贵阳市“教育思想讨论”参考资料二

# 中小学教育评价

（内部使用）

贵阳市教委“教育思想讨论”办公室编

一九八七年八月

教育不是主观的东西，想怎么办就怎么办。它一方面受经济条件的限制，另一方面要受社会实践的检验，所以衡量教育的好坏必须以其社会实践的效果为检验标准。我们研究教育思想也好，研究教育的任何其他问题也好，都不能离开这个基本的观点。

接受社会实践检验是很复杂的，要考虑到我们与世界交往，要面向世界；要预测未来，要面向二十一世纪；但基点是面向今后一个时期社会主义物质文明和精神文明建设的需要。

——何东昌

（引自《中国教育报》1986年12月16日）

# 目　录

# 第一章　教育评价概述

评价，是人类社会生活中的一项基本活动，是人们按照一定的标准对客观事物作出的价值判断。为了判断自己当前或过去活动的价值，人们每时每刻都在作出评价。然而人们的评价活动并不全都是有意识的和科学的。只有科学的系统的评价才能对被评价的活动有指导作用；并且人们也只对那些被认为是重要的活动，才进行有意识的科学评价。

## 第一节　教育评价的产生与发展

教育评价的产生与发展是和教育测量的产生与发展分不开的。自有教育以来，人们总要以一定的方式考核教育的成果。在古代，口试是主要的考查方式。后来，由于学校教育的发展和学生数量的增加，对考查的方式也提出了新的要求。鉴于口试既难于取得统一的评定标准，又不能大规模地同时施行，所以口试逐渐为笔试所代替，这是考试方法上的一大进步。一千三百多年前，我国唐太宗时期开始的科举制度是世界上公认的最早的笔试。直到18世纪后期和19世纪，英国和美国才在学校里使用笔试。早期的笔试大都采用论文式试题，评分上免不了带有主观性，并且不能测量出学生的知识广度。为了矫正考查方法上的这个弊病，力求考试客观化，本世纪初在美国逐渐兴起了教育测验运动。当时保证考试客观化的手段是使测验标准化。虽然标准化大大提高了教育测验的客观性，但是当时的教

育测验有个最大的缺陷，即它只偏重于测量记忆性的知识，对于人的能力虽然也有些涉及，但很难全部测定，尤其是忽略人的心理品质，并且不能顾及教育目的及价值观等。对于人的社会态度、实际技术、创造力和兴趣等在教育中起重大作用的领域，教育测量是不能把握的。到1930年前后，由美国经济大恐慌而引起的教育危机，使一些教育学家开始对教育的根本问题进行探讨，教育观点也随之改变，从重知的教育转而主张全人的教育。当时的进步主义教育联盟乃以全面发展人的才能为教育目标，首先对课程的内容进行了长达八年之久（1933——1940）的实验研究。与此同时，为了检查、比较两种教育形式的效果，推举著名教授泰勒等人组织了一个评价委员会，该委员会一方面批评教育测验中存在的问题，另一方面用教育评价的思想去研究考查教育效果的方法。认为只有评价的思想与方法，才能使新课程达到目标，实现理想，于是悉心设计了一套教育评价法。随着上述各项活动的实施，美国的教育测验运动于四十年代初转向教育评价。于是很多人把1942年泰勒关于教育评价的文章（General Seatemen on Psyclology）的发表看作是教育评价发展史上的里程碑。

五十年代，由于控制论、系统论、信息论思想的影响，教育评价在实现教育目标方面可以发挥反馈、调节、控制作用的思想进一步得到明确。六十年代是教育评价发展史上的一个重要转折期，一方面主张教育机会均等的思想代替了过去以优秀儿童为中心的个别差异教育的思想，另一方面由于社会生产力的迅猛发展，对科学文化提出了新的要求。与此相应，在评价与测验方面，也从过去注意人与人之间的差异和选拔优秀转为关心每个学生的全面发展。认为教育评价的目的已不仅仅是适应教学的需要，主要应为教育的管理、决策服务，通过科学的

管理和决策促使每个学生都能更好地掌握知识、发展能力。

在这种时代精神的要求下，六十年代的教育评价思想有了新的发展。出现了一些具有指导意义的理论和方法。例如1963年R·Glaser倡导高效标参照测验。这种测验以教学大纲具体规定的学生必须掌握的知识、技能为编制标准，然后以学生在这个测验中所得的分数与标准相比较，以了解学生达标的情况。一般说如果能达到80%——90%的话，便认为是通过而可以进入下一单元的学习。若达不到这一程度，则需要针对他们存在的问题进行补救指导，直到达标为止。

到了六十年代后期，教育评价的应用范围也逐渐从局部的、个别的评价转向大规模的国家教育系统的评价。例如美国，那时几乎所有政治家和公众舆论都关心美国的普通教育质量是否下降的问题。于是联邦政府决定实行“全国教育进展评定计划”。在全国范围内对9岁、13岁、17岁的学生和部分年轻人进行周期性的评价。通过这样的评价研究发现了许多有意义的结果。例如，最近的结果表明9岁学生的阅读能力比10年前有所提高，但图表阅读能力不及10年以前。中学生对命题进行书面逻辑论证的能力比5年以前有所降低等等。

教育评价发展到现在，它的主要目的是对整个教育活动、教育系统的工作情况进行分析判断，用以制定决策、改进工作。而教育测量则主要是定量地描述所要实现的教育目标或对此目标的实现程度。

我国是世界上最早使用心理测验的国家之一，在教育评价方法的发展史上作出过重大贡献。陶行知先生从事教育评价工作就早于美国。他于1920年以前视察了一些乡村小学。这年九月他参加了江苏省教育会讨论“乡村标准学校”的会议，于年底制定了乡村小学教育评价的“量表”——《乡村小学比赛

表》。这一量表分十一类一百零四项，充分体现了定量分析和定性分析相结合的原则。它不仅注重检查学校为普及初等教育、为生产建设和改造社会服务的质量，而且检查学校改革教育、教学方法，为学生德、智、体、美全面发展的培养目标而服务的质量。此外，陶先生还撰文提出了“专家”参与评价学校，提出了“客观的、系统的调查”的评价原则。

新中国建立以来，由于极左路线的数度冲击，教育屡遭劫难，评价自然也无从谈起。粉碎“四人邦”以后，随着教育事业的蓬勃发展，教育评价引起了党和国家的高度重视。1985年5月，《中共中央关于教育体制改革的决定》指出，要“组织教育界、知识界和用人部门定期对高等教育的办学水平进行评估，对成绩卓著的学校给予荣誉和物质上的重点支持，办得不好的学校要整顿以至停办”。赵紫阳总理在《关于第七个五年计划的报告》中更加明确地指出：“要加强教育事业的管理，逐步建立系统的教育评价和监督制度。”

几年来，我国的教育评价工作有了迅速的发展。1984年5月，我国正式加入了国际教育成就评价学会(简称IEA)；国内关于教育评估的讲座和研究会，已举行了不下于二十次，自发的和有组织的学会也纷纷成立；华东师大与加拿大国际教育发展协会合作研究“中国教育评价问题”，目前正从事一些国外有关著述的翻译及教材的编写。

中小学学校教育领域的教育评价探索，尤其令人注目。

1982年，上海市教科所开展的“初中平面几何成绩诊断性评定研究”，已获显著成果；1985年开始，该市教育局、教科所组织120人，进行中学教学的诊断性评价。

1984年年底，IEA组织“第二次科学研究”，选定北京、天津、河北、山西等省市作为试点，进行中学学校管理与学科

教学的评价试验，并准备将来在全国或更大范围内推广进行。

北京市教科所1984年开始，进行了“基本理论”、“学校评价”、“教师评价”、“德育”、“体育”、“教学评价”六个专题的研究；北京市教育局于1986——1987年在全市进行学校评价实验，力争尽快建立符合北京实际的教育评价制度。

天津市教育局制定了《评价试行方案（讨论稿三）》，要求区、县教育局所属中、小学开展全面评价工作。该市和平区1982年起对学校工作与学生德智体美劳进行全面评价，目前已形成一整套目标体系并付诸实施。

湖北黄石市教委发出在全市中小学开展评价的通知，并成立“中小学评价工作指导委员会”，编纂评价手册。

重庆市沙坪坝地区教育局，1984年也组织全区五十多所中学，一百多名校长、教导主任、教研组长开展中学教育质量评估活动。

此外，武汉、无锡、大连、青岛、福州等地也都积极开展了学校评价的工作。

## 第二节　教育评价的定义和作用

什么是教育评价呢？自1929年美国教育学家泰勒首次提出这个科学概念以来，它的对象和目标不断发展。从改进教学效果，发展到改善教育管理，甚至成为制定教育计划和决定某些教育方针和政策的基础。目前，对于教育评价尚无一个公认的定义，但我们认为教育评价的基本精神是注重教育效果的价值观。教育评价因对象不同或评价者的侧重点不同而有多种定

义。比较广义的定义可以是：教育评价是为教育决策提供信息的过程。这个过程，是运用科学的方法，有步骤地对有关教育现象进行数量测量、性质描述，并作出价值判断的系统过程。这一定义可适用于各种教育评价，如教育方案评价，教育规划评价，教学方法评价，教材评价，学校水平评价，教师工作评价和学生质量评价等等。例如，在教学改革实验中，为了了解各种教学方法的效果，需要在教学过程中不断检查预定的教学目标在哪些方面已经实现，为此可以进行一个或几个标准化测验，搜集有关数据和资料加以整理，并据此对各种教学法的效果进行分析比较，最后为决策者提供决策的依据，这个系统的过程就是教育评价（教法方面的）过程。

教育的评价与测量、评价与研究都有类似之处。如评价与测量都使用测验，比较系统地收集测试对象的有关资料；评价与研究都要对收集来的资料作出分析，推出结论。但它们存在着根本区别：评价与测量的根本区别在于测量只对事物作数量化的描述，评价则要在定量、定性的基础上进一步作出价值判断；测量只是评价的一种手段。评价与研究的主要区别在于研究的结论往往要求具有较大的概括性，不一定能直接用于实践，而评价的结论则是特定的，直接为教育决策服务。

教育评价的应用范围日益广泛，评价的目的也多种多样。如：对照教育目标检查教学效果；判断教材、教法甚至教学大纲的适用性；提供教育机构的一些基本信息，便于与家长或其他方面合作；评定学校的工作效率和教师的工作能力；甚至衡量地区、国家的教育成就，改进教育过程；在更广泛的范围上说，它甚至涉及到评定教育投资的合理性，人才资源的品质以及估价教育对国家、社会的影响等等。

从评价结果所起的作用考虑，也可将教育评价分为形成性评价和总结性评价两大类。前者在教育过程中多次进行，用以了解教育对象对给定任务的掌握程度和确认出哪些部分未被掌握；后者则在给定任务完成后进行，用以标定某种任务所造成的影响。如学期末给学生评定成绩或评定等级名次。

## 第三节　教育评价的一般程序

教育评价的程序一般说来可以分为教育目标与评价对象的确定、评价方法的选择以及结果的分析与解释。

### 一、教育目标与评价对象的确定

评价对象的确定依赖于教育目标的确定。多数学者认为，如果教育目标能够用一种测量的术语描述得清晰、客观和具体时，既有助于教育效果的提高，也有助于评价对象的确定。反之，如果这些目标不可测量或意义含混不清，则会给客观评价带来很大的困难。确定出表达清晰、客观、具体的教学目标的方法很多，最常用的是任务分析法（task analysis）。任务分析法最早用于工业心理学之中。其内容是：先对某种任务（或工作）所要求的基本知识或技能进行分析，并用操作性的术语表述之，然后以这些基本的要求作为标准来选拔工作人员，或对工作人员的工作效率、成绩等进行评定。这个方法在用于教育评价中、确定评价目标时，先对教育目标提出总的要求，然后分析它可以分解为哪些基本的成份，并用操作性的术语表述这些基本的成份。这些成份将通过测验结果等反映出来，然后可以和预期的目标进行对照比较，这样便可以获得有关教材、教学方法、教育计划的有效性等多方面的信息。一般地来

说，评价的目标表述得越清晰、客观，评价对象就越容易确定。

六十年代中期，美国心理学家B·S·布卢姆等人研究提出了一个教育目标的分类系统，正是适应了教育评价在这方面的需要。他们认为整体的教育目标一定体现在教学活动中。所以他们把各种教学活动的目标分为：认知的、情感的和心理运动的三个主要方面。认知目标常常被认为是大多数教育目标之所在。他们又进一步把教学活动中的认知目标从简单到复杂分为知识、理解、应用、分析、综合和评价六类。这一分类法已经在国外的教材编写和教学评价及测验编制等领域发生了重要影响。

关于情感的教育目标目前还没有象上述认知目标那样明确、清楚的分类，因而也没有象认知目标分类那样对教育及其评价活动产生重要的影响。其原因之一就是由于情感还缺乏客观的测量指标。

目前，许多国家的教育工作者和政策制定者都已普遍接受了B·S·布卢姆等人提出的这套教育目标分类系统，并把它们作为评价各种教育活动、教育计划成就的主要标准。

二、教育评价方法的选择

由于评价是为决策提供信息的过程，因此评价的方法主要是收集资料的方法，一般可分为定性与定量两大类。常用的方法各式各样，主要有调查、观察、分析、测验四种。这些方法各有所长，调查可以了解学生的观点、兴趣、态度；观察可以了解学生的动作技能与行为；分析可以了解学生某些不太外显的技能与情感；测验可以了解学生的学业成绩与学习能力。

表一：四种评价方法的重要特点

| | 调查 | 观察 | 分析 | 测验 |
|---|---|---|---|---|
| 可获得的信息种类 | 意见；自我意识；主观判断；情感（尤其是态度）；社会意识。 | 操作或某些操作结果；情感（尤其是感情反应）；社会性的相互作用；动作技能；典型性行为。 | 学习过程中的学习成果（中间目标）；认知与动作；技能；某些情感产物。 | 态度与学业成绩；最终目标；认知产物；最高的作业水平。 |
| 客观性 | 高度主观；带有偏见与误差。 | 主观，但仔细组织与实施的话，可以客观化。 | 客观，但因时而发生变化。 | 最客观。 |
| 代价 | 省钱，但可能费时。 | 省钱，但十分费时。 | 不太费钱，准备时间长，但很关键。 | 最费钱，但每单位时间可得到最多信息。 |

在大规模教育评价中测验是最常用的。不同的评价对象、不同规模的评价要求采用不同的特定测验。因此，从事教育评价工作的人必须对测验的基本特点及其有关知识具有一定的基础。例如基本的统计学知识，有关测量的基本知识，如信度、效度、常模、测验的分类及各类测验的特点，测验编制的基本步骤等等。限于篇幅，下面仅就测验的分类作一简单介绍。

测验按其标准化程度可分为标准化测验和非标准化测验。

标准化测验具有以下特点：（1）所有的考生所做的试题、时限等，施测条件完全相同，计分手段及分数解释方法也完全相同；（2）测验是由专门机构编制，并且往往是根据某种测验理论编制的；（3）有信度、效度资料可查，并有常模资料；（4）测验使用的规模大，可以在整个地区、整个国家、甚至在国际上统一使用。与标准化测验相比，非标准化测验具有以下特点：（1）施测题目、时限、计分手段等对不同的被测者可能不同。（2）测验往往由使用者本人编制。所以，在教育界使用的非标准化测验往往被称为“教师编制的测验”。这种测验一般是根据经验而不是根据理论编制的。我国各级学校中目前使用的各种测验（考试）基本上属于此类测验。（3）一般无信度、效度资料，也无常模。当然，这并不是说非标准化测验就不可能具有信度、效度和常模资料。它们之所以没有，往往是因为时间、经费等条件的限制而未收集这些资料。再则，这种测验的信度、效度以及分数的意义有时是可以由编制者（教师）根据经验进行估计的。（4）测验规模小，这种测验往往只在本班、本校范围内使用。

另外，测验还可以分为成就测验或能力倾向测验或简称能力测验。成就测验是用以测量学生在做测验时已经掌握了哪些知识技能的；能力测验有助于预测学生经过一段时间的学习或训练后将可能掌握哪些知识和技能。

评价按其规模来讲可以分为大规模评价和小规模评价。例如IEA目前在我国所进行的中学科学教育成就评价试点研究，涉及北京、天津、河北、山西的114所中学，3000多名学生，就是一项大规模评价，而教师对一个班级或一个年级所进行的评价属于小规模评价。

从理论上来讲，无论大规模评价还是小规模评价，都应该

使用标准化的测验。这是因为标准化的测验经过精心的编制或修订，信度和效度较高，分数的解释方法也比较科学。但是在实际上，小规模评价的评价对象比较具体和多样，而现有的标准化测验的测量对象却比较一般，品种较少，所以标准化测验往往难以符合小规模评价的使用要求。另外，由于标准化测验对编制的技术、测验的信度、效度等质量要求较高，所以要花费的时间、经费较多。因此，在小规模评价中常常不使用标准化测验。

在确定使用何种测验时，还要确定分数表示法和分数意义的解释方法。确定分数表示法的原则是：要求分数能表示出一定的意义，例如表示答对题数占总题数的百分比，表示考生在考生团体中的顺序位置等等。具体地说，可使用百分数、Z分数，经过转换的标准分数等。确定分数解释方法的原则是：分数意义的解释既要有科学的统计学依据，又要有经验资料加以证明。

三、结果的分析

测验结果的分析重点不在于测验的结果、名次，而在于产生这一结果的原因及其对实现教育目标的价值。例如在IEA科学教育成就的研究中采用理科标准成就测验的成绩代表各国学生的学习成绩，此外还要通过学生、教师和学校问卷收集关于学生家庭背景、教师经历、学校条件等各方面的资料，然后用多元分析的方法确定这些因素对于学生成绩的影响，为制订政策提供科学的依据。六十年代匈牙利通过IEA的研究，发现教学效率不高、教学方法过时、教科书内容不足是造成小学生阅读能力低的主要原因。这就导致了匈牙利小学课程的重要改革。在小学强调默读的训练，在8年基础教育中持续进行阅读技能的教学。

## 第四节　教育评价的模型和教学评价的类型

所谓“评价模型”，与通常所说的数学模型、经济计量模型相差甚远，有人甚至反对使用这个字眼。但是，作为一种抽象，评价模型集中地体现了评价者所依据的评价理论，所希望达到的评价目的，从而也大致确定了评价所使用的方法，是我们学习研究教育评价的重要内容。

由于各种评价的意图与情景不同，因而形成了各种评价模型。这里只介绍其中三种。

1.泰勒的目标达到模型。其特点是为方案制订出一系列期望达到的大目标，每一个大目标都用可以测量的行为目标来表示。方案实施后，进行测量，根据目标达到的程度来判断方案成功与否。此模式用于评价教学的结果或目标达到的程度。

2.斯克里文的形成——终结模型。斯克里文认为，不能只检查目标实现的程度，还应进一步评价目标本身是否有价值。所谓形成性评价是指在方案实施中进行的“中途”评价，它提供的反馈信息可使方案制定者及时修改或调整原有的方案；而终结性评价的含意与泰勒目标达到程度的评价相似。

3.CIPP模型。是斯塔弗尔比姆提出的评价模型。因为包括背景关系(Context)、投入资源(Input)、过程(Process)和结果（Produot）四方面评价而得名。在这一模式中，评价者与方案无直接联系，评价者只是为决策者描述、收集、提供决策所需的信息。在制订方案时进行背景关系评价，把现有方案与各种可能的其他方案加以比较。在设计教学时进行投入资源

评价，以便最合理地使用各方面的资源。在实施教学时进行过程评价，以便及时了解、反馈实施情况。在进入下一个循环时进行成果评价，以便就继续执行、修改或中止方案作出决策，可见这是一种比较综合的模型。

教师在教学过程中常用的评价有四种基本类型：测定性、形成性、诊断性和终结性。

测定性评价是在教学前进行的，分为两种：一是关心学生对教学的准备状态，即是否具有必要的预备性知识和技能。这种评价采用的是掌握性测验，试题难度较低，只限于最基本的要素，常常属于标准参照性的测量（即学生能完成的学习任务）。另一种则关心教学计划对班级的适合程度，采用检查性测验。试题总难度属中等，难度范围较广，常属于常模参照性的测量，以便了解学生的实际水平和差异。

形成性评价是在教学各阶段中进行的，目的是了解学生的学习进程，向师生提供反馈，强化正确的反应，指出需要改进的不足。采用掌握性测验，试题应根据有关学习任务的要求，可难可易，常属于标准参照的测量，测验的结果主要用于改进教学，不注重对学生进行分等。

诊断性评价在需要时（大多在形成性评价后）即可进行，目的是了解错误的详细原因。一般是根据学生存在的最常见的错误原因来编制试题，测验限于有限的内容，难度较低。

终结性评价在课程与单元结束时进行，结果主要用于对学生作出鉴定或分等，了解教学目标达到的程度，评价教学的有效性。期中与期末的考试常属于此类，采用常模参照性的测验。涉及面广，试题难度不等。

如果说，模型是教育评价方法的质的规定性，那么，评价方法的量的确定性则常用硬与软来表示，来衡量评价的信度和