实用教育评价理论与技术

刘五驹 著

苏州大学出版社

实用教育评价理论与技术

刘五驹 著

苏州大学出版社

图书在版编目(CIP)数据

实用教育评价理论与技术/刘五驹著. —苏州：苏州大学出版社，2008.12
ISBN 978-7-81137-198-7

Ⅰ.实… Ⅱ.刘… Ⅲ.教育评估 Ⅳ.G449

中国版本图书馆 CIP 数据核字(2009)第 011695 号

实用教育评价理论与技术
刘五驹 著
责任编辑 杨 华

苏州大学出版社出版发行
（地址：苏州市干将东路 200 号 邮编：215021）
丹阳市教育印刷厂印装
（地址：丹阳市西门外 邮编：212300）

开本 880mm×1 230mm 1/32 印张 8.625 字数 264 千
2008 年 12 月第 1 版 2008 年 12 月第 1 次印刷
ISBN 978-7-81137-198-7 定价：18.00 元

苏州大学版图书若有印装错误，本社负责调换
苏州大学出版社营销部 电话：0512-67258835

目录

第十章　高考改革探索

第一章
教育评价的理论基础

本章主要阐述教育评价的基本概念、教育评价发展的历史进程、教育评价的主要类型和教育评价的主要模式。

第一节　教育评价的概念

一、评价

评价作为人类的一种行为活动，自古以来就伴随在我们的日常生活之中。“评”就是议论是非高下，“价”指价格高低，价值几何。《宋史·戚同文传》有：“市物不评价，市人知而不欺。”评价在这里是议论货物价格之意，即所谓讨价还价。《辞海》(1999 年版)指出，评价今亦泛指衡量人物或事物的价值。

人们凡要理性地行为，总是在评价的基础上进行的，这包括行为的条件和可能、行为的方式和结果。有了比较清醒的认识，然后行为才可能具有理性。人类对评价的概念有一个逐步产生和发展的过程，但是，人类的评价行为活动却是古已有之。特别是对人才的评价与选拔，中国在人类历史发展的长河中有着无可替代的地位与作用。尧舜禹的禅让，就是人才评价的实践活动，5 000 年中国的文明史也是一部人才选拔评价的发展历史。

作为一种活动，评价一词与英文 evaluate 相对应。从词汇学角度分

析，这个词由三部分构成：前缀 e-(加强意义)＋词干 valu[e](估价)＋动词后缀-ate。显然，evaluate 的核心词义离不开价值的判断。但是，evaluate决不仅仅是面对最终结果作出价值判断的活动。这个词在数学中还是一个广泛使用的词汇，即数值的计算、求解的活动。即使在讨价还价的过程中，对事物本身特征的把握，也是以价值衡量为前提和依据的。因此，严格地讲，评价的概念应包含两层含义：对事物本质特征或相关特征的把握，以及由此作出的价值判定。

二、教育评价

简而言之，教育评价就是对教育现象本身特征的测定并作出价值判断。在教育评价理论的发展历程中，由于人们对教育评价本身认识的逐步深入和观察视角的不同，教育评价的概念有着多种不同的解说。比较典型的说法有以下几种。

1. 把教育评价等同于教育测验，甚至认为其主要内容就是考试。

2. 强调教育评价是一种“专业判断”，更看重评价者的经验和综合素养。

3. 认为教育评价是把人们的行为及其结果与应有的理想状态或既定的目标相比较的过程。

4. 教育评价是一种有系统、有目的地搜集资料、协助决策的过程。

研究者们对教育评价的功能、目的也有着完全不同的看法。布鲁姆(Benjamin S. Bloom)在其著名的《教育评价》一书中开宗明义，第一句话就是说：“教育的功能：挑选还是发展?”由此，是发展的评价还是选拔的评价成了许多人评判各种教育评价好坏优劣的分水岭。

教育评价概念的外延也有一个逐步扩张、演变的过程。在 20 世纪 30 年代，随着著名的“八年研究”教育评价学的诞生，人们对教育评价的研究主要集中在学校教学评价上。之后，随着研究的深入，研究的范围也逐渐扩展到学校教育的各个方面。近 20 年来，我国理论界对教育评价的研究领域已经涉及校内校外几乎所有与教育有关的现象。

教育评价的概念有一种比较简明扼要的表述，并在国内外文献中被广泛引用：

$$\text{评价}=\text{行为方式的描述}\begin{cases}\text{量的描述(测量)}\\ \text{或}\\ \text{质的描述}\end{cases}+\text{行为方式的价值判定}$$

人们开展教育评价活动有不尽相同的目的，人们在教育评价活动中各自扮演不同的角色，人们观察教育评价有不同的角度和侧重，所以对教育评价作出不同的解说是完全正常、合理的现象。上述各种不同的解说都是言之有据的，都是从某种角度描述了教育评价某一方面的特征。综合这些认识，可以使我们更全面地了解教育评价。各种不尽相同的解说也在一定程度上反映了人们对教育评价功用的不同预期，同时它也反过来促进教育评价活动多方位的扩张和发展。

如果我们不去过多地考虑教育评价的目的、意义、功能等因素，而只考虑教育评价自身的活动内容，教育评价可以概括为：对教育现象及其相关因素的外显客观特征进行测量(定性或定量的)，并且作出价值判断的过程。

之所以这样来定义，首先是因为试图用一段文字全方位地把教育评价的内涵和外延作出全面的界定是非常困难的，甚至是不可能的。而且这样必然使文字冗长，且无实际价值。对一个基本概念的理解是一个随着学习、研究逐渐深入的过程，而且只要学习研究不断，理解认识会不断深入。

其次，教育评价的对象包括教育目的、内容、方式方法，教师、学生、教育行政人员、教育教学的管理、总务管理等所有教育现象。同时还包括影响教育的各种非教育因素，如校园的周边环境等。

作为一种科学的评价，评价的依据、教育评价对象的特征必须是客观的、外显的，而不能是主观臆测的。这既是教育评价活动的基础所在，更是教育评价研究长期的奋斗目标。对各类重要的和基本的教育现象本质特征的把握，是教育评价研究的艰巨任务。

从教育评价活动性质的差别来看，教育评价活动可以分为前后相继的两个阶段：测量阶段和价值判断阶段。测量阶段是对评价对象相关信息的搜集、整理，以及关键特征参数水平的测定。价值判断阶段则是在测量阶段的基础之上，根据特定的评价目的对评价对象作出价值评判。

测量阶段需要使用各种质的或量的信息搜集办法，要借助于统计

学、测量学的各种技术。测量阶段是解决评价对象“是什么”的问题，即获取评价对象关键特征的客观数据。因此，客观性是测量阶段教育评价的基本特性。在测量阶段还有一个相对客观性而言更为重要、也更为困难的要求，即有效性。所谓有效性，是指评价选取的这些关键特征是否反映了评价目的所需要的评价对象的本质，测量获得的数据是否代表了这些关键特征参数的真实水平。

价值判断阶段客观性是相对的。在理想状态下测量阶段的结果应该是客观、唯一的。但是，同样的测量结果，其价值判断却可能是多样的，甚至得出完全相反的结论。其原因在于评价者对测量结果认识程度的不同和价值取向的不同。这是教育评价与一般评价的一个重要区别。对人的评价，特别是对成长变化中学生的评价是一件极其复杂的事情，它往往不存在唯一的价值标准。

例 1-1：学生在学校中的学业成绩，无疑是评价学生学习状态和学业成就的重要指标。在中国恐怕很难找到一位老师或家长不希望自己的学生或孩子不进一步提高自己的学业成绩。在最近十几年愈演愈烈的生源竞争大战中，学业成绩，甚至仅是某种考试成绩成为选拔最基本、也往往是最终的标准。那个号称公平、公正的最后一道屏障的高考也是以成绩作为评价的最重要的标准。但是，所谓“前十名现象”早已是教育界普遍关注的现实问题。如果你是学生的老师或孩子的家长，你会采取什么样的评价标准呢？

显然，测量是作出价值判断的基础，价值判断是测量的最终目的。作出有效的测量本身就是一件不容易的事，而要在此基础上作出正确的价值判断，则更为困难。严格地讲，除了极端状态之外，对一个尚在成长、发展中的个体学生的某种素质作出绝对的价值评判是不可能的，我们所能做的仅仅是一种特定条件下的相对价值判断。这里隐含着哲学性的问题，值得我们深思。

三、相关术语分析

在我国，与教育评价概念有着密切关系的词汇有两个：教育测量与教育评估。

从上面的分析可以看出，教育测量本身就是教育评价的重要组成部分，是进行价值判断的前提和基础。从学科发展的历史来看，教育测量学的发展在前，教育评价学的产生在后。关于教育和心理测量的研究，在19世纪中后期已经在西方展开。如果考虑到考试研究，则人类关于教育测量的活动已经有了数千年的历史。而教育评价学的产生则是20世纪30年代的事。当时，人们发现各种孤立的某种心理或教育现象的测验或测量并不能正确、全面地分析教育现象，并促进学生的发展，因而需要一种综合多种因素、在一定价值观念指导之下才能作出更有益的评价。不过测量是一门领域非常广泛而内容差异很大、专业性要求很高的方法性和技术性学科，这就不是教育评价学所能完全包容的。教育评价学本身对测量的研究是在一般性的层面上进行的，它需要大量各种不同专业的测量学的深入研究，作为方法和技术上的支持和补充。

关于教育评价与教育评估，也有学者认为两者在概念上没有什么不同，完全一样。不过从现实应用来看，还是存在一些区别的。首先，在使用习惯上，两词各有相对固定的群体。政府和社会团体习惯使用教育评估，而学校内部和学者则更倾向于使用教育评价。其次，在现实活动中，对两个词的取舍不同，在侧重于客观事实的测定时，往往使用评估一词，而在强调评价的主观价值判断时，更多使用评价一词。当然，这种区别也仅仅是现实中的一种习惯差异，并无一定之规。

教育评价的系统理论来源于西方。然而在英文中，评价一词似乎比中文还要乱一点。在1933年至1940年标志着教育评价诞生的“八年研究”中，泰勒(R. W. Tyler)等人正式提出并使用了educational evaluation(教育评价)一词。但是，由于历史使用习惯和现实的原因，有几个意义相近的词往往在使用中相互混淆。这里根据W. 詹姆斯·波帕姆(W. James Popham)的看法略作区分。

measurement一词的基本含义是测量，强调在定量形式上对评估对

象的测定。过去教师们往往把各种测试中对学生评分的测量行为误认为就是评价。grading 一词也有类似情况，其基本的含义是等级的确定。它的本质仍然是一种测量行为，但缺少对评估对象价值的判断过程。

assessment 和 appraisal 这两个词是最经常与 evaluation 相混淆使用的。在我国，前者往往就译为“评价”，后者常被译成“评估”。assessment 似乎既包含了测量的意义，也包含了价值判断的意义。因此有人使用评价这个词时实质仅仅是表达测量的意思，但有人用它却包含了价值的评判。而 appraisal 则更倾向于价值评判，往往作为 evaluation 的同义词。

第二节 教育评价的历程及借鉴

教育评价作为一种教育活动现象古已有之，有教育活动就会自然而然地出现教育评价活动。最早的、也是最基本的教育评价活动形式就是对教育对象的观察，通过观察了解教育对象的现状以便于决定如何进行教育或了解教育的效果。对教育评价的历程，根据不同的标准可以有不同的阶段划分。一般以教育测量学的兴起和教育评价学的兴起作为两个划分的界线，把教育评价的发展划分为三个大的历史时期：教育评价产生和发展的萌芽时期、教育测量的蓬勃发展时期以及教育评价兴起时期。当然，在每个大的时期中，又可以分出多个发展阶段。

一、教育评价产生和发展的萌芽时期(1864 年以前)

古代的教育评价主要集中在对人才的选拔和早期的教育考试制度。在这方面，中国有着无可替代的历史地位。中国古代教育评价的发展以隋唐时期科举制度的产生为界，又可以分为两个时期。

（一）科举以前的时代

从某种意义上讲，有人群的共同生活，对人的评价就自然产生。每当人们对各种人或事物作出自己的判断时，评价行为就已经蕴含其中。

中国文明的历史，大抵是从尧、舜、禹的传说开始的。在那个时代对

部落联盟的首领选任就有所谓“九德”的标准：“宽而栗，柔而立，愿而恭，乱而敬，扰而毅，直而温，简而廉，刚而塞，强而义。”(《尚书·皋陶谟》)

西周时期，建立了视学制度和选贤贡士制度。视学包括两种情况：一是天子象征性的视学，二是督导性的视学。督导性的视学隔年进行一次，视察的内容包括德行和道艺两个方面。选贤贡士制度是由诸侯和地方官吏选拔德行、道艺兼优者贡于天子，或升入大学。这是目前我国文献所知最早的人才选拔制度。不过，真正的贡举名额很少，目的在于引导社会风尚，实现化民成俗的目的。《礼记·学记》对周代的学制及评价标准有一段著名的描述：“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，强立而不反，谓之大成。”这说明当时的学校是每年招生，隔年进行一次考核，第七年考核合格者可谓达到小成，第九年考核学生，如果达到触类旁通、坚信不疑的程度，则达到大成。

春秋战国官学衰败，私学兴起，文化随文人散落民间。一时间诸子百家兴起，各持不同的思想，各有不同的教育制度及评价标准。这一时期对人才选拔评价最大的特征就是无一定之规。不论家庭出身背景如何，不论身在哪个社会阶层，不论地域国别，只要有一德，或有一技、一智、一才能或一学识，或风云际会为诸侯所用，或著书立说自立门户，广招门徒成一代宗师，都能成就一番事业，影响一时，甚至流芳百世，并因此创造了中国历史上一段群星璀璨、思想勃发、最为耀眼的文明。其中显然有值得我们深思的道理。这一时期还出现了关于学生学习、生活行为规范的重要文献《管子·弟子职》，它被认为是“当时齐稷下学宫之学则”(郭沫若、闻一多等《管子集校》)。这是目前所知中国最早的学生守则。通过它我们也可以看出，稷下学宫的管理体现了那个时代的特点，松而不散，自由而有序。

汉推崇儒学，建立了比较完备的中央和地方两个相对独立的官学体系。汉代人才选拔采用察举选取士制度。汉高祖十一年(公元前196年)，刘邦下诏求贤，要求郡守劝勉贤士应诏。汉文帝则亲自出题策问。“策问”即皇帝就治国政务提出试题，写在竹简上，称作“策题”，应试者在竹简上写出回答称为“对策”。这是中国历史上正规的人才选拔考试，并且是采用笔试形式的开端，也被认为是世界上最早有记载的笔试。汉武

帝时期，察举制度作为一种固定的选士制度被确立。每年州举“秀才”，郡举“孝廉”，历代沿习。察举分为两大科目：一是每年由州郡按规定向朝廷推荐人才，为常科，也称作岁举；另一类是皇帝根据需要直接指定选士标准和名目的科目，为特科，也称诏举。常科包括孝廉和秀才（东汉因避讳改称茂才）两科。特科有贤良方正、明经、童子等科。察举制度有两大特点：一是举荐与考试相结合；二是选才与用人相结合。这两点在当时并对后世产生了深远的影响。举荐与考试相结合使人才的选拔相对更规范、更合理，选才与用人相结合激励了整个社会重教向学之风，其影响绵延了2 000余年。但是，由于举荐在考试之前，使得弄虚作假、攀附权贵之风渐行。特别是统一规范化的评价标准，使社会选人、用人的标准。甚至整个社会的价值取向都渐趋单一化，由此造成对人们思想的禁锢，至今余毒未消。

魏晋南北朝时期，察举制逐渐退居次要地位，“九品中正”制度成为主要的选士制度。所谓九品中正制就是由朝廷在州、郡设立大、小中正官，负责考察士人的家世和德才表现，据此将士人评定为九个品级（上上、上中、上下、中上、中中、中下、下上、下中、下下）逐级上报，吏部选择前三个品级者授予官职，又称九品官人法。九品中正制是以“唯才是举”的宗旨问世的，其关键在于中正官是否“中正”。为此，刘劭奉诏作《都官考课》七十二条。但是，在当时的政治与文化环境下，九品中正制最终演变成以世家门第为品评士人的唯一标准，导致“上品无寒门，下品无世族”（《晋书·刘毅传》）的结果。

（二）科举时代

隋文帝在开皇年间正式废除“九品中正制”，依察举之制选拔人才。隋炀帝大业二年（606年）始置进士科，标志着科举制度的创立。唐承隋制，以分科考试选拔人才，渐成定制，世代相承，成为在中国历史上推行了1 300年之久的科举选士制度。科举制度不仅对我国的政治和教育制度产生了重大影响，它对民族的精神和文化也产生了深远的影响。

唐代的科举考试依汉代察举制，也分为常科和制举两类。制举为天子特诏举行，通常所说的科举是指常科。常科开有秀才、明经、进士、明法、明字、明算等科。秀才科要求最高，取人最少，唐太宗之后便名存实亡。明法、明字、明算科选择的人数不多，学子们热衷于明经和进士两

科。考试方法主要是帖经、墨义、对策和诗赋，主要是对“五经”的记忆，虽有对策考试，最终还是落入对经典的套用。隋唐科举制度的创立，是中央集权的政治需要，也是察举制度发展的结果。科举制打破了自汉魏以来人才选拔依门第或推荐而定的规定，对稳定社会，以及改变人才退化、政府无能的状态产生了重大影响，也是后世文官和公务员考核录用制度的重要思想渊源，并由此逐步形成了中国社会的价值取向和公平观念的变化。

唐代官学有比较严格的考核制度，如国子监“六学”考试分为三种：旬考、岁考和毕业考。另外，唐代对教师和教育行政官员实行定期考核制度，称为考课制度。每年一小考，三至五年一大考，内容包括业务、品德和教学效果，并由此决定升迁和奖励。不过，随着科举的发展，教育渐渐沦为科举的附庸。

宋朝执行“兴文教，抑武事”(《续资治通鉴长编》卷十八)的治国方针。北宋就有三次称为兴学运动的教育改革。第一次兴学是由范仲淹领导的庆历兴学，改科举考试以策论为主，出现了胡瑗的“苏湖教法”，提倡明体达用，采用分斋教学法，试图把兴学与科举结合起来。第二次兴学是王安石领导的熙宁、元丰兴学，试图建立教、养、取、任相配套的人才制度，并实行“三舍法”，学生需要通过考试逐渐晋级。第三次兴学是在宋徽宗倡导下进行的，提出废止科举，代以学选。三次兴学的目的都在于振兴官学，改变教育沦为科举附庸的状况，培养有真才实学，能够经国济世的人才。三次兴学都在复杂的政治斗争中宣告终止，但是，他们的探索和尝试，留下许多至今仍值得反思的经验、教训。

明清两代科举制度受到高度的重视，制度趋于完备，考试更标准化、规范化，科举呈现高度程式化的特征。明清两代科举常科只有进士一科，每三年举行一次。考试分为三级：乡试、会试和殿试。明朝规定，参加科举考试必须有州、县学生员资格。也有人把获得参加科举考试资格的“童试”算做一级，则科举考试共有四级。乡试在各省城举行，由学政主持，考场称“贡院”，考中者统称“举人”，第一名称“解元”。会试由礼部主持，于乡试的第二年在京师举行，考中者称“贡士”，第一名称“会元”。会试一个月后举行殿试，由皇帝主持。殿试无淘汰，仅把应试贡士排出名次。所有应试者均赐出身“进士”，第一名称“状元”。明清科举考试最

大的特点就是以八股取士。试题只能取自《四书》、《五经》，答题只能代圣贤立言，不能做任何发挥，文体只能是八股文（包括破题、承题、起讲、起股、中股、后股、束股和大结几个部分）。八股正文必须有四段对偶的句子，每段两股，合起来为八股。明清的科举形成了一整套从考场到批阅试卷等规范严谨的制度。如入场的搜检，试卷的“糊名”，阅卷时的“朱卷”等。明清两代科举制度的完备化是以学子的思想禁锢为代价的。科举已经成了具有完备游戏规则的一场竞赛，而与社会对人才的实际需求并无多大关系。1905 年，科举制度终于走到了它的尽头。

（三）考试制度的演变

在古代，中外教育考试都有一个由“口试”向“笔试”的演变过程。其原因有二：

首先，口试是人类自然而然形成的最简便的考试形式。语言是人类能够系统、完整、深入地进行信息沟通的基本工具，教师通过交谈把知识传授给学生，同样还是通过交谈，教师可以了解学生对知识的掌握情况。我国先秦诸子百家的教学大抵如此，古代希腊的先哲们也是如此进行教学的，如著名的“产婆法”。当然，交谈的方式，亦即口试的方式可以有多种形式。如师生一对一的交流，一师对多生的交流，更为正规的是多师对一生的口试。

其次，笔试是需要一定的技术与资源为前提的。在造纸术普及之前，书写是一件比较麻烦，甚至是有一点奢侈的事。不要说羊皮、布帛价格昂贵，即使是竹简也难以在一般的考试中广泛使用。

在西方，根据有关文献（林昌华《学校教育评价》）记载，正式的口试首先于 1219 年在波兰大学的法学考试中使用，1702 年英国剑桥大学开始用笔试取代口试。

从口试发展到笔试是历史的进步，也是教育评价发展的需要。不过口试和笔试都各有利弊，在学校教育实践中，两种形式始终都在使用。只是在正式的、较大规模的考试中笔试才成为基本的考试形式。

口试的优点在于可以直观地、综合地考察个人的语言、思维及应变能力。口试的缺点则在于它过于费时、费工（费师），且对学生心理压力较大，受主考者主观影响较大，在缺少多媒体的时代是无法复验的。

笔试的最大优点就是它的高效率，一份试卷可以让成千上万的学生

同时使用，且考试原始信息便于保存、复验。但是，用纸、笔的形式所能测量到的人的能力是有限的。特别是在大规模考试中，由于公正的需要，必须有统一的答题标准，笔试往往考察了学生记忆能力和“死”的知识，较难考察人的创造性能力的发挥。虽然笔试的评分相对口试来说要客观一些，但是对同一份试卷，不同教师评判下来，理科相差 10 分左右，文科相差二三十分是经常的事。于是，如何提高测量的精确、有效程度，标准化、客观化成为其后研究的重要方向。

二、教育测量的蓬勃发展时期(1864—1940)

1864 年费舍(George Fisher)搜集了许多学生成绩样本，汇集成《量表集》(*Scale Book*)。这被认为是开创客观化、标准化测量的发端。其后，1879 年冯特(Wundt)在德国莱比锡建立了第一个心理学实验室。1882 年，高尔顿(Francis Galton)在英国伦敦建立了人类学测验实验室。1895 年，比纳(Alfred Binet)设计了第一套智力测验方案，并在 1905 年与西蒙(Simon)共同设计了著名的比纳-西蒙智力量表。

1897 年，莱斯(J. M. Rice)发表了拼字实验研究的结果，在社会上引起较大的影响，并引发争论。拼字实验选取了 20 所学校的 3 000 余名学生，在 8 年中，每天花 45 分钟时间进行拼字练习的学生与每天进行 15 分钟练习的学生相比较，结果是没有明显差别的。该实验引起了社会对教育测验的关注，推动了教育测量的发展。

1904 年，被誉为教育测量鼻祖的美国心理学家桑代克(E. L. Thorndike)出版了《心理与社会测量》(*Mantaland Social Measurements*)，标志着教育测验运动的开始。桑代克在书中有一句名言：“凡是存在的东西都有数量；凡有数量的东西都可测量。”

20 世纪 20 年代，美国的教育测验运动进入了蓬勃发展时期，各种测量量表、各类标准化测验层出不穷。但是，随着测量学的广泛运用，人们开始注意到仅靠测量特别是纸、笔的测量不可能对复杂的教育现象作出客观、全面的评价。于是，20 世纪三四十年代，教育评价应运而生了。不过教育测量仍然是教育评价的重要基础，20 世纪 40 年代后，教育测量进入深入发展阶段，产生了一系列新的测量理论与方法，并形成了许