认知诊断评估中的题目-属性关系

喻晓锋 著

江西高校出版社
JIANGXI UNIVERSITIES AND COLLEGES PRESS

认知诊断评估中的题目-属性关系

喻晓锋 著

江西高校出版社
JIANGXI UNIVERSITIES AND COLLEGES PRESS

图书在版编目(CIP)数据

认知诊断评估中的题目－属性关系/喻晓锋著. --南昌:江西高校出版社,2023.11
ISBN 978-7-5762-4316-1

Ⅰ. ①认… Ⅱ. ①喻 Ⅲ. ①心理测量学—研究 Ⅳ. ①B841.7

中国国家版本馆CIP数据核字(2023)第213303号

出版发行	江西高校出版社
社　　址	江西省南昌市洪都北大道96号
总编室电话	(0791)88504319
销售电话	(0791)88522516
网　　址	www.juacp.com
印　　刷	江西新华印刷发展集团有限公司
经　　销	全国新华书店
开　　本	700mm×1000mm　1/16
印　　张	16.25
字　　数	264千字
版　　次	2023年11月第1版
	2023年11月第1次印刷
书　　号	ISBN 978-7-5762-4316-1
定　　价	68.00元

赣版权登字－07－2023－809

前言

“认知诊断评估中的题目-属性关系”是江西省高校教改重点课题“心理统计测量方向本科生编程课程教学模式研究与实践(JXJG-19-2-13)”、国家自然科学基金项目“复杂测验情境下认知诊断的关键技术问题研究(32360208)”和江西省教育科学“十四五”规划2021年度课题“智慧学习和智慧测评关键技术研究(21YB027)”的研究成果。该研究旨在帮助教育统计与测量专业的学习者、使用者和研究人员系统地学习现代测量新技术的理论、方法和技术,促进现代测验新技术的应用和发展。

现代心理与教育测量理论和技术有了长足的发展,在很多方面有重大和关键的革新。认知诊断作为心理与教育测量理论的新发展,具有经典测量理论和项目反应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不足,如理论假设很难实际界定和操作、参数严重依赖于样本、项目特性与被试特性之间没有建立内在联系,等等。项目反应理论在被试针对性指导方面也做得不够好,而认知诊断则很好地解决了这些问题,因此在实践指导和应用中具有更强的生命力。

然而,认知诊断在实际测量中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则不容易掌握和使用这门方法和技术。这使得认知诊断的优势无法转变成现实,其推广应用范围不能与它本身的优势相匹配。

笔者希望能够将近些年在心理和教育测量领域的理论和实践与大家分享;希望能够将现代测量学理论、方法和技术用更通俗的语言进行推广,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:

一、内容基础性和与时俱进。所有内容为认知诊断测验有关的基础性内容,但是又与时俱进,为读者理解认知诊断理论中的模型、技术和方法,打下良好的基础。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。

二、内容叙述过程更加通俗易懂。力求把认知诊断的原理、技术、方法通过文字形式和更加简单的式子呈现给读者。尽量从初学者、未接触认知诊断的心理学或教育学研究者的角度,来剖析认知诊断中的各种原理、技术和方法,并通过日常实例进行讲解。

三、实践应用性。通过多个具体的研究,讲解认知诊断技术在实际应用中的使用过程,并给予详细的说明,让读者在理解内容的同时,能够实际动手操作,加深印象。

四、内容要与国内外发展接轨,概念能反映国际的统一界定。

全书共包括9章内容:第一章主要讲述贝叶斯网与测量模型,及

贝叶斯网作为测量模型的优势；第二章主要介绍与贝叶斯网测量模型有关的理论基础，具体介绍各种常用模型；第三章介绍利用贝叶斯网结构学习得到属性之间的层级结构；第四章介绍贝叶斯网分类器在认知诊断分类中的应用及关于贝叶斯网作为诊断模型的进一步讨论；第五章介绍基于贝叶斯网模型的多级计分诊断测验分类研究；第六章主要讲述测验中 $\boldsymbol{Q}$ 矩阵的验证与估计；第七章主要讲述 $\boldsymbol{Q}$ 矩阵估计研究；第八章介绍多级计分认知诊断评估中的 $\boldsymbol{Q}$ 矩阵验证方法与应用研究；第九章介绍认知诊断测验中的被试拟合研究。

在本书行将付梓之际，特别感谢我的家人在书稿撰写过程中给予我的支持和动力。

目录 CONTENTS

第一章　贝叶斯网与测量模型

1.1　认知诊断

随着社会的发展、教育的普及，教育已经逐步由“精英教育”转化为“普及教育”。教育者不但关注教育结果，而且关注教育过程。在强调测验选拔功能的同时，教育测验的辅助教学与诊断功能逐步受到重视。教育测量者不仅仅只希望从测验中得到被试的一个总分或能力值，他们想要看到每个被试对于知识或技能掌握状态的详细描述。因此，测验应该加强诊断的功能，为教师、被试提供有关被试掌握知识的详情，使教师可以有针对性地编制测试题，进行补救教学；被试也可以有针对性地进行学习，从而提高学习效率。及时准确地对被试进行认知诊断（Cognitive Diagnosis，CD）是教学过程中不可或缺的重要环节，也是智能教学系统研究的主要内容之一。认知诊断评估（Cognitive Diagnosis Assessment，CDA）的目标是测量学生的特定知识结构和操作技能，并提供关于学生认知上的优点和不足的信息。一些研究者已经断定 CDA 是 21 世纪新的测量模式，并呼吁大力开展对 CDA 的研究和使用。被试的知识状态是不可直接观察的，如何准确地对被试进行诊断评价引起国内外教育、心理和人工智能等领域专家的兴趣，美国政府甚至通过立法要求教师提供对学生的认知诊断报告。

认知诊断的本质是从被试的作答数据中去挖掘出被试对相关知识的掌握详情，为教师和被试提供教学和学习的指导依据。一般的做法是：首先找出被试对知识的所有可能掌握情况（Tatsuoka 的规则空间模型中称为理想被试的属性掌握模式，Leighton 等人的属性层次方法中称为期望被试的属性掌握模式），一种知识掌握情况对应一类被试；然后将参加测验的所有被试分别归到某种可能的知识掌握情况上。认知诊断的关键在于如何根据被试的作答数据准确地推断出其知识掌握详情，也就是选择一个尽可能准的分类方法。

贝叶斯网络是用来表示变量间连接概率的图形模式，它提供一种自然的表示因果信息的方法，用来发现数据间的潜在关系。近几年来，贝叶斯网络已经成为数据挖掘和知识发现的一个重要工具，在分类、聚类、预测和规则推理等方

面取得了良好的应用效果。贝叶斯理论给出了信任函数在数学上的计算方法，具有稳固的数学基础。在数据挖掘中，贝叶斯网络可以处理不完整和带有噪声的数据，它用概率测度来描述数据间的相互关系，语义清晰、可理解性强，有助于利用数据间的因果关系进行预测分析。贝叶斯方法正以其独特的不确定的表达形式、丰富的概率表达能力、综合先验知识的增量学习特性等成为当前数据挖掘众多方法中最引人注目的一个。

根据贝叶斯统计技术，贝叶斯网络很方便地将领域知识和数据结合起来。如果要对一个实际问题进行分析，先验或领域知识是至关重要的，特别是当数据是不完全的或数据很难获得的时候。

1.2 认知诊断研究的国内外现状

教育测量的过程就是分析被试作答数据和测验项目，并从中得到有价值的信息的过程。在教育测量学界，学者们提出了不少诊断测验模型。有统计数据表明，到2006年为止，至少已有62种认知诊断模型被开发并被用于认知诊断。这些模型通过被试在试题上的作答反应推测被试的知识状态或心理特质。Tatsuoka的规则空间模型(Rule Space Model，RSM)是较早提出且最有影响力的认知诊断模型之一，它从被试的作答反应入手，推测出被试内部知识结构，从错误反应中得知被试的知识缺陷，但它的 $\boldsymbol{Q}$ 矩阵理论(Q-Matrix Theory)和求理想项目反应模式的理论存在不足。规则空间模型的分类方法较复杂，是否有其他既简便，分类效果又更好的分类方法呢？这很值得探讨。RSM从20世纪80年代开始研究，模型不断完善，后来有好几种认知诊断模型都是应用它的一些概念而发展起来的，如联合(统一)模型(Unified Model)、融合模型(Fusion Model)、DINA模型(Deterministic Input，Noisy "And" Gate Model)、NIDA模型(Noisy Input，Deterministic "And" Gate Model)等。Leighton等人的属性层级模型(Attribute Hierarchy Method，AHM)是RSM的一种变体，该模型将认知心理学和心理测量学相结合，便于开发和分析教育与心理测验。AHM给出不同于规则空间模型的分类法，包括IRT分类方法(方法A、方法B)和非IRT分类方法(人工神经网络法)。陈德枝等(2009)对AHM与RSM诊断准确率进行了比较研究。

Tatsuoka的RSM中的理想反应模式、Leighton等人提出的0/1评分AHM中的期望反应模式和DINA模型中的理想反应模式等都有一个基本假设，即当且

仅当被试掌握该项目所涉及的所有属性，该被试才能答对该项目。这样，掌握一个属性的被试和只差一个属性没掌握的被试在该项目的得分都是0分。这种评分方式容易造成诊断信息的损失，而多级评分可克服这一缺陷。Bolt和Fu认为，精确的认知诊断要求具有更丰富的有关被试在项目上的作答反应的信息。

许多研究者提出实现认知诊断测量的框架。已经有ETS的大批研究者对Mislevy等人的概念上的测量框架“以证据为中心的设计(Evidence-Centered Design，ECD)”和相应的操作框架“四过程模型(Four-Process Model)”进行深入研究。

Millan等人研究利用贝叶斯网对被试进行认知诊断。因为学生的知识状态随学习过程发生变化，所以Reye和Millan等研究利用动态贝叶斯网对学生建模。

迄今为止，国外对AHM的研究都基于0/1评分模型。祝玉芳等对多级评分的AHM进行研究，并且提出新的分类方法，这是与我国测验中某些项目(如证明题、计算题)的多级评分现状相适应的。Mislevy等人将贝叶斯网和Samejima的等级评分模型进行综合研究。Bolt和Fu开发了多级评分的融合模型(Fusion Model)，但是融合模型中未知参数估计特别复杂，且Bolt和Fu报道的诊断准确率不高。

1.2.1　二级计分(0/1计分)的认知诊断模型

线性逻辑斯蒂克特质模型(Linear Logistic Trait Model，LLTM)是较早被应用于认知诊断研究的模型，它是在Rasch模型的基础上发展而来的。Tatsuoka在1983年提出的规则空间模型强调的是$\boldsymbol{Q}$矩阵理论，注重建立起项目和属性之间的关系，将不可观察到的认知属性和被试的内在心理加工过程转化为理想项目反应模式。在此之后，$\boldsymbol{Q}$矩阵理论成为诊断评估研究的重要部分，后续研究者们开发的很多认知诊断模型是基于$\boldsymbol{Q}$矩阵理论而构建的。LLTM和RSM被认为是认知诊断模型中两个具有基础性的模型，之后的研究者们开发的认知诊断模型很多是以这两个模型为基础的。比如多成分潜在特质模型(Multicomponent Latent Trait Model，MLTM)以及一般潜在特质模型(General Latent Trait Model，GLTM)等多个潜在特质模型均是由LLTM发展而来的；而统一模型(Uni-

fied Model,UM)、融合模型(Fusion Model,FM)和属性层级模型(Attribute Hierarchy Model,AHM)则是在 RSM 的基础上发展而来的。除了上述提到的几个认知诊断模型外,常见的认知诊断模型还有 DINA 模型、高阶 DINA 模型(High-Order DINA Model,HO-DINA 模型)、DINO 模型(Deterministic Inputs,Noisy "Or" Gate Model)、LCDM,以及 ACDM(Additive Cognitive Diagnostic Model)等。相对于 RSM 和 AHM 而言,DINA 模型是通过一个更为简洁的参数化模型实现对被试属性掌握模式的诊断,它假设属性间相互独立,这个条件并不总是满足。基于此,de la Torre 于 2011 年提出了一种广义的 DINA 模型(Generalized DINA Model),它是在 DINA 模型的基础上,通过放宽部分假设条件而建构起来的,使模型能够更好地拟合实际数据的情形。

1.2.2 多级计分的认知诊断模型

比较可惜的是,上述这些模型基本上仅适用于二级计分(0/1 评分)数据(Dichotomous Data)。在我们实际的测验情境中,测验中题目的形式往往丰富多样,除了有选择题这样的客观题外,还有论述题、简答题、作文题等主观题。这些题型的数据基本是多级的,这就造成上述的 0/1 计分的认知诊断模型不适用,大大限制了认知诊断在实际中的应用,也限制了认知诊断的进一步推广和发展。

基于此,多级计分的诊断模型的开发得到了国内外许多研究学者的关注。国内外研究者们已经开始将一些 CDMs 拓展到多级计分的情形中去。已有的多级计分认知诊断模型包括基于有序类别属性编码模型(The Model Based on the Ordered-Category Attribute Coding,OCAC),多属性的 R-RUM 模型(Templin,2004)和多属性的 GDMs(The General Diagnostic Models)。Almond、DiBello、Moulder 和 Zapta-Rivera 在 2007 年的研究中指出贝叶斯网络可以定义复杂的任务结构,能够应用于多级评分数据,但未能给出实际应用方法;Bolt 和 Fu 于 2004 年对 0/1 评分的 Fusion 模型(Hartz 等,2002)进行了拓展,但是由于 0/1 评分的 Fusion 模型过于复杂,多级评分的 Fusion 模型就更为复杂,未知参数估计比较困难,从而限制了该方法的进一步推广;Chiu、Douglas 和 Li 在 2009 年提出了属性合分的 0/1 计分 K-Means 聚类诊断法,并通过模拟研究考查了其判准率,发现与参数模型不相上下。多级计分的认知诊断模型在国内同样受到了广

大研究者的关注。祝玉芳和丁树良于2009年提出了基于等级反应模型的属性层次方法,这一方法拓宽了传统0/1评分的AHM,实现了多级计分的AHM;田伟和辛涛2012年将规则空间模型(RSM)应用到了多级计分的项目上,并在此基础上开发了基于MATLAB软件的规则空间模型软件;涂冬波、蔡艳和戴海琦2010年将多级计分模型与DINA模型相结合提出P-DINA模型,它适用于各种评分的数据资料并对HO-DINA模型进行拓展,将模型拓宽至可用于多级评分环境,采用MCMC(Markov Chain Monte Carlo)算法实现了对新模型参数的估计;2003年Chen和de la Torre提出了一种新模型pG-DINA,检验了该模型参数在不同条件下的估计能力,并将其分类精度与改进的G-DINA模型进行了比较,评估该模型的可行性;2013年Sun等人基于项目反应理论的认知诊断方法和基于广义距离指数识别考生的理想反应模式下提出了一种基于广义距离判别的多变量响应检测方法(Generalized Distance Discriminating Method for Test with Polytomous,GDD-P),通过属性模式与理想掌握模式之间的关系来识别属性模式;康春花、任平和曾平飞在2016年为了找出更为合适的评估方法和更为多样化的计分方式,将0/1计分的K-Means聚类诊断法扩展为多级计分聚类诊断法,并通过模拟和实证研究考查了其精确性、稳定性以及影响因素;2016年蔡艳、苗莹和涂冬波在0/1评分的CD-CAT基础上,拓展出了适合多级评分CD-CAT的认知诊断模型及选题策略,为实现多级评分CD-CAT提供了方法支持;2016年Ma和de la Torre提出了一种新的多级计分认知诊断模型——顺序加工的多级诊断模型(sequential GDINA模型,S-GDINA)。该模型是基于G-DINA模型开发而来的,更深入和细致地研究了被试个体的内在心理加工过程。这些模型的提出为丰富认知诊断模型和解决现实教育测量任务做出了贡献。

1.2.3　sequential GDINA模型

为了验证贝叶斯网分类器在多级计分认知诊断模型中的分类效果,本研究在这里着重介绍Ma和de la Torre于2016年提出的多级计分诊断模型S-GDINA(sequential GDINA)。

S-GDINA模型也称为顺序过程模型(Sequential Process Model,SPM),它主要是处理考生在问题解决过程中需要的一系列认知属性。现实的评估测验情况中存在着一些结构性作答项目,比如一些主观题,或是需要多个步骤才能完

成的计算题等，它们往往都是多级计分的。基于此，Ma 和 de la Torre 在使用 G-DINA 模型作为每个类别上的链接处理函数的基础上，提出了 S-GDINA 模型。

假设完成一道题目需要多个步骤，每个步骤都包含一些属性，对被试的评分是根据他们成功完成的连续步骤来评分的。具体来说，假如一道题目总共需要 H 步来完成，那么如果被试第一步就失败了，则被试就属于 0 类（没掌握这道题的任何属性）；如果被试正确作答了第一步，但未通过第二步，则被试被划分为第一类；如果被试正确完成了第二步，但未通过第三步，则被试属于第二类。以此类推，对于这道完成需要 H 步的题目来说，总共有 $H+1$ 种有序类别，即类别 0 到类别 H。

一个测验，如果考查了 K 个属性，那么被试最多会被分为 2^K 种潜在类别（Latent Classes），每一种潜在类别都具有自己独特的属性掌握模式，即 $\alpha_c = (\alpha_{c1}, \alpha_{c2}, \cdots, \alpha_{ck})$，$c = 1, \cdots, 2^K$。如果 $\alpha_{ck} = 1$，表示被试类别 c 掌握了属性 k；如果 $\alpha_{ck} = 0$，则表示被试类别 c 没有掌握该属性 k。借鉴 Samejima（1995）提出的等级反应模型（The Grade Response Model，GRM），定义对于属性掌握模式为 α_c 的被试，将他/她在项目类别上正确作答的概率用公式表示为：

$$S_j(h \mid \alpha_c) = \begin{cases} 1, if\ h = 0 \\ 0, if\ h = H_j + 1 \end{cases} \qquad \text{公式 1-1}$$

其中，H_j 表示项目 j 的总类别数，取值范围为 $(0, 1, \cdots, H_j)$，则被试在 j 项目上得 h 分的概率为：

$$P(X_j = h \mid \alpha_c) = [1 - S_j(h+1 \mid \alpha_c)] \prod_{x=0}^{h} S_j(x \mid \alpha_c) \qquad \text{公式 1-2}$$

$S_j(x \mid \alpha_c)$ 表示属性掌握模式为 α_c 的被试在 h 类别上得分的概率，它通常可以被处理成一个广泛使用的通用的认知诊断模型，也被称之为处理函数，例如 DINA 模型或者 G-DINA 模型。Ma 和 de la Torre 在这里使用了 G-DINA 模型来作为处理函数。

在 G-DINA 模型中，如果项目 j 考查 K_j^* 个属性，那么被试会被分为 $2^{K_j^*}$ 种潜在类别。像 G-DINA 模型一样，在 S-GDINA 模型中，对于每个得分类别 h（即被试在该项目上得 h 分），如果项目 j 考查了 K_{jh}^* 个属性，那么被试的潜在类别会有 $2^{K_{jh}^*}$ 种，K_{jh}^* 表示项目 j 在 h 类别上需要考查的属性个数。α_{ljh}^* 表示在项目 j 上达到 h 类别需要的属性缩减向量，其中 $l = 1, 2, \cdots, 2^{K_{jh}^*}$，比如 α_{1jh}^* 表示被试正确作答第一步的属性掌握模式。则对于属性掌握模式为 α_{ljh}^* 的被试，能够正确作

答 h 类别的概率函数可以用公式表示为：

$$S_j(h \mid \alpha_{ljh}^*) = \Phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \Phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*}\sum_{k=1}^{K_{jh}^*-1} \Phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \Phi_{jh12\cdots K_{jh}^*}\prod_{k=1}^{K_{jh}^*}\alpha_{lk} \quad \text{公式 1 - 3}$$

在上述公式中，Φ_{jh0} 代表的是截距参数，表示被试在项目 j 上未掌握该项目所考查的任何一个属性而答对该题的概率；Φ_{jhk} 代表的是属性 k 的主效应，表示被试在项目 j 上掌握了属性 k 而达到该项目 h 类的概率；$\Phi_{jhkk'}$ 表示被试在项目 j 上达到 h 类别上属性 k 和属性 k' 的交互效应；$\Phi_{jh12\cdots K_{jh}^*}$ 表示的是项目 j 上考查的所有属性之间的交互作用。

下面，还以前面的例子来具体解释 S-GDINA 这个模型。假设求解“5 ×6 ÷3 +4 -2 =?”这样一道小学数学的四则混合运算题，该题考查了小学数学中的四则运算的优先级、乘法、除法、加法和减法五个属性。以这道小学数学算术题为例，Φ_{jh0} 表示的就是被试在五个数学运算属性一个都没有掌握的情况下答对这道题的概率；当 $k=1$ 时，Φ_{jh1} 表示的是当被试掌握第一个属性（被试掌握了四则运算的优先级）的时候，正确作答这道题的概率；Φ_{jh12} 表示的是当被试同时掌握乘法或除法时正确作答该题的概率。以此类推，$\Phi_{jh12345}$ 表示的则是当被试把该题所考查的所有认知属性全都掌握了的情况下，正确作答该题的概率。

1.3　基于贝叶斯网模型的诊断测验研究

RSM 中的 $\boldsymbol{Q}$ 矩阵是请学科专家从已编制的测验中抽取出属性后给出测验的关联 $\boldsymbol{Q}$ 阵，丁树良等人于 2009 年发现由 Tatsuoka 在 1991、1995 年所介绍的由测验项目抽取属性层级关系是不可靠的，当属性层级关系定义不准确时，会给诊断分类造成混乱，导致整个诊断都难以正确进行。AHM 的方法是在测验编制之前便要求确定属性间的层级关系。本研究将利用贝叶斯网学习从被试的属性掌握模式中得出属性间的层级关系，这对 RSM 的不足来说是一个很好的补充，也可以对 AHM 中的属性层级关系进行验证；另一方面，将贝叶斯网分类器应用到现代教育测量的认知诊断分类中，将 0/1 计分与 AHM 中典型的分类方法进行比较，将多级计分与祝玉芳 2008 年的方法进行比较，研究何种方法更有利于提高归准率。

第二章　与贝叶斯网测量模型有关的理论基础

2.1　项目反应理论(IRT)

2.1.1　简介

项目反应理论(Item Response Theory,IRT),亦被称作潜在特质理论(Latent Trait Theory),从20世纪60年代提出以来得到很大的发展。随着计算机技术的发展,IRT得以迅速推广和应用。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,以及一些诊断模型,如规则空间模型、属性层次模型等也使用IRT作为分析的理论依据。

2.1.2　项目反应理论的基础模型

不管是经典测验理论(Classic Test Theory,CTT)还是项目反应理论,它们的核心都是数学模型,所有这些模型都是建立在一定假设基础之上的,是反映被试在测验中观察不到的能力水平和观察到的反应之间的数学函数关系。CTT采用的是线性模型(依赖于被试团体),IRT的理论体系则构建于更复杂的数学模型之上,它采用非线性模型,建立被试对项目的反应(观察变量)与其潜在特质(潜变量)之间的非线性关系,这一点更符合事实。项目反应理论中基础模型的发展日趋完善,0/1评分的单维评分模型有线性逻辑斯蒂克特质模型、正态肩形模型、拉希模型等;多值记分单维模型有等级反应模型(The Graded Response Model,GRM)、评定量表模型(The Rating Scale Model,RSM)、称名反应模型(The Nominal Response Model,NRM)、部分评分模型(Partial Credit Model,PCM)和拓广部分评分模型(Generalized Partial Credit Model,GPCM)。此外,项目反应理论的基础模型还有多维IRT模型和非参数IRT模型等,极大地丰富和完善了项目反应理论。