

JIAO YU CE LIANG JIAO YU CE LIANG

● 教育测量

郭述平 王景英 编著

东北师范大学出版社

教育测量

郭述平 编著
王景英

东北师范大学出版社

教 育 测 量
JIAOYU CELIANG
郭述平 王景英 编著

责任编辑：有 华 封面设计：甘 莉 责任校对：路 遥

东北师范大学出版社出版
(长春市斯大林大街110号)

吉林省新华书店发行
长春树勋印刷厂印刷

开本：850×1168毫米1/32

1988年12月第1版

印张：8

1988年12月第1次印刷

字数：196 千

印 数：0001—7000册

ISBN 7-5602-0229-2/G·80压膜 定价：2.50 元

序

教育测量学不是一门新兴科学，但自从建国以来，由于某种原因，它并未得到进一步发展和应用。近十数年来，研究这门科学的人日渐增多，可谓一件大好事。辽宁郭述平同志和吉林王景英同志根据工作需要，对教育测量学进行了广泛研究和实践。同时，吸取了这门科学的最新成果，合写出《教育测量》一书。

在该书问世之前，我有幸得以先睹为快，从中学到不少的新东西。我认为这本书突出的一个特点，是理论与实践密切结合，富有实践性。他们无论在参加考试改革的实践中，还是在教育统计、教育测量的教学实践中，都是以教育科学（包括教育测量学）理论为指导的，从而取得了丰硕成果，并把所获得的第一手材料上升到理论高度。“实践出真知”。从郭述平、王景英二同志的实践中，从他们合著的《教育测量》的诞生中又可证明这个真理的正确性。

这本书的另一特点，是作者把辩证唯物主义观点贯穿于全书的始终。不论主观性试题和客观性试题及其评分的阐述，不论常模参照考试分数和目标参照考试分数的解释，不论信度和效度的估计方法，都在充分肯定他们各自的优点和适用条件的同时，又指出它们的缺点和局限性。这就避免片面性，克服了形式主义。

这本书的第三个特点是贯彻学用一致的原则，尤其注重应用。作者对于每一理论的阐述不仅清晰、透彻，一看就懂；而且把它的应用条件和技术写得条分缕析，极易掌握。就考分的合成与试题的编制技术而论，作者不仅讲明科学道理，而且对于试题如何编制，分数怎样合成，更以浓厚的笔墨详加介绍。

这本书优点甚多，恕不一一备述。总之，它的体系比较完整，结构相当严谨，观点明确，分析清楚，堪称这门科学中的一本好书。它可以作为大专学校教材，可以作为行政干部和教师参考用书，也可以作为有志之士自学用书。相信它的出版必将有助于教育测量知识的普及，有助于我国教育测量学的发展，有助于我国现行考试制度的改革。

写了这些，就算是个序吧。

曹延亭

1987年5月中旬于长春

目 录

序	(1)
第一章 教育测量的历史简介	(1)
第一节 教育测量的发展简史	(1)
第二节 中国教育测量的曲折道路及现状	(9)
第二章 教育测量的一般理论概述	(11)
第一节 教育测量的概念	(11)
第二节 教育测量的性质与功能	(22)
第三节 编制测验的一般原则	(31)
第三章 考试分数的统计分析方法 (一)	(34)
第一节 次数分布	(34)
第二节 集中量数	(40)
第三节 差异量数	(46)
第四章 考试分数的统计分析方法 (二)	(57)
第一节 正态分布	(57)
第二节 相关	(64)
第五章 试题的编制技术	(75)
第一节 客观性试题的编制技术	(75)
第二节 论述题的编制技术	(91)
第六章 试题分类	(98)
第一节 试题分类的依据	(98)
第二节 试题分类	(99)
第七章 题目分析	(117)
第一节 题目的难度	(117)

第二节	题目的区分度	(123)
第八章	施测过程	(129)
第一节	主试的误差因素	(129)
第二节	被试的误差因素	(132)
第九章	评分方法	(139)
第一节	客观性试题的评分	(139)
第二节	主观性试题的评分	(143)
第十章	考试分数的合成与解释	(151)
第一节	分数的合成	(151)
第二节	常模参照考试分数的解释	(159)
第三节	目标参照考试分数的解释	(178)
第四节	解释与应用考试成绩应注意的几个问题	(182)
第十一章	考试的信度	(185)
第一节	考试信度的概念	(185)
第二节	估计信度的方法	(187)
第三节	影响信度的因素	(198)
第十二章	考试的效度	(202)
第一节	考试效度的概念	(202)
第二节	效度的种类与计算方法	(204)
附 表		(215)
1	正态分布表	(215)
2	范氏项目分析表	(219)
3	检验相关系数 $\rho = 0$ 的临界值(r_{α})表	(247)
4	等级相关系数(r_{ρ})的临界值	(248)
主要参考书目		(249)
后 记		(250)

第一章 教育测量的历史简介

学业成就的考查与评定，是教育过程中的一个重要环节，教育测量以学业成就的考评作为自己的研究对象已有悠久的历史。要使学业成就的考评科学化，更好地发挥其在教育中的重要作用，有必要考察教育测量的历史，分析其现状与未来。

第一节 教育测量的发展简史

教育测量，是人们对教育过程中某些变量进行测定。它是随着心理科学和教育统计学的发展以及整个教育科学的发展、需要而逐渐发展起来的。

自产生学校以来，就有了教育测量活动。有关学者考察结果证明，最早的学生成绩考核办法产生于中国。据史书记载，在我国西周时期，设在国都的大学，就有了“比年入学，中年考校”的制度。到了唐代的科举制度，已有成套的考核办法。中国这种通过考试选拔人材的思想、方法也影响到国外。可以说，我们的先人在教育测量方面积累了丰富的经验，也取得了很大的成就。但是教育测量作为一门独立的学科，还是在20世纪以后，它随着西方心理测验的兴起和教育测验运动的进展而发展起来的。

一、西方心理测验的产生与发展促进了教育测验的发展

心理测验产生于社会需要。19世纪末20世纪初，由于当时西方资本主义大生产的迫切需要，急需大量的劳动力，要对雇工进

行能力高低的鉴定。同时，由于工业生产的发展，社会分工日益精细，需要大量的熟练工人及各种专门人才，这就要对工人进行专业训练或职业辅导等，这些需要是促进测验产生和发展的重要因素。

促进测验产生和发展的另一个重要因素是个别差异的研究。实验心理学家在研究中发现，同一刺激各人的反应是不同的，要研究这种个别差异，就必须有测量工具，这便引起测验运动。

在心理测验方面，最早尝试者是英国生物学家和心理学家高尔顿(F·Galton)。他认为人的能力由遗传而来，因此有必要测量那些有否亲缘关系的人们的特性，来确定他们的相似程度，并在1884年国际博览会上设立一个人类测量实验室，以测量人的身体素质及各种机能。他创造应用一些简单的量表，用统计方法对个别差异资料进行分析。高尔顿是这个时期心理测验的代表。

到了19世纪90年代，美国心理学家卡特尔(J·M·Cattell)，将实验心理学和刚兴起的测验运动结合起来，在自己的实验室内编制测验50多种。1890年发表《心理测验与测量》论文，总结了实验并首次提出“心理测验”这个术语。

20世纪初，心理测验的理论开始和实际相结合，这一结合促进了心理测验进一步发展。法国的比纳(A·Binet)对此做出了突出贡献。比纳长期热衷于智力测验，研究个别差异。为了探讨测量的理论与方法，他做过许多尝试，并决心将测验的理论研究变为实际应用。1904年法国教育部组织一个由医学家、教育家、科学家组成的专门委员会，研究公立学校中低能班的管理法，比纳是其中一员，他极力主张用测验法去辨别心理缺陷的儿童。经细心研究，终于在次年与其助手西蒙(T·Simon)在《心理学报》上发表了题为《诊断异常儿童智力的新方法》一文，介绍了他们编制成功的第一个智力量表。

这个量表虽然是极简单的，但量表中包含30多个测验项目，

可用来测量各种各样的能力，这是以前心理学者的实验所没有的。此量表经1908、1911年两次修订，使之进一步完善。比纳对心理测验的贡献是巨大的。目前世界上的智力测验众多，但其基本原理和主要方法都是建立在比纳的基础之上的，他被称为智力测验的鼻祖。

比纳智力量表对心理学和教育学的影响既深且广。本世纪初，在心理学方面智力测验已成为心理研究、心理诊断不可缺少的工具；在教育方面，智力的测量引起教育家对个别以及适应个别的种种设施的注意，并编制出简单的教育测验。因此，心理测验的发展对教育测量的发展起到了促进的作用。

二、教育测量的发展

心理测验的理论与方法，深刻地影响着学校教育。在心理测验蓬勃发展时，美国及西欧传统的学校考试也开始了改革。虽然教育测验和心理测验所测量的内容有所不同，但所依据的原则及方法是相同的，因此教育测验和心理测验的发展必然是相互影响的。

教育测验以美国发展最早，最发达。纵观美国教育测验理论与实践的发展史，可见教育测量历史之一斑。

1. 贺瑞斯·曼及菲奢的研究

美国教育家贺瑞斯·曼(Horace Mann)早在1845年研究口试与笔试的利弊中，提出许多测验的观点。他提出统一的笔试比个别的口试有许多优越性，提倡使用统一的笔试。他认为：笔试比个别的口试公平，儿童接受同样的试题。统一笔试，在相同时间可以比口试多考许多学生，不但经济，而且比口试可靠，因为考试时间越长，题目越多，考试结果越可靠；同时笔试不象口试那样易使考生慌乱而影响考试的可靠性。贺瑞斯·曼的这些思想虽然在当时的教育测验中没有明显的影响，但是他的关于测验的

思想却是十分有价值的。他提出的统一笔试，提高考试可靠度的思想和现在的标准化考试的原则极相似。

1864年，英国教师菲奢在他自己的研究中提出了标准测验的观念。他收集了许多学生成绩样本编成《量表集(Scale Book)》一书，用来作为考查学生作品的标准。他对量表集中的各科学生作品样本都评定一定的分数，当要评定某一学生的作品时，就用其作品和各种样本比较，判其优劣，给相应的分数。虽然各种样本分数是其主观评定的，不够精确，但是这种思想符合后来的编造书法量表和作文量表的思想。

贺瑞斯·曼和菲奢的研究虽没有引起当时教育界的注意，但他们提出上述测验的思想和方法是很宝贵的。

2. 来斯的影响

19世纪末，美国教育界关于学校课程设置引起一场争论。当时的美国学校比较守旧，学校课程设置株守传统的读法、算法、拼法、书法，教学方法注意练习、背诵，极为死板。这些引起了教育界有识之士及家长的不满，他们要求改革，主张学校应增设一些实用学科；而固守传统学科的守旧派当然怕这样会影响课程的学习，为此两种思想展开争论。1894年，来斯(J·M·Rice)提出解决争论的客观方法：用统一的测验去考查教授拼法时间不同的学校的学生，看教授时间的多少，与学习成绩有无关系。他选用50个字作为拼法测验，测验结果表明，教授时间的多少，与成绩的优劣没有关系。8年之中每天用15分钟学习拼法的学生成绩并不比每天用45分钟学习拼法的学生成绩差。这一结果引起教育界的震惊。虽有各种怀疑存在，但终为有思想的教育家所注意，这一事实不但促进了当时的课程改革，而且引起更多的人对教育测验的关注和研究。

虽然来斯并没有编制什么标准测验，但是他用客观的考试方法来衡量教育效果、研究教育问题的思想方法，对测验运动的影

响是深远的，曾被称为教育测验的创始者。

3. 桑代克的贡献

桑代克 (E·L·Thorndike) 作为教育测量理论的主要创始人，于1904年发表了《心理与社会测量》一书，介绍了统计方法以及编造测验的基本原理。这是世界上第一本有关社会科学方面的测量学专著。它开始了现代教育测量的历史。至今仍是十分有价值的测量学、统计学书籍。

1908年在桑代克指导下，斯东 (C·W·Stone) 编造的算术推理测验，是一种最早的标准测验，而1909年桑代克编造的书法量表是教育测量历史上第一个用科学方法制成的教育测量的工具。后来他又编造了作文量表、图画量表和书法量表等。桑代克在教育测量运动中的贡献是卓越的。爱里斯 (L·P·Ayres) 曾说过：“我们既称来斯为教育测验的创始者，则对桑代克应称为教育测验运动的鼻祖。”在桑代克的指导及其测验理论影响下，陆续有各种量表及标准测验问世。1907年在美国开始的大规模教育调查中，应用了教育测验这一工具，对教育测验的发展和应用的推广都有很大的影响。

4. 测验汇集与中学各科测验的编造

1918年测验汇集的出现，标志着教育测验运动的蓬勃发展。1918年以前编造的教育测验，都是单独学科的测验，如书法测验、拼写测验、算术测验等。而测验汇集则是几种学科甚至包括智力测验等组成的测验。这种测验在选择测验及施测等方面都比单科测验便利，可用于学生测验，尤其在普通的教育调查中更为有用，宾特纳1920年发表的教育调查测验是最早出现的测验汇集。

随着测验逐步推广，教育测验由初期局限于小学课程的各个学科，发展推广到中等教育和高等教育阶段。各科测验已成为学校教学和行政上不可缺少的工具。据统计，到1928年，美国的标准测验总数约在1300种左右，其中绝大多数是教育测验，可见教

育测验发展之迅速。

概括美国教育测量的历史，大约经历了三个历史时期。从19世纪末20世纪初到30年代初，这是“客观测验”、“标准测验”兴起和繁荣时期。在这一时期中，专家们编写了大量的教育和心理测验，促进了教育测量理论与方法的迅速发展。第二个时期是30—40年代教育测验运动在理论和技术上进入反省和改进的时期。因为教育测量所借助的客观测验毕竟不能测得人的发展情况的全部，那些简单的测量目标，可以通过测验用数量表示出来，但是象社会态度、技术、能力、兴趣等都难以用当时的测验方法测定，因而当时的测验受到怀疑和指责。美国以艾钦为首的一批学者，潜心研究8年，于40年代设计了一套教育评价的方法，使测量与评价紧密联系教育目标，全面考察教育与学习的各个方面，以改进当时教育测验理论与方法之不足，获得了成功。第三个时期是自50年代以来，在新的多因素理论的推动下，教育测验进入向多元化方向的稳步发展时期。特别是第二次世界大战以后，各国对教育测验更加重视，尤其是美国，在大规模范围内大力提倡教育测验。这一时期兴起了目标参照测验，它的兴起是教育测验深入发展的重要标志。

美国有庞大的考试机构，如教育考试服务处（Educational Testing Service），主要举办高等学校入学考试（SAT）、研究生资格考试（GRE）和“托福”（TOFEI）考试（即外国学生到美国学习进行的英语考试）。除此之外，还有各种专业的考试委员会，负责全国本专业的发放证书等资格考试。

到了70年代末80年代初，除了美国以外的其他各国的教育测验也有新发展。比如苏联，也从50年代的片面否定测验转到重视教育测量，为了确保高等学校教育质量，在高等和中等专业教育部门设立了专门的高等学校国家检查局，并详细规定了学生考试评分标准；教育工作者开始注意研究西方标准化测验。

随着社会对考试问题的重视程度的日益提高，不少国家也相继建立了专门的考试机构，而且由考试竞争带来的非理想性效果，已引起世界上许多国家的重视。1980年3月，在英国、日本、斯里兰卡等国的教育学家、心理学家的倡导下，召开了国际性的科伦坡考试问题讨论会，讨论了“以考试为目标的学校教育所引起的非认知性的后果”等。总之，教育测量的发展，无论在理论上还是在实际应用上，尽管在世界的范围内还不够平衡，而且存在不少实际问题，但是人们总是在不断地改革和完善它，使它成了教育科学研究及教育实际工作不可缺少的工具。

三、教育测量的发展趋势

教育测量的发展，离不开时代的要求。在强调发展能力的当今世界，教育必然要求把对学生知识水平的测量同对学生能力水平的测量结合起来。就是说，通过测量，不仅要了解学生的知识，而且要了解学生的各种实际能力；不仅要明确学生之间的个别差异，而且要清楚教学目标的完成情况；不仅要检查学生对知识的掌握，而且要对整个教学效果进行评价，等等。

和这一要求相适应，教育测量发展的动向表现出三个明显特点。一是教育测量向着内容的整体化方向发展，其重要标志是论文式测验和行为评定重新受到重视；二是从强调常模参照性测验到同时提倡目标参照性测验；三是重视高层次学习结果的测量，并以多层次来评量学生的能力与成就。

教育测验实际活动的每一发展，都是在吸取前一阶段的全部发展成就并在克服其不足的基础上进行的。西方的客观测验本来是从批判旧式的论文测验而产生的，但是客观测验的最大限制是学生不能自由发表自己的思想见解，而这一点又恰是论文式测验的长处。论文式试题能综合反映学生的各层次的能力水平，其作用是客观试题所不能代替的。在强调发展能力的今天，要求测验

不但要考查知识，而且要考查能力，不仅要编制种种只有单一答案的求同式思维题，而且要编制有多种答案的求异式思维题。客观测验很难全面完成这些任务，论文式测验和行为评定恰能为学生创造性地发表自己的见解提供机会，有利于考查学生的能力、态度、品德等高层次智能，因此论文式测验和行为评定重新受到重视。当然对论文试题评分不客观的弊端应采取科学的评分程序加以控制。

目标参照测验是在苏联第一颗人造地球卫星上天后，美国教育界受到震动，美国教育专家们认识到教育测验必须紧密联系教育目标，否则不利提高教学质量，同时加上第二次世界大战后在新的智力结构理论及新的教学论、课程论体系的刺激等情况下应运而生的一种新的测验。在这之前的测验一般都是常模参照测验，常模反映普通水平，是以相对的目标标准来确定学生间的个别差异。目标参照测验是以绝对的目标（预定的学习结果）为标准来测定学生对目标的达到程度，它强调测验和教学过程的有机统一，以提高教学效果及考查学生具体的行为目标。西方现在一方面仍大量使用各种常模参照性测验，同时在教学体制的许多改革中又都广泛应用目标参照测验，尤其在需要严格贯彻循序渐进、依靠自学或自动调节进度的各种教学体制中，以及在计算机辅助、计算机管理的教学体制中，目标参照测验的应用效果更为显著。可以说，在未来的教育测验中，目标参照测验将会有较大发展。

在教学中，教师期待的学生的学习结果有低层次的，也有高层次的。但以往的测验，大多以知识的测量为主，或较多测量的是较低层次的学习结果。在当今时代，这种只能测量低层次学习结果的测验已不能完全适应社会的要求。人们总想知道一个人的能力如何，是否有创造力，因为高层次学习结果更明确地反映着一个人的能力。因此，教育测量不仅要重视学生知识的获得，而

且要强调学生理解、应用、分析、综合、评价等能力的获得与评量。另外，由于因素分析等多元统计分析方法的应用，以及近代心理学家重视个别差异和多重能力理论的影响，教育测量则开始重视用多种测量技术和统计方法，去评定一个人的各种层次的能力，以便量材用人。

分析西方教育测量发展的历史、现状和未将，是为进一步研究推进我国教育测验的发展，为提高我国教育科学研究水平、提高教学质量服务，为我国的四个现代化多出人才、早出人才服务。

第二节 中国教育测量的曲折道路及现状

西方教育测验之兴起，其影响也波及我国。1918年俞子夷编造的小学国文毛笔书法量表，是我国最早的标准化测验。1920年北京高等师范学校和南京高等师范学校建立了我国最早的心理实验室。廖世承与陈鹤琴先生在南京高师开设测验课程，并用心理测验对投考该校的学生进行考查，随后出版了他们俩人合著的《智力测验法》，这是我国最早介绍心理测验的书籍。

1922年，应中华教育改进社的聘请，美国教育测量专家麦柯尔（W·A·Mccall）来华讲学。在他的指导下，北京师范大学、北京大学、燕京大学、北京女子高等师范、东南大学等校教授和学生开始编造测验，共编出测验40多种。麦柯尔曾评价这些测验的水平“至少都与美国的标准相等，有许多种甚至比美国的为优”。

1923年在教育改进社的主持下，进行全国小学教育调查，包括22个城市和11个乡镇，测验了9.2万个儿童，这次调查引起了教育界对测验的注意。1924年，陆志苇先生发表了《订正比纳西蒙智力测验说明书》。1931年中国测验学会成立，1932年《测

验》杂志创刊。30年代，教育测量作为一门学科在我国已经受到一定的重视。我国有不少学者从事教育与心理测量方面的研究和教学，曾出版了陈选善先生著《教育测验》、王书林主编《心理与教育测量》及廖世承、陈鹤琴合编的《测验概要》、艾伟先生的《小学儿童能力测量》、孙邦正的《心理与教育测验》等，约20多种版书，并编制或修订数十种各类测验及各种量表。当时教育测量作为一门必修课已在师范院校开设，教育测量得到一定发展。

50年代，在苏联片面否定测验观点的影响下，我国教育测量课停开了，研究工作中止了，使我国教育测验工作远远落后于先进国家。近30多年来，师范院校的毕业生缺乏教育测量方面知识的学习和训练，这对教育科学研究的发展及教育质量的提高都受到不同程度的影响。

粉碎“四人帮”以后，我国从闭关锁国中解放出来，与西方广泛的文化交流，促进了我国教育测量学的新生。1979年我国学者参考国外资料，编制了少年儿童学习能力测验。1980年教育和心理学工作者开始修订中国的比纳量表和韦氏量表。近几年来，在运用教育测量理论分析高考试卷、试题，研究改革考试方法，改革学生成绩管理等方面都取得了很大成果。全国大规模的选拔考试在题型上，多数采取了一部分或大部分客观试题；我国高考在外语等学科上都进行着不同程度的标准化考试试点，其他学科也都较大部分地采取客观试题，逐渐扩大标准化考试的范围和提高标准化的水平。教育测量知识逐渐为更多的人所了解、掌握和应用。

目前，教育测量这门课程已陆续在师范院校和教育学院开设，很多人在进行教育测量的理论和方法的研究。1985年4月，我国教育部主持召开了标准化测试及选拔考试国际讨论会，参加会的有来自美国、英国、日本、澳大利亚、新西兰、瑞典等国考试机构的负责人或测量专家，各国代表就标准化测试、高等教育招生考试制度，使用考试及改进课堂教学等问题进行了学术研讨。