

万卷方法

研究方法·基础应用

量表编制：理论与应用

Scale Development: Theory and Applications 3Ed

第3版

罗伯特·F.德威利斯 (Robert F. DeVellis) 著

席仲恩 杜珏 译



重庆大学出版社

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



万卷方法

研究方法·基础应用

量表编制：理论与应用

Scale Development: Theory and Applications 3Ed

第 3 版

罗伯特·F·德威利斯 (Robert F. DeVellis) 著

席仲恩 杜 珏 译

重庆大学出版社

Authorized translation from the English language edition, entitled SCALE DEVELOPMENT: THEORY AND APPLICATION, 3rd edition by Robert F. DeVellis, published by Sage Publications, Inc., Copyright © 2012 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 © 2013 by Chongqing University Press.

量表编制:理论与应用,第3版,作者:罗伯特·F.德威利斯。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(2013)第284号。

图书在版编目(CIP)数据

量表编制:理论与应用:原书第3版/(美)罗伯特·F.德威利斯(Robert F. DeVellis)著;席仲恩,杜珏译.—重庆:重庆大学出版社,2016.10

(万卷方法)

书名原文:Scale Development: Theory and Applications

ISBN 978-7-5689-0172-7

I. ①量… II. ①德…②席…③杜… III. ①社会测量—基本知识 IV. ①C91-03

中国版本图书馆 CIP 数据核字(2016)第236348号

量表编制:理论与应用

(第3版)

罗伯特·F.德威利斯 著

席仲恩 杜珏 译

责任编辑:林佳木 版式设计:林佳木

责任校对:郭小梅 责任印制:赵晟

*

重庆大学出版社出版发行

出版人:易树平

社址:重庆市沙坪坝区大学城西路21号

邮编:401331

电话:(023) 88617190 88617185(中小学)

传真:(023) 88617186 88617166

网址: <http://www.cqup.com.cn>

邮箱: fxk@cqup.com.cn (营销中心)

全国新华书店经销

自贡兴华印务有限公司印刷

*

开本:940mm×1360mm 1/32 印张:7.25 字数:209千

2016年10月第3版 2016年10月第6次印刷

印数:13 001—17 000

ISBN 978-7-5689-0172-7 定价:34.00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

作译者简介

罗伯特·F.德威利斯 美国北卡罗来纳大学教堂山分校公共卫生学院卫生行为和卫生教育系教授,他拥有30多年心理和社会测量方面的实践经验。他是美国国立卫生研究院《患者结局报告测量信息系统》路线图项目的活跃成员。他曾担任美国心理协会第38分会心理卫生分会的理事,曾获得2005年度美国风湿病专业委员会颁发的突出贡献学者奖。此外,他还担任过《关节炎护理与研究》学刊的副主编以及二十几本学术刊物的特约主编、特约副主编、审稿人等。目前,德威利斯博士的研究兴趣有:配偶及其他亲密关系对病情的不良影响,与健康 and 病情有关的社会变量及行为变量的测量。

席仲恩 博士,重庆邮电大学教授,主要从事心理测验和教育测量研究、翻译理论与翻译教学研究、国际学术论文写作规范与教学研究等工作,有丰富的学术翻译、学术写作及论文润色和修改经验。关于心理测验和教育测量,席博士近期的主要研究兴趣是其中的计量学原理。

杜珏 重庆邮电大学教师,主要从事语言测试和外语教学研究。

前 言

自从开始考虑写一本关于量表编制的书,我就为自己设定了这样一个目标:让非常复杂的信息,以明了的方式呈现出来,以帮助广大读者理解测量工具的创造、使用及评价背后的逻辑,使读者对于量表的作用机理有一种感觉。为了揭去蒙在测量之上的那层神秘面纱,只要有可能,我就联系大家都熟悉的经历,强调对测量概念的理解,弱化对测量的严格数学意义上的理解。我还尝试着把我自己多年在测量工作中的经验体会提炼出来,和大家一块儿分享,其中有些是我的老师和同事传授给我的,有些是自己在长期的测量实践中,从一开始对概念的错误理解中摸索出来的。这本小书的前两版出版之后,受到多个领域工作者的广泛欢迎,我颇感欣慰。我之前的学生们谈到,他们曾在根本没有想到的地方看到了《量表编制》一书,包括美国国家宇航局科学家的桌面上,以及很多远在其他大洲的国家。这本小书之所以如此受欢迎,我想可能是我用的方法浅显易懂,且不需要多少预备知识。这一点,对于那些之前没有多少社会测量经验的读者尤为重要。因此,这一点依然是《量表编制》第3版所坚持的。为此,在第3版中,我在更新多个议题时,保留了那些在前一版中被证明行之有效的地方,并对其他一些议题做了扩充,增加了一些新的且更加明确的阐明概念的例子。

第1章是关于测量的历史渊源。本版对该章做了扩充,新增加了几个历史视角。特别是,增加了关于早期科学家(包括艾萨克·牛顿)如何看待观察结果不一致问题的内容。这个课题就是我们今天称作测量误差的问题。在本章和其他多处,我都明确地指出劳德和诺维克对社会测量学的具体贡献。因为最近,他们合写的那本出版于1968年的著作《心理测验分数的统计学理论》,经过几十年的沉睡和短缺,终于变得随处可得。

第2章是关于潜变量的介绍。在这一章中,我也增加了新的内容,包括如何考虑环境的方方面面以及受测人对于这些方面做出的反应之间的差别问题。环境和受测人对于环境所做出的反应是两个不同的变量,也是不时被混淆的概念。我还进一步澄清了真分数等价和真分数本质上等价这两个有微妙差别的模型。

第3章是关于信度的。我扩展了关于信度问题的背景,即把信度视作观察到的现象所反映出的某个过程的一致性程度。在此宽阔的背景之下,组内相关系数(ICC)就被看作是表达信号与信号加噪音之比的典型方式,从而在概念上把组内相关系数和其他表达信度的具体方式联系起来,包括克伦巴赫 α 的计算与科恩 κ 系数的计算。在讨论评分者间信度时要用到科恩 κ 系数。尽管在最狭窄的意义上讲,科恩的 κ 系数才与量表信度搭上关系,但是,学生经常不能把关于观察者的信度与关于项目的信度联系起来。我坚信,把这两种概念的信度联系起来,能加深我们在更广泛意义上对信度概念的理解。本章我还讨论了最近学界关于克伦巴赫 α 系数的批评。

第4章是关于效度的。在这一章,我增加了一些例子,以便读者更好地理解效度是如何确立的。我还增加了关于在概念上区分标准效度和构念效度以及内容效度和表面效度的内容。此外,我还增加了一些关于相关系数衰减及其对于效度评估影响的讨论。

第5章依然是关于量表编制步骤的。这一章,我扩充了关于项目

冗余度的讨论,更清楚地区分了对于给定量表而言的“好”冗余与“坏”冗余。

在第6章中,我扩充了因素分析的范围,对其中的多个例子做了广泛修订。这些例子都配了新的图解。我相信,这样会进一步阐明因素分析中的有关概念,例如因素旋转。我还增加了一部分讨论,讨论如何用平行分析确定需要保留的因素个数。

我对第7章做了大量的扩展,以更好地反映过去几年业界对于项目反应理论(IRT)的日益重视。尽管这一章是对项目反应理论的简介,而不是对它的详细分析,我还是增加了关于如何从众多的IRT模型中进行选择的内容,以及关于解释IRT分析图像结果的图示说明。此外,我还扩展了经典测试理论相对于项目反应理论的作用的讨论,总结了对这两种理论进行直接比较的研究的结果。

在第8章中,我对现用量表的有关信息进行了更新,并增加了一些关于《患者结局报告测量信息系统》路线图的信息。这项计划由美国国立卫生研究院发起,其目标是,建立一套更为系统的测定有关健康变量的方法。在国立卫生研究院第一轮的工作中,我有幸参与了研究工作,并主持了社会卫生组的工作。该项目中所研发出的一系列工具,是健康研究人员的重要资源。因此,我向读者推荐了一些有关量表详细研究的文献。我还圈点出了部分出自美国国立卫生研究院的关于测量研究方法的论文,包括一些定性项目评论的方法过程。这些,对于读者可能会有用处。

总之,《量表编制》第3版更新且扩充了前版的内容,剔除了一小部分被实践证明不如我预想的那么对读者有帮助的材料,并增加了对新近出现课题的讨论。我相信,这些变更会使本书更加清晰易读,更加实用。我希望,你们能同意我的看法。

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	6
心理物理学的作用	6
测量的后继发展	7
基本概念的演进	7
智力测验的演进	8
心理统计方法领域的扩展	8
测量在社会科学中的作用	9
理论与测量的关系	9
理论量度与非理论量度	10
量 表	11
量表并非个个造来平等	13
劣质测量的代价	14
小结与预览	15
2 解读潜变量	17
构念及其量度	17
作为项目得分假设原因的潜变量	19
路径图	20
图示约定	20
量表编制过程中的路径图	21

测量模型的进一步讨论	23
经典测量模型的假定	23
平行测试	24
其他模型	27
3 信 度	31
基于方差分析的方法	32
连续值项目与二值项目	33
内部一致性	34
阿尔法系数	35
协方差矩阵	36
多项目量表的协方差矩阵	37
阿尔法系数与协方差矩阵	37
另外一个阿尔法系数公式	41
对阿尔法系数的批评	44
基于量表分数相关程度的信度	45
信度系数的复本进路	45
信度系数的折半进路	46
信度系数的项目成绩标准化进路	48
评分者间一致性	49
信度系数的历时进路	51
信度系数与统计力度	54
概化理论	55
4 效 度	60
内容效度	61
效标关联效度	63
效标关联效度与正确性	64
构念效度	66
构念效度与效标关联效度的区别	67
相关系数的衰减	69
相关系数多高才算展现了构念效度	70
多特质—多方法矩阵	70
表面效度又是怎么回事儿	72

5 量表编制指南	75
第1步:明确你到底要测量什么	75
理论有助于明确所测内容	75
特定性有助于明确所测内容	76
明确量表应包括的内容	78
第2步:建立一个项目池	79
选择反映量表目的的项目	79
项目冗余	81
项目数量	83
开始编写项目	84
优良项目与劣质项目的特性	85
正面表述的项目与负面表述的项目	88
小 结	90
第3步:决定项目形式	90
瑟斯顿治标法	90
古特曼治标法	92
由等权项目构成的量表	94
备择反应选项的最佳个数	94
反应形式的具体类型	98
第4步:请专家评审最初项目池中的项目	107
第5步:考虑把效验性项目包括进去	109
第6步:在样本身上施测项目	110
第7步:评价项目	112
对项目表现情况的初步检查	112
因素分析	117
阿尔法系数	118
第8步:优化量表长度	121
量表长度对信度的影响	121
“差”项目对量表的影响	121
调整量表长度	123
分裂样本	124
6 因素分析	126
因素分析概貌	128
因素分析概念类比	129

因素分析的概念	134
提取因素	134
因素旋转	143
因素解释	156
主成分与共同因素	158
成分与因素的异同	158
确认性因素分析	161
量表编制中因素分析的使用	164
样本大小	167
7 项目反应理论概述	169
项目难度	173
项目区分度	175
猜测度	176
项目特征曲线	178
IRT 应用于多反应项目	181
IRT 的复杂性	187
8 广阔研究背景下的测量	194
编制量表之前	194
寻找现存工具	194
在受测总体背景下审视构念	196
决定量表的施测模式	198
在其他量表或程序的背景下考虑所开发量表	198
量表施测之后	200
数据分析	200
数据解释	200
概括性	201
最后的思考	201
参考文献	203
附录:现行效度理论的外延和内涵	211

在广阔的社会研究领域,测量至关重要。下面先看几个假设的情境:

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的是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。焦点小组(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。但其他的焦点小组分析表明,营销组可能很难构造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指导。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,二者间的边界并非常常清晰可察。测量过程中,当需要探测或凝练一个新的对象时,或者,当一种量化某个现象的方式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。而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,是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。在行为科学和社会科学领域,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^①。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

^① 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 psychometry,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

前就得以出现。毫无疑问,自史前以来,有些测量形式一直是我们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。

测量结果中可能包含误差且可以通过一定的做法来缩小误差的思想是一个更加晚近的洞见。布赫瓦尔德(Buchwald, 2006)在其关于测量结果偏差及其对知识的影响的评论中指出,在1660年代后几年和1670年代头几年,还是二十几岁的艾萨克·牛顿就显然首次使用了平均多次观测结果的方法。牛顿的目的是,在自己关于天文现象的观察值存在差异时,用各次观测结果的平均值代表观测结果,以得到一个更为准确的测量结果。有趣的是,牛顿在其原初报告中并没有记述自己使用平均值这一做法,而且一直隐瞒了几十年。这一隐瞒,与其说是学术诚信问题,不如说是当年人们对于误差及其在测量中作用理解的局限问题。在评论另一位近代天文学家隐瞒自己观测结果的偏差时,艾尔德(Alder, 2002)指出,即使到了1700年代后期,隐瞒观测结果偏差这一做法“不仅普遍,而且被看作是智者的特权;而误差却被看作是道德的欠失”(P. 301)。同样,布赫瓦尔德(Buchwald, 2006)也指出:

[17世纪和18世纪科学家的]流行做法是,观测结果的差异不被看作是测量过程本身的不可避免的伴随品,而是工作的失败或技艺的不足;测量中的误差与测量中的任何错误行为,其潜在

的威胁并无多大的差别:它可能引起一些道德上的恶果,因此,必须妥当处理。(P. 566)

在1600年代晚期和1700年代早期,需要对自然现象进行系统观测的科学家当中不仅有天文学家,还有其他方面的。1660年代,在根据英格兰罕不什尔郡的洗礼及葬礼记录编制当地的出生率和死亡率时,约翰·格朗特(John Graunt)就使用了平均值(这种方法现在已不常用)来总结自己的发现。根据布赫瓦尔德(Buchwald, 2006)的记述,格朗特使用平均值的动机,是要捕获住那瞬息可变的“真”值。当时他的想法是,出生率与死亡率之比遵循某种自然法则,但是,那些发生在任何一个年度的不可预测事件可能会掩盖那一基本事实。这种关于观察是通往自然真理的有缺陷窗口的观点间接表明,当时人们对于测量的看法已经变得越来越成熟:除了观察者的局限之外,其他因素也可能败坏经验信息;因此,对观测值的适当调整处理,可能会更准确地反映出所感兴趣自然现象的真实情况。

尽管有这些早期的洞见卓识,但只是在牛顿首次使用平均值一个世纪之后,科学家才开始广泛地认识到,凡测量皆有误差,平均值能使该误差降到最小(Buchwald, 2006)。根据物理学家及科普作家蒙洛迪诺(Mlodinow, 2008)记述,在18世纪后期和19世纪早期,天文学和物理学的发展,迫使当时的科学家更加系统地对待随机误差问题,因此也导致了数理统计学科的诞生。到了1777年时,丹尼尔·伯努利(更著名的雅克布·伯努利的侄子)对天文观测结果的分布和射箭飞行轨迹的分布进行了对比,发现两者都是围绕着某个中心分布,距中心越近分布的密度越大,距中心越远分布的结果越稀少。尽管关于该观察结果的理论处理在某些方面是错误的,但它标志着对测量误差进行形式分析的开始(Mlodinow, 2008)。布赫瓦尔德(Buchwald, 2006)指出,18世纪的这种对于测量误差的解读,存在一个基本缺陷。那就是,未能区分随机误差和系统误差。直到19世纪初,人们才更深刻地理解了随机性问题。随着人们对随机性理解的加深,测量也有了长足的进展。随着测量的发展,科学也向前迈进了一步。

统计方法的出现和智力测验的作用

农纳利(Nunnally, 1978)支持这样的观点:对于随机性、概率以及统计学的更加深入的了解,是测量学得以繁荣的必要条件。农纳利指出,尽管系统的观察方法一直在进行,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。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。正如我们前面看到的那样,在天文学和物理科学之中,测量问题普遍存在,因此也受到艾萨克·牛顿爵士的关注(Buchwald, 2006)。心理物理学存在于物理学和心理学的结合部,因此它关注的问题是刺激的物理属性以及刺激是如何被人类感知的。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。纳仁思和卢斯(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