现代心理测量理论与技术丛书

京师心理研究 JINGSHI XINLI YANJIU

项目反应理论基础

罗照盛 著

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

项目反应理论基础

罗照盛 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

项目反应理论基础／罗照盛著.—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 978-7-303-14693-2

Ⅰ.①项… Ⅱ.①罗… Ⅲ.①心理测量学－研究 Ⅳ.①B841.7

中国版本图书馆CIP数据核字（2012）第125303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：10.5
字　　数：190千字
版　　次：2012年9月第1版
印　　次：2012年9月第1次印刷
定　　价：24.00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。

張华华

2012年3月16日

于美国伊利诺伊大学香槟校区

前 言

项目反应理论作为心理与教育测量理论的新发展，具有经典测量理论无法比拟的优势。经典测量理论在指导实践的过程中暴露出许多矛盾，如理论假设很难实际界定和操作、参数依赖于样本、项目特性与被试特性之间没有建立内在联系，等等。项目反应理论则很好地解决了这些问题，因此在指导研究实践中具有更强的生命力。

然而，项目反应理论的推广和应用却受到许多因素的阻碍，以至于只有专门从事测量学研究的学者们才能较好地理解和掌握它，而其他心理学和教育学研究者只好继续使用经典测量理论的方法和技术。这使得项目反应理论的优势无法转变成现实。影响项目反应理论推广和应用的主要因素有：(1)该理论的模型相对比较复杂，包含有许多数学函数式子，这让许多学者望而却步；(2)该理论的技术和方法实现过程比较复杂，包含一系列的参数估计过程，这个过程又涉及许多数理统计的知识，这让许多学者又望而却步；(3)相对于经典测量理论可以使用如 SPSS 等流行软件，项目反应理论各种参数和技术的应用没有易用的计算机软件支持；(4)测量学研究者与心理学、教育学其他领域之间缺乏足够的交流。以上因素使测量学研究者与其他领域的研究者之间相互隔离，无法相互学习、相互提高。

本书的写作目的就是希望能够拉近测量学研究者与心理学、教育学其他领域研究者之间的距离，使测量学的最新研究成果能够用于心理学和教育学的研究和应用实践中，而心理学和教育学的研究和应用实践又反过来促进测量学理论和技术的发展。为了达到这个目的，本书的写作原则为：

(1)基础性和启发性。所有内容为项目反应理论的基础性内容，为读者理解现代测量理论，包括认知诊断理论中的模型、技术和方法，打下良好完备的基础。许多内容的选择是为了启发读者去进一步理解该领域其他相似的内容。

(2)通俗平白。力求把项目反应理论的原理、技术、方法通过文字形式和更加简单的式子呈现给读者。尽量从一个未接触项目反应理论的心理学或教育学研究者的角度，来剖析项目反应理论中的各种原理、技术和方法。尽量通过日常实例进行讲解。

(3)实践应用性。涉及需要进行计算的内容，将用世界上比较常用的软件进行数据分析的实际演练，并给予详细的说明，以让读者在理解内容的同时，能够实际动手操作，加深印象。

(4)与国内外发展接轨。概念能反映国际的统一界定。

本书共包括9章内容：第一章绪论部分主要讲述项目反应理论的发展，及其相对经典测量理论的优势；第二章模型部分主要讲述项目反应理论模型的提出过程、模型假设、各种常用模型的具体介绍；第三章假设检验部分主要针对项目反应理论模型的假设，讲述如何验证模型与测验数据之间的拟合性问题；第四章参数估计部分讲述了项目反应理论模型项目参数与被试参数估计的原理，及各种常用参数估计方法；第五章等值部分首先讲述了国内外文献中的各种量表化术语及其含义，之后重点讲述了等值的设计和方法，最后还专门介绍了非常实用的纵向量表化设计和方法；第六章信息函数部分主要讲述了信息量的概念及其与信度和测量误差之间的关系，具体讲述了项目和测验信息函数的特点；第七章应用部分主要讲述了项目反应理论及其技术在题库建设、测验编制中的应用，测验编制分为常模参照测验、标准参照测验与计算机化自适应测验三个内容；第八章发展部分主要讲述了多级记分模型、多维项目反应理论模型、认知诊断理论和项目功能差异；第九章数据分析实践主要介绍了国际流行软件在数据分析实践中的具体操作和应用。

本书的写作，是基于本人近20年的学习、研究和实践的个人理解写成。在本书的写作过程中，始终得到江西师范大学戴海琦教授的关心、指导和激励，从写作大纲到最后内容的审订，倾注了戴老师的许多辛勤劳动和无微不至的照顾之心。丁树良老师也为本书的内容提出了许多指导性意见和建议。同时，本人在美国伊利诺伊大学访学期间，还得到张华华教授的许多睿智想法的启发。

在本书即将完成之际，特别感谢在本人的学术生涯中的各位指导老师：张厚粲教授、漆书青教授、戴海琦教授、丁树良教授、张华华教授，他们在我求学的不同阶段给了我悉心的指导和关心。师恩永不忘！

罗照盛

2012年春于美国伊利诺伊大学香槟校区

(University of Illinois at Urbana-Champaign)

目 录

第一章 绪 论

心理测量学研究的是心理测量活动中的一般理论、方法和技术问题，而不是某种具体的心理现象的测量问题。在20世纪初，随着智力测验的蓬勃发展，用于指导测验编制和分析的理论——经典测量理论(Classical Test Theory，CTT)开始得到发展，其核心内容包括真分数、信度、效度等概念。由于皮尔逊统计理论(Pearson statistics)在那个时代的统治地位，CTT的概念和计算方法基本是基于相关的概念发展起来的。在这个阶段，大家关注的焦点是被试在测验上的总分，即被试在测验中的一般性表现，至于被试在每个具体项目上的表现并未受到足够的重视。

第一节 经典测量理论回顾

经典测量理论历史上为心理测量学的发展和应用作出了巨大贡献，直至现在，它还在心理与教育测量领域中占据着重要地位，今后，经典测量理论仍将在指导测验开发、分析和应用方面继续发挥它的重要作用。然而，经典测量理论由于其理论体系上的先天不足，影响了它的深入发展与应用。经典测量理论是在随机抽样理论基础上建立的一套心理与教育测量理论体系，它的核心概念及其计算方法基本是基于相关的概念发展起来的。而所有数据分析的基础是被试在项目上作答的观察分数，在经典测量理论的核心定义中认为，观察分数(observed score)等于真分数(true score)加上误差分数(error score)，可是，在实际数据分析中，真分数总是无法获得，因此，在用观察分数对被试特质或项目特性进行评价时就不可避免地掺杂了大量的误差因素。经典测量理论主要的局限表现在以下五个方面。

一、观察分数等权重线性累加的不合理性

在经典测量理论中，许多测量学指标都是基于观察分数的直接累加得到。比如，标志被试水平的总分是通过观察分数的直接累加得到，而项目的难度指标，同样也是由观察分数累加后进行简单转换得到。对于被试总分来说，分数的累加过程是很不合理的，因为项目之间难度有高低之分，同时每个题目的区分能力也不尽相同，答对不同难度和区分度的题目就应该给予不同权重的分数，而不是在相同权重基础上的简单累加。比如，在1、0记分的题目中，一个低水平被试答对一个容易的题目，与一个高水平

被试答对一个高难度题目的得分权重是一样的。虽然高水平被试可能答对更多题目，而低水平被试只能答对更少的题目，但这种等权重的分数累加方式，不能真正地反映不同被试之间的能力水平差异程度。对于项目难度指标的计算，这种情况是一样的。

在测验编制时，虽然不同题型题目可能赋予不同分数，然而，赋分高的题目，其难度却并不一定是高的，在1、0记分选择题中，就经常会有许多题目难度超过赋分更高的简单问答题，而且，仅凭经验也无法根据题目区分度进行赋分加权。另外，虽然经典测量理论也可以基于项目参数进行加权记分，然而，由于经典测量理论项目参数本身的缺陷，使这种加权过程也显得非常粗糙。

二、测验对被试的评价依赖于测验的具体项目组合和项目数量

在经典测量理论中，测验对被试的评价指标主要是测验总分，而测验总分是被试在各个项目上的观察分数的组合。在用测验总分评价被试时，不同被试之间水平的比较只能在他们考了同一份测验的情形下才能进行，而如果不同被试参加了题量不同，或是题量相同而题目不同的测验，我们立即就会指出被试的这些总分之间是不可比的，除非这两个测验是严格平行的测验。这种现象被称为测验分数的解释依赖于测验的具体项目组合和项目数量。虽然也有基于经典测量理论的测验分数等值方法，然而，这些等值的设计需要测验项目和被试样本满足非常严格的假设，否则，等值结果就会很不稳定。经典测量理论框架下对被试进行比较被迫限制在同一份试卷上，这使得测验分数的应用受到很大的限制，缺乏足够的拓广性。在大规模测试评价中，使用同一份试卷就可能存在以下一些问题：(1)必须在同一时间内测试，否则测试内容就很难保密，这使测试管理和试卷安全保密难度加大；(2)无法根据不同类型考生设置不同的试卷；(3)如果需要进行发展性评价，即评价被试在不同时间的水平发展状况，经典测量理论就无法满足要求，而这种发展性评价现在越来越受到人们重视。

三、测验及项目的性能指标的估计依赖于具体的被试样本

经典测量理论有一套刻画测验及项目各方面性能的指标，但在经典测量理论中，项目指标同样也是通过观察分数的简单转换得到，这些性能指标的估计严重依赖于参加当前测试的具体被试样本的特性。最明显的例子就是，项目难度估计值会随着参加测试的被试样本群体水平的变化而变化。项目区分度、测验的信度和效度，由于它们本质上都是通过基于观察分数的相关系数进行估计，而我们知道，相关系数的计算同样都是依赖于具体的被试样本群体。结果是，同一测验在不同被试样本群体上施测，所估计的性能指标会不一样，对测验性能的解释也就不一致。

由于测验及项目性能指标的这个特性，使得对测验和项目的性能分析

结果只能是相对当前测试样本有效。研究者们为了使这些性能指标具有更好的推广性，就会花很多的心思去抽取能够完全代表总体的被试样本进行测试。于是，我们会发现，许多基于经典测量理论开发编制的心理测验量表，在项目编制和分析过程中以及在最后的测验信度、效度质量分析中，测验开发者会组织大量的人力、花大量时间、通过精心培训去抽取能够代表总体分布的样本进行测试，而且，测试还需分为预测试和正式测试。

四、被试能力与项目难度两个指标含义的非统一性

经典测量理论中，被试能力水平是用测验总分表示的，测验总分则是由全体测验项目计算得到，被试得分是被试在项目集上的成功比例。项目难度的参照系则是全体被试，项目难度值是在该项目上作答成功的被试的比例。因此，这两个指标的参照系是不相同的，是不可比较和相通的。比如，以0、1记分题目为例，如果以得分率作为被试能力水平的指标，而以错误率作为项目难度的指标，进行如下比较是没有意义的：能力水平为0.6的被试答对难度为0.5的项目的概率很大，这从直观上看是有道理的，因为被试能力水平高于项目难度水平，然而，如果在经典测量理论下进行解释就可能存在问题，能力水平为0.6的被试意味着他答对了60%的题目，但参加这次考试的被试中大部分答对率在80%以上，而项目难度为0.5意味着只有50%的被试答对了这个项目，而这50%的被试里极有可能并不包含答对率只有60%的那些被试。被试能力与项目难度这种貌合神离的特点，使得经典测量理论这套测验性能指标的使用价值非常有限。

五、测量误差估计的不精确性和笼统性

测量的目的是为了获得对象的精确测值。任何测量过程都是有误差的，在得到测值的同时，获得测值的误差估计也是非常重要的。在经典测量理论中，测量的随机误差值是通过测量的信度估算的，但经典测量理论中测验的信度估值本身却是不精确的、笼统的。首先，经典测量理论的信度并不是按“真分数方差与观察分数方差的比”的定义公式求取的，而是基于“平行测验”的假设，通过计算相关系数估计得到的，但是“平行测验”建构本身可能就存在很大误差，这样估得的信度肯定是不精确的，从而导致测验误差的估值也是不精确的。其次，在经典测量理论中，测验的误差估值只有一个。经验告诉我们，同一测验在估计不同能力水平的被试时，其误差是不一样的。因此，以同一误差估计值来评价所有被试测值的精确性是非常笼统，也很不精确的。

经典测量理论的这些局限性限制了它在实践中的应用，随着人们对经典测量理论这些局限性的认识的深化，人们感到必须建立一套新的测量理论，以适应心理测量实践发展的需要，项目反应理论(Item Response Theory，IRT)就是在此背景下发展起来的一种全新的现代测量理论。

第二节　项目反应理论的基本架构及其优良性质

我们都知道，影响被试在项目上的作答结果的主要因素有两个方面：一方面是被试本身的能力水平；另一方面是项目的计量学属性，如项目难度、区分度、猜测性。按照一般经验来说，在同一个项目上，能力水平越高的被试，答对这个项目的可能性就越大；而对于同一个被试来说，越容易的项目也越可能被答对。

作为现代测量理论重心的项目反应理论(IRT)，它的特点是以概率函数的形式来描述项目作答反应结果是如何受到被试能力水平和项目特性联合作用的影响，具体来说，就是依据被试在各个项目上的实际作答反应结果，经数学模型的运算，统一估计出被试的能力水平(abilities)或潜在心理特质水平(latent traits)，以及项目的计量学参数。描述被试能力水平、项目参数与项目作答结果之间关系的数学模型称为项目特征函数(item characteristic function，ICF)，以图形表示则称为项目特征曲线(item characteristic curve，ICC)。下图 1-1 为一典型的 ICC：横轴表示被试的能力水平，纵轴表示概率，例如，在曲线上有 A、B、C、D、E 五个点，它们分别代表了 5 位不同能力水平被试在该项目上的答对概率，由图可知能力值(θ)越高，答对该项目的概率(p 值)就越大。

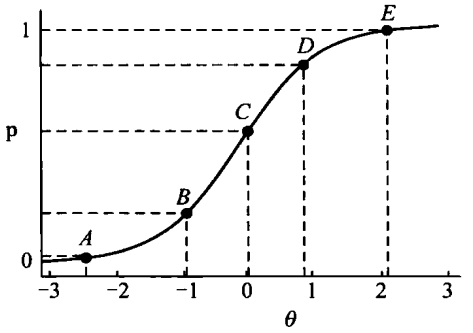

图 1-1　典型的项目特征曲线

IRT 将项目视为测量被试能力水平的基本单位，项目的属性通过项目参数来描述。项目一般包含以下三个计量学参数：a 参数：即区分度参数，它的值越大表示项目对不同被试能力水平的鉴别力越强；反之，则鉴别力越弱。在 ICC 图中，a 参数反映了项目特征曲线(ICC)的斜率，其理论值范围介于$-\infty\sim+\infty$，但在实际应用中的取值范围一般为 0～+3。b 参数：即难度参数，它的值越大表示项目越难，在 ICC 图中，它反映了项目特征曲线(ICC)位于能力量尺上的位置，因假定被试的能力值范围为$-\infty\sim+\infty$，所以 b 参数的理论值范围亦然。不过，在实际应用中被试能力值取

值范围一般介于－3～＋3。c 参数：就是猜测参数，代表了被试仅凭猜测答对项目的可能性，它的值越大表示不论被试能力水平高低，均更容易答对这个项目；值越小，则不易光凭猜测答对这个项目。c 参数反映项目特征曲线(ICC)的左下渐近线(lower asymptote)的高度，其理论值范围介于0.0～1.0，但是，在实际应用中，过高 c 参数的项目经常不被接受。

项目反应理论(IRT)以项目特征函数(ICF)来描述项目作答反应结果与被试能力水平及项目参数之间的关系，因所包含的参数个数不同，函数可被区分为不同的模型，常用的数学模型有单参数模型、双参数模型及三参数模型等三种，各模型之项目特征函数如公式(1-1)至公式(1-3)所示。

$$\text{单参数模型：}P_{ij}(\theta_j)=\frac{1}{1+e^{-D(\theta_j-b_i)}} \tag{1-1}$$

$$\text{双参数模型：}P_{ij}(\theta_j)=\frac{1}{1+e^{-Da_i(\theta_j-b_i)}} \tag{1-2}$$

$$\text{三参数模型：}P_{ij}(\theta_j)=C_i+(1-C_i)\frac{1}{1+e^{-Da_i(\theta_j-b_i)}} \tag{1-3}$$

公式中：D 为常数 1.7；e：自然对数之底；j：被试编号；θ_j：第 j 位被试的能力值；i：项目编号；a_i，b_i，c_i：分别表示第 i 题的区分度参数、难度参数、猜测参数；$P_{ij}(\theta_j, a_i, b_i, c_i)$：表示能力值为 θ_j 的被试 j 答对第 i 题的概率。当然，也可以将函数式以图形表示，称为项目特征曲线(ICC)。

因为心理学研究的各种心理现象中，潜在心理特质水平(latent trait)与外部行为表现之间的关系大多是非线性的，所以，在描述被试能力水平与项目作答反应之间的关系上，IRT 以基于概率的数学模型来表示，较之经典测量理论以线性关系来表达，更能契合心理学研究中关于心理现象本质规律的描述。

项目反应理论较之于经典测量理论，有以下一些特征。

一、被试能力参数与项目参数具有不变性的特征

关于参数不变性(parameter invariance)的性质，国内外有两个角度的理解。

首先，是从同一总体的角度进行理解，即来自同一总体的不同样本所估计得到的总体参数是不变的。具体来说，只要测试同一特质的测验项目的参数具有足够宽的覆盖，也就是测验中既有难的题目，有中等难的题目，也有容易的题目，那么，不管项目分布形态如何，被试能力参数的估计就不依赖于具体的项目；同时，只要在同一维度上被试的能力水平分布全距足够宽，那么，不管被试分布形态如何，项目参数的估计也不会依赖于具体的被试样本群体及其分布形态。在这里，“足够宽”的含义，对于被

试能力参数的估计来说，指的是在测验项目组成中，既要有部分被试能答对的项目，同时又要有一些被试无法答对的项目；对于项目参数的估计来说，指的是在被试样本组成中，既要有部分能答对该项目的被试，同时又要有一些无法答对这个项目的被试。在考虑样本组成时，不必关心它是否服从某种分布形态，如正态分布等。

其次，是从不同总体的角度进行的理解。比如，如比和祖波(Rupp & Zumbo，2006)就指出，许多研究者在理解项目反应理论中关于参数不变性这一特性时存在误解。正如前面说到的关于“被试能力参数与项目参数具有样本不变性的特征”这一内容一样，我们认为，参数不变性指的是，使用同一总体内不同样本(题目样本或被试样本，下同)所估计的相同被试或相同项目的参数是不变的。而如比和祖波(Rupp & Zumbo，2006)却认为，参数不变性指的是，对于项目参数来说，使用不同被试总体估计同一批项目时得到的参数是不变的，比如，参数不变性应该检验以下情形：一个人事选拔测验在应用于男性总体时估计得到的参数与应用于女性总体时估计得到的参数是否相同，或者，一个智力测验，用加拿大中学生被试总体得到的参数和美国中学生被试总体得到的参数是否相同。而对于被试参数来说，不同测量条件(measurement conditions)下获得的被试参数应该也是不变的，在这里，如比和祖波(Rupp & Zumbo，2006)没有明确指出测量条件指的是什么，可能包括时间、地点、测试方式(如问卷、面试等)等，不过，对于题目条件来说，至少应该保证是属于同一个领域的内容或同一特质范畴。

当然，参数不变性性质必须剔除随机误差因素的影响。另外，由不同样本所得到的参数估计值由于单位量表(scale)不同的原因，数据不可能完全一致，因此，需要首先通过本书第五章讲到的量表化方法使不同样本估计得到的参数量表统一，然后再进行比较。

正是由于项目反应理论具有参数的不变性特征，这使它在指导题库建设、计算机化自适应测验中能够发挥经典测量理论无法达到的优势。

二、被试能力参数与项目难度参数具有统一的量表

根据项目反应理论模型估计出来的被试能力参数与项目难度参数具有统一的量表，即被试参数与项目参数可以被标定在同一个参照尺度上，这样，被试位置参数与项目位置参数就可以直接在这个尺度上进行比较和解释。比如，在项目反应理论中进行如下比较和解释是很合理的：能力水平估计值为 0.6 的被试答对难度估计值为 0.5 的项目的概率大于答错的概率，而答对难度估计值为 0.7 的项目的概率则小于答错的概率。而在实际应用中，用于测试能力水平为 0.6 的被试的最佳项目的难度也应该在 0.6 左右，离开 0.6 太远的项目要么太容易，要么太难，以至于浪费被试的作答时间，

就像要求大学生作答“1+1”等于几一样。

被试能力参数与项目难度参数具有统一量表的特性，为项目反应理论的实际应用带来极大方便，特别是在指导测验编制及进行自适应测验中，可以有针对性地选择适合被试能力水平的项目，以使测试更加高效和精确。

三、可以针对不同被试精确估计每个项目及测验的测量误差

测量误差是评价测验质量的主要指标，其中，用于标志随机误差造成的测验结果一致性问题的指标叫做信度，信度与测量标准误差之间存在反比关系。在经典测量理论中，对于所有的被试，测验只提供一个统一的信度指标，这是它的一个不足之处。

在项目反应理论中，每个项目和测验为每个不同被试特质水平的估计提供了独立的信度指标，及相应的被试特质参数估计误差指标。在项目反应理论中，用信息量来代替信度的概念。信息量通过信息函数(information function)来计算，信息函数是项目反应理论中的一个非常重要的概念，它分为项目信息函数和测验信息函数，测验信息函数是项目信息函数的累加。信息函数反映了项目或测验在估计被试特质水平时所提供的信息量大小关系。不同项目对同一被试特质水平可以计算独立的信息量值，同时，同一项目(测验)可以对不同特质水平计算独立的信息量值，这些信息量值之间可能很不相同，具体要看项目难度与被试特质水平的匹配程度及项目的区分度等参数指标特性。

有了不同项目对不同被试单独计算信息量值的技术和方法，我们就可以对每个被试的特质水平估计误差进行主动控制，从而更加有利于指导测验的编制。

第三节 项目反应理论发展简史

在项目反应理论发展的早期，大家都称之为潜在特质理论(latent trait theory)，因为这个理论就是想探讨影响人们作答行为的内在特质的特征。后来有一段时间，大家又倾向于称之为项目特征曲线理论(item characteristic curve theory)，因为这个理论的核心概念就是项目特征曲线。然而，由于项目特征曲线从概念和表现方式上反映的是二值(如1、0)记分数据，这个名称在多值记分模型中就显得有点别扭，因此，现在更多的人称之为项目反应理论(item response theory)，这很好地反映了该理论的应用和发展主要是基于对被试在项目上的反应为基础的事实。

1916年，推孟(Terman)在对比奈--西蒙(Binet-Simon)智力量表进行分析时，用图形绘制了不同年龄段被试在项目上的正确作答比例，这是项

目特征曲线的雏形。1952 年 Frederic M. Lord 在大量数据分析基础上，提出双参数正态肩形曲线模型及其参数估计方法，一般这被认为是项目反应理论正式创立的标志。1960 年，Georg Rasch 提出了 3 个项目反应模型，由于其模型建构过程与其他模型不一样，而自成一派，后来的代表性人物为 Wright。1957 年、1958 年，Birnbaum 提出了比正态肩形曲线模型更加易用的 logistic 模型，打开了 IRT 用于实际的大门。1968 年，Lord 和 Novick 出版著名的《心理测验分数的统计理论》一书，其中有 5 个章节深入介绍了项目反应理论模型。1969 年，Samejima 提出了适合多级记分题型的等级反应模型，突破了项目反应理论仅用于 0、1 评分题型的限制。1978 年，Sympson 提出多维三参数模型，突破了项目反应理论仅用于单维测验的限制。

以上这些历史是关于项目反应理论发展的基础性回顾，通过这些学者们的努力，建立了项目反应理论的基本概念和理论框架。从那以后至今关于项目反应理论的发展，主要表现在三个方面：(1)模型的修正和发展，提出了适应于人格测验的理想点模型(ideal point model)或展开模型(unfolding model)，提出了适应各种测验数据的更多的多级记分模型，以及适应更多认知任务特点的各种多维模型。(2)基于项目反应理论各种技术和方法的研究，如等值研究、计算机化自适应测验研究，以及各种参数估计方法的研究等。(3)基于项目反应理论的应用研究，如测验质量分析、被试特质分析、等值分析与题库建设、计算机化自适应测验在实际考试中的应用，等等。

以上关于项目反应理论的发展简史，主要是参考了 Hambleton 和 Swaminathan (1985)的书《项目反应理论：原理与应用》(*Item Response Theory: Principles and Applications*)，参考文献中并不具体标明这里提到的每个文献的出处，请有兴趣的读者自己查阅。

当然，项目反应理论的发展，除了各种模型的提出之外，非常重要的是对其参数估计方法的研究及相应计算程序的开发，一定意义上说，没有后者的支持，仅有模型是没有实用价值的，模型也就不可能得到发展。所以，项目反应理论的发展，同时伴随着许多学者对其模型性质和参数估计技术的研究和程序开发的过程。在参数估计方法和技术方面，影响比较大的有，Kolakowski 和 Bock(1970)实现的联合极大似然估计方法，以及 Bock 和 Aitkin(1981)提出的基于 EM 算法的边际极大似然估计方法，而马尔可夫链蒙特卡罗(Markov Chain Monte Carlo，MCMC)方法则在估计复杂模型参数方面显示出了很大优势。当前应用比较广泛的著名计算机软件有适应 1、0 记分模型的 BILOG 程序，适应多级记分模型的 MULTILOG