7
	4NPL-GRM	M_E	0.011 0	0.002 9	−0.004 6	0.002 7	0.000 7	−0.000 5	−0.002 8
		M_{AE}	0.051 4	0.015 3	0.015 2	0.008 6	0.003 2	0.003 3	0.008 3
		R_{MSE}	0.072 8	0.022 5	0.025 3	0.014 4	0.004 6	0.005 2	0.015 1
3 号	GRM	M_E	−1.417 0	−0.236 0	0.2831				
		M_{AE}	1.417 0	0.312 4	0.344 1				
		R_{MSE}	1.578 2	0.399 0	0.444 6				
	4PL-GRM	M_E	−0.356 0	−0.139 6	0.174 2	−0.080 3	0.005 6	−0.002 5	0.089 1
		M_{AE}	0.356 4	0.139 8	0.175 0	0.080 4	0.013 8	0.012 9	0.089 1
		R_{MSE}	0.492 6	0.182 8	0.258 8	0.101 4	0.019 4	0.018 4	0.117 0
	4NPL-GRM	M_E	0.003 7	0.001 4	−0.001 1	0.001 3	0.000 1	−0.000 4	−0.001 6
		M_{AE}	0.041 8	0.018 2	0.015 7	0.008 9	0.003 3	0.003 4	0.007 7
		R_{MSE}	0.055 3	0.033 0	0.024 3	0.016 3	0.005 2	0.006 5	0.012 7

考察 1 号测验估计结果的偏向性,比较各模型的区分度和难度参数的 M_E 和 M_{AE} 后发现,$|M_E|$ 均明显小于 M_{AE},这表明 GRM、4PL-GRM 和 4NPL-GRM 对 1 号测验数据的参数估计无显著偏向.

考察 2 号测验估计结果的返真性,使用各个模型的 M_{AE} 与 R_{MSE} 指标进行比较,得到 4PL-GRM 最大平均绝对误差 $M_{AE}(a)$ = 0.049 2,最大均方根误差 $R_{MSE}(a)$ = 0.070 2;得到 4NPL-GRM 最大平均绝对误差 $M_{AE}(a)$ = 0.051 4,最大均方根误差 $R_{MSE}(a)$ = 0.072 8,这些评价指标均在正常范围内,返真效果优秀. 以上结果说明 2 号测验数据适宜被 4PL-GRM 和 4NPL-GRM 同时拟合,而 GRM 是最大平均绝对误差 $M_{AE}(a)$ = 1.530 0,最大均方根误差 $R_{MSE}(a)$ = 1.700 2,这些评价指标显著超出正常范围,结合估计参数与模拟数据进行卡方检验(显著性水平 0.05),检验通过率为 0%,这说明 2 号测验数据完全无法被 GRM 拟合. 比较 4PL-GRM 和 4NPL-GRM 的区分度和难度参数的返真效果,4PL-GRM 的 M_{AE} 与 R_{MSE} 整体小于 4NPL-GRM,因此对 2 号测验数据表现为 4PL-GRM > 4NPL-GRM.

考察 2 号测验估计结果的偏向性,比较各模型的区分度和难度参数的 M_E 和 M_{AE} 后发现,4PL-GRM 和 4NPL-GRM 各估计参数的 $|M_E|$ 均明显小于 M_{AE},这表现出 4PL-GRM 和 4NPL-GRM 对 2 号测验数据的参数估计的无偏向性,而 GRM 的 $M_{AE}(a)$ = 1.530 0,$M_E(a)$ = −1.530 0,GRM 表现出对 2 号测验数据的区分度估计有显著偏向,且偏向为负,即 GRM 区分度估计值相较于真值明显偏低.

考察 3 号测验估计结果的返真性,使用各个模型的 M_{AE} 与 R_{MSE} 指标进行比较,得到 GRM 最大平均绝对误差 $M_{AE}(a)$ = 1.417 0,最大均方根误差 $R_{MSE}(a)$ = 1.578 2,这些评价结果显著超出正常范围,卡方检验通过率仅为 1%. 因此,3 号测验数据完全无法被 GRM 拟合;4PL-GRM 最大平均绝对误差

$M_{AE}(a)$ = 0.356 4, 最大均方根误差 $R_{MSE}(a)$ = 0.492 6, 卡方检验通过率为 19%. 因此,3 号测验数据可以被4PL-GRM 部分拟合. 而4NPL-GRM 最大平均绝对误差 M_{AE} = 0.041 8, 最大均方根误差 $R_{MSE}(a)$ = 0.055 3,这些评价结果明显均在正常范围内,卡方检验通过率为 91%. 比较以上模型,可以说明 3 号测验数据仅适宜被 4NPL-GRM 拟合.

考察 3 号测验估计结果的偏向性,比较各模型的区分度和难度参数的 M_E 和 M_{AE} 发现,4NPL-GRM 的各估计参数 $|M_E|$ 均明显小于 M_{AE}, 这表现出 4NPL-GRM 对 3 号测验数据的参数估计的无偏向性. 而 GRM 的 $M_E(a)$ = −1.417 0, $M_{AE}(a)$ = 1.417 0; 4 PL-GRM的 $M_E(a)$ = −0.3560, $M_{AE}(a)$ = 0.3564, $M_E(b_1)$ = −0.139 6, $M_{AE}(b_1)$ = 0.139 8, $M_E(b_2)$ = 0.174 2, $M_{AE}(b_2)$ = 0.175 0;GRM 表现出对 2 号测验数据的区分度估计的显著偏向,4PL-GRM 表现出对 2 号测验数据的区分度与难度估计的显著偏向,即 GRM 的区分度和4PL-GRM 的区分度与难度 b_1 的估计值,相较于真值明显偏低,4PL-GRM 的难度 b_2 的估计值相较于真值明显偏高.

2.2 潜在能力条件估计

使用估计的项目参数,进行潜在能力的条件估计,可以更直观地描述各个模型. 估计方法采取使用较为广泛的后验期望估计法,该方法的优势是对于极端的反应数据表现比较稳定且容易计算. θ 的先验分布选取为标准正态分布,根据积分理论,求积节点的个数越多,所求得的估计值就越准确,在本实验中取为 60.

图 1 展示了 3 种测验数据的 GRM、4PL-GRM 和 4NPL-GRM 的 $M_E(\theta)$ 随 θ 的变化曲线(分 200 组作均匀光滑). 经仔细观察,可以发现以下现象:

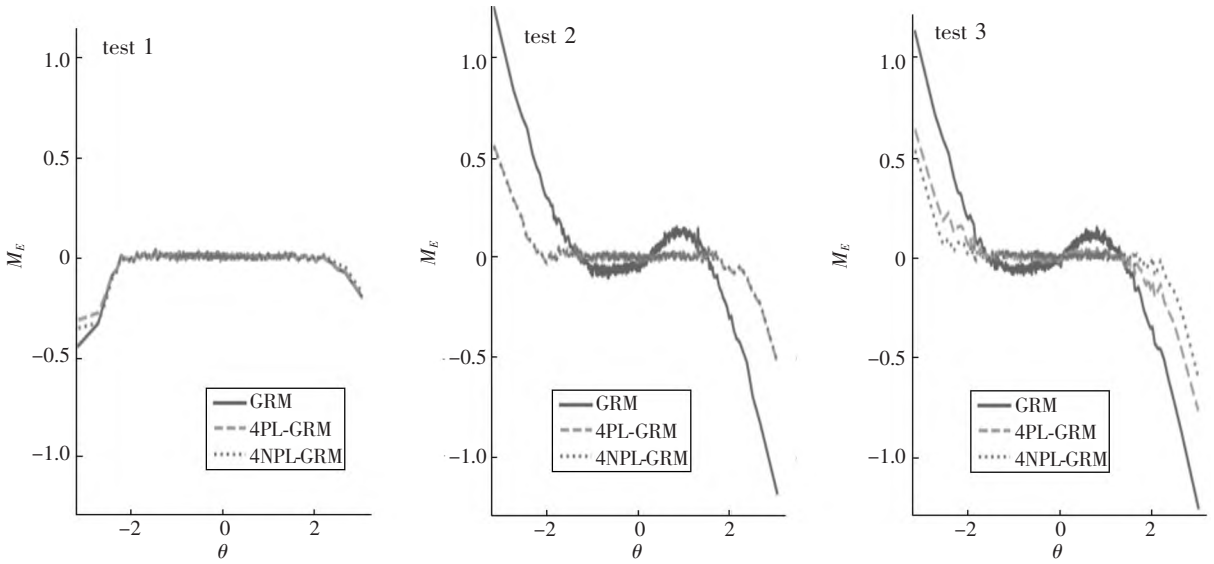

图 2 在 3 种测验数据下各模型的 M_E 变化曲线

1) 在1 号测验数据中,各模型曲线近乎重合,并具有相同的变化趋势; 当 $|\theta| < 2$ 时,曲线稳定在 0 值,将这种区间称作稳定区间,即此时估计值约为真值; 当 $|\theta| > 2$ 时, $M_E(\theta)$ 开始趋于负值,即估计值较真值逐渐偏小.

2) 在2 号测验中,4PL-GRM 和4NPL-GRM 曲线重合并整体呈现单调递减的变化,当 $|\theta| < 1.8$ 时有稳定区间; 而当 $\theta > 1.8$ 时,它们的 $M_E(\theta)$ 趋于负值; 当 $\theta < -1.8$ 时,它们的 $M_E(\theta)$ 趋于正值. 此时, GRM 曲线不具备稳定区间并呈现振荡的特性,从 θ =0 开始向右移动, $M_E(\theta)$ 开始时缓慢趋于正值,之后快速趋于负值,从 θ = 0 开始向左移动, $M_E(\theta)$ 开始时缓慢趋于负值,之后快速趋于正值,这表明 GRM 对潜在能力估计是不稳定的. 当 $|\theta| > 2$ 时, 2 类 曲线趋势相同,可以计算曲线之间间隔平均为 0.517, 即对比 GRM、4PL-GRM 和 4NPL-GRM 使高能力和低能力被试估计值得到了有效的纠正,矫正值为 0.517.

3) 在 3 号测验中,GRM、4PL-GRM 和 4NPL-GRM 曲线表现为与 2 号测验相似的变化趋势,稳定区间缩小为 $|\theta| < 1.5$. 当 $|\theta| > 1.5$ 时,曲线分离,可以计算 GRM、4PL-GRM 曲线间隔平均为0.382. 4PL-GRM 和 4NPL-GRM 曲线间隔平均为0.161, 即 4PL-GRM 在 GRM 的基础上平均矫正高能力和低能力被试估计值为 0.382,4NPL-GRM 继续在 4PL-GRM 的基础上平均矫正估计值为 0.161.

4) 综合 3 种测试条件可以发现,GRM 对 1 号测

验的估计最具优势,4PL-GRM 对 2 号测验的估计最具优势,但整体来说,4NPL-GRM 表现出了最好的稳定性与估计无偏性.

3　4PL-GRM 的估计偏差

当 4PL-GRM 拟合带有猜测和失误等级参数差异性的项目时,区分度和难度参数会整体偏移,为了进一步分析等级参数差异性导致的估计参数的变化趋势,此实验分为 2 类. 第 1 类实验探究猜测参数的差异性引起的估计偏差,故固定 $a = 1.5, \gamma_1 = \gamma_2 = 1, c_2 = 0, c_1$ 由 0 到 0.4 以 0.04 的间隔递增. 第 2 实验探究失误参数的差异性引起的估计偏差,故固定 $a = 1.5, c_1 = c_2 = 0, \gamma_1 = 1, \gamma_2$ 由 1 到 0.6 以 0.04 的间隔递减. 考虑到难度参数对实验的重要影响,因此取 3 个水平,$b_1 = -1.5$ 且 $b_2 = -0.5$,$b_1 = -0.5$ 且 $b_2 = 0.5$,$b_1 = 0.5$ 且 $b_2 = 1.5$.

图 2 和图 3 展示了在 3 种难度水平下的 4PL-GRM 的区分度和难度参数估计的偏差随等级参数差异值的变化曲线. 经仔细观察, 可以发现以下现象:

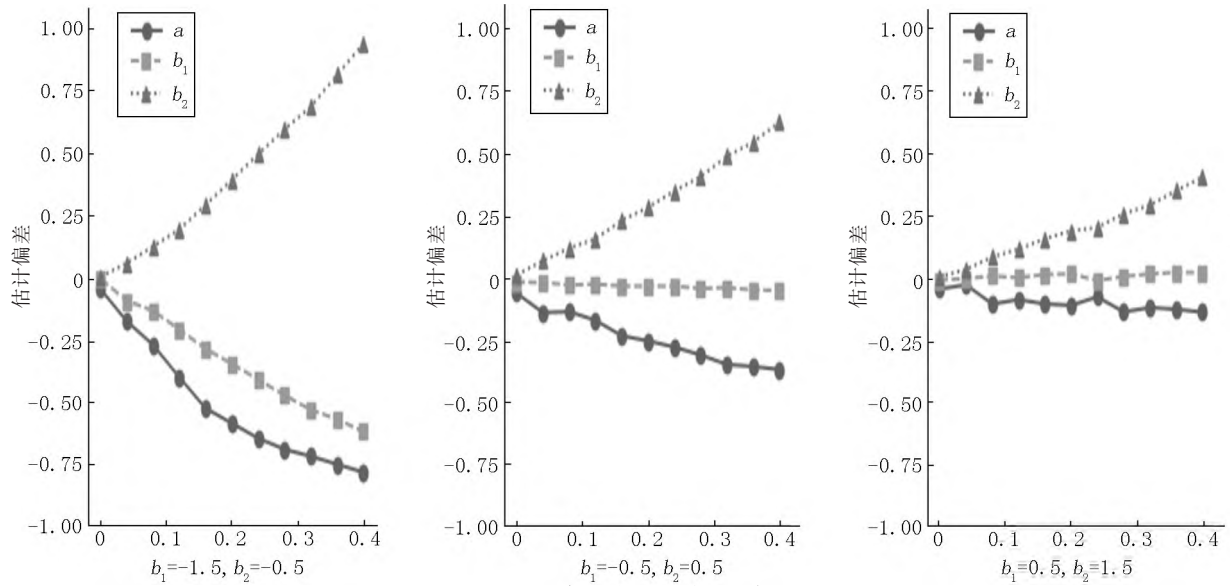

图 4　失误参数变化下的估计偏差

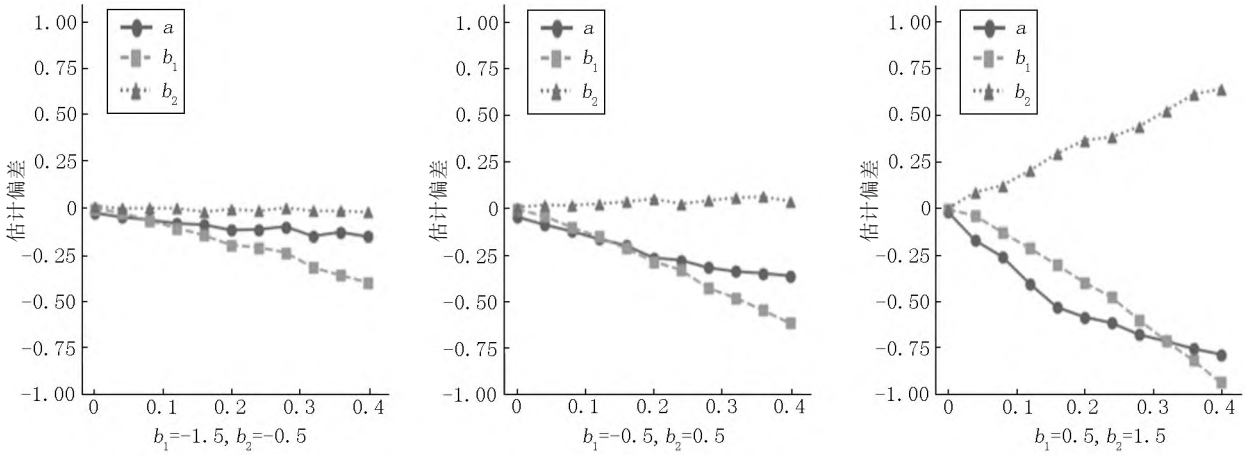

图 5　在猜测参数变化下的估计偏差

1) 结合图 1 和图 2 的共同特点,发现等级参数的差异性和难度的移动变化均会改变区分度和难度的偏差,使之单调变化,但单调性不变. 自猜测或失误参数的差异开始增加,估计值 a、b_1 的偏离值单调不增,b_2 偏离值单调不减,这与估计项目参数实验的数据与结论保持一致.

2) 图 1 和图 2 表现不同的是,失误参数差异增大的情况下,随着项目难度增大,a、b_1、b_2 的曲线表现愈来愈平缓. 相反,在猜测参数差异增大的情况下,随着项目难度增大,a、b_1、b_2 的曲线表现愈来愈陡峭,这说明对于难度越低的 2 等级项目,4PL-GRM 对失误参数差异性的变化越敏感,参数的估计值也

越接近真值,而4PL-GRM对猜测参数差异性的变化越迟钝,a、b_1、b_2参数的估计值也越远离真值.

3)对比b_1、b_2,即使同是难度参数,偏离方向和程度都不同,在图1中b_2较b_1的曲线更加陡峭,在图2中b_2较b_1的曲线更加平缓.这说明,不同等级参数的差异性对每个等级的难度参数估计的干扰也不相同,失误参数的差异性对b_2的影响比b_1更加大,猜测参数的差异性对b_1的影响比b_2更大.

4 实测分析

4.1 实测方法

在实践中,能力参数和项目参数都未知,需要同时对能力和项目参数进行估计,在这种情况下可采用R软件平台的mirt包提供的MCEM算法估计项目参数与能力参数.尽管 mirt 包并没有直接提供4PL-GRM模型和4NPL-GRM模型,但由于等级反应模型本质是由 Logistic 模型组合而成,因此一个多级计分类型项目数据也可以逆向转化为多个0-1计分类型项目数据(以表3为例,项目由0,1,2,3,4计分),利用 mirt 包已经提供的4PL模型,同时限制参数条件,可以间接得到4PL-GRM模型和4NPL-GRM模型的估计结果.为了检验估计方法的质量,并分别对比4NPL-GRM与4PL-GRM,GRM的性能,需要进行模拟测验的估计.

表3 得分转换表

得分	转换后			
0	0	0	0	0
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4	1	1	1	1

设计3类由0-1评分题,2等级题,4等级题和6等级题各5道,和10 000个被试$\theta \sim N(0,1)$组成的测验.在1类测验中,$a \sim U(0.7,3.5)$,$b_t \sim N(0,1)$,$b_1 < b_2 < \cdots < b_N$,2类测验是在1类测验的基础上增加项目猜测参数$c \sim \mathrm{Beta}(3,17)$和失误参数$\gamma \sim \mathrm{Beta}(17,3)$,3类测验是在1类测验的基础上增加等级猜测参数$c_t \sim \mathrm{Beta}(3,17)$,$c_1 \geqslant c_2 \geqslant \cdots \geqslant c_N \geqslant 0$和等级失误参数$\gamma_t \sim \mathrm{Beta}(17,3)$,$1 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_N$.模拟实验重复10次,取得估计参数相应的评价结果均值见表4.

表4 模拟数据的评价结果

数据	模型	指标	a	b	c	γ	θ
1	GRM	M_{AE}	0.032 4	0.012 4	0.000 0	0.000 0	0.110 6
		R_{MSE}	0.051 0	0.017 2	0.000 0	0.000 0	0.149 6
	4NPL-GRM	M_{AE}	0.048 1	0.018 6	0.004 6	0.004 9	0.111 0
		R_{MSE}	0.065 6	0.035 2	0.019 6	0.019 3	0.150 4
2	4PL-GRM	M_{AE}	0.102 5	0.024 6	0.006 9	0.007 1	0.176 7
		R_{MSE}	0.139 5	0.034 9	0.013 1	0.012 5	0.236 8
	4NPL-GRM	M_{AE}	0.117 2	0.049 6	0.016 5	0.015 8	0.177 1
		R_{MSE}	0.159 4	0.099 9	0.036 7	0.033 7	0.237 7
3	4NPL-GRM	M_{AE}	0.117 0	0.037 2	0.014 7	0.017 1	0.174 4
		R_{MSE}	0.157 7	0.063 4	0.031 6	0.044 3	0.235 2

从表4的结果来看,MCEM算法配合4NPL-GRM对3类测验均能有效估计,但如果与GRM或4PL-GRM对比会发现,4NPL-GRM的各项参数估计R_{MSE}与M_{AE}更大,而且10次模拟实验估计结果都表明:GRM对1类测验,4PL-GRM对2类测验的估计优势是4NPL-GRM无法企及的.这种现象的出现可以被认为是由于4NPL-GRM追求模型的普适性,模型参数的增加导致估计结果出现不可避免的精度损失.

4.2 实测数据

对2020年某省的地理测验进行实测数据分析,数据包含51 238名考生,17道0-1评分题(包括15道多项选择题)和4道2等级题,2道3等级和4等级题,1道6等级和8等级题.对数据的分析条件进行检验,KMO检验统计量为0.960,Bartlett球型检验$p = 0.000$,提取出的第1个因子特征根为6.67,第2个因子特征根为1.21,第1个因子与第2个因子特征根比值为5.51,说明该试卷符合单维性假设,取得实测参数估计结果的均值见表5.

实测数据的具体估计结果显示,在 10 个多级计分项目中,包括 36 个等级猜测参数和失误参数,其中猜测度低于 0.004 的参数有 25 个,占总体的 69.4%,等级项目猜测参数整体均值为 0.026 6,而猜测现象主要集中在多等级项目前 2 个等级中,第 1 等级的猜测参数均值为 0.051 3,第 2 等级为 0.033 3,可以认为多级评分项目的猜测度较低,远小于 0-1 计分项目的 0.201 2,这与陈青等[13]研究一致. 观察发现相邻等级猜测参数差异值共 26 个,差异值最大前 10 平均为 0.061 0,整体最大为 0.167 6,因此猜测参数的差异性不可以忽略.

表 5　实测数据估计结果均值

	a	b	c	γ
0-1 评分项目	1.651 4	-0.974 9	0.201 2	0.931 9
多级评分项目	2.161 8	-0.160 2	0.026 6	0.857 1
所有项目	1.840 5	-0.440 2	0.081 0	0.883 0

而多级计分项目中失误现象较 0-1 计分项目更加明显,其中失误参数低于 0.94 的有 15 个,低于 0.90 的有 9 个,分别占总体的 41.7% 和 25.0%,失误现象主要集中在多等级项目最后 2 个等级中,最后等级的失误参数均值为 0.781 7,另一等级为 0.876 6,参数整体均值为 0.857 1,小于 0-1 计分项目的 0.931 9. 观察发现相邻等级失误参数差异整体平均为 0.075 7,差异值最大前 10 均值为 0.178 2,整体最大为 0.510 4,因此失误参数和其差异性也不可以忽略.

另外参数估计结果还显示了多级评分项目各个等级的评价质量,其中 8 级计分项目的失误参数分布为 1.000 0、1.000 0、1.000 0、1.000 0、1.000 0、0.999 9、0.999 8、0.998 6,而 6 级计分项目的失误参数分布为 0.892 2、0.834 9、0.324 6、0.280 5、0.048 3、0.035 4,发现失误参数估计值异常低,检查原始得分数据,按总分排名并筛选出前 5.6% 的被试,统计得分占比(由低向高):1.0%、1.0%、34.5%、5.5%、48.1%、2.1%、7.5%,高潜力被试的高分比例不仅低,而且得分比例较分散,项目各等级区别明显,因此在地理测验中,此 6 级计分项目使用 4NPL-GRM 是必要的.

5　讨论

由于在实际测验中多级反应项目十分复杂,比如当项目各选项考察的内容不一样时,本文基于 GRM 提出了适用于等级参数不一致的 4NPL-GRM. 根据模拟研究的结果表明:与 GRM 和 4PL-GRM 相比,4NPL-GRM 表现出更加优良的统计性质. 首先,4NPL-GRM 在 3 次测验中估计参数没有出现明显误差,4PL-GRM 仅在具备等级参数差异性的 3 号测验中出现较大误差,而 GRM 的较大误差同时出现在 2、3 号测验中. 其次,4NPL-GRM 具有优秀的估计无偏性,保证估计具有良好的精度.

相比之下,4PL-GRM 的表现较差,据改变等级参数差值的模拟研究的结果表明:等级猜测参数和失误参数之间差异性越大,偏离程度越大,并且难度较低的项目对失误参数差异性表现明显,难度较高的项目对猜测参数差异性表现明显. 因此,当测试的项目存在明显的等级猜测参数和失误参数差异性时,不宜选用 4PL-GRM. 而使用 MCEM 算法同时对能力和项目参数进行估计时发现,4NPL-GRM 为追求模型的普适性,模型参数的增加导致估计结果出现不可避免的精度损失. 因此,3 种等级反应模型各有优缺,需要根据实际情况谨慎选择.

最后,在实际的地理测验下,可以发现在等级反应项目中猜测度较低,但失误现象明显,并且存在猜测参数与失误参数各等级之间的差异性较大的情况,不可以忽略. 因此,使用 4NPL-GRM 才能更加全面地反映项目的各个等级的特性,评价各等级质量,做出有效的测验编制和更加精确估计的潜在能力估计.

6　参考文献

[1] HUNG S P, LIAO Yihan, ECCLESTON C, et al. Developing a shortened version of the dementia knowledge assessment scale (DKAS-TC) with a sample in Taiwan: an item response theory approach [J]. BMC Geriatrics, 2022, 22(1): 1-10.

[2] DONGWOO C, KYUNGSOO P. An item response theory based integrated model of headache, nausea, photophobia, and phonophobia in migraine patients [J]. Journal of

Pharmacokinetics & Pharmacodynamics, 2018, 45(5): 721-731.
[3] 刘叶,鲁杰,李顶春,等. 基于经典测量理论和项目反应理论对慢性病毒性肝炎患者生命质量量表的评价[J]. 临床肝胆病杂志,2022,38(11):2470-2477.
[4] 李建生,冯贞贞,谢洋. 基于临床调查的慢性阻塞性肺疾病稳定期证候疗效评价量表的初步形成[J]. 中医杂志,2022,63(13):1235-1242.
[5] 夏雨霏,霍增辉. 上市公司精准扶贫能力估算及差异特征:基于等级反应模型 GRM 的实证研究[J]. 商业会计,2022(13):29-33.
[6] 霍增辉,张玫,吴海涛. 基于项目反应理论的农户相对贫困测度研究:来自浙江农村的经验证据[J]. 农业经济问题,2021(7):57-66.
[7] 王保鲁,JUNG H B. 基于项目反应理论的纺织服装企业新技术接受期望研究[J]. 北京服装学院学报(自然科学版),2020,40(2):82-87.
[8] 杨尚剑. 基于项目反应理论的运动员组织公民行为量表的修订及在凝聚力与满意度中的中介作用[J]. 沈阳体育学院学报,2019,38(2):52-57,64.
[9] 孙文树. 基于 CTT、IRT、FT 的体育明星代言人信源模型量表研究[J]. 哈尔滨体育学院学报,2019,37(1):18-27.
[10] 刘玥,刘红云. 四参数 Logistic 模型和传统模型对被试作答拟合能力的比较研究[J]. 心理学探新,2018,38(3):228-235.
[11] 金英姿,王佶旻. 四参数 Logistic 模型与双参数、三参数 Logistic 模型在语言测验中的拟合比较及睡眠现象检验:以来华留学生预科结业考试为例[J]. 中国考试,2022(8):57-65.
[12] SAMEJIMA F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika Monograph, 1969,34(17):1-7.
[13] 陈青,丁树良,朱隆尹,等. 3 参数等级反应模型及其参数估计[J]. 江西师范大学学报(自然科学版),2010,34(2):117-122.
[14] 陈青,丁树良. 三参数等级反应模型及其信息函数的应用[J]. 考试研究,2009,5(2):77-84.
[15] 简小珠,戴海琦. 4 参数 GRM 对猜测现象和失误现象的纠正[J]. 江西师范大学学报(自然科学版),2016,40(2):140-144.
[16] 胡小芳. IRT 中参数估计的新方法:三点法[D]. 重庆:西南大学,2016.

The Improved Four-Parameter GRM and Its Application

ZENG Guang[1], ZHANG Yuling[2], XIE Xiaoyao[1], LI Ruiyuan[1*]
(1. Guizhou Key Laboratory of Information and Computing Science, Guizhou Normal University, Guiyang Guizhou 550001, China;
2. Guiyang Institute of Educational Sciences, Guiyang Guizhou 550001, China)

Abstract: Due to the complexity of multi-level response items in the actual test, there may be inconsistencies in the guess parameters and error parameters of each level. Therefore, an improved model of GRM is proposed, which has reasonable assumptions and maintains the characteristics of GRM, and has better universality and accuracy. Taking the second-class response item as an example, through the simulation data to test the model, it is found that the error of estimating parameters using the 4PL-GRM model increases with the inconsistency of guess parameters and error parameters, and the improved model has better stability. In the actual geography test, it can be found that the degree of guessing in the grade response items is low, but the error phenomenon is obvious, and the difference between the parameters is large, which can not be ignored.

Key words: item response theory; GRM; 4PLM

(责任编辑:冉小晓)