文章编号:1000-5862(2016)02-0140-05

4 参数 GRM 对猜测现象和失误现象的纠正

简小珠[1,2],戴海琦[2*]
(1. 井冈山大学教师教育研究中心,江西 吉安　343009;
2. 江西师范大学心理学院,江西省心理与认知科学重点实验室,江西 南昌　330022)

摘要: 将 c、γ 参数加入到 Samejima 等级反应模型中形成 4 参数等级反应模型(4 参数 GRM),该模型包含了两级记分 1、2、3、4 参数 Logistic 模型、Samejima 等级反应模型. 4 参数 GRM 适合测验中的多级和两级记分试题,也可以适合两级记分试题的猜测现象和失误现象. Samejima 等级反应模型下,被试作答的猜测现象会导致能力高估现象,失误现象会导致能力低估现象. 在 4 参数 GRM 下,被试能力高估现象和低估现象均得到了有效的纠正.

关键词: 项目反应理论; 等级反应模型; 4 参数等级反应模型; 猜测现象; 失误现象
中图分类号: B 842.1　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2016.02.06

0　测验中 IRT 模型选择的困境

在测验分析时往往会发现被试作答情况存在着猜测现象和失误现象,简小珠等[1]概述了以往对猜测现象和失误现象的研究. 在实际测验中一般包含了两级和多级记分试题,而且两级记分试题往往存在着猜测现象和失误现象. 此时选择数学模型存在如下问题:(i)若选择两级记分 4 参数 Logistic 模型,可以适合两级记分试题存在着的猜测现象和失误现象,但该模型无法适合测验中的多级记分试题;(ii)若选择 Samejima 等级反应模型,能适合多级和两级记分试题,但无法适合两级记分试题中存在着的猜测现象和失误现象. 以上是许多研究者在进行测验分析与选择 IRT 模型时经常遇到的一个困境.

猜测现象往往是指低能力被试在相对高难度试题上得分的现象,既包括被试在高难度的两级记分试题上的得分情况,又包括在高难度的多级记分试题上得到一部分得分情况. 低能力被试在相对高难度试题上得分的原因有多种:(i)被试在选择题上的猜测作答行为;(ii)低能力被试通过押题的方式压中了某道试题;(iii)低能力被试抄袭书本或其他考生的答案,或与已参加测试的考生分享试题而答对相对高难度试题[2];(iv)试题的内容较偏,有利于一部分低能力被试(命题内容偏差造成的);(v)多级记分试题的评分标准宽松,评分时"挨到边就给分"等. 这些原因都可能使得被试在难度相对较大的试题上得分,或在多级记分试题上得到一部分得分. 本文对被试在相对较难试题上得分,且期望概率小于 0.05 时而得分的现象定义为猜测现象(guessing phenomenon),用 c 参数表示[3],称为猜测参数. "相对较难"是指试题难度大于被试能力水平 2 个单位以上,即 $b-\theta>2$,此时被试答对的期望概率 P 小于 0.05.

这里沿用"猜测现象"的命名,需补充说明的是猜测现象不一定是被试猜测作答行为所导致的,前面已论述了有多种原因可能会导致猜测现象.

失误现象是指高能力被试在容易试题上失分的现象,B. D. Wright 等[4]将这种现象称为睡眠现象(sleeping phenomenon),有些研究中也称为失误现象(slipping phenomenon),K. Rulison 等[5]概述了高能力被试在容易试题上失分的多种原因. C. Schuster 等[6]论述了答案转录错误(transcription errors,从试题转录到答题卷/卡时)也可能造成高能力被试在容易试题上的失分现象. 本文将被试在相对容易的试题上答错失分,且期望概率大于 0.95 而答错或失分的现象定义为失误现象,用 γ 参数表示,γ 参数又称失误参数. 这"相对容易"是指试题难度小于被试能力水平 2 个单位以上,即 $b-\theta<-2$,此时被试答对的期望概率 P 大于 0.95.

收稿日期:2015-07-03
基金项目:江西省社会科学规划青年项目(13JY47)和江西省高校人文社会科学课题资助项目.
通信作者:戴海琦(1947-),男,上海人,教授,博士生导师,主要从事心理测量学和考试学的研究.

1　4 参数 GRM 的提出

F. Samejima[7] 提出的等级反应模型（GRM）是普通使用的一种多级记分模型. 然而该模型不含 c、γ 参数，因而无法反映测验中的猜测现象和失误现象. 陈青等[8] 将 c 参数加入到 GRM 中以反映多级记分模型下的猜测现象，本文在 Samejima 等级反应模型的基础上加入 c、γ 参数，发展出等级反应模型的改进模型. 该模型论述如下：用 c、γ 参数反映多级记分试题上的猜测现象和失误现象，并在各个得分等级上体现，而且 c、γ 参数体现在多级记分试题的各个得分等级上的项目特征函数上，c 参数的概率均匀分配到各个得分等级中的项目特征函数中. 在 GRM 加入 c、γ 参数后的数学模型具体为：满分为 m_j 的多级记分试题，被试得 0 分及 0 分以上的概率为 $P_{\alpha j,0}=1$. 被试得 1 分及 1 分以上的概率为 $P_{\alpha j,1}=c_j+(\gamma_j-c_j)(1+\exp(-1.7a_j(\theta_\alpha-b_{j,1})))^{-1}$，被试得 2 分及 2 分以上的概率为 $P_{\alpha j,2}=(m_j-1)c_j/m_j+(\gamma_j-c_j)(1+\exp(-1.7a_j(\theta_\alpha-b_{j,2})))^{-1}$，其他得分依此类推. 被试得 m_j 分的概率为 $P_{\alpha j,m_j}=c_j/m_j+(\gamma_j-c_j)(1+\exp(-1.7a_j(\theta_\alpha-b_{j,mj})))^{-1}$.

令被试得 m_j+1 分的概率为 $P_{\alpha j,m_j+1}=0$，同时在 Samejima 等级反应模型下令被试得 m_j+1 分概率为 $P_{\alpha j,m_j+1}=0$[9]. 由此进一步得出，被试恰好得 t 分的概率为 $P^*_{\alpha j,t}=P_{\alpha j,t}-P_{\alpha j,t+1}$（其中 t 为 $0,1,2\cdots,m_j$），那么被试恰得 0 分的概率为 $P^*_{\alpha j,0}=1-P_{\alpha j,1}=1-(c_j+(\gamma_j-c_j)(1+\exp(-1.7a_j(\theta_\alpha-b_{j,1})))^{-1})$. 被试恰得 1 分的概率为 $P^*_{\alpha j,1}=P_{\alpha j,1}-P_{\alpha j,2}=c_j/m_j+(\gamma_j-c_j)((1+\exp(-1.7a_j(\theta_\alpha-b_{j,1})))^{-1}-(1+\exp(-1.7a_j(\theta_\alpha-b_{j,2})))^{-1})$. 被试恰得 2 分的概率为 $P^*_{\alpha j,2}=P_{\alpha j,2}-P_{\alpha j,3}=c_j/m_j+(\gamma_j-_j)((1+\exp(-1.7a_j(\theta_\alpha-b_j,2)))^{-1}-(1+\exp(-1.7a_j\cdot(\theta_\alpha-b_{j,3})))^{-1})$，…，被试恰得 m_j 分的概率为 $P^*_{\alpha j,m_j}=c_j/m_j+(\gamma_j-c_j)(1+\exp(-1.7a_j(\theta_\alpha-b_{j,m_j})))^{-1}$. 被试在各个得分等级上的概率总和为 $P^*_{\alpha j,0}+P^*_{\alpha j,1}+P^*_{\alpha j,2}+\cdots+P^*_{\alpha j,m_j}=1$.

以上是 c、γ 参数加入 Samejima 等级反应模型后形成的新模型，称为 4 参数等级反应模型（4 参数 GRM）. 本文将 F. Samejima 提出的等级反应模型称为 GRM 原模型. 以上公式，当 $m_j=1$ 时，4 参数 GRM 就变成了两级记分 4 参数 Logistic 模型. 从数学角度，1、2、3 参数 Logistic 模型是 4 参数 Logistic 模型的特例，而 4 参数模型是 4 参数 GRM 的特例. 因此，4 参数 GRM 包含了 GRM 原模型，两级记分 1、2、3、4 参数模型. 因此，选择 4 参数 GRM 可以解决前面所论述的 IRT 模型选择困境，可以适合多级和两级记分试题，同时也可以适合两级记分试题的猜测现象和失误现象.

2　应用样例

2.1　测验设计

本文采用固定测验试题参数的设计方式，设计一个由 35 道试题组成的多级记分试题测验，满分均为 4 分，4 个难度参数，a 参数 $\log a$ 服从 $N(0,1)$，$b_1\sim b_4$ 参数服从 $N(0,1)$. 设计一个中等能力被试在此测验上作答，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，得分情况见表 1. 测验中这 35 题的试题参数、以及被试得分情况产生后，在后面的测验分析时一直为固定值. 在被试完成该测验后，再给被试额外增加一道试题的测试，并设计该试题的难度参数在 $[-3.6,3.6]$ 之间变化. 额外增加的这道试题的位置，可以在测验前 35 题中的任何一个位置插入，本文为了叙述方便，将此题的序号命名为第 36 题. 设计 21 种测试 case，每种测试情境在第 36 题安排一道难度不同的试题，这 21 种情境下的区分度都为 1.0，满分为 5 分，5 个难度参数且间隔为 0.3，其中每道试题第 3 个难度参数 b_3 从 -3 至 3 由依次递增，如表 2 所示. 被试在第 36 题的得分情况分别有 0、1、2、3、4、5 分，既包括了中等能力被试答对高难度试题的情况（猜测现象），也包括了答错容易试题的情况（失误现象）.

2.2　能力估计

在 GRM 原模型下使用 Multilog7.0 对被试作答情况进行能力估计（在 Multilog7.0 的程序可参考文献[10]），由于测验的试题参数已知，被试作答情况也已知，则属于被试能力条件估计：（i）中等能力被试完成了 35 道试题时，使用 Multilog 估计能力值为 0.003，并记为 θ'，作为被试作答第 36 题后能力估计值变化的参照点（参照值 θ'）；（ii）中等能力被试在第 36 题的 21 个测试案例的得分为 0 至 5 分时，使用 Multilog 得到能力估计值 θ，再减去参照值 θ'，得到被试能力步长 $\theta-\theta'$，本文将能力变化值 $\theta-\theta'$ 记为“能力步长”. 由被试在在测验 case_1 至 case_{21} 时的能力步长绘制成图 1 所示. 由 GRM 原模型的能力估计公式可知，被试得 0 分时使用 $1-P(b_1)$ 参与能力估计，绘图时横坐标使用 b_1；被试得 1 分时使用

$P(b_1) - P(b_2)$ 参与能力估计,绘图时横坐标使用 $(b_1 + b_2)/2$;被试得 2 分时使用 $P(b_2) - P(b_3)$ 参与能力估计,绘图时横坐标使用 $(b_1 + b_2)/2$ …被试得 5 分时使用 $P(b_5)$ 参与能力估计,绘图时使用 b_5. 根据 IRT 的基本假设[11],项目与项目之间相互独立,项目与被试作答相互独立,被试作答与被试作答之间相互独立. 因此,$\theta - \theta'$ 能力步长可认为是被试完成第 36 题后产生的.

表 1 测验前 35 道试题项目参数和得分情况

编号	区分度	难度参数				得分情况
		b_1	b_2	b_3	b_4	
1	0.85	-3.05	-0.49	-0.01	0.79	1
2	0.63	-2.48	-1.71	-0.68	0.52	2
3	0.89	-1.96	-1.61	0.73	0.97	3
4	0.55	-1.46	-0.38	0.09	0.31	2
5	1.13	-1.48	-0.19	0.55	1.17	3
6	0.51	-1.28	-0.28	-0.05	1.77	0
7	0.64	-2.20	-1.17	-0.18	0.52	4
8	0.87	-1.50	-1.29	-0.92	0.81	2
9	0.79	-2.17	-0.67	0.07	1.66	3
10	0.77	-2.15	-1.08	-0.67	0.48	2
11	0.66	-0.76	-0.21	0.48	2.43	3
12	0.58	-0.76	-0.37	0.15	1.04	2
13	1.36	-1.12	-0.57	0.09	1.03	2
14	0.99	-0.38	0.33	0.69	1.11	3
15	0.78	-0.74	-0.34	0.50	0.87	2
16	0.95	-2.73	-1.37	-0.06	0.93	3
17	0.57	-2.62	-2.11	-0.26	1.80	2
18	1.09	-1.52	-0.30	0.26	1.59	1
19	0.77	-0.79	0.29	0.54	1.84	2
20	0.89	-1.46	-0.10	0.21	1.11	3
21	1.25	-0.63	-0.35	0.49	0.95	2
22	0.63	-1.71	-0.18	0.01	0.92	2
23	0.72	-0.76	-0.47	0.24	0.90	1
24	0.61	-1.45	-1.07	0.00	0.84	2
25	0.56	-0.33	-0.06	0.45	1.15	3
26	0.76	-1.12	-0.72	-0.26	0.12	2
27	0.56	-0.76	-0.52	-0.02	0.22	2
28	0.74	-1.91	-0.53	0.64	0.89	1
29	0.77	-0.78	-0.05	0.61	0.84	2
30	0.75	-1.27	-0.08	0.39	2.80	3
31	1.28	-0.22	0.11	0.39	2.22	3
32	0.94	-0.97	-0.81	-0.24	0.29	1
33	0.52	-1.43	0.23	0.53	1.02	1
34	0.61	-0.44	0.89	1.16	1.67	3
35	0.87	-1.61	-0.70	1.31	1.61	1

注:中等能力被试在测验前 35 道试题能力估计值为 0.003.

表 2 第 36 题在 21 个测试 case 下的项目参数

case	第 36 题难度参数				
	b_1	b_2	b_3	b_4	b_5
1	-3.6	-3.3	-3.0	-2.7	-2.4
2	-3.3	-3.0	-2.7	-2.4	-2.1
3	-3.0	-2.7	-2.4	-2.1	-1.8
4	-2.7	-2.4	-2.1	-1.8	-1.5
5	-2.4	-2.1	-1.8	-1.5	-1.2
6	-2.1	-1.8	-1.5	-1.2	-0.9
7	-1.8	-1.5	-1.2	-0.9	-0.6
8	-1.5	-1.2	-0.9	-0.6	-0.3
9	-1.2	-0.9	-0.6	-0.3	0.0
10	-0.9	-0.6	-0.3	0.0	0.3
11	-0.6	-0.3	0.0	0.3	0.6
12	-0.3	0.0	0.3	0.6	0.9
13	0.0	0.3	0.6	0.9	1.2
14	0.3	0.6	0.9	1.2	1.5
15	0.6	0.9	1.2	1.5	1.8
16	0.9	1.2	1.5	1.8	2.1
17	1.2	1.5	1.8	2.1	2.4
18	1.5	1.8	2.1	2.4	2.7
19	1.8	2.1	2.4	2.7	3.0
20	2.1	2.4	2.7	3.0	3.3
21	2.4	2.7	3.0	3.3	3.6

2.3 结果与分析

2.3.1 在 GRM 旧模型下被试能力步长的结果与分析 在 GRM 原模型下,被试作答第 36 题后所得到的能力步长,将依据 2 个标准来判断被试能力步长估计情况,是否存在相对的能力高估或低估现象:(i) 被试在该得分上的正确作答概率;(ii) 试题对被试所提供的项目信息量;并参照其它测试 case 下的能力步长情况,来分析判断被试能力估计值变化是否高估或低估.

1) 得分为满分时的能力步长情况. 图 1 中得分为 5 时曲线的左半部分,从 $case_1$ 至 $case_{11}$ 的能力步长随着第 36 题的项目信息量增大而增大,即被试能力步长与项目信息量的变化趋势基本上是一致的. 得分为 5 时曲线的右半部分,$case_{21}$ 时被试答对并得 5 分($b_5 = 3.6$) 的概率 P 值为 0.000 43,$P < 0.05$,即得 5 分的概率很小的,被试得 5 分具有很大的偶然性,称为猜测现象. 从该试题的项目信息量来看,b_5 值远大于被试能力值,对被试能力估计提供的信息量很小,但从图 1 来看,相对于 $case_1 \sim case_{20}$ 的能力步长来说,$case_{21}$ 时的能力步长却是最大的,即在 $case_{21}$ 时被试能力估计值被高估了,即存在着能力高估现象. 同理,在 $case_{20}$,$case_{19}$,$case_{18}$ 时的被试能力估计值也存在不同程度的能力高估现象.

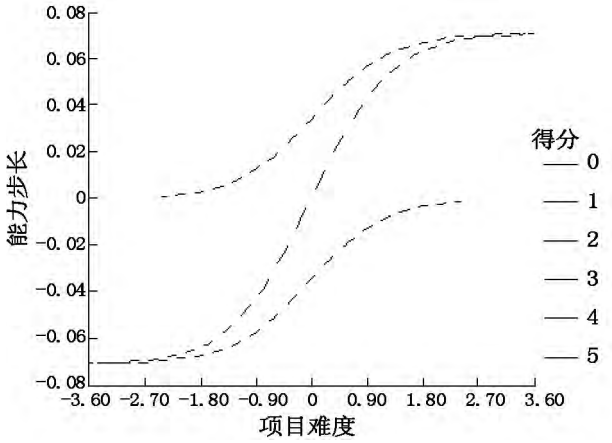
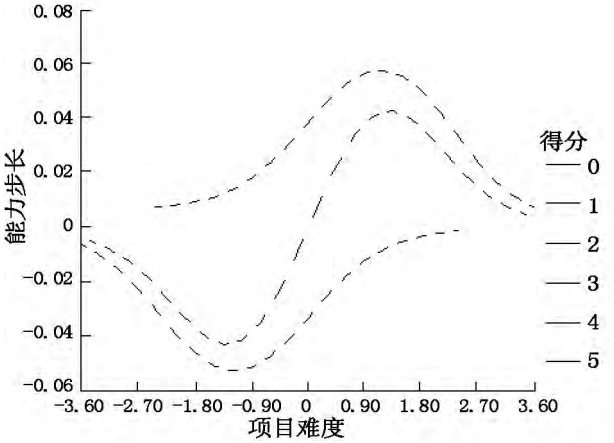


图 1　在 GRM 原模型下的能力步长

2) 得分为 0 分时的能力步长情况. 图 1 中得分为 0 的曲线左半部分，$case_1$ 时被试在第 36 题上得 1 分($b_1 = -3.6$) 的概率 P 值为 0.998 75，$P > 0.95$，该被试得 0 分，称为失误现象，被试得 0 分是具有较大的偶然性，这时应该给予被试能力步长后退的幅度应该比较小. 该题 b_1 值远小于被试能力值，对被试提供的信息量很小，但在 $case_1$ 时的负步长绝对值却是最大的，即在 $case_1$ 时被试能力估计值被低估了，即存在着被试能力低估现象. 同理，$case_2$，$case_3$，$case_4$ 时的被试能力估计值存在着不同程度的能力低估现象. 分析得分为 0 的曲线右半部分，从 $case_{12} \sim case_{21}$($b_1 = -0.6$ 至 $b_1 = -2.4$) 的能力步长绝对值的大小，随着所作答试题的项目信息量减小而减小，变化趋势基本一致.

3) 得分为中间得分时的能力步长情况. 图 1 中得分为 4 的曲线左半部分，在 case1 时被试在第 36 题上得 4 分($b_4 = -2.7$) 的概率 P 值为 0.995 5，$P > 0.95$，被试得 4 分的概率 L 较大，这时被试能力步长后退的幅度应该比较小. 由图 1 可知，得分为 4 的曲线左半部分与图 1 中得分为 0 的曲线左半部分的情况很相似，存在被试能力低估现象. 分析得分为 4 曲线的右半部分，在 $case_{21}$ 时作答第 36 题且得 4 分($b_4 = 3.3$) 的 P 值为 0.000 98，$P < 0.05$，即被试在该题得 4 分的概率较小. 由图 1 可知，得分为 4 的曲线右半部分与得分为 5 分的曲线右半部分的情况很相似，存在被试能力高估现象. 图 1 中得分为 3、2、1 的曲线，与得分为 4 的曲线有相同的趋势，存在类似的被试能力高估现象和低估现象.

2.3.2　在 4 参数 GRM 下被试能力步长的结果与分析　本文赋予 4 参数 GRM 下第 36 题的 c_j 参数为 0.10，γ_j 参数为 0.98；同时赋予测验前 35 道题的 c_j 参数都为 0，γ_j 参数都为 1. 使用 Visual Basic 语言，编写了能力条件估计程序，能力估计算法使用极大似然估计方法. 使用能力估计程序再次估计被试能力步长，如图 2. 由图 2 中被试在 $case_{21}$，$case_{20}$，$case_{19}$ 时(存在猜测现象)，被试能力高估现象得到了纠正；而被试在 $case_1$，$case_2$，$case_3$ 时(存在失误现象)，被试能力低估现象得到了有效的纠正.



注：图中得分 4 的曲线从 -2.55 至 3.45，得分 3 的曲线从 -2.85 至 3.15，得分 2 的曲线从 -3.15 至 2.85，得分 1 的曲线从 -3.45 至 2.55.

图 2　在 4 参数 GRM 下的能力步长

在纸笔测验模拟情境下，在 2 参数模型下发现被试的猜测现象会导致能力高估现象[12]，失误现象会导致能力低估现象；而两级记分 4 参数模型能同时纠正被试能力高估现象和低估现象. 本文的举例论述了 4 参数 GRM 在多级记分题下纠正能力高估现象和低估现象，可以说是 4 参数模型的进一步拓展.

陈青等[13] 运用 EM/MMLE 算法估计出了 3 参数 GRM 的项目参数，E. Loken 等[14] 使用贝叶斯方法实现了对 4 参数模型的项目参数估计. 而在 4 参数 GRM 下如何实现该模型的项目参数估计，是使用 EM/MMLE 算法，还是贝叶斯估计方法，或马尔可夫链蒙特卡洛(MCMC，) 方法进行项目参数估计，有待于进一步探讨.

3　结论

在 Samejima 等级反应模型加入了 c、γ 参数后形成 4 参数 GRM. 4 参数 GRM 包含了以往的两级记分 1、2、3、4 参数 Logistic 模型，和 Samejima 等级反应模型原模型. 4 参数 GRM 可以解决实际测量中 IRT 模型选择困境，可以适合测验中的两级和多级记分试题，同时也适合两级记分试题的猜测现象和失误现象. 在应用举例中，在 GRM 原模型下，被试作答猜测现象时会导致能力高估现象；失误现象会导致能力低估现象；在 4 参数 GRM 下，被试能力高

估现象和低估现象得到了有效的纠正.

4 参考文献

[1] 简小珠,焦璨,Reise,等. 4 参数模型对被试作答异常现象的拟合与纠正 [J]. 心理科学进展,2010,18(3):537-544.

[2] Yi Qing,Zhang Jinming,Chang Huahua. Severity of organized item theft in computerized adaptive testing: a Simulation study [J]. Applied Psychological Measurement,2008,32:543-558.

[3] Barton M A,Lord F M. An upper asymptote for the three-parameter Logistic item response model [R]. Princeton,NJ: Educational Testing Service,1981.

[4] Wright B D. Solving measurement problems with the Rasch model [J]. Journal of Educational Measurement,1977,14:97-116.

[5] Rulison K,Loken E. I've fallen and I can't get up: can high-ability students recover from early mistakes in CAT? [J]. Applied Psychological Measurement,2009,33(2):83-101.

[6] Schuster C,Yuan K. Robust estimation of latent ability in item response models [J]. Journal of Educational and behavioral Statistics,2011,36(6):720-735.

[7] Samejima F. Estimation of latent ability using a response pattern of graded scores [J]. Psychometrika Monograph Supplement,1969,34(4):100-114.

[8] 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.

[9] Embretson S E,Reise S P. Item response theory for psychologists [M]. NJ: Lawrence Erlbaum Associates,2000.

[10] 简小珠. IRT 模型中 c、γ 参数对被试能力高估和低估现象的纠正 [D]. 广州: 华南师范大学,2011.

[11] Bock R D,Aitkin M. Marginal maximun likelihood estimation of item parameters: an application of a EM algorithm [J]. Psychometrika,1981,46:179-197.

[12] 简小珠,戴海崎,彭春妹. IRT 中 Logistic 模型的 c、γ 参数对能力估计的改善 [J]. 心理学报,2007,39(4):737-746.

[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] Loken E,Rulison K L. Estimation of a four-parameter item response theory model [J]. British Journal of Mathematical and Statistical Psychology,2010,63:509-525.

Four-Parameter GRM and the Countermeasure to Sleeping and Guessing Phenomena

JIAN Xiaozhu[1,2],DAI Haiqi[2*]

(1. Education Research Centre,Jinggangshan University,Ji'an Jiangxi 343009,China;
2. School of Psychology,key Laboratory of Psychology and Cognition Science of Jiangxi Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: There exist the sleeping phenomena that the high-ability examinees make wrong response on the easy item,and the guessing phenomena that the low-ability examinees make correct response on the difficult item,in the Paper and Pencil Test and CAT. The authors add c parameter and γ parameter into the Samejima's graded-response model,and get a new model,4P-GRM. Rasch model,two-parameter logistic model,three-parameter logistic model,four-parameter logistic model all are the special case of the 4P-GRM. A Polytomous test is designed,and an average examinee have been arranged to tested on the test. After the examinee finish the test and can get the estimated ability θ using the program of MULTILOG. An extra item with difficulty parameters arranged from big to small in different test case was given to the examinee. . The author estimates the new ability θ'. The author calculates the ability steplength($\theta-\theta'$). According the outcome of the ability steplength: (i) When getting full score,the examinee will be overestimated when the examinee makes correct responses on the difficulty items. (ii) When getting zero score,the examinee will be underestimated when the examinee makes wrong responses on the easy items. (iii) When getting middle score,there is the ability overestimation phenomenon and underestimation phenomenon meanwhile. Furthermore,under the 4P-GRM,the result is that: (i) When the score is full score,the overestimation phenomenon on the low ability examinee can be rectified. (ii) When the score is zero,the underestimation phenomenon on the high ability examinee can be rectified. (iii) When the score is middle score,the overestimation phenomenon and underestimation phenomenon can also be rectified.

Key words: IRT; GRM; 4P-GRM; guessing phenomenon; sleeping phenomenon (责任编辑: 冉小晓)