

万卷方法

# 量表编制： 理论与应用

(第2版)——校订新译本

SCALE DEVELOPMENT:  
THEORY AND APPLICATION

罗伯特·F.德维利斯 著

魏勇刚 席仲恩 龙长权 译

李红 校



重庆大学出版社

<http://www.cqup.com.cn>

# 量表编制： 理论与应用

(第2版)——校订新译本

SCALE DEVELOPMENT:  
THEORY AND APPLICATION

罗伯特·F.德维利斯 著

魏勇刚 席仲恩 龙长权 译

李 红 校

重庆大学出版社

Authorized translation from the English language edition, entitled SCALE DEVELOPMENT: THEORY AND APPLICATION, 2nd edition by Robert F. DeVellis, published by Sage Publications, Inc., Copyright © 2003 by Sage Publications, Inc. All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher. CHINESE SIMPLIFIED language edition published by CHONGQING UNIVERSITY PRESS, Copyright © 2003 by Chongqing University Press.

量表编制:理论与应用,第2版,作者:罗伯特·F.德维利斯。原书英文版由 Sage 出版公司出版。原书版权属 Sage 出版公司。

本书简体中文版专有出版权由 Sage 出版公司授予重庆大学出版社,未经出版者书面许可,不得以任何形式复制。

版贸渝核字(2003)第10号。

#### 图书在版编目(CIP)数据

量表编制:理论与应用/(美)德维利斯  
(DeVellis, R. F.)著;魏勇刚,席仲恩,龙长权译。—  
2版。—重庆:重庆大学出版社,2010.4  
(万卷方法)

书名原文:Scale Development: Theory and  
Applications

ISBN 978-7-5624-5285-0

I. ①量… II. ①德…②魏…③席…④龙…  
III. ①测量—基本知识 IV. ①P2

中国版本图书馆 CIP 数据核字(2010)第 018019 号

#### 量表编制:理论与应用

(第2版)

罗伯特·F.德维利斯 著  
魏勇刚 席仲恩 龙长权 译  
李红 校

责任编辑:吴文静 版式设计:雷少波  
责任校对:张洪海 责任印制:赵 昆

\*  
重庆大学出版社出版发行

出版人:张鸽盛

社址:重庆市沙坪坝正街174号重庆大学(A区)内

邮编:400030

电话:(023) 65102378 65105781

传真:(023) 65103686 65105565

网址: <http://www.cqup.com.cn>

邮箱: [fxk@cqup.com.cn](mailto:fxk@cqup.com.cn) (营销中心)

全国新华书店经销

重庆升光电力印务有限公司印刷

\*

开本:940×1360 1/32 印张:5.75 字数:163千  
2004年10月第1版 2010年4月第2版 2010年4月第3次印刷  
印数:6 001—10 000

ISBN 978-7-5624-5285-0 定价:25.00元

---

本书如有印刷、装订等质量问题,本社负责调换  
版权所有,请勿擅自翻印和用本书  
制作各类出版物及配套用书,违者必究

# 目 录

---

1 概 貌 .....	1
测量概述 .....	2
社会科学中的测量史源 .....	3
几个早期的例子 .....	3
统计方法的出现和智力测验的作用 .....	4
心理物理学的作用 .....	4
测量的后继发展 .....	5
基本概念的演进 .....	5
智力测验的演进 .....	6
心理统计方法领域的扩展 .....	6
测量在社会科学中的作用 .....	6
理论与测量的关系 .....	6
理论量度与非理论量度 .....	8
量 表 .....	8
量表并非个个造来平等 .....	10
劣质测量的代价 .....	12
小结与预览 .....	12
2 解读潜变量 .....	14
建构及其量度 .....	14
作为项目成绩造因的潜变量 .....	15
路径图 .....	16
图示约定 .....	16
量表编制过程中的路径图 .....	17
测量模型的进一步讨论 .....	19

经典测量模型的假定 .....	19
平行测试 .....	20
其他模型 .....	22
练习 .....	25
3 信度 .....	26
连续值项目与二值项目 .....	26
内部一致性 .....	27
阿尔法系数 .....	27
协方差矩阵 .....	28
多项目量表的协方差矩阵 .....	29
阿尔法系数与协方差矩阵 .....	30
另外一个阿尔法系数公式 .....	33
信度系数与统计力度 .....	36
基于量表分数相关程度的信度 .....	37
信度系数的复本进路 .....	37
信度系数的折半进路 .....	38
信度系数的项目成绩标准化进路 .....	39
信度系数的历时进路 .....	41
概化理论 .....	42
小结 .....	44
练习 .....	45
4 效度 .....	46
内容效度 .....	47
效标关联效度 .....	48
效标关联效度与正确性 .....	49
建构效度 .....	51
建构效度与效标关联效度的区别 .....	51
相关系数多高才算展现了建构效度 .....	52
多特质—多方法矩阵 .....	52
表面效度又是怎么回事儿 .....	54
练习 .....	56
5 量表编制指南 .....	57
第1步:明确你到底要测量什么 .....	57
理论有助于明确所测内容 .....	57

特定性有助于明确所测内容 .....	58
明确量表应包括的内容 .....	60
第2步:建立一个项目池 .....	61
选择反映量表目的的项目 .....	61
项目冗余 .....	62
项目数量 .....	63
开始编写项目 .....	64
优良项目与劣质项目的特性 .....	64
正面表述的项目与负面表述的项目 .....	68
小 结 .....	69
第3步:决定项目形式 .....	69
瑟斯顿治标法 .....	70
古特曼治标法 .....	71
由等权项目构成的量表 .....	73
备择反应选项的最佳个数 .....	73
反应形式的具体类型 .....	77
第4步:请专家评审最初项目池中的项目 .....	85
第5步:考虑把效验性项目包括进去 .....	87
第6步:在样本身上施测项目 .....	88
第7步:评价项目 .....	90
对项目表现情况的初步检查 .....	90
因素分析 .....	95
阿尔法系数 .....	95
第8步:优化量表长度 .....	98
量表长度对信度的影响 .....	98
“差”项目对量表的影响 .....	99
调整量表长度 .....	100
分裂样本 .....	101
练 习 .....	102
6 因素分析 .....	103
因素分析概貌 .....	105
因素分析概念类比 .....	105
因素分析的概念 .....	109
提取因素 .....	110

因素旋转 .....	115
因素解释 .....	125
主成分与共同因素 .....	126
成分与因素的异同 .....	127
确认性因素分析 .....	129
量表编制中因素分析的使用 .....	132
样本大小 .....	135
结 论 .....	136
<b>7 项目反应理论概述 .....</b>	<b>137</b>
项目难度 .....	139
项目区分度 .....	141
猜测度 .....	142
项目特征曲线 .....	143
IRT 的复杂性 .....	146
何时使用 IRT .....	148
等级性项目 .....	148
项目功能差异 .....	149
结 论 .....	150
<b>8 广阔研究背景下的测量 .....</b>	<b>153</b>
编制量表之前 .....	153
寻找现存工具 .....	153
在总体背景下审视建构 .....	154
决定量表的施测模式 .....	156
在其他量表或程序的背景下考虑所开发量表 .....	156
量表施测之后 .....	157
数据分析 .....	158
数据解释 .....	158
概括性 .....	159
最后的思考 .....	159
参考文献 .....	160
附录:现行效度理论的外延和内涵 .....	164

在广阔的社会研究领域,测量至关重要。下面先看几个假设性情境:

1. 健康心理学家普遍面临这样一个难题:他所需要的测量工具显然不存在,而他目前的研究正需要一种尺度,以便把个体看医生时的所要和所期(望)区分开来。这位心理学家发现,先前的研究疏忽了对这两个概念的区分,现存的量表对这两个概念的区分又不完全和他的要求一致。尽管他可以编写几个问题来探测所要和所期之间的差别,但他还是担心,用这些“编造”的项目来测量这两个概念,既不够可信,也不够有效。

2. 流行病学家正在犹豫接下来的工作该如何做:他正在对一次全国健康调查的大量数据进行二次性分析(secondary analysis),想看一看感知到的心理压力的某些方面和健康状况之间有何关系。尽管在原初的调查表中并没有说明哪些项目旨在测量心理压力,但几个原本旨在测量其他变量的项目看上去可能与心理压力的内容有关。因此,有可能把这些项目汇总在一起,从而构成一个可信、有效的心理压力量表。不过,如果这些汇总在一起的项目构成一个糟糕的压力量表,那么研究者就可能据此得出错误的结论。

3. 某营销组试图策划一个关于高价婴儿玩具的商业活动,但失败了。焦点组(focus groups)分析表明,父母的购买决策受此类玩具是否对儿童具有明显的教育意义的影响非常强烈。营销组猜想,对婴儿教育和职业有高期望的父母,受这组新玩具的吸引最大。因此,营销组想根据一个更大的、地理上更分散的父母样本来估计父母的期望值。但其他的焦点组分析表明,营销组可能很难构成一个足够大的消费者样本。



在以上每个情境中,对具体领域感兴趣的研究人员在研究一开始便都遇上了测量问题。尽管他们谁也不是主要对测量本身感兴趣,但是,他们中的每一位在处理主要的研究目标之前,都必须找到一个能量化特定现象的方法。可是对于每一种情况,“现成的”测量工具要么不合适,要么不存在。几位研究者都认识到,如果他们随便采用某种测量方法,就要冒获得的数据不准确这个风险。因而,似乎唯一的选择就是:自己动手,编制自己的测量工具。

许多社会科学研究者都遇到了类似的难题。对于这类难题常见的反应是:依赖现有的、可能不合适的测量工具,或者假定那些新近编制的“看起来”不错的问卷项目可以用来进行测量。常见的借口是:编制可信、有效测量工具太难,不熟悉测量工具的制作方法,或者无法得到如何编制测量工具的实用信息。为了获得量表编制技能,研究者很可能要么去阅读那些非常艰深、主要为测量学专家编写的资料,要么阅读那些过于笼统不可使用的材料。本书旨在为这类研究人员再多增加一个选择的对象。

## 测量概述

测量是一个基础性的科学活动。我们通过观察获得关于人、物、事件和过程的知识。要弄清楚这些观察结果,常常需要对它们量化,即要求我们测量那些我们有科学兴趣的事物。测量与其所服务的广泛的科学问题相互作用,二者间的边界常常察觉不清。测量过程中,当需要探测或凝练一个新的对象时,或者,当怎样量化一个现象可以对认识该现象提供新的启示时,二者间的交互作用就发生了。例如,史密斯、厄普及德维利斯(Smith, Earp & DeVellis, 1995)调查了妇女对受虐的感受。根据理论分析建立的先验模型暗示,受虐的感受有六个不同的侧面。旨在编制一个测量这些感受的量表的实证工作显示,在受虐和未受虐的妇女中,一个简单得多的概念贯穿其中,透彻地解释了研究对象为什么对测量所用的40个项目中的37个作出这样的回答。这一发现暗示,对于研究人员看作是变量复合体的概念,该社区中的妇女感受到的却是一个单一的、广泛的现象。于是,我们在设计关于妇女受虐感受的测量方法的过程中,发现了关于感受结构的新东西。

邓肯(Duncan, 1984)认为,测量植根于社会活动,这些活动以及活动中的测量实际上都先于科学,“所有测量……都是社会测量。物理测量也是服务于社会目的的”(p. 35)。在论及最早的形式化社会测量时邓肯指出,像投票、人口普查以及工作晋升等,“原本都是为了满足人们的日常需要,绝非仅仅为了满足科学好奇心而进行的实验”(p. 106)。他进而指出,同样的例子也“可以从物理学史中拿出:古人在解决社会和实用问题的过程中,成功地实现了对长度(距离)、面积、体积、重量以及时间的测量,物理科学就是建立在这些成就的基础之上的”(p. 106)。

不管最初的动机是什么,科学的每一个领域都发展出了自身的一套测量程序。例如,物理学发展出了特定的方法和设备来探测亚原子粒子。在社会科学和行为科学领域,一个专门研究心理和社会现象的测量问题的分支——心理统计方法(psychometrics)发展了起来<sup>①</sup>。典型的测量程序是问卷调查,所调查的变量是一个更广泛的理论框架的组成部分。

## 社会科学中的测量史源

### 几个早期的例子

常识和历史记载都支持邓肯的观点:社会需要使得测量在科学之前就得以出现。毫无疑问,自史前以来,有些测量形式一直是人类技能中的一个部分。最早的人类必须对物体、财产以及对手作出评量,评量的基础通常是像大小这样的特性。邓肯(Duncan, 1984)引用圣经上的记载说明了早期人类对测量的关注程度(例如,缺两上帝憎,足斤上帝悦。《旧约·箴言》,第11节第1句),并指出,亚里士多德的著述中提到了专司度量衡的官员。阿纳斯塔西(Anastasi, 1968)指出,古希腊时所使用的苏格拉底式探

<sup>①</sup> 在我国,有些学者把 psychometrics 翻译成“心理计量学”,这是错误的,因为,计量学的英语是 metrology。目前,英语词汇中还没有和汉语“心理计量学”相对应的术语,就算通过类比创造一个英语新词的话,也应该是 psychometrology,而不是 psychometrics。在社会(包括心理)测量领域至今还未谈计量,因为一谈到计量,必然涉及量的单位及量纲问题,没有单位,量的值是无法表示的。至今,社会测量领域还没有设计出自己的基本单位(如米、千克、秒),更不用说导出单位(如牛、伏特)了。——译者按

究理解的方法,在某种程度上可以看作是知识测验。迪布瓦博士(P. H. DuBois)在他 1964 年的论文中记述到,早在公元前 2200 年,中国就有了“公务员”测验(Barnette, 1976)。赖特(Wright, 1999)也举了其他一些关于古代准确测量的重要例子,包括“七度(weight of seven)”这种七世纪的穆斯林征税原则。他还指出,有人把法国革命爆发的部分原因和农民受够了当时不公正的度量衡制度联系起来。

### 统计方法的出现和智力测验的作用

衣纳利(Nunnally, 1978)指出,尽管系统的观察方法仍在继续,但由于没有可用的统计方法,关于人类能力测量的科学直到 19 世纪下半叶才得以出现。邓肯(Duncan, 1984)也指出,在几何学以外的大部分数学领域,学科基础的形式化滞后应用的时间要以千年计,即数学其他分支基础的形式化在 19 世纪才发生。达尔文在进化论方面的工作以及他对物种间系统变异的观察和测量,使得适当统计方法的发展在 19 世纪终于启动。他的堂弟高尔顿男爵(Sir Francis Galton)把对差异的系统观察扩展到了人类——高尔顿主要关注的是解剖特质和智力特质的遗传问题。被誉为“统计学奠基人”(例如,Allen & Yen, 1979, p. 3)的卡尔·皮尔逊(Karl Pearson)是高尔顿的一个晚辈同事,他发展出了能系统考量变量间关系的数学方法,其中包括以他的名字命名的积矩相关系数。于是,科学家便能够量化可测特性间相互关系的程度。查尔斯·斯皮尔曼(Charles Spearman)继承前辈的研究传统,为 20 世纪初因素分析的发展和普及奠定了基础。值得一提的是,许多形式化测验的早期贡献者(其中包括 20 世纪初在法国发展出智力测验的阿尔弗雷德·比纳(Alfred Binet))都对智力测验很感兴趣。因此,许多早期心理统计方法方面的成果都应用到了“智力测验”中。

### 心理物理学的作用

现代测量学的另一个历史根源是心理物理学。把物理学的测量程序应用于感觉研究的尝试,引起了关于测量本质的长期争论。纳仁思及卢斯(Narens & Luce, 1986)总结这段争论时指出:19 世纪晚期赫尔姆霍茨(Hermann von Helmholtz)发现,像长度和质量这样的物理属性具有和正实数一样的内部数学结构。例如,时间或长

度可以和普通数一样进行排序和相加。20世纪早期,争论还在继续。英国科学促进会委员会(The Commission of the British Association for Advancement of Science)认为,心理变量的基本测量还没有可能,因为在对感官知觉结果排序或相加时要遇到其固有的问题。斯蒂文斯(S. Smith Stevens)辩解道,适用于长度或质量这类量的严格可加性并不是必要的,并指出,个体可以对声音强度作出较为一致的比率性判断(例如,人可以判断一个声音的大小是另一个声音大小的两倍或一半),而这种比率属性使得来自这些测量中的数据可以进行数学运算。斯蒂文斯还把测量结果分为定类(nominal)、定序(ordinal)、定距(interval)和定比(ratio)这几个水平。他还指出,响度判断结果符合定比量的要求(Duncan, 1984)。大约就在斯蒂文斯为心理物理性质量化的合法性辩护的同时,瑟斯顿(Louis L. Thurstone)却在发展因素分析的数学基础。瑟斯顿的兴趣横跨心理物理学和智力学科两个领域。根据邓肯(Duncan, 1984)的说法,斯蒂文斯认为,瑟斯顿应用心理物理方法对社会刺激进行了量化。可见,瑟斯顿的工作代表了两个不同历史渊源的心理学分支的汇合。

## 测量的后继发展

### 基本概念的演进

尽管斯蒂文斯在社会测量史上影响很大,但他提出的一系列测量学概念绝不是定论。他把测量定义为“根据规则对物体或事件赋数”(Duncan, 1984)<sup>①</sup>。邓肯(Duncan, 1984: 126)向这一定义提出了挑战,他认为,斯蒂文斯对测量的定义犹如“把弹钢琴定义为根据一定模式敲击乐器的键盘‘一样不够完备’,测量不单单是赋数等,它还是按照物体或事件的质(quality)……或属性(property)的不同度进行赋数”。纳仁思和卢斯(Narens & Luce, 1986)也指出了斯蒂文斯关于测量概念的局限,并提出了许多改进

<sup>①</sup> 斯蒂文斯对测量的定义几乎成了社会测量学中的公认定义,但需要指出的是,这个定义和计量学中的定义差别很大。国际计量部门对测量的规范定义是:测量是一个通过实验对量的值进行确定的过程。根据这个定义,定名水平的测量就不能算作测量了,因此测量结果也就只有三个水平了。——译者按

意见。尽管如此,他们都强调了斯蒂文斯的基本观点:在英国科学促进会委员会认定的测量模型之外,还存在其他的模型,而这些其他模型,使得测量方法不仅运用到物理科学,也运用到了非物理科学。这些对量数(measures)基本属性的讨论,在本质上确立了社会科学领域中诸测量程序的科学合法性。

### 智力测验的演进

尽管“智力测验”(现在通常叫“能力测验”)一直是心理统计方法领域中颇具活力的传统部门,但它并不是本书的主要讨论对象。如果测量的目标是能力之外的其他特征,那么,心理统计方法领域中的许多新进展(包括项目反应理论)就不那么常用,也不那么容易应用。几十年过去了,新方法在能力测验之外的可用性已经变得更加显而易见,因此,我将在本书的后半部分,用专门的一章简单介绍项目反应理论。本书中,我还是重点讨论“经典”理论,因为,这一理论在能力之外的心理和社会现象的测量中基本占据着统治地位。

### 心理统计方法领域的扩展

邓肯(Duncan, 1984)指出,心理统计方法在社会科学中的影响,已经超越了最初的感觉测量和智力测量领域。现在,它已经成为一种独立的方法论范式。邓肯用三个例子来支持他的观点:①心理统计方法中对信度和效度的定义被广泛使用;②因素分析在社会科学研究中备受欢迎;③采用心理统计方法编制的量表所测量的变量的范围已经远远超过了它最初所关注的那几种(p. 203)。在以后的章节中,我们将一一讨论心理统计方法中的概念和方法在各种心理和社会现象测量中的应用问题。

## 测量在社会科学中的作用

### 理论与测量的关系

在社会科学中,我们试图测量的现象常常由理论衍生。因此,在构念测量问题的过程中,理论扮演了一个很重要的角色。当然,科学中的很多领域都测量理论衍生出的东西——在亚原子微粒被

测量确认之前,它仅仅是一个理论建构。不过,心理学与其他社会科学中的理论和物理科学中的理论很不相同。在社会科学中,科学家倾向于依赖大量的只关于狭小范围内现象的理论模型;在物理科学中,科学家使用的理论较少,但理论适用的范围却很广。例如,费斯廷格(Festinger, 1954)的社会比较理论就只关注人类经验中一个相当狭小的范围,即人们通过与他人比较来评价自己的观点或能力的方式。相比之下,物理学家不断努力,试图建立一个大一统的场理论,这个理论可以在一个框架中囊括自然界所有的基本力。此外,社会科学不像物理科学那么成熟,因而,理论演进很快。待测现象无形无色、难以捉摸,而且衍生于多个变化不定的理论,这种测量向社会科学研究者提出了严峻的挑战。在这种学科背景下,高度关注测量程序,并完全认识它们的优缺点,就尤为重要。

研究者对所关注的现象、假设建构中的抽象关系以及可用的定量工具了解得越多,就越具备去编制可信、有效、可用的量表的条件。其中,对具体现象的细节的了解,可能最为重要。例如,社会比较理论有许多侧面,每个侧面蕴涵着不同的测量策略。甲研究问题可能要求把社会比较操作化地定义为:关于较高或较低他人信息的相对参照标准;乙研究问题则可能要求把社会比较操作化地定义为:多维度自我评定所参照的“典型人”。不同的量度从不同的侧面刻画同一现象(如“社会比较”),可能导致测量结果的不一致(DeVellis, et al., 1991)。尽管给变量冠以相同的名称,但所评估的却是几个本质上不同的变量。因此,编制一个最适合于具体研究问题的量表,需要理解理论的微妙之处。

不同的变量要求不同的评估策略。例如,从储罐中取出硬币,其值可以直接观察到。然而,社会和行为科学家感兴趣的绝大多数变量,其值是不能直接观察到的。信念、动机状态、期望、需求、情感以及社会角色觉察等,不过是众多其值不能直接观察的变量中的几个例子。有些不能直接观察的变量的值,也不能通过问卷调查确定,需要通过其他研究程序。例如,认知研究人员虽然不能直接观察个体怎样在其自我图式中构建性别信息,但他们可以通过回忆程序(recall procedures)推断出个体在思想上是怎样构建自我和性别的。不过在许多情况下,只能用纸笔测试评估社会科学变量,其他方法既不可能,也不实际。当我们关注的测量对象是理

论建构时,情形虽不总是这样,但常常是这样。因此,一个对测量雌雄同体感兴趣的研究者可能会发现,用一套精心编制的问卷测量,要比用其他方法容易得多。

### 理论量度与非理论量度

至此,我们应该承认,本书的重点虽然在理论建构的测量方面,但并非所有的纸笔测试都需要理论建构,如性别和年龄就不需要理论建构。这两个变量可以成为理论模型的组成部分,也可用来描述被试,具体使用要取决于具体的研究问题。医院通过让病人以纸笔形式回答一系列问题来了解其饮食偏好的做法,就没有理论基础。有些情况下,研究可能以非理论始,但却以形成理论终。例如,市场研究者可能让父母列一张他们给孩子买的玩具类型的清单,但随后,该研究者可能会探究这些清单中的关系。根据所发现的消费模式,研究者会发展出一个消费行为模型。其他的相对非理论测量的例子就是各种各样的民意调查。例如,询问人们使用哪种品牌的香皂,或者,准备投哪个候选人的票,这类问题很少用来探索深层的理论建构。因为,研究者所关注的只是反应本身,而不是反应有可能折射出的其他特性。

有时,很难区分理论测量情境和非理论测量情境。例如,欲通过探寻投票者对总统候选人的偏爱来预测选举结果,就会要求调查对象报告各自行为的意向。调查者可能要问调查对象在投票决策过程中怎样做到不从个人利益出发,但目的只是预测最后的投票结果。但如果问这同一问题的背景是探究对特定问题的态度如何影响投票人的投票偏向时,那么研究深层就很可能有一个精当的理论在支撑。在这种情况下,有关投票的信息并不是为了揭示投票人将会投谁的票,而是为了探明个体的有关特征。在上述两个例子中,量度与理论有关与否,完全取决于研究人员的意图,而不是所用的程序。对如何编制非用于调查理论建构问卷感兴趣的读者,可参考下面的著作:肯维斯及普莱斯(Converse & Presser, 1986)、查迦及布莱尔(Czaja & Blair, 1996)、迪尔曼(Dillman, 2000)、芬克(Fink, 1995)、福勒(Fowler, 1993, 1995)以及韦斯伯格、克劳里克及博文(Weisberg, Krosnick & Bowen, 1996)。

### 量 表

所谓量表,就是这样一种测量工具,它由多个项目构成,形成

一个复合分数,旨在揭示不易用直接方法测量的理论变量的水平。有时,根据理论知识我们推断某些现象是存在的,而且我们想测量这些现象,但无法直接进行,这时,我们就需要编制量表<sup>①</sup>。例如,我们可能会用抑郁或焦虑来解释所观察到的行为。绝大多数的理论家都认为,抑郁或焦虑并不等同于我们所观察到的行为,但潜伏在这些行为的背后。我们的理论暗示,这些现象存在并影响着行为,但却无形。有时,通过它们引发的行为来推测其存在也可能是合适的。然而,在其他情况下,我们可能没有办法得到关于行为的信息(譬如只能用邮寄方式调查时),也不肯定如何解释手头的行为样本(譬如在遭遇某一事件时,绝大多数人都反应强烈,而某个人却不动声色),或者,我们可能不愿意假定行为与其背后的建构同构(譬如我们怀疑哭是因为高兴而不是因为悲伤)。在我们不能把行为当作现象的表示来解释的情况下,采用一个精心编制、严格效验过的量表进行测量也许会非常有用。

即使对于从理论中衍生出的变量,也存在一个隐连续统(implicit continuum),一端是相对具体的可观测现象,另一端是相对抽象的不可观测现象。并不是所有的现象都要用由多个项目组成的量表来测量,年龄和性别就是两个例子:它们确实和许多理论有关,但却不需要用多项目量表来精确测量。多数情况下,这些变量都是和具体的、相对清楚的性质(如形态)或事件(出生日期)关联的。除特殊情况(如神经受损)外,调查对象是不会不知道自己的年龄和性别的。这类问题,他们可以高度精确地回答。相比之下,种族倒是一个较为复杂和抽象的变量。典型情况下,它是体貌、文化和历史因素的综合,因此比性别或年龄要更无形,更像一个社会建构。定义一个人的种族固然复杂、费时,但是,绝大多数人都能自我定义,稍加思索便能报告自己的种族。因此,在大多数情况下,一个变量就足以确定种族。不过,许多其他的理论变量都需要被试对一些不太容易得到的信息进行重建、解释、判断、比较或者评价。例如,问已婚的人如果选择不同的配偶他们的生活会怎样的不同。这样的问题回答起来就得费一番脑筋,而且一个项目可能涵盖不了所关注现象的复杂内容。对于这种情况,量表

---

<sup>①</sup> 这里的论述不够严谨。为了严谨,还须在“这时”之前再加上条件“也没有其他现成的量表可用”。——译者按



也许是最合适的测量工具。多项目可能会抓住此类变量的本质,达到单项目所不能达到的精确程度。就是这种不能直接观测又需要被试思考的变量,最适合用量表评估。

也应该把量表和其他获取复合分数的多项目量度加以对照。不同多项目量度之间的差别,无论是对于理论,还是对于实践,都非常重要。关于这些,在后面的章节里我们会加以揭示。在本书中,“量表”包含了伯伦(Bollen, 1989: 64-65; 也可参见 Loehlin, 1998: 200-202)所谓的“效果指示(effect indicators)”——由潜伏建构(下章中称作“潜变量”)引起的项目值。抑郁感的量度通常和量表的性征一致,因为对每个项目的反应都有一个共同的因,即作答者的情感状态。因此,一个人对诸如“我感到很悲伤”和“我的生活很乏味”这类问题的答法,很可能主要取决于他当时的感觉。我将用指标(index)来描述作为“原因指示(cause indicators)”的项目集,即那些确定建构水平的项目。例如,总统候选人的魅力,就可能是一个符合指标性征的量度。项目可能会估量候选人的住址、家庭大小、体貌吸引力、激励竞选工作人员的能力以及可能的经济资源。尽管这些性征可能没有共同的因,但它们可能有共同的果——增加总统竞选成功的可能性。对有可能汇集在一起产生复合分数的项目集团,有一个更一般性的名称——应急变量(emergent variable, 例如, Cohen, Teresi, Marchi & Velez, 1990)<sup>①</sup>,这类变量包括所有可以汇集在一起的共享某个性征、可以用一个类别名称概括的项目单位。把性征组成团未必隐含某种因果联系。例如,把以少于5个字母的单词开头的英语句子归在一起就很容易,尽管它们之间既没有共同的因,也没有共同的果。应急变量的“突然出现”仅仅是因为在研究过程中某人或某事(如数据分析程序)发现了项目间的某种相似性。

### 量表并非个个造来平等

遗憾的是,并非所有的量表都是精心编制的。许多量表,用“拼凑”可能比用“编制”更为合适。研究者往往把一些自以为能组

---

<sup>①</sup> 这是一个事先没有设定的变量,只是把量表中的几个项目组合起来另外产生一个分数时才出现。由于临时决定把几个项目组合起来产生另外一个分数是为了应研究中的紧急需要,因此译作“应急变量”。——译者按