教育测量与评价

范晓玲　杨志明　主编

中南工业大学出版社

前　言

这是一部专门为小学教育专业自考学员编写的教育测量与评价的教材。本教材不仅可以作为自考学员的自学教材，而且也可以作为广大的教育工作者从事教育教学、教育管理和教育科学研究的工具书，同时还可以作为教育工作者拓宽视野、扩大思路的参考书。

教育测量与评价是现代教育科学与新兴学科不断发展及与其他各门学科大融合、大渗透的结果。其历史源远流长，若从本世纪初强调测验客观性和科学性的教育测量运动算起，到30年代以后教育测量逐渐受到批判而教育评价运动逐渐兴起，再到今天人们对教育测量与评价的不断深入和发展，也已经历了近一个世纪的变革和不断的进步。在近百年的发展中，教育测量与评价的观念深入人心，教育测量与评价的方法与技术被普遍使用，大到一个国家掌握其教育的发展状况与趋势，小到一名教师了解自己的教学效果、调整教学措施，都无不使用到测量与评价的知识与技术。因此，教育测量与评价不仅是现代教育科学发展的使然，而且也是未来教育科学发展的必需。

由于众所周知的历史原因，我国教育测量与评价的研究起步较晚，落后于一些发达国家。但是，随着我国政治经济、科学技术的迅猛发展，国际化、全球化、信息化的呼声越来越高，教育科学与教育事业的发展要适应这一要求，尽快与国际接轨，不仅需要对教育教学工作进行有效的测量与评价，而且还需要在教育实践中不

断总结经验，形成具有中国特色的、在国际上具有影响力的教育测量与评价的理论与方法。而作为一名教师，要在未来社会中成为一名合格的教育工作者，不仅要教好书、育好人，而且还要懂得对教育现象的测评，提高自我决策的水平。

本书的编写一方面着重于基本概念、原理的介绍，另一方面则尽力反映国内外最新的有关教育测量与评价的发展动态。全书十二章，第一章为总述，介绍教育测量与评价的概况，第二章到第六章为教育测量，介绍教育测量的原理与方法，第七章至第十二章为教育评价，介绍教育评价的理论与技术。

本书由范晓玲、杨志明任主编，杨志明负责第一、二、三、四、五、六章的设计并编写了其中的第一、二、三章，范晓玲负责第七、八、九、十、十一、十二章的设计及编写工作，贺革编写了第四章，谭林斐编写了第五章，陈建文、杨志明编写了第六章，陈景才参编了第三章第五、六两节的工作。全书最后由主编统稿。

由于编者水平有限，加之时间仓促，书中仍有不尽如人意的地方，恳请专家、读者不吝赐教。

本书的编写得到了省自考办、湖南师范大学教育科学学院的有关领导和同事们的大力支持和帮助，同时参考和引用了国内外同行的有关材料，谨此表示衷心的感谢！

编　者

1998 年 6 月于岳麓山

目 录

第一章 绪 论

第一节 什么是教育测量与评价

测量与评价是我们在日常生活中经常要遇到的事情。比如，我们去小摊上买苹果，就一定要解决好两个问题，即它有多重（测量）、它值多少钱（价值判断）。同样，许多教育问题的解决也常常要求进行定量分析和价值判断，教育测量与评价就是为这种分析和判断提供理论和方法的一门课程，它是教育测量学与教育评价学的合并。

一、什么是教育测量

测量（Measurement）就是对客观事物的某种属性，依据某种法则赋予某个数值的过程。这里，客观事物的属性指的是测量对象，法则指的是测量工具及其操作规范，数值是测量的结果，三者是一个测量的三个方面，缺一不可。例如，用秤称某人体重达 60 公斤是一个完整的物理测量，其中，“重量”（即法定计量单位中的“质量”）是客观事物（人）的一种属性、“秤及其操作规程”是所遵循的法则，“60”则是所赋予的数值。

一般说来，客观事物的属性，不仅包括事物的物理特性，而且还包括那些不被人的感官直接感觉到的心理属性，只不过前者比

较直观，后者比较抽象。例如，物体的长度、重量等等比较直观，而人的智力、个性等等比较抽象。对于事物的物理属性的测量，我们可以采用直接测量法，而对于人们的心理特性的测量，则只能采用间接测量的办法。

关于测定事物的法则，我们可以从两方面进行理解：其一是测量工具，其二是工具的使用方法或者说操作规程。一个测量是否准确在很大程度上取决于有没有一个好的测量工具，有了好的工具以后则有赖于工具的使用是否恰当。一般来说，良好的测量工具必须满足两个条件·其一是要有一个恰当的零点，其二是要有一个良好的单位系统(刻度系统)。在物理测量中，测量工具的制定相对而言比较容易(如，秤的零点和刻度较易确定)，而人的心理特性的测量工具却不太容易制定(如，人的性格的测量工具就不好定零点和单位体系)。

关于测量的数字，主要是用来描述事物属性多少的标志。它可以有四种不同层次水平的理解 其一是将它理解为名称量表意义上的数字，如学生的学号、考号等。这种数字只是事物的一种标记，并无大小之分，更不可进行加、减、乘、除运算；其二是将它理解为顺序量表意义上的数字，如体育比赛中的第一名、第二名等。这种数字只表明事物间的先后顺序，没有相同的单位，也没有绝对零点，不能进行加、减、乘、除运算，其三是将它理解为等距量表意义上的数字，如温度等，这种量表具有相等的单位，但只有相对零点(如摄氏温度为 0 并不表示没有温度)，所以这种数字只能进行加、减运算；其四是将它理解为比率量表意义上的数字，如物体的高度、质量等。这种量表不仅有相等的单位，而且具有绝对零点(如，重量为零即表示没有重量)，故这种数字可进行加、减、乘、除运算。一般地，物理测量的数字大都是等距量表和比率量表上的数，而心理特性测量的数字则大都是顺序量表上的值。

什么是教育测量(Educatıonal Measurement)呢？从广义上讲，教育测量就是依据一定的法则，对教育活动中的某一现象给予数量化的描述。它涉及的范围很广，凡是需要并能够测量的与教育有关的活动均在研究之列，如教育投入、教育效果的测定等等。从狭义上讲，教育测量指的是依据一定的法则，对学生的学习能力、学业成绩、兴趣爱好、思想品德等一系列属性进行数量上的测定。本书主要指的是狭义上的教育测量。

与物理测量相比，教育测量具有几个明显的特点 第一，教育测量具有间接性。科学发展到今天，我们还无法直接测量人的某些心理特性(能力、个性等)，只能测量人的外显行为。也即是说，我们通常只能是通过一个人对测验题目的反应来推论出他的能力水平。第二，教育测量具有相对性。在对人的行为做比较时，没有绝对的标准(即没有绝对零点)，只有一个连续的行为序列。教育测量只是将某一个个体的能力水平在这种行为序列上找到一个位置而已。某人学得好或不好常常是将他与大多数人进行比较而言的。第三，教育测量具有艰巨性。这种艰巨性主要表现在两个方面：其一是教育现象本身复杂多变，不如物理特性简单稳定。其二是教育测量的工具不易制定，即使是制定好了测量工具，其操作规程也往往比较复杂(如标准化考试就涉及到四大环节)。第四，教育测量的结果(数字)通常不具有加、减、乘、除的特性。例如，某次考试中张三得 30 分、李四得 60 分并不能认为二者的水平相差一倍(顺序量表上的值没有绝对零点，不能做除法)。又如，某次考试中的原 60 分与 50 分以及 95 分与 85 分虽然都相差 10 分，但这两个 10 分的“含金量”是不同的(顺序量表上的值没有相同的单位，不可以做减法运算)。第五，教育测量具有可行性。著名教育测量学家桑代克和麦柯尔在几十年前曾先后提出：“凡客观存在的事物都有其数量”，“凡有数量的东西都可以测量”。教育现象是应当可

测的(姑且将此观点作为公理看待)。事实上,随着科技的进步,人的能力、性格、学业成绩等都已有了较高水平的测量方法。

二、什么是教育评价

评价(Evaluation)是人们按照一定的价值观对人和事物所进行的一种价值判断过程。这一概念包含两个方面的内容 其一是人们的价值观,其二是人和事的实际状况,评价则是将事实情况与价值标准进行比较并作出结论的过程。简单地说,评价就是进行事实判断和价值判断的过程。

那么,什么是教育评价(Educational Evaluation)呢? 教育评价就是评价者按照一定的价值标准,运用科学可行的方法,对某种教育现象进行价值判断,并为教育决策提供信息的过程。这一定义我们可以从以下几个方面来理解 第一,教育评价是一种比较正式的活动过程,它包含着一系列的步骤和方法。第二,教育评价是有明确的目标的,这种目标的具体化就是评价的"价值标准",评价乃是以实际状态与理想目标相比较的过程。第三,教育评价要有理论作指导,要采取多种手段和技术来收集处理信息。第四,教育评价的最终目的和归宿,是用一定的价值观对各种状态进行价值判断,以评定效益、帮助决策。一句话,教育评价就是一个明确目标、测定效果、判断价值和帮助决策的过程。

当然,由于看问题的角度不同,教育评价的概念往往具有不同的表述方式,但主要思想还是接近的。例如,美国教育评价专家泰勒(R W Tyler)认为 "教育评价就是将教育实际表现与理想目标加以比较的过程。"另一专家斯塔弗尔比姆(D L Stufflabeam)则认为 "教育评价是描绘、获取与提供有用的资料,以便在诸种可行的途径中,判定决策途径的过程。"此外,布卢姆(B S Bloom)在《教育评价》一书中又指出 "评价乃是系统收集证据用以确定学习

者实际上是否发生了某些变化，确定学生个体变化的数量或程度。”我国学者林昌华认为，教育评价乃是“以教育为对象，根据一定的目标，采用一切可行的评价技术和方法，对教育现象及其效果进行测定，分析目标实现程度，作出价值判断的过程。”

三、什么是教育测量与评价

简单地说，教育测量与评价就是教育测量与教育评价的合称。具体地说，教育测量与评价就是针对某一教育现象，依据一定的法则给它赋予一定的数值，然后再根据某一价值标准，对该教育现象进行价值判断。其中，教育测量侧重于事实判断，强调的是教育现象的数量方面。教育评价侧重于价值判断，强调的是教育现象的质的方面。一般情况下，测量是评价的基础，评价是测量结果的进一步解释。例如某生数学考试得 80 分，这是一个测量过程。如果进一步考察该生数学学得好还是不好（通常是以团体的平均值为参照），则是一个评价过程。

四、教育测量与评价的内容

教育测量与评价主要包括两大内容，即教育测量学的基本原理和方法以及教育评价学的基本原理和方法。具体地说，它包括如下一些内容

①教育测评的基本问题 它包括教育测评的含义、发展历史与现状、目的、类型与功能等方面。

②教育测评的理论问题 它包括教育测量的理论体系（信度、效度等）、教育评价的基本模式、教育评价方案的结构等。

③教育测评的方法问题 它包括学生学业成绩的考试方法、评价目标的确立方法（评价指标的确立方法）、评价指标权重的确立方法等。

④教育测评在实际工作中的应用问题 它包括学生评价、教师评价、课程评价、学业评价等等。

第二节 教育测量与评价的历史渊源

一、教育测量与评价活动起源于中国

一般认为,比较正规而系统的教育测量与评价活动起源于中国。早在2500多年前,孔子就曾说过“中人以上可以语上也。中人以下不可以语上也。”比他稍晚的孟子也曾提到.“权,然后知轻重;度,然后知短长。物皆然,心为甚。”又据《学记》记载“古之教者,家有塾,党有庠,术有序,国有学。比年入学,中年考校。一年视离经辨志,三年视敬业乐群,五年视博习亲师,七年视论学取友,谓之小成,九年知类通达,强立而不返,谓之大成。”这段话的大意是 古代的教育制度,每20户设一个私塾,500户的县设一个学堂,12500户的行政区设学校,国都设大学。大学每年都招生,每隔一年考查一次。第一年考查学生分析课文的能力和志趣,第三年考查学生的专业思想是否巩固,同学之间能否相亲相助,第五年考查学生的知识是否广博,对教师是否敬爱,第七年考查学生研究学问的本领和识别朋友的能力,合格的就叫“小成”,第九年考查学生对学业是否已能触类旁通、其行动是否能坚定不移,若合格,则称他为“大成”。这充分说明我国古代早已有了对学生各种能力与个性特征的教育测评活动。这种活动发展到后来,便形成了一套比较完整的测评制度。例如 汉朝有“察举”制、魏晋南北朝时期有“九品中正制度”、自隋以后则有了更为著名的科举制度(公元606~1906年)。当年科举的帖经、墨义、策论、诗赋等,就是我们当今

的填空、简答、论述和作文等题型的源流。此外，它在考务管理等方面也创立了一整套极有价值的方式方法。

与我国相比，西方的比较正式的教育测评活动却起步较晚，而且受中国科举制度的影响较大。例如，一本介绍过明代科举考试的内容和方法的书(《伟大的中国》，作者·胡安·贡萨雷期·德万多萨(葡萄牙修士，1583 年)就曾被译成多种文字在西方广为流传。从 1570 年到 1870 年间用英文出版的有关明清政治制度的 70 多种书中，都曾把科举制度当作重要的内容加以介绍。法国资产阶级启蒙思想家伏尔泰曾高度评价说："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的考试的。"孙中山先生在考察西方一些发达国家后也曾说过："现在各国的考试制度，差不多都是学英国的，穷流溯源，英国的考试制度，原来还是从我们中国学去的。"不过，由于种种原因，我国近代的教育测量与评价水平反而落后于西方一些发达国家了，我们国家并没能从科举之类的重大教育测评活动中发展出众多的现代教育测量与评价理论。

二、教育测量与评价在国外的早期活动

在西方，教育测量活动也很早就有，但主要是口试，没有出现过像中国科举考试那样的严格正规的笔试。一般认为，1720 年英国剑桥大学施行的笔试测验是西方学校最早的笔试。而美国的笔试测验则开始得更晚(1845 年由 Horace Mann 倡导的在美国波士顿市举行的毕业生笔试测验)。

笔试引入之后，教育测量活动的客观性有所提高，但仍有大量问题，于是在 1864 年出现了世界上第一个教育测量的量表〔英国人乔治·费希尔(George Fisher)的《量表集》〕。该量表对学生的书

法、拼字、算术、语法、作文、历史、自然、图画、法文等学科的不同水平(样品)分别评定了一种分数。其他教师则可根据量表中所列出的有关样品的得分标准去评价学生作业。显然,用制定量表的方法来进行教育测量是具有开创性意义的(尽管当时的量表水平不高,它也未能引起有关当局的足够重视)。

在教育测量发展史上,美国学者莱斯(J Rice)的拼字测验具有一定地位。在19世纪末,美国教育界发生了一场争论,争论的一方主张改革当时的课程,加入实用科学,而另一方则反对这种作法,认为加入新学科后会挤掉学生学习旧有学科的时间,影响原有的教学方法(重练习和背诵的方法)。为了解决这场争论,1894年莱斯选定了50个字设计成测验去测了20所学校的16000名学生的拼字能力,同时对各校每周教授拼法的课时数进行了调查。测验结果表明,8年中每天用45分钟学习拼法的学生,较之每天花15分钟学习拼法的学生成绩没有多大差别。莱斯的这一结论在当时引起了广泛的重视。

在此时期,实验心理学和心理测量学的兴起促进了教育测量的发展。1879年,实验心理学的鼻祖冯特(W Wundt)在德国的莱比锡创立了世界上第一个心理学实验室。他在该室中从事的关于人的个体差异方面的研究被认为是具有划时代意义的(引起人们对测量工具研究的兴趣)。不过,首先指出并进行心理测验的是英国生物学家和心理学家弗朗西斯·高尔顿(Francis Galton),他设计了许多测验进行心理测量,例如,他在1884年的国际博览会上设立了一个人类测量实验室,参观者付三个便士就可测量自己的某些身体素质和视、听觉的敏锐性、肌肉力量、反应功能。博览会闭幕后,该实验室又迁到伦敦的南圣顿博物院,开办了6年之久。同时他还出版了有关书籍(《遗传的天才》,1869年,《人类才能及其发展的研究》,1883年),把数学方法(统计方法)引入了心理测量,

并对测量中的问卷法、自由联想法以及等级量表应用等方面作出了贡献。

1890年,美国心理学家卡特尔(J M Cattell)发表了《心理测验与测量》一文,在文中他明确提出:"如果我们有一个普遍的标准,使在不同时间和地点得到的结果可以比较的话,那么测验的科学性和实用性的价值必然会大有提高。"事实上他主张建立常模、制定标准,并首次提出了"心理测验"这个术语,为教育测量的发展开辟了新的途径。

19世纪90年代至1904年之间,各种测验蓬勃发展,研究心理与教育测量的人也越来越多。例如,1889年厄恩(Oehrn)就曾编制过知觉、记忆、联想和运动机能方面的一组测验;1892年波尔顿(T L Bolton)发明过勾消测验,施行过记数测验,1896年艾宾浩斯(Ebbinghaus)曾用算术运算、记忆广度、句子填充测验测量过小学生;1896年皮尔逊(Karl Pearson)改进了他的老师高尔顿的相关系数的计算;1903年克来(Kelly)已有了以智龄为常模的思想;1904年斯皮尔曼(Spearman)在其《普通智力》一文中开始提出智力二因素论。这都为以后科学的心理测量和教育测量的产生和发展起了重要作用(尽管此时心理测量标准化程度低、缺乏理论指导等等)。

值得一提的是 在20世纪30年代以前,现代教育评价概念尚未形成,教育测量与教育评价几乎是同义词。事实上,人们此时关心的主要是个体学习状况,对学校乃至整个教育的评价并未正式提出来。当然,个别的教育评价活动还是有的。例如,19世纪的英国已有了许多评价教育、公共卫生等事业的活动(这种活动通常是由政府指定成立的委员会来调查大家所关注的地区的一些现象),并于1839年开始实行了视导制度。在教育界,这种制度要求视导员每年要通过对学校的访问和调查,提交关于学校现状和学

生的报告，作出评价结论。在美国，对学校成就进行评价的最早尝试是1845年。当时，波士顿教育董事会把笔试引入文法学校后，就用学生的测验成绩来评价学校，并根据评价结果来决定校长的任免。到19世纪后期，“美国北部大学和中等学校中心协会”成立，学校评价采取了同行判断的方法。发展到后来，便有了类似于“目标参照性测验”的测验活动，有了对经费使用、学生退学率、升级率等方面进行的更为综合的评价活动。

第三节　外国的现代教育测量与评价

一、现代心理测量的产生和发展

著名学者波林(E G Boring)指出：在测验领域中，“19世纪80年代是高尔顿的10年，90年代是卡特尔的10年，20世纪头10年则是比奈的10年。”事实上，世界上第一个现代标准化意义的心理测量量表是由比奈(A Binet)创立的(1904年)。

1904年，法国教育部曾委派许多医学家、教育家与科学家组织一个委员会，专门研究公立学校中低能班的管理办法。比奈是该委员会成员之一，他与助手西蒙(T Simon)提出了《诊断异常儿童智力的新方法》，向世人介绍了第一个智力量表——比奈－西蒙量表(1904年)。该量表经过多次修订(1905年，1911年)，至今仍有很大影响。

自比奈－西蒙量表问世以后，心理测验运动便逐渐兴起，20年代开始狂热，40年代达到高峰，50年代后转向稳步发展。在此期间，测验的编制不仅有文字测验，而且有非文字测验(宾特纳所编的非文字量表最好)，不仅有个体智力测验〔推孟(L M Terman)

编的斯坦福－比纳量表最著名〕，而且有团体智力测验〔奥蒂斯（A S Otis）编的陆军甲、乙两种测验最著名〕；不仅有能力测验，而且有能力倾向测验〔韦氏儿童、成人、学前智力量表（Wechsler，1949、1955、1967年）最出名〕；不仅测量人的认知，而且测量人的情感、人际关系、动机、兴趣、态度、性格等人格特点〔如罗夏（Rarshach）墨迹测验、卡特尔16PF、艾森克EPQ等〕，同时关于智力的概念，先后出现了斯皮尔曼（C Spearman，1904年）的智力二因素理论（一般智力g和特殊智力s）、瑟斯顿（L L Thurstone，20世纪30年代）的智力群因素论（七种），以及吉尔福特（J P Guilford，20世纪60年代）的智力多维结构理论等。此外，认知心理学的兴起也促进了人们关于认知成分以及元认知的测验研究。

二、现代教育测量的产生和发展

正当心理学家们忙于发展智力测验的时候，传统的学校考试也正在进行一场改革，卡特尔的学生桑代克（E L Thorndike）等人，利用心理测量原理，编造了第一批标准化的教育测验。1904年，桑代克出版了《心理与社会测量》一书，介绍了心理统计方法及编造测验的基本原理。这是世界上第一本社会科学方面的测量专著。1908年，在桑代克指导下，斯通（C W Stune）编造了一个算术推理测验，这是一种最早的标准化测验。1909年，桑代克发表了书法量表，这是世界上第一个用科学方法制成的教育测量工具。自此以后，各种标准化测验和量表日渐增多，由单科测验发展到成套的一般学绩测验，由常模参照测验发展到目标参照性测验；由小学扩展到中学、大学，由用于调查和选人发展到用于诊断和促进教学；由对知识能力的测量扩展到对学习态度、兴趣以及品德、性格等方面的测量。因此，有人（L P Ayres）说："我们既称莱斯为教