



*Social Science Research Methodology Series*

社会科学  
研究方法  
丛书

# 测评的 概化理论 及其应用

杨志明 张雷 著

Generalizability Theory  
and Its Applications

Zhiming Yang & Lei Chang

教育科学出版社

Educational Science Publishing House

杨志明 张雷 著

# Generalizability Theory and Its Applications

## 测评的概化理论 及其应用

教育科学出版社

· 北 京 ·

责任编辑 韦 禾  
版式设计 尹明好  
责任校对 徐 虹  
责任印制 滕景云

### 图书在版编目 (CIP) 数据

测评的概化理论及其应用 / 杨志明, 张雷著. —北京:  
教育科学出版社, 2003.5

(社会科学研究方法丛书 / 张雷, 侯杰泰主编)

ISBN 7-5041-2448-6

I. 测... II. ①杨... ②张... III. 考试-研究  
IV. G424.7

中国版本图书馆 CIP 数据核字 (2003) 第 012327 号

---

|      |                    |       |                        |
|------|--------------------|-------|------------------------|
| 出版发行 | 教育科学出版社            | 市场部电话 | 010-62003339           |
| 社 址  | 北京·北三环中路 46 号      | 编辑部电话 | 010-62003343           |
| 邮 编  | 100088             | 网 址   | http://www.esph.com.cn |
| 传 真  | 010-62013803       |       |                        |
| 经 销  | 各地新华书店             |       |                        |
| 印 刷  | 保定市印刷厂             |       |                        |
| 开 本  | 787 毫米×960 毫米 1/16 |       |                        |
| 印 张  | 16.5               | 版 次   | 2003 年 5 月第 1 版        |
| 字 数  | 266 千              | 印 次   | 2003 年 5 月第 1 次印刷      |
| 定 价  | 29.00 元            | 印 数   | 1-3 000 册              |

---

如有印装质量问题, 请到所购图书销售部门联系调换。

## 作者简介



杨志明，1983年毕业于江西宜春师专数学系，1989年毕业于湖南师范大学教育系，获硕士学位，2003年毕业于香港中文大学，获博士学位。先后访学于江西师范大学教育系（1983—1984）、北京师范大学心理学系（1991—1992）和美国大学考试中心（ACT，2002.7—8）。现为湖南师范大学心理学系教授，兼任中国教育学会教育统计与测量分会副秘书长、理论部副主任。2002年入选湖南省首批青年社会科学研究人才百人工程。主要从事教育与心理统计和测量的教学和研究工作，曾主持过教育部考试中心等部门的多项科研课题，发表论文40余篇，主编或参编学术著作6部。



张雷，1988年和1991年先后在美国南加州大学获得硕士和博士学位，主要研究方向为心理统计、心理测验、发展心理等。曾先后执教于美国南加州大学教育心理系、中佛罗里达大学教育学院，现为香港中文大学教育心理系教授。

# 总序

对现代科学尤其是社会科学而言，研究方法的发展在很大程度上能够起到推动整个学科发展的作用，研究方法的落后必然会限制学科的发展。在我国现阶段，同学科专业的建设及发展相比，社会科学研究方法的建设和普及却显得相对薄弱，很多学科的研究方法，尤其是量化方法，还远远落后于世界先进水平，这无形中妨碍了有关学科专业的进一步发展。

社会科学研究方法的落后，首先表现在缺乏一大批热心于研究方法的学者，多数优秀学者往往致力于专业研究，而较少顾及到研究方法的研究。其次表现为此领域相关书籍的匮乏，特别是介绍、讲解现代研究方法的专著不仅数量有限，而且内容陈旧，更没有一套可以系统介绍社会科学研究方法的教材，很多学科苦于找不到合适的方法课用书。而在先进国家，每一种主要的统计分析方法在每一学科中都可找到多种专著及更多的普及读物，以供学者、学生选用。针对这样的情况，我们决定编辑这套社会科学研究方法丛书。

在我国学科建设飞速发展的 20 年里, 社会科学研究方法在西方也经历了一场空前的革命, 20 世纪 70 年代 EM (Expectation Maximization) 算法的出现和计算机技术的发展, 带来了新一代的统计和测量理论及方法。传统的回归及方差分析和经典测量理论, 虽然仍广泛应用, 但已不再是主要的、更不是惟一的研究方法系统, 也无法应付由新一代研究方法带动下的学科专业发展。

新一代统计分析方法最突出的发展是结构方程的发展和应用。到了 80 年代, 结构方程这套新的数据分析系统已经成熟, 目前正为广大社会科学研究人员所接受, 并成为各类社会科学学科研究生的必修课。结构方程的路径分析思想拓展了社会科学研究的思路, 很多学科的专业课题已由过去的只研究单变量转变成研究多变量, 由分析主效应到同时分析交互效应, 由对单指标和直接观测变量进行研究到对多指标和潜变量进行研究。与此同时, 结构方程的技术日趋专业化、深入化和复杂化, 不少学者以此为研究专业, 并且也有专门的学术期刊专注于结构方程及其相关技术的发展和运用。

统计分析方法的另一大突破性发展是多层分析的理论和方法。多层和嵌套分析的思想由来已久, 但直到 20 世纪 90 年代才发展成一套完整而系统的理论和方法, 并逐渐推广和为研究人员所接受。多层分析技术系统地解决了困扰社会科学半个多世纪的生态谬误 (Ecological Fallacy) 问题。在教育、管理、经济以及社会学、心理学等领域的研究中, 取样往往呈嵌套结构, 例如学生嵌套于学校, 学校又嵌套于社区, 这种嵌套型的样本用传统的 OLS 回归方法分析会导致估计误差。多层分析方法不仅可以减少这种统计误差, 而且可以避免由人为选择分析单位而可能出现的错误。在多层分析中各层样本均可作为分析单位, 而且还可以研究他们之间的交互作用, 从而拓宽了各专业的研究范围, 深化了各专业的研究思路。目前多层分析方法日趋成熟, 并在新一代统计分析方法中处于前沿位置。

新一代统计方法的发展还体现在对追踪数据和发展模型的处理上。这一进展的主要特点是借助和延伸已有的方法, 建构一套处理重复数据的具体步骤。目前最为广泛接受的是基于结构方程的追踪数据分析方法和基于多层分析技术的发展模型方法。这些方法通过对结构方程与多层分析的巧妙应用, 而各自形成了一套独立的系统。

此外, 在上述统计方法系统发展的同时, 很多专门的数据处理技术也迅速

改进。分析类别变量 (categorical variable) 的方法就是其中一个例子。传统的统计方法多建立在对连续变量进行分析的基础上, 而对建立在分类资料基础上的社会科学问题则不能很好地加以解决, 分析类别变量的方法则解决了这一难题。另一个困扰研究者的统计问题是如何对缺失值 (missing value) 进行处理, 新一代的基于 bootstrap 等的方法则能更有效、更完整地处理缺失值。

随着新一代统计分析理论及方法的发展、推广与应用, 测量方法也有了质的飞跃。与自然科学相比, 社会科学的主要难题在于难以直接、精确地对人类行为加以测量, 而要依靠测量理论对其进行推断。然而, 传统的经典测量理论不能同时对项目难度和考生能力进行估计, 无法应付大规模自适应考试的要求。随着计算机技术的发展, 经典测量理论在教育 and 心理测量领域的应用逐渐让位于项目反应理论 (Item Response Theory, IRT)。美国和欧洲的著名测量机构均把 IRT 应用于设计、测试集成、测试校准、建构测试题库以及其他的测验发展过程。计算机自适应测验 (computer adaptive test) 的普遍应用更是 IRT 在测量应用方面的一大贡献。

另一个新一代的测量理论和方法是概化理论, 概化理论最近的发展是同人们对考试认识的改变分不开的。越来越多的学者和教育工作者对标准化、多重选择类的考试方法提出了质疑, 表现性评价 (performance assessment) 的思想则逐渐生根。然而, 随之出现的问题是由这种直接操作的测试方法所产生的多种误差无法用传统测量理论来解决, 于是概化理论在这种新环境中的作用逐渐被人所认识并得到发展。此外, 人力资源测评、管理咨询等也推动了概化理论的发展和应用。

这套丛书将对上面提到的新一代数据分析和测量评估的理论及方法陆续进行介绍, 旨在推动我国社会科学研究方法的发展, 改变该领域发展滞后的现状。每一本专著的作者都是该领域颇有建树的专家, 在写作风格上既强调数理及专业技术方面的严谨性, 也兼顾语言和介绍方式上的通俗化, 从而适合具有不同数学背景读者的需要。丛书的另一个特点是对方法应用性的强调, 旨在满足大多数社会科学专业研究人员对应用研究方法的需要。为了加强其应用性, 丛书要求作者从实际操作入手, 以具体研究为例, 准确清楚地介绍各研究方法的操作步骤。丛书的第三个特点是重点突出, 一般只抓住某一方法和理论的中心, 尽量不涉及偏难及纯技术的, 尤其是在理论上尚有争议或技术上不成熟的问题,

从而使初学者能够掌握到要领而又不致过于肤浅，或者感到信息量太大而难于接受、消化。最后，这套丛书的内容均是作者结合自己的研究实践亲笔撰写而就，并非翻译、编纂之举。

这套丛书可以作为教育、心理以及其他社会科学学科的研究生和部分本科生的教材，也可以作为从事社会科学研究的人员的参考书、工具书。我们希望这套丛书的出版能起到抛砖引玉的作用，唤起更多热心于方法研究的学者多出书、出好书，多参与我国社会科学研究方法和理论的建设工作，把我国的社会科学研究方法推向一个新起点。

张 雷 侯杰泰

2002 年 12 月于香港



# 序 1

在心理与教育测评领域中，控制测量误差一直是人们追求的目标之一。中国是考试的故乡，大规模的科举考试历经了 1300 年之久，它在控制测评误差的方式方面积累了丰富的经验，可惜一直没有上升到理论的高度。到 20 世纪初，西方开始有了测验标准化的概念和方法，并逐渐建成了经典测验理论。

经典测验理论是采用标准化和随机化的方法来控制测评误差的。但实际的测评活动往往受到其他一些条件的影响，而很难达到标准化的要求。因此，几乎在经典测验理论提出的同时，Cronbach 等人就开始尝试使用把测评的条件因素作为模型参数的所谓统计调整技术，并十分重视测评误差与决策需要之间的关系，初步形成了测量的概化理论。以后又经过 Brennan 等许多学者的努力而逐步使之走向成熟和规范。

杨志明和张雷博士的这本书是我国第一本探讨概化理论

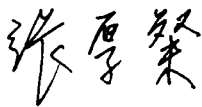
及其应用的著作，具有以下主要特点。

1. 深入浅出，易于理解。由于概化理论引入了诸如测评侧面、观测全域、概化全域等大量的新概念和新名词，所以初学者往往对之有畏难情绪。作者采取与经典测验理论相比较的方式来阐述概化理论的基本概念和方法，有助于读者抓住要点，理解实质。

2. 结合理论，介绍方法。在阐述理论模型的同时，介绍了常用的测量设计中所用数据估计方法的 SPSS 命令语句，这样做既不妨碍对理论模型的准确表述，又方便了读者的学习。

3. 联系实际，内容全面。作者详细讨论了一元概化理论的原理和方法，还结合实例对在智力测验、普通话面试、人才测评以及教学评价等多方面的应用作了分析，并初步介绍了多元概化理论。这样，既突出了重点，又有助于读者了解概化理论的全貌和最新发展。

杨志明教授自 1991 年在北京师范大学做访问学者开始，一直钻研概化理论，并尝试将它应用于教育考试、心理测验、人才测评等诸多领域。自 1999 年开始，他又到香港中文大学攻读博士学位，师从张雷教授，继续从事有关概化理论的深入研究，很快取得了良好的成绩，写出这个领域的第一本著作，愿为做序，以示祝贺。



(国际心理科学联合会副主席，  
北京师范大学心理学系教授)

2002 年 9 月

## 序 2

Foreword

by

Robert L. Brennan

University of Iowa

for

***Generalizability Theory and  
Its Applications***

by

Zhiming Yang and Lei Chang

In 1972, Wiley published a monograph by Cronbach, Gleser, Nanda, and Rajartnam entitled *The Dependability of Behavioral Measurements*. That book incorporated, systematized, and extended their previous research into what come to be called *generalizability theory*, which liberalizes classical test theory, in part through the application of analysis of variance procedures that focus on variance components. Generalizability theory perhaps the most broadly defined measurement model currently in

existence, and the Cronbach et al. (1972) treatment of the theory represents a major contribution to psychometrics. However, their book is 30 years old, and is out of print.

In 1983, ACT, Inc. Published my monograph entitled *Elements of Generalizability Theory*, with a slightly revised version appearing in 1992. That treatment is considerably less comprehensive than Cronbach et al. (1972) but still detailed enough to convey much of the richness of the theory. In 1991, Sage published an excellent monograph by Shavelson and Webb entitled *Generalizability Theory* that provide a very readable introduction to the theory. In 2001, Springer published my book entitled *Generalizability Theory*. It provides a comprehensive and up-date treatment of generalizability theory that covers all of the major topics that have been discussed in generalizability theory, as well as some new ones.

A distinctly limiting feature of all these references, however, is that they are available only in English, but increasingly generalizability theory is being used throughout the world. There a few non-English treatment of the theory including a monograph in France by Cardinet and Tourneuer entitled *Assurer la mesure*, but to the best of my knowledge there has not been a Chinese monograph or book written on generalizability theory until this one.

I am delighted that Yang and Chang, two dedicated scholar, accepted the challenging task of writing this book, and I believe the authors have chosen an excellent set of topics and examples to cover. In the past two decades a number of Chinese students and scholars have contacted me with numerous questions about generalizability theory, and I have often been at a loss to help them very much because of the English-Chinese language barrier. I am hopeful, therefore, that this book will bridge that gap and serve to expand the use of generalizability theory in areas where the Chinese culture dominates. Perhaps this will even help to facilitate greater communication among scholars in various areas of scientific inquiry.

Iowa City, IA  
October, 2002

**Robert L. Brennan**

## (序 2 译文)

1972 年, Wiley 公司出版了 Cronbach, Gleser, Nanda 和 Rajartnam 的《行为测量的可靠性》专著。该书在系统总结和拓展他们前期一些研究成果的基础之上,把概化理论组合成了一个整体。从一定程度上讲,通过应用方差分析技术分解变异分量的方法,概化理论使经典测验理论获得了新生。或许,概化理论是当今存在的最广义的一种测验理论,该理论本身以及 Cronbach 等人 (1972) 关于它的研究,可以看做是对心理计量学 (Psychometrics) 的一项重大贡献。不过,他们的专著是 30 多年前出版的,而且已经绝版了。

1983 年,美国大学考试公司 (ACT, Inc.) 出版了我的专著《概化理论纲要》(Elements of Generalizability Theory),后略经修改于 1992 年再版。这两本书比 Cronbach 等人的著作 (1972) 略微简略了一些,但它们对概化理论的丰富内涵仍然作了足够详尽的阐述。1991 年, Sage 出版公司出版了 Shavelson 和 Webb 合写的《概化理论》。这是一本极为出色的、通俗易懂的概化理论入门著作。2001 年,我的专著《概化理论》由 Springer 出版公司出版。该书涵盖了概化理论研究领域中讨论过的所有重大课题,反映了概化理论的最新研究成果。

不过,所有这些研究有一个显著的缺点,这就是它们全是英文文献,可概化理论的应用范围已经遍及全球。据我所知,目前仅有 Cardinet 和 Tourneuer 用法语写的《Assurer la mesure》等少量的非英语版本的概化理论专著。在中国,除了本书之外,目前还没有一本概化理论方面的专著。

令人欣慰的是，杨志明和张雷两位学者选择了这一挑战性的工作，写出了这本中文版概化理论著作。相信他们在书中选择了很好的主题和实例。在过去的二十多年中，许多中国的学生和学者曾经与我进行过联系，询问过很多有关概化理论的问题，但是，由于中文与英文之间沟通上的困难，我常常帮不上他们多少忙。因此，我希望这本书可以起到在英文与中文之间进行沟通的桥梁作用，使得概化理论能够在讲中文的地区得到更好的应用和发展。或许，它还有助于不同领域内从事科学调查研究的学者之间进行交流与合作。

罗伯特·L. 布瑞南

2002 年 10 月

于美国衣阿华大学

# 前 言

在评价人的能力或某种素质时,必须使用科学的方法。考试方法(包括笔试和面试)是一种常用的方法,此外还有评价者中心法、无领导小组法等等。这些方法在我国的高等学校入学考试、企业和事业单位的人才招聘以及竞争上岗等选拔性测评工作中,在中学生的学业会考、企事业单位员工的绩效考评以及各种各样的证书考试等达标性测评工作之中,都被大量采用。

对于这些测评工作,人们通常关注的问题主要有两个。其一,测验分数能不能有效地把应试者(被试)的真正水平区分开来(体现在选拔性测评工作中)。其二,测评分数是不是应试者真正水平的一个恰当估计(体现在达标性测评工作中)。为研究方便,测量学把用于第一种目的的测验称为常模参照性测验(Norm-referenced Test),把用于第二种目的的测验称为标准参照性测验(Criterion-referenced Test)。

无论使用哪种方法,如何保证测评质量的问题都是其首要问题。若因为测评质量不高,造成人才选拔不当或者员工的绩效考核不公等后果,则会给国家和企事业单位的工作带来极大损害。

一般来说,保证测评质量的主要方法是控制测评误差。控制测评误差的方法主要有三种,即:匹配或标准化技术、随机化技术以及统计调整的技术。其中,标准化技术旨在从测量工作的各个环节上控制测量误差,使得“测量误差不会干扰被测目标群体内个体间先后顺序的排定工作”。随机化技术旨在控制测量的系统误差。统计调整技术则是把误差变量作为模型的参数来对待,从而清楚地了解误差变量对测量目标的影响程度,以便在一个可接受的误差范围内来解释测评结果。

在各种控制测评误差的方法中,标准化技术用得最多。如,在各种大规模的常模参照性测验之中,以经典测验理论(Classical Test Theory, CTT)为指导,通过大量采用像托福(TOEFL)考试一样的测验标准化技术,较好地控制了测评误差,达到了考试的目的。当然,这种方法付出了高昂的代价:费时、费力建题库,并牺牲了对创造性思维的测评,等等。但是,对于一些没有题库或没有必要建设题库的临时性的常模参照性测验,以及所有的标准参照性测验来说,标准化方法很难实施。这就使得测量误差的控制工作变得十分困难,因此需要寻找新的方法。

概化理论(Generalizability Theory, GT或G理论)是一种能够同时达到区分考生与评估应考者真正实力之目的,并较好地控制测评误差的现代测量理论。它控制测量误差的办法主要是增加了统计调整技术。即,把干扰测验分数的无关变量或因素引入测量模型之中,然后用统计技术分别估计出这些因素或因素之间的交互作用对测验分数的影响程度。剔除这些影响后,被试之间水平的真正差异便能得到非常清晰的显现,即达到了控制测评误差的目的。

概化理论的基本原理是:首先运用实验设计的思想,分析影响测验分数变异的各种来源(如,被试水平的差异、题目的难度、评分者的评分标准等)。接着,运用方差分析的技术,分别估计各种变异来源对分数总变异所作的贡献(通常用方差分量作为指标)。然后,根据不同的研究目的的需要,分别考察研究目标在测验总分变异中所占的比重。一般地,当测量目标引起的分数变异所占比重较大时,测量被看做是具有较高信度的。不过,测量信度的概念在概化理论中为概化系数(Generalizability Coefficient)或可靠性指数(Dependent Index)所代替。

概化理论的一个突出特点是:对同一次测量,可以根据研究目的的不同提供多个测量信度。事实上,当测量目标发生变化(如,测量目标由被试水平改



为评分者素质)时,或者当测量结果被推论的范围不同时,测量信度都会发生变化。为了理解这一特点,我们来简要考察以下一个实例。

众所周知,王勃是一位才华横溢的文人,因为他曾经写下过《滕王阁序》这样的不朽名篇。要是一位中学生也写出了一篇让老师感到非常出色的记叙文,比方说其语文老师给他的作文评定了98分,我们是否可以说他也是才华横溢呢?显然,仅仅凭任课老师给学生的一篇记叙文打了98分就来推论其才学水平很高是太夸张了点。

这里实际上涉及到测评活动中行为样本与行为总体之间的关系问题。事实上,在测评活动中,被评价者的部分行为表现(一篇文章)与他的总体表现(有才气的文人)并不是一回事,个别人的一时一地的评价(任课教师)与大众的无论何时何地的认可也不是一回事。即,行为样本与行为总体、条件样本与条件总体之间的关系并不一定是完全匹配的。这种样本(包括被测的行为和测评时的条件)与总体之间的匹配关系会直接影响测评工作的准确性和科学性。这些关系要是处理得好,测评结论就能够令人信服,否则,结论便会被怀疑甚至遭到否定。

王勃由于他能写出得到世人广泛赞誉的《滕王阁序》,所以我们可以把他评价为杰出的文人。这种评价用概化理论的思想来剖析就是:第一,行为样本(写出《滕王阁序》)是行为总体(杰出的文人)的一个很好的代表;第二,对行为样本作评价时的条件样本(被当时在场众人誉为佳作)与对行为总体作评价时的条件总体(后人的广泛赞誉)的关系是相匹配的。假若我们以此(行为样本)为根据把他说成是杰出的政治家和军事家(行为总体),则会显得荒唐可笑。因为该行为样本(写出《滕王阁序》)与行为总体(杰出的政治家和军事家)之间是不匹配的。其推论的范围过大,(测评)误差大得让人难以接受!同样,假若该文仅仅在当时受到赞誉(在条件样本上作出的评价)却遭到了后人的广泛批评(条件总体上的评价),则说明当时的赞誉本身就是不恰当的。这也同样意味着(测评)误差过大。

其实,在各种各样的测评活动中,把测量结果概括到不同的程度,其测量精度是会大不相同的,二者之间存在着此消彼长的关系。即,测验分数推论的范围越大,其测量精度就会越小(测量误差增大,测量效度降低)。反之,测量精度就会提高,误差就会缩小。例如,当我们把那个中学生的98分仅仅解释为写记叙文水平时,测量误差相对较小。当我们把这个分数推论到更大范围(写