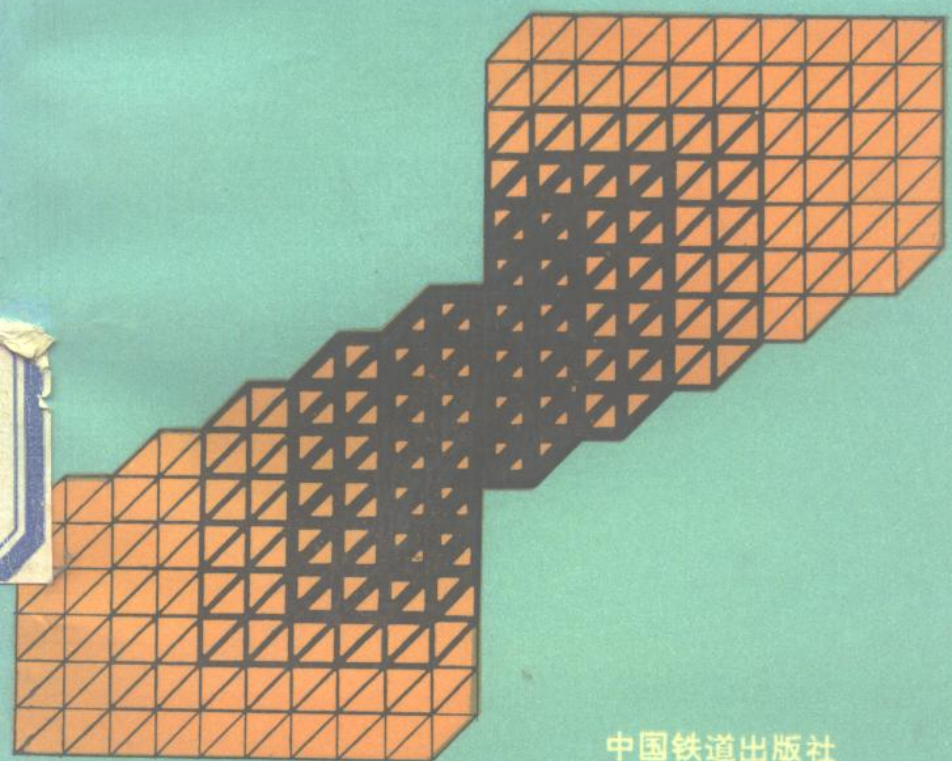
高等学校教学参考书

教育测量学基础

北方交通大学　许茂祖　张桂花　编



中国铁道出版社

高等学校教学参考书

教育测量学基础

北方交通大学 许茂祖 张桂花 编
北方交通大学 蒋焕文 主审

中国铁道出版社
1995年·北京

(京)新登字 063 号

内 容 简 介

本书介绍了教育测量的基本概念、主要内容及其发展历史，分析了教育测量理论对考试与考试改革、考试管理及教学评价等工作的指导作用及其与提高考试科学化、规范化水平的关系；研究有关考试的基本概念、功能、考试的科学分类及对各类考试命题的基本要求，试题的分类及各类试题的命题方法；以概率与数理统计理论为工具，研究了考试分数的收集、整理方法，以考试结果的正态性检验、一致性检验等方法评价试题质量和考试质量，并且介绍了如何利用计算机进行试题与考试质量的分析与评价(配套软件主编单位有售)，书中还介绍了标准化考试的概念、基本特征及如何实施标准化考试，如何对标准化考试进行组织与管理，并对标准化考试作了较为客观的评价。为便于读者自学，书中每章后面还附有思考题及习题。本书可作为高等师范学校学生的教材或教学参考书，也可作为向各级各类学校的广大教师及教学管理人员进行继续教育的培训教材。

高等学校教学参考书
教育测量学基础
北方交通大学　许茂祖　张桂花　编
中国铁道出版社出版、发行
(北京市东单三条 14 号)
责任编辑　倪嘉寒　封面设计　翟　达
各地新华书店经售
北京市燕山联营印刷厂印刷

开本：787×1092 毫米 1/32　印张：8.25　字数：181 千
1995 年 10 月　第 1 版　第 1 次印刷
印数：1—2000 册

ISBN7-113-02009-7/G・39　定价：10.70 元

前　言

考试是每个学校、每个教师经常要做的工作，因而也是每个教师十分熟悉的工作。但是，不少学校、不少教师所从事的考试工作，从命题、考试实施、阅卷直到对考试结果的解释和分析以及对考试质量的评价，还缺乏科学理论的指导，而主要是凭经验来进行的。当然，教师在教学实践中积累起来的丰富经验，对搞好考试工作是十分必要的，但是，仅凭经验而不以科学的理论去指导考试工作，势必使考试工作带有很大的盲目性，难以做到科学、客观、准确，更无法实现考试的标准化。这样的例子是很多的。例如，著名教育家斯太奇曾做过如下的实验：把同一份语文试卷复制142份，分别交给142位中学语文教师评分，结果共给出了35种成绩，最高为98分，最低只有50分；把同一份几何试卷复制116份，分别交给116位中学几何教师评分，结果给出了60多种成绩，最高为92分，最低仅28分。在国外，还有这样一件事例，某年夏季，许多大学教授在一起评阅历史试卷，有一位教授为工作方便，自己写了一份标准答案，由于工作疏忽，这份标准答案与其它试卷混到了一起，被另一位教授评阅，所给的分数竟是不及格；为慎重起见，这位评卷的教授又将它拿给其他教授评阅，结果各人所给分数相差极大，最高为90分，最低为40分，如果连同写这份标准答案的教授在内，最高分数应为100分。上面几个例子都是教师自编考试的情况。事实上，即使是在由专门机构组织的大规模考试中，也不乏这样的例子。例如，在我国的高等学

校入学考试中，就有类似的情况。1983 年在高等学校入学考试阅卷前，原教育部学生管理司曾从北京市的试卷中，随机抽取了语文、政治、数学、物理 4 科各 5 份试卷，复印后分发到全国除西藏、台湾以外的 28 个省、市、自治区，由这些地区的阅卷组分头评分，其结果如表 1 所示。从表中可以看出，不同地区对同一份试卷评分的结果，相差很大，最大达 33 分，最小也有 6 分。同一年，还请某省从所有科目中随机抽取 1 份试卷，复印后由同一科目的不同阅卷组分头评分，其结果如表 2

表 1　全国各地区对语文、政治、数学、物理每科 5 份试卷评分的差异

科　目	语　文					政　治				
试　卷	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
最高分	45	87	92	85	83	85	74	66	84	78
最低分	26	35	64	56	50	69	51	54	70	63
最大差异	19	32	28	29	33	15	23	12	14	15
科　目	数　学					物　理				
试　卷	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
最高分	109	119	114	118	114	91	80	90	80	69
最低分	94	107	103	110	103	85	71	77	74	62
最大差异	15	12	11	8	11	6	9	13	6	7

表 2　某省各科目不同阅卷组对同一份试卷评分的差异

科　目	语文	政治	数学	物理	化学	生物	地理	历史	英语
最高分	85	77	78	58	86	28.6	73	56.7	59.5
最低分	62	70	67	56	79	18.2	69	50.0	58.0
最大差异	23	7	11	2	7	10.4	4	6.7	1.5

所示。从表中可以看出，同一地区不同阅卷组对同一份试卷的评分结果相差也是比较大的，最大的达23分，即使是采用所谓客观题为主的科目——英语，各阅卷组的评分结果也相差1.5分。这几个例子说明，缺乏科学理论指导的考试，是很难准确地测量学生的学习情况和教师的教学情况的，即很难用它对教育进行科学、准确、客观的测量的。

因此，高等学校的教师和教学管理人员，有必要学习一些教育测量的基本知识，以指导考试、考试改革和考试管理工作，逐步使其科学化、标准化。为此，我们编写了本书。本书的第一、二、四、五、七、八章由许茂祖执笔，第三、六章由张桂花执笔，第八章的程序由崔铸提供。北方交通大学教务处长蒋焕文教授审阅了全部书稿并提出了宝贵的修改意见，在这里谨致以衷心的感谢。由于作者水平所限，书中缺点错误在所难免，欢迎批评指正。

编　者

1994年10月

目　录

第1章 绪 论

1.1 教育测量的概念

教育是一种以人才培养为目的的社会现象。专门研究教育现象的教育科学，也和其它任何一门科学一样，要经历从定性描述到定量分析的发展过程。要作定量分析，就离不开测量(*Measurement*)。教育测量学的创始人之一、美国心理学家桑代克(E·L·Thorndike 1874～1949)指出："凡是存在的东西都有数量，凡有数量的东西都可以测量。"教育作为一种社会现象，也是社会上客观存在的一种事物，它也是有数量的；例如，学习完一门课程，学生究竟掌握了课程教学内容的百分之几十；在一定时间内，一个学生可以做几道习题；一个实验，学生需要用多少时间来完成；一个学生在学习结束后，其学业成绩在一个学生群体中占有什么样的位置等等，都是教育这一事物在数量上的表现。既然教育是有数量的，那么，对它也是可以进行测量的，这种测量就是教育测量(*Educational Measurement*)。

可见，教育测量就是对教育的特征、属性及其运动、发展规律的定量描述，它主要用于对学生的精神特性进行数量化测定。具体地说，教育测量就是对学生的学习能力、学业成绩、兴趣爱好、思想品德、身体素质以及教育措施上的许多问题的数量化测定。

一般物理量的测量工具是各种量具和仪器。由于教育测量的对象是学生的精神特性，而不是具体的物体的物理特性，

因此，不可能利用一种具体的量具或仪器来直接进行测量，而只能通过出考题给学生解答来作间接的测量。也就是说，教育测量是这样实现的：以考试(*Examintion*)与测验(*Test*)引起学生的某种行为，并对行为的结果作定量描述，这一定量描述就是教育测量的结果。这样看来，考试与测验是教育测量的工具。

这里所谓的考试，是指教师凭自己的经验进行命题和评分的这样一种测试，也就是传统概念上的考试，或称为旧式考试，而测验则是指经过比较细致的科学分析来编制试题、按严格的程序进行的这样一种测试。因此，严格地说，考试与测验的含义是不同的，但是它们之间并无一条不可逾越的界限，在实际的教育测量工作中，这两个概念常常互相不加区别地通用。本书将一律使用"考试"这个概念。这样，也可以说，考试是教育测量的工具。

应当指出，考试是教育测量的一种工具，但它不是教育测量的唯一工具，还有许多其它工具与手段如调查、观察、对事物进行数量化分析等，都可以对教育这一事物进行数量化测定，即进行教育测量。尽管如此，由于教育测量的对象是一种精神的、心理的特性，考试仍不失为它的一种最主要、也是年复一年地使用最多的一种手段。本书也只研究采用这种手段所进行的教育测量。

从教育测量的定义和其实施可以看出，从本质上说，教育测量是一种比较活动，即通过一定的方法，将被测量的事物与作为测量标准的参照事物相比较，以便对被测事物赋值。这一点，与一般物理量的测量是一样的。通过比较与赋值，就得到了测量结果——量值，它由两部分组成，一部分为所赋的数值，另一部分就是作为测量标准的参照事物，称为测量单位。

显然，测量单位是很重要的，没有测量单位，就无法进行测量。对测量单位有如下要求：具有客观性、稳定性和等值性。如果以不具备这些特性的事物作为测量单位，测量工作是不可能准确、真实和客观地进行的。然而，传统考试所使用的测量单位——分数是不具备上述特性的，因此传统考试是不够科学的，要进行科学的考试就必须解决这个问题。

综上所述，测量对象即被测量的事物、测量方法与测量工具、测量单位就构成了教育测量的三个要素，这三个要素，对于完成教育测量工作，都是必不可少的。

通过测量，可以实现对被测量的事物的数量化描述，即得到量值。量值是出现在一个有参照点和单位的连续体上的，这个连续体称为量表。根据如在学校考试或考查中常用的百分制记分法或五分制记分法(1、2、3、4、5 或优、良、中、及格、不及格)所采用的量表，就分别称为百分量表或五分量表。

根据量表本身的属性在本质上的不同，它可以分为五大类。

1. 名义量表，又称为分类量表，即用分类的方法，对被测事物进行定性的描述。显然，这一与分类密切联系的量表是一种低水平的量表，但却是在教育测量中应用最为广泛的一种量表。例如，在考试中，选择题这种所谓的客观试题，就是这样一种量表，因为对答案只能按对与错来分类。当然，名义量表也不限用于二分变量(即只有两个参照标准)，也可以有多分变量(即有多个参照标准)，如高等学校的教师，按专业技术职务可以分为教授、副教授、讲师、助教等。

2. 位次量表，是用反映事物的相对顺序关系的数值来表示的一种量表，比如，按学生的考试成绩来确定第一名、第二名、第三名等等。这种量表中的数值，是不能进行加减运算的，

但可以用于作次序的统计。

3. 间距量表，以确定的单位间的距离来描述被测量的事物，但变量不具有零点的量表称为间距量表，用这种量表表示的数值，可以互相作加减运算，但不能表示倍数关系，因此，不能作乘除运算。

4. 比率量表，是具有等距、等质性且具有零点的量表，用这种量表表示的数值，不仅可以作加减运算，还可以作乘除运算。与前面三种量表相比，这是一种高水平的量表。

5. 模糊量表，用来描述模糊关系的量表称为模糊量表。由于在教育活动中有许多事物只具有模糊的关系，而不具有明确的非此即彼的关系或精确的数量关系，比如人的科学道德、团结互助精神等，就具有一定的模糊性，只能用高、不太高、不高等模糊的语言来描述，这时所用的量表就是模糊量表。事实上，在教育测量中用得最多的还是这种模糊量表。

上述各种量表，虽然各具不同的属性，但这种分类只是相对的，事实上，这几种量表之间还是有一定的关系的，这种关系主要表现在如下两个方面。

第一，各种量表之间以一定的依存性互相联系着。例如，在前三种量表中，每一种量表都包含了它后面的各种量表，譬如位次量表就包含在名义量表中，因为按照某一属性对客体进行分类后，如在此基础上对其中可以排出顺序的属性，按顺序将各客体进行排序，名义量表就发展成了位次量表。又如，对一组客体，已经按位次量表排了名次，在此基础上，为了进一步作定量的描述，还可以用很好、好、较好、差、很差等模糊语言对这些已排出顺序的客体作模糊描述，然后再对这些模糊语言赋值，位次量表就转化成了模糊量表。

第二，随着人们对于教育活动中各种事物及其属性的认

识的深化和教育测量手段的发展，对同一事物属性的描述，就可以由只能使用低水平的量表进而发展到可以使用高水平的量表。例如，在1916年引入智商的概念以前，对儿童智力的测量，只能按位次量表进行，而引入智商的概念以后，人们对智力的认识出现了质的深化，于是就可以用间距量表来对儿童的智力进行测量了。

1.2 教育测量的特点

教育测量是一种对社会现象的测量，多数人对它是比较生疏的。但一般人对一般物理量的测量即对自然现象的测量是比较熟悉的。既然这两种测量都属于测量的范畴，那么它们必然有共同的特征和属性；既然这两种测量的对象有本质的不同，那么它们也必然有本质上不同的特征和属性。通过对这两类测量的特征与属性的比较，将有助于我们对教育测量的特点有更清楚的认识。

1. 教育测量与一般物理量测量的共同特点。教育测量与一般物理量的测量都是对客观的事物的存在、特征、属性及运动规律作定量描述，或说是进行数量化测定的过程。这一测定过程就是对客观的事物的存在、特征、属性及运动规律赋予一定的量值。显然，无论是教育测量，还是一般物理量的测量，都是一种事实判断的过程。

为了对客观事物的存在、特征、属性及运动规律赋值，无论是教育测量，还是一般物理量的测量，都要利用一定的工具，采用一定的客观的尺度作标准，通过测试(*Test*)的手段来实现。

由于受到人们对客观事物认识水平的限制，无论是教育测量，还是一般物理量的测量，在测量结果中都不可避免地会

存在测量误差，而且按测量误差的来源与性质可以分为系统误差和随机误差，这种测量误差影响着测量结果的有效性和可信程度。在实际测量中，必须设法减小测量误差。

以上几点，就是教育测量和一般物理量测量的共同之处。这些共同之处反映出这两种测量的基本特性是相同的，因此，它们都被称之为"测量"，在许多方面，遵循着共同的或者是相似的规律。

2. 教育测量与一般物理量测量的不同点。尽管教育测量与一般物理量的测量具有上述相同的基本特性，但它们毕竟不是一回事，它们在许多方面也表现出不同的特点。

(1)测量对象的特点不同：一般物理量测量是对客观事物的物理特征与属性进行的测量，而教育测量是对学生的心理特性(精神特性)进行的测量。这两种测量对象各自具有不同的特点。客观事物的物理特征与属性是比较具体也比较稳定的，比较容易控制，比如，物体的质量、长度，今天具有的量值与几个月以后具有的量值，应该是基本相同的。但学生的心理特性一般地说是比较抽象的，而且由于学生是活生生的人，其心理特性时刻在发生着变化，这种变化又不易控制。比如一个学生，今天在一个小时内可以做 10 个数学习题，也许明天在一个小时内就只能做 9 个或能做 11 个同样的数学习题，因为他做习题的速度不仅取决于题目的难度和工作量的大小，而且与他的精神状态、身体状态也有很大的关系。测量对象的不同特点，决定了教育测量比一般物理量的测量要复杂得多。

(2)测量尺度的客观性不同：一般物理量的测量尺度，无论是测量质量的克、千克，测量长度的米，还是测量时间的秒等等都是比较客观的，不会因测量者和测量条件的改变而改变。但是，教育测量的尺度，是含有一定的主观因素的，即各个

测量者心目中的尺度并不完全相同，即使是同一个测量者，在对同一个对象进行多次测量时，由于测量条件的改变，他心目中的测量尺度也会改变。这样，尽管我们要求教育测量的尺度是客观的，但事实上其客观性远不如对一般物理量测量的尺度。

(3)测量结果的客观性不同：在一般物理量的测量中，由于被测事物的物理特征与属性是比较稳定、比较容易控制的，而且测量的尺度又是客观的，所以，在排除了测量误差的影响之后，测量的结果将是客观的、唯一的，不因测量者的不同而不同，也不因测量条件的变化而变化。但是，在教育测量中，由于被测对象是时刻在变化的、不容易控制的学生的心理特性和精神特性，而且测量尺度又带有一定的主观性，表现在不同的测量者对教育活动中的同一事物进行测量，往往会得到不同的测量结果(如不同的教师评阅同一个学生的同一份试卷，往往得到的分数是不同的)；即使是同一个测量者，在对教育活动中的同一事物进行多次测量时，往往也会得到不同的结果(如用同一个教师出的几份等量的试卷对同一个学生进行几次考试，往往得到的分数也是不相同的)。

应当指出，教育测量的尺度和测量结果的客观性不如对一般物理量的测量，但绝不意味着教育测量的尺度与测量的结果是主观的，也不意味着教育测量没有客观的尺度和测量结果。事实上，教育测量的尺度是客观存在的，测量者即测量的主体也都力求客观地掌握测量尺度，因此，测量的结果也是客观存在的。但是，由于教育测量的尺度是要由测量的主体来掌握的，在测量的实施过程中，测量的主体是将被测量的事物与他心目中的标准(如标准答案等)进行比较后进行赋值的，所以，测量结果在一定程度上取决于测量主体及其心目中的

测量尺度，我们称这种现象为教育测量具有主体性。

综上所述，我们可以看出，教育测量具有以下几个特点：

1. 具有一般测量的共同特征与属性；

2. 测量对象与测量标准具有社会性；

3. 测量尺度具有较大的不确定性，测量单位具有较大的近似性；

4. 测量标准及测量结果具有主体性。

鉴于教育测量的上述特点，为使测量结果能比较真实地、准确地、客观地反映出被测对象的特征与属性，就必须对教育测量的标准物有一定的要求，或说教育测量的标准物应该满足以下一些基本条件。

1. 不低于构成被测对象稳定状态的最低限度；

2. 具有实用意义上的合理性；

3. 在测量过程中单位的不变性，即在一次测量过程中，不能使用不同的单位。

以上 1、2 两个条件是容易满足的。第 3 个条件在对一般物理量的测量中是不难满足的，但在教育测量中有时却是不容易满足的。例如，在测量科研成果时，可以用完成鉴定的科研成果的项目数或发表的科研论文的篇数来测量，但这一项成果和那一项成果，虽然都是一项，但它们很难是等值的；这一篇论文和那一篇论文，虽然都是一篇论文，也很难是等值的。在这里，测量过程中测量单位就不具有不变性。而在上述这类测量中，要找到一种具有不变性的测量单位，却是非常困难的。

只有正确地认识教育测量的这些特点，才能正确地掌握教育测量的规律和技术，正确地进行教育测量工作。