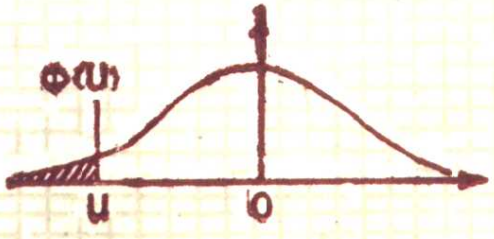
教育测量与统计方法基础



卢正勇

福建教育出版社

教育测量与统计方法基础

卢正勇

福州大学图书馆藏书印

福建教育出版社
一九八七年·福州

内容提要

教育测量学是研究心理与教育的测量和评价的科学，它以现代教育学、心理学和统计数学为基础，运用各种测试手段和数理统计等工具，探讨对教育效果进行测量和评价的原理和方法。

本书以教育测量的知识为主线，结合介绍教育统计的必要知识。内容包括测验及试题质量的衡量，信度、效度及难度、区分度的计算试题试卷的编制方法，测验分数的评价。

正确运用教育测量的知识，能使教育测量在教育工作的许多方面发挥应有的作用。本书可供大中小学教师及教研人员阅读参考。

教育测量与统计方法基础

卢　正　勇

福建教育出版社出版
（福州大梦山7号）
福建省新华书店发行　　福安县印刷厂印刷
开本787×1092　1/32　7.5印张　字数156千　插页3
1987年12月第一版　　1987年12月第一次印刷
印数：1—3,100

ISBN7—5334—0126—3　　书号：7159·1312
G·83　　定价：1.45元

绪　论

教育测量学是研究心理与教育的测量和评价的科学。它以现代教育学、心理学和统计数学为基础，运用各种测试手段和数理统计等数学工具，探讨对教育效果进行测量和评价的原理和方法。其内容包括智力测量、学业成绩测量和品德测量等方面。

正确运用教育测量的知识，能使教育测量在教育工作的许多方面发挥其应有的作用，如改革入学考试，合理分班；估量学生的学业成绩，诊断学生学习缺陷，确定学生学习难点，测量学生智力水平，发现特殊才能，预测学生的发展；分析学生兴趣爱好，指导学生升学就业；评定学生道德品质；评价师生教与学，进行教学改革，估量教育行政工作效率等。教育测量所掌握的事实与数据，是学校与教育行政领导对教学、教育作出重大决策的依据。例如根据学生学习困难诊断测验的结果，在教学上作出有针对性的补救措施的决策；根据教学效果的测量与评价，作出改革课程、教材、教法的决策；根据毕业考试、高考或职工文化考试成绩与评价，作出录取新生、招工、招干的决策等。

我国是文明古国，在教育测量方面有着悠久的历史。隋唐以来以论文方式选拔文官的科举制，是世界上最早施行的论文式测验，这在当时的历史条件下是曾经有过积极意义

的。法国思想家伏尔泰曾经这样评价："人类精神，肯定想象不出比这样的政府更好的政府。在这个政府里，重要的衙门彼此统属，任何事情都在那里决定，而其成员都是先经过几场严格考试的。"（至于科举考试发展到明清，流于以八股文取士等弊端，另当别论。）但是，教育测量作为一门独立的科学，是廿世纪初形成的。当时，心理学界对个别差异的研究、实验心理学的研究及统计数学的长足进步，为教育测量的发展提供了必要的前提；教育科学化的运动，促进了教育测量理论的研究与测量工具的改进。这样，教育测量的研究迅速发展起来，并逐渐形成一门科学，高等院校也开设了相应的课程。当前，许多国家和地区还建立了大规模的考试专业机构，从事教育测量的研究并编制各种类型的标准测验。各级各类学校的教师可以自编测验，也可以根据不同需要购买标准测验试卷。大学或研究生院可以要求考生参加考试机构举办的各种专门测验，并根据其成绩和其它信息决定是否录取。最著名的考试专业机构，美国的教育考试服务社（Educational Testing Service简称ETS），于1948年1月1日开始工作，拥有2700名工作人员，主要举办高等学校入学考试（STA）、研究生入学资格考试（GRE）和著名的英语"托福"（TOFEL）考试等。

我国在卅年代亦曾开展教育测量的研究，教育测量学曾作为高等、中等师范院校的必修课开设。解放后，受到苏联教育思想的影响，全国师范教育系统完全停开了这门课程；"文革"中，教育又备受摧残。三十多年来，高等、中等师范院校学生缺乏教育测量知识的教育，致使目前在教育第一

此为试读，需要完整PDF请访问：www.ertongbook.com

线的不少教师、教育行政干部缺乏教育测量的基本知识。有的教师对当前考试改革中出现的新事物很不理解；从家长、教师直至教育行政领导干部，对分数的解释与评价都存在不少错误，这典型地表现在有些地方举行的检查教学质量的统考之后，简单地以各校平均分数的高低，评价各校教育工作的优劣，评价教师教学水平的高低，从而产生不良的影响。由此可见，广大教育工作者学习教育测量的基本知识，对理解与推动当前的教育改革、促进教育质量的提高是十分必要的。

教育测量学的内容十分丰富，本书限于篇幅，拟以学业成绩测量为主，介绍教育测量的基本知识，内容大体可以分为以下四个方面：

①教育测量的基本概念（第一章）；

②如何测量学业成绩（第一章）；

③编制测量工具的原理与方法（第二、三、四章）；

④测量结果的评价（第五章）。

教育测量学是一门交叉学科，它与教育学、心理学特别是数理统计学、模糊数学等有密切联系。例如教育测量学中的试卷效度、信度及试题区分度分析，必需使用数理统计学中的相关系数等知识；教育测量结果的解释，如量表的编制、常模资料的计算等，必需用到统计学中的均值、标准差、正态分布、统计假设检验等知识。本书在处理教育测量与统计两者关系时，以教育测量的知识为主线，结合介绍教育统计的必要知识。此外，各章后都附有一些参考文献，以便于读者进一步学习；书后附有练习题，供读后练习。

目　录

绪　论

第一章　教育测量与评价概述

§1　测量、测验与评价的概念……（1）

§2　学业成绩的测量与评价的作用……（5）

§3　测验的分类……（6）

§4　学业成绩的测量方法……（11）

第二章　如何衡量测验及试题的质量

§1　衡量测验质量的标准……（15）

§2　衡量试题质量的指标……（47）

第三章　信度、效度及区分度的计算

§1　总体与样本、统计抽样法……（54）

§2　统计特征数及其计算……（60）

§3　相关系数……（72）

§4　试卷信度的计算方法……（81）

§5　试卷效度的计算方法……（89）

§6　试题区分度的计算方法……（94）

§7　试卷、试题质量指标间的关系及提高试卷信度、效度的途径……（99）

第四章　试题与试卷的编制方法

§1　主要题型的优缺点及编制技术……（107）

§2　标准化测验的概念……（134）

§3 标准化测验试卷编制的一般程序……（136）
第五章 测验结果的评价
§1 测验分数的整理，正态分布……（147）
§2 评分体制和不同体制下分数的解释…（157）
§3 分数的组合……（172）
§4 学生某科成绩进步的评价
与各科成绩关系的评价……（176）
§5 统计检验方法……（184）
§6 学校（班级）间同科统考平均成绩的比较
与评价……（198）
§7 同年级（班级）不同学科间统考平均成绩
的比较与评价……（200）
§8 学校（班级）统考平均成绩与常模团体平
均成绩的比较与评价……（203）
习 题
答 案
附：常用数理统计用表
表一 由ρ值求r的对照表……（220）
表二 相关系数临界值表……（222）
表三 正态分布表……（223）
表四 正态分布的双侧分位数（u_α）表……（229）
表五 t分布的双侧分位数（t_α）表……（230）
表六 x^2分布的上侧分位数表……（232）
表七 随机数表

第一章　教育测量与评价概述

§1　测量、测验与评价的概念

一、测量与教育测量

测量是根据某种规则给被测对象指派某个数值，并称它为该被测对象的测量值。或者说，测量是根据某种规则，用数值描述事物的量、质或发生次数、类别等，测量过程是对事物某一性质的数量化过程。

例如我们要测量一条绳索的长度，这里绳索的长度是绳索的一种属性，我们要给它指派某个数值，如何指派呢？人们公认的规则是，把绳索拉直，并与长度单位（例如厘米）进行比较（即用刻有长度单位的尺来量），如果该绳索是长度单位1厘米的200倍，则指派200厘米作为这绳索的测量值，测量时使用的刻有长度单位的尺，便是测量物体长度的一种工具。

教育上的测量，意义与此类似。它是根据某种规则，给学生达到教育目标的程度指派某个数值，测验是教育测量的一种工具，所谓测验，是指编制试卷、施测和评分等的一系列系统程序，它以一组试题施测于考生，引起考生的反应，

并以此来估计考生的学业、智力等。人们还常使用“考试”一词，在英文中译时，习惯上把“*test*”译为“测验”，而“*examination*”译为“考试”。在学业成绩的测量中，常把两者视为同义词，本书对此两词亦不加区别。

二、量表及其分类

用温度计测量气温，温度计上标有零点、刻度（温度单位）它是测量气温的一种量表。类似地，用百分制评定学生的学业成绩，最低分为0分，最高分为100分，并以1分为单位，因此它也可以看作是测量学业成绩的一种量表。测量不同对象，使用不同的量表，斯蒂文斯（S. S. Stevens.）把量表归为四种类型：

（1）比率量表

这种量表具有绝对零点与等距的单位，而且单位可以细分。其测量值是一个实数，如物体长度、重量的测量量表都是比率量表。比率量表的测量值倍数关系是有意义的。例如某甲体重是60公斤，某乙体重是30公斤，我们可以说，甲的体重是乙的2倍，物理测量中的量表大都是比率量表，而教育测量中的量表几乎都不是比率量表。

（2）区间量表

这种量表有等距的单位，而且单位也可以细分，测量值也可用实数表示，但它与比率量表不同的是它没有绝对零点，只有相对零点。例如摄氏温度计就是物理测量区间量表的典型例子，它的零点是相对的，0℃不表示没有温度；设甲、乙物体的温度分别为40℃与20℃，我们可以说甲的温度

比乙高20℃，但不能说甲的温度是乙的2倍．也就是说，区间量表的测值的倍数关系是没有意义的。

教育测量中的百分制评分，通常近似地看作一种区间量表。为什么说是“近似”呢？因为区间量表的单位是等距的，而百分制评分中的每1分所代表的量是否相等是不清楚的。这是因为教师在命题时往往使学生得低分容易而得高分难，所以百分制把学生成绩分为100个单位，它们只是近似相等的。这样，百分制评分，至多只是一种近似的区间量表，百分制评分中的0分与温度计中的零度一样，只是一个相对零点。考0分的学生并非对所考核的知识一无所知，只是相对于这份试卷的试题内容他一无所知。此外，得60分的学生知识也并非是考30分学生的2倍，这是由于零点是相对的，如果零点向下移动10分（即考试标准降低一些），那么得60分与30分的学生在理论上将变成得70分与40分，他们的分数的比值将是7：4而不是2：1。

（3）顺序量表

这种量表的测值只反映事物的先后次序、级别，没有相等距离单位与绝对零点．例如把全班学生成绩排名次，名次就是顺序量表的测值，我们可以说第一名比第二名好，第二名比第三名好，但第一、二名间的差异与第二、三名间的差异程度是否相等是不管的，又如苏联的五级记分制，考试成绩的5分、4分、3分等也只是表示成绩优劣的顺序关系。对顺序量表的测值进行四则运算都是没有实际意义的。

（4）称名量表

这种量表数据的不同，只说明事物的类别或属性不同，

此为试读，需要完整PDF请访问：www.ertongbook.com

对学生编考号，也是对不同学生的一种数量化方法。其间没有大小关系，不能进行运算，例如1号学生“加”2号学生并不能等于3号学生。

了解各类不同的量表，对测量结果的解释具有重要意义。上面说过，常用的百分制评分所得的分数，只能近似地看作区间量表的测值。因此，某生的分数是另一学生分数的二倍并不说明他们知识间也有这种倍数关系。此外，分数的参照零点是相对的，而且通常各试卷参照零点并不相同。这样，原则上只能直接比较同一份卷施测的学生成绩，如要比较不同卷测验的结果，就需要寻找共同的参照零点与单位。不了解这一点，往往是人们对分数产生误解的重要原因。

三、教育评价

教育测量与教育评价是经常联系在一起的。教育测量是对学生所达到教育目标的程度加以数量化，而教育评价则是根据测量结果等各方面信息作出教育价值的评论。例如某生在某科考试中获得95分，这里95分是一个测量值。如果问该生该科学得怎样，假设这份试卷是参照教学目标要求编制的，则我们根据该生获得95分的成绩，评论他该科成绩优异，这就是给该生该科教育效果的一个评价。又如某市施行某科统考后，根据甲、乙两校考生的统考分数（测量结果）通过统计分析后，推断甲校该科的教学效果优于乙校，这也是一种评价。

测量结果是评价的重要依据，但不是唯一依据，例如教师对学生的评价，不仅通过考试分数，还应通过平时观察或

其它途径作出综合评价。

§2 学业成绩的测量与评价的作用

学业成绩测量与评价的作用可归纳为以下主要的三方面：

①促进学生学习：学业成绩测量与评价是通过测验来进行的。测验是教学过程的重要环节。它能促使学生在考前对教学内容进行复习、巩固，并对知识进行系统的综合整理。测验的基本功能是教育行为的反馈，对学生来说是学习行为的反馈，测验能使学生知道自己哪些内容真正学会了，哪些内容模糊不清，哪些内容根本不懂，并知道自己学习态度与学习方法存在的问题，从而调整今后的学习。

②促进教学工作的改革：测量与评价能为教师及学校领导提供教学工作效果与效率的反馈信息，提供包括教学计划、大纲、教材、教学思想、方法在内的一系列改革的方向。

③甄别学习水平，指导教学管理及因材施教：测量与评价能甄别学生的学习水平，对学校的教学管理（如补考、升留级、分班、分组等）提供决策依据。此外，通过测量和评价还能发现特殊才能的学生或差生，进行因材施教。

总之，教育测量与评价，当前已逐渐成为教育决策的重要依据之一。

§3 测验的分类

依照测验的使用规模，可把测验分为学校教育过程中使用的测验与社会上举行的大规模考试两类；依测验分数解释的参照标准，又可分为目标参照与常模参照考试两类。

一、学校测验与社会上的考试

1. 学校教育过程中使用的测验

这类测验，依测验的目的和作用，又可分为如下三种：

(1)配置性测验

配置性测验一般在各学年、各学期开始前或开始时举行，其目的是了解学生是否具有新的教学目标所要求的预备知识与能力。换句话说，配置性测验是一种摸底测验，摸清情况，以便更恰当地编班、分组，根据学生水平安排教学计划与进度，提出恰当的教学要求并使用适当的教学方法。这种测验最常在初中、高中、大学一年级新生入学后举行。例如某农村中学一位初一数学教师，在初一学生入学时，编制了一套小学数学的水平测验题，分为基本概念、四则运算、解应用题等几个大部分，以了解入学新生对各部分数学知识掌握的实际程度与存在问题。于是把一部分程度很差的学生编为一班，并根据该次测验中发现的问题，例如分数四则运算生疏等，有针对性地用一个月时间进行补课，使他们真正达到小学毕业水平，而后学习有理数。这样做，这个班进度虽然慢于其它班，但学习基础补好了，两年内就跟上另一

班，取得良好效果。

(2)形成性测验

这种测验一般是在教与学的过程中进行的。目的在于了解教学的效果，诊断学生学习中的缺陷，探究教学中存在的问题，以便对教学工作进行调整。这种测验即通常所谓的平时考试或阶段（单元）测验，通过这种测验，能分析学生对所学诸方面内容的掌握情况、困难所在，还能分析同一个学生在不同学科中的成绩，据以提供对学生学习的指导。根据多次测验的结果，还可分析学生变化、进步的情况。总之，它是通过对教学结果的了解来调节教学工作的一种测验。由于它既能诊断学生学习上的困难，又能评价学生的进步情况，故又称为诊断——进步测验。

(3)总结性测验

这种测验一般在学期（学年）末进行，目的是为了解学生通过一学期（学年）的学习，是否达到教学目标的要求，故亦称成就考试，总结性测验也可以提供下学期（学年）编班分组的资料，因此，总结性测验亦具有配置性测验的功能。

2. 社会上举行的大规模考试

社会上举行的大规模考试，通常可以分为三种类型：

(1)水平考试

这种考试是测量考生现有的知识水平和对某门课程的掌握程度。人们预先定下某门课程的特定水平标准，只要考生达到这个标准，就承认其学历，发给通过这种水平考试的证书，而不管考生以前是否系统学过该门课程和以什么方式学

习这门课程。

这类考试的例子可以举出许多，例如我国实行的高等教育自学考试就是这类考试。它是通过考试，从广大社会自学者中选拔达到普通高校相应本、专科水平的人才，国家制定了每一专业的考试计划，而每一门课程的特定水平标准则体现在国家公布的每门课程的考试大纲上，一个考生通过了考试计划中的某门课程考试，成绩及格以上，便发给该科结业证书。考试计划规定的全部课程及格了，便发给该专业的专（本）科毕业证书，国家承认其学历。此外，我国用于考核出国留学生英语水平的EPT考试，美国用于考核进入美国学习的外国留学生英语水平的“托福”考试，都属于水平考试。

(2)学能考试

这类考试是测量考生将来有无学习某种专业的能力。例如高考、研究生入学考试等。以高考来说，一个人可以因为受到了精心教育而从中学毕业，但高等学校想挑选的是将来进入大学某专业后能够学好，而不光是过去已学好的人。因此这种性质的考试带有预测性，它更侧重于将来学习时所必需的某些能力的测量。

(3)竞赛性考试

这类考试旨在发现、选拔具有某方面特殊才能的人，以便进一步培养。例如数学竞赛、化学竞赛、物理竞赛等。

二、目标参照性考试与常模参照性考试

本世纪七十年代，人们按评价测量结果所参照的标准的

不同，把考试区分为以下两大类：

1．目标参照性考试

这种考试是以某种目的的需要而确定的标准为依据来进行命题和解释分数的。其及格的参照点是最低教学要求所达到的水平，其分数解释是完成目标情况和能力水平，即达到目标的“完满程度”。

视力测验是这类测验的最简单例子。一个人的左眼能看清视力表上某一行的符号，表明了他左眼的视力水平，完满标准应是1.5。如果一个人左眼视力为1.0，说明他与标准相比较还差0.5。

在使用程序教学法教学时，其学习的各小阶段自我检查测验，大都采用目标参照测验的办法来编制，即针对某个教学目标，拟定一组最低要求的检查题。学生能顺利解答这些问题，则认为这个教学目标已达到基本要求，可以继续学习下一单元；否则，需再学与答错问题有关的本单元内容，直至达到基本要求后才往下继续学习。

2．常模参照性考试

这是依照测验集体的常模数据（例如平均数、标准差等）来解释分数的考试。这类考试的目的是把个人的成绩与其他人的成绩作出比较，着眼于把人们的成绩区别开来，其分数解释是：分数大小表明一个人在该集体中的相对地位高低。这类考试的命题要有利于甄别考生的水平，应使分数有效大幅度的差异，以便于比较。

竞赛与升学考试（如高考）是这类考试的典型例子。例如数学竞赛的主要目的，是从应试者中挑选少数优胜者。考

此为试读，需要完整PDF请访问：www.ertongbook.com