四参数 Logistic 模型研究进展及其评析*

简小珠[1,2],张敏强[1],彭春妹[2]
(1.华南师范大学 心理应用研究中心,广州 510631;2.井冈山大学 教育学院,吉安 343009)

摘 要:在测验中存在着低能力被试答对高难度试题的猜测现象,和高能力被试答错容易试题的睡眠现象,此时可以使用四参数模型来分析测验数据。Barton 和 Lord 认为应用四参数模型的实践意义不大,但结论的依据不充分。近年来研究者从测验项目拟合,改善被试能力估计等方面进行了分析,认为在四参数模型下可以有效纠正被试能力高估或低估现象,认为单、两、三参数模型是四参数模型的特例,建议使用四参数模型。

关键词:IRT;四参数模型;猜测现象;睡眠现象
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2010)03-0069-05

1 四参数 Logistic 模型的测量涵义及其研究进展

1.1 测验中的猜测现象和睡眠现象以及四参数 Logistic 模型

在测验时,低能力被试凭猜测或者其它原因答对了高难度试题的现象,叫做猜测现象(guessing phenomenon)。此外在测验中还存在高能力被试答错容易试题的现象,Wright 将其称为睡眠现象[1](sleeping phenomenon)。有研究者在测验分析时,发现了测验中存在着睡眠现象。Reise 和 Waller 在分析人格测验 MMPI-2 时,发现了一些试题存在着睡眠现象的作答情况[2]。简小珠,戴海崎,彭春妹(2007)在分析测验时[3]也发现了一些试题同时存在着猜测现象和睡眠现象,或单独存在猜测现象和睡眠现象。或许有研究者会提出疑问:试题存在睡眠现象,是不是试题质量存在问题?已有研究[4]已经分析了测验中存在睡眠现象的试题,发现这些试题的测量性能都良好,认为睡眠现象与试题是否存在质量问题之间没有必然的联系。另外从题型的角度来看,如果测验中一些填空题,试题难度较大,高能力被试未必能全部答对,那么存在试题作答概率的上渐近线(即睡眠现象);而对于低能力被试来说则很难答对,猜测度就有可能为 0,这时可以用三参数模型 γ 型(含 a,b,γ 参数)来反映[4,5]。如果将此填空题改为选择题的形式,低能力被试群体就可能存在猜测现象,那么就同时存在睡眠现象和猜测现象,这时可以使用四参数模型来反映[6,7]。

1.2 四参数模型的研究进展

在过去 IRT 研究中,四参数模型有关的研究相对较少。关于四参数模型在实际测验中的应用价值存在着两种的观点,即不提倡使用和建议使用四参数模型。下面按研究文献发表的时间顺序论述。

McDonald 最早提出四参数模型[6],建议用 γ 参数反映高能被试答错容易试题的现象。Barton 和 Lord 的研究中[7],在三参数模型的基础上增加 γ 参数后,对比分析在三参数模型与四参数模型下的测验极大似然估计值变化情况,以及被试能力估计值的整体变化情况。Barton 和 Lord 通过分析,得出三个论据:1)在三参数模型增加 γ 参数后,测验极大似然估计值没有显著增加;2)被试能力估计值在整体上没有显著的变化;3)四参数模型增加了计算的复杂性,参数估计费时,因而认为使用四参数模型在实践中的应用意义不大,不提倡使用四参数模型。

在 Barton 和 Lord 之后的近二十年里,关于四参数模型的研究论文几乎没有,四参数模型只在一些教材中被提及。Hambleton 和 Swaminathan 提及了四参数模型[8],认为四参数模型在测验分析中没有实际价值。漆书青,戴海崎的编著中也提及了四参数模型[9]。而在 IRT 经典著作中,比如 Embretson 和 Reise[10]、Baker[11]、漆书青,戴海崎和丁树良[12]等著作,都没有提及四参数模型。在此期间的 BILOG、MULTILOG、LOGIST 等软件都没有四参数模型程序模块。

* **基金项目**:教育部省部共建人文社会科学重点研究基地项目(2009JJDXLX006),广东省自然科学基金项目(9151063101000002),江西省社会科学规划"十一五"学科共建项目(09JY226)。

通讯作者:张敏强,E-mail:zhangmq1117@yahoo.com.cn。

直至最近几年，研究者逐步开始关注四参数模型和测验中的睡眠现象。2003 年，Reise 和 Waller 在分析人格测验 MMPI－2 时[2]，发现了一些试题存在着上渐进线(即睡眠现象)，认为睡眠现象也是试题的一个属性，建议可以使用四参数模型来拟合测验数据。在论文的结尾部分，Reise 和 Waller 认为在四参数模型能较好兼容下渐近线不是 0，和上渐进性不是 1 的作答情况，因而可能成为未来研究的热点方向之一。2004 年，Hessen 在论述测量模型时[5]，以四参数模型为基础，认为三、两、单参数模型是四参数模型的一个特例。在随后的研究中，Hessen 把四参数模型改写成一个非参数模型的形式[13]，对非参数 IRT 模型的一些数学性质等进行了探讨，并应用于项目功能差异分析。

2005 年，戴海崎和简小珠提出使用四参数模型[14]，以纠正被试答错容易试题时能力估计偏低的问题。2007 年，简小珠，戴海崎，彭春妹设计了一个纸笔测验和中等能力被试作答情况[3]，得出在单、两参数模型下存在着第一、第二未契合现象(被试能力高估和低估现象)；在四参数模型下则可以有效纠正第一、第二未契合现象。2009 年，Rulison 和 Loken 使用 CAT 模拟的方法[15]，在四参数模型下，在测试开始阶段额外增加两道中等难度的试题并让被试答错，高能力被试的最后能力估计值能顺利到达模拟初值，没有受到答错容易试题的影响。通过一系列的 CAT 模拟分析，Rulison 和 Loken 认为使用四参数模型可以纠正被试在 CAT 测验开始时答错容易试题时造成的能力低估现象。

2 对四参数 Logistic 模型的两种观点的评析

2.1 对四参数模型的两种观点的比较

综合以上对四参数模型的研究，主要有两种观点：1)不提倡使用的观点，以 Barton 和 Lord 为代表；2)建议使用的观点，以 Reise 和 Waller、Hessen、简小珠，戴海崎，彭春妹、Rulison 和 Loken 为代表，认为四参数模型能够更好的拟合测验项目，可以有效纠正被试能力高估和低估现象，单、两、三参数模型是四参数模型的一个特例。为什么有两种截然相反的观点？论文认为是由于研究方法、研究角度不同，以及历史条件限制而造成的。不提倡使用四参数模型的研究角度是：Barton 和 Lord 在四参数模型下使用固定 c、γ 参数来估计试题的项目参数(现在是让 c、γ 参数自由估计)；而且从整份测验来分析测验极大似然值，和分析被试群体能力估计值的整体变化状况，没有具体分析被试作答情况，也没有单独分析被试作答是否存在猜测现象或睡眠现象，以及对能力估计的影响。因此 Barton 和 Lord 得出的结论具有局限性，论据粗略。建议使用四参数模型的研究角度是：1)Reise 和 Waller 从模型对测验项目拟合的角度来分析的，建议使用四参数模型来拟合测验中存在的睡眠现象。2)Hessen 是从数学函数的角度来分析模型，认为单、两、三模型是四参数模型的特例；3)简小珠，戴海崎，彭春妹是在纸笔测验形式下，额外增加一道试题的测试，单独分析被试作答出现猜测现象和睡眠现象时的能力估计情况。4)Rulison 和 Loken 使用 CAT 模拟方法，额外增加两道试题的测试(包括被试作答的猜测现象和睡眠现象)，分析被试能力估计模拟返真情况。由此可知，建议使用四参数模型的研究中，研究角度较为全面，或从项目拟合的角度，或从数学函数的角度，或从额外增加猜测现象和睡眠现象来分析被试能力估计值的角度(包括纸笔测验形式和 CAT 测验模拟方法)。

2.2 对 Barton 和 Lord 的研究报告的探讨分析

只有 Barton 和 Lord 的研究认为四参数模型的实践意义不大，不提倡使用四参数模型。由于 Lord 是 IRT 的奠基人之一，而且 Barton 和 Lord 的论文是 ETS 的研究报告，因此论文观点的影响力较为深远，影响后来研究者对四参数模型的观点。然而，Barton 和 Lord 的研究结论是否有确切的论据？研究方法是否恰当？下面对 Barton 和 Lord 的三个主要论据进行分析。

论据一，总测验的极大似然值没有显著增加。此论据受到以下几个方面的质疑：首先，Barton 和 Lord 固定 c、γ 参数进行项目参数估计，参数估计方法落后。在分析四批测验数据时，Barton 和 Lord 预先固定所有试题 c 参数，γ 参数固定为 0.99 或 0.98，再估计四参数模型的 a、b 参数，计算测验极大似然估计值。由于 γ 参数仅由 1 减小至 0.99，测验极大似然估计值的变化幅度很小，所以测验极大似然估计值的变化也往往是相应很小了。而目前的项目参数估计软件往往是让 c、γ 参数自由估计，c、γ 参数浮动范围会比较大，那么总测验的极大似然值就可能会有较大的变化。其次，测验极大似然值公式为 $\ln L=\sum_{a=1}^{N}\sum_{j=1}^{m}[u_{aj}\cdot\ln P_{aj}-(1-u_{aj})\cdot\ln Qu_{aj}]$，测验极大似然值与被试数量有关，也与试题数量有关，因此测验极大似然值变化是否显著，需要进行卡

方检验，才能进一步得出是否存在显著性结论，然而 Barton 和 Lord 没有进行卡方检验。第三，在分析数学模型是否适合测验数据时，测验极大似然值仅是其中一个参考指标，而且目前依据的参考指标主要是项目拟合指数、被试残差等。第四，Barton 和 Lord 仅由 4 批实测数据得出的结论，是否具有广泛的代表性，这也值得质疑。

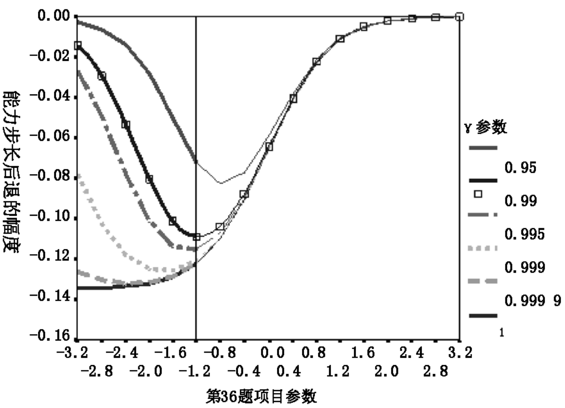

图 1　中等能力被试答错第 36 题后的能力步长曲线

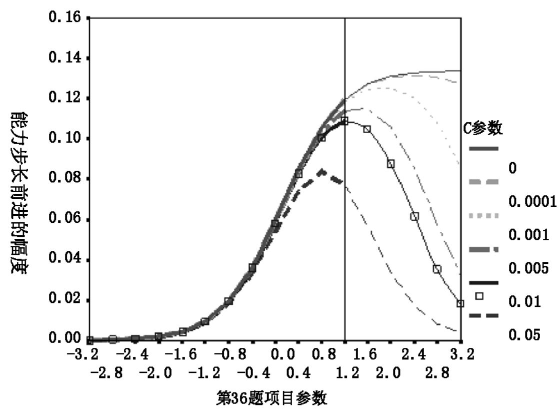

图 2　中等能力被试答对第 36 题后的能力步长曲线

注：图 1、图 2 的数据来源于简小珠硕士论文（详见参考文献 4）。由该论文中表 9 的数据绘制成图 1，由表 8 的数据绘制成图 2。其中，图 1 中增加 γ 参数为 0.995、0.999、0.9999 时的能力步长曲线；图 2 中增加 c 参数为 0.01、0.005、0.001、0.0001 时的能力步长曲线。

论据二，被试能力估计值在被试整体上没有显著的改变。Barton 和 Lord 使用散点图表示三、四参数模型下的能力估计值，发现被试群体在测验上的能力估计没有显著改变。而最近研究认为 γ 参数在 0.99 或 0.98 时能够有效纠正高能力被试答错容易试题时的能力低估现象[3,15]。这个观点是不是矛盾？文章认为，由于 Barton 和 Lord 在散点图形上是从被试群体的被试估计值的整体变化情况来分析的；而简小珠等的研究[3]是从单个能力被试角度来分析能力估计值相对变化情况，Rulison 和 Loken 也是仅从单个高能力被试的角度[15]（多次模拟的能力估计值平均值，使用 Bias 和 RMSE 指标），分析高能力被试答错容易试题后能力估计值的变化情况。

为什么分析全部被试与单独分析高能力被试的能力估计值，会有不同的结果？文章根据已有的研究[4]，从 c、γ 参数对被试能力估计值的影响的角度来分析。

由简小珠（2006）的论文中表格 9 可知[4]，当被试答对试题时，该试题的 γ 参数大小对被试能力估计值影响很小，能力估计值的变化都在 0.001 以下。由图 1 可知：在被试答错试题时，γ 参数的大小对被试能力估计的影响，因所答错试题的难度大小不同而不同：1）先分析在横坐标 $b=-1.2$ 的右边部分。在 $b=-1.2$ 位置，$\gamma=0.99$ 的曲线与 $\gamma=1$ 的曲线相差很小，相差只有 0.015，即被试答错试题而且试题难度 $b-\theta>-1.2$ 时，γ 参数由 1 减小至 0.99 时对被试能力估计值的影响很小。2）再分析在横坐标 $b=1.2$ 的左边部分。随着被试答错容易试题的难度减小，$\gamma=0.99$ 的曲线与 $\gamma=1$ 的曲线相差的距离越来越大，当 $b=-3.2$ 时，能力步长相差为 0.13 左右，即被试答错容易试题而且试题难度 $b-\theta\leqslant-1.2$ 时，γ 参数为 0.99 时能有效纠正此时的能力高估现象。而且由图 1 可知，γ 参数在 0.95 至 0.995 这个区间，都能够有效的纠正被试能力估计值高估现象。综上所述，γ 参数对改善被试能力估计值的情况具有较强的针对性，仅对被试答错容易试题而且试题难度 $b-\theta\leqslant-1.2$ 时出现的能力低估现象具有较好的纠正作用；而对于被试答对试题的作答情况，或者被试答错试题而且试题难度 $b-\theta>-1.2$ 时的作答情况，γ 参数对被试能力估计值的影响很小。同理，由图 2 可得 c 参数对改善被试能力估计值的情况也具有较强的针对性，仅对被试答对高难度试题而且试题难度 $b-\theta\geqslant-1.2$ 时出现的能力高估现象具有较好的纠正作用；而对于被试答错试题的作答情况，或者被试答对试题而且试题难度 $b-\theta<-1.2$ 时的作答情况，c 参数对被试能力估计值的影响很小。

假设某一被试即使答错了 6 道试题而且试题难度 $b-\theta=-1.2$ 时，那么该被试能力步长后退的幅度，可以根据公式来计算[4]：作答相同的 k 题后的能力步长 $\approx k\times$ 作答 1 题后的能力步长（$k\leqslant6$）。被试能力步长将后退的步长幅度约为 0.09 左右，可见被试答错 6 道试题而且试题难度 $b-\theta=-1.2$ 时，对

被试能力估计值的影响相对很小。而且，只有被试答错了 $b-\theta<-1.2$ 的试题时，对被试能力估计值影响才会逐渐增大。Barton 和 Lord 的研究是分析全部被试群体的情况，在增加 γ 参数为 0.99 后，大多数被试的能力估计值都变化不显著，这容易掩盖其中一小部分高能力被试和中等能力被试群体，答错了容易试题而造成的能力低估现象。由 Barton 和 Lord 的论文中的表 1 和图 1 可知，当 γ 参数从 1 变为 0.99 时，还是有一小部分被试的能力估计值发生了较大的变化。遗憾的是，Barton 和 Lord 没有进一步探讨这部分被试的作答中是否存在睡眠现象（答错容易试题 $b-\theta<-1.2$ 的情况）。由被试群体的能力估计值整体变化不大，从而认为四参数模型对被试能力估计的影响作用不大，Barton 和 Lord 得出此结论的依据不够充分。

论据三，是四参数模型增加了计算的复杂而且费时。1981 年项目参数估计方法落后，计算机运行速度很慢，四参数模型的项目参数估计是一件非常困难的事情。目前，项目参数估计方法已经有了 MMLE/EM 算法，MCMC 方法，最新的 IRT 软件 WINSTEPS 已经包含了四参数模型的项目参数估计模块[16]，实现了四参数模型的项目参数估计。

综上所述，Barton 和 Lord 的研究存在以下不足：1）项目参数估计方法落后，使用固定 c 参数和 γ 参数的方法估计四参数模型的项目参数，对测验极大似然估计值的差异没有进行显著性检验；2）仅对整个被试群体做了粗略的分析，没有单独的分析高能力被试群体的能力估计值情况，也没有具体分析高能力被试的作答情况。因而，Barton 和 Lord 不提倡使用四参数模型的依据是不充分的。

3 小结与展望

综合对四参数模型的以往研究，特别是对 Barton 和 Lord 研究报告的分析与探讨，可以得到以下结论：1）四参数模型可以提高模型与测验项目拟合性能；2）四参数模型下可以纠正低能力被试答对高难度试题时的能力高估现象，和高能力被试答错容易试题时的能力低估现象；3）单、双、三参数模型是四参数模型的一个特例；4）由于 Barton 和 Lord 的研究方法不恰当，不提倡使用四参数模型的结论是不恰当的。总之，在实际测验分析中，可以应用四参数模型来反映并分析测验数据中的猜测现象和睡眠现象。WINSTEPS 软件中包含了四参数模型的程序模块，为四参数模型在心理与教育测量中的广泛应用奠定了基础。

四参数模型在其他学科的应用也较为广泛，比如：1）在生物学与医学中，在研究生物族群的数量发展规律时，往往是采用四参数模型来分析测量数据[17]；2）WHO 组织认为在分析医学测量数据时[18]，建议使用十种数学模型进行分析（其中包括四参数模型）。当然在不同学科的情况下，四参数模型所反映的测量意义可能不同，只要数学模型能更有效的描述客观世界，那就应当被应用到实际测量分析中。另外，为进一步提高测量数据的拟合性，有研究者提出五参数 Logistic 模型来分析测量数据[19]，并编写了相应的软件。

参考文献

1 Wright B D. Solving measurement problems with the research model. Journal of Educational Measurement，1977，14：97－116.

2 Reise S P，Waller N G. How many IRT parameters does it take to model psychopathology items? Psychological Methods，2003，8(2)：164－184.

3 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. 心理学报，2007，39(4)：737－746.

4 简小珠. Logistic 模型 c、γ 参数对被试作答的拟合能力. 硕士论文. 南昌：江西师范大学，2006.

5 Hessen D J. A new class of parametric IRT models for dichotomous item scores. Journal of Applied Measurement，2004，5(4)：385－397.

6 McDonald R P. Non－linear factor analysis. Psychometric Monographs，1967：15.

7 Barton M A，Lord F M. An upper asymptote for the three－parameter Logistic item response model. In：Research Bulletin. Princeton，NJ：Educational Testing Service，1981：81－20.

8 Hambleton R K，Swaminathan H. Item response theory：Principles and applications. Boston：Kluwer－Nijhoff，1985：48－49.

9 漆书青，戴海崎. 项目反应理论及其应用研究. 南昌：江西高校出版社，1992.

10 Embretson S E，Reise S P. Item response theory for psychologists. Mahwah：Lawrence Erlbaum Associates，Inc，2000.

11 Baker F B. Item response theory：parameter estimation techniques. 2nd eds. New York：Marcel Dekker，nc. 2004.

12 漆书青，戴海崎，丁树良. 现代教育与心理测量学原理. 北京：高等教育出版社，2002.

13 Hessen D J. Constant latent odds－ratios models and the Mantel－Haenszel null hypothesis. Psychometrika，2005，70(3)：497－516.

14 戴海崎，简小珠. 被试作答的偶然性对 IRT 能力估计的影响研究. 心理科学，2005，28(6)：1433－1436.

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

16 Linacre J M. A user's guide to winsteps ministep rasch－model computer programs. Retrieved from：http://199.236.93.8/winman/index.htm? asymptote.htm. 2009－6－28.

17 吴德福. 四参数 Logistic 函数的简化求解. 中华核医学杂志，1990，10(1)：48－50.

18 Edwards P S. The who immunoassay program，Version 5.2. London：Middlesex Hospital，1985.

19 Brendan. Retrieved from：http://www.brendan.com/5pl.html. 2009－7－03

The Advances of Four－Parameter Logistic Model and Its Comments

Jian Xiaozhu[1,2]，Zhang Minqiang[1]，Peng Chunmei[2]

（1. Center for Psychological Application，South China Normal University，Guangzhou 510631；

2. Education School，Jinggangshan University，Ji'an 343009）

Abstract：In the paper－pencil test and CAT，there exist the phenomena that the high－ability examinee makes wrong response on the easy item(guessing)，and that the low－ability examinee makes correct response on the difficult item(sleeping). Many researches have been made on the topic of the guessing or ceiling phenomenon under four－parameter Logistic model(4PM). Barton & Lord didn't agree to urge the use of 4PM. In recent years，however，many researchers have made the researches and proposed the use of 4PM. It has been demonstrated that：(1)4PM can improve the good－fitness of mode－data；(2)4PM can rectify the underestimation phenomenon when there exist the sleeping phenomenon，and 4PM can rectify the overestimation phenomenon when there exist the guessing phenomenon；(3)one－parameter，two－parameter and three－parameter Logistic model are the special case of 4PM. The new program WINSTEPS has published. The user can estimate the item parameter under 4PM using the WINSTEPS. WINSTEPS will help to popularize the 4PM.

Key words：IRT；Four－Parameter Logistic Model；guessing；sleeping