辽宁省高等教育自学考试指定教材

教育评价与测量

辽宁省高等教育自学考试委员会　组编

郭庆科/编 著

辽宁师范大学出版社

辽宁省高等教育自学考试指定教材

教育评价与测量

辽宁省高等教育自学考试委员会组编
郭庆科　　编　著

辽宁师范大学出版社
大连

图书在版编目(CIP)数据

教育评价与测量/郭庆科编著．—大连：辽宁师范大学出版社，2004.5
ISBN 7-81042-974-4

Ⅰ.教… Ⅱ.郭… Ⅲ.教育评估-高等教育-自学考试-教材 Ⅳ.G449.7

中国版本图书馆 CIP 数据核字(2004)第 024403 号

责任编辑：杨人格
责任校对：陈　冲
封面设计：张　环
版式设计：孟　冀

出 版 者：辽宁师范大学出版社
地　　址：大连市黄河路 850 号
邮　　编：116029
营销电话：(0411)84206854　84215261　84259913(教材)
印 刷 者：金城印刷厂
发 行 者：新华书店

幅面尺寸：140mm×203mm
印　　张：9.5
字　　数：280 千字

出版时间：2004 年 6 月第 1 版
印刷时间：2004 年 6 月第 1 次印刷
定　　价：14.50 元

本书编委会

内容简介

教育测量学和教育评价学是研究如何客观评量学生发展和教育教学效果的基础性学科，前者主要涉及学生评价，后者则涉及对整个教育工作的评价。基层教育工作者掌握这两门学科的知识是非常必要的。由于这两门学科都需要教育统计学的知识，掌握起来可能有一定难度。考虑到基础教育工作者的实际情况，本书尽量避开了与统计学有关的内容。在测量部分，讲述了信度、效度的基础知识以及各类教育心理测验的原理，项目分析和测验编制两部分的可操作性强，可帮助研究者发展出高质量的测验。教育评价部分介绍了有关领域的原理、思想和方法，有助于教育工作者形成系统的评价观，从素质教育的角度理解学生评价，并能指导教学和管理工作。本书适合广大中小学教师、自学考试考生、教育行政管理人员、师范院校师生学习和阅读。

目　录

第一章　教育测量概述

在教育研究中，人们有时需要对学生的心理特点加以区分，如智力、能力、品德以及兴趣爱好等。这就要对这些心理特点进行量化，专门研究如何将人的心理特点进行量化的学科称教育或心理测量学。而教育测量学则是在心理测量学的基础上产生和发展起来的，二者一脉相承，从研究范围上讲也是基本相同的，因此对这两个概念本书不做区分。本章所介绍的是心理与教育测量的一些基本知识。

第一节　教育与心理测量的历史

一、中国古代的心理测验

中国古代的学者早就注意到人与人之间存在能力和个性方面的差异，并尝试使用有效的方法将不同心理品质的人加以区分。因此很早就产生了心理测量的思想，并开始了心理测量的实践。中国古代的医书《黄帝内经》中曾把人分成太阴、太阳、少阴、少阳、阴阳和平五种，这一分类可看做心理测验的前奏。孔子按照能力和知识将人分成上、中、下三等，并说“惟上智与下愚不移”，“中人以上，可以语上也，中人以下，不可以语上也”。这种划分包含了等级评定法的基本思想。

孟子曾对齐宣王说:“权,然后知轻重;度,然后知长短,物皆然,心尤甚,王请度之。”就包含了将人的能力、品格等心理特性加以量化的思想,这与心理测量学家们的信条“凡物之存在必有其数量”(桑代克,1918)、“凡有其数量的事物都可以测量”(麦克尔,1939)表达了同样的思想。

三国时刘劭著有《人物志》一书,将人分成圣贤、豪杰、傲荡、拘栗四种:“心小志大者圣贤之人也,心大志大者豪杰之隽也,心小志小者拘栗之人也”。包含了等级评定的思想。1937 年施罗克(Shryock,J. K.)曾将《人物志》翻译到美国,书名为《人类能力之研究》。刘劭认为“众人之察不能尽备”,只能“观其感变以审常度”。用今天的话来讲,就是说,人的所有行为是不能完全被观察到的,只能通过观察其有代表性的行为来预测其整个的人格。这与现代心理测验中对“行为样组”的测量是一致的。

在测验的实践方面,中国也是远远早于西方的。公元前 11 世纪,西周采用“试射”的办法选拔文官,不仅测查射箭的命中率,还要观察试射者的行动是否合乎礼义,举止是否合乎乐律。这可以视为人类历史上最早的测验。到了汉代,文官考试开始使用笔试,其内容包括法律、军事、农业、税收和地理。至公元 606 年,中国开始了科举取士的制度,用一系列严格的考试逐级选拔人才,这比 18 世纪后欧洲运用笔试录取文官的制度要早一千多年。任何一个心理测量年表都要以中国的科举考试为开端。

二、西方对教育与心理测量的早期探索

1. 冯特

冯特是实验心理学的创始人,他首次将实验方法引入心理学研究。冯特在实验中证实了个别差异的存在,并发明了测量思维敏捷性等方面个别差异的工具。早期的实验心理学和心理物理学为心理测验的发展奠定了基础。

2. 高尔顿

第一个直接推动心理测验产生的人是英国生物学家高尔顿。高尔顿开创了个别差异心理学的研究，并引入了定量研究方法。1869年，高尔顿出版《遗传的天才》一书，提出了能力遗传决定论。在1883年出版的《人类能力及其发展的研究》一书中，高尔顿首次提出了“心理测量”的概念，列举了大量有关人类个别差异的资料，并多次重申个别差异可以测量的观念。

高尔顿认为所有智力行为都与人的内在感觉能力有关，因此运用心理物理学方法，通过测量人的运动反应时和感觉辨别能力就能测量到人的智力。在这一思想指导下，他进行了大量的人类测量学研究。1884年，高尔顿在伦敦国际博览会上建立一个“人类测量实验室”，此后的六年中，共搜集了9337人的生理、感知觉方面的个人资料，这些资料包括身高、体重、头颅大小、手背跨度、中指和前臂长度、握力、呼吸力、拉力、视力、听力以及视觉和听觉反应时等。高尔顿曾试图运用他的学生皮尔逊发展出的相关程序计算这些测量与各成就指标间的相关，但结果并没有证实他的设想。

3. 卡特尔

卡特尔(Cattell, James M.)是推动心理测验产生的另一个重要人物，1890年，卡特尔发表了一篇有重要意义的论文《心理测验与测量》，第一次使用了“心理测验”这一概念，他认为，“心理学者不立根基于实验与测量上，决不能有自然科学的精确”，“如果我们规定一个一律的手续使在异时、异地得出的结果可以比较、综合，则测验的科学和实用的价值都可以增加”。这些观点可作为测验编制的重要指导思想。

卡特尔发展了高尔顿的方法，提出10项心理测验，包括如下一些内容：(1)握力测量，用握力计测肌肉力量；(2)动作速度测量；(3)触觉两点阈测量；(4)引起痛觉的最低点测量；(5)对辨别重量最小差别的能力的测量；(6)对声音反应时的测量；(7)说出混在一起的四种

颜色的名称;(8)将50厘米的线平分;(9)对10秒时间判断的测量;(10)复述听过一次的字母。

卡特尔也认为生理能量与心理能量密不可分,因此对智力的测量也主要是对感觉辨别力、反应时、动作过程等的测量。卡特尔的学生维斯勒(Clerk Wissler,1901)进行了一项开创性的研究工作,他搜集了哥伦比亚大学和伯纳德大学300名学生的智力测验分数与学业成绩,试图以智力测验分数预测他们的学习成绩,这可视为对智力测验进行的最早的效度研究。但维斯勒的研究结果令人失望,因为二者的相关很低。例如,学业成绩与数字记忆测验的相关为0.16,与握力测验相关为-0.08,与颜色命名相关为0.02,与反应时相关为-0.02。与数字记忆的相关虽达到了显著水平,其对学业成绩的预测能力却是微乎其微。维斯勒的工作证明用简单心理过程测量智力的方法是错误的,改变了智力测验发展的方向,启发了以后比奈—西蒙智力测验的产生。

冯特、高尔顿和卡特尔对心理测量的研究主要包括人的基本心理能力,如感觉辨别力、反应时、动作过程,虽然没有得到预期的结果,但却为心理测量的产生打下了基础,树立了"人的心理属性可以测量"的理念。

三、心理测验在西方的产生

法国心理学家比奈是第一个智力测验的创立者,他因此被称为心理测验的鼻祖。

比奈(1857—1911)生于法国尼斯,早年学过法律、医学,以后致力于心理学的研究。从青年时期就对人们行为的个别差异及不同年龄阶段行为的发展变化产生兴趣。

比奈将智力定义为:选择并保持一个明确方向的倾向,为达到预定目标而做出适应的能力及自我评价的能力。比奈认为,智力的基本技巧包括预测一个行为或事件的结果,具体体现在以下方面:检查

自己行动的结果；对自己当前行为的监督；现实性测量；各种协调、控制性的行为；学习及解决问题。比奈开创性地发现智力与复杂的任务有关，尽管很难明确说明这些任务测量了什么，但被试在这些任务上的表现确能预测其智力行为。这些为他编制第一个智力测验打下基础。

1904 年，法国教育部委派教育家、医学家组成一个委员会，研究公立学校低能儿童的教育方法，比奈也是委员之一，他主张用心理测验的方法鉴别智力落后的儿童。比奈与其助手西蒙合作，编制成了历史上第一个智力测验，并在 1905 年的《心理学年报》上发表文章《诊断异常儿童智力的新方法》一文介绍此量表，史称 1905 年量表。此后比奈又在 1908 年和 1911 年对此量表进行修订，产生 1908 年量表和 1911 年量表。

1905 年量表共包括 30 个由易到难的问题。特点是测验项目种类繁多，能测量智力多方面的表现。测验包括一些测量感知觉的题目，但语言部分占了很大比例，主要涉及推理、判断、理解等高级心理过程。测验适用于 3～14 岁的低能儿童，也可用来对正常儿童做某种程度的区分，以通过的题目数来衡量智力的高低。

1908 年，比奈和西蒙根据 1905 年量表的使用情况和问题对测验进行了第一次修订。1905 年量表主要是用来鉴别智力落后儿童的，测题偏易，不适用于正常儿童，1908 年量表删去了这些不适当的题目，增加了一些较难的新题目，使测题总数达到 59 个，测题按年龄分组。组别为 3～13 岁，共 11 个年龄组，每组的测题都保证该年龄组儿童有 80%～90%的人能通过，使用智力年龄表示儿童的智力发展水平。该测验使用了现代智力测验的一些重要概念，如效度、常模等。

1911 年，比奈又对 1908 年量表进行了修订，这次修订版的变化不大，主要是删改了一些项目，并重新划分了年龄组，包括：删掉了 9 个旧项目，增加了 4 个新项目，使测题总数达到 54 个，增加了 15 岁

组和成人组，除4岁组仅有4个项目外，其余组项目数均为5个。不幸的是，比奈于当年去世，年仅54岁。他的去世是心理测验运动的重大损失。

四、教育与心理测量的发展

1. 智力测验的发展

比奈—西蒙量表问世后，立即引起世界范围内的广泛关注，并被翻译成多种文字。1916年，美国斯坦福大学的推孟将比奈—西蒙量表进行了修订，形成著名的斯坦福—比奈智力量表。推孟对原量表进行了较大的改动，保留了比奈—西蒙量表的51道题目，自己又编制了39道，使测验总长度变为90题，适应的年龄组为3～14岁组儿童，另加普通成人组和优秀成人组，并首次使用了智商（即IQ）的概念。斯坦福—比奈量表被广泛地应用于教育和临床背景，是最有影响的智力测验之一。此后它分别在1937年、1960年、1972年和1986年四次被修订。

但斯坦福—比奈量表也存在一些缺点，突出表现在离差智商的概念上。美国心理学家韦克斯勒看到了这一缺点，并从1934年起着手编制智力量表。1939年，韦克斯勒发表了第一个用于测量16～60岁成人智力的韦克斯勒—贝尔韦量表（W—BI），这个量表以后发展成为韦克斯勒成人智力量表（WAIS，1955）。韦克斯勒在W—BI的基础上于1949年发表第一个儿童智力量表（WISC），测量对象为6～16岁的儿童。1967年出版韦氏幼儿智力量表（WPPSI），1974年发表WISC修订本WISC—R，1981年发表WAIS修订本WAIS—R，1991年，WISC第三版（WISC—Ⅲ）出版。韦克斯勒改进了智商的概念，提出了离差智商，这是一个巨大的进步，他还扩展了智商的概念。从言语、操作及总体三方面衡量智力的高低。

1938年，英国心理学家瑞文（Raven，J. C.）出版瑞文标准推理测验，这是一个著名的非文字智力测验。既可用于个别测验，也可进行

团体测验。1947年瑞文又出版彩色推理测验和高级推理测验。

2. 标准化教育测验的发展

在教育测验方面，虽然在公元606年隋朝就开始了科举考试，但欧洲学校一直没有笔试的传统，直至1702年英国的剑桥大学才开始使用笔试录取学生。1845年，在美国著名教育家贺拉斯·曼(Horace Mann)的倡导下，麻省波士顿市教育委员会开始用笔试的办法考查该市所属学校的毕业生。这以前的教育测验都是以主观性试题为主，评分缺乏客观性。

客观的标准化教育测验的最初尝试者是英国格林尼治医院的教师费舍(George Fisher)。他深感对学生成绩的评定缺乏统一的标准，因此试图建立评定学生各科成绩的统一量尺。他搜集很多学生的书法、拼写、算术、语法、作文、历史、自然、图画、法文等科的成绩，汇编成《量表集》一书，作为衡量学生各科成绩的标准，费舍对收入量表集的学生的成绩评定了等级，其他学生的各科成绩与这些标准相比较，就能判断其优劣等级。用今天的话说，收入量表集的学生就相当于常模样本组，而量表集就相当于常模量表。费舍的工作是教育测验历史上的重大进步。但是，《量表集》中学生的各科测验分数及其等级是费舍本人主观决定的，缺乏客观的依据。

1894年，莱斯(Rice, J. M)博士主张用统一的测验去评定不同学校学生的成绩。1895至1905年间，他先后编制了算术、拼字、言语等测验，测试了数万名学生，引起了教育界对教育测验的关注，莱斯因为在测验上的突出贡献而被称为教育测验的开创者。

1904年，美国心理学家桑代克(Thorndike, E. L.)出版《心理与社会测量导论》一书，这是关于测验理论的第一部著作，该书系统地介绍了统计方法及测验编制的基本原理，为测验的发展奠定了基础，并提出"事物的存在必有其数量"的著名观点。1908年桑代克发表书法量表，这是第一个书法量表，在教育测验历史上有重要意义。1909年，他又根据统计学"等距原理"为测验量表确定了单位。此外

他还编写了拼写量表、作文量表、图画量表等。桑代克在测验理论及实践研究中的贡献使他成为测验运动的中心人物，被称为教育测量的鼻祖。

在桑代克的推动下，一些标准教育测验相继出现。1908 年，桑代克的学生斯通(Stone，C. W.)编制了算术测验，1911 年，希莱格斯(Hillegas，M. B.)编制了作文量表，白根汉(Buckinghan，B. R.)编制了拼写量表，柯蒂斯(Courtis，S. A.)编制了算术标准化测验，纽约、波士顿等大城市设立测验研究部，组织测验专家进行大规模的教育测验。这一时期的标准化测验得到了极大发展，特点是注重测验评分的客观化，并用常模对测验进行解释。

经过桑代克等人的早期探索，标准化教育测验逐渐走向成熟。早期的教育测验都是单科测验，如拼法、算术测验等，为适应调查的需要，于是出现了综合测验。综合测验能同时考查学生在几个主要学科上的学业成绩。1920 年，宾特纳(Pintner，R.)发表的《测验汇编》，孟禄(Monroe，W. S.)和白根汉合编的综合测验是最早的综合测验。1923 年，凯利(Kelley，T. L.)、鲁驰(Ruch，G. M.)和推孟编制了斯坦福成就测验(Standford Achievement Test)，已具备现代测验的性质，在当时非常流行。

1917 年以前，教育测验主要限于小学科目，1917 年后，用于中等以上学校的教育测验得到了发展。心理测验的理论和方法对教育测验产生了巨大的推动作用，越来越多训练有素的心理测量学家和教育测量学家投入到教育测验的开发和应用中，使教育测验得到深入发展。由此，教育测验进入了兴盛期。30 年代后，教育测验不仅用于评定学生的学业成绩，而且还广泛应用于会计、律师、医生、商业、交通、军队等职业领域，进行人才选拔、安置、培训等。

3. 能力倾向测验的产生与发展

能力倾向又称为特殊能力。对特殊能力测验的研究始于斯皮尔曼对一般能力和特殊能力的划分。一些心理学家(尤其是坚持因素

分析的心理学家)则认为智力并不是单一的能力,而是由彼此相互独立的特殊能力构成的。因此,不少心理学家将兴趣转向了能力倾向测验的编制。

1915 年,西肖尔(Seashore)编制的音乐能力测验是最早的特殊能力测验。此后又有罗杰斯(Rogers)编制的教学能力测验等。30 年代得到广泛应用的因素分析方法推动了特殊能力测验的发展。人们发现不同的职业需要不同的特殊能力,特殊能力测验在职业选拔、职业咨询方面得到了广泛的应用。

4. 人格测验的产生与发展

人格测验的产生要迟于智力测验,但对人格测验的探索却可以追溯到古代。中国古代很早就开始使用观察法、等级评定法评定人格,但古代对人格的评估往往与知识、能力相混同,且过于侧重品德的一面,因而难以达成科学的认识。

最早尝试用科学方法研究人格的是高尔顿。他认为人格的测量也应该和智力测量一样,应遵循客观的方法,他还尝试编制了评定品格的量表。

克雷丕林从心理诊断的角度对人格测验做出重要探索。1892 年,他使用自由联想法研究精神病人,得出不少有意义的结论。此后,自由联想法一直是一种重要的临床诊断方法。而以后人格测验的产生,也主要是出于对病理诊断的需要。人格测验先是被应用于临床,而后才应用于评量正常人的人格。由于克雷丕林的重大贡献,他被认为是人格测验的先驱。

伍德沃斯(Woodworth,1917)编制了第一个现代意义上的人格问卷,即伍德沃斯个人资料调查表,用来鉴别不能从事军事工作的精神病患者。问卷包括 100 多个关于神经病病状的问题,让被试根据自己的情况回答,这被称为自陈问卷(Self-report inventory)。战后伍德沃斯的测验被多次修订,被用于测量学生和正常人的学习适应和社会适应,此后人格测验被广泛应用。

自陈量表被认为是客观化和标准化的人格测验，在人格测验中占据主导地位。著名的人格自陈问卷有艾森克的EPQ问卷，卡特尔的16PF问卷，明尼苏达多项人格量表(MMPI)，加利福尼亚心理调查表(CPI)等。

与自陈量表相对的是投射测验，瑞士精神病医生罗夏(Hermann Rorschach)注意到被试对墨迹图的反应可用来区分正常人和精神分裂症患者，且能区分不同人格类型的正常人。1921年，他发表了著名的罗夏墨迹测验。另一个著名的投射测验是由莫瑞(Murray)和摩根(Morgan)于1935年发表的主题统觉测验(TAT)，此外，还有句子完成测验、绘画测验等。投射测验先是在临床诊断方面被广泛采用，此后又被用于测量正常人的人格、动机等。

此后，哈商(Hartshorn，H)和梅(May，M. A.)开创了品德测量的情景测验法，这种方法是通过观察被试在特定情境中的行为以对其品德和人格进行评量。这也是一种研究人格的重要的客观性方法。

五、中国的教育与心理测量

辛亥革命后，中国学者很快吸收了西方先进的测验理论和方法，开始了现代教育测验的探索。1916年，樊炳清首先将比奈—西蒙智力量表介绍到我国。1918年，俞子夷模仿桑代克的书法量表编制了《小学生语文毛笔书法量表》，这是我国最早的标准化教育量表。

1920年廖世承和陈鹤琴在南京高等师范学校首次开设心理测验课程，并对学生进行心理测验，这是科学心理测验的开始。1921年二人合著出版《心理测验法》一书。1922年，费培杰将比奈—西蒙量表译成中文，同年美国著名教育测量专家麦柯尔(McCall，W. A.)应“中华教育改进社”的邀请来华讲学并主持编制测验，使用了他倡导的“TBCF”分制，共编制测验40多种。麦柯尔对当时编制的测验评价很高，认为它们达到了美国的水平，有的测验项目还优于美国。