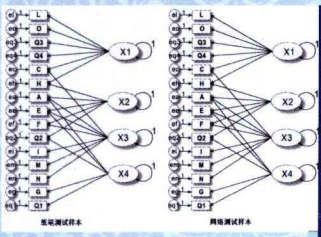
北京航空航天大学人文社会科学文库



基于结构方程模型的测验分析方法

吴瑞林 著

北京大学出版社
PEKING UNIVERSITY PRESS

北京航空航天大学人文社会科学文库

基于结构方程模型的测验分析方法

吴瑞林 著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

基于结构方程模型的测验分析方法/吴瑞林著. —北京:北京大学出版社,2013.3

(北京航空航天大学人文社会科学文库)

ISBN 978-7-301-21990-4

Ⅰ. ①基… Ⅱ. ①吴… Ⅲ. ①教育心理-心理测验-研究 Ⅳ. ①G449.1

中国版本图书馆 CIP 数据核字(2010)第 015398 号

书　　名:基于结构方程模型的测验分析方法

著作责任者:吴瑞林　著

责 任 编 辑:闵艳芸

标 准 书 号:ISBN 978-7-301-21990-4/G·3576

出 版 发 行:北京大学出版社

地　　址:北京市海淀区成府路 205 号　100871

网　　址:http://www.pup.cn

新 浪 微 博:@北京大学出版社

电 子 信 箱:minyanyun@163.com

电　　话:邮购部 62752015　发行部 62750672　编辑部 62752824

出版部 62754962

印 刷 者:三河市博文印刷厂

经 销 者:新华书店

965 毫米×1300 毫米　16 开本　9.25 印张　135 千字

2013 年 3 月第 1 版　2013 年 3 月第 1 次印刷

定　　价:22.00 元

前　言

《基于结构方程模型的测验分析方法》是在本人博士学位论文研究的基础上开始写作的。但与学位论文不同，这本书没有描述具体的技术细节、公式推导和仿真实验数据。它更希望从文献综述的角度，系统介绍使用结构方程模型帮助进行测验质量分析和测验开发的方法，尤其是近年来该领域研究的最新进展，本书前三章的内容是结构方程模型和顺序数据分析的一般知识。

第四章到第七章，分别介绍了使用结构方程模型进行因素分析、信度分析、题目的难度和区分度分析、以及测量一致性检验的原理和方法。并且每章中都配有一个实例，更加直观的展示了这些方法的操作过程和结果。书中的例子均使用结构方程模型软件 Mplus 5.1 和 EQS 6.0 完成分析。为了便于读者更好地理解分析的过程和方法，Mplus 和 EQS 软件的指令说明被附在书后。但本书无意于写成一本关于结构方程模型的教科书，那不是作者对本书设定的目标。

本书第八章中，样本量、嵌套模型、报告模型的结果以及打包操作，这几个结构方程模型中的技术性问题被讨论。这几个问题是使用结构方程模型中时必须注意的问题，也是在研究者中存在争议的问题。希望笔者对这些问题的讨论，能够对于推广和规范使用结构方程模型有参考价值。

本书写作时，参考了大量的中外论文和专著，凡是涉及引用了他人的成果和思想时，都尽可能的在文中按照 APA（美国心理学会）写作格式进行了标注。在这里，先要对被引用作品的原作者表达我深深的感谢和敬意，没有你们创造性的工作，就不可能有这本书的诞生。同时，我依然不

能保证没有遗漏掉某些作者的重要论述，对此我表示深深的歉意。

这本书得以出版，首先要感谢北京航空航天大学人文社会科学学院的领导，没有你们的关心和督促，本书介绍的学术成果和思想不可能在如此短的时间内与读者见面。最后，还要感谢我攻读博士期间的三位导师：美国 Notre Dame 大学心理系的 Ke-Hai Yuan 教授，北京航空航天大学高等教育研究所的张彦通教授，北京航空航天大学心理与行为研究所的王建中教授。

吴瑞林

2011 年 8 月

目　录

第一章
测验与测验质量分析

在心理学和教育学以及其他社会科学研究和实践中，测验是一种普遍使用评价和收集数据的方法。本章将就测验的一些基本概念进行说明，并说明为什么要进行测验质量的评估。

1.1 测量与测验

测量的概念出现得很早，甚至早于数学学科的诞生，出于政治和建筑工程的需要，人类早在数千年前就已经开始使用测量的手段（Wright，1997）。英国物理学家、测量学的奠基人之一 Champbell 定义测量为“依据一定的法则使用测量工具对事物的特征进行定量描述的过程”（Stevens，1946）。

心理测量就是依据一定的法则用数字对人的行为加以确定，依据一定的心理学理论，使用一定的操作程序，给人的行为确定出一种数量化的价值。而教育测量则对教育效果或过程加以确定的过程，它以现代教育学、心理学和统计学为基础，运用各种测试方法和手段，对教育现状、教育效果、学业成就及其能力、品格、心理素质等方面进行科学的测定。

提到教育测验，往往容易让人想到考试，显然，考试是一种教育测验。但这样的定义来自于狭义的教育测量角度，它只关注于教育活动的效果，也就是认为只有对学业成就水平的测验（考试）是教育测验。从广义的角度来看，教育测验不仅包括对学生学业成绩和知识水平的测量，还包括

教育领域中其他教育现象的测量(顾海根,2008)。比如对教师教学水平的测量、对学校管理水平的测量、对学生心理健康情况的测量。

当然,任何教育测量都是对人的某种心理特质或态度的测量,所以所有的教育测量也都可以被看作是心理测量。很多的时候,教育测量和心理测量两个概念也很难被严格的区分开,本书所讨论的测验既包括教育测量中的测验,也包括心理测量中的测验,甚至很多方法适合于社会科学研究中的各种量表和问卷。

最后,要说明的是,心理测量和心理测验两个概念并不完全等价,心理测量的概念更加广泛,凡是涉及人的心理活动和心理属性的测量都可以被称为心理测量,但并不都是依靠心理测验的方式进行,比如生物性的皮肤电、脑电波的测量。反之,测验也不一定都是测量,如果测验是一种主观定性描述,这样的测验就不是测量。

1.2 教育和心理测量的特点

无论是心理测量还是教育测量,它们的测量对象都是人的心理活动,这就决定了教育和心理测量比对物体进行物理特征的测量困难很多。因为心理活动隐于体内,不能被直接观察和度量,而且心理现象本身十分复杂、容易发生变化。所以,心理上的测量具有很强的间接性和相对性,它通过观察、评估人们对测验项目反应(行为)的强度或倾向来间接地推测心理变化,而且没有绝对的标准(郑日昌, 孙大强,2005)。这类测量与自然科学的测量相比,受到更多主客观因素的影响,也就包含更多的测量误差。

测量通常包括参照点和单位两个要素,教育和心理测量也不例外,参照点是计算事物量的起点,也就是尺子的零点,而单位是测量工具的基本要求,即尺子的刻度(scale)。20 世纪 40 年代,哈佛大学的心理学家斯蒂芬斯(Stevens)根据测量的精确程度用命名(nominal)、顺序(ordinal)、等距(interval)和等比(ratio)四个词来给心理上的测量尺度分类。著名的《科学》杂志发表他的这一观点(Stevens,1946)后,对测量的这种分类方

法几乎被其后所有的心理和教育测量的教科书所引用。这四种测量尺度所表示的含义分别为：

（1）命名量表，指测量的结果只具有名词意义的属性，这里虽然也用数字对事物进行分类，但此时的数字只是事物的符号，并没有任何数量上的含义。

（2）顺序量表，指按照事物的大小、等级、程度而排列数字的量表。比如比赛的排名，测定物质硬度的莫氏硬度，就是这类测量。顺序量表中，数字只表示等级或程度，它没有相等的测量单位，也没有绝对零点。

（3）等距量表，顾名思义，要求测量单位间的距离相等，也就是说每一个测量刻度所代表的意义相同；所以，等距量表具有可加性，这也是构成真分数测验理论的基础。

（4）等比量表，这是测量尺度的最高形式，它既要求测量单位间等距，又要求具有绝对起点，比如测量长度、重量等物理量时就是这种尺度。但等比量表在进行教育和心理时是无法被达到的。

本质上讲，心理与教育测量的测验属于顺序量表（戴海琦，张锋，陈雪枫，2007；张敏强，1998）。这是因为：（1）从所使用的参照点来看，教育和心理测量的参照点均为相对参照点。例如，学生考试的成绩一般被限定在 0 到 100 的范围里，但学生即便得了零分，也不能认为他的知识水平或者智力水平为“零”。这就决定了心理与教育测量无法达到等比量表的程度。（2）从使用的测量单位看，教育和心理测量的单位远没有达到物理测量单位的成熟和精确。教育和心理测量的测量单位意义很难定义，语文考试中的 1“分”无法与数学考试中的 1“分”在意义上等同。另外，教育和心理测量的单位也不等值，例如，学生答对两道不同难度的选择题，会分别获得 1“分”，但两者的意义却并不完全相等。因此，教育和心理测量的尺度也不满足等距量表的要求。

然而，由于顺序量表不具有可加性，更不能进行乘除操作，为物理学、生物学等自然科学研究开发的成熟统计方法不能直接应用于顺序量表，这将导致教育和心理测验在理论研究和实际应用中受到极大地限制。为了克服这些缺陷，大多数心理与教育测量的分数解释工作被放在等距量

表上进行，也就是说它们获得的测量结果被当作等距数据来处理了，但这并不能从根本上改变教育和心理测量中单位不等距的本质，本书第三章将对该问题有更为深入的讨论。

1.3 教育和心理测验的功能

教育测量与心理测量是两个有所区别，却又相互关联的概念。一般认为教育测量是心理测量的一个分支，教育测验也是心理测验的一个子集。但也有学者认为教育测验与心理测验是完全不同的两个分类，并无从属的关系。

朱德全和宋乃庆(2007)在《教育统计与测评技术》一书中对教育测量进行了如下定义："教育测量是根据教育目标的要求，按照一定的规则对教育活动的效果加以数量化测定的过程，它主要用于对学生精神特性的测量。"显然，这样的定义来自于狭义的教育测量角度，它只关注于教育活动的效果，也就是认为只有对学业成就的测验是教育测量。从广义的角度来看，教育测验既包括对学生学业成绩和知识水平的测量，也包括对教育领域中各种教育现象的测量。比如对教师教学水平的测量、对学校管理水平的测量、对学生心理健康情况的测量。

教育和心理测验具有两种基本功能：一是预测，即通过测验了解个体间的差别，并据此预测不同的人在将来活动中可能表现出来的不同。二是诊断，重点在于测量同一个人的不同特征间的差异，如通过测验判断一个人是语言能力强还是逻辑思维能力强。正是测验的这两大功能，决定了其在社会各领域中的广泛应用。具体来说，又可以将测验的功能分成以下四种：

(1) 选优

随着现代科学的发展，对各种行业的心理适应性和操作准确性的要求都越来越高，仅凭个人经验来选拔人才已经不能满足实际的需要。尤其是选择拔尖创新性人才，智力和心理素质的测评是必然需要进行的项目。测验作为一种科学方法在人事管理、入学考试、优秀人才选拔

中得以广泛应用，它可以预测一个人从事各种活动的适当性，提高人才选优的效率和正确性。常见的各级各类考试往往是出于这种目的而设置的。

（2）“汰劣”

“汰劣”是企业进行招聘中的一个专有名词，它指使用测量的方法首先淘汰那些智力因素、行为风格、职业能力等方面与企业需求不符的人员，然后再对未淘汰的应聘者进行面试、焦点中心活动以及无领袖讨论等方法选择合适的人才。在目前的企业招聘中，这样的方法被普遍使用，而且效果明显。

（3）安置

通过测验可以科学地了解一个人的能力、性格等特点，从而为因材施教或人尽其才提供依据。学校可以对入学的学生按能力分班，工厂也可以将工人安置到与其性格和职业能力相匹配的岗位。

（4）诊断

测验能够将一个人的行为在许多方面进行比较，从而确定其相对长处和短处，找到行为变化的原因。例如，在教育实践中，测验可用于对学生进行诊断，发现差生学习适应不良和学习苦难的原因。诊断功能还体现在临床和咨询过程中的诊断和鉴别，比如诊断情绪不稳定的学生。

1.4 测验质量与测验质量的评估

测验的质量将直接决定测量结果的好坏，所以，在教育和心理测量与评估中，必须追求高质量的测验，好的测验可以提供科学、可靠地教学评价结果，为教育决策提供高可信的依据。高质量的测验必须具有高的可靠性（信度）和有效性（效度），信度和效度是测验结果可靠性和客观性的衡量标准。只有高信度和高效度的测验才有推广的价值，才能发挥测验的效能。因此要先通过试用而对测验的基本特征信度、效度进行鉴定，确认其质量合格后再加以推广。实际上测验就像产品一样，不仅要向使用

者保证质量，而且要在测验使用说明书中对测验的功用、适用范围、使用方法等给出详细的说明。

在经典测验理论中，信度指测量的可靠性或一致性，指观测到的变量中真实值所占的比例；而效度指的是测量的有效性或正确性，即测量到的变量在多大程度上是真正想要测量的内容，即如何解释测验所获得的结果。举例来说信度指尺子的刻度制作的是否准确和一致；而效度表明是在用尺子测量长度，而不是在测量重量。所以，美国教育研究协会、美国心理学会和全美教育测量学会联合制定的《教育与心理测验标准》(*Standards for Educational and Psychological Testing*)(燕妮琴，谢小庆，2003)中要求，规范的测验必须报告其信度和效度的情况。那么，准确的对测验(或测量)的信度和效度进行评估就成为一个关键问题，对测验信度和效度的分析也始终处于教育和心理测量学研究的中心位置。

另外，测验的难度和区分度也是测验质量的重要指标，确切的说，他们是评价每道题目质量的指标。所以，题目难度和区分度的分析也被称为项目分析(或题目分析)。还要指出的是，题目的公平性问题在20世纪的后几十年中被心理和教育测量界大量提及，并被《教育与心理测验标准》和一些专业的考试机构列为必要的检测项目，本书将在第七章中详细地讨论这一问题。

传统的统计方法，如相关分析、回归分析、因素分析和结构方程模型等，都将测量到的顺序数据假定为服从多元正态分布的等距数据来处理(如图1.1所示)。经典测验理论中的传统信度和效度分析方法也同样是建立在连续数据或者等距数据的假设之上。然而，这种假设不符合教育和心理测量非等距数据的属性，给测验的分析引入了更多的误差，降低了测验信度和效度估计的准确性，甚至有可能导致完全谬误的结论。

从教育和心理测量的本质来看，等距数据的假设是不存在的，有必要回归于心理测量数据的本质，测验获得的结果应该被看作顺序数据(ordinal data)，或者被完整的称为有序分类数据(ordinal categorical data)。分类指每道题目有若干个离散的选项，每个选项表示一个分类(category)，这组选项是按照某种态度强弱程度的等级排列。那样，心理和教育测量

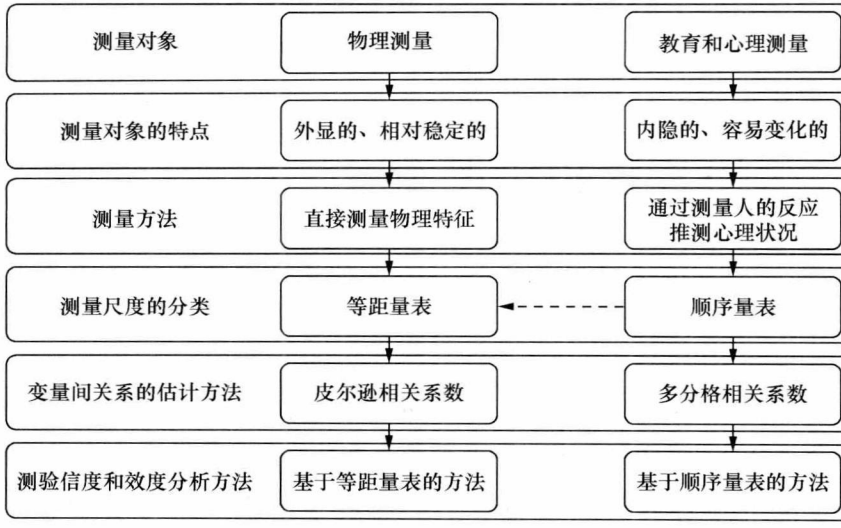


图 1.1　物理测量与教育和心理测量的比较

研究工作者就应该开发出一套能够基于顺序量表估计测验信度和效度办法。

如果能回到教育测验数据是不等距的顺序量表这个事实，就能够更加准确的估计出测验的信度和效度。这样做，在测验的分析技术上是一个大的进步，将提高测验信度和效度评估的准确性和可靠性，从而提高测验的规范化水平和科学性。

第二章
结构方程模型基础

结构方程模型(Structural Equation Modeling,通常简写为 SEM)是近30年来社会科学和行为科学研究中最为重要的多元统计方法之一。本章将简单介绍其基本原理和特点,希望能让不熟悉 SEM 的读者在短时间内对其有一定程度的感性认识。

2.1 什么是结构方程模型

结构方程模型也被称为协方差结构模型[①](Covariance Structural Modeling),其核心概念诞生于20世纪70年代初,在80年代快速发展并走向成熟。结构方程模型的出现与20世纪社会科学研究中的两项统计技术(因素分析和回归分析)有着密不可分的联系,Jöreskog 等人利用数学矩阵巧妙的将这两种统计技术整合于一体,并使用计算机技术,开创出一个崭新的量化研究范式,正式宣告结构方程模型时代的到来(邱皓政和林碧芳,2009)。Jöreskog 与其同事 Sörbom 不仅在理论上提出了结构方程模型的概念,还开发出名为 LISREL(Linear Structural Relationships)的统计软件,将其提出的理念和技术转化为研究者可以直接使用的工具,积极促进了结构方程模型的发展和应用。此后,有关结构方程模型的讨论和使用便蔚为风潮,成为社会科学和行为科学研究者必备的专门知识之一。

① 当模型不考虑均值部分,而只关注模型的协方差和方差部分时,可以被称为协方差结构模型;否则,被称为均值-协方差结构模型。

结构方程模型的核心是提出了“潜变量(latent variable)”的概念,教育科学和心理学研究中涉及的变量,多数不能被直接观测,比如智力、学习动机、家庭社会经济地位等。只能用一些外显的指标(manifest indicator)去间接测量这些潜变量。在以往的统计中,如何处理这些潜变量是个难题,为了简化,观察到的外显变量往往被当作要测的潜变量直接统计。与传统方法相比,结构方程模型则将外显变量与潜变量分开处理,其优势被 Bollen 和 Long(1993)总结为下面五点:

(1) 能同时处理多个自变量和因变量。在以前的统计中,可以允许存在多个自变量,但因变量只能有一个。需要检测多个因变量时要忽略其他因变量的存在而逐个考虑自变量与因变量之间的多对一的关系。结构方程模型可以同时将多个自变量和多个因变量放在同一个统计模型内,考虑的是多对多的复杂关系。

(2) 允许自变量和因变量含有测量误差。由于潜变量概念的引入,每个自变量或因变量可以由多个外显指标来测量。每个测量指标都允许有自己的测量误差,这样,测量误差就被剔除在路径分析之外,路径分析的参数估计值也就更为准确。

(3) 同时估计因子结构和因子关系。在包括偏最小二乘法(Partial Least Square,PLS)在内的统计方法中,测量模型与结构模型相分离,需要先计算潜变量与指标间的因素载荷(factor loading),根据这些载荷计算得到潜变量的测量值,再分析这些潜变量之间的结构和关系。而结构方程模型的方法将这两步同时进行,一次估计获得这两类模型的全部参数。

(4) 允许更大弹性的测量模型。传统上,测量模型只允许每一个题目(或指标)从属于一个因子,在结构方程模型中不再存在这样的限制,一道题目可以同时从属于多个因子,比如 GRE 考试(美国研究生入学考试)中的数学题就可能同时测试数学和英语两种能力。这种复杂的从属关系还可以用于因素分析中存在高阶因子的情况。

(5) 检验整个模型的拟合程度。传统的路径分析中,只能估计单条路径(两个变量之间的关系)的强弱,是否显著在结构方程模型中则是对整个模型与样本数据间的拟合程度进行估计,从整体上来判断哪个模型

更符合数据所呈现的复杂关系。

从统计方法的角度看，结构方程模型是一般线性模型（general linear model）的扩展，它几乎整合之前所有线性关系的统计方法。比如线性回归分析、方差分析、探索性因素分析和验证性因素分析、路径分析、中介效应检验、增长曲线模型、高阶因素检验、跨组分析和纵向追踪数据的分析。除了前述这些线性关系的分析外，它甚至还被用于一些非线性的统计（Lee & Song，2003）。

结构方程模型近年来的应用领域正在逐步扩大，已经成功地用于解决很多社会科学和行为科学研究中的问题。比如，在宏观经济政策研究中，分析两代人之间的职业联系，就业中的种族歧视和性别歧视问题，收入与住房的关系；在医学中，药物使用的后果和不良反应的研究；政治学中，投票选举行为，以及基因和文化对投票影响的研究；市场研究中，消费者购买行为等诸多现象的调查。当然，结构方程模型在教育和心理上的应用更为普遍和广泛。

2.2 结构方程模型的数学原理

完整的结构方程模型包括测量模型和因果模型两部分。遵循 Bollen（1989）所撰写的经典教材（*Structural Equations with Latent Variables*）中的符号，矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别表示自变量和因变量的观测指标，它们与潜变量 $\boldsymbol{\xi}$ 和 $\boldsymbol{\eta}$ 的关系模型为：

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \\ \boldsymbol{Y} &= \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \end{aligned} \tag{2.1}$$

$\boldsymbol{\delta}$ 和 $\boldsymbol{\varepsilon}$ 是测量模型的误差项，$\boldsymbol{\Lambda}_x$ 和 $\boldsymbol{\Lambda}_y$ 分别是观测变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的因素载荷矩阵。潜变量因果关系的模型为：

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2.2}$$

这里，$\boldsymbol{B}$ 是因变量的系数阵，表示因变量间的相互关系；矩阵 $\boldsymbol{\Gamma}$ 表示自变量 $\boldsymbol{\xi}$ 和因变量 $\boldsymbol{\eta}$ 间的关系；$\boldsymbol{\zeta}$ 是潜变量间的随机误差项。

这样，一个完整的结构方程模型的参数由八个矩阵组成：$\boldsymbol{\Lambda}_x$，$\boldsymbol{\Lambda}_y$，$\boldsymbol{B}$，