王孝玲 编著

JIAOYU CELIANG

教育测量

修订版

华东师范大学出版社

教育测量

（修订版）

华东师范大学出版社

Jiaoyu Celiang

王孝玲 编著

图书在版编目(CIP)数据

教育测量/王孝玲编著. —修订本. ——上海:华东师范大学出版社,2004.3

ISBN 7-5617-3727-0

Ⅰ.教... Ⅱ.王... Ⅲ.教育测量 Ⅳ.G449

中国版本图书馆 CIP 数据核字(2004)第 024571 号

教育测量(修订版)

编　　著　王孝玲
责任编辑　张　捷
责任校对　邱红穗
封面设计　黄惠敏
版式设计　蒋　克

出版发行　华东师范大学出版社
　　市场部电话　021-62865537
　　门市(邮购)电话　021-62869887
　　门市地址　华东师大校内先锋路口
业务电话　上海地区　021-62232873
　　华东　中南地区　021-62458734
　　华北　东北地区　021-62571961
　　西南　西北地区　021-62232893
业务传真　021-62860410　62602316
http://www.ecnupress.com.cn
社　　址　上海市中山北路 3663 号
　　邮编　200062

印 刷 者　上海市印刷三厂
开　　本　890×1240　32 开
印　　张　13
字　　数　368 千字
版　　次　2004 年 7 月第一版
印　　次　2004 年 7 月第一次
印　　数　5100
书　　号　ISBN 7-5617-3727-0/G·2036
定　　价　19.00 元

出 版 人　朱杰人

前　言

本书第一版面世至今已有十多年，在此期间国外现代教育测量的新兴理论不断传入，国内教育测量学者们的研究成果也不断涌现，所以有必要加以补充和修改。

修订版除了对第一版原有内容加以修改之外，还新增加了三章，即第十一章“标准参照测验及其鉴定”；第十二章“项目反应理论及其在标准参照测验中的应用”；第十三章“测验举例”，并为各章练习题（计算性的题目）提供了答案。

修订版第一至第十章主要论述的是，在经典测验理论指导下常模参照测验的编制原理及方法。第十一章论述的是标准参照测验。第十二章论述的是现代教育测验理论——项目反应理论，及在其指导下标准参照测验的编制原理及方法。

常模参照测验是一种选拔性、竞赛性的测验。这种测验过去一直以有悠久历史的经典测验理论为依据。标准参照测验是一种达标性的测验，如各级各类学校平时教学中区分学生掌握与否的测验；判断初、高中学生是否达到毕业水准的测验；各行各业招聘录用测验；任职资格证书测验；从业执照测验等。其应用范围比常模参照测验广泛。但是由于经典测验理论基本不适合于标准参照测验，所以标准参照测验理论及操作技术的产生和发展都远比常模参照测验迟得多。项目反应理论是现代教育与心理测验的新兴理论。它的理论与操作都大大优于经典测验理论，而且既适合于常模参照测验，又适合于标准参照测验。从而，标准参照测验寻到了适合它的理论及编制

的方法途径。作者为了反映教育测量学的这一发展现实，特补充第十一和十二章有关内容。根据项目反应理论的优越性和经典测验理论的局限性，虽然可以推测出项目反应理论有代替经典测验理论的可能，但是由于项目反应理论目前应用最为广泛的还只是二值记分的反应模型，再加上它的计算手续繁复，所以曾经起了很大作用的经典测验理论现在仍然发挥着它的作用。这也是本书仍以介绍经典测验理论为主的原因。

本书第十一章第四节由安徽师范大学教育系赵必华老师撰写，其他各章均由华东师范大学王孝玲撰写。

由于作者的水平有限，本书难免会出现错误或不妥，敬请同行、读者批评指正。

王孝玲

2004 年 1 月

目 录

第一章　教育测量的基本原理

第一节　测量的概念

史蒂文斯(S. S. Stevens)于1951年曾给测量下了这样的定义："从广义而言，测量是根据法则给事物分派数字。"这一定义概括了物理测量、社会测量和心理测量的共性。

例如，测量学生的体重时，学生只能身穿极少量的衣服，赤脚自然站立在体重计上，这时体重计上所指示的数字，就是该生的体重。在这里，学生的"体重"是我们所要测量的属性，而"身穿极少量的衣服，赤脚自然站立在体重计上"，是测量体重所依据的规则。体重计上所指示的"数字"，就是我们用来描述学生体重的数。

又如，教师要用1、2、3、4、5五个等级对学生的道德品质进行评定，道德品质最好者评为5，最差者评为1，其他依道德品质的不同程度评为4、3、2。在这里，学生的"道德品质"是教师所要测量的属性，1、2、3、4、5五个等级是用来描述学生道德品质优劣的数字，而"教师对不同道德品质的学生，予以评定不同的等级"这种主观规定，就是规则。

从史氏对测量下的定义可以看出，测量包括三个要素：第一，事物的属性；第二，数字；第三，规则。

下面对这三个要素加以具体分析。

一、事物的属性是测量的对象

我们对事物进行测量，确切地说，测量的对象是事物的某种属性。例如，物体的长度、重量、体积、温度以及一个事件发生的时间长

短等，都是事物的物理属性。它们的存在形式比较具体，大多可以被人的感觉器官所直接感觉到，如看得见、听得到、摸得着、尝得出、嗅得到。但是，我们还往往需要测量人的心理属性，如学生的智力、个性、品德、知识、技能、习惯、能力、态度、兴趣、爱好等。它们的存在形式比较抽象，大多不能被人的感官直接感觉到。

二、数字是描述事物属性的符号

数字在未被用来表示事物的属性之前，它仅仅是一个符号，它本身没有量的意义。当数字被合理地用来描述事物的属性时，我们才赋予它以量的意义，即从数字变成了数。

数的特性为逻辑运算提供了许多可能性。数的系统是非常合乎逻辑的。

数的系统（指自然数）有以下几个特性：（1）同一性和区分性。所谓同一性就是指每一个数的独特性。例如，用同一个数字表示的事物必定是相同的。既然每一个数都是独特的，那么就没有任何一个别的数与它完全相同。这就是数与数之间的区分性。是 1 就不是 2，是 2 就不是 1，用 1 和 2 分别表示的事物是不相同的两个事物。数的同一性和区分性是一个问题的两个方面。（2）等级性或位次性。这是指若干个数之间按其大小所形成的次序关系。如 $3>2>1$。若用数的等级性描述事物，那么，事物之间必有位次可循。（3）等距性。若第一个数与第二个数之差，等于第二个数与第三个数之差（例如，1、2、3 三个数，$3-2=1$，$2-1=1$），那么，这三个数具有等距性。（4）等比性。若一个数可以表示为是另一个数的倍数，如桌子的长是宽的 2 倍，这类数具有等比性。

上述数的特性从低到高排列。一个数若具有较高的特性，则必具有较低的特性。

在实际测量中，由于测量的需要以及所欲测量的事物属性的不同，有时并不需要让数的各种特性同时具备。当然，能多具备一些更好，因为测量中运用数的效果，确实也与这些数所包括的特性多少相关联。

假如我们能用数合理地描述事物的属性，并且在允许的条件下，对数进行运算，我们就可以通过运算的结果，对所要测量的属性进行推测。如果事物的属性和数的系统之间，在性质上或形式上存在着高度的类似性，我们就可以用数来描述事物的真实情况。

三、规则是给事物属性分派数字的依据

测量中最关键且最困难的事情就是制定规则。所谓规则就是指导我们如何测量的一种准则或方法，即在测量时给事物的属性分派数字的依据。例如，有一种规则可描述为：对色盲者，分派数字 0，对非色盲者，则分派数字 1。如果有一个集合 $M\{M_1, M_2, M_3, M_4, M_5\}$ 共有 5 个人，其中有两个（M_2，M_4）色盲，三个（M_1，M_3，M_5）非色盲。假如我们另有一种能够清晰地界定色盲与非色盲的既定规则，那么我们就可以给集合 M 中的色盲者分派数字 0，非色盲者分派数字 1。假如把 0 与 1 也视为一个集合 N，则 $N=\{0, 1\}$。这类测量的原理可以用集合 M 与集合 N 的关系来剖析（见图 1.1）。

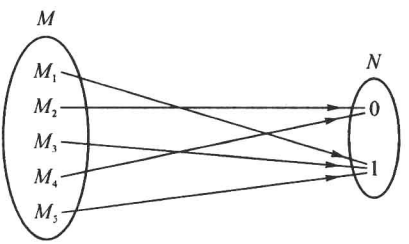

图 1.1　集合 M 与集合 N 的关系

从上图可以看出，集合 M 中的每个成员，仅能分派到集合 N 中的一个数字而已，呈现出一定的函数关系。这是一种有顺序配对的集合。其实，在数学上函数就是把某一集合中的事物分派到另一集合事物之上的规则。可以说，任何测量都呈现函数关系，而任何函数关系都是建立一种顺序配对的集合。因为被测量的成员可当作一个

集合，而分派到每个成员上的数字可当作另一个集合。我们可将测量的程序写成如下的一般公式：

$$f = \{(X, Y);\ X = \text{任何事物},\ Y = \text{一个数字}\}$$

这个公式可以解释为：函数 f 等于有顺序配对(X, Y)的集合，而 X 是一种事物，其相对应的 Y 是一个数字。

例如，被测量的事物(X')是 5 个学生，而数字(Y')为 1、2、3、4、5 五个等级。假使 f 是这样一种法则：对于学习最好的学生 X_4 给予等级 5，对于学习较好的学生 X_1 给予等级 4，……对于学习最差的学生 X_2 给予等级 1。那么，这种测量的程序可以用集合 X 与集合 Y 的关系图来表示，如图 1.2 所示。

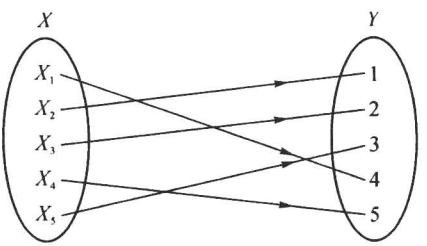

图 1.2　集合 X 与集合 Y 的关系

当测量的其他条件相同时，使用不同的规则，会产生不同的测量效果：使用好的规则，可以得到正确可靠的测量结果；使用差的规则就会得到无效或偏倚的测量结果。

规则的好坏，一方面取决于制定规则的程序；另一方面取决于所欲测量的事物属性本身是否易于建立规则和便于操作规则。一般来说，具体且稳定的事物属性，如性别、身高、体重等，其测量的规则就易于建立和使用；抽象且易变的事物属性，如人的智力、知识、技能、人格、态度等心理属性，其测量规则就难以制定和使用。

第二节　教育测量的可能性及其特点

一、什么是教育测量

从广义来说，教育测量就是对于教育领域内的事物或现象，根据一定客观标准，作缜密的考核，并依一定的规则将考核的结果予以数量的描述。

在教育工作中，根据科学研究的需要和改革教学方法、提高教育质量的需要，对学生的思想品德、学习成绩、健康状况的测量，对教师的教学效果的测量，对教学经费、物资设备以及行政管理效率的测量，都属于教育测量的范围。

从狭义来说，教育测量是指对学生某些学科经过学习和训练之后，所获得的知识、技能的测量，又称成就测量、学业成绩测量或学科测量。它是按教育测验的规则，对学生掌握某些学科的知识、技能予以数的描述。

二、教育测量的可能性

从教育测量的广义概念来说，它所测量的属性，虽然也包括了物理属性，如学生的身高、体重等，但它测量的主要对象是心理属性，如学生的知识、技能等。

物理属性的测量由来已久，其测量的可能性被大家所公认。例如，用尺量长短，用秤称重量，用钟表计时间，用温度计量温度，等等。这些都是客观的测量方法。其测量工具已达到相当精确的程度。

心理属性是否也可以客观地进行测量呢？

尽管教育测验在教学过程中已成为教师考核学生学习成绩不可缺少的工具，但是由于人的心理属性是抽象的，不易捉摸，实现客观的测量比较困难，因此有人对心理属性测量的可能性就产生了怀疑。

其实，心理属性与物理属性一样，都是可以测量的。恩格斯说：

“数学的对象是现实世界的空间形式和数量关系。”在20世纪初，美国心理学家和测验学者已为心理属性测量的可能性提供了以下两个理论基础。(1)任何现象，只要是存在的，总有数量。这个原则是由美国心理学家桑代克(E. L. Thorndike)在1918年提出的。他说：“凡物的存在必有其数量。”人的心理现象虽然看不见，摸不着，但它是客观存在的现实，是脑这块高级物质的属性，它也有数量的差异。例如，人的智力有高低之分，学生的学习成绩有优劣之别。这高低之间、优劣之间，存在着程度的不同。所谓程度不同，就是数量的不同。(2)凡有数量的现象，都可以测量。这个原则是由美国测验学者麦柯尔(W. A. McCal)于1923年提出的。人的心理属性也是可以测量的，虽然我们不能用尺来量它，用秤来称它，但是它必定会反映在某种活动之中，或表现在某种行为之中，于是我们就可以通过对人的行为的测量来推测他的某种心理属性。当然实现这种测量是很困难的，到目前为止，对于某些心理属性，如智力、创造力、知识、技能、习惯、品德、理想、兴趣、态度等，我们尚不能一一加以测量或测量得还不十分准确可靠。这是因为测验学的发展历史还很短，许多测量工具还没有发明，已发明的测量工具还不十分完善。但是，我们不能因为某种心理现象的测量工具还没有发明，就说这种现象是无法测量的。

三、教育测量的特点

教育测量与物理测量虽然都具有可能性，但测量的方法却不同。物理测量的对象是物质实态的性质、功能或组织。它们中的大部分是可以被感觉得到的。其测量的单位一般仅根据一定的空间或时间就可以确定，所以可以直接测量。此外，物理现象的变化甚小，引起变化的因素也较少，所以测量的实现既容易又精确。例如，一幅布，今天量为8米，几天或几个月，甚至几年之后，再量仍为8米。

教育测量的对象多属于人的心理属性，它是不能作为物质实态来操作的结构概念。这种结构概念，不能直接测量，只能从测量与这个结构概念有关的或从反映这种结构概念的(可测量的)因素着手，

对这个结构概念进行间接的测量。例如，要想测量学生算术运算能力，就得让学生完成一套有关的算术作业或测验题目，以引起他们的行为反应。这时，学生的算术运算能力必定会表现在他完成算术作业或测验的行为反应之中。这些行为反应正是他们算术运算能力所引起的结果。因此，我们可以通过对学生完成算术作业或测验行为的测量来间接地估计和推测学生的算术运算能力。这就是教育测量常用的间接测量方法。对学生测验结果的记分，虽然属于一种直接测量，但是，即使测量分数反映了学生的算术运算能力，它也不是算术运算能力的本身。

在物理测量中，有时也采用间接测量的方法。例如，测量室内气温的高低，就是通过观察温度计上水银柱的高低来确定的。因为水银柱的高低是气温变化引起的结果。

人的心理属性，往往是难以明确规定的，有些甚至缺乏公认的定义。另外，它们易受条件的影响而发生变化，制约它们变化的因素也甚多。因此，测量的实现较为困难。例如，一个学生在某种条件下，可以做出某种类型和某种难度的代数题目，而在另一种条件下，却做不出同一种类型和同等难度的代数题目。这是因为一个人的能力和作业成绩，易受心理动机、态度、情绪、发育、健康、睡眠、光线、气压、温度等因素的影响，难以得到准确可靠的测量结果。

然而，无论物理测量还是教育测量，其正确性都不是绝对的，而是相对的。因为即使是物理现象也会随着条件的变化而变化。例如，空中的高压电线会随气温的上升而伸长，随气温的下降而缩短。因此，在教育测量中一味地追求绝对的测量是徒劳无功的。

在20世纪60年代所创立的“模糊概念定量表示法”，把普通集合中元素对集合的绝对隶属关系（非0即1），用模糊集合隶属度（从0到1之间的许多实数）的思想来代替。所谓隶属度就是把对象属于某个事物的程度用“0，1”之间的一个实数来表示。隶属度中的0与1是两种极端，0表示最差，1表示最好，其他的情况处于0与1之间。这种定量表示法可能恰恰比较客观地描述了人的心理属性复杂程度的模糊性。

第三节　四种测量量表

由于事物属性不同，以及所制定的规则不同，致使用数的特性来描述事物属性所达到的程度也不同。这就产生了不同的测量水平。史蒂文斯将测量的水平分成四种，每一种测量水平都产生与其相应的测量量表。

一、名称量表

名称测量是测量中最简单的形式——分类。即属于同一类的事物用同一个数字表示，属于另一类的事物用另一个数字表示。用来描述各类事物的数字仅仅是事物的名称。它只具有相同与不同的特性，没有数量大小的含义。用这类数字表示的量表叫名称量表。例如，某市升学统一考试，学生准考证号码上，前两个数字是各个区的代号。如 03 是 A 区的代号，表示 A 区的学生；04 是 B 区的代号，表示 B 区的学生。又如，将学生按性别进行分类，凡男生用 1 表示，女生用 2 表示。如果既按性别分类，又按对某门学科喜欢和不喜欢两个标准进行分类，喜欢用 1 表示，不喜欢用 0 表示。于是男生喜欢者可表示为 11；男生不喜欢者可表示为 10；女生喜欢者可表示为 21；女生不喜欢者可表示为 20。在这里，用来描述事物的数字仅仅是代表事物的符号。它只能区分事物的类别，没有数量的大小、多少、位次和倍数关系。也就是说，它只具有数的同一性和区分性，而不具有等级性、等距性和等比性。因此，不能将之进行加减乘除四则运算。对于名称测量结果的数据所进行的统计处理，不是用来描述事物的数字本身，而是归入每一类中个体的数目（频数）。对这类点计数据所允许和适用的统计方法，有比率（相对频数，即某一类的频数与总频数之比）、百分比、Φ 相关系数、χ^2 检验。

二、等级量表或位次量表

对于事物的属性按一个标准进行分类，用来描述各个类别的数字，不仅具有区分性，而且还具有等级性（位次性），这些数字之间能表示事物大小的位次关系，但不具有等距性和等比性。用这样的数字表示的量表叫等级量表或位次量表。例如，将学生的口头表达能力分成甲、乙、丙三个等级。甲等用 3 表示，乙等用 2 表示，丙等用 1 表示。于是对于学生口头表达能力的评定构成了 $3>2>1$ 的位次关系。但是这些数字只能确定事物相等或不等的关系。在不等的情况下，只能确定大于或小于的关系，如 $3>2$、$2>1$，则 $3>1$ 的关系，却不能确定甲等的 3 比丙等的 1 大多少个相等的单位。因为 3 与 2 和 2 与 1 之间的差距是不相等的。因此对于量表上的这些数字不能进行加减乘除的运算。它们所能适用的统计方法，有中位数、百分位数、等级相关系数、肯德尔和谐系数（多列等级相关），以及符号检验、秩次检验、秩次方差分析。

三、等 距 量 表

有相等单位和人定参照点的量表叫等距量表。这种量表上的数值不仅具有区分性、等级性，还具有等距性。但是量表上的参照点（读数的起点）不是绝对零点，而是人定的参照点。例如，用摄氏温度计测量的温度，9℃与 6℃之差等于 6℃与 3℃之差。即 $9°-6°=3°$，$6°-3°=3°$。但是，这并不意味着 9℃是 3℃的 3 倍。这是因为摄氏温度表是以冰点作为人定参照点。摄氏零度并不意味着没有温度，而摄氏温度表上的绝对零点在零下 273℃，即 −273℃。时间量表上的参照点也是人定的。钟表上的零点，并不意味着没有时间。这类量表上的数值只能作加减运算，不能作乘除运算。它们所能适用的统计方法有算术平均数、标准差、积差相关系数以及 Z、t、F 检验等。