现代学校管理评价指南

GUANLIPINGJIAZHINAN

本书编委会　编

中国科技文化出版社

第四卷

中国科技文化出版社

第一章　测验和评价在教学中的作用

对学生的学业进行评价需要运用许多学绩测量手段，但评价不仅仅是这些手段的集合。它是一个系统的过程，对于有效的教学有重要的作用。评价始于对教学目标的确认，止于对目标达成程度的判断。

我们对于测验和测量提出了大量的外部要求。显而易见，这些要求的结果会使得美国的学生接受大量的测验和测量。不过学生在教师自编测验以及教师从出版物中选取的测验上所花费的时间，要远远多于他们在标准化测验上所花费的时间。测量和评价是对教师自编测验和自选测验的补充，甚至这一说法也不能充分概括测量和评价所包含的范围。另一种说法进一步扩展了对这一范围的理解，即测量和评价是对学生的答题纸、家庭作业、操作任务以及学期考试的正式评价。但这种说法依然是不全面的，因为对随时发生的课堂活动有重要作用的多数评价是非正式的。

课堂中的非正式观察可以指导教学决策。例如，学生的口头提问可能表明需要对材料进行全面复习，课堂讨论有助于教师发现一些必须纠正的错误理解，对某个话题的兴趣也许表明应在该话题上花费更多的时间。同样的，在观察个别学生时，教师可能做出这样一些判断：桑德拉在写作上需要帮助，比尔应该多练习些数学题，胡安和贝蒂的语音应当矫正，应当鼓励玛丽亚读些更有挑战性的书。

在教学过程中教师要不断做出诸如此类的教学决策：一些基于学生的口头回答，一些基于对某项技能的实际操作，另一些基于学生困惑的表情、语调或身体动作。而所有这些都以教师随时进行的观察为基础。尽管这些观察是非正式的，但它们在有效教学中具有不可或缺的作用。

针对学生学业的测量和其他评价方法不是为取代教师的非正式的观察和判断而设计的，而是为了补充和扩展教师的这些非正式手段，其目的在于获取与学生有关的信息。教师仍然是观察者和决策者，测量和评价方法仅仅是为教学决策提供更加系统、客观的依据。

第一节　教学决策需要评价信息

教师要做出大量的决策，因而就需要更系统地测量学生的态度、成绩和个性的发展，以补充对学生的非正式观察。虽然我们不可能将教师所要做出的全部决策毫无遗漏地罗列出来，但是我们可以确定那些较为一般的决策。下表列举了一些教师在教学过程中可能会遇到的教学决策，同时括号中还列出了解决这些问题的最有效的测量和评价方法。

1. 教学计划是否适合学生？（学习能力倾向测验、以往成绩的记录）

2. 怎样对学生分组才能促进更有效的学习？（教师自编的测验、以往成绩的记录）

3. 学生是否做好了接受下一阶段学习任务的准备？（对所需技能的预测、以往成绩的记录）

4. 学生在多大程度上达到了教学目标？（教师自编的测验、课堂作业、提问，观察）

5. 在满足基本要求后，学生的进步达到何种程度？（教师自编的测验、综合成就测验、课堂作业、提问、观察）

6. 什么时候（对哪个知识点）进行复习最为有效？（阶段考试、提问、观察）

7. 学生有哪类学习困难？（诊断测验、观察、提问、学生成长记录袋、学生咨询）

8. 应当建议哪些学生参加咨询、特殊班级或治疗项目？（学习能力倾向测验、成就测验、诊断测验，观察）

9. 哪些学生缺乏自我了解？（自我评定、学生会议）

10. 怎样给学生评定适当的分数？（综合所有的评价信息）

11. 学生的哪些进步是应该告知父母的？（回顾所有的评价信息）

12. 我的教学效率如何？（成就测验、学生的评定、上级的评价）

这些问题都表明了教师在教学过程中需要多种类型的信息。在教学过程中需要回答大量的问题，而且在不同决策间存在一定的交叉。对于一个特定的评价情景，我们可能会用到多种多样的评价信息。教与学的过程中包含一连串互相关联的、以促进学生的学习为目的的教学决策。我们的观点是，在很大程度上，教学的有效性取决于那些作为决策基础的信息的性质和质量。

第二节　评价、测验和测量

评价、测验和测量这三个术语很容易被混淆，因为它们可能被包含在同一个过程中。评价是一个更为一般化的术语，它包括获取与学生学业有关信息的所有方法（观察、表现或项目评价、纸笔测验），也包括对学生学业进步的价值判断过程。测验是评价的一种特定的形式，通常是由一组要求在固定时间内完成的题目组成，并在相同的条件下对所有学生施测。不过，尽管测验只是评价的一种形式，我们有时还会同时使用这两个词。这时候，“评价”这个词所强调的，其实是那些不会由“测验”一词所联想到的表现或项目评价。测量是指依据特定的规则对测验或其它评价方式的结果进行量化的方法和过程（如计算正确回答的数目，给一篇文章的特定方面打分）。在下面的框图中，我们对每个术语的特定含义进行了小结。

与测量或测验相比，评价是一个更全面的、涵盖面更广的术语。测量限于对学生的定量描述，即测量结果总是用数字来体现（如，玛丽答对了 40 道数学题中的 35 道）。它既不包括定性的描述（如，玛丽的作业很整齐），也不含有对所得结果的价值判断。而评价则不同，它可以包括对学生的定量描述（测量）。和定性描述（非测量）两个方面。此外，评价还总是包含对结果的价值判断。图 8－1－1 体现出了评价的全面性以及测量和非测量手段在评价过程中的作用。如图所示，评价不一定要依据测量的结果；当以测量为基础时，它也超越了单纯的定量描述的范围。

术　语　表

评价：在获取关于学生表现的信息时所使用的各种方法的总称。它包括传统的纸笔测验、开放性问题（如论述题）以及对真实性任务的操作（如实验室实验）。评价所要回答的问题是：“个人的表现如何?”

测验：在相同的条件下，通过施测同一套问题来测量一个行为样本的工具或系统的方法。因为测验是评价的一种特定形式，测验也回答“在与其他人比较时，个人的表现如何”这一问题。

测量：对个体具有某一特征的程度进行量化描述的方法。测量所回答的是“程序”问题。

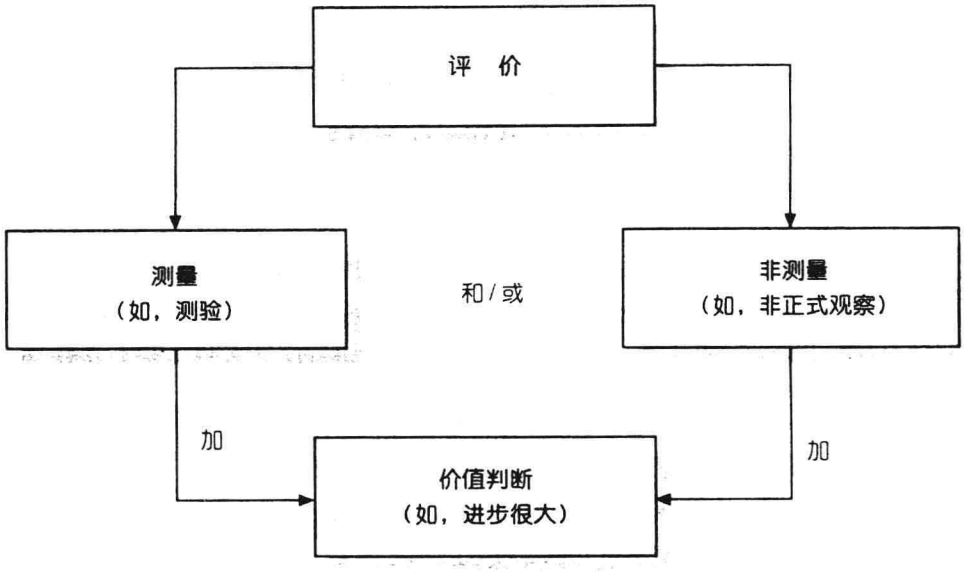


图 8-1-1　评价过程

第三节　评价的一般原则

评价是评定学生学习及发展的质与量的全过程。要想使得这一过程取得最佳效果，还需考虑以下原则：

1. **首先要明确界定评价目标。**有效的评价不仅取决于所用评价方法的技术含量，也取决于对所要评价的东西的仔细描述，这两个方面同等重要。因此，在选择或形成评价方法之前应当具体说明测量目标的特性。在对学生的学业进行评价时，就意味着在选择评价方法前明确界定预期的学习目标。

对内容标准或课程目标的总体描述，是一个有益的切入点。但在大多数情况下，为了使评价有效，教师应当做得更加具体。例如，历史课的内容标准应当是学生“在历史背景中去理解思想和文献”。对于在一般水平上进行陈述的内容标准，也许可以采用选择、简答或论述题。为了给这个标准确定一个适当的评价顺序，教师需要回答这样一些问题：“思想是什么？文献又是什么？历史背景指的是什么？评价学生理解程度的可靠指标是什么？”在对标准的一般陈述中，找不到这些问题的答案，但为了设计评价的方法，明确也好、含蓄也罢，总之必须做出回答。

2. **评价方法的选取应依其是否适用于所测量的特性或表现而定。**在选择评价方法时，我们经常得考虑客观性、精确性及便利性这几个因素。尽管这些标准很重要，但相

比于这一主要标准仍是次要的：这种方法能否有效测得想要考查的学习或发展目标？在随后的章节中所讨论的每种方法都有其特定的使用范围。例如，在评价学生的成就时，我们就需要使评价任务类型与既定的学习目标间达到良好的匹配。如果学习目标是训练观点的组织和写一篇完整作文的能力，那么就得对不同条件下完成的习作（如，随堂作文测试、写作作业和学期论文）进行分析，而这是关于写作方法的选择题所无法代替的。

3. **综合评价需要使用多种方法。**对于学校的教学体系中强调的所有的学习及发展目标而言，还没有哪一种工具或方法能够独挑起评价的重任。选择和简答测验适于测量知识、理解和应用，而论述测验和其它写作任务适合评价学生组织和表达观点的能力。一些需要学生提出问题、从图书馆查找资料或收集数据（如通过访谈或实验室观察）的任务，适于测量学生提出并解决问题的特定技能。观察技术适合评价学生的操作能力及表现的诸多方面，而自我报告技术适用于评价兴趣和态度。因此，要想完整地描述学生的学习及其发展情况，就需要使用多种不同的评价方法。

4. **要充分认识各种评价方法自身的局限性。**从十分完善的测量工具（如，标准化的态度和成就测验）到非常粗糙的评价手段（如，观察和自我报告技术），目前的评价方法良莠不齐。而且，即便是采用现在最好的教育和心理测量工具，其结果也易受多种测量误差的影响。

我们不可能在一个测验或评价中提出所有相关的问题，以实现对与课程目标或内容标准有关的全部知识、技能和理解的全面考查。相反，被采用的只是所有这些问题的一个样本。即使是一个相对较小的范围，如对光合作用或分数加减法的理解，也有大量的问题可被采用，但任何的测验或评价都只抽取这些问题的一小部分。因此，在教育和心理测量中，抽样误差是一个普遍存在的问题。一个成就测验可能不是某一特定教学内容的充分的取样，一个用于评价学生社会适应性的观察工具也可能没有抽取足够数目的行为样本，来作为该特质的可靠指标。但庆幸的是，严格遵循已有的测量程序可以使抽样误差得到控制。

第二种误差来源是影响评价结果的随机因素，如客观测验中的猜测、论述测验中评分的主观性、对观察工具的错误选择、自我报告工具中不一致的回答（如态度量表）。由于这些问题的存在，我们不能因为教育测验中的几分之差就给学生划等级。事实上，在教育或心理评价中，没有任何分数可以算做对所测特征的精确无误的度量。只不过谨慎地使用评价方法可以使测量中的误差控制在最低水平。

对测量结果的不正确的解释是误差的另一个重要的来源。有时候，教育评价的使用者们高估了结果本身的精确度，或者在解释时无中生有，将评价结果作为某些评价本身

并未涉及的特性的指标。如，有时候学习能力倾向测验的结果就被当做对；先天固有的能力、而不是具有可塑性的能力的度量；或者被认为测的是个人价值，而不是言语和数学推理能力。这种错误解释测量结果的现象十分普遍，而且它是影响评价有效性的主要因素之一。为了避免错误的解释，应当切实注意测验实际测量的内容及其所能达到的精确程度。

不过，评价方法的这些局限并不能抹杀测验及其它评价方式的价值。正确认识这些局限性可以促进对评价的有效使用。请记住，评价工具越粗糙，它的局限性越大，因此在使用时也就要更加谨慎。

5. **评价本身并不是目的，它只是达到目标的一种手段。**某种评价方法被使用就意味着它一定服务于某种目的，并且使用者也应当明确地意识到这一点。盲目地收集并堆积有关学生的信息，是对时间和精力的浪费。最正确的态度是将评价看做是为教学决策收集信息依据的过程。

第四节　评价与教学过程

课堂教学的主要目的是帮助学生达到一系列既定的学习目标。一般而言，这一目标应当包括学生在智力、情感和生理等方面的积极变化。当从这样的视角来看待课堂教学时，评价就成为教学过程中不可或缺的一个部分。相应的教学目标决定预期的学习成果，有计划的学习活动促成了学生可喜的进步，而学生的学习进展由测验或其它工具定期地予以评价。虽然教与学的相互依赖性是人所共知的，但是教、学及评价之间的这种相互依赖性却较少被人认识到。

下面框图中的内容说明了这样一个事实，即评价作为有效教学中不可或缺的一个部分，其重要作用已被教学革新中的领袖人物所认识。在下述教学进程的步骤中，我们可以清楚地认识到这三个教育要素之间的相依关系。

测量与评价——有效教学的必要是组成部分

州际新教师评价与支持协会（INTASC，Interstate New Teacher Assessment and Support Consortium）的创建目的，是为新教师建立有效教学的标准，并合作编制州教师资格认证的评价方法。他们提出："教师需要理解并使用正式和非正式的评价策略，以评估并确保学习者智力、社会性和生理的持续发展。"

国家职业教师标准组织（NBPTS，National Board of Professional Teaching Standards）是

一个为有卓越成就的职业教师提供认证的组织。他们提出："教师有责任管理及监督学生学习。我们所要认证的教师应当知道如何创造、丰富、维持和改变教学结构，来激发并保持学生的兴趣，并使用多种方法测量学生的成长与理解力。

一、确定教学目标

不论是教学还是评价，它们的第一个步骤都是确定课堂教学所要达成的学习成果。在一段学习经历结束以后，学生该如何思考和行动？他们该具备怎样的知识和理解能力？他们应该展示出何种技能？他们应该发展了怎样的兴趣和态度？他们的思维、感受和行为习惯应该发生怎样的变化？简而言之，我们正在力图达到怎样的特定变化，并且当我们成功地引起这些变化时，学生们应有怎样的表现？由州或学区确立的内容标准和课程指导为教学目的的说明提供了有益的参考。但是，为了更具体的界定学生的学习目标，并正确引导制定评价的具体过程，还需要对这些标准和目标进行更详细的说明。在本书的第 3 章中，提供了确定教学目的及目标的指南。

二、预先评价学习者的需要

一般来说，一旦明确界定了教学目标，就应当对学生与预期的学习成果相关的需要进行一些评估。学生具备进入下一步教学所必需的能力和技能吗？学生是否已经具备了目标技能或达到了期望的理解程度？在教学开始时对学生的知识和技能进行评估，我们就能够回答此类问题。这些信息对于给缺少必要的准备技能的学生制定计划，以及为了适应学生需要而修改教学计划，都是非常有用的。

三、提供恰当的教学

恰当的教学是指把课程内容及教学方法整合进有计划的教学活动中去，以帮助学生取得预期的学习结果。在这个教学阶段，测量与评价是监测学习进步与诊断学习困难的手段。因此，教学中的定期评价就可以提供一种反馈—改正的方法，有助于不断地对教学进行调整以适应个人及群体的需要。

教师可以将许多评价紧密地整合进教学活动，以便能够监控和调整教学。例如，当解决一个科学问题时，教学活动可以以小组合作的方式进行。在小组活动过程中，教师可能会观察到绝大部分时间是埃里克在发言，并操纵着仪器，而小组内多数其他人只是被动的观察者。这样的观察使得教师能够随着教学进展而不断做出调整。另外，小测试或答疑时间可以用于检查单个学生在小组活动中的收获。

四、评价期望的学习成果

教学过程的最后一步是确定学生达到学习目标的程度。这一目的可以通过为测量期

望的学习成果而特别设计的测验或其它评价方法来实现。理想的状况是内容标准和教学目标上就指明了学生应有的变化，以及适合度量和描述这些变化的评价工？具。将一系列的评价方法与预期的学习成果相匹配，突出重要的知识点，这是有效的课堂评价的基本内容，对此我们将在下面的章节中进行详尽的讨论。

五、结果的运用

经常有人认为学生评价实质上服务于教师和管理者的利益。这种观点忽略了评价给学生带来的直接利益。正确使用评价方法，可在以下几方面直接促进学生的学习：（1）阐明期望的学习成果；（2）提供近期目标；（3）提供学习进程的反馈；（4）提供必要的信息，以帮助学生克服学习困难并选择今后的学习内容。虽然教学过程中的定期评价可能是实现这些目标最好的手段，但是对各种期望的学习成果的学期评价也是有必要的。

从精心编制的测验和其它评价方式中所获得的信息也可以促进教学。这些信息有助于教师判断：（1）教学目标的恰当性及可行性；（2）教学资源是否有用；（3）教学方法的有效性。因此，评价方法不仅有助于教学过程自身的改进，也直接有助于学生学习的进步。

当然，评价结果也被用于打分以及向家长汇报学生的进步。正如将在第十二章中讨论的，学生的成长记录袋可以有效地将学生的进步传达给家长，这种方法不仅给传统的分数和成绩赋予了更多的意义，而且还能向父母说明可在哪些领域对学生提供帮助。大量的评价方法的系统运用可以客观、全面地报告每个学生的学习进步。除了打分和报告，评价结果也用于学校的各种管理和指导职能。评价结果有助于课程开发，帮助学生做出教育和职业决定，以及评价学校项目的有效性。图 2.2 的教学模型总结了教学过程的基本步骤，并说明了教、学及评价之间的相互依赖关系。

第五节　评价方法的种类

评价过程可能会涉及到许多不同的方法。依据所采用的参照体系，可以从许多不同的角度对这些方法进行分类和描述。这里我们将呈现那些对于理解和使用评价方法最有用的分类基础。尽管这些类别并非严格分开，但却提供了一个关于评价方法的基本框架，并介绍了一些有用的基本术语。这些用于课堂评价的具体方法将在以后的章节中予以描述和说明。

一、最佳表现与典型表现

根据测量的性质，测验和其它评价方式可以分为两大类：对最佳表现的测量和对典

型表现的测量（Cronbach，1990）。第一类方法用于确定一个人发展的能力或成就。这类方法关心在鼓励学生尽可能取得最好的成绩时，个体所能做出的最佳表现。总之，此类测验结果表明了当做出最大努力时个体的水平。能力倾向测验与成就测验属于这一类。这两类测验的区别通常是就测验结果的使用，而不是测验本身的性质而言的。能力倾向测验主要用于预测个体在将来的学习活动中的成功，而成就测验则用于度量个体在过去的学习活动中成功的程度。

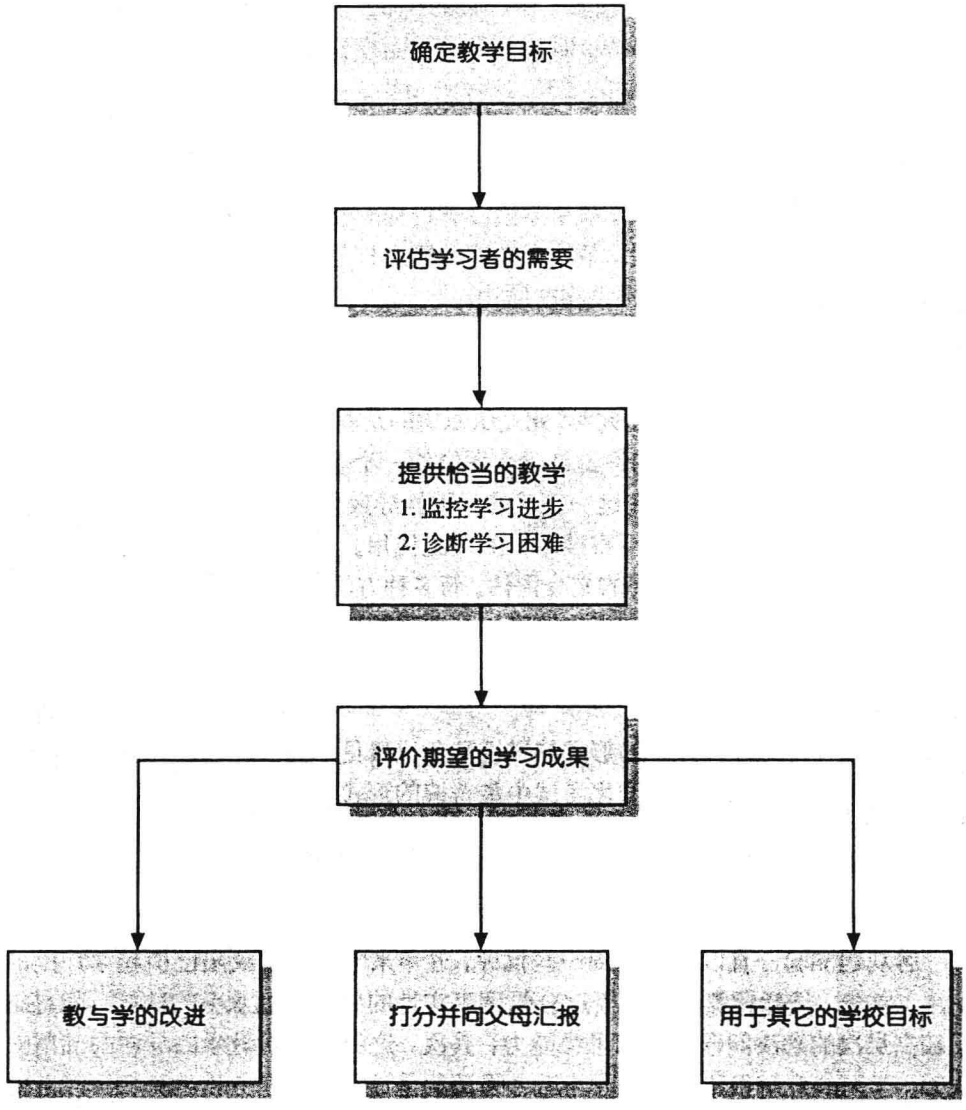


图 8-1-2　简化的教学模型

不过，一些测验可能同时服务于两种目的。那么很明显，它们的区别主要在于侧重点的不同。比如，在课程结束时为测量成就而设计的代数测验，也许就可以用来预测学

生在今后数学学习中的成功。这样的功能重叠妨碍了分类的明确性，但能力和成就这两个术语为能力测量问题的探讨提供了有用的名称。

虽然成就测验的目的是测量“最佳表现”，但只有当学生确实尽到其最大努力时，这一目的才能够真正实现。如果学生没有这样的动机，此类测验的结果显然会低估了其最佳表现水平。因此，最佳表现这一概念只是就评价的意图而言的，而不是指能够从学生的分数中得出可靠的结论。

此种分类中的第二类由那些测量一个人典型表现的方法所组成。这一类方法所关心的是个人“将”做什么，而不是“能”做什么。用于评价兴趣、态度、适应性和各种人格特征的方法都属于此类。此处所强调的是获取具有代表性的表现而不是最佳表现。虽然对典型表现的评价是学生评价的一个非常重要的领域，但这一领域却存在很多困难。测验工具在这一领域的局限性导致其它手段（如面试、问卷、轶事记录法及其它自我报告和观察技术）的广泛使用。单独采用上面的任何一种方法，都不足以实现对典型表现的充分评估。而多种方法相结合，就可以使教师对学生在这些方面所取得的进步和变化做出相当准确的评价。

二、客观题测验和复杂—表现性评价

近年来，对测验和评价的形式的辩论和争议屡见不鲜。在过去的半个多世纪里，选择题已经成为美国标准化考试中最普遍的形式。选择题和其它各种客观题（如判断题和匹配题）都能够使学生在短时间内对较多题目作答，因而都是非常有效率的。同时这些题型可由机器进行记分。评分客观、一定时期内的高信度，以及低成本高效率是这种题型的主要优点。

客观题测验一直以来就受到一些批评，近年来它的一些缺陷已引起了广泛的关注。首先，选择题测验倾向于过分强调事实性的知识和低水平的技能，而忽略了较高层次的解决问题能力和思维能力；其次，这种测验对教学的导向与当前对认知、学习的理解很不相符，因为后者强调的是让学生自己去建构知识，以及他们自己的理解，而不是对零散的事实及程序性技能的积累（Resnick & Resnick，1992）。

20 世纪 90 年代，有一股支持采用一种十分不同的测评方法的情绪猛然高涨，这种方法以扩展性任务和对学生的复杂行为表现的分析为基础。人们期望表现性评价任务能够切实地反映长期的教学目标。这些任务要求学生解决课堂之外的一些重要的问题，或是让学生按照自己认为正确的方法来操作。短文写作（written essay）就是复杂操作任务的一个例子，它比选择测验更能反映出有效交流这一教学目标。这样的题型还包括要求进行扩展性回答的数学问题、在实验室进行的科学实验、一件艺术品的创作、口头表达、设计和作业的展示。

选择测验和复杂表现评价代表了一个连续体的两极，要求做出简短回答的测验处在这两极之间。但是如果只给学生很短的时间，不能自己选择题目，也没有机会修改，那么即使是作文测验也无法实现复杂表现评价的意图。

为了强调在评价学生行为表现的同时，鼓励学生按照自己的判断解决问题和积累学习经验，表现性评价常被称为真实性评价。然而在鼓励学生解决真实的问题的意义上看，并不是所有的表现性评价都是“真实”的。

与选择测验相比，表现性评价的施测及评分都更费时。人的主观评判是评分的一个重要部分，这要求评价者具有高度的专业知识并接受过专业训练。对评价学生成就而言，客观题测验、复杂表现评价以及一系列的中介手段，都是非常有用的。正如我们将在以后的章节中所讨论的，所有的这些评价方法都用得上，而且必须依据特定的评价目的，以及评价对教学的影响来选择合适的方法组合。

三、安置性、形成性、诊断性和总结性评价

测验和其它评价方法也可根据它们在课堂教学中的用途进行分类。这一分类系统遵循的是评价方法在课堂中使用的时间次序（Airasian & Madaus，1972）。这一分类系统如下所列：

1. 安置性评价：在教学开始时确定学生的表现。
2. 形成性评价：在教学过程中检测学生进步。
3. 诊断性评价：在教学过程中诊断学习困难。
4. 总结性评价：在教学结束时评价成就。

尽管有时一种工具可用于多个目的（如，同时用于形成性和总结性评价），但每种课堂评价通常需要使用一些为特定用途而专门设计的工具。

1. **安置性评价**。与学生学习开始时的表现有关，并且总是关注如下问题：（1）学生是否具备了进行下一步学习所需的知识和技能？如，学生的阅读理解能力是否已达到了能够独立阅读一个历史课单元的水平？或者，刚刚开始学习代数的学生是否熟练掌握了关键的几个数学概念？（2）对于下一步教学目标中的理解力和技能，学生已经发展到何种水平？如果学生的理解力和熟练程度已经很高，就表明尽可以跳过某些单元或以更高级的课程代替。（3）学生的兴趣、学习习惯、及个性特征是否表明一种教学模式比另一种更合适（如，小组教学与个别学习相比）？回答这些问题需要运用各种手段：以往成绩的记录，对课程目标的预测，自陈量表，观察技术等。总之，安置性评价的目的就是确定每个学生在教学进程中的位置、以及最有效的教学模式。

2. **形成性评价**。用以监测教学过程中的学习进展，其目的是为学生和教师提供关于学会与否的连续的反馈。给予学生的反馈强化了正确的学习，并可以发现具体的学习

错误和需要改正的错误观念。教师得到反馈后，就可以调整教学，更好地指导小组和个人的工作。形成性评价极大地依赖于为每个教学部分（如单元、章节）特别准备的测验和评价。形成性评价中所使用的测验和其它评价任务大都是由教师自编的，不过也可以向教科书或其它教学材料的出版商预定测验。当然，观察法对监测学生的进步、识别出学习错误也是很有用的。由于形成性评价是直接用于改善教学的，所以其结果通常不用于打分。

3. **诊断性评价。**诊断性评价的专门化程度很高。它与那些顽固不化或反复出现的学习困难有关，而这些学习困难是形成性评价所不能解决的。如果一名学生在教师调整了教学方法之后，还是在阅读、数学或其它科目上不断受挫，那么这就表明需要对该学生进行更加仔细的诊断。运用一个医学上的类比，可以说形成性评价为简单的学习问题提供急救治疗，而诊断性评价则寻求急救治疗不能奏效的根本原因。因此，诊断性评价更加全面而细致。它不仅要使用特殊的诊断测验，还需要运用多种观察技术。严重的学习障碍可能还需要教育、心理以及医学专家的帮助，而鉴于诊断的恰当性，还要为该学生制定个别教育计划。诊断测验的目的就是查明那些持久的学习问题的成因，并且制定出矫正计划。

4. **总结性评价。**总结性评价通常在教学课程（或单元）结束的时候进行，被用来确定教学目标达成的程度。主要是用于给学生的表现打分，或证明学生对预期的学习目标的掌握情况。总结性评价中所使用的方法是由教学目标决定的，它们通常包括教师编制的成就测验，对各种行为表现的评定（如实验、口头报告），以及对作品的评价（如作文、绘画、研究报告）。这些关于学生成就的各种信息资源可系统地收集到成长记录袋中，以总结或展示学生的成绩和进步。虽然总结性评价的主要目的是评分，或证明学生的成就，但它也为判断课程目标是否恰当、教学是否有效提供了信息依据。

四、常模参照与标准参照测量

对测验和其它评价结果的解释也为评价的分类提供了依据。解释学生表现情况的基本方法有两种。常模参照解释描述的是学生的表现在已知群体中的相对位置（如，打字水平优于班中90%的人）。标准参照解释描述学生具体的行为表现（如，每分钟打40个字，一个错也不出）。如果将解释限定为获取具体目标（如，用大写字母书写一些名词），它也可称为目标参照。目标参照属于标准参照的一种，但它所包含的任务范围不如标准参照所采用的任务范围广泛。对这几个概念更为具体的界定请见下面的框图。

目前，基于标准的评价是标准参照解释中最主要的一种。设计与具体内容标准相匹配的基于标准的评价方法，就可以根据确定的行为表现标准，在少量的几个行为表现水平上进行报告。比如，评价结果可在3～5个表现水平上进行报告，如较熟练、熟练和

精通三个水平。学生是否达到熟练标准与其他同学的表现无关。相反，这一参照系只与划定的熟练标准的效标或分数线有关。

术　语　表

常模参照评价：根据个体在某个特定群体中的相对位置，来解释个人表现的测验或评价方法。

标准参照评价：依据明确界定的学习任务范围来解释个体行为表现的测验或评价方法。

还有一些述语用得较少，不过与“标准参照”意义相近：基于标准、目标参照、内容参照、范围参照和一般参照。

常模参照解释依靠的是将学生的表现与常模群体的平均表现作比较。依据评价结果的不同用途，可以选择依据当地、州或全国的常模。例如，使用全国常模时，我们可将某学生在一项词汇测验上的表现描述为“达到或超过全国六年级样本中76%学生的水平”。标准参照解释可以有不同的形式。例如，我们可以：（1）描述学生所能胜任的具体的学习任务（如从1数到100）；（2）指出学生正确完成任务的百分比；（3）将测验中的表现情况与一套表现标准相比较，判断是否达到既定标准（如熟练操作）。

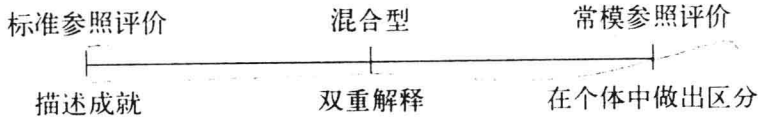
尽管在这两种解释方法中都使用了“百分比”一词，但其含义是截然不同的。通过确定群体中与被评个体得分相同或更低者所占的百分比（称做百分位数），常模参照解释指出了学生在群体中的相对位置。标准参照解释则关注答对题目的百分比（正确率）。百分位数和正确率之间的差别体现了常模参照解释和标准参照解释之间的根本差异，因而是非常重要的。

严格地讲，“常模参照”和“标准参照”只是就解释结果的方法而言的。不过，当测验和其它评价方法是专门为适合某一种解释而设计时，这种对解释作明确区分的意义和价值就体现出来了。因此，使用常模参照和标准参照将测验和评价方法分为两大类别是合理的。

仅仅通过观察测验本身（见框图“常模参照测验和标准参照测验的比较”），就想把为厂适合一种解释类型而特别编制的测验识别出来是不可能的。只有在编制和使用测验的过程中，我们才能注意到它们的区别。常模参照测验的一个特征，是它选择的都是中等难度的题目，并且将所有学生都能答对的题目剔除掉。在这一方法下，分数的分布范围很广，这样就可以将学生们不同的学绩水平区分开来，从而有助于根据学生的相对成就水平做出决策，如选拔、分组和相对分数的评定等。与此相反，标准参照测验中的题目均与所测学习成果直接相关，但不一定能区分出学生的不同水平，同样也不必删除容

易的题目或改变题目的难度。如果学习任务简单，那么就设计简单的题目。标准参照测验的目的在于对各个学生特定的知识和技能进行描述，这些信息将有助于计划小组及个别化的教学。

我们最好将这两种评价看做一个连续体的两极，而不要把它们截然分开。正如下图所示的连续体，标准参照测验强调对成就的描述，而常模参照测验强调在个体中做出区分。



为了兼容二者的优点，测验编制者尝试将常模参照测验编制得更具有描述性，这样就可以同时采用常模参照和标准参照两种解释。同样的，测验编制者也已经为标准参照解释的测验加入了常模参照解释。在已出版的测验中，双重解释的测验越来越多，许多测验已移向连续体的中央。虽然这会导致测验编制上的折衷趋势、以及测验解释中的谨慎，但是变通性的提高有助于测验使用效率的改善。

第六节　评价种类小结

表 2.1 中对课堂测验及其它评价方法进行了基本的描述。在以后的章节中将进一步对这些评价种类进行讨论。

其它描述性术语

下列用于描述测验的术语以对比的方式出现，它们实际上是某个连续体的两端（如速度与难度测验）。

非正式测验与标准化测验。非正式测验是由任课教师编制的，而标准化测验是由测验专家设计，并在标准的条件下实施、评分及解释的。

个别测验与团体测验。一些测验以口头提问的方式一对一地施测（如个别智力测验），而另一些测验可对一群人施测。

掌握测验与调查测验。一些成就测验测量学生对有限的学习成果的掌握程度，而另一些则在大范围内测量学生的一般成就水平。虽然，掌握测验通常采用标准参照解释，而调查测验则较多强调常模参照解释，但对那些精心设计的调查测验，也可进行一些标准参照解释。

常模参照测验（NRTs）和标准参照测验（NRTs）的比较

NRTs 与 CRTs 的共同点

1. 都需要说明要测量的成就范围。

2. 都需要一个恰当的且具有代表性的测验题目的样本。

3. 都使用相同的题型。

4. 都使用相同的题目编写原则（除了题目难度）。

5. 都采用同样的质量评价指标（信度和效度）。

6. 都可用于教育评价。

NRTs 与 CRTs 的区别

请记住，区别只在于强调的重点不同。

1.NRTs：通常测量较大范围的学习任务，而对每个具体任务只用几个题目测量。

CRTs：通常关注于有限范围内的学习任务，而对每个具体任务采用相对较多的题目测量。

2.NRTs：强调用相对学习水平对个体做出区分。

CRTs：强调对个体能否完成学习任务进行描述。

3.NRTs：多采用中等难度的题目，通常不用非常容易和非常难的题目。

CRTs：题目难度与任务难度相匹配，而不改变题目难度或删除过易和过难的题目。

4.NRTs：结果的解释需要一个定义明确的群体。

CRTs：结果的解释需要一个有限的具定义明确的成就范围。

建构测验与选择测验。一些测验要求受测者提供答案（如论述题测验），而另一些则要求他们从已知选项中选出正确答案（如选择题测验）。

速度测验与难度测验。速度测验测量个体在指定时间内所能完成的题目数，而难度测验测量受测者在充裕的时间内所能达到的行为水平。难度测验中的题目，按照由易到难的顺序排列。

客观测验与主观测验。在客观测验中，能力相当的受测者将得到相同的分数（如选择题测验）。而在主观测验中，分数受到评分者的观点或主观判断的影响（如论述题测验）。