现代心理测量理论与技术丛书

京师心理研究 JINGSHI XINLI YANJU

项目反应理论新进展专题研究

丁树良 罗 芬 涂冬波 等 著

Some Topics of
Advanced Development in
Item Response Theory

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

现代心理测量理论与技术丛书

项目反应理论新进展专题研究

丁树良　罗　芬　涂冬波　等 著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

项目反应理论新进展专题研究／丁树良，罗芬，涂冬波等著.—北京：北京师范大学出版社，2012.9
（现代心理测量理论与技术丛书）
ISBN 978-7-303-14889-9

Ⅰ.①项… Ⅱ.①丁…②罗…③涂… Ⅲ.①心理测验-研究 Ⅳ.①B841.7

中国版本图书馆CIP数据核字（2012）第141119号

营销中心电话 010-58802755 58800035
北师大出版社职业教育分社网 http://zjfs.bnup.com.cn
电子信箱 bsdzyjy@126.com

出版发行：北京师范大学出版社 www.bnup.com.cn
北京新街口外大街19号
邮政编码：100875
印 刷：北京市易丰印刷有限责任公司
经 销：全国新华书店
开 本：170 mm × 240 mm
印 张：15
字 数：275千字
版 次：2012年9月第1版
印 次：2012年9月第1次印刷
定 价：32.00元

策划编辑：陈红艳 责任编辑：陈红艳
美术编辑：高 霞 装帧设计：高 霞
责任校对：李 菡 责任印制：孙文凯

“现代心理测量理论与技术丛书”编委会

序

心理与教育测量是评价个体心理特质发展水平状态的重要手段。以项目反应理论为代表的现代测量理论的发展，为指导心理与教育测量研究及实践提供了强大的理论与技术支持。在项目反应理论基础上的参数估计、等值、信息量评价、项目功能差异甄别等技术保证了测验开发更加科学，而计算机化自适应测验的理论和技术，更为测验的发展提供了一个广阔而光明的前景。最近十几年蓬勃兴起的认知诊断评价理论，则将测量理论与技术推向了更加精细化的评价水平上。

不过，在国内的心理与教育测量实践中，大多数研究和实践仍然主要是基于经典测量理论基础上的。许多试图使用现代测量理论为指导的研究者由于担心无法很好地把握该理论的原理和方法而望而却步。为了让现代测量理论的发展研究成果能够更多地用于指导研究和实践工作，测量学研究者应该做出更多的努力。

江西师范大学心理与教育测量学研究团队在戴海琦、丁树良、漆书青等教授的带领下，从20世纪80年代初开始对现代测量理论进行深入研究，取得了许多理论和实践的研究成果，研究团队也进一步发展壮大。随着研究的深入以及研究领域的进一步拓展，加之现代测量理论受到越来越多研究者的关注，江西师范大学现代测量理论研究团队顺应形势和发展需要，基于自身近30年的理论研究和实践积累，出版一套关于现代心理测量理论与技术的丛书，这是心理与教育研究领域的一件有益之事，也必将进一步推动心理与教育测量理论与技术在中国的发展。这套丛书包括项目反应理论和认知诊断评价理论，具体内容从理论原理、技术方法到应用实践技术，内容全面、结构完整，是读者全面深入了解和掌握现代测量理论与技术很好的参考书。

值此丛书即将付梓出版之际，作为与江西师范大学心理与教育测量学研究团队交流合作多年的同行，我备感欣慰，特作此短序以示祝贺，并希望他们在今后取得更多的研究成果和更大的发展。

张华华

2012年3月16日

于美国伊利诺伊大学香槟校区

前 言

项目反应理论(IRT)是一种现代教育与心理测量学理论，它能够解决一些经典测量理论(CTT)解决不了的问题，因此自其问世以来，便备受关注，显示勃勃生机。但是学习 IRT 需要比学习 CTT 更多的数理知识，这使得不少学习者望而却步。

本书是 IRT 新进展专题研究。本书将我们在近 30 年学习和使用 IRT 的过程中的一些体会，分成几个专题进行介绍。一般而言，专题研究与研究者的兴趣爱好、学术背景、社会需求关系密切。

本专题研究展现了多人的研究成果，由于兴趣不同、背景各异、社会需求发生变化，所以本书内容可能比较分散，这一点在本书的目录上也有反映。但是本书大体上还是可以分成理论和模型、参数估计、IRT 的应用这三个方面。具体而言，对现有的且对某些学者较陌生的模型、方法的介绍，包括多维 *IRT* 模型、拓广等级展开模型、多侧面 *Rasch* 模型、群体水平项目反应理论、*MCMC* 估计方法等；新模型、新方法的开发和应用：包括带有猜测参数的等级反应模型(3*PLM-GRM*)、多题多做模型、新的参数估计方法(如双重两步迭代估计)等；还有 IRT 的典型应用——计算机化自适应测验(CAT)、题库建设(包括我们认为题库建设的新思路，即在线校准)、项目反应模型和 Q 矩阵相结合开发认知诊断模型。众所周知，如上三方面的内容不能截然分开，不用说新模型的开发可能用新的参数估计方法，纵使是对现有模型的介绍中也有新的研究和新的应用，可谓是旧瓶装新酒；当然应用中又涉及不少新的方法。

IRT 博大精深，囿于我们的见识和精力，我们不仅涉及的面有限，而且体会难免粗浅，之所以鼓起勇气整理出来奉献给读者，一是抛砖引玉，欲就教于方家；二是我们是过来人，在 IRT 这条川流不息、咆哮奔腾的大河里呛过水，知道学习过程中的一些暗礁险滩，希望能够为读者清除一些障碍，为传播 IRT 尽一份绵薄之力。

本书一共有十章，有的章节由几位作者撰写，甚至同一个模型(3*PLM-GRM*)给出两种不同的研究方式、不同的参数估计方法，目的是让读者和作者一样从不同角度看待同一个问题；当然对同一个模型的未知参数使用不同的估计方法，并且进行比较，这其实是很通常的做法。有不少同事参与了本书的编写工作，依照各自负责的章节顺序排列如下：

毛萌萌、丁树良撰写第一章第一节；

丁树良撰写第一章第二节；
罗芬撰写第二章第一节；
涂冬波撰写第二章第二节；
罗芬、丁树良撰写第二章第三、四节；
黄华彩、罗芬撰写第二章第五节；
涂冬波撰写第三章第一节；
丁树良、罗芬、朱玮、戴海琦撰写第三章第二节；
蔡艳撰写第四章；
邓稳根撰写第五章；
涂冬波撰写第六章；
汪文义、刘铁川撰写第七章；
熊建华撰写第八章第一、二、三节；
游晓锋、丁树良、刘红云撰写第八章第四节；
罗芬、丁树良、王晓庆撰写第九章第一、三、四节；
程小扬、丁树良撰写第九章第二节；
汪文义撰写第十章；
熊建华对第八章统稿，罗芬对第九章进行统稿；
熊建华、罗芬对本书进行初步编辑。

如上所示，本书凝集许多人员的心血，集思广益比一两个人思路开阔，这是优点；但是各人表述方法有所差异，但愿这一点不至于影响对本书的阅读兴趣。

全书的目录提纲经过戴海琦先生的修改，最后由丁树良统稿。统稿过程中得到了戴海琦先生、熊建华老师、罗芬老师和北京师范大学的刘玥老师的大力帮助，还得到了江西师范大学心理学院、计算机信息工程学院各位领导及老师的支持，在此一并表示感谢！

限于时间及能力，本书仍有许多不足之处，恳请广大读者批评指正（邮箱：ding06026@163.com）。

编者
2012年5月4日
于江西师范大学心理统计与测量中心

目　录

第一章　项目反应理论简介

制约人的行为的心理品质称为心理特质，到今天还不能证明这些心理特质存在于人类的物理或生理结构之中，因此它们又称为潜在0－1特质。心理测量的首要任务就是要针对人的某种行为探清制约这种行为的潜在特质结构，包括它的结构成分和各种成分的性质，然后将其量表化；其次就是要根据被试的行为判定被试在这些潜在特质量表上的确切位置；最后根据被试的潜在特质水平预测被试的行为。由于特质的潜在性，我们不能像测量物体长度或重量那样对潜在特质进行直接度量，即我们无法直接完成探清潜在特质结构的任务，只能借助于一些可观察的间接变量来鉴别和定义这些潜在特质，包括它的类别和作用大小等。为了方便将潜在特质量表化，心理测量学定义了潜在特质空间(latent trait space)，表示对于人的某一种行为起制约作用的若干潜在特质的集合，记为 $\theta=(\theta_1, \theta_2, \theta_3, \cdots, \theta_t, \cdots)$，其中 θ_t 为空间中的第 t 个特质，该特质空间为 k 维。

如果某个特质空间已经包含了制约某一行为的所有潜在特质，就称为全特质空间。若是单维的全特质空间，那么该任务行为就具有单维性。

测量就是依据一定的法则使用工具对事物的特征进行定量描述的过程。通过获取人们对一系列测验任务(项目)进行反应的信息(比如用分数评估所完成任务，分数就是所获得的信息)，然后对这种信息加以处理，以估计做出这种反应的被试在潜在特质空间的位置。

潜在特质理论实质上是一切心理测量理论的基础，只是在应用潜在特质理论时，各种测验理论的角度和起点不同，精细和粗放有别，对潜在特质探明的程度各异而已。

所有测量都存在误差，为了得到较为准确的测量结果，就应该控制误差。一般来说，对测量误差的控制有三种方法：配对或标准化、随机化、统计调整。配对或标准化技术的应用使得误差变量的影响不能解释测量结果的差异，随机化技术的应用可使误差变量的影响不能在测量结果上形成系统误差，统计调整技术建立在数学模型基础上，将误差变量的影响参数化，从而在测量中调整参数估计值，减少误差变量的影响。

第一节　经典测量理论与项目反应理论

一、经典测验理论及其局限性

经典测验理论，也称真分数理论(True Score Theory)，模型始于斯皮尔曼(Spearman，1904)，而由洛德和诺维克(Lord，Novick，1966)做出了最终的公理化形式。CTT 假定在测验水平上，观察得分等于真分数(即特质分数)与误差之和($X=T+E$)，对于测验而言，由此得到两个重要的推论：

(1)真分数等于观察分数的平均数；

(2)在一组测量分数中，观察分数的变异数(即方差)等于真分数的变异数与误差分数的变异数之和。

经典测量理论在真分数理论假设的基石上构建起了它的理论大厦，主要包括信度(Reliability)、效度(Validity)、项目分析(Item Analysis)、常模(Norm)、标准化(Standardization)等基本概念。经典测验理论对心理与教育测量理论和实践的贡献都是巨大的，并且还将在测量实践中继续发挥它的作用。但是它的理论体系先天不足，从而主要存在以下局限性：

(1)测验结果拓广的有限性：经典测量理论主要应用配对或标准化技术和随机化技术控制测量误差。然而，使用配对或标准技术的测量结果仅仅能在相同的测量条件下成立，却不能将其拓展到非标准化的环境中去，使得测量结果的应用受到很大的限制。

(2)测量分数的测验依赖性：经典测量理论应用标准化技术控制误差，但其标准化的对象是测验的各种外部变量，对测验项目的“性质”却没有也不可能实现标准化。这就使得测量相同能力的两个不同测验上的分数，即使其测量的外部条件都已标准化，其结果一般都是不可比的。这造成了测验分数对具体测验的依赖性，迫使经典测验理论要么使用统一试卷，要么使用实际上并不平行的所谓“平行试卷”。这样做，不是给实际操作带来困难，就是给结果解释带来较大的误差。

(3)统计量的样本依赖性：经典测量理论以测验的信度、效度和测验项目的难度、区分度等参数来刻画测量各方面的特性。这些参数的估计对样本的依赖性是很大的，经典测量理论总是强调样本对总体的代表性，实际上即使经典理论应用随机抽样，偏差总是存在的，有时还会很大；更何况，受客观条件的限制，有时还难以做到真正的随机抽样。样本依赖性使得估计出来的参数对测验的分析的限价非常有限。

(4)信度估计的不精确性：测量的重要目标就是降低测量误差，提高测量的精度。在经典测量理论中，信度被定义为真分数的变异在总变异(观测分数)中所占的比率。可见经典测量理论对所有被试的信度估计值均相

同。经典测量理论中假定对不同能力水平的被试来说，测量的误差是相同的。事实上，一份测验只有在施测于能力水平与测验难度相当的被试时容易获得比较高的测量精度。当测验施测于能力水平高于(或低于)测验难度的被试时就容易产生较大的测量误差。而且测量误差值会随着被试水平与测验难度距离的增加而变大。

另外，真分数的方差是无法求得的，误差的方差也无法计算。为了估计信度，经典测量理论就提出了平行测验的概念，并在此基础上推演出了若干个信度估计公式。但是严格的平行测验是不存在的，等价测验也很难获得，在此基础上估计的测验信度很难达到比较高的精确程度。

(5)能力量表与难度量表的不一致性：经典测量理论的项目难度参数是该项目的通过率或平均得分。它随测验施测的被试群体的变化而变化，参照组是考生集合；被试能力参数是由被试在这个测验上所得真分数表达的，参照组是项目集合；所以项目难度参数与被试能力参数定义在毫不相干的两个度量系统上，它们之间无法进行比较。因此，一份所有项目参数均已知的测验施测于一个能力水平参数已知的被试，其在各个项目上的反应情况如何，结果分数将会是多少以及测量的误差将会有多大，根据经典测量理论都是事先无法估计的，即不能进行预报。这表明经典测量理论的参数指标对测验编制活动的指导价值是相当有限的。

经典测验理论的某些局限性在概括力理论(Generalizability Theory，GT)中得到一些改善，而大多数依然存在。原因是，概括力理论与经典测验理论同属于随机抽样理论，经典测量理论的项目参数系统并没有改善，因此经典测验理论的主要局限性依然存在。

另外，还有人认为经典测量理论的最大局限是，它只能处理两级评分，对于多级评分项目，经典测量理论实际上是主观给分(Bock & Moustaki，2007)。

二、项目反应理论

与经典测量理论在测验水平上进行分析，即根据被试在整个测验的反应行为确定被试在潜在特质空间的位置不同，在心理测量理论家族中，其中有一部分理论是研究测验中单个项目上的反应行为与被试在潜在特质空间位置的关系。这些心理测量理论称为项目反应理论(Item Response Theory，IRT)。IRT 是在分析与克服经典测量理论的局限性的基础上发展起来的一种新兴的心理与教育测量理论，它不仅在项目水平(而不是测验水平)上对被试反应与被试潜在特质之间的关系进行描述，从而比经典测量理论更加深入细致；而且这种描述的数学形式也更加丰富和细腻，使得许多数学工具都可以应用，从而得到更加广泛深入的结论。项目反应理论的发展首先建立在潜在特质理论的基础之上，其主要内容就是揭示被试在测

验项目上的反应行为(作答)与测验所测的被试潜在特质(即制约人的行为的心理品质)之间的关系，被试在这些项目上的反应过程是一种非线性的单调过程，这种关系的函数描述称为项目特征曲线(Item Characteristic Curve，ICC)。项目特征曲线的形态确定后，再配上使项目特征曲线能够成立或者使项目特征曲线能够确定(比如未知参数能够估计)的基本假设后，则构成项目反应理论模型。

三、经典测验理论与项目反应理论的对比

(一)项目反应理论的一些假设

与经典测量理论相比，大部分 IRT 是建立在强假设基础上，主要有特质空间的单维性假设(被测量的测验结果只取决于一种能力，其他能力的影响都可以忽略)、局部独立性假设(已知能力和项目参数条件下，假设被试答对某一项目的概率独立于其他项目)、项目特征曲线假设(被试对某项目的正确反应概率与其能力之间的关系可以用一个关于能力单调上升的函数来表示)。

当然随着研究的深入与拓展，出现了许多其他的模型，它们将这些假设进行修改，以便更好地描写客观现象。比如，出现了多维项目反应理论(Multidimensional Item Response Theory，MIRT)，它就不满足单维性假设；出现了题组反应理论(Testlet Response Theory，TRT)，它就打破了局部独立性假设；出现了展开模型(Unfolding Model，UM)，它就不要求单调性假设。这些新兴模型的建立，使项目反应理论的应用范围更加广阔，这也使得项目反应理论比经典测量理论更加灵活。

(二)项目反应理论的优点

与经典测量理论相比，项目反应理论具有以下优点：

(1)项目反应理论深入测验的微观领域，将被试特质水平与被试在项目上的行为关联起来并且将其参数化、模型化。若模型成立并且项目参数均已知，可生成独立于测验项目性质的特质水平测量结果，这是项目反应理论建立项目反应模型的最大优点，也就是通常所说的被试能力估计不依赖于测验项目的特殊选择。

(2)IRT 模型项目参数的估计独立于被试样本。项目特征曲线是被试作答正确的概率对其潜在特质水平的回归。而回归曲线并不依赖于回归变量本身的频数分布。所以，在求取项目特征曲线的各种参数时，由于回归函数的形状、位置都不依赖于被试的分布，所以它的参数，包括难度、区分度和猜测参数也都是不变的。

(3)能力参数与项目难度参数的配套性，亦即项目难度参数与能力参数是定义在同一个量表上的。这样，对一个能力参数已知的被试，配给一个项目参数已知的试题，便可以通过模型预测被试正确作答的概率。如果

估出被试的能力，我们可以在题库中选出难度与其能力相当的项目进行新一轮的测试，使得能力估计更为精确。这一特点为自适应测评奠定了基础。

(4)通过模型测得的被试能力水平，可以估计每一个被试的测量误差。项目反应理论中给出 Fisher 信息量的概念，通过 Fisher 信息量，可以度量对每一个被试的测量误差，它包括项目信息量和测验信息量，并且在局部独立性假设之下，测验信息量等于项目信息量之和，后来张华华和殷子良(Chang & Ying，1996)又引入 KL (Kullback-Leibler)信息量的概念，他们证明 Fisher 信息量是局部信息量(local information)，而 KL 信息量才是整体信息量，或者称为全局信息量(universal information)。

项目反应理论除了在自身的基本理论系统、模型种类、数据模型拟合检验方法和参数估计方法的发展之外，在实际应用方面也有很大成就，主要表现在三个方面：一是指导测验编制，通过建立测验信息目标函数来评估测验的误差，从根本上改善了测验编制的指导思想。在此基础上发展起了多种测验编制方法，特别是对目标参照性测验编制的指导，而此前经典测验理论对此无能为力；二是计算化自适应测验(Computerized Adaptive Testing，CAT)的兴起；三是项目反应理论认知测量模型的出现，将测量导向与认知心理学相结合的方向，应用测量模型直接探索人的认知结构。

虽然项目反应理论较之经典测量理论更显复杂深奥，且更趋数学化，但随着计算机技术的发展越来越普及，使得计算上便捷许多，因此许多大型测验的编制已逐渐采用项目反应理论。目前一些大型的考试 TOEEL，GRE 等，都相继采用了以项目反应理论为基础的计算机化自适应测验(CAT)，一些传统的智力测验如比奈测验、韦氏智力测验、瑞文测验等也使用项目反应理论作为分析的理论依据。国外一些权威性的测评机构如“美国教育研究联合会(AERA)”和“国家教育测量委员会(NCME)”都已相继开展了对项目反应理论的专题讨论，在理论上证实了其优越性。

有人认为项目反应理论的特色是参数不变性和信息函数(余宁民，2009，pp. 56－57)，鉴于信息函数的重要性，在本节末尾我们给出信息函数的一般计算公式。

IRT 与 CTT 之间的一个重大区别是 CTT 对所有被试的测量误差都用整个测验的误差来概括。这显然是十分粗糙的做法。IRT 认为不同能力的被试的测量误差不同。给定一个能力的估计值，IRT 用测验的 Fisher 信息函数在这个能力估计值处的值(Fisher 信息量)的倒数近似表示测量误差的平方。在数量统计中，能力的最大似然估计的分布渐近服从正态分布，其均值是能力真值，其方差是测验信息量的倒数。这里所说的“渐近”服从正态分布是指当测验长度 L 趋于无穷时，以正态分布为其极限分布。

IRT 有一个基本假设，即给定能力条件下，被试对项目的反应是独立的，这是一种条件独立性，称为局部独立性(Local Independency，LI)。在局部独立性定义下，测验信息量是构成测验的所有项目信息量之和。

设被试 i 对测验的反应向量(得分向量)为 $X_i=(x_{i1}，\cdots，x_{im})$，对于0－1评分情形，在局部独立性假设之下，其对数似然函数为

$$l_i=\ln L(x_i)=\sum_{j=1}^{m}[x_{ij}\ln P_{ij}+(1-x_{ij})\ln Q_{ij}]$$

Fisher 信息量定义为 l_i 对 θ_i 的二阶偏导数的期望值的相反数，即

$$I(\theta_i)=-E\frac{\partial^2\ln l_i}{\partial\theta_i^2}$$

由于

$$\frac{\partial\ln l_i}{\partial\theta_i}=\sum_{j=1}^{m}\left(\frac{x_{ij}}{P_{ij}}-\frac{1-x_{ij}}{Q_{ij}}\right)\frac{\partial P_{ij}}{\partial\theta_i}=\sum_{j=1}^{m}\frac{x_{ij}-P_{ij}}{P_{ij}Q_{ij}}\frac{\partial P_{ij}}{\partial\theta_i}$$

$$\frac{\partial^2\ln l_i}{\partial\theta_i^2}=\sum_{j=1}^{m}\left\{\left[\frac{\partial}{\partial\theta_i}\left(\frac{x_{ij}-P_{ij}}{P_{ij}Q_{ij}}\right)\right]\frac{\partial P_{ij}}{\partial\theta_i}+\left(\frac{x_{ij}-P_{ij}}{P_{ij}Q_{ij}}\frac{\partial^2P_{ij}}{\partial\theta_i^2}\right)\right\}$$

$$=\sum_{j=1}^{m}\left[\frac{-1}{P_{ij}Q_{ij}}\left(\frac{\partial P_{ij}}{\partial\theta_i}\right)^2-\frac{(x_{ij}-P_{ij})(1-2P_{ij})}{(P_{ij}Q_{ij})^2}\left(\frac{\partial P_{ij}}{\partial\theta_i}\right)^2+\left(\frac{x_{ij}-P_{ij}}{P_{ij}Q_{ij}}\frac{\partial^2P_{ij}}{\partial\theta_i^2}\right)\right]$$

注意到 $Ex_{ij}=P_{ij}$，故对上式取期望后再取相反数，便得到

$$-E\frac{\partial^2 l}{\partial\theta_i^2}=\sum_{j=1}^{m}\frac{1}{P_{ij}Q_{ij}}\left(\frac{\partial P_{ij}}{\partial\theta_i}\right)^2=I(\theta_i)$$

而 $I_j(\theta_i)=\frac{1}{P_{ij}Q_{ij}}\left(\frac{\partial P_{ij}}{\partial\theta_i}\right)^2$，称为项目 j 的信息量。

再将 P_{ij} 的具体形式代入 $I(\theta_i)$ 或 $I_j(\theta_i)$，便可以计算出测验信息量或项目信息量。但这里 θ_i 为被试 i 的真实能力，无法知晓，若用 θ_i 的估计值 $\hat{\theta}$ 代入，便可以近似计算关于 θ_i 的测验误差方差的近似值。

对于多级评分方式，仍然用上式 X_i 表示被试 i 的得分向量，并引入得分 x_{ij} 的指示变量

$$u_{ijk}=\begin{cases}1，若\ x_{ij}=k\\0，否则\end{cases}$$

设项目 j 满分值为 f_j，则得到 X_i 的似然函数

$$L(X_i)=\prod_{j=1}^{m}\prod_{t=0}^{f_j}P_{ijt}^{u_{ijt}}$$

对数似然函数为

$$l_i=\sum_{j=1}^{m}\sum_{t=0}^{f_j}u_{ijt}\ln P_{ijt}$$

Fisher 信息量仍定义为 $-E\frac{\partial^2\ln l_i}{\partial\theta_i^2}$

由于 $\frac{\partial l_i}{\partial\theta_i}=\sum_{j=1}^{m}\sum_{t=0}^{f_j}\frac{u_{ijt}}{P_{ijt}}\frac{\partial P_{ijt}}{\partial\theta_i}$

$$\frac{\partial^2 \ln l_i}{\partial \theta_i^2} = \sum_{j=1}^{m} \sum_{t=0}^{f_j} \left\{ \frac{u_{ijt}}{P_{ijt}} \frac{\partial^2 P_{ijt}}{\partial \theta_i^2} - \frac{u_{ijt}}{P_{ijt}^2} \left(\frac{\partial P_{ijt}}{\partial \theta_i} \right)^2 \right\}$$

取期望，注意到 $Eu_{ijt} = P_{ijt}$，则有

$$E \frac{\partial^2 \ln l_i}{\partial \theta_i^2} = \sum_{j=1}^{m} \sum_{t=0}^{f_j} \left\{ \frac{\partial^2 P_{ijt}}{\partial \theta_i^2} - \frac{1}{P_{ijt}} \left(\frac{\partial P_{ijt}}{\partial \theta_i} \right)^2 \right\}$$

又注意到 $\sum_{t=0}^{f_j} P_{ijt} = 1$，知 $\frac{\partial}{\partial \theta_i} \left(\sum_{t=0}^{f_j} P_{ijt} \right) = 0$

$$\sum_{j=1}^{m} \sum_{t=0}^{f_j} \frac{\partial^2 P_{ijt}}{\partial \theta_i^2} = \sum_{j=1}^{m} \frac{\partial^2}{\partial \theta_i^2} \left[\sum_{t=0}^{f_j} P_{ijt} \right]$$

$$= \sum_{j=1}^{m} \frac{\partial}{\partial \theta_i} \left[\frac{\partial}{\partial \theta_i} \sum_{t=0}^{f_j} P_{ijt} \right] = 0$$

$$\text{故} \quad -E \frac{\partial^2 \ln l_i}{\partial \theta_i^2} = \sum_{j=1}^{m} \sum_{t=0}^{f_j} \frac{1}{P_{ijt}} \left(\frac{\partial P_{ijt}}{\partial \theta_i} \right)^2$$

注意到对于 0－1 评分时，$P_{ij0} = Q_{ij}$，$P_{ij1} = P_{ij}$

则上式可以写成

$$-E \frac{\partial^2 l_i}{\partial \theta_i^2} = \sum_{j=1}^{m} \left[\frac{1}{P_{ij0}} \left(\frac{\partial P_{ij0}}{\partial \theta_i} \right)^2 + \frac{1}{P_{ij1}} \left(\frac{\partial P_{ij1}}{\partial \theta_i} \right)^2 \right]$$

$$= \sum_{j=1}^{m} \left[\frac{1}{Q_{ij}} \left(\frac{\partial Q_{ij}}{\partial \theta_i} \right)^2 + \frac{1}{P_{ij}} \left(\frac{\partial P_{ij}}{\partial \theta_i} \right)^2 \right]$$

$$= \sum_{j=1}^{m} \frac{1}{P_{ij} Q_{ij}} \left(\frac{\partial P_{ij}}{\partial \theta_i} \right)^2$$

这种推导过程似乎比针对 0－1 评分方式的 Fisher 信息函数的推导还更简洁。

以上是单维 IRT 的信息函数，这时对 0－1 评分模型，其信息函数的计算可以表示成

$$\sum_{j=1}^{m} \left(\frac{\partial P_{ij}}{\partial \theta_i} \right)^2 / (P_{ij} Q_{ij})$$

$\frac{\partial P_{ij}}{\partial \theta_i}$是一阶偏导，对于 MIRT，$\theta_i$ 是一个向量而不是一个标量，并要考虑 θ_i 与各个坐标轴之间的夹角，这时对向量的导数应该改为方向导数(directional derivative)，即梯度(gradient)，记 α 是 θ 向量和各坐标轴之间的夹角构成向量，则 0－1 评分 MIRT 的关于项目 j 的项目信息函数定义为

$$I_a^{(j)}(\theta) = [\nabla_a P_i(\theta)]^2 / [P_i(\theta) Q_j(\theta)]$$

对于项目反应曲面的方向导数 $\nabla_a P_j(\theta)$ 由下式给出

$$\nabla_a P_j(\theta) = a_1 P_j(\theta) Q_j(\theta) \cos\alpha_1 + \cdots + a_k P_j(\theta) Q_j(\theta) \cos\alpha_k$$

$$= P_j(\theta) Q_j(\theta) \sum_{t=1}^{k} \alpha_t \cos\alpha_t$$

于是

$$I_a^{(j)}(\theta) = P_j(\theta) Q_j(\theta) \left(\sum_{t=1}^{k} \alpha_t \cos\alpha_t \right)^2$$

而对于0－1评分MIRT的项目信息量和测验信息量的关系类似与项目特征曲面(item characteristic surface)和测验特征曲面(test characteristic surface)的关系。但是要注意在某个特殊的方向上测验信息量是项目信息量之和①。

第二节　常见的 *IRT* 计量模型

项目反应理论用数学模型揭示和描述被试在项目上的作答反应与被试潜特质之间的关系。数学模型是事物本质的抽象，而抛弃了一些非本质的现象；数学模型要能够求解，必须加上一些条件或一些假设，这些条件(这些假设)一方面使得模型能够应用；但另一方面也就限制了该模型的一些应用。比如说"在不受外力条件下，物体保持原有运动状态，不是静止的便是匀速运动"，这个规律仅仅在"不受外力"这一条件下成立。如果受到外力，则结论不成立。模型中一些假设，也可能在当时条件限制下，为了能够将模型中的未知参数求出来而不得不增加上去的，随着时间推移，寻找到了新的数学工具，可以解决更复杂的问题，模型的一些限制条件也就放宽了；或许是在一些研究条件下，对模型加上了一些限制，但是有研究者对这些限制提出了质疑，提出了挑战，甚至找到了不满足这些假设的实例，即出现了新的情况，原有模型无法描述这种新情况，为了研究这些新情况，新的假设和新的模型应运而生。

在项目反应理论的发生发展过程中，开发了许许多多的模型。在这里我们只介绍一些常用的模型。

项目反应理论中的模型有多种分类标准，比如按潜在特质的维度的分类，有单维($m=1$)及多维($m\geqslant 2$)；按项目反应函数形态分有Logistic模型、正态模型和其他模型；按记分方式有双歧评分(即0－1评分)模型及多级评分模型；这些分类又可以进行交叉分类，比如按评分方式及潜在特质的维度这两个标准出现0－1评分多维项目反应模型(dichotomous，MIRT)及多级评分多维项目反应模型(Polytomous，MIRT)；依照评分方式及潜特质维度及反应函数的形态又出了0－1评分Logistic MIRT及0－1评分正态MIRT，由这种交叉划分方式可以将项目反应模型划分更细。

当然，上述模型划分还比较粗，比如0－1评分Logistic MIRT，又由心理特质是否可以相互补偿还是部分补偿开发出相应的补偿型及部分补偿型模型。

特别值得注意的是上述模型，答对项目的概率依据被试能力的升高而变大，这种性质称为单调性，甚至依照潜在特质的单调性有一类非参数项

① C. R. Rao. Handbook of Statistics. Vol. 26. 2007, pp. 626－627.