沈阳师范大学学术文库系列丛书

多维项目反应模型应用理论

付志慧 著

科学出版社

沈阳师范大学学术文库系列丛书

多维项目反应模型应用理论

付志慧 著

科学出版社

北京

内 容 简 介

本书主要研究多维项目反应模型的理论及其应用. 首先，基于潜在变量理论，给出项目反应模型的详细定义，并利用贝叶斯统计理论和拟似然理论的相关工具，基于数据扩充技术，详细推导出多维 Logistic 项目反应模型及多维等级反应模型的 Gibbs 抽样方法. 在此基础上，针对缺失反应数据问题，对不同的缺失过程建模，同时给出多维等级反应数据模型和缺失指标模型的后验估计. 其次，本书将多维项目反应模型应用到纵向反应数据中，并采用成对建模的方法来拟合数据，无需考虑能力的维数问题，对各个时刻点的能力之间的相关程度，可以直接给出估计. 该模型还可用于各学校的纵向教学质量评估或同类院校之间的阶段性教学横向比较. 最后，将 Copula 方法引入项目反应理论中来分析相依反应数据，解决项目反应理论中局部相依性假设问题.

本书可作为从事教育统计及心理测量方向研究人员的参考资料，也可以作为教育测量专业的研究生教材.

图书在版编目(CIP)数据

多维项目反应模型应用理论/付志慧著. —北京: 科学出版社, 2017.3

ISBN 978-7-03-051481-3

Ⅰ. ①多… Ⅱ. ①付… Ⅲ. ①多维分析-数学模型-研究 Ⅳ. ①O572.11

中国版本图书馆 CIP 数据核字 (2016) 第 322335 号

责任编辑: 李 欣 / 责任校对: 邹慧卿

责任印制: 张 伟 / 封面设计: 陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码: 100717

http://www.sciencep.com

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 3 月第 一 版 开本: 720 × 1000 B5

2017 年 3 月第一次印刷 印张: 7 3/4

字数: 150 000

定价: 49.00 元

(如有印装质量问题, 我社负责调换)

前　言

统计在教育与心理学研究中的重要应用价值已达成共识, 尤其是近年来项目反应理论 (item response theory, IRT) 显示出其绝对优势, 其核心在于: 针对被试能力与被试者对测验项目“正确回答概率”之间关系建模. 它能够解决一些经典测验理论 (classical test theory, CTT) 解决不了的问题, 已成为教育与心理研究领域的热点问题之一. 然而项目反应理论也并非完美的测量理论, 它也存在一些有待完善的问题. 这些问题主要来自于项目反应理论三个非常严格的基本假设, 即单一维度假设、局部独立性假设和单调递增假设.

项目反应理论的单维性假设与许多心理或教育测验的实际是不相符的, 因为测验数据的多维性与人在完成一项测验任务时需要多种能力的共同配合是相符的, 很少有测验只测查单一的能力或特质. 多维项目反应理论 (multidimensional IRT, MIRT) 正是近 20 年来测验理论发展的主要新进展之一. 多维项目反应理论为测验中所涉及的每个维度引入能力和项目区分度参数, 进而模拟测验题目和被试之间的交互作用. 目前多维项目反应理论仍然处于早期发展阶段, 有许多问题需要处理.

另外, 在我国考试、测评中, 许多测验或量表中均有题组类型的项目, 如汉语考试和外语考试中的阅读理解题、完形填空题、听力短文理解题; 人才测评中的情景判断测验题及数学试卷中的计算题, 同一题目中不同子问题之间都具有相关性. 因此在实际测验问题中, 项目反应理论的局部独立性假设往往不能满足, 应对原有项目反应模型进行改进.

项目反应理论在实际应用中存在的核心问题在于参数估计的复杂性, 参数大致有两类: 一是潜在特质参数 (如学习能力、心理指标等), 二是项目或题目参数. 每一个被试至少匹配一个潜在特质参数 (多维时参数个数成倍增多), 而每一个项目至少匹配两个参数以上 (如区分度、难度等, 有时候还要加上猜测系数), 从而模型中待估参数数目庞大, 模型结构也比较复杂. 随着现代统计学及数学的不断发展, 参数估计的方法也不断发展, 其估计方法主要有条件极大似然估计、联合极大似然估计、边际极大似然估计与 EM 算法及边际贝叶斯估计. 1992 年统计学家 Albert 首先将马尔可夫链蒙特卡罗方法 (MCMC 方法) 应用到项目反应理论参数估计研究中. 自 MCMC 方法引入心理计量学领域后, 心理计量领域中的许多复杂、高维模型的参数估计成为现实, 它是一种全新的参数估计方法. 现有的多维项目反应理论参数估计方法是从 FA 和结构方程模型中移植过来的, 其普遍应用的软件有 NOHARM 和 TESTFACT. TESTFACT 利用极大似然估计, 只能对多维项目反应

模型进行探索性的分析, NOHARM 应用最小二乘法, 允许对高维度的无限量项目进行探索性和验证性模型的项目参数估计, 但是 NOHARM 不能估计被试的能力水平. 另外, 在多级评分多维模型的参数估计方面, 尽管 DIMTEST 软件的其中一个程序 Poly-DIMTEST 及 TESTFACT 对多分项目拓广的 POLYFACT 程序可以用来处理多分数据的参数估计, 但是 POLYFACT 程序只能处理探索性的分析. 另外, 在调查问卷、心理测验等资料分析中经常遇到数据缺失问题, 它给数据分析与应用带来很多困难. 如果单纯地删除缺失数据, 那么就得到一个有偏的样本, 从而得到有偏的论断. 因此, 对缺失数据的正确处理就变得尤为重要. 因此, 还需要更多精细的研究来充实和完备多维项目反应理论的参数估计过程. 本人在读博期间, 针对三参数 Logistic 模型及多维等级反应模型, 基于数据扩充技术, 分别给出了高效简单的 Gibbs 抽样方法, 并对缺失反应数据进行合理建模, 给出了贝叶斯后验估计.

本书以正在进行的教育统计理论及方法的研究为基础, 与信息科学相结合, 针对主要测量方法在项目反应模型中存在的弊端提出改进, 重点放在教育统计研究方法的创新及应用上. 本书包含 6 章, 集中研究以下几个内容: ①多维项目反应模型的参数估计问题; ② 多维项目反应理论参数估计程序的开发; ③对缺失数据及纵向数据进行建模及估计; ④采用 Copula 连接函数法对相依项目反应数据建模.

本书的写作目的是希望能够拉近教育测量学、心理学与统计学领域研究者之间的距离, 针对现有项目反应模型和软件的局限性, 开发出简单易操作的算法及程序, 从而将理论知识更好地应用到实践当中. 由于本人的精力、见识及知识面有限, 仅希望能为读者清除一些障碍, 为项目反应理论尽一份绵薄之力.

本书的编写基于本人近十年的学习、研究和实践. 其中获得我尊敬的博士生导师东北师范大学的史宁中教授和陶剑教授的启发、指导和大力帮助, 他们为本书提出了很多指导性意见和建议. 在此, 特别感谢在本人的学术生涯中的各位指导老师: 史宁中教授、陶剑教授、张华华教授、郭建华教授、高巍教授、王德辉教授, 他们对待科学问题的严谨和认真, 以及执着的探索精神, 给我留下深刻的印象, 使我获益终身. 最后, 限于作者水平有限, 文中难免会有不当或疏漏之处, 敬请诸位不吝批评和指正.

付志慧

2016 年 12 月

目　　录

第1章 绪 论

1.1 背景介绍

心理和教育测量是进行心理和教育研究的一种重要手段. 它在心理和教育的咨询、诊断、评价及人员的分类选拔中具有不可替代的重要作用. 19 世纪末, 心理学脱离哲学而独立, 心理和教育测量学随之正式诞生, 称为心理计量学 (psychometrics). 它是一门包括量化心理学 (quantitative psychology)、个别差异 (individual differences) 和心理测验理论 (mental test theories) 等研究范围的学问.

经过近百年的发展, 心理计量学领域已形成三种理论并存的局面 (Allen & Yen, 2002). 分别是: ① 真分数理论 (true score theory)(Gulliksen, 1950; Lord ,1965); ② 概括化理论 (generalizability theory, GT); ③ 项目反应理论 (item response theory, IRT); (Lord, 1980; Hambleton, 1989; Hambleton et al., 1985, 1991; Baker & Kim, 2004). 前两者又称为随机抽样理论 (random sampling theory), 真分数理论是历史上出现时间最早、发展时间最长、对实际工作影响广泛的一种心理计量学理论. 一般将真分数理论称为经典测量理论, 将概括化理论特别是项目反应理论称为现代测量理论 (modern test theory). 项目反应理论在 20 世纪下半叶得到迅猛发展, Warm(1989) 称 "项目反应理论对经典测验理论好比爱因斯坦相对论对牛顿定律", 可见其影响之深远. 随着计算机技术的发展, 项目反应理论得以迅速推广应用. 目前一些大型的考试 (如 TOFEL、GRE、GMAT 等), 都相继采用以项目反应理论为基础的计算机化适应性测验 (computerized adaptive testing, CAT) (Wainer, 1990), 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. 然而项目反应理论也并非完美的测量理论, 它也存在一些尚待完善的问题, 这些问题源于该理论的三条基本假设: ① 单维性假设 (unidimensionality), 即假定测验的所有项目只考察单一的能力或特质, 这也是人们对项目反应理论提出质疑的关键所在; ② 局部独立性假设 (local independence), 即当控制所测量的能力或特质之后, 被试在不同题目上的反应是相互独立的; ③ 单调性 (monotonicity), 即被试对题目正确反应的概率随其能力水平的增加而单调递增. 鉴于上述弊端, 亟待提出一些新的评价模型和方法. 近年来, 项目反应理论主要研究三个方面的问题: 多维项目反应理论、非参数项目反应理论、认知诊断理论. 我国的教育与心理定量评价方法处于刚刚起步阶段, 现有考试以测试学生所掌握的知识为

主, 基本上没有反映出考生的学习能力, 主要原因是我国对考试理论和心理计量学的研究非常薄弱, 研究考试问题的学者比较少 (辛涛, 2005), 而关于较复杂的多维项目反应模型的相关文献更未见报道. 本书主要讨论多维项目反应模型的参数估计问题.

项目反应理论在实际应用中存在的核心问题在于参数估计的复杂性, 随着现代统计学及数学的不断发展, 参数估计的方法也不断发展, 其估计方法主要有 (Baker & Kim, 2004): 条件极大似然估计 (conditional maximum likelihood estimation, CMLE) (Andersen, 1972, 1973; Wright & Douglas, 1977), 联合极大似然估计 (joint maximum likelihood estimation, JMLE)(Birnbaum, 1968; Wingersky et al., 1982), 边际极大似然估计与 EM 算法 (marginal maximum likelihood estimation and an EM algorithm, MMLE/EM 算法) (Bock & Aitkin, 1981; von Davier & Sinharay, 2007), 边际贝叶斯估计 (marginalized Bayesian estimation) (Mislevy, 1986), 马尔可夫链蒙特卡罗 (Markov chain Monte Carlo, MCMC) 方法 (Albert, 1992; Patz & Junker, 1999; Sahu, 2002). 1992 年统计学家 J.H. Albert 首先将 MCMC 方法应用到项目反应理论参数估计研究中, 大大简化了项目反应理论中参数估计的复杂度, 并且估计精度较好. MCMC 方法源于物理学研究, 20 世纪末引入心理计量学领域, 它是一种动态的计算机模拟技术, 是根据任一多元理论分布, 特别是根据以贝叶斯推断为中心的多元后验分布, 来模拟随机样本的一种方法. 其基本思想是通过模拟服从某一分布也即平稳分布 (一般是待估参数的联合后验分布) 的马尔可夫链, 然后根据模拟的马尔可夫链上的样本点对待估参数进行估计. 当项目反应模型中的参数的个数或维度过多时, 传统的 EM 算法一般难于或无法实现模型的参数估计. 自 MCMC 方法引进心理计量学领域后, 心理计量学领域中的许多复杂、高维模型的参数估计成为现实, 它是一种全新的参数估计方法. 在国外, 该算法已被广泛应用于项目反应理论下的各种模型的参数估计. Patz 和 Junker (1999) 应用 MCMC 方法估计项目反应理论下的 Logistic 模型、分部评分模型 (partial credit models) 及 GLLT 模型 (generalized linear logistic test model) 的参数估计, 拓广了 MCMC 方法在项目反应理论参数估计的实际应用; Bradlow 等 (1999) 运用 MCMC 方法估计两参数 Logistic 的相依题组模型 (testlet model), 成功实现了题组内项目间存在相依项目反应模型的参数估计; Jimmy 和 Douglas (2004) 使用 MCMC 方法估计高维的认知诊断模型 (higher-order DINA model). 因此, 在项目反应模型参数估计中, MCMC 方法实现了 EM 算法难于或无法解决的问题, 更好地服务于实际. 但在我国 MCMC 算法在项目反应理论参数估计的研究却很少 (涂冬波等, 2008).

本书主要将 MCMC 方法应用到多维项目反应模型中. 其次, 在调查问卷、心理测验等资料分析中经常遇到数据缺失问题, 给数据分析与应用带来很多困难. 如果单纯地删除缺失数据, 就得到一个有偏的样本, 从而得到有偏的论断. 因此, 对不完

全数据 (incomplete data) 的正确处理就变得尤为重要 (Little & Rubin, 2014). 本书在项目反应理论框架下, 对不同缺失机制进行建模, 并采用 MCMC 方法给出模型的贝叶斯后验估计. 再次, 在教育与心理研究中, 通常要通过对被试在几个时间点的反应值进行纵向分析, 以得到被试相关能力随时间的变化情况. 例如, 如果要研究学生对数学知识掌握的进展情况, 那么我们就在几个时间点对学生进行一次或多次数学测验. 可以应用项目反应理论将学生对题目的反应值与一个潜变量 (数学能力) 联系起来 (Lord, 1980). 纵向设计中, 常常假定学生的能力是随时间变化的, 重复测量分析的复杂性就在于不同时间点的反应数据并不是独立的. Andersen (1985) 建议每个时刻点的二值反应数据用一个单维的推广 Rasch 项目反应模型来拟合, 且各个时刻的潜变量是相关的. 本书采用题目间多维 (between-item-multidimensional) 模型来分析重复测量的纵向数据, 对于各个时刻点的潜变量之间的相关程度, 可以直接给出估计. 考虑到各时刻能力参数之间的相依性, 我们假定各个时刻的潜在变量生成的能力向量服从多维正态分布, 分布的均值向量和协方差矩阵分别体现了能力随时间的变化情况和不同时刻能力之间参数的相关程度. Andersen (1985), Andrade 和 Tavares (2005) 给出了该模型的边际极大似然 (MML) 估计. 然而 MML 估计的局限性是高维积分的计算问题, 一般我们采用象限积分法来近似计算, 然而潜变量的维数越高, 需要的象限点越多, 这就给计算带来了很大困难. 我们给出两种解决方法, 首先给出 MCMC 方法; 其次给出成对似然 (pairwise likelihood) 的方法, 该方法属于拟似然范畴, 为复合似然 (composite likelihood)(Lindsay, 1988) 的特例. Cox 和 Reid 在 2004 年的论文中讨论了该拟似然方法的优点. Fieuws 和 Verbeke (2005, 2006) 将成对拟合的方法应用到混合线性模型中, 无需考虑潜变量的维数问题, 并通过模拟计算表明该方法产生了无偏估计和有效的标准差.

1.2 本书结构安排

第 1 章: 绪论.

第 2 章: 项目反应理论简介.

第 3 章: 多维 Logistic 项目反应模型的贝叶斯估计. 介绍了多维 3PLM 的贝叶斯推断过程, 并基于数据扩充技术, 给出参数的 Gibbs 抽样方法, 并通过模拟实验验证了该方法的可行性.

第 4 章: 多维等级反应模型下不可忽略缺失数据的贝叶斯估计. 在第 3 章的基础上, 针对缺失反应数据问题, 对不同的缺失过程建模. 当缺失机制不可忽略时, 引入缺失模型, 我们用一个二级评分反应模型来拟合缺失指标, 从而减小由于忽略缺失数据估计参数时产生的偏差. 然后基于数据扩充技术的 Gibbs 抽样方法, 同时给出多维等级反应数据模型和缺失指标模型的后验估计.

第 5 章：讨论多维项目反应模型在纵向反应数据中的应用.

第 6 章：讨论相依项目反应数据的 Copula 建模法.

本书收录了作者近年来所撰写的部分论文, 部分已发表, 部分已经投到相应的刊物上.

第 2 章　项目反应理论简介

2.1　测 验 理 论

2.1.1　测验理论发展历程

心理与教育测量理论的发展经历了三个阶段：20 世纪 50 年代之前是真分数理论起作用, 称为经典测验理论阶段, 20 世纪 50 年代以后, 概化理论 (generalizability theory, GT) 和项目反应理论等多种测量理论发展起来. 经典测验理论是历史上的第一个测验理论, 也是测验的最一般、最基本的理论, 应用极为广泛, 它在测验发展中有着特殊的地位. 现代测验理论大多是在经典测验理论的研究基础上, 针对它在某方面存在的问题发展起来的. 概化理论是针对经典测验理论的信度问题发展起来的. 而项目反应理论, 是克服经典测验理论中题目参数等指标的变异性发展起来的.

经典测验理论是心理学研究者所熟悉的, 主要是指真分数理论, 其基本思想是把测验的得分 (通常称为测验的观察分) 看作真分数和误差分数的线性组合, 可归结为如下简单数学模型：$X = T + e$, X 是观测分数, T 是真分数, e 是误差分. 传统信度、效度、项目分析的原理与方法均建立在这一模型上.

概化理论又称为概括力理论或拓广理论, 它将试验设计及其分析、方差分量模型等统计工具应用到教育与心理测量学, 对经典测验理论中的一个重要概念 (信度) 进行了推广, 即结合测量的情境关系对经典测验理论给出的笼统误差进行探查和分解, 辨明误差的不同来源, 并且在一定范围内变动测量的情境, 考查这种变动引起的误差的相对变化, 从而达到对误差方差进行控制, 提高测验 “信度” 的目的. 项目反应理论在国外发展很快, 最近一段时期国内研究人员也开始应用项目反应理论解决国内测验问题. 它的基本思想与心理学中关于潜在特质的一般理论有关. 项目反应理论假设被试在测验时对试题的反应受某种心理潜在特质支配, 于是我们就可对这种特质进行界定, 然后据此估计出该被试这种特质的估计值, 并根据其估计值来预测、解释被试对项目或测验的反应. 因此项目反应理论主要用于建立各种与数据拟合的模型, 以此确定被试的潜在特质值和他们对项目的反应之间的关系.

2.1.2　经典测验理论的局限性

通过一定的测量工具 (如测验量表和测量仪器) 进行测量, 在测量工具上直接

获得的值 (读数), 叫观测值或观察分数. 由于有测量误差存在, 所以观察值并不等于所测特质的真实值, 即观察分数中包含真分数和误差分数. 而要获得真实分数, 就必须将测量的误差从观察分数中分离出来. 为了解决这一问题, 真分数理论提出了三个假设:

第一, 真分数具有不变性;

第二, 误差是完全随机的;

第三, 真分数与误差分数的和是观测分数.

在上述三个基本假设的基础上, 真分数理论有以下两个重要推论: 第一, 真分数等于实际得分的平均数; 第二, 在一组测量分数中, 真分数的变异数 (方差) 与误差分数的变异数 (方差) 之和等于实得分数的变异数 (方差). 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、效度、项目分析、标准化等基本概念. 经典测验理论的理论体系很完善, 是其他测验理论赖以产生的基石. 优点主要有以下四点:

第一, 理论方法体系相对完整;

第二, 前提假设比较弱;

第三, 所涉及的数学模型及参数的概念和估计方法易理解和掌握;

第四, 标准化技术在控制测验误差等方面有明显的效果.

但是, 经典测验理论在理论体系和方法体系方面存在许多其本身难以克服的缺点, 具体表现为: 测验结果拓广的有限性. 心理与教育测量都是间接测量, 心理与教育测量的方法又宛如实验: 测量中主试记录下被试的反应向量, 这些反应向量虽不是直接测量, 但提供了推断测值的依据. 与其他实验一样, 要获得对测值的有效推断, 必须抑制影响实验变量的各种误差变量. 实验设计理论列出了三种应对实验误差的基本方法: 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实验结果的差异. 随机化技术的应用可使误差变量的影响不能在实验结果上形成系统误差. 统计调整技术建立在数学模型的基础上, 将误差变量的影响参数化, 从而在实验中调整参数估计值, 减小误差变量的影响. 经典测验理论主要应用的是配对或标准化技术和随机化技术. 然而, 使用配对或标准化技术的测量结果仅能在相同的测量条件下成立, 却不能将其拓广到非标准化环境中, 这使得测量的应用受到很大的限制.

测验分数的测验依赖性. 经典测验理论控制误差应用标准化技术, 但其标准化的对象是测验的各种外部变量, 对测验的内部变量即测验的项目的 "性质" 这一变量却没有也不能实现标准化. 这就使得设计用来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, 其值一般都是不等的. 因为每一个测验包括了它独特的项目集, 并且每一项目的性质也不相同: 从测量学观点看, 项目的这种性质是实验中的 "噪声" 或者是逃脱了标准化的误差变量. 这一事实造成了

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". 这种处理方法, 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.

统计量的样本依赖性. 经典测验理论构造了一个完整的理论体系, 同时设计了一套参数指标来刻画测量的各方面特性. 这些指标中最主要的就是测验的信度、效度和测验项目的难度、区分度. 要施行高质量的测验离不开对这四个 "度" 的准确估计. 但是在经典测验理论中, 这些参数的估计对样本的依赖性是很大的, 最明显的例子就是项目难度. 对于同一项目, 若样本的群体水平较低, 就有较高的难度估计值; 若样本的群体水平较高, 则又会有较低的难度估计值. 项目区分度从本质上讲是样本群体的项目分数与测验总分之间的相关系数. 众所周知, 相关系数的估计受样本全距的影响很大. 样本全距宽, 相关系数值大; 样本全距窄, 相关系数值小. 测验的信度和效度也主要通过相关分析估计, 因此同样受到样本全距的影响. 经典测验理论为避免抽样偏差对参数估计的影响, 特别强调样本对总体的代表性. 但经典理论应用的是随机抽样, 随机抽样的偏差总是存在, 有时还会很大. 更何况在实际操作中, 由于客观条件的限制, 有时还做不到真正的随机抽样. 参数估计值的这种样本依赖性使得所估参数对测验的分析仅具有限价值.

信度估计的不精确性. 信度是测验质量的重要指标, 从本质上讲, 信度是测量随机误差的指标. 心理与教育测量不能没有测量误差的估计指标. 但在经典测验理论中, 测验信度的估计却是很不精确的. 批评经典测验理论信度估计不精确有两重含义: 其一是指估计方法的不精确. 在经典测验理论中, 根据真分数理论, 将原始分数分解为测验真分数和误差分数两部分, 并且进一步假设误差分数与真分数是相互独立的, 从而导出测验信度是真分数方差与原始分数方差之比. 且不说这些假设的真实性, 就按此结果, 这样定义的测验信度实际上是不能计算的. 因为在这个定义中除原始分数方差实际可得之外, 真分数方差与误差分数方差都是无从求取的. 为实际估计测验信度, 经典测验理论又提出了平行测验概念或条件稍弱的 τ 等价测验概念, 从而推演出若干信度估计公式. 但是严格平行的测验是不存在的, τ 等价的测验也是很难获取的, 在此基础上估计的测验信度很难达到比较高的精确程度. 测量的重要目标就是要提高测验质量, 降低测验误差, 而作为测验误差指标的测验信度在经典测验理论中却首先得不到精确估计, 应该是一个缺陷. 批评经典测验理论信度估计的不精确性的第一重含义是指估计值的笼统性. 在经典测验理论中, 每个测验都只有一个信度值, 换句话说, 每个测验都相同的测验误差, 相同的测验施测于不同的被试也会有不同的测量误差. 事实上, 一份测验只有施测于能力水平与测验难度相当的被试时才容易取得比较高的测量精确度. 当测验施测于能力水平高于 (或低于) 测验难度的被试时样品就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. 对测量活动中这一客观现象, 经典理论

只用一个测量误差值来解释, 是难以令人满意的. 显然, 对大多数被试来说, 一个误差值对他们的描写, 不是偏高, 就是偏低. 另外, 测验由项目组成. 各项目在测验总信度中作用如何, 经典测验理论也无法问答, 所以是笼统值.

能力量表与难度量表的不一致性. 测验工作者应用测验试题测量被试水平, 显然应该选择最适合被试能力水平的试题才有针对性. 但是在经典测验理论中被试能力量表是测验的卷面总分, 项目的难度量表是被试群体的得分率. 被试卷面总分的参照系是测验的全部项目. 被试得分 80 表示该被试答对了全部项目的百分之八十. 项目难度量表的参照系是被试群体, 项目难度 0.80 表示有百分之八十的被试答对了该项目. 由于两个量表的参照系数完全不同, 我们无法判断难度为 0.80 的项目是否就与得分 80 的被试水平相匹配. 换句话说, 在经典测验理论的参数指标中, 找不到验证某个项目是否恰好匹配被试某种能力水平的计量方法, 这就使得在测量选题时带有一定的盲目性, 究其原因, 就是难度量表与能力量表的不一致性, 虽然两个指标各自的意义都非常清楚, 但是由于没有把它们定义在同一个参照系上, 所以失去了精确指导测验编制的作用. 更深入的研究可以发现, 经典测验理论的所有项目参数与被试能力参数之间的关系都是非常含混泛化的. 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将如何, 结果分数将会是多少及测量的误差将会有多大, 都是事先无法估计的. 这种现象说明, 经典测验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.

经典测验理论的这些局限性在概括力理论中仅有个别得到了一些改善, 而大多数还依然存在, 原因为: 从本质上讲, 概括力理论与经典测验理论同属于随机抽样理论, 而概括力理论并未改良经典理论的微观结构, 也就是没有改善经典理论的项目参数系统. 概括力理论更多的是从整个测验的宏观结构及其与外部测验条件的关系上作了较深入的计量分析. 因此经典理论的一些主要局限性依然存在.

随着政治、经济、文化的发展, 当代社会需要开发出功能更为齐全, 适应面更为广泛, 测量精度更高的测验. 对于这种需求, 经典测验理论已经显得力不从心. 项目反应理论就是在这种需求的刺激下, 在批评了经典测验理论的局限性的基础上发展起来的一种现代测验理论. 项目反应理论比较成功地应用了实验误差控制的统计调整技术, 在测验的较为微观领域即测验项目上开展研究, 建立项目反应模型, 将测验项目 "性质" 对测量的影响参数化, 再通过模型控制这些参数, 从而达到控制测量误差的目的. 项目反应理论的兴起和发展受到了测量理论和实践工作者的高度重视, 对心理与教育测量的发展起到了很大的促进作用.

2.2　潜在特质理论简介

在日常生活中不难发现, 人们的行为举止就好像处于某些心理品质的定量控

制之中, 甚至觉得好像是这些心理品质实际上决定了他的一切行为, 这是诱惑心理学家研究人类心理品质的起因. 但是至今没有任何迹象证明这些心理量存在于人的物理或生理知觉之中. 心理学上把这类制约人的行为的心理特征称为心理特质, 同时这种心理特质并没有明确它的物理与生理属性, 因此又被称为潜在特质 (latent trait). 如此定义的潜在特质仅是一种统计结构, 并不能说明它是一种物理的或生理的实体.

心理与教育测量的任务就是要定量地估计个体在每一种这样的潜在特质量表上的位置, 然后又根据所估个体的特质位置去解释或预测个体在类似境况下将会产生的行为反应. 在认知测量中, 潜在特质通常被称为被试能力 (应该注意到它与理论心理学常用的能力概念的区别). 但是, 人类的这些心理特征或直接称其为潜在特质的潜在性 (即物理、生理属性不明), 至今还未被它的主体直接探明, 这就给心理与教育的测量带来很大的困难. 测量学家只能借助于一些可观察的间接变量来鉴别与定义这些潜在特质, 并且也只能用同样的方法来探查: 在约束已知行为发展的过程中, 有哪些潜在特质起了比较重要的作用? 用这样的方法来考察某种潜在特质将对人的哪些行为发展产生重要影响?

以上所述构成心理测量研究中潜在特质理论的基本内容.

心理测量学进一步将潜在特质数学模型化. 心理测量学将其定义为: 对于某一特殊行为的发展起作用的所有潜在特质的集合称为潜在特质空间 (latent trait space). 在潜在特质空间中, 互相独立的潜在特质的个数, 称为这个特质空间的维度. 潜在特质空间可能是多维的, 也可能是单维的. 一个 k 维的潜在特质空间可以用向量的形式表示为

$$\boldsymbol{H} = (\theta_1, \theta_2, \theta_3, \cdots, \theta_k).$$

包含决定某一行为发展的所有潜在特质的特质空间称为全特质空间. 全特质空间的维度也是有高低的, 其数值完全取决于所研究行为的性质. 特质空间的维度越高, 研究越困难.

心理测量学者首先关心的是能否查明潜在特质空间的维度, 查明各维特质在决定人的行为时所作的贡献的大小. 心理测量学者更关心的是能否估计出个体在这些潜在特质上的位置, 并且能否预测具有特定的特质位置的个体其行为发展的方向和水平. 这些任务实际上是心理测量学研究的主要内容, 潜在特质理论实际上是一切心理测量理论研究的基础, 只是在应用潜在特质理论时各自的角度和起点及其结果的明晰度不同罢了.

项目特征曲线 (item characteristic curve, ICC) 以认知领域的测量为例. 我们都有这样的经验: 一道编制得较好的试题, 被试在其上正确作答的概率会随着被试的测验总分的增大而提高. 也就是说被试总分高, 在试题上正确作答的概率也高; 被

试总分低, 在试题上正确作答的概率也低. 试题正确作答的概率与被试总分之间呈正相关. 如此, 沿着被试总分的由低到高, 对试题正确作答的概率形成一条不降曲线 (或称单调上升曲线), 这就是试题正确作答率对测验总分的回归曲线. 应用实验数据资料是很容易获得这样的一条曲线的, 但是应该注意到, 被试测验总分是一种随测验特性而变的分数量表, 同时也很容易受到抽样的影响, 是一个很不稳定的被试水平描写量. 因此, 根据被试卷面总分求得的这一条回归曲线复杂、易变, 不是对测验项目特性的良好刻画. 若用能稳定反映被试水平的潜在特质量表分数代替被试卷面总分作为回归曲线的自变量, 就可求得被试在试题上正确作答概率对潜在特质分数的回归曲线, 这条线就被称为项目特征曲线.

下面来分析一下项目特征曲线的形态特点. 首先, 人的潜在特质量表应该是定义在正负无穷的区域内的. 尽管特质地位处于极端水平的被试极少, 但理论上还是存在的. 其次, 被试在试题上正确作答的概率, 记为 $P(\theta)$, 无论其处于什么特质水平上, 取值都在 [0,1] 之内. 再次, 如果试题的测验质量较好, 则被试正确作答概率应随被试特质水平的提高而提高. 因此曲线的图像不可能是一条直线, 而只能是一条从负无穷到正无穷的递增曲线. 图 2.2.1 是几种可能的项目特征曲线形态. 其中 (a) 图是一种阶梯形曲线, 其概率取值只有 0 和 1 两点, 因此也称“全或无”曲线. 1944 年戈特曼 (Guttman) 提出的“无误差模型”也称为“理想量表模型”的图形, 就是 (a) 图形状, (a) 图形状是项目反应特征曲线的雏形. 图 2.2.1 中的 (b) 图形状有三段, 两头的两段与 (a) 图意义相同, 中间一段是一倾斜的线段. 此图描写的是被试正确作答概率在某个范围内随值 θ 连续按比例递增, 而在该范围之外, $P(\theta)$ 的取值在低段为 0, 在高段为 1. 如果斜率值取 0, 表示试题答对概率与我们所定义的潜在特质没有任何关系. (c) 图是一条 S 形光滑曲线, 下渐近线为 $P(\theta)=0$, 上渐近线为 $P(\theta)=1$. 在曲线中部其斜率达到最大值, 越往两端, 曲线斜率越小, 类似于一条概率分布累积曲线. 该曲线的变化有两种, 一是上下渐近线的取值变化, 二是曲线中部的斜率变化. S 形曲线是最常见的项目特征曲线形态, 项目反应理论中一些基础反应模型都采用 S 形曲线.

以认知测量为例, 无论是测验编制者还是测验使用者都有这样的经验, 那就是对于一道编制质量好的题目, 全卷总分较低的被试在试题上的正确作答概率较小, 而全卷总分较高的被试在该题目上的正确作答概率相应较大, 这种伴随着总分的由低到高, 题目正确作答概率由小到大的变化基本上是一种连续性变化, 因此形成了一条从低分到高分的不降曲线, 这就是题目正确作答率对测验卷面总分的回归曲线, 测验卷面总分是一种随测验特性而变的分数量表, 使得题目对总分的回归曲线形态趋向复杂, 形成不了对题目性质的独立描写. 不同形态的项目特征曲线应该有不同的项目特征函数. 有时即使是同一种形态, 也有不同的解析式.