



提升教师专业实践能力译丛

首席学术顾问◎顾泠沅
主 编◎胡庆芳 程可拉

邓妍妍 彭春艳◎译
程可拉◎校

有效的课堂 评价手册

〔美〕罗伯特·J.马扎诺 著
(Robert J. Marzano)

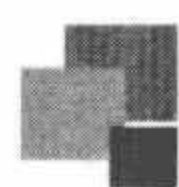
旁征博引诠释课堂评价的方法与技术 绘声绘色解说课堂评价的实践与案例

Classroom Assessment & Grading that Work



教育科学出版社

Educational Science Publishing House



提升教师专业实践能力译丛

首席学术顾问◎顾泠沅
主 编◎胡庆芳 程可拉

邓妍妍 彭春艳◎译
程可拉◎校

有效的课堂 评价手册

[美] 罗伯特·J.马扎诺 著
(Robert J. Marzano)

旁征博引诠释课堂评价的方法与技术 绘声绘色解说课堂评价的实践与案例

教育科学出版社

· 北 京 ·

责任编辑 王峥媚
责任校对 徐虹
责任印制 曲凤玲

图书在版编目 (CIP) 数据

有效的课堂评价手册/ (美) 马扎诺著; 邓妍妍, 彭春艳译. —北京: 教育科学出版社, 2009. 6

(提升教师专业实践力译丛/胡庆芳, 程可拉主编)

书名原文: Classroom Assessment & Grading that Work
ISBN 978 - 7 - 5041 - 4311 - 2

I. 有… II. ①马…②邓…③彭… III. 课堂教学—教学
评议—中小学—手册 IV. G632. 421 - 62

中国版本图书馆 CIP 数据核字 (2009) 第 043666 号

北京市版权局著作权合同登记 图字: 01 - 2007 - 3719 号

出版发行 教育科学出版社

社址	北京·朝阳区安慧北里安园甲 9 号	市场部电话	010 - 64989009
邮编	100101	编辑部电话	010 - 64989449
传真	010 - 64891796	网 址	http://www.esph.com.cn

经 销 各地新华书店

印 刷	莱芜市圣龙印务有限责任公司	版 次	2009 年 6 月第 1 版
开 本	700 毫米×1 000 毫米 1/16	印 次	2009 年 6 月第 1 次印刷
印 张	14	印 数	1 - 6 100 册
字 数	195 千	定 价	26.80 元

如有印装质量问题, 请到所购图书销售部门联系调换。

提升教师专业实践力： 教师专业化的核心价值与永恒魅力

自 20 世纪 60 年代以来，教师专业化开始成为一种强劲的思想浪潮。吹响教师专业化的号角无疑是在 1966 年，联合国教科文组织和国际劳工组织在“关于教师地位的建议”中庄严陈述：“应把教育工作视为专门的职业，这种职业要求教师经过严格的、持续的学习，获得并保持专门的知识 and 特别的技术。”

自 20 世纪 80 年代以来，教师专业化运动开始从追求教师职业的专业地位与权利向教师的专业发展转向，教师的专业发展开始越来越清晰地成为教师专业化的方向和主题。人们越来越明确地认识到，提高教师专业地位的有效途径是不断改善教师的专业教育，从而促进教师的专业发展。只有不断提高教师的专业水平，才能使教学工作成为受人尊敬的一种专业，成为具有较高社会地位的一种专业。其间一系列的报告会聚成了时代的主旋律：1986 年，美国卡耐基基金会和霍尔姆斯小组相继发表“国家为培养 21 世纪的教师作准备”和“明天的教师”两个报告，明确提出以教师的专业性作为教师教育改革和教师职业发展的目标；在 1989—1992 年的四年间，经济合作与发展组织（OECD）发表了诸如“教师质量”等一系列有关教师教育改革及教师专业发展前景的研究报告。

教师专业化的浪潮极大地推动了世界许多国家教师教育新理念和体制的建立。在“教师专业化”从理念到实践的广泛而深入的推进历程中，世界上大多数发达国家和地区的教师教育先后经历了从中等教育水平的师范学校教育到高等教育程度的师范学院教育，

从师范学院的独立培养到综合大学的本科教育以及大学后专门的教育课程培养的转变，并逐步形成了教师教育完整学位序列的培养体制。

但是，理念与行动、愿景与现实总是存在着不尽如人意的差距。2006年9月18日，美国哥伦比亚大学教育学院前院长阿瑟·莱文（Arthur Levine）领衔的“教育学院计划”（The Education Schools Project）历时四年，调查了全美1206所教育学院，最后正式发表研究报告“中小学教师的教育”（Educating School Teachers），其“美国教师的摇篮千疮百孔”的隐喻式结论石破天惊，在美国掀起轩然大波，并引起全世界对教师教育的关注和反思。莱文在报告中尖锐地指出，尽管越来越多的证据显示了教师质量的重要性，但美国教师教育“与课堂现实脱节”，大多数教育学院抓不住问题的要害，课程混乱，教职员远离课堂，未能“跟上人口、技术以及全球竞争的变化步伐，无法应对提高学生学习成绩的压力”。

问题的严重性也许还不止于此。美国国家教育统计中心2005年的数据显示，接近三分之一的新教师在他们开始从教的第一个三年任期内离职，在第一个五年任期内的离职率甚至达到50%。其中主要的原因就包括许多在校任职的新教师因无法胜任教学工作而不得不离开教师岗位。中国的情况也同样不容乐观。据2007年广东省对本省中西部的调查，25.1%的新教师有机会就会选择“跳槽”，9.8%的新教师迫切想要离开本单位。其中一个重要的原因也是教师在职专业发展的机会与能获得的支持不够，不能胜任学校分配的工作。放眼全球，教师专业发展的策略与支持毫无疑问成为了世界各国面向21世纪应对本国教师质量提高的必答题。

本着促进教师专业实践力提升的良好初衷，我们在繁荣的国际学术出版市场上努力精选了一些理论创新性和实践指导性兼优的精品著作，力求及时高质量地翻译和推介给国内同行。首批将推出《提升专业实践力：教学的框架》《有效的课堂教学手册》《有效的课堂管理手册》《有效的课堂评价手册》《有效的课堂指导手册》《有效的教师领导手册》《创造有活力的学校》和《优化测试，优化

教学》。

非常感谢教育科学出版社对我们过去所做的比较教育研究工作的密切关注以及对承担本译丛主编工作的充分信任！我们也非常荣幸地邀请到了全国著名的教育家顾泠沅教授、著名的比较教育学专家李其龙教授、课程与教学论专家高文教授、著名的基础教育研究专家周卫研究员，以及上海市教育科学研究院教师发展研究中心特聘研究员张玉华女士，出任本译丛的学术顾问。

我们还要衷心感谢各位译者在繁忙的教学科研之余，保持高度的合作热情，远离浮躁与功利，安心于书斋，孜孜不倦于对异域文化和思想智慧的传递！

我们真诚地期望正在阅读这套译丛的广大专家和同行多与我们联系，给我们提出宝贵的意见和建议，Closetouch@163.com 永远期待着您智慧的声音！

胡庆芳 程可拉

目 录

第一章	课堂评价案例 / 1
第二章	国家课标的作用 / 13
第三章	测量连续性学习的量表 / 31
第四章	课堂评价的设计 / 61
第五章	促进学习的评价方法 / 95
第六章	最后综合分数与等级 / 111
第七章	成绩单与基于标准或主题的教育前景 / 133
术语注释 / 153	
附录 A:	半分分值与整数分值的对照 / 177
附录 B:	语言艺术测量主题样本的评分细则 / 179
参考文献	/ 189
作者简介	/ 204
译后记	/ 205

图 表 目 录

表序	表名
表 1.1	教师课堂评价技能的提高对学生成绩的影响 / 4
表 1.2	不同反馈类型的效果 / 5
表 1.3	与 15 个星期的评价次数相关的提高 / 10
表 2.1	测量主题“要旨阅读”的协变范畴 / 23
表 2.2	测量主题样本 / 25
表 3.1	五位教师在不同题项上的评分分布 / 33
表 3.2	在 100 分量表中五位教师给十位学生的总分 / 33
表 3.3	教师 2 和教师 3 是如何给学生 2 评分的 / 35
表 3.4	正态分布 / 41
表 3.5	三个学生的答题情况 / 42
表 3.6	代表测量主题进步的评分量表 / 43
表 3.7	100 分的分布 / 44
表 3.8	简化版评分量表 / 47
表 3.9	简化版评分量表的快速参考指南 / 48
表 3.10	简化版量表的评分流程图 / 49
表 3.11	完整版评分量表 / 51
表 3.12	简化版和完整版评分量表之间的关系 / 52
表 3.13	完整版量表的评分流程图 / 53
表 3.14	具体的测量主题评分量表 / 56
表 3.15	适用 NAEP 表现层次的完整版量表 / 58

表序	表名
表 4.1	9 周内对 6 个测量主题进行评价的总计划 / 62
表 4.2	基本细节 / 68
表 4.3	规律与原理 / 69
表 4.4	运用于三类信息题和任务的认知过程 / 71
表 4.5	一类题目和任务的心理程序 / 74
表 4.6	总体宏观程序 / 76
表 4.7	强迫选择题的类型 / 80
表 4.8	林肯一道格拉斯辩论节选 / 84
表 4.9	知识型、题型和任务型及评价表格 / 87
表 4.10	生活技能主题的一般评分量表 / 91
表 5.1	学生进步图表 / 96
表 5.2	班级图表 / 97
表 5.3	学生版评分量表 / 99
表 5.4	一段时间内一个学生在一个主题上的得分柱形图 / 101
表 5.5	含有平均分数线的得分柱形图 / 101
表 5.6	对学习的幂函数的描写 / 103
表 5.7	包含幂函数分的柱形图 / 103
表 5.8	观察分数、平均分数与幂函数算得分数之比较 / 104
表 5.9	供证据收集法使用的成绩簿 / 106
表 6.1	供输入测试成绩的品尼高加法系统表 / 113
表 6.2	最合适的分数 / 115
表 6.3	品尼高加法系统分数报告单 / 118
表 6.4	测量主题一览表 / 119
表 6.5	生活技能主题作业完成情况的评分量表 / 122
表 6.6	课堂小测验的评分量表 / 124
表 6.7	分配给每个主题的权数 / 126
表 6.8	加权平均值 / 127
表 6.9	两个联合评级系统 / 131
表 7.1	含有等级成绩和主题分数的成绩单 / 135

表序 表名

表 7.2 不含字母等级成绩的成绩单 / 137

表 7.3 传统科目和相关的跨课程主题 / 140

表 7.4 成绩单样本 / 141

表 7.5 巧克奇的学生获得的全国百分位分 / 151

表 7.6 阿拉斯加州中学质量和基准考试通过率 / 151

表 TN 5.1 A 模拟学习的函数 / 163

表 TN 5.1 B 各种函数的方差百分比 / 163

表 TN 6.1 基于顺序和时间的绩效 / 166

第一章 课堂评价案例

从19世纪90年代初期开始，提高中小学学生的学术成绩就一直是美国教育工作者关注的重点，当时工业上的领军人物、政治家、家长和社会普遍认为受教育的人数对一个国家的美好未来而言是最重要的保证（Ravitch, 1983）。从那以后，美国进行了一系列大范围的教育改革试验，所有的改革都是为强化学生的成绩而设计的。教育工作者进行的试验有课程改革、降低生师比、提高技术的可行性和使用率，等等。这些改革都取得了成效。但是，即使是最好的改革也无法体现出在教育体制上对学生成绩影响最大的重要变量——任课教师的重大作用。

事实上，对任课教师在教育学生过程中的作用的每一项研究都得出相同的明确结论：一个教学效果好的教师比学校教育的其他任何可控因素更能强化学生的学习。为了更好地说明这一点，保罗·怀特、桑德拉·鸿和威廉·桑德斯（S. Paul Wright, Sandra Horn, William Sanders, 1997）等研究员经过分析6万多名三年级至五年级学生的测验分数，得出了以下观察意见：

这个研究结果证明，影响学生学习的最重要因素是教师。同时，结果还表明了教师教学效果上存在很大变异。这个结果清楚地表明，与其他单一因素相比，我们能够更多地通过提高教师的教学效果来改进教育。教学效果好的教师对各种水平的学生似乎都有效，而无论这些学生在课堂上的水平差异有多大。假如学生在教学效果差的教师指导下，不管他们的学习成绩是否相同，学业进步都很小。（p. 63）

其他研究也证明了怀特、鸿和桑德斯的结论（参见 Nye, Konstantopoulos, Hedges, 2004）。卡缇·黑罗克（Kati Haycock, 1998）通过比较一个学生分别与“教学效果最好的教师（most effective teacher）”和“教学效果最差的教师（least effective teacher）”在一起一年时间有什么进步来生动地表达任课教师的影响（要讨论“教学效果最好的教师”和“教学效果最差的教师”怎样定义，参见术语注释 1.1）。黑罗克解释说，教学效果最好的教师在学生的成绩上能提高 52 个百分点（percentile points），而教学效果最差的教师在成绩上只能提高 14 个百分点，相差了 38 个百分点。人们发现，据估计，学生仅从长大一岁以及日常生活中获取新知识和技术就可在学业成绩上大约获得 6 个百分点（Cahen, Davis, 1987; Hattie, 1992）。教学效果差的教师几乎比不上学生自己的生活经历。

由于对任课教师的影响潜能给予如此强烈的肯定，按照逻辑提出一个问题：教学效果好的教师能做什么？答案有很多，大多数集中在教学和管理策略上（参见 Hattie, 1992; Marzano, Pickering, 2003; Marzano, Pickering, Pollock, 2001）。这些答案强调了策略的运用，如精心设计实践活动、对比活动和交际学习目标，以及使用图像、图表和图片来展示知识。事实上，尽管“高产（high-yield）”的教学策略和课堂管理策略是高效教学的关键部分，但该书只强调教学中在讨论如何提高学生成绩的途径中常常被忽视的一个方面：课堂评价（classroom assessment）。

使一些教育工作者惊讶的是，课堂评价效果研究的主要结论表明，课堂评价可能是教师手中最有力的武器。为了说明这一点，保罗·布拉克和戴兰·威廉（Paul Black, Dylan William, 1998）综合了 250 多项研究的结论，把有效的课堂评价影响描述如下：

这里所报告的研究归纳显示，形成性评价确实促进学生学习。成绩的提高看起来相当大，在教育方面至今所报告的成绩提高仍是最大的。作为展示这些成绩有多大提高的指标，“效应值（effect size）”为 0.7（参见描述效应规模的术语注释 1.2），假如能在全国范围内获得这个效应，

它将等同于把一个“中等”国家，如英国、新西兰或美国的数学成绩分数提高到继太平洋圈内的新加坡、韩国、日本和中国香港之后的“前五名”。(p. 61)

值得关注的是，布拉克和威廉（1998）的评论都赞同形成性（formative）评价，反对终结性（summative）评价。下一节再讨论这个区别。为了更好地理解布拉克和威廉的结论，参见表 1.1（参见描述如何获得表 1.1 的术语注释 1.3）。该表的上部分记述了运用课堂评价技巧排在第 50 百分点位的一位教师，以及她班上一位成绩排在第 50 百分点位的学生。随后，她在运用课堂评价技巧上的教学效果提高到第 84 百分点位。按照布拉克和威廉的结果，可以预测，该学生的成绩将能提高到第 63 百分点位。该表的下部分展示了更为戏剧性的情形。假如教师运用课堂评价的技巧从第 50 百分点位提高到第 99 百分点位，可以预测学生的成绩会提高到第 78 百分点位。

表 1.1 中报告的结果是很有意义的——课堂评价对学生的成绩有很大的影响。按照这个结果，人们可以得出这样的结论：更多地评价学生会促进其自主学习。但这样的结论可能是错误的。像教育中的许多方面一样，课堂评价只能在一定的条件下促进学生的成绩。有幸的是，研究提供了有关这些条件的指南。

对课堂评价研究的简要评论

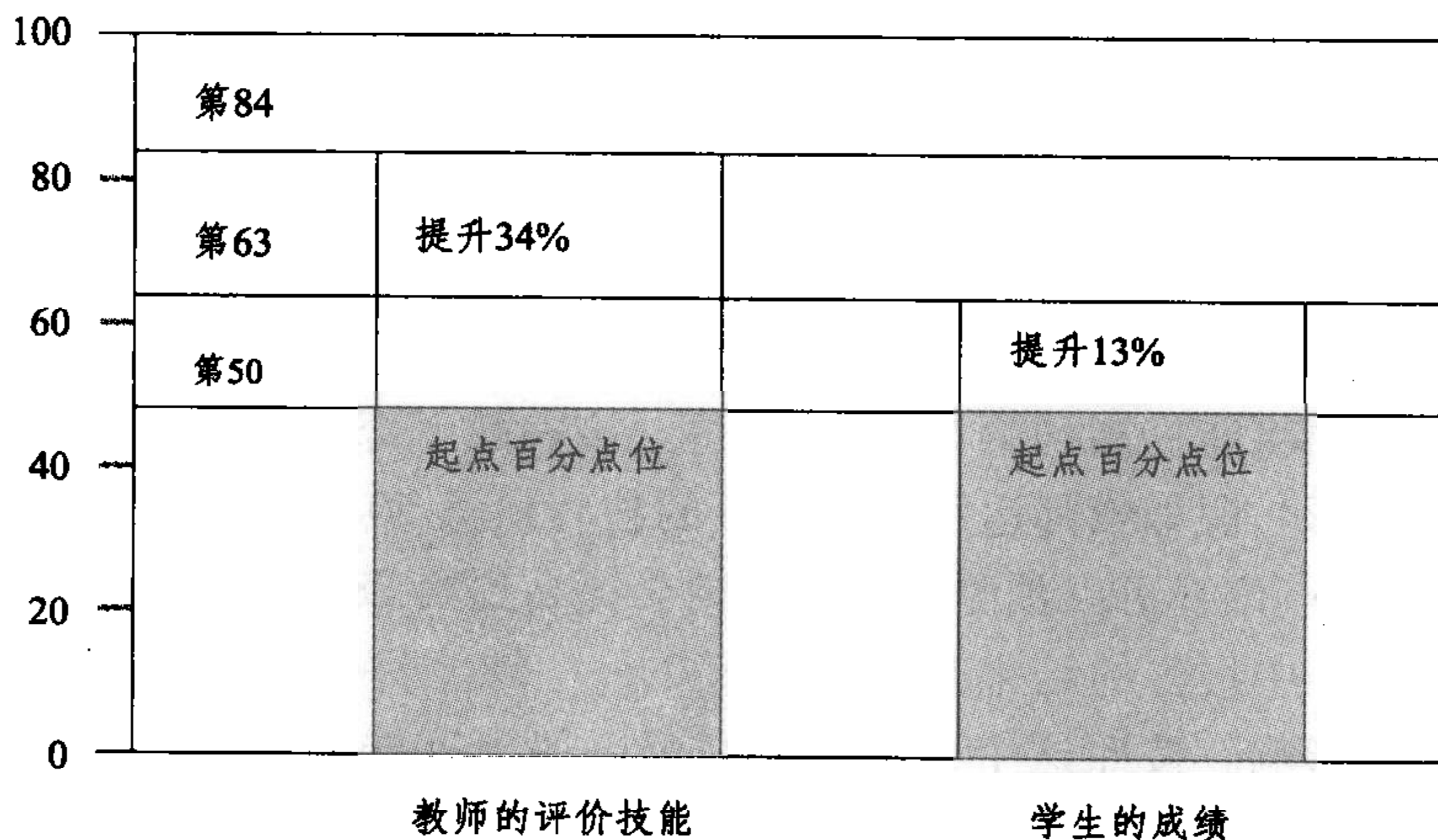
学者们对课堂评价研究做了很多评论（Natriello, 1987, Fuchs, Fuchs, 1986; Crooks, 1988; Bangert-Drowns, Kulik, Kulik, 1991; Bangert-Drowns, Kulik, Kulik, Morgan, 1991; Kluger, DeNisi, 1996; Black, William, 1998）。这些评论引发了许多结论，为有效的课堂评价拓宽了视野；但本书只着重提出四个原则：

- 课堂评价的反馈应该为学生清晰地描述学习目标上的进步和怎样再提高。
- 课堂评价的反馈应该鼓励学生进步。

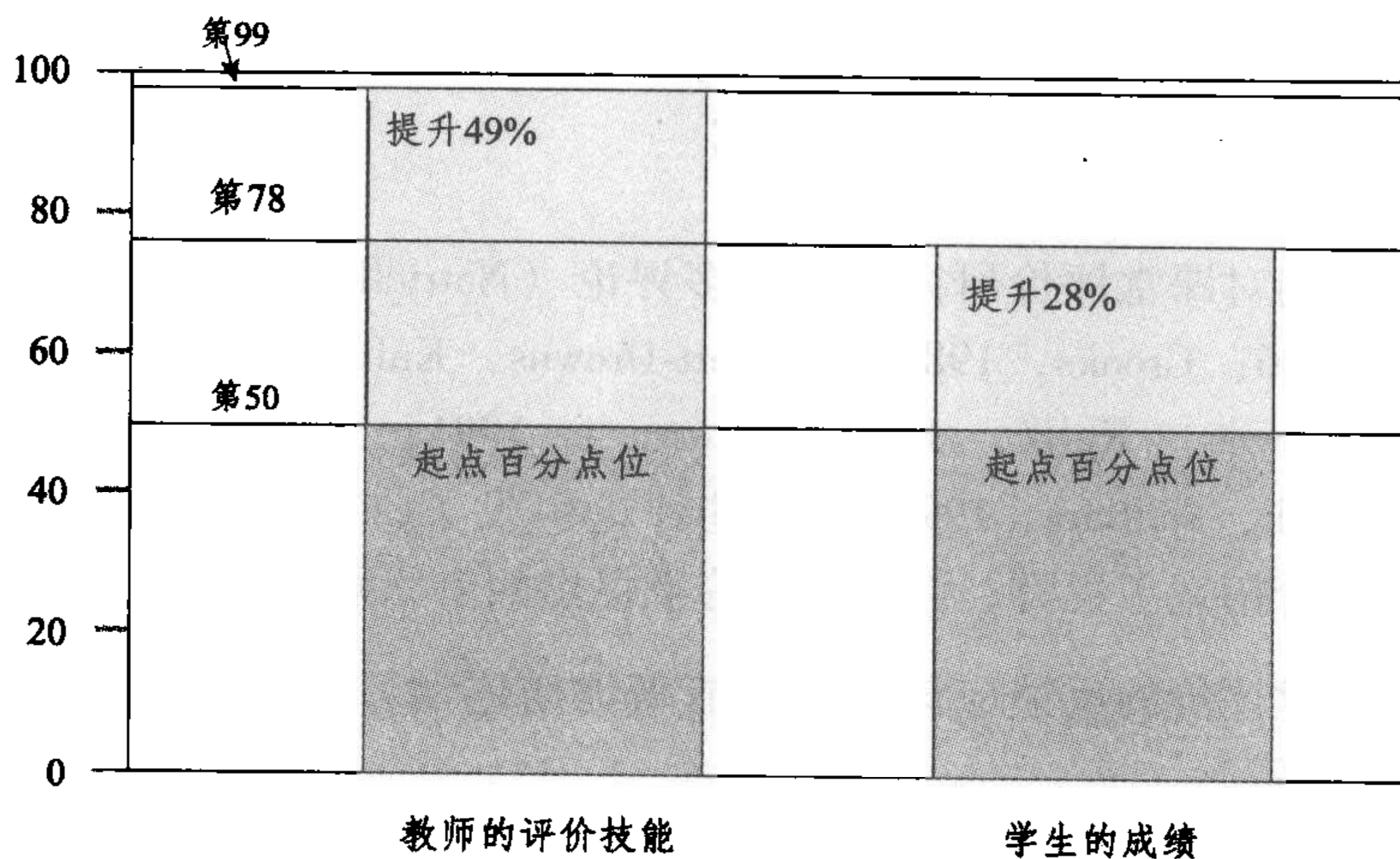
- 课堂评价在本质上应该是形成性的。
- 形成性课堂评价应该是经常性的。

表 1.1 教师课堂评价技能的提高对学生成绩的影响

当教师的课堂评价技能从第 50 百分点位提高到第 84 百分点位时，预测学生成绩的提高。



当教师的课堂评价技能从第 50 百分点位提高到第 99 百分点位时，预测学生成绩的提高。



对学生学习进步和提高提供清晰的描述

在基础层面，课堂评价是一种涉及学生进步的反馈（feedback），它能解释反馈如何强化学习的理由。根据约 8 000 例研究的评论结果，研究员约翰·哈迪尔（John Hattie, 1992）做了如下的评述：“提高学习成绩的最有力的改进方法是反馈。改进教育的最简单的处方必须是‘一定程度的反馈’”（p. 9）。

正如哈迪尔所说的那样，并不是所有形式的反馈都同样有效。事实上，有些形式的反馈可能会妨碍学习。为了说明这一点，要看表 1.2 显示的研究成果。该表展示了来自两个元分析（meta-analytic）研究的成果，一个是罗伯特·班格特德劳恩斯、陈林·库利克、詹姆斯·库利克以及玛丽·特里萨·摩根（Robert Bangert-Drowns, Chen-Lin Kulik, James Kulik, Mary Teresa Morgan, 1991）的研究，他们观察了 40 例课堂评价；另一个是林恩·富克斯和道格拉斯·富克斯（Lynn Fuchs, Douglas Fuchs, 1986）的研究，他们观察了 21 例课堂评价。从表 1.2 这两个综合研究中得出的结论能帮助人们理解有效课堂评价的第一原则。

表 1.2 不同反馈类型的效果

来 源	课堂评价的反馈特点	研究数量*	效应规模	学生成绩得失百分点位
Banger-Drowns, Kulik, Kulik, Morgan (1991)	正/误	6	-0.08	-3
	提供正确答案	39	0.22	8.5
	学生理解和不理解的标准	30	0.41	16
	解释	9	0.53	20
	重复到正确为止	4	0.53	20
Fuchs, Fuchs (1986)	用图表展示结果	89	0.70	26
	用规则评价（阐释）	49	0.91	32
* 表明研究员计算效应规模而验证的研究数量。参见术语注释 1.2				

表 1.2 的前五行是从罗伯特·班格特德劳恩斯、陈林·库利克、詹姆斯·库利克以及玛丽·特里萨·摩根的综合归纳开始的。第一行表明，若学生得到课堂评价的反馈仅仅是告诉他们答案的对错，学习就会受到负面的影响，被扣去 3 个百分点。但是，当学生的答案得到肯定时，学习的效应是正面的。如第二行所示，这个做法可使学生的成绩提高 8.5 个百分点。

表 1.2 的第三行指的是学生是否清楚用于判断他们答案的标准。评分标准表达清晰占学生成绩的 16 个百分点。第四行报告了一个特别有趣的结果——假如学生能解释为什么他们的答案对或错，在成绩上就能增加 20 个百分点。第五行表明，让学生反复回应评价，直到他们能正确回答问题，就可获得 20 个百分点。

表 1.2 的第六、七行来自林恩·富克斯和道格拉斯·富克斯（1986）的研究。第六行用图解的形式展示了对学生结果的影响。正如我们将在以下的章节中看到的，用图解的形式展示评价效果能长期帮助学生控制他们自己的学习。而且，这种做法也能帮助教师更准确地判断学生的理解力和技能水平，在学生成绩中获得 26 个百分点。很有可能的是，看到用图解展示的学生分数会给教师提供更精确具体的参数，从而决定下一步的教学。最后，第七行说明教师阐释评价结果的方式。假如用一套“规则”来阐释的话，学生的成绩会提高 32 个百分点。在第三章我们将深入探讨这个问题。简言之，这个结果指的是一所学校或一个学区的教师应该有严格和统一的方法来阐释课堂评价的效果。

鼓励学生进步

研究文献中有一点令人困惑，传达给学生的反馈方式极大地影响着它对学生的成绩起正面还是负面作用。这是阿夫拉姆·克卢格和安杰洛·德尼西（Avraham Kluger, Angelo DeNisi, 1996）所进行的元分析的主要结论之一。分析了涉及 23 000 多名学生的 607 项实验/控制比较后，研究人员发现，在他们检验的研究中，有 33% 项是反馈对成绩起了负面作用。他们用一个随意的因素验证这个似是而

非的效果：究竟反馈会鼓励学生还是给学生泼冷水。克卢格和德尼西发现，当评价反馈打击学生时，得到的效应规模是负 0.14，这转换成学生成绩即降低了 5.5 个百分点（参见讨论效应规模的术语注释 1.2）。

当然，这个结果中的关键问题是，什么构成了鼓励型（encouraging）反馈和打击型（discouraging）反馈？克卢格和德尼西告诫说这个问题没有简单的答案，但研究提供了一些强有力的指导。要理解鼓励型反馈和打击型反馈的作用，我们必须考虑动机理论（motivation theory）的两个基础——激励（drive）理论和归因（attribution）理论。

激励理论要求人类的大多数动机必须能被解释为两种竞争力的作用，激励——既能争取成功也会担心失败（Atkinson, 1957, 1964, 1987; Atkinson, Raynor, 1974）。人们逐渐形成了不同的激励趋势——向往成功（success oriented）或避免失败（failure avoidant）。当这些趋势形成习惯后，它们会转变成对新任务——特别是对学生有挑战性的任务的强烈期待。

向往成功的学生一般会被挑战所激励，因为他们期望伴随着成功的积极情感。避免失败的学生会因挑战而泄气，因为他们预测到了伴随失败的消极情感。事实上，避免失败的学生可能会运用阿 Q 精神（self-handicapping strategies 自我缺陷策略），使他们的失败显得有理由，而不是由于缺乏能力。这些策略包括拖延（procrastination）（Rothblum, Solomon, Murakami, 1986; Solomon, Rothblum, 1984）；制定无法实现的目标，从而使失败得到认可（Snyder, 1984）；以及承认一些不足（weaknesses）或缺陷（handicaps）来为失败寻找借口——建构“学术木头腿（academic wooden leg）”（Covington, 1992; Covington, Omelich, Schwarzer, 1986）。

归因理论从另一个角度阐释了鼓励型反馈和打击型反馈。它认为学生看待失败和成功的态度会鼓励或打击他们（Weiner, 1972, 1974; Weiner et al., 1971）。通常，人们会把成功归结为四个原因：能力（ability）、勤奋（effort）、幸运（luck）和任务的难度（task

difficulty)。其中，勤奋是最具鼓励性的。正如马丁·科温顿（Martin Covington, 1992）所解释的：

归因理论最重要的特点之一是它强调勤奋在成绩中的作用。有几个理由能证明勤奋的重要。其一，假如学生相信他们的失败是因为不够努力，他们对未来的成功会抱有更加乐观的态度。其二，勤奋会提高成功的自豪感（pride），并补偿失败的内疚感（guilt）。另外，最重要的是，强调勤奋在成绩中的作用得到了证实，因为人们普遍认为学生的勤奋是随教师的行动而变化的。（p. 16）

在归因理论和激励理论之间存在着非常直接的关系。向往成功的学生尤其会相信勤奋的作用。他们期待勤奋努力会带给他们成功。因此，即使要面对挑战性的任务，他们也能找到通往成功的道路。归因理论中一个更有鼓励性的方面是，那些不相信勤奋可以获得成功的学生会随着时间的推移而了解他们能够成功。马丁·塞利格曼（Seligman, 1975；Seligman, Maier, Greer, 1968）强调，假如学生亲身体会过勤奋使他们成功的经历，他们甚至能养成一种勤奋倾向（effort oriented）的风格。塞利格曼把这种动力称之为“学习的乐观主义（learned optimism）”。

激励理论和归因理论对为什么评价反馈会鼓励一些学生而打击另一些学生做出了有道理的解释。假设对一个学生的评价很糟糕，如果学生是个失败的回避者，负面的结果将使更加相信他无法完成有挑战性的任务，消极的情感会伴随着这些任务。这种情况对学生的打击最大。然而，假如学生是成功倾向的话，测验中不理想的表现则不会令他感到沮丧，因为他有改进的策略——更勤奋地学习。

总而言之，激励理论告诉我们，鼓励型的课堂评价必须具备两个特点。首先，教师必须引导学生理解，分数低并不意味着失败。反之，回避失败的学生若得不到高分就会不断地失去信心。其次，教师必须向学生证明他们的勤奋一定会取得更好的成绩。

形成性评价和终结性评价的意义

形成性 (formative) 评价和终结性 (summative) 评价经常被用于对教育评价的讨论。事实上,形成性评价和终结性评价的概念在刚形成时与课堂评价,甚至与学习没有什么联系。

形成性评价和终结性评价之间的区别最早为人们所认识是在1967年,迈克尔·斯克里文 (Michael Scriven) 在美国教育研究协会 (American Educational Research Association) 研究评价的专题系列文章中提出的。斯克里文最初的观点是,应该区分正在形成的项目和已经进入最后状态的项目。结果,评价呈现出不同的特点,对形成性情景和终结性情景的阐释各异。这种区别很快被应用到对学生的评价中。形成性评价专门用于学习知识的过程中,而终结性评价则用于某段学习的最后阶段——如课程结束 (参见 McMillan, 2000)。彼得·艾拉西恩 (Peter Airasian, 1994) 更正式地将形成性评价定义为“互动的、主要用于形成和改变中的过程或活动。相反,过程或活动最后出现的评价叫终结性评价,这时已很难改变或纠正已经发生的一切” (pp. 135 - 136)。

虽然形成性评价和终结性评价都被广泛应用于课堂评价的文献中,但是形成性评价在研究中得到更多的关注。尤其是形成性课堂评价成为几乎所有试图综合分析课堂评价研究的方法。回顾布拉克 (Black) 和威廉 (Wiliam) 对 250 多项研究的归纳发现,与终结性评价对立的形成性评价对学生的学习会产生更加重大的影响。特伦斯·克鲁克斯 (Terrance Crooks, 1988) 在研究评论中报告说,终结性评价的效应规模一直比形成性评价的效应规模低。简言之,形成性评价具有坚实的研究基础来支持它对学习的影响力。

然而,在研究文献中,形成性评价并没有统一的定论。正如布拉克和威廉 (Black, Wiliam, 1998) 所说的“形成性评价并没有被严谨界定和广泛认可的定义” (p. 7)。在这本书中,我运用的是布拉克和威廉提出的定义,即形成性评价为“所有由教师和/或学生进行的活动,并以这些活动提供的信息作为反馈去修正他们正在进行

的教学活动和学习活动”（pp. 7 - 8）。这个定义的范围很广，能用于作为评价和限定活动时间的两种活动类型。按照定义，形成性课堂评价能够并应该在某段学习过程之初就立即开始，并跨越整个期间。除此之外，形成性课堂评价可以有各种很宽泛的形式，有正式的（如书面测验）和非正式的（如与学生讨论）。

经常性评价的意义

研究得出的重大成果之一是评价的经常性与学生的学业成就有着紧密的联系。这一发现在罗伯特·班格特德劳恩斯（Robert Bangert-Drowns）、詹姆斯·库利克（James Kulik）和陈林·库利克（Chen-Lin Kulik）（1991）的元分析中得到了充分地展示。他们分析了 29 项研究经常性评价的结论，结果见表 1.3。

表 1.3 与 15 个星期的评价次数有关的提高

评价次数	效应规模	百分点的提高
0	0	0
1	0.34	13.5
5	0.53	20.0
10	0.60	22.5
15	0.66	24.5
20	0.71	26.0
25	0.78	28.5
30	0.80	29.0

注释：效应规模来自班格特德劳恩斯、库利克、库利克（1991）报告的数据。

为了说明这个表，假设我们正在考核一位正在修读一个 15 周课程的学生的学习（若要讨论这表是如何建构的，参见术语注释 1.4）。这个表描述了在 15 周学习期间运用形成性评价的不同次数对学习的提高情况。假如教师进行 5 次评价，学生成绩可提高 20 个百分点；假如教师进行 25 次评价，学生成绩可提高 28.5 个百分点；以此类推。罗伯特·班格特德劳恩斯、陈林·库利克和詹姆斯·库利克（1991）对这个发现做了多方面的评述。首先，他们关注表

1.3 描述的单项评价中较强的效应——13.5 个百分点；其次，他们强调了经常性评价的效果会随着时间的推移逐渐减弱的事实。如表 1.3 所示，评价效果从 0 次到 10 次有大幅跳跃，然后趋于平稳。罗伯特·班格特德劳恩斯、陈林·库利克和詹姆斯·库利克并不推崇在 15 周内教师要进行 30 次评价，而是说教师应该系统地运用课堂评价作为一种反馈形式。

林恩·富克斯和道格拉斯·富克斯（Fuchs, Fuchs, 1986）在他们对 21 项受控研究的元分析中也报告了相同的现象。报告说每周提供两次评价的结果是效应规模为 0.85，或获得 30 个百分点。虽然在一个单元的教学中或者在评分期间没有固定要进行多少次评价，但研究得出的信息是清晰的：每周或更频繁地进行系统的课堂评价对学生成绩会起到很有效的正面作用。

小结和归纳

研究证明形成性课堂评价是教师在课堂上使用的最有效的手段之一。形成性评价被定义为任何能对学生学习提供有效反馈（sound feedback）的活动。有效反馈的特点包括：频繁的，能对学生学习的进步及提高做出清晰的、鼓励性的指导。在以下的章节中，这些原则将被运用到设计有效的课堂评价综合体系上。

第二章 国家课标的作用

第一章所讨论的基于研究的第一原则证明了评价的意义是为学生清晰描述如何根据学习目标取得进步。这一章主要讨论各州及美国国家课标在清晰的学习目标执行上所发挥的作用。

毫不夸张地说，课标运动（standards movement）渗透到了美国的基础教育中。罗伯特·格拉泽和罗伯特·林（Robert Glaser, Robert Linn, 1993）解释道：

在我们国家的教育改革运动中，20 世纪的最后 10 年毫无疑问被确认为强烈呼吁国家教育标准出台的时期。联邦和各州立法者、总统和州长候选人、教师和学科问题专家、顾问、政府部门和民间社团都努力呼吁建立标准。（p. xiii）

格拉泽和林是在 20 世纪后期做出这个评论的。课标运动在 21 世纪初期不会减弱。49 个州（只有艾奥瓦州除外）已经开发了用于设计指导学校教学内容的标准。可以证明，即使当前的课标运动受到阻碍，各州都需要证明学生应该懂得什么知识以及能够在具体的学科领域做什么（讨论参见 Lauer et al. , 2005）。

关于课堂评价，可以推断，课标表明了学校应该评价什么。即，课标运动确定学生应该懂得什么以及能够做什么。如第一章列举的研究，每个地区最重要的策略是对各州课标的执行进行恰当的课堂评价，从而确保不让一个孩子掉队。然而，课标中有两个问题严重影响了教师进行有效的课堂评价：（1）内容太多；（2）缺乏一维性（unidimensionality）。

课标内容太多

强劲有力的全美课标运动可能有解决问题的办法，但又会产生很多问题。其中一个最突出的问题是，课标文献的内容太繁多杂乱。为了说明这点，美国中部大陆教育和学习研究机构（以下简称 McREL）的研究人员论证了覆盖 14 个学科领域的国家级和州级约 200 个课标和 3 093 个基准点（benchmarks）的文献（Kendall, Marzano, 2000）。研究人员询问教师需要多长时间才能完成这些课标和基准点上的内容。研究人员比较课标的教学内容需要的时间量和一般课堂教学需要的时间量后，发现需要用比现在多 71% 的教学时间去完成课标文献中规定的内容（Marzano, Kendall, Gaddy, 1999）。另一种观察的方法是，按照目前的情形，学校教育要从幼儿园延伸到 21 年级或 22 年级去适应国家文献中所有的课标和基准点。这根本不可能。事实上，校方根本不可能给每学年增加时间，即使几天也不可能。如赫伯特·沃尔伯格（Herbert Walberg, 1997）提到的，一个中等大小的学区每学年即使增加几天的成本都是昂贵的。

即使每学年都增加一些时间，也教不完国家级和州级课标文献上规定的所有内容。虽然一开始听起来是可以承受，但比较美国的课标与其他国家的课标得出的必然结论是，我们证明了我们基础教育体系的内容太多。如第三届国际数学和科学研究（TIMSS）（Schmidt, McKnight, Raizen, 1996）上所报告的，美国的数学课本想要覆盖德国课本 175% 的主题、覆盖日本课本 350% 的主题。同样，美国科学课本想要覆盖德国课本 9 倍的主题、覆盖日本课本 4 倍多的主题。然而，虽然德国和日本的课本所覆盖的主题少很多，但这两国的学生在数学和科学的表现上却大大胜过美国学生。做这种细致的调查很有意义。假如课程主题太多超过了时间的允许，那么这些主题仅是一种虚构的形式。

学区或学校应该做些什么？最直接但不是最简单的答案是，大力删减要求教师在课堂上讲授的内容。因此，学校必须区分哪些内

容对全体学生是必须学习的，哪些是可以删除的。许多课程研究人员和理论家都已表达了这个观点。例如，芬威克·英格利希（Fenwick English, 2000）建议学校要认真清查从而决定要完成课程规定的内容需要多少时间。当学校确定内容太多而难以在规定的时间内完成时，必须调整内容适应时间的分布。道格拉斯·里夫斯（Douglas Reeves, 2002）和拉里·安斯沃思（Larry Ainsworth, 2003a, 2003b）提出了“公布（unpacking）”课标文献概念，确认哪些内容是必需的，并把必要的内容组成一小套“重要的课标”。在格兰特·威金斯和杰伊·麦克泰伊（Grant Wiggins, Jay McTighe）《设计而理解》（Understanding by Design）一书中，他们同样强调通过持久的理解形式来重修国家课标的观点。这些是值得重视的建议，但需要时间和技术。在本章，我将阐述我对里夫斯、安斯沃思、威金斯、麦克泰伊以及其他建议的活动效果的看法。

课标缺乏一维性

即使调整内容来适应所规定的教学时间，课标文献中的必修内容仍要重组，根据形成性课堂评价来修改。这个要求来自测量法理论的基本原则——一维性原则。用简单的话说，一维性指的是测试中一个单独的分数代表被评价的一个方面或特点。这个概念支撑教育和心理学中测量法理论中几乎所有的论点。为了说明这一点，弗雷德里克·洛德（Frederick Lord, 1959）在一篇论述测量法理论基础的文章中解释说，测试“是任务的集成，受测试者在这些任务上的表现被看成是他在某种心理方面所在位置（standing）的指数（index）。”实际上，洛德的论述指的是，任何通过一个单独分数来描述学生表现的测试应该仅测量一个特点。有趣的是，课堂评价、标准化评价和各州评价经常破坏了这个假设。事实上，约翰·哈蒂（John Hattie, 1984, 1985; Hattie, Krakowski, Rogers, Swaminathan, 1996）等研究人员记录了设计一个一维性测试的难度，以及一维性假设是如何经常被破坏的（参见术语注释 2.1）。

为了证明一维性假设被破坏的后果，我们来看一份有 20 道题目的测试。有 10 道题测试一个具体特点或某个方面，如对文学作品不同流派的理解。我们把这叫做 A 类题。另外 10 道题测试与 A 类题目没有联系的另一方面，如编辑语法的能力。我们称之为 B 类题。下面，我们看看参加这次测验的三个学生的成绩。

	A 类题	B 类题	总分
学生甲	2	10	12
学生乙	10	2	12
学生丙	6	6	12

三个学生的总分一样，都是 12 分，但他们应对这两类题目的个人能力差别很大。学生甲解决 B 类题的能力强，但对 A 类题表现差；学生乙则相反，解决 A 类题的能力强，但对 B 类题表现差；学生丙解决 A、B 类题目的能力相同，在知识上可称为“均等 (fair)”。在后面讨论怎样进行课堂评价评分时我们还要回到一维性这个概念上。我们将会看到，在一份有两类题目以上的测试中评一个总分没有意义。反而，一维性原则是对每类题都应该给一个总分。

在各州和国家课标文献上，一维性原则指的是课标上的各类知识或特点都应该清晰地描述出来。然而，却很少这样做。为了说明这一点，请看美国国家数学教师协会 (NCTM) 出版的数学课标中的标准基点陈述，里面标明了五年级结束时，学生应该懂得什么知识以及能够做到什么：

- 能够熟练进行整数的加、减、乘、除。(NCTM, 2000, p. 392)

基准点上的信息和技巧是相互联系的，因为它们都涉及整数计算。但基本程序是不一样的，事实上差别很大。认知心理学家已做了明确的归纳，他们证明了加、减、乘、除认知过程的实际步骤（参见 Anderson, 1983）。仅这个基准点就可能覆盖四个单独的类型：

- 数的相加过程。
- 数的相减过程。

- 数的相乘过程。
- 数的相除过程。

这个例子提供的信息表明课标文献能够包含多少学科内容。美国国家数学教师协会的课标文献仅包含从幼儿园到十二年级的 241 个基准点。人们会认为美国国家数学教师协会的文献会注明 241 类内容。然而当我打开说明这些程序的美国国家数学教师协会基点陈述时，我发现了 741 个独特的要素（unique elements）（Marzano, 2002b）。观察其他的课标文献，可以看出美国国家数学教师协会的材料很典型。

以前和最近撰写的课标文献几乎没给教师提供里面所要求的类型指南。没有这样的指南，很难把课标文献作为完整的形成性课堂评价体系的基础。

克服阻碍

虽然阻碍看起来很难克服，但如果学校愿意在课标文献上重组（reconstitute）知识结构的话，阻碍是可以克服的。已经有很多研究人员、理论家和顾问提出重组国家课标和基准点的要求，他们与学校一起努力来实施基于课标的教育（参见 Ainsworth, 2003a, 2003b; Reeves, 2002; Wiggins, McTighe, 2005）。准确地说，怎样重组课标文献并不像怎样去实施的要求那么清晰可行。在科学上，美国高级科学协会（2001）建议国家科学标准应重组成“群体（clusters）”，如物质结构、细胞、物质和能量的流动以及生命的进化。里维斯（Reeves）和爱因斯沃德（Ainsworth）提出“力标准（power standards）”的建构，威金斯和麦克泰伊提出拱形（overarching）或持久理解（enduring understandings）的概念。约翰·肯德尔（John Kendall, 2000）建议课标和基准点应重组为“主题（topics）”。他解释说：

有 49 个州出版了基础教育的课标。大多数的州文献都是以相同的形式组织教学内容，学区和学校也发表了效仿

的课标文献。这种组成是直接明了的。被称为“课标”的最高、最广的层面是能够把主题领域分切成可操作的板块目录。这些目录的标签从单字到长句不等，但都是为了同一个目的。比如在数学上，目录标签标明了度量、计算、几何等领域。最细小低微的层面是课标的核心，它有很多称谓，如基点（benchmark）、目的（objective）或指示体（indicator）。这个结构层面描述学生应该懂得什么和能够做什么。（p. 37）

而肯德尔提醒到，这两个层面——课标的广度和基点、目的或指示体的具体层面——对教育家来说并不是很好操作的实践工具。他提出了一个“过渡（interim）”结构，称之为“主题（topic）”。按照他的说法，主题可以被定义为“介于课标的广度和基准点的具体性之间的某种东西”（p. 38）。他进一步解释说主题有三个基本的目的：

- 给教师提供更简易的方法为教学找到恰当的信息。
- 在学科领域内和跨学科领域的基准点之间建立更清晰的联系。
- 给学生提供具体有用的反馈。

肯德尔的主题关注教学。我觉得以课堂评价为目的的话，作为关系密切的范畴目录，用“测量主题（measurement topic）”这个术语更好。我推荐以下三个设计测量主题的步骤。

步骤 1：公布课标文献的基准点

设计有效的测量主题的第一个步骤是“公布（unpack）”课标文献的基准点，像在美国国家数学教师协会（NCTM）基准点陈述事例中显示的那样，公布基准点的专门草案（technical protocol）在马扎诺的书（2002b）中描述。虽然这是证明每个基准点陈述的信息和技能的独特要素的简单问题，我认为学科专家相当专业，在这方面做得非常好。因此，学区仅需要集中有经验的数学教师和课程专家去公布数学课标，集中有经验的科学教师和课程专家去公布科学课

标，等等。

这次使用内容范围不是数学，而是美国国家科学教育课标中特意
为从幼儿园到四年级的学生列举的科学基准点（美国国家研究委员会，
1996）：

- 光在遇到物体之前是以直线传播的。光能被镜子反
射，能被透镜折射，或被物体吸收。
- 热的产生有许多途径，如燃烧，摩擦或混合两种物
质。热还能从一种物质传递到另一种物质。
- 电路中的电能产生光、热、声音和磁场效应。
- 电路需要一个完整的回路使电流能够通过。
- 磁能够互相吸引和排斥，也能吸引和排斥其他物质。

(p. 127)

这些基准点涵盖了五个方面，每点一个。这里的科学例子和它
前面的数学例子来自美国国家课标文献。各州的课标文献在他们的
基准点陈述中展现了相同的问题（多维度）。

再举一例，参看以下五年级基准点的“测量”标准，选自俄亥
俄州题为“十二年级的数学学业内容标准”的课标文献（俄亥俄州
教育部门，2001）：

1. 识别和选择适当的单位来度量角，如度数。
2. 识别坐标网格或坐标平面点之间的路径；比较路径
的长度，如，最短的路径，等长的路径。
3. 展示和描述覆盖一个三维物体的表面（面积）和填
充其内部（体积）的差别。
4. 展示对线性单位、平方单位和立方单位之间差异的
理解。
5. 演算时在同一测量系统中进行换算。
6. 运用策略来确定出三角形、长方形、平行四边形的
周长和面积以及矩形棱柱体的体积公式。
7. 用基点角度（如，45 度、90 度、120 度）来评估角

度的测量。用工具测量和绘制角度。(pp. 72 - 73)

同样，这些简单的基准点包括了许多方面——至少涉及了七个，若公布每个陈述还会更多。如，第六条陈述了用公式证明三角形矩形、平行四边形的周长和面积，以及矩形棱柱体的体积。其中的每一个都能代表一个单独的方面，因为学生可能只胜任其中一个方面，而不是其他方面。

列举美国国家数学和科学标准以及俄亥俄州的数学标准并不是为了批评这些具体的文献。实际上，使用任何学科领域的国家和各州的文献都会产生相同的结果。国家和各州的课标文献并不是为简单地应用于课堂评价而设计的。所有的文献都必须“公布 (unpacked)”来说明其具体的知识范围。

步骤2：论证所有学生都必须学习的知识范畴

当标准基准点陈述被公布时，学区或学校可能会发现那比他们原来想象的内容要多很多。回想起公布美国国家数学教师协会 (NCTM) 出版的课标文献中 241 个基准点时，我发现有 741 个方面。必须大幅度地削减大量的内容才能符合教学时间的安排。只有熟练的教师和课程专家能有效地做到这点。

请参看我在前面提到的用数学范畴演绎的研究 (Marzano, 2002b)。具体而言，给 10 位数学教师派发了美国国家数学课标中的 741 个数学范畴 (NCTM, 2000)，要求他们论证哪些是所有学生必须学习的内容，而不管这些学生未来的志愿是什么。每个教师要较快地独自分别评价这所有的 741 个范畴。把这些评价集中起来排列出 404 个基本的范畴。换句话说，原来的范畴被减少了 46% (参见 Marzano, 2002b)。按照这种基本程序可以解决内容太多的问题，学区或学校只需召集学科专家，让他们来论证哪些内容是学生必须学习的。

熟练的教师和课程专家经过删减确定必学内容后，学区可以推广到整个社区。如，在 20 世纪 90 年代早期，来自 McRET 的顾问们为当地一个学区论证本学区所有学生必须要掌握的内容。在当地报

纸的一个增刊上公布了初步的基本项目 [为了社区群众能更容易理解, 学区使用的术语是基本项目 (essential elements), 而不用范畴 (dimensions) 或特征 (traits)]。每个项目后都有一个问題, 征求社区群众的意见, 问他们是否认为所陈述的知识内容真正应该让学生学习。在增刊上填写了意见的社区群众把它交到学区办公室。按社区群众的意见统计每个项目被同意的百分比, 并把这些意见排列成表。该学区的学科专家们根据社区群众的意见恰当地调整了基本项目。在最后的分析中, 教师和社区群众都认为这些基本项目能代表当地学区学生基础教育的需要 (参见 Marzano, Kendall, 1996)。

步骤3: 把范畴组织到相关的信息和技能目录中

设计测量主题的第三个步骤是把确认为重要的范畴编排到相关要素的目录中。当然, 这个步骤的关键问题是相关术语的意思要清晰。这个问题可由协变性 (covariance) 概念阐明。

协变性或称协方差在术语注释 2.2 中有更严谨的讨论。简言之, 协变性指的是当一个方面的能力提高时, 另一个方面的能力也会随之提高 (Shadish, Cook, Campbell, 2002)。范畴的协变性是教学功能的一部分。为了说明这一点, 要再次参考前面讨论的包含在美国国家数学教师协会基准点中的范畴:

- 数的相加过程。
- 数的相减过程。
- 数的相乘过程。
- 数的相除过程。

要确定这些范畴是独立的, 因为学生可能能够熟练地进行整数相加, 但进行整数相减就不一定熟练。另一方面, 也要确定整数加减有重叠的步骤, 更重要的是, 它们的讲授是相继进行的。它们是重叠并被同时讲授的事实意味着它们是协变的。一方面的技能提高时, 另一方面的技能也提高了。因此, 联合形成测量主题的“相关 (related)”范畴在它们的组成部分重叠起来, 通常被同时讲授, 或

在讲授时相互联系。

要很好地理解包含协变范畴的测量主题，可参见表 2.1，上面列出了名为“要旨阅读（Reading for Main Idea）”的测量主题的协变范畴 [表 2.1 没有显示测量主题的最终公式。第三章展示了“基于标题（rubric-based）”公式，其中的测量主题应该得到论述]。测量主题说明在九年级和十年级，当阅读到适合本年级的材料时，学生能辨别并阐明文章观点的主要形式，如：

- 复杂的因果关系。
- 带有复杂的支撑系统的论点。
- 带有复杂的解决办法的问题。
- 带有多重故事线索的复杂情节。

这个测量主题的核心是证实篇章信息类型的能力。据说这种能动性在具体操作上限定了“要旨阅读”。假如学生阅读《红色英勇徽章》（The Red Badge of Courage）一书，理解该书中心大意的重要部分是辨别各个不同的情节和分辨不同的故事线索（如，辨别带有多重故事线索的复杂情节）。假如学生阅读一篇论述强化环境法规重要性的社论，理解社论中心大意的同时要分辨作者暗指和明述的论点的基本内容（如，证明带有复杂的支撑系统的论点）。

归纳这些相同测量主题的范畴，可以在确认学生辨别涉及复杂的情节和多重故事线索的信息类型的能力提高的同时，他们证明涉及复杂论点信息形式的能力也提高了——还有他们辨别涉及复杂的随意关系的信息类型的能力也同时提高，等等。

如将在第三章所述，表 2.1 所示的测量主题能使之相对更容易发挥形成性课堂评价。它清晰地描述了教师从一个年级层次到另一个年级层次要强调的内容。参见从一年级到九、十年级协变要素的提升，为九、十年级列举了四个要素，也为八年级列举了四个要素，有些要素是相同的。如，两个层面都涉及辨别复杂的情节、多重的故事线索和复杂的解决办法的问题的范畴。然而，九、十年级学生还要证实复杂的支撑系统的论点，而在八年级，他们仅需辨别基本

的论点。

从一个年级到另一个年级，测量主题中的协变范畴变得更复杂。而且某些范畴会在某一年级终止。如，辨别年代的范畴在六年级末就停止了。这表明在学区的课程计划中，年代学在三年级到六年级讲授，三年级开始讲简易的年代学到六年级讲完复杂的内容。简言之，测量主题应该反映学区的教学价值和实践。

表 2.1 测量主题“要旨阅读”的协变范畴

中 轴	当阅读适合该年级的材料时，学生能辨别并阐明课文观点的主要类型，如：
九年级 (低级区间)	• 复杂的因果关系 • 带有复杂支撑系统的论点 • 带有复杂解决办法的问题 • 带有多条故事线索的复杂情节
八年级	• 复杂的因果关系 • 基本论点 • 带有复杂解决办法的问题 • 带有多条故事线索的复杂情节
七年级	• 复杂的因果关系 • 基本论点 • 带有复杂解决办法的问题 • 带有单一故事线索的情节
六年级	• 复杂的因果关系 • 基本论点 • 复杂的时间次序 • 带有基本解决办法的问题 • 带有单一故事线索的情节
五年级	• 复杂的因果关系 • 复杂的时间次序 • 带有基本解决办法的问题 • 带有单一故事线索的情节
四年级	• 基本的因果 • 简单的时间次序 • 带有基本解决办法的问题 • 带有单一故事线索的情节

续表

三年级	• 基本的因果 • 简单的时间次序 • 带有基本解决办法的问题 • 带有单一故事线索的情节
二年级	• 基本的因果 • 带有单一故事线索的情节
一年级	• 带有单一故事线索的情节
幼儿园	• 不能运用

设计测量主题的综合系统的指南

如前面的讨论所述，设计测量主题的综合系统是需要多方努力的。但是，如本书后面章节所示，一个陈述清晰的主题系统必要的前提条件是运用课堂评价去跟踪学生的进步。马扎诺和海斯迪德提供了语言文学、数学、科学和社会学的测量主题范例。另外，以下的建议将帮助各学区设计有效的测量主题系统。

把每个学科领域和年级层次的测量主题限制在 20 个以内

由于实施课标的一个主要障碍是课标包括的内容太多，要证实太多的测量主题是反生产力的。我建议每个学科的测量主题不超过 20 个，最理想是每年级 15 个左右。

表 2.2 为语言文学、数学、科学和社会学提供了测量主题清单。要注意表 2.2 列举的仅是范例。每个学区必须制定自己的主题，以反映他们州的课标内容、他们的教师以及社区的优势。与许多学区的教育者一起研究时，我发现他们给测量主题起了各种各样的名称，但他们论证居于这些测量主题中的范畴却是相似的。

语言文学列举了 12 个主题，数学有 18 个，科学有 13 个，社会学也有 13 个。表 2.2 没有显示年级的跨度。实际上，测量主题不能跨越所有的年级层次。如，加、减、乘、除的数学测量主题从幼儿园

表 2.2 测量主题样本

语言艺术	18. 结果的效度
阅读	科学
1. 识字与词汇	科学本质
2. 要旨阅读	1. 科学史
3. 文学分析	2. 科学法
写作	3. 技术
4. 语言规约	科学应用
5. 组织和中心	4. 数学和科学语言
6. 研究和技术	5. 科学交流
7. 评估和修正	6. 共同主题
8. 写作技巧应用	物理科学
听与说	7. 物质和能源
9. 理解	8. 力的本质
10. 组织和演说	9. 生命的多元性
11. 口头媒体的分析和评估	10. 人类
12. 说话技巧应用	11. 生命的相互依赖性
数学	地球科学
数字意义	12. 地球及其形成的过程
1. 数字	13. 宇宙
2. 运算关系	社会研究
3. 估计	公民/政府/民主
计算	1. 权利、责任和政治进程参与
4. 加/减	2. 美国和各州宪法
5. 乘/除	3. 民事和刑事司法系统
6. 运算	文化和文化多元性
7. 比率/比例/百分比	4. 文化影响
代数和函数	5. 文化的趋同与差异
8. 代数概念	生产、分配与消费
9. 绘制坐标平面	6. 经济体制的本质和功能
几何	7. 本地、各州、国家和全球经济
10. 线/角、图形	8. 个人经济
11. 运动几何	时间、持续和变化
测量	9. 改变历史的人物和事件
12. 实际应用	10. 当前的事件和现代世界
13. 维数测量	11. 过去、现在和未来的影响
数据分析/概率	人民、地方和环境
14. 直观表示	12. 空间思维和表格、地图及图形的使用
15. 统计	13. 地理对人民和事件的影响
16. 概率	
问题解决	
17. 策略/解题	

开始到五年级结束，比率、比例、百分数以及曲线坐标平面等则到六年级才讲授。在效果上，即使数学有 18 个测量主题，但每个年级都应少于 18 个。主题是在目录中分类组合的，学区和学校会使用不同的目录术语，如“组成部分”，“题目”，甚至用“标准”。

假如某个年级的主题数量太少，有些主题可以在一个学年内重复讲授；有些主题仅讲授一个季度，而其他主题可能每个季度都讲授。如，三年级每个季度都讲授加、减、乘、除的数学测量主题，而线段、角度和图形则只讲授一个季度。

一套表达清晰的测量主题要完成我所指的“有保证的、可行的课程（a guaranteed and viable curriculum）”，还有很长的路要走（Marzano, 2003）。这个概念在《学校的作用》（What Works in Schools）一书中有深入的介绍。但是，有这种课程的学区可以保证不管谁教某个课程或某个年级，某些主题必须得到恰当的讲授。很明显，教师必须严格按照具体的测量主题来达到要求。要使这样的保证生效，学区必须有足够的测量主题来确保掌握主题的过程是可行的——教师要在规定的时间内完成。

包含生存技能的测量主题

近几十年来的许多研究和报告都关注“生存技能（life skills）”的重要性——信息和技能对传统的学科专业并不特别，但对在各种情景下都能取得成功却是很重要的。例如，由获取必要技能的秘书委员会（SCANS）1991 年拟定的题为《对学校工作的要求：为美国的 2000 年》的报告提到，作者们用了 12 个月“与生意人、公务员、雇员管理者、联邦官员以及工人进行在线或面对面交谈。我们与他们在商店、政府办公室和生产车间交谈”（p. v）。这个大范围的研究得出了关键的工作技能或生存技能的综合清单。包括的行为有努力、合作、坚持公司政策、解决问题、思考和陈述理由等方面表现出色。

针对 SCANS 报告的一份补充文献《工作单位基础：雇主要求的基本技能》（Carnevale, Gainer, Meltzer, 1990），是由美国训练和发展社团（ASTD）资助的。这份报告总结了“大约 5 万名职业者、经

理、行政管理者、教师 and 人类资源发展领域的研究员”的意见 (p. xiii)。报告再次证实了守时、尊重他人、诚实工作等品质的重要性。

SCANS 报告和《工作单位基础》关注的是雇主们。而通过民意测验对父母和监护人调查的研究报告了相同的结果。如，公众民意调查机构资助的报告《首要的第一件事：美国人期望公立学校做的事情》(Farkas, Friedman, Boese, Shaw, 1994) 提到，受访者中有 88% 的人认为学校应该培育与工作相关的能力，如：自律、守时和可靠。在一份由 McREL 研究员指导的、盖洛普 (Gallup) 公司操作的对美国成年人做的普通调查中，受访者把生存技能排列在数学、科学、历史、语言文学和体育等所有学生在高中毕业前必须学习的 13 个学科领域之上 (Marzano, Kendall, Cicchinelli, 1998)。最后，“21 世纪技能伙伴”网站 (the Partnership for 21st Century skills: www.21stcenturyskills.org) 确认，既然生存技能对在 21 世纪获得成功如此重要，在基础教育中就必须大大强调生存技能的培养。

由于这些生存技能的重要性以及各种工作上的强制要求，要在学校讲授和强化生存技能，我提议应设计测量主题来强调它们。以下的生存技能通常被学区证明是非常重要的：

- **参与 (Participation)** 指学生努力参与班上的活动、积极应对分配给他们的任务的表现。

- **完成作业 (Work completion)** 指学生坚持遵守分配给他们的任务要求的表现。包括学生要及时上交作业、遵守教师制定的规定 (如：对报告格式的考虑)。

- **行为 (Behavior)** 指学生的言行举止要合乎规范的程度。包括教师个人的要求和学校范围制定的规章制度。

- **集体合作 (Working in groups)** 强调学生为了达到小组目标而积极参与的程度。这个目录不包括学生在小组中的行为，它包括在行为目录 (category behavior) 中。该目录关注学生参与小组目标完成的表现，反对仅关注他们自己的目标。

像学术测量主题一样，协变范畴也要求每个年级表明各自的生存技能测量主题。但是，随着年级的提高，生存技能测量主题范畴会有更多的重叠。如，学区可以证实生存技能中参与（participation）主题的范畴或要素：

- 要尽量努力回答教师的提问。
- 自愿发表意见而不需要被点名。
- 注重演示。

虽然已为每个年级列举了不同的范畴，但必须证明它们对中学的所有年级都是重要的。

在高中层次改变测量主题的结构

以上所描述的测量主题途径对从幼儿园到八年级、甚至到高中前两年的九年级和十年级都发挥了很好的作用。因此，由于不同年级协变因素变化而确定逐渐变得更复杂的测量主题是很有意义的。但这个途径在大多数高中所用的课程结构上并没发挥作用，因为高中课程在内容上更多元化。例如，高中通常开设的课程有代数 I、代数 II、几何学、三角学等，而小学只提供数学。解决的办法是分别考虑高中的“低级区间（lower-division）”课程和“高级区间（upper-division）”课程。

虽然使用了不同的术语，低级区间和高级区间课程的区别成为讨论高中课程改革的特色有一段时间了。比如，初级证书（CIM）和高级证书（CAM）的概念最早是由美国劳动力技能委员会在 1990 年的题为《美国的选择：高技能还是低工资》（Rothman, 1995）的报告中提出的。一般来说，CIM 表示期待 16 岁学生应该懂得的和能够做的事情。它类似所谓的“普通文字标准（general literacy standards）”——即要求普通大众都能掌握的信息和技能（参见 Marzano, Kendall, 1996）。相反，CAM 指对那些将继续在某个学科领域深造的学生的期望。这个概念与所谓的“高级标准（advanced standards）”或“世界级标准（world-class standards）”类似。如，1989

年9月在弗吉尼亚州卡罗特斯市的第一届教育高级会议上，所确定的6个目标中的目标4（Goal 4）明确表示美国的学校需要世界级标准：

目标4：到2000年，美国学生在科学和数学上的成就要排在世界第一。（美国国家教育目标小组，1991，p. 4）

把美国学生的操作能力提高到能与别国学生媲美甚至超过他们的程度，这种对高级标准的呼吁在纽约的帕利塞利斯（Palisades）1996年召开的第二届教育高级会议上得到更大的反响：“作为州长，我们对发展和确立具有国际竞争力的学术标准负有责任”（美国国家州长协会，1996，p. 4）。

实际上，高级标准和基础标准的区别以及CIM和CAM的区别意味着，对高中的入门课程来说，学校或学区应该使用对幼儿园到八年级所要求的相同测量主题。这就是表2.2列举的从幼儿园到十年级的协变因素。为九年级和十年列举的协变因素要求范畴将成为语言文学低级区间的焦点。对高级区间或高级课程而言，要发布全新的测量主题，反映这些课程更高级的内容。

允许教师选择测量主题

我对学区和学校的最后一个具体建议是每个年级的每个学科都允许有“可供教师选择的（teacher-choice）”测量主题。假如学区或学校能真正有效地设计测量主题，就应该有选择的空间。正如名称所指，教师选择主题包括教师认为重要的、但在学区和学校发布的测量主题中没有反映的内容。事实上，这种选择使教师能弥补学区和学校课程的不足，展现他们在学科领域的独特经验和背景。如，一个教十年级的英语教师在新闻学方面可能有深厚的背景，但学区和学校在测量主题上却没能反映新闻学的这些技能：

- 报道事件时要客观。
- 焦点放在故事上而不是报道者身上。
- 事件的见证必须是确凿的。

学区和学校在测量主题系统上要有很大的灵活性去适应所聘教

师各自的特长，这是很重要的。

小结和归纳

学区和学校需要以测量主题的形式重构各州课标的知识体系。要做到这点，就要运用三步程序（three-step process），即公布课标、证实要学生学会的基本范畴、把范畴编进与信息 and 技能相关的目录中。按照各种建议确保测量主题系统灵活，以适应全方位的基础教育体系。若没有重构各州的课标文献，教师就没有机会有效地运用课标作为形成性课堂评价的基础。

第三章 测量连续性学习的量表

第一章所展示的研究文献有四个普遍原则，其中之一是课堂评价应该自然形成。根据定义，形成性评价可以测量学习的进步。那么，教师怎样用评价方式测量学生学习上的进步？本章将主要讨论这个问题。我们先以批判的眼光审视分数体系，这是许多教师常用的评价体系。

为什么分数体系有缺陷

首先，“教师怎样用评价方式测量学习的进步？”这个问题似乎没有前提依据（non sequitur）。人们认为，教师所要做的就是在评分过程中对相应的测量主题进行一系列的评价，以及不断地检验评分形式。例如，假如教师要跟踪学生某个科学测量主题一个季度的学习进步情况，就要建构许多测量主题的评价。总之，教师要对该主题进行四个测试——初期的预测试（pretest）、第三周和第六周分别进行的两次中期测试（mid-tests）以及期末的后期测试（post-test）。假如教师用百分制（100-point）、百分比（percentage）或量表（scale）对测试进行评分，跟踪学生的学习就是检验每个学生四次测试得分进步的一个简单问题。这个评分形式能表明一个学生学到了很多：56、60、75 和 86。但是，另一种评分形式却无法表明学生学到了很多：65、68、70 和 71。

正如这个体系刚出现时所看到的，它存在着一个重要的缺陷——各个测试的评分对学生的某个测量主题的理解力和技能无法

比较。即，仅仅因为学生在第一次测试得了 56 分、第二次得了 60 分并不能说明她的理解力和技能提高了 4 个百分点。事实上，研究表明，学生在测试中得到的分数更多地取决于谁评分、他们怎样评分，而不取决于学生懂得什么和理解什么。要说明这一点，可参考检验八年级科学测试的研究（参见 Marzano, 2002a），它设有六个答案的题项——那些题项要求学生解释他们的答案，而不是多项选择题。

有 10 个学生参加了测试，他们的答题分别由 5 位教师进行评分，他们都是对测试内容非常熟悉并富有经验的八年级科学教师。然而，在给这 10 个学生评分前，每个教师都认真阅读了试题，并根据每道题强调的内容重要性对题项的分数进行分布。当然，这是在课堂评价的许多测试中推荐的常规做法（参见 Airasian, 1994; Brookhart, 2004; McMillan, 2000）。

虽然这种做法看起来很有道理，但在解释和比较学生的分数方面还是会产生混乱，因为不同的教师在题项分量的分配上会有不同。参见表 3.1，它显示每位教师对每项题目分配的分数。在每个分数的标示下是该分数所占的百分比。如，第一位教师给 6 个题项分配的总分是 50，1、3 题各 10 分；第 2 题 15 分；4、5、6 每题 5 分。10 分的题目占总分的 20%，15 分的题目占总分的 30%，每个 5 分题各占总分的 10%。

教师分配的题项的分量表明限定的形式。所有的教师对 1、2、3 题分配的分数都比 4、5、6 题多。然而，教师们分配的分数并不一致。题目分数分量形式差别最大的是第三位教师，他为第 1 题项分配了 45 分，第 2、3 题分配了 15 分。按照他的计算，在学生得到的总分中这三道题就占到了 50%，各自分别都是 16.7%。而第一位教师分配的前三道题分数分别占到总分的 20%、30% 和 20%。

教师对每个题项的分数和分量的分配差异在某种程度上解释了学生最后的测试总分的多样性。参见表 3.2，它报告了每位教师给每位学生计算的总分。检验图表时，要留意每位学生的总分已经被转换成百分比或 100 分的量表。要突出不同的教师给每位学生的最后

总分的差异，可参看第二位教师与第三位教师计算第二位学生的总分分别是 91 和 50。这相差 41 分是该研究中教师之间评分的最大差

表 3.1 五位教师在不同题项上的评分分布

教 师		题 项						总 分
		1	2	3	4	5	6	
1	分数	10	15	10	5	5	5	50
	%	20.0	30.0	20.0	10.0	10.0	10.0	
2	分数	25	25	25	5	5	5	90
	%	27.8	27.8	27.8	5.6	5.6	5.6	
3	分数	45	15	15	5	5	5	90
	%	50.0	16.7	16.7	5.6	5.6	5.6	
4	分数	20	15	15	10	10	10	80
	%	25.0	18.8	18.8	12.5	12.5	12.5	
5	分数	20	20	20	10	10	10	90
	%	22.2	22.2	22.2	11.1	11.1	11.1	
资料来源：Marzano (2002a)								

表 3.2 在 100 分量表中五位教师给十位学生的总分

学 生	教 师				
	1	2	3	4	5
1	65	75	80	50	60
2	78	91	50	86	74
3	90	70	82	100	85
4	65	70	50	70	60
5	82	92	72	72	100
6	71	61	82	75	60
7	87	100	85	81	72
8	76	86	70	50	60
9	73	84	72	62	75
10	80	92	100	80	76
资料来源：Marzano (2002a)					

异，而且由于在权重分配中的不同，这意义很大。重新考查表 3.1，我们发现第二位学生从第二位教师那里得到最后总分比是 91，因为该生在 6 个题项的分数中总共获得了 82 分：

- 第 1 题 25 分中得 20 分。
- 第 2、3 题 25 分得了满分。
- 第 4、5、6 题的 5 分题中各得了 4 分。

90 分中得了 82 分转换成百分得分比是 91。

第二位学生从第三位教师那里得到的最后总分比是 50，见以下的分数分配：

- 第 1 题 45 分中得 20 分。
- 第 2、3 题 15 分各得了 10 分。
- 第 4、5 题 5 分各得了 2 分。
- 第 6 题 5 分得了 1 分。

90 分中得了 45 分转换成百分得分比是 50。

该表还显示了另一个多样性：存在于分数体系中教师的判断力——教师对学生的应答满意程度的期待。即，在给学生的每一题项评分时，教师的头脑中已存在能够表明学生的理解或技能展示的相关答案。假如学生的应答符合教师的想法，学生就能得满分。假如学生的应答不符合教师的想法，教师就会酌情给一些分数。达到四分之三要求的应答可得 75%，达到一半要求的应答可得 50%，等等。如杰弗里·史密斯、莉萨·史密斯和理查德·德利西（Jeffrey Smith, Lisa Smith, Richard DeLisi, 2001）所解释的，教师会“对正确答案给满分，然后根据学生错误的程度扣分”（p. 52）。

在这项研究中给科学测试评分时，教师对学生题项的应答符合理想答案的程度有不同的看法。让我们认真比较一下第二位和第三位教师给学生 2 的每个题项评分的具体区别（参见表 3.3）。教师 2 分别给学生 2 的 6 个题项的总分百分比是：第 1 题总分比是 80%；第 2、3 题总分比是 100%；第 4、5、6 题总分比是 80%。因此，该学生第 1 题 25 分中得 20 分；第 2、3 题 25 分得了满分；第 4、5、

6 题 5 分中各得了 4 分。总分 90 分中得了 82 分，转换成百分比为 91%。相反，教师 3 分别给学生 2 的 6 个题项的总分百分比是：第 1 题总分比是 44%；第 2、3 题总分比是 67%；第 4、5 题总分比是 40%；第 6 题总分比是 20%。因此，该学生在第 1 题 45 分中得 20 分；第 2、3 题 15 分各得了 10 分；第 4、5 题 5 分各得了 2 分；第 6 题 5 分得了 1 分。总分 90 中得了 45 分，转换成百分比为 50%。实际上，教师对学生怎样才算答好题的看法的差异在他们对每个题项分配的权重或分数不同的基础上，进一步加大了总分的差异。很明显，教师在运用分数体系进行评价打分时在许多方面都有所不同；学生从分数体系中得到的总分由于教师不同而无法比较。

表 3.3 教师 2 和教师 3 是如何给学生 2 评分的

	题 项						总分	学生得分 的百分比
	1	2	3	4	5	6		
教师 2								
题项的分数	25	25	25	5	5	5	90	
学生的得分	20	25	25	4	4	4	82	91%
该题项得分 的百分比	80%	100%	100%	80%	80%	80%		
教师 3								
题项的分数	45	15	15	5	5	5	90	
学生的得分	20	10	10	2	2	1	45	50%
该题项得分 的百分比	44%	67%	67%	40%	40%	20%		

分数体系的起源

尽管有自身的缺陷，但由于分数体系是对课堂评价打分的可选方法，探究它的起源还是有指导作用的。测评专家达雷尔·博克 (Darrel Bock, 1997) 将分数体系追溯到第一次世界大战，为了快速有效地鉴定成千上万个新兵的能力，当时美国军队设计并运用了字

母等级测试法 (Alpha Test)。测试的目的是评价新兵的能力，从而把他们安排到最适合发挥他们的能力的岗位上。这种测试需要一种基于多项选择题的对错评分的快速、有效的评分体系。选对得 1 分，选错不得分。测试的概括性总分很容易计算，即用正确的题数除以总题数再乘以 100 就得出百分数。总的来说，字母等级测试对快速和有效地评价成千上万个新兵是很成功的。评分简易的字母等级测试的成功应用使多项选择题和取得概括性总分的百分数方法得到了普及。

基于正确答案比率的多项选择形式和概括性总分在 20 世纪 40 年代得到了批准，当时美国大学入学考试董事会 (CEEB—the College Entrance Examination Board) 任命心理学家卡尔·宾哈姆 (Carl Bingham) 开发学术能力测试 (SAT—the Scholastic Aptitude Test)。学术能力测试用于预测学生在大学的成就。由于对多项选择题评分容易，考试中的写作部分在 1942 年被取消了。在那以前，写作部分在测试中占的比重最大。1947 年，多项选择形式成为固定的模式在很大程度上是由于有了机器评分的答题卡。如博克 (Bock) 所解释的：

由于早期的设备仅能计算用铅笔做的所选题项的正确题数，正确题数的分数成为了教育测量理论研究的唯一数据来源。它是测试理论研究的重点。(p. 23)

多项选择形式和正确率的概括性总分的用途很快普及到所有能够用 1 分或 0 分表示对错的题项中——包括正误 (true/false) 题、配对 (matching) 题、填空 (fill-in-the blank) 题。事实上，测量理论的学说，例如信度 (reliability)、效度 (validity) 以及广度 (extent)，这些能区分学生在测试中做得好不好的题项 (被称为题项区分度) (item discrimination) 源自于被评为对错的假设 (参见 Gulliksen, 1950; Lord, Novick, 1968; Magnusson, 1966; Nunally, 1967)。

正如人们预测的，以测试结构主题为基础的课程和课本采用了

对/错的启发式概念来给测试评分，用百分数作为选用的概括性总分。这样的话，若对那些无法用对或错来评分的题项和任务进行分数分布，这是一个欠妥的做法，比如写作、口头演示等。若认识不到这一点，基础教育界很快就会被分数或百分比体系所困扰。

理性看待评价

以上的讨论使我们认识到，目前在课堂上使用的分数体系对有效的形成性评价任务是不够的。要了解如何改进分数体系，我们首先必须思考课堂评价的本质。更具体地给那些已经并将继续在本书中使用的术语下定义是有必要的。阅读各个课堂评价专家的相关著作（McMillan, 1997; O'Connor, 1995; Stiggins, 1994, 1997; Terwilliger, 1989），就会发现以下的定义：

- 评价（Assessment）——指能提供学生对一个具体的测量主题的理解力和技能运用的相关信息的有计划的活动。
- 测试（Test）——发生在某个具体时间，并最常用纸笔形式进行的一类评价。
- 评估（Evaluation）——在评价的基础上对学生的理解力和技能做出判断的过程。
- 测量（Measurement）——根据一套明确的规则对评价打分。
- 评分（Score）——通过测量的过程用数字或字母对评价打分；与另一个术语“mark”同义。

这些术语对课堂评价提供了有趣的观点，因为它们指的是一整套行动。评价是所有能给学生提供在一个具体的测量主题中的理解力和技能运用的相关信息的有计划的活动；测试是众多评价形式中的一种。不管使用哪种评价，都要通过评估对每个学生的理解力和技能做出判断。这些判断通过使用测量程序转换成分数。

人们会推测，尽管我们很关注学生在评价中得到的最终成绩，我们也应该重视产生那个成绩的过程，确保它经过了有效的评价、

评估和测量。所有这些整体过程的构成就是真实分数（true score）的概念。

真实分数的概念

真实分数的概念是测量理论的中心。真实分数在术语注释 3.1 中有更严谨的表述。尽管如此，我们思考一下测评理论家戴维·马格努森（David Magnusson, 1966）的以下论述：

某个测试测量的特点会由潜在的连续体表现出来，每个个体都会在能力量表上占有一定的位置。个体在能力量表上的位置决定着他在测试上的真实成绩以及他在真实成绩量表上的位置。（p. 63）

通俗地说，马格努森的论述表明，学生在测试题上的表现仅被认为是学生通过测试测量主题理解力和技能的标记。学生可能在应对所测量的主题的技能和信息方面非常熟练，但可能会因为疲倦而漏掉某些题目、误读题目、在答题卡上填上反应错误的“泡影（bubble）”，等等。反过来，学生可能在应对所测量的主题的技能和信息方面做得不好，但可能由于诸如猜测、作弊、侥幸等因素在某些题目上选择了正确答案。总之，学生在某个特殊的评价上得到的分数总会被看成是对他某个主题的真实分数的评估。

在古典测试理论（CTT—classic test theory）中，学生在测试上取得的分数和他在试题上的真实分数之间的关系由以下的方程式表现出来：

$$\text{监测的分数} = \text{真实分数（得分）} + \text{误差分数（失分）}$$

这个方程式表明，学生在评价上的监测分数（即教师评定的最后成绩）包括两部分——真实分数和误差分数。学生的真实分数代表他对相关主题的真实理解力和技能水平；误差分数是除了学生的真实理解力和技能水平以外、与其他因素相关的监测分数的一部分。如我们在前面的讨论中所了解的，用分数进行评价的打分特别容易出现误差，因为教师可能会评分过高或者过低。

这个争论存在已久的分数体系问题并不是没引起测评专家的关注。早在1904年，著名的教育和心理学测评先驱爱德华·桑代克（Edward Thorndike）就评论过拼写测试上的问题。桑代克（1904）观察到：

即使人们想测评如此简单的拼写问题，也会由于没有测评单位而受到阻碍。人们可以任意组织一些词汇，通过正确拼写的数量来观察其能力。但假如要检验这些词汇，就会由于单位的不均等而无法做到。基于词汇之间均等的所有结果都是很不准确的。（p. 7）

桑代克高度归纳了一个事实：即使是像拼写那样看似明了的学科领域，即使不是完全不可能，也很难以合理的方式确定每个题项的分数。每道题1分是没有意义的，因为有些词汇比另一些词汇更难拼写。但每个词汇不是1分也不严谨。假如正确拼写 gift 得1分，那么正确拼写 memento 该得多少分？

题项应答理论

尽管桑代克发出了警告，这种评分方法一直没人质疑，也没改变，直到20世纪50年代中期，一个新的测评理论基础——被称为题项应答理论（IRT—item response theory）问世了。苏珊·恩布莱特森和史蒂文·赖泽（Susan Embretson, Steven Reise, 2000）对IRT的历史进行了详尽的讨论。虽然对IRT理论的初步讨论通常追溯回艾伦·伯恩鲍姆（Allen Birnbaum, 1957, 1958a, 1958b）的著作，然而，最早综合评述IRT的人是弗雷德里克·洛德和梅尔文·诺维克（Frederick Lord, Melvin Novick, 1968）。目前，IRT作为最主要的体系用于对国家课标测试和标准化测试这种大规模评价的设计和评分上。

IRT并不是简单地给个体学生的测试加分来确定成绩，而是通过恩布莱特森和赖泽（2000）称为“诊断推测（clinical inference）”的途径。他们说，“在诊断推测过程的方法中，用诊断或推测来评估

可能性。即，所表现的行为（包括测试行为）怎样判断是一种诊断”（p. 54）。由于与评价打分有关，IRT 方法被描述为回答以下的问题：若学生这样答题，关于学生对这些题项所测量的理解力和技能特点，什么是最有可能的推测？这个途径与早前对评价、评估、测量和评分等术语所下的定义相当吻合。书面测试是一种评价——是收集学生对相关主题的理解力和技能水平信息的方法。诊断型推测必须由学生的监测反应形式来确定。学生的反应形式是恩布莱特森和赖泽所指的“表现行为（presenting behavior）”。评估是诊断型推测形成的过程。仅仅把正确的应答相加再除以总分不是评估，因为这样做没有包括判断。

IRT 提供的观点在为课堂评价打分的意义上很有说服力。首先，它增强了判断的可信度，设计一个可信的计划使教师能够用来为题项分布分数（对所有的实际目的）是不可能的。其次，它意指一个教师能够运用的基本策略，用于课堂评价的设计和打分，尽可能诚信地告诫人们，没有完全可信的方法给评价打分，不管它是由教师还是学区、州或测试公司构建的。

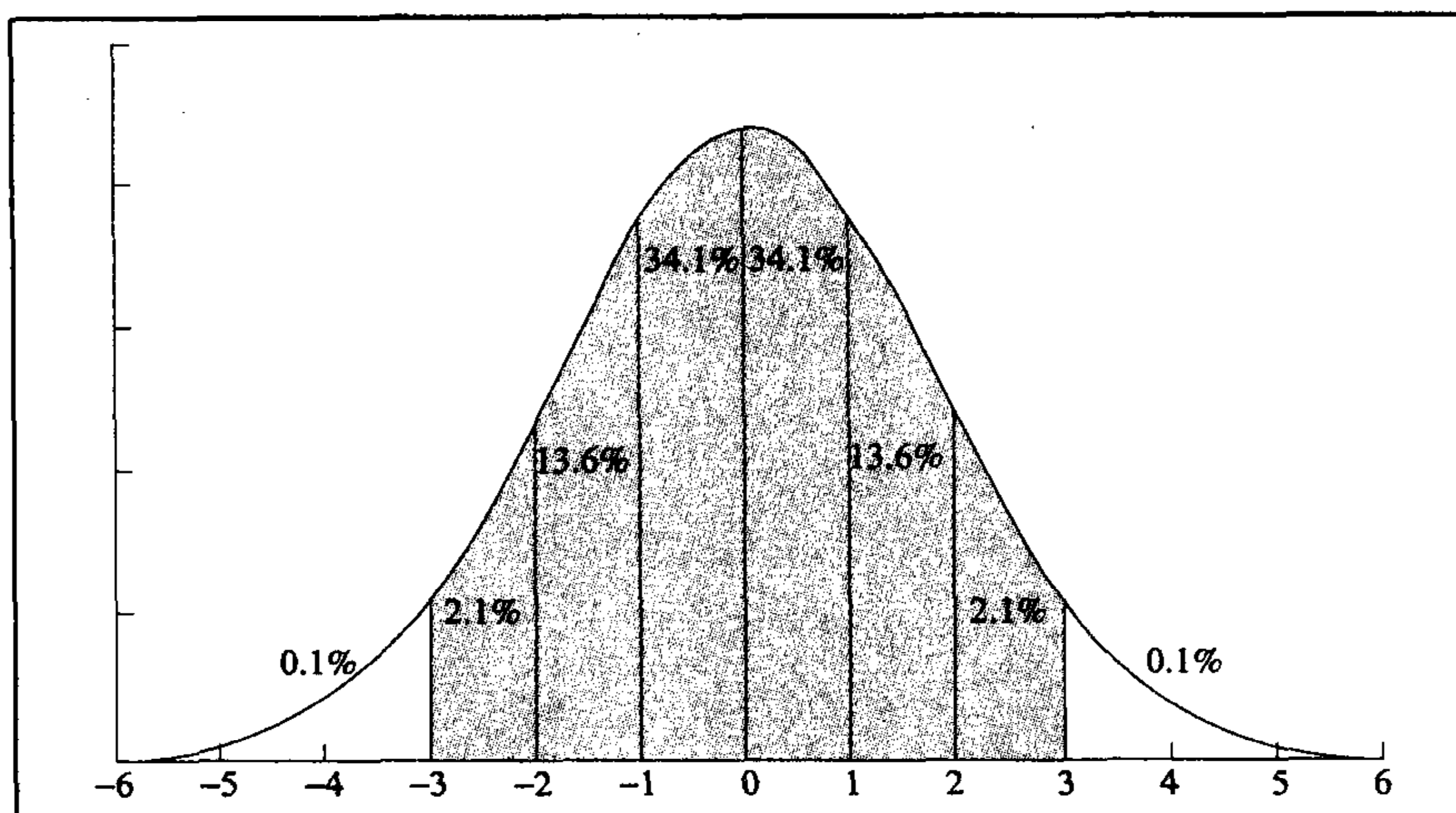
了解 IRT 逻辑

假如题项应答理论（IRT）模式不是把学生的测试成绩相加，那么它是运用什么程序？它怎样让教师用来给课堂评价打分？我们先谈谈 IRT 模式运用的程序。

IRT 模式基于一些基本的设想进行操作（参见术语注释 3.2）。首先，它们假设测试所测量的主题并不是直接观察的。IRT 理论家特别讨论了“延迟（latent）”的特征。其次，假设受测量的延迟特征模仿了如表 3.4 所描述的正态分布（normal distribution），这与许多教育家以及公众的假设相同（参见术语注释 3.3 讨论正态分布）。在这个描述中，平均分数是 0，分数的范围从 -3 到 +3。因此，学生在基于 IRT 的测试中的得分可以是 +3，+2，+1，0 等，被称为“特征分数（trait score）”。这些特征分数又被转换成一些其他的公

制 (metric)。如，其公制可能是 1 000 分。因此，+3 的成绩可能被转换成分数 1 000，+2 的成绩可能被转换成分数 750，0 的成绩可能被转换成分数 500，等等。

表 3.4 正态分布



为了确定学生的特征分数，IRT 模式运用了精确的数学分析题目应答的形式（参见 Embretson, Reise, 2000）。例如，假设用 IRT 理论来设计一份有 9 个题项的测试（实际上，在 IRT 测试发展中运用了更多的题项）。同时也假设，当测试题被运用时，在数学上确定 1、2、3 题包含的是被测量特征的简易信息；4、5、6 题为中等难度的信息；7、8、9 题为相对更难的信息。

用这个信息作为基础，9 道测试题的应答形式都能分布在特征分数上。如，表 3.5 描述了三个学生对这 9 道测试题的应答形式。在表 3.5 中，1 表示学生答对了该题，而 0 则表示学生没有答对。要注意的是三个学生都答对了 6 道题。然而，考虑到每道题难度的信息时，IRT 模式能计算每个形式的可能性，因为它与表 3.5 描述的每个特征分数相关。设定监测形式的话，它分布到可能性最大的特征分数上。如以学生甲为例，IRT 模式计算了对特征分数的应答形式的

可能性是 -3.0 , -2.5 , -2.0 , -1.5 等。若设定应答形式的话, IRT 模式的数学确定 $+0.5$ 的特征分数有最大的可能性。而相同的分析逻辑和数学分别给学生乙和丙分布了 $+1.0$ 和 $+1.5$ 的特征分数。

表 3.5 三个学生的答题情况

题 项	学 生		
	A	B	C
1	1	1	1
2	1	1	0
3	1	0	1
4	1	0	0
5	1	1	1
6	0	1	0
7	1	1	1
8	0	0	1
9	0	1	1
总分	6	6	6
假设的 IRT 特征分数	$+0.5$	$+1.0$	$+1.5$

虽然这个例子是根据 IRT 程序的总体过度简化设计的, 它说明了 IRT 测试发展的逻辑——分布最大可能性或最合理的特征分数, 假设了解每个题项的难度以及学生在这些题项上的应答形式。要运用 IRT 逻辑给课堂评价打分, 必须具备两个要素:

- 能代表对已有特性连续操作的量表。
- 能把评价的应答形式转换成量表上的分数的过程。

以下介绍量表的使用。

代表连续操作的量表

IRT 模式假设, 按照表 3.4 所示的正态分布对延迟现象进行操作, 该模式用复杂的数学把学生的应答形式转换成分布上的特征分数。很明显, 任课教师在评价打分时无法使用复杂的数学计算。因

此，他们需要一种量表，该量表不依据正态分布，而是依据具体的测量主题里理解力和技能的逻辑进步。表 3.6 展示了这种量表。

表 3.6 代表测量主题进步的评分量表

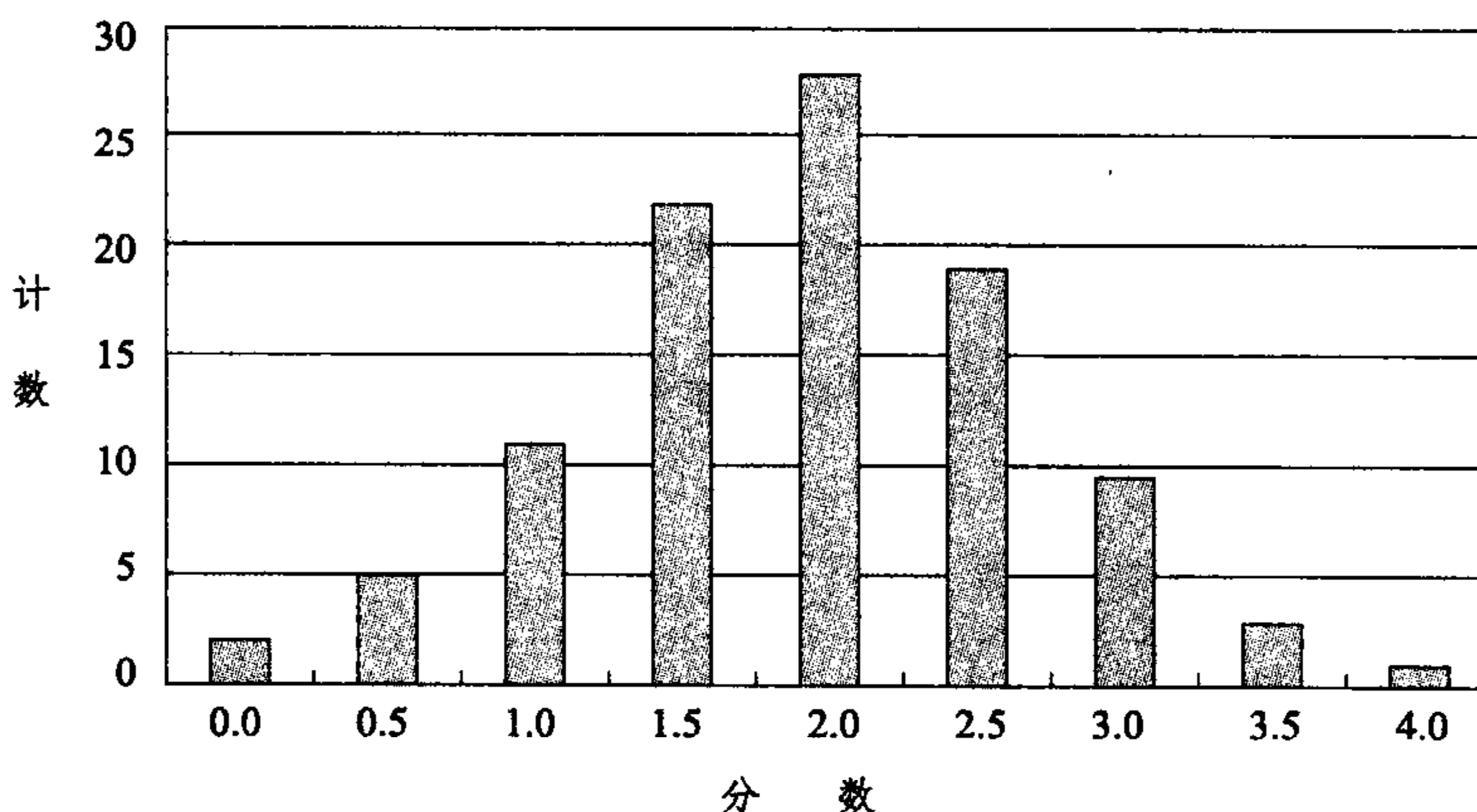
主题分数	量表描述的情况
4.0	除了 3.0 分要求的能力，能更深入地推理和应用超出讲授的内容
3.5	除了 3.0 分要求的能力，能在一定程度上推理和应用超出讲授的内容
3.0	在明确讲授的（简单或复杂的）信息和/或程序中没有重大错漏
2.5	在较简单的细节和程序中没有大的错漏，掌握一些较复杂的观点和程序
2.0	在较简单的细节和程序中没有大的错漏，但在较复杂的观点和程序上有很多错漏
1.5	掌握一定的简单细节和程序，但在较复杂的观点和程序上有很多错漏
1.0	借助帮助能掌握一定的简单细节和程序，以及某些较复杂的观点和程序
0.5	借助帮助能掌握一定的简单细节和程序，但无法理解较复杂的观点和程序
0.0	即使借助帮助也一窍不通
资料来源：Marzano（2004c）	

为了说明表 3.6 的量表，请看关于遗传（heredity）的科学测量主题。这种量表的最低分值是 0.0，表示不懂得该主题；学生即使借助帮助，也显示不出与遗传主题相关的理解力和技能。1.0 分表示，学生借助帮助就能对简单的细节和程序以及某些较复杂的观点和程序有所了解。给学生在评价上提供帮助的概念是量表很重要的特征，将在本章的后面进行深入的阐述。若得到 2.0 分，说明学生对简单的细节和程序已能独立地理解，并掌握了相关技能，但对复杂的观点和程序还做不到这一点。3.0 分意味着学生对所教过的无论是简单

还是复杂的内容的理解力和技能都不错。而 4.0 分则意味着学生的推理和应用能力超过了班上所教的水平。这将在第四章进行深入阐述的量表的重要特征。这里仅把量表作为基于表 3.4 正态分布的 IRT 连续性的逻辑选项来介绍。

有趣的是，当把表 3.6 的量表用于对学生评价打分时，分数的分布与正态分布非常相似。表 3.7 说明了这一点，它显示了用表 3.6 所示的量表对 100 名学生评价分数的分布。与正态分布相似的现象在一定程度上表明，量表能反映学生掌握某个主题分数的真实多样性。即，假如一组学生的实际成绩按正态分布的话，与正态分布相似的量表的分布结果能反映出学生的真实分数（参见术语注释 3.3）。

表 3.7 100 分的分布



把评价的应答模式转换成量表的分数

第二个题项需要运用 IRT 的逻辑对课堂评价打分，这是把评价的应答模式（pattern of responses）转换成量表的分数的过程。理解这一过程的最好方法是在具体的课堂评价情景下进行介绍。如：

你在考虑租一辆汽车，了解了四家公司的租金，每家公司都有一套日常租金和一定的免费里程数。但假如你用完了这些公司的免费里程数，他们就把每英里的租费加进日常租金中。

	红鸟出租公司	舒适出租公司	信誉出租公司	M & A 出租公司
日常租金	\$ 43.00	\$ 27.50	\$ 40.00	\$ 35.25
免费里程	1200	500	900	800
每英里租费	\$ 0.22	\$ 0.32	\$ 0.25	\$ 0.20

第一部分：

1. 哪家公司的日常租金最高？

答案_____

2. 哪家公司的免费里程最多？

答案_____

3. 假如每家公司的日常租金和免费里程相同，哪家公司租金最便宜？

答案_____

4. 假如每家公司的免费里程和每英里的租费相同，哪家公司租金最昂贵？

答案_____

5. 假如你用完了免费里程数而且每天行程 100 英里，哪家公司收取的费用最少？

答案_____

第二部分：

6. 假如你每天行程 100 英里，哪家公司收取以下天数的费用最少？

5 天：答案_____

10 天：答案_____

15 天：答案_____

20 天：答案_____

制作表格或图表来说明每家公司对以上的四个选项（5 天、10 天、15 天和 20 天）各自的费用，并解释你答案的计算方法。

第三部分：

7. 这四家公司分别在什么条件下是“最划算的交易（best deal）”？给每家公司都描述在什么情形下是最好的选择。在你的答案和解释中用上日常租金、免费里程数以及用完免费里程数后每英里的租费。

这份测试首先要关注的是它包括了不同类型的题项。第 1 题到第 5 题相当简单，要求学生阅读表格并做一些简单的计算。我称为第一类的题目——所涉及的细节和程序对学生面言相对容易。第 6 题比 1—5 题都复杂多了，学生要进行复合运算，并要比较运算的结果。我称这为第二类题——这些涉及复杂观点和程序的题目对学生的难度大多了。这类题要求学生回答的东西不是很明显。在这种情景下，学生必须计算每家公司不同的租用时间量的总费用，再对其进行比较。最后，第 7 题要求学生教师未在课堂上讲授的具体内容进行推测或应用。我把这称为第三类题。

这个评价涵盖的一、二、三类题并不是偶然的。对表 3.6 的量表进行检验表明，第一类题项用于确认学生能否获得 2.0 分，这说明他们能否胜任简单的细节和程序。第二类题项用于确认学生能否获得 3.0 分，这说明他们能否胜任复杂的观点和程序。第三类题项对于确认学生能否获得 4.0 分很有必要，这说明他们能否超越班上所学的东西。

简言之，要设计能运用表 3.6 的量表评分的课堂评价，必须包括三类题项：

- 第一类题项：指对学生来说相对容易的基本细节和程序。
- 第二类题项：指对学生来说难度较大的复杂思想内容和程序。
- 第三类题项：能超越课堂上教师所教的内容。

在第四章，我们将深入讨论如何设计这三类题项。但我觉得，即使教师没接受过这三类题项的有关特点的培训，他们也能自觉理解并能熟练设计这些题项。

运用简化版的评分量表

假如一个评价包括以上讨论的三类题项，教师把应答模式转化为测量主题的评分量表就相对要容易些。教师第一次运用这种体系时，我一般建议他们先使用表 3.8 这种量表的简化版本。表 3.8 被称为“简化版量表 (simplified scale)”，因为它仅包含 5 个整分值 (whole-point value) ——4.0, 3.0, 2.0, 1.0 和 0.0。虽然这个量表不如半分 (half-point) 量表 (还包括 3.5, 2.5, 1.5 和 0.5) 精确，但它能很好地显示把应答模式转化为评分量表的程序。另外，在某些情形下，半分量表很难分辨，没有很大意义。

表 3.8 简化版评分量表

主题分数	对量表的描述
4.0	除了 3.0 分要求的能力，能更深入地推理和应用超出讲授的内容
3.0	在明确讲授的（简单或复杂的）信息和/或程序中没有重大错漏
2.0	在较简单的细节和程序中没有大的错漏，但在较复杂的观点和程序上有很多错漏
1.0	借助帮助能掌握一定的简单细节和程序，以及某些较复杂的观点和程序
0.0	即使借助帮助也一窍不通
资料来源：Marzano (2004c)	

为了说明怎样运用简化版量表，假设学生能正确回答第一类题（从第 1 题至第 5 题）和第二类题（第 6 题）的所有题目，但没做第三类题（第 7 题），学生能得到 3.0 分，因为他在题项上的应答模式表明他能理解简单的细节和复杂的思想观点，但无法对超出教授的

内容进行推理和应用。

表 3.9 简化版评分量表的快速参考指南

	学生答题类型				
一类题	+	+	+	有一定理解 借助帮助	0 借助帮助
二类题	+	+	0	有一定理解 借助帮助	0 借助帮助
三类题	+	0	0	0 借助帮助	0 借助帮助
简化版（5级） 量表的分数	4.0	3.0	2.0	1.0	0.0
注：+表示回答正确；0表示不正确或没回答。资料来源：Marzano（2004c）					

表 3.9 提供了应用简化版量表给评价打分的快速参考指南。在该表中，标志 + 表明学生能正确应答具体的题项（一、二、三类）；标志 0 表明学生的应答不正确或没有应答；“借助帮助有一定理解”表明教师给学生提供指导和暗示，学生展示了题项要求的某些内容知识。因此，表 3.9 表明，学生答对了所有的题项能在评价上得到 4.0 分。答对第一、二类题、但没答对第三类题的学生得到 3.0 分。答对第一类题、但没答对第二、三类题的学生得到 2.0 分。假如学生独立做题时没有答对任何题，但借助帮助在第一、二类题上得到一些分数，他们就有 1.0 分。但如果学生独立做题时没有答对任何题，借助帮助也没法做对，就只有 0.0 分。

表 3.9 再次强调了 0.0 分和 1.0 分的重要区别。为了辨别学生究竟该得 1.0 分还是 0.0 分，教师必须与学生互动，因为在这两个分值上学生都没有答对评价中的题项。第一眼看上去，对得到 1.0 分和 0.0 分的学生的评价是相同的，都没答对一道题。但是，学生通过教师的帮助和指导能答对一些题，他们就有 1.0 分。但假如学生借助帮助也没有答对任何题，就只有 0.0 分。

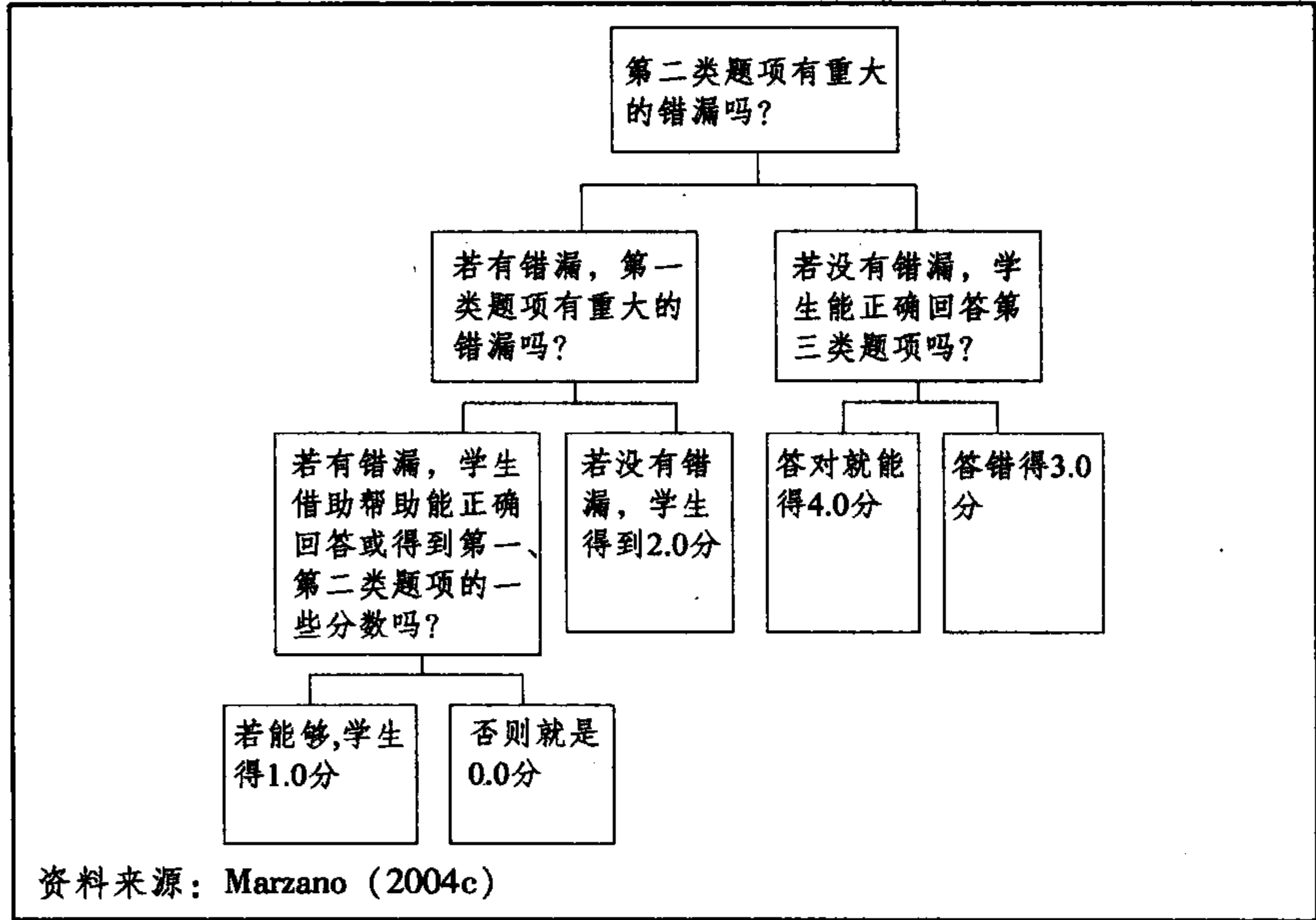
当我表示要与个别学生交谈测试应答的想法时，常常有教师提出质疑，他们教了 150 多学生，怎么与学生交谈？答案很简单，即

他们仅与那些没答对任何题、甚至不愿答题的学生互动。如表 3.7 显示的（展示 100 名学生的测试分数的分布），很少的学生仅得到 0.0 分和 1.0 分。总之，教师一定要与一些学生交谈——仅指那些没有答题或全部答错的学生。

如前面所提到的，我发现一旦教师熟悉了表 3.8 描述的简化版评分量表，他们就能很快并很准确地对课堂评价打分。因此，教师要阅读每个学生每道题的答案，答对的题目用简单的加法符号“+”表示，答错的题目或没有回答的题用“0”表示。回到我们的样板试题上，教师用“+”或“0”给每道题打分，假设学生从第 1—5 题都得到“+”，而第 6—7 题都得到“0”。要记住第 1—5 题是第一类题，第 6 题是第二类题，第 7 题是第三类题，教师就要用表 3.9 表示的快速参考指南来阐释应答模式。

此外，有些教师发现表 3.10 的评分流程图很有用。它表示教师评价打分时可能做出的决定。对测试题进行对错编码后，教师首先

表 3.10 简化版量表的评分流程图



要询问在第二类题项中是否有严重的错误或遗漏。假如答案是肯定的，教师再询问在第一类题项中是否有严重的错误或遗漏。若答案是否定的话，学生至少得 2.0 分，照此类推。虽然这个方案在量表中有提及，我发现该流程图显示了这个评分系统的基本“逻辑(logic)”。回顾第一章的表 1.2，运用一套规则进行课堂评价打分是与占学生成绩 32 百分点联系起来的 (Fuchs, Fuchs, 1986)。表 3.10 的流程图描绘了一个非常严谨的规则体系。

运用完整版量表

若是第一次试验这章描述的系统，从简化版量表开始是非常好的。但是，测量理论告诉我们，量表的分值越多，测量就越精确。关于 IRT 模式，恩布莱特森和赖泽 (2000) 解释说：“缩小分数之间的差别能提高评估水平的精确性” (p. 56)。假设教师只运用两种级别——及格/不及格——给测试评分；又假设学生必须正确回答 60% 以上的题目才能及格。在这种情况下，答对所有题目的学生就会与只能答对 60% 题目的学生得到相同的成绩（及格）。同样，所有题目都没答对的学生也会与只答对 59% 题目的学生得到相同的成绩（不及格）。测试中这种及格/不及格的成绩无法给教师或学生提供很多该学生的表现信息。照此类推，我们可以归纳出，带有半分的完整版量表能提供比只有整数分值的简化版本更精确的测量。

表 3.11 再次展示了完整版量表，以公式的形式进行评价打分更有用。半分的分数被分开在右边，表示它们描述整数的分数之间的应答模式以及能更精确地进行评价打分。半分能把部分分数分布到题项中。请看 3.0，2.5 和 2.0。3.0 分表示学生能正确回答第一类题（简单的细节和程序）和第二类题（复杂的观点和程序）的所有题目；2.0 分表示学生能正确回答第一类题，但没有回答第二类题。然而，若学生能正确回答第一类题的所有题目以及第二类题的部分题目，或在第二类题目上得到部分分数，那成绩应如何确定？使用简化版量表的话，就定为 2.0 分；使用完整版量表的话，就可定为 2.5 分，从而得到更精确的测量。

表 3.11 完整版评分量表

主题分数	对量表的描述
4.0	除了 3.0 分要求的能力，能更深入地推理和应用超出讲授的内容
3.5	除了 3.0 分要求的能力，能在一定程度上推理和应用超出讲授的内容
3.0	在明确讲授的（简单或复杂的）信息和/或程序中没有重大错漏
2.5	在较简单的细节和程序中没有大的错漏，掌握一定的较复杂的观点和程序
2.0	在较简单的细节和程序中没有大的错漏，但在较复杂的观点和程序上有很多错漏
1.5	掌握一定的简单细节和程序，但在较复杂的观点和程序上有很多错漏
1.0	借助帮助能掌握一定的简单细节和程序，以及某些较复杂的观点和程序
0.5	借助帮助能掌握一定的简单细节和程序，但无法理解较复杂的观点和程序
0.0	即使借助帮助也一窍不通
资料来源：Marzano（2004c）	

如前面所提到的，有时很难确定半分成绩。如，对学生在舞蹈课上的专业技巧能力评分时，很难考虑半分成绩。假如学生能表现基本的技巧，但对复杂的技巧有困难的话，很难就复杂技巧使用“部分分数（partial credit）”——她可能做到，也可能做不到。同样的限制会出现在评价学生在机械课中使用专业类工具的能力或在体育课中表现专业技巧的时候。

完整版量表是简化版量表的逻辑扩展。教师能自由轮替使用两种量表。当评价要求确定半分成绩时选择完整版量表；当评价难以确定半分成绩时就使用简化版量表。

表 3.12 描述了两种评分量表之间的关系。如前所示，标志

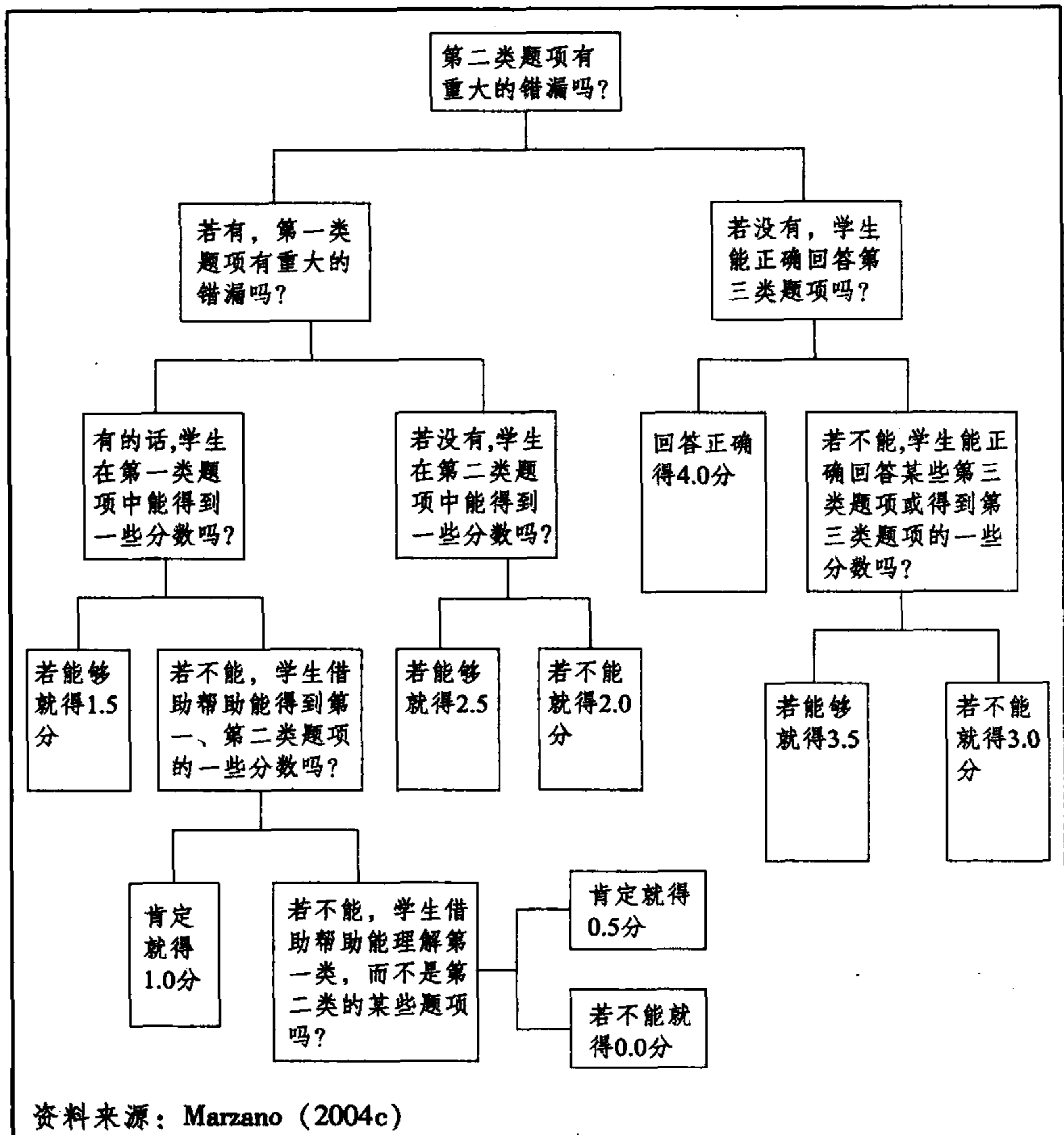
“+”表明学生能正确回答具体的题项（一、二、三类）；标志“0”表明学生的回答不正确或没有回答；“部分”表明学生能答对某些题项或答对题项的某些部分（附录 A 包含了对完整版量表和简化版量表的另一种比较，虽然表 3.12 对大多数人来说已足够）。

表 3.12 简化版和完整版评分量表之间的关系

	学生答题类型								
第一类题项	+	+	+	+	+	部分	部分 借助帮助	部分 借助帮助	0 借助帮助
第二类题项	+	+	+	部分	0	0	部分 借助帮助	0 借助帮助	0 借助帮助
第三类题项	+	部分	0	0	0	0	0 借助帮助	0 借助帮助	0 借助帮助
完整版（9级） 量表的分数	4.0	3.5	3.0	2.5	2.0	1.5	1.0	0.5	0.0
简化版（5级） 量表的分数	4.0	3.0		2.0		1.0		0.0	
资料来源：Marzano（2004c）									

如前所示，要用完整版量表对课堂评价打分，教师要检验每个学生的回答，答对的题目用符号“+”表示，答错或没有回答用“0”表示，部分正确用“部分”表示。回到原来的样板测试题，教师用“+”、“0”和“部分”的符号对每道题进行评分。假设学生在第一类题所有题目中都获得“+”，第二类的题目分别得到“+”、“0”和“部分”，他就会得到 2.5 分。另一个例子是，假设学生所有题目都没答对，但教师给了学生一些提示，学生仅在第一类题得到部分分数，第二类题没得分，他仅得到 0.5 分。一些教师发现表 3.13 的流程图有助于展示使用完整版量表评价打分的逻辑。要使用流程图，教师必须先询问：在第二类题项中是否有严重的错误或遗漏？假如答案是肯定的话，教师再询问在第一类题项中是否有严重的错误或遗漏；若答案是否定的话，教师再问：学生是否在第二类题得到部分分数？假如答案是肯定的话，学生得到 2.5 分；假如答案是否定的话，则学生得到 2.0 分，以此类推。

表 3.13 完整版量表的评分流程图



假如学生的回答没有依据逻辑形式该怎么办?

偶尔会出现的问题是, 学生的回答模式不符合简化版量表或完整版量表的预期形式。即, 学生给出不合逻辑的回答形式, 如, 能正确回答第二类所有题目, 但遗漏了第一类题的某些题目; 或者能

正确回答第三类所有题目，但遗漏了第二类题的某些题目；等等。首先要注意的问题是，这种现象甚至在运用 IRT 模式严格设计的标准化测试中也会出现（参见 Embretson, Reise, 2000）。按照 IRT 的说法，不合逻辑的形式指“异常的形式（aberrant pattern）”。其次要注意的问题是，不合逻辑的形式会有很多原因，包括以下这些：

- 试题中的题项在某些方面有缺陷。
- 学生尽力回答某些题项而忽略了其他题项。
- 教师对学生回答的评估不准确。

教师在这种情形下会采取很多措施，包括删除一些被认为不合理的题项、重新思考具体题项的分类、与那些表现出不合逻辑的回答形式的学生交谈。这个问题将在第四章有更深入的阐述。但在这里，我要简单地提醒，评价打分的最终目的是用完整版或简化版量表尽可能准确地评价学生在所评估的主题中取得的真实成绩。要意识到没有任何评价能提供评价学生的完美、准确的数据，教师应通过收集更多的信息超越所观察到的回答模式。

人们必须超越简单地计算正误去阐释测试分数的概念，这是测评专家经常讨论的问题。事实上，2005 年在美国国家教育测评委员会上做的主旨讲话中，戴维·弗里斯比（David Frisbie, 2005）解释说，几十年来，教育家们一直错误地把信度作为具体测试的特性。他说，信度被看成是解释分数的方式更合适（参见术语注释 3.4 能获得更详尽的弗里斯比的评论）。从表面价值看，弗里斯比论述的含义是，试题上的分数本身是不可信的。分数没有意义时，教育家们必须透过分数来获取学生真实成绩的最准确评价。

课堂评价评分方法的准确度

这一章描述的评价打分方法的一个重要的逻辑问题是：这种方法的准确度（精度）有多大？如前面提到的，术语注释 3.4 更深入地阐明了信度的概念。简言之，信度被看成是对学生评价打分如何

准确地评估他们的真实成绩。具体而言，测试要计算某种信度系数 (reliability coefficients)，这些信度系数从 0 到 1.0 排列，1.0 是个最理想的信度。

为了检验这章描述的信度系统，请 5 位教师给这章前面讨论的同样的 10 位学生所做的科学测试评分。这 5 位教师对测试内容的熟悉程度与那些运用分数系统对测试评分的教师一样。同时，第二组的 5 位教师则运用表 3.8 所示的早期简化量表版本。结果是，前 5 位教师通过分布分数评分，另 5 位教师运用简化版量表评分（参见 Marzano, 2002a）。在分数系统和运用简化版量表系统中都计算信度系数 [在术语上称为普遍性系数 (generalizability coefficient)]。用分数系统计算的信度系数是 0.294，而运用简化版量表系统计算的信度系数是 0.719。另外，我们发现，若两位教师用简化版量表单独对学生测试打分，对这两个独自比率的混合分数有一个评估的信度（普遍性）是 0.822。若有 4 位教师分别单独对学生的评价打分，这 4 个独自比率的联合分数 (combined score) 的信度是 0.901。同样的结果出现在迈克尔·弗利塞克 (Michael Flicek, 2005a, 2005b) 的报告中。

这些发现很有价值，因为它们表明，在同等条件下，教师做的评价具有与标准化测试所报告的信度一致的信度。贾森·奥斯本 (Jason Osborne, 2003) 发现，心理学杂志报告的平均信度是 0.83。而卢及其同事 (Lou and his colleagues, 1996) 报告的标准化成绩测试的信度是 0.85，学术成就的非标准化测试的信度是 0.75。

自初期研究指导使用简化版量表以来，其他研究则显示完整版量表产生的效果更好 (Marzano, 2006)。另外，这些研究表明，当完整版量表被重新写进来证实 2.0 分、3.0 分和 4.0 分的测量主题的具体内容时，完整版量表更有用、更准确。请参见表 3.14。表中的量表是为五年级的测量主题“要旨阅读”而制作的。具体的因素已提供给分值 4.0、3.0 和 2.0。要注意的是表 3.14 的例子仅是个案。运用了许多方案来定义和确认分值 4.0、3.0 和 2.0 的具体因素（参见 Marzano, Haystead）。在这个事例中，3.0 分表示学生能够辨别结构句型的具体类型，无论它们在测试中是外显的还是隐含的。分值

2.0 表示学生能够辨别相同类型的句型，假如它们清晰地测试中呈现；但若这些句型是隐含的，学生就会犯错误或遗漏。分值 4.0 表示除了辨别外显和隐含的句型外，学生还能解释和陈辩这些形式的推断。

表 3.14 具体的测量主题评分量表

学科范围：语言艺术		
测量主题：要旨阅读		
五年级		
4.0 分	除了 3.0 分要求的能力，学生能超出所教授的内容，如： · 解释哪些课文内容是明确的、哪些需要推理。 · 解释并对有关内容的推理进行陈辩。	
	3.5 分	除了 3.0 分要求的能力，能在一定程度上对超出所教的内容进行推测和知识的应用。
3.0 分	在阅读适合该年级的材料时，学生能够确定课文的要旨，如： · 外显和暗指的复杂因果关系。 · 外显和暗指的复杂时间次序。 · 能用基本方法解答的外显和暗指的问题。 · 外显和暗指的简单故事情节。学生没有明显的重大错漏。	
	2.5 分	在处理简单的细节和程序中没有大的错漏，掌握一定的较复杂的思想观点和程序。
2.0 分	学生在较简单的细节中没犯大的错漏，比如，能辨别： · 外显的复杂因果关系。 · 外显的复杂时间次序。 · 能用基本方法解答的外显的问题。 · 外显的简单故事情节。但是，学生在比较复杂的思想观点和程序上有较大的错漏。	
	1.5 分	掌握一定简单的思想观点和程序，但在比较复杂的观点和程序上有较大的错漏。
1.0 分	借助帮助，学生掌握一些简单的思想观点和程序和一些较复杂的观点和程序。	
	0.5 分	借助帮助，学生掌握一定简单的思想观点和程序，但不理解较复杂的观点和程序。
0.0 分	即使借助帮助学生也一窍不通。	

我认为,表 3.14 的公式是构建测量主题最理想的方法。附录 B 展示了图表中不同年级的语言艺术测量主题,以作为学区和学校模仿的普通细则。若需要一个语言艺术、数学、科学和社会研究测量主题的完整目录,请参见《让课标有助于任课教师》(Making Standards Useful to Classroom Teachers)(Marzano, Haystead)。学区和学校能采取的最有力的行动之一是以表 3.14 的公式发布每个年级各自的测量主题。这个行动不仅要求学区和学校为每个测量主题严格定义操作层面,还给教师提供评价打分的明确指南。

实操能力标准的问题

为了有效地用于跟踪学生在测量主题上的进步,本章展示的量表必须与基于目前课标的规则一致。其中的规则之一是确认学生成绩的具体操作能力标准(performance standards)。然而,具体操作能力标准的概念经常被误解。这个术语是由目标 3 和目标 4 的课标评论规划组在 1993 年的美国国家教育目标专门小组报告中(1993)启用的。根据规划组组长雪莉·马尔科姆(Shirley Malcom)姓氏命名的马尔科姆报告,它把具体操作能力标准定义为“多好才够好(how good is good enough)”(pp. ii - iii)。自从报告发表以来,全美都遵从的规则是用 4 个级别定义学生的表现:高级的(advanced)、熟练的(proficient)、基本的(basic)和低下的(below basic)。这个方案在美国国家教育进步评价著作中有其基础。如詹姆斯·波普汉姆(James Popham, 2003)指出的:

美国教育家按照美国国家教育进步评价(NAEP)使用的描述性目录在不断地建立实操能力标准,这是在联邦政府的赞助下进行的周期性测试。NAEP 的结果使学生的实操能力可以在参与评价的州之间进行比较……自从 1990 年以来,NAEP 的结果一直是用 4 种能力级别描述:高级的、熟练的、基本的和低下的。50 个州的大多数州现在都用这 4 个级别或与它们很相似的级别。例如,假如学生参加一个

全州范围的考试，里面有 65 道多项选择题，测试的具体操作标准必须确定 65 道题中有多少题答对学生才被界定为高级的、熟练的，等等。(p. 39)

这一章介绍的完整版量表很容易适应来遵循这个规则。参见表 3.15。该表中使用的分类方案是直接的。高级能力指学生能理解超出班上所教的内容，用 4.0 和 3.5 的分值表示。熟练能力指学生在相关的简单和复杂的观点和程序中没有错误，或是在简单的细节和程序上没错误，在复杂的观点和程序上掌握一定的知识，用 3.0 和 2.5 的分值表示，等等。

表 3.15 适用 NAEP 表现层次的完整版量表

高级的	4.0 分	除了 3.0 分要求的能力，更深入地推理和应用超出讲授的内容	
		3.5 分	除了 3.0 分要求的能力，能在一定程度上推理和应用超出讲授的内容
熟练的	3.0 分	在明确讲授的（简单或复杂的）信息和/或程序中没有错漏	
		2.5 分	在处理较简单的细节和程序中没有大的错漏，掌握一定的较复杂的观点和程序
基本的	2.0 分	在处理较简单的细节和程序中没有错漏，但在较复杂的观点和程序上有很多错漏	
		1.5 分	掌握一定的简单细节和程序，但在较复杂的观点和程序上有很多错漏
低下的	1.0 分	借助帮助能掌握一定的简单细节和程序以及某些较复杂的观点和程序	
		0.5 分	借助帮助能掌握一定的简单细节和程序，但无法理解较复杂的观点和程序
	0.0 分	即使借助帮助也一窍不通	
资料来源：Marzano（2004c）			

虽然表 3.15 的逻辑是比较直接的，但你要明白每个能力表现层次的“减分（cut point）”仍是很随意的。波普汉姆（2003）再次

提出：

我的观点是具体操作能力标准是可伸缩的，你很难知道“基本的”级别意味着什么，直到你阅读了那个能力层次精细的描述。如，“不让一个孩子掉队”法令呼吁各州至少建立三个学术成就标准层次（高级的、熟练的和基本的），还表明国家条例要求那些被认为是“熟练”或更高层次的学生比率不断提高……但是，各州都允许自己确定“熟练”的程度。若没有足够的学生在 NCLB 测试中达到熟练程度，这些学校会被亮红牌，因此，在一些州，他们所确立的“熟练”程度就有很大的水分。（p. 40）

波普汉姆的观点意味着，学区和学校建立的具体操作标准要反映其对学生的期望，而且学区和学校也要认为这些期望是合理和有价值的。例如，学区和学校可以决定低下的级别应得到的分值是 1.5 而不是 1.0，或者是 0.5 而不是 1.0。

归纳和小结

作者在不赞成使用分数对评价打分的基础上提出了论点。所选的方法是基于题项回答理论逻辑。运用这个方法时，教师把学生的回答模式转化成量表分数，它代表了学生对测量主题的理解力和技能的进步。量表的简化版本仅使用整数成绩，而完整版本则可以有半分成绩，因此完整版本在测量上更精确。

第四章 课堂评价的设计

第三章描述了课堂评价打分的过程，这与用传统的评分法进行打分大不相同。在新的评分系统中，学生的题目回答模式是确定学生在评分标准表上的分数的依据，而这一分数是学生在测量主题的学习上取得预期进步的标志。本章将阐述如何设计适应这种新的评分方法的课堂评价。本章还将讨论教师应该怎样为整个评价打分阶段都进行的形成性评价进行策划。第一章讨论了以研究为基础的课堂评价的四个原则。本章的讨论将有助于实施第三和第四个原则——课堂评价应该是形成性的；课堂评价应该是经常性的。让我们首先讨论形成性评价系统的设计问题。

形成性评价系统的设计

教师首先要回答的问题是，在评价打分期间将涉及多少个测量主题，以及每一个测量主题应该进行多少次评价？例如，在学年的第一学季（quarter），一位七年级的理科教师要跟踪学生在下列6个主题方面的进步：

1. 物质与能量
2. 力与运动
3. 繁殖与遗传
4. 科学调查与通讯
5. 地球与地球形成的过程
6. 生命的适应性与相互依存

大概各学区与学校已经确定这6个测量主题为七年级前9周的教学重点。教师很可能还希望或被要求跟踪学生的一些生活技能主题（参与活动，完成作业，小组活动和行为）方面的进展情况，我们将在本章讨论这些主题。

对每个科学主题进行单独评价，效果会相当差。对每一个主题进行单独评价也很可能与教师的实际教学不相吻合。也就是说，在学季的前几周，教师要讲授“物质与能量”和“力与运动”这两个主题的内容，而只评价其中一个主题的内容。对两个主题同时进行评价比评价单个主题更有意义。每次评价评出两个分数，每个主题一个分数。

然而，设计阶段性评价系统需要确定哪些主题将要进行评价，什么时候评价，一次评价是否涉及一个以上的主题。尽管教师不必确定每个测量主题在评价打分期间的评价时间，但是草拟一个总的评价计划是很有用的。例如，表4.1呈现一个9周内对6个科学主题进行评价的总体方案。如该表所示，该教师前三周将教学重点放在主题1，2和3。在此期间，她会对每个主题进行两次评价。第一周她会进行一次测试来检查学生对主题1和主题2的学习情况，题目涉及这两个主题。第二周她会进行一次测试或对主题2和主题3

表4.1 9周内对6个测量主题进行评价的总计划

周次	主题					
	1	2	3	4	5	6
1	×	×				
2		×	×			
3	×		×			
4				×		×
5					×	
6				×	×	×
7	×	×	×			
8				×	×	×
9	×	×	×	×	×	×

分别进行一次测试来检查学生对这两个主题的学习情况。第三周她会对主题1和主题3进行评价。在以后的三周里，她会讲授测量主题4、5和6。她又会对每个主题进行两次评价。注意在最后一周，她也许会进行一次综合性的终结性考试来检查学生学习所有6个主题的情况。

尽管表4.1呈现了一个为期9周的评价计划，但常见的是教师将各个主题组合成教学单元，这些单元的教学是在整个评价打分阶段进行的。本杰明·布卢姆（Benjamin Bloom，1976）可能是第一个描写教师这样做的人。他注意到在一个学年里，学生要面对150个独立的教学单元。每个单元大约需要7个小时。假设每个教学日上5门学术课程，那么一学年里，学生（每门课程）需要学习30个单元（或一学期内大约每门课程15个教学单元）。这种单元法很适合用表4.1所示的评价计划，因为其中有些主题可以自然地串在一起形成教学单元。例如，前3个主题可以串在一起形成一个四周半的教学单元，最后3个主题可以串在一起构成一个与前面类似的单元。

在美国越来越普及的活动就是学校或学区为具体的测量主题进行评价设计，所有的教师都必须这样做。拉里·安斯沃思和唐纳德·维格特（Larry Ainsworth，Donald Viegut，2006）记载了这样的评价设计与使用情况。学校或学区设计的评价有时候称为“统一评价”或“基准评价”，通常在评价打分阶段的末期进行（例如，在9周末或每个学期末进行）。例如，由理科教师构成的校委会已经为提前规定的6个科学主题进行了评价设计，准备在评价打分阶段的末期使用。已经拟订了表4.1中的评价计划的教师可以将校委会设计的那些评价用于第9周的终结性考试中，而不需要自己再进行期末考试设计。

这种方法有许多优点。首先，教这些主题的理科教师可以使用相同的期末考试试题，这样可以确保所有的教师使用相同的评分方案。其次，如第三章所述，两位教师用本书介绍的评分方法单独给每份试卷评分，然后将两者评出的分数综合起来，发现所得分数的信度高达0.82（参见Marzano，2002a）。这样，如果两位理科教师单

独给每次期末考试评分，而每个学生的考试成绩是由两位教师评出的分数平均后得出，那么那些通过联合评分的考试的信度可以比得上一些国家级考试的信度——可以说，进行联合评分是教师可用来促进教学发展的最有效的方式之一。总之，对测量主题进行统一评价可以成为本书所介绍的形成性评价系统的有效补充。

教师对课堂评价的设计

课堂评价方案一旦设计好了，教师就必须进行课堂评价设计（当然，统一的、学校或学区的评价设计除外）。课堂评价设计不必在单元教学或评价打分阶段的开始就完成。多年来，我观察到教师一般提前一周进行设计。

如第三章所讨论的，理想的课堂评价涉及三种类型题目或任务：

- 一类题目（item）或任务（task）：涉及基本细节与程序，学生做起来比较容易。
- 二类题目或任务：涉及更为复杂的思想内容与程序。
- 三类题目或任务：要求学生课堂上没有教过的知识进行推理或运用。

注意这里使用了“item”和“task”这两个词来描写上述三种类型的题目和任务。在本文的其余地方，我将交替地使用这两个词，尽管“题型”这个词一般是指试卷上经常出现的传统题型：单项选择、填空、是非判断、简答。“任务”这个词在这里指的是一种收集学生信息的更为有效的方法，也就是收集有关他们对某一特定主题的学习状况的信息，如表演任务、演示，甚至教师在非正式场合提出的问题。

我发现教师可以对这三类题目和任务进行非常直觉式的设计。有这样一个例子，一位六年级的教师想设计一份评价题来检查学生对“繁殖与遗传”这个科学主题的学习情况。设计这个主题的一类题目时，该教师会自问自答下面这个问题：对于这个主题，专心听

课的学生应该了解或者能够轻松地回答出来的细节与程序是什么？对于这个主题，该教师判定所有的学生都应该能够理解课堂上教过的一些基本术语。因此，她会设计一些如下的简单的一类题目：

简要解释下列每一个术语：

1. 遗传
2. 后代
3. 有性繁殖
4. 无性繁殖
5. 基因

这些术语涵盖的意思非常复杂，而该教师让学生解释这些术语是想确定学生是否对它们有一个总体了解。

设计这个主题的二类题目时，该教师会自问自答下面的问题：对于这个测量主题，专心听课的学生应该理解的、比较复杂的内容和程序是什么？他们能够轻松操作的比较复杂的程序是什么？这类知识更为复杂，因为它涉及的范围更广，意思更含蓄，组成成分更多。针对这个具体的主题，该教师会设计这样一道题：

解释该现象：如果在田野间生活的昆虫和鸟儿突然大量死去，而又没有别的昆虫和鸟儿来补充，长满开花植物的田野会发生什么呢？

人们认为二类题目涉及的信息比较复杂。因为它已不再是对术语特征的简单回忆了。而要回答上述问题，学生必须了解开花植物是有性繁殖，它们需要昆虫和鸟儿将花粉从一朵花儿传到另一朵花儿上去。因此，开花植物是依赖昆虫与鸟儿生存的。如果经常在某一块田地上活动的昆虫和鸟儿的数目降到一定的程度，有些植物就会死去。这种理解实质上是生成性的，因为它需要学生生产信息——既然这样，如果昆虫和鸟儿的数目骤降，情况会怎样呢？要回答这道二类题目，教师跟学生在课堂应该已经讨论了这个事实：开花植物依赖鸟儿与昆虫传播花粉，在不同的地点进行有性繁殖。这样，专心上课的学生就能够正确回答这个问题，尽管这道题需要的不仅

仅是对信息的回忆。

设计三类题目时，该教师就会自问自答下面这个问题：对于这个测量主题，学生能够进行什么样的推理和运用，尽管这种推理和运用是课堂上还没有涉及的？对于这个具体的主题，该教师可以设计这样一道题：

解释遗传特征与环境造成的特征之间的差异，然后列举一些你自己拥有的遗传特征与环境造成的特征，最后解释为什么你的行为更受遗传特征的影响或更受环境特征的影响。

这个三类任务有些特征类似于有关开花植物的二类任务，显然，它也要求学生产出信息。而那个产出的信息就含有说明遗传特征与环境特征的具体例子。三类题目与二类题目的区别就在于这个三类任务超出课堂所教的知识范围，要求学生将他们对遗传特征与环境特征的理解运用到自己的行为上去。重要的一点就是如果该教师在课堂上已经教过学生将课堂上出现过的信息运用到他们的行为上去，那么这就不是一个三类任务。因此，三类任务的一个关键特征就是它要求学生对课堂上没有教过的内容进行全新的推理和运用。

总之，这三类题目的设计可以是非常直觉的。教师只要提出下列问题并加以回答就行了：

- 对于这个测量主题，如果学生专心听课，他们应该理解的基本细节和程序是什么，或能够轻松操作的程序是什么？
- 对于这个测量主题，如果学生专心听课，他们应该理解的比较复杂的思想内容是什么，或能够轻松操作的程序是什么？
- 对于这个测量主题，学生能够做出的推理和运用是什么，尽管这些推理和运用是课堂上没有涉及的？

尽管采用直觉的方法设计一、二、三类题目效果好，但若对不同类型的知识有个彻底的了解会设计出更好的题目来。这里让我们讨论三类主题知识和一、二、三类题目或任务在每类知识中的表现方式。要全面讨论这些类型的主题知识，请参见马扎诺和肯德尔

(已出版)。这三类主题知识是：(1) 信息；(2) 心理程序 (mental procedures)；(3) 心理性肌肉运动程序 (psychomotor procedures)。

信息

第一类主题知识目录是信息，更技术性的提法是陈述性知识 (declarative knowledge)。例如，下列的主题知识在本质上就是信息：

- 数学：真分数与假分数的特点。
- 科学：伯努利原理。
- 语言艺术：不同语言风格 (genres) 的特征。
- 社会学科：美国内战的主要事件。
- 健康与体育教育：解释吸毒成瘾的特征。
- 艺术：五线谱的音符。
- 技术：计算机零部件。
- 外语：表示常见概念的词语。

信息的一类题目或任务

一类题目和任务的核心信息涉及与测量主题相关的基本细节和程序。基本细节以术语、事实和时间顺序的形式出现 (参见表 4.2)。从学习的角度来看，信息是相当直接的。实质上，信息不具有生成性，因此，细节知识并不蕴涵新信息的产生。例如，知道“小说人物罗宾汉第一次出现在 19 世纪早期英国文学中”这一信息并不有助于学生得到新知识，也不意味着学生能这样做。

学生通常是通过识别和回忆呈现知识。例如，下面就是一道识别题，教师可以用来检查学生对阿拉莫战役细节的了解情况：

在参加过阿拉莫战役的人物的左边填上“X”：

- _____ Sam Houston
- _____ David “Davy” Crockett
- _____ Jim Bowie
- _____ Daniel Boone

_____ Lt. Col. William Barret Travis
 _____ Capt. George Kimbell
 _____ Col. James Fannin
 _____ Kit Carson

另外，教师还可以使用下面的记忆题来检查相同的细节：

写出参加过阿拉莫战役的主要人物的名字。

表 4.2 基本细节

术语 (Vocabulary terms) 是一种普通的基本细节。信息是否算得上术语取决于在教学上怎样处理它。例如，遗传、后代、有性繁殖、无性繁殖、基因都涉及复杂信息，但它们都可视为术语。若情况如此，要求学生展示的是准确但又有些浅显的理解，而不是深层次的知识。

事实 (Facts) 是一种信息性知识，涉及具体人物、地点、生物、非生物、事件、具体事件原因的细节。

- 具体真实人物与虚构人物的特征：

尼尔·阿姆斯特朗于1966年完成了历史上载人宇宙飞船第一次轨道对接的任务——完成了双子星8号的部分任务，后于1969年7月登上月球，完成了阿波罗11号宇宙飞船的部分任务，成为历史上第一个登上月球的人。

查尔斯·狄更斯根据他自己一生中经历的部分事件撰写了《大卫·科波菲尔》这部小说。

- 具体地点的特征：

丹佛位于科罗拉多州，其海拔高度为5280英尺。

- 具体生物的特征：

赛马海饼 (Seabiscuit) 曾在经济大萧条时期激励许多人，因为它能够在比赛中战胜自身的伤痛夺得冠军。

- 具体非生物的特征：

罗塞塔石碑 (Rosetta Stone) 有助于历史学家明白《圣经》中许多章节的来历。

- 具体事件的特征：

2001年的9·11事件使美国陷入了长期的冲突之中。

- 造成事件发生的原因：

波士顿倾茶事件促使美国殖民地人民反抗英国的统治和《独立宣言》的宣布。

时间顺序 (Time sequences) 涉及两个不同时间发生的事件。例如，1966年11月22日，肯尼迪总统遭遇刺杀与同年11月25日举行葬礼这两件事表现出时间的先后顺序。

信息的二类题目和任务

二类题目和任务的核心信息通常涉及规律和原理（参见表4.3）。根据定义，规律和原理比细节更复杂，因为它们在本质上具有生成性。规律衍生实例，原理产生推论。事实上，要理解一个规律或原理，学生必须具有超越单纯记忆或识别它的能力，学生必须要有举例说明的能力或预测能力。例如，学生仅仅能够复述伯努利

表 4.3 规律与原理

规律 (Generalization) 是人们能够为之提供例证的陈述。例如，“伟大的领袖人物通常产生于危难之中”这句话表达的就是一个规律。某些规律与某些事实是很容易混淆的。事实确定了具体人物、地点、生物、非生物、事件以及引起事件发生的原因的特征。而规律确定同类信息的种类特征。例如，不经过代表讨论通过的征税法是导致美国革命战争爆发的原因之一，这句话传达出的信息就是一个事实。有关战争的普遍原因的信息形成一个规律。

- 真实人物或虚构人物类的特征：
早期殖民地人民历经千辛万苦。
类似独角兽的人出现在许多文化的神话中。
- 地点类的特征：
大城市犯罪率高。
- 生物类的特征：
金毛猎犬是优良猎犬。
- 非生物类的特征：
美国各地山脉的纬度和地形特征差别很大。
- 事件类的特征：
地震发生的强度和持续时间是各不相同的。
- 因果关系类的特征：
酸雨是由未经处理的工业污染与具体的天气情况结合而造成的。
- 抽象事物的特征：
多数社会而不是所有的社会都重视职责和义务。

原理 (Principles) 是一种处理因果关系的特殊规律。如同规律一样，原理是可以为之提供实例的陈述。但是，原理更倾向于预测在一定的情况下要发生的事情。例如，下面这句话就是一个原理：“妇女肺癌患者的增加是与妇女抽烟人数的增加直接成正比的。”一旦学生明白了这个原理，就可以进行具体的推算。例如，如果他们知道在某一具体环境里妇女抽烟人数增加的百分比，就可以推算出妇女患肺癌人数增加的百分比。

原理并不意味着对该原理的理解。然而，理解伯努利原理要求学生具备准确地预测一定情况下所发生的事情的能力。真正理解伯努利原理的学生能够准确地预测两种机翼设计方案中哪个能产生更大的提升力。

信息的二类题目与任务通常是开放性的。例如，下面是一道二类题目，教师可用来考查“细胞膜具有选择性渗透特征”这一规律。

课堂上，我们学习了细胞膜具有选择性渗透特征——让某些物质通过，而将其他的阻挡在外。举例说明细胞膜让什么进去，而又将什么挡在外面。对于所举的每一个例子，解释细胞膜为什么允许它进去或将它阻挡在外。

信息的三类题目和任务

如前所述，信息的三类题目和任务实际上是超出课堂上所教授的知识的。因此，要正确回答三类题目需要进行推理。虽然二类题目也需要推理，但三类题目中所要进行的各种推理和运用是课堂上没有涉及的。创建三类信息题目和任务的一个方法就是运用表 4.4 所描述的认知程序（要详细了解认知程序，参见马扎诺，2004b）。例如，讲授过减数分裂与细胞有丝分裂知识的教师可以让学生描写两个过程的相似点与不同点来创建一个比较任务。

只要向学生呈现了一组相关信息，分类任务就容易创建了。例如，一位艺术教师已经教过下列有关知识：重叠、雕刻加法、明暗法、减法、尺寸大小变化、色彩变化、混色和透视法。要创建一个三类任务，该教师可以让学生将上述各项分成两类或更多类。还可以让学生解释分类的规则，并说明为什么各项都要放在各自的类别里。

表 4.4 运用于三类信息题目和任务的认知过程

比较 (Comparing) 指的是确定两个或三个事物或概念之间的相似点与不同点的过程。严格说来,“比较”指的是确定相似点,而“对比 (contrasting)”指的是确定不同点。但是,许多教师用“比较”这个词兼指两者。

分类 (Classifying) 就是将相似的事物依据其特点分组归类的过程。

创造比喻 (Creating metaphor) 是将表面上或字面上不相关的信息联系起来的一般或基本模式的确定过程。

创建类比 (Creating analogies) 是确定两组事物之间的关系的过程——换句话说,就是确定关系中的关系的过程。

错误分析 (Analyzing errors) 是识别信息呈现或运用过程中的错误并加以纠正的过程。

创造比喻涉及将表面上不相关的信息联系起来。一个典型的比喻就是“爱情是玫瑰”。表面上看,爱情与玫瑰似乎没有关系,但是,在深层次,就可以找到联系。两者都迷人,两者都可以造成痛苦。创建一个涉及创造比喻的三类任务,教师可以让学生描述表面上看起来没有关系的两个事物之间的相似点。要求学生解释“细胞是怎样跟工厂相似的”就是一个创造比喻任务的典例。

创设类比涉及确定关系中的关系。类比的形式是“A 与 B 的关系如同 C 与 D 的关系”。下面就是一位理科教师可能创建的类比任务:

解释以下的每种关系:

- 氧气与人类的关系如同二氧化碳跟植物的关系。
- 原子核与原子的关系如同地核与地球的关系。
- 牛顿与力的关系如同伯努利与空气压力的关系。

或者说,一个三类任务要求学生创设类比而不是解释教师教过的那些东西。

错误分析涉及识别错误的信息并加以解释。为了说明这个问题,假定教师已经在课堂上讲授了太阳以及太阳跟地球的关系的知识。她可能会创建一个如下的错误分析的任务:

萨利知道，如果她在上午11点到中午1点这段时间出去，她很可能被晒伤。她问了6个朋友为什么会这样，得到了6种回答。判断下列回答哪些是错误的，并对每个错误加以解释：

- 回答1：我们在中午比在上午和下午更靠近太阳一些。

- 回答2：中午的太阳造成的灼伤比早上或下午的太阳造成的灼伤更严重。

- 回答3：阳光直射在某一表面上时，被吸收的能量比它们斜照在某一表面上要多些。

- 回答4：太阳当顶时，其光线穿过的大气比太阳降落在天边时穿过的要少些。

- 回答5：空气通常在中午的时候比在一天的其他任何时候都暖和一些。

- 回答6：阳光中的紫外线是造成晒伤的主要原因。

总之，信息的一类题目和任务主要涉及基本细节，如术语、事实、时间顺序等。二类题目和任务主要涉及规律和原理。三类题目和任务要求学生课堂上没有教过的知识进行推理和运用。三类题目和任务有时涉及比较、分类、创设比喻、创造类比与错误分析这些认知活动。

心理程序

第二类主题内容就是心理程序。用技术性术语说，心理程序是一种程序性知识（procedural knowledge），而不是陈述性知识（declarative knowledge）或信息。关于陈述性知识与程序性知识，心理学家罗伯特·斯诺和戴维·洛曼（Robert Snow, David Lohman, 1989）说：“陈述性知识与程序性知识的区别，或更简单地说，内容知识（content knowledge）与过程知识（process knowledge）的区别”，就指导教育实践来说，是最基本的区别之一（p. 266）。心理

学家约翰·安德森 (John Anderson, 1983) 已经用称之为“生产网络 (production networks)”的“如果—那么 (IF—THEN)”结构描写了程序性知识的基本结构。下面的例子 (Marzano, 2001) 展示了用于多位数减法的心理程序的一部分生成环节。

- 1a. 如果 (IF) 目标是做多位数减法,
- 1b. 那么 (THEN) 计算最右边的数。
- 2a. 如果 (IF) 答案已经记录在当前位上, 左边还有一位数,
- 2b. 那么 (THEN) 计算左边的那位数。
- 3a. 如果 (IF) 目标是计算这位数, 而下位无数, 或者说下位数为零,
- 3b. 那么 (THEN) 就将上位数记录下来作为答案, 等等。(p. 24)

像这样的生成环节是人工智能科学家用来编写计算机运行程序的, 这种运行程序在很大程度上模仿了人的思维。它们不是作为教具设计的。但是, 这个例子确实呈现出心理程序的基本特征。心理程序是一系列能够导出产品的步骤。在此例中, 这个产品就是这个多位数减法问题的答案。

乍一看, 有些教师对这样看心理程序会持否定态度, 因为这似乎意味着操作程序只有一个方法, 而且意味着各种程序应该按部就班地教学。事实上, 情况并不是这样。心理学家告诉我们, 学习一个心理 (或心理性肌肉运动) 程序涉及几个阶段。第一阶段称为“认知”阶段 (“cognitive” stage) (Fitts, Posner, 1967)。在这个阶段, 实际上, 学习者并不能有效地操作这个程序, 但他们可以对这些程序的各个步骤有个总的了解。教师可能已经将这些步骤教给了学生, 或者学生自己通过尝试与犯错已经建构起了这些步骤。这两种方法都是正统的方法, 尽管有些认知心理学家认为, 不幸的是, 教师偏爱发现法而不喜欢给学生展示各个步骤的方法 (参见 Anderson, Greeno, Reder, Simon, 2000; Anderson, Reder, Simon,

1995, 1996)。

不管学生是怎样获得对一个程序各步骤的初始理解的, 这些步骤都会在下一个学习阶段, 即“联想”阶段 (“associative” stage) 得到更改或修正。在这个阶段, 初始步骤中的错误被确认、删除, 被更有效的步骤取而代之。即使一套步骤没有瑕疵, 学习者也可能用一套与其拥有的背景知识相一致的步骤, 或者用他们喜欢的风格的步骤去换掉它。在第三阶段, 即“自动化”阶段 (“autonomous” stage), 程序几乎无需加以思索就可操作 (LaBerge, Samuels, 1974)。

心理程序的一类题目和任务

心理程序的一类题目和任务主要涉及能够稳定地、无变化地产出产品的直接程序。例如, 执行多位数减法的心理程序: 只要程序执行正确, 一个正确答案就会产生。构成一类题目和任务的核心心理程序包括单一规则、运算法则和策略 (参见表 4.5)。

通常情况下, 构成一类题目核心的单一规则或运算法则, 是一个更为复杂的程序的组成成分 (将在下一节有关心理程序的二类题目和任务里讨论)。例如, 运用单一大写规则的心理程序内嵌在更为复杂的结构 (mechanics) 编辑的程序中。

表 4.5 一类题目和任务的心理程序

单一规则 (Single rules) 或叫做构成最简单类型的心理程序的小规则。例如, 大写规则就是单一规则的心理程序: 如果一个词在句首, 那么就大写该词的第一个字母。

运算法则 (Algorithms) 是一旦习得, 运用起来没有变化的心理程序。多位数加法和减法就是运算法则的范例。一旦学生学习了这些程序并达到自动化的水平, 他们通常会以相同的方式操作这些程序。

策略 (Tactics) 在运用过程中比单一规则、运算法则的运用表现出更多的变化, 因为人会根据不同的情况使用不同的策略。例如, 人会依据情况和地图的类型改变看地图的程序。

心理程序的一类题目和任务通常以练习的形式出现（参见 Frisbie, 2005），要求学生执行高度结构化的心理程序。例如，下面就是一些数学教师会使用的一类题目，用以评价学生操作整数多位数乘法程序的能力：

$$32 \times 15 =$$

$$17 \times 47 =$$

$$132 \times 25 =$$

$$99 \times 14 =$$

$$121 \times 134 =$$

心理程序的二类题目和任务

二类题目和任务一般处理更加复杂的心理程序，技术上称之为宏观程序（macroprocedures）（参见 Marzano, Kendall, 1996）。之所以称之为宏观程序是因为它们含有嵌套步骤。例如，写作的宏观程序涉及收集信息、整理思路、写初稿、修改、编辑、出版这些步骤。另外，嵌在宏观程序里的步骤通常会涉及内嵌要素。例如，编辑的次程序涉及内套步骤，如总逻辑编辑、语法编辑、常规编辑（editing for conventions）。最后，宏观程序还有这样的特点：它们能产出多个有效的产品。例如，针对同一个题目的两个学生会写出两篇不同的作文，两篇作文都体现出对写作程序的有效运用。

宏观程序在数量上没有单一规则、运算法则和策略那么多。除写作之外，阅读可算得上宏观程序，还有使用文字处理器、上因特网、作曲、做饭、定计划、讨论、使用复杂的计算器、演音乐剧，等等都可算得上宏观程序。表 4.6 呈现了一套带有普遍性的、重要的总体宏观程序。

表 4.6 中的宏观程序有时候内嵌于具体的学科里。例如，解题的程序嵌在数学里，实验调查的程序嵌在自然科学里，做决策的程序嵌在社会学科里，等等。然而这些程序是跨学科的，因为它们涉及多个学科。例如，技术分类涉及解决问题的程序，健康教育和体育教育涉及做决策的程序。

表 4.6 总体宏观程序

做决策 (Decision making) 是制定标准并运用标准从大致相同的选项中进行选择的过程。

解决问题 (Problem solving) 是克服目标追求途中的各种束缚或条件限制的过程。

实验调查 (Experimental inquiry) 是解释所观察到的现象并加以验证的过程。

调查 (Investigation) 是确定和解决困惑和矛盾的过程。

发明 (Invention) 是创造独一无二的能够满足人们需要的产品或程序的过程。

心理程序的二类题目通常比一类题目更开放。例如，这是一个戏剧教师布置学生去完成的解决问题的任务：

你要排演戏剧《我们的城市》，但又没有钱布置背景。实际上，你只可以用盒子做舞台背景材料。画一个草图表明你准备怎样布置这个特殊场景，并解释你是怎样按照场景的需要使用这些盒子的。

心理程序的三类题目和任务

心理程序的三类题目和任务用于检查学生在一定的语境中运用课堂上没有教过的心理程序所能够达到的程度。例如，假设某教师已经以下列形式教过了多位数减法：

$$\begin{array}{r} 375 \\ -291 \\ \hline \end{array}$$

要为这个心理程序设计一个三类任务，该教师仅需改变一下呈现心理程序的基本信息的形式和情景。新的题目形式是一道文字题：

你打算存钱买辆自行车，要花 375 美元。你已经存了 291 美元，那你还要存多少钱？

使这道题成为一个三类任务的不是任务本身的特点，而是看该任务对学生来说是不是新的。如前所述，心理程序的三类题目和任

务要求学生课堂上没有直接教过的知识进行全新的运用。因此，如果该教师在课堂上已经教过多位数减法的文字题，那么这个例子就不是一个三类任务。再举一个例子，假设有一位教师想为写作的心理程序设计一个三类任务。如果该教师已经教过写作程序并且用短篇故事巩固了这个程序，那么该程序的一个三类任务也许应该是写一首诗。

这种对心理程序三类题目的讨论以及前面对信息的三类题目的讨论有助于区分第三章中提到的完全评分标准表中的分值4.0与3.5的意义。让我们回顾前面所谈，能够对课堂上没有教过的知识进行推理和运用的学生可评4.0分。推理涉及信息，而运用涉及心理和心理性肌肉运动程序（下节讨论）。给推理和运用能力评4.0分的主要理由是两者都是处理课堂上没有明显涉及的新环境和新情景。因此，分值4.0表明学生所学已经超过课堂所教的知识。从表3.7中的分数分布情况，我们可以看到只有少部分学生会不同一般地展示出这种能力水平。评分标准表对这种高级水平的表现应该有所反映。事实上，“高级的（advanced）”这个词（很多地区用它表示一个成绩水准）的主要意思是“进步，更大的进步”。分值3.5也表明学生的所学超过课堂明确所教的知识，不过学生进行的推理和知识的运用只有部分是准确的。

心理性肌肉运动程序

顾名思义，心理性肌肉运动程序就是个人因工作或娱乐需要进行身体活动的身体运动程序。篮球中的防守、打棒球、跳民间舞、唱咏叹调、演戏、在城市里开车等，这些活动都属于心理性肌肉运动程序。

跟心理程序一样，心理性肌肉运动程序储存为“如果/那么（IF/THEN）”的结构或生产网络。而学习心理性肌肉运动程序划分的阶段跟学习心理程序划分的阶段相似（参见 Anderson, 1983, 1995; Gagne, 1977, 1989）。尤其是在第一阶段（认知阶段），学生

对所涉及的步骤能够逐步理解，但不能操作。在第二阶段（联想阶段），学生对最初接触或发现的一套步骤进行补充、修改，使得这套步骤用起来更有效，操作起来更得心应手。在第三阶段（自动化阶段），学生进行操练，并优化这个程序以便能够比较轻松自如地操作它。

心理性肌肉运动程序的一类题目和任务

跟心理程序的一类题目一样，心理性肌肉运动程序的一类题目通常注重更为复杂的心理性肌肉运动程序中的各个步骤或技能的训练。例如，在“篮球中打防守”这个高难度程序中，体育教师会集中训练快速横向运动这一心理性肌肉运动技能——将身体横在篮球和被防卫的选手之间。同样的，在训练“在城市中开车”这一复杂的心理性肌肉运动程序时，驾驶教练会集中训练侧位停车。

心理性肌肉运动程序的一类任务不需要用书面测试进行成绩评定。最有效的评价方法就是用身体演示或表演。很显然，学生必须用身体操作心理性肌肉运动程序以展示其运动能力。为了评估学生打防守时表现出的操控心理性肌肉运动程序的能力，体育教师会让学生防守一个试图投篮得分的进攻队员。为了评估学生操控侧位停车的心理性肌肉运动程序的能力，驾驶教练会让学生表演侧位停车。

心理性肌肉运动程序的二类题目和任务

心理性肌肉运动程序的二类题目和任务一般涉及由许多子程序组成的复杂的程序。如前所述，在适当的防守位置做横向运动这一心理性肌肉运动程序的子程序嵌在更复杂的心理性肌肉运动程序中，即嵌在篮球赛中的防守的程序中。侧位停车的子程序嵌在城市闹市区中驾车这一更复杂的程序中。再就是，二类任务依赖于身体对动作的演示。在打防守的情形中，学生的程序操控能力是通过参加篮球比赛来评估的。在城市闹市区驾车的情形中，学生的程序操控能力是通过在城市中实际驾车来评定的。

心理性肌肉运动程序的三类题目和任务

心理性肌肉运动程序的三类题目涉及学生在特定的情况下操控课堂上没有直接学过的心理性肌肉运动程序的程度。例如，如果防守一种特别进攻是二类任务的中心任务，那么防守一种课堂上没有直接学过的进攻便构成了一个三类任务。侧位停车的三类任务就是在非常小的空地上或斜坡上演示侧位停车的程序，前提是这些情形是学生以前没有遇到过的。

题型和任务类型

因此，本书的讨论表明不同类型的知识要求不同的题型和任务类型。也就是用来检查信息最有效的题型和任务类型不同于用来检查心理程序或心理性肌肉运动程序最有效的题型和任务类型。在这一节里，我们讨论一般的题型、任务类型与信息、心理程序和心理性肌肉运动程序之间的关系。在此，我们讨论 5 种题型和任务：强迫选择，书面简答，问答，口头回答与报道，演示与表演。

强迫选择题和任务

强迫选择题和任务是一种教育评价手段。如第三章所述，20 世纪早期的测试活动依赖于可以评判为对或错的试题。这就是强迫选择题的特点。

一般说来，强迫选择题可以分成 6 类（参见表 4.7），强迫选择题很适合于用来检测信息的掌握情况，而且很有效，尤其适合于用来检查术语、事实和时间顺序的掌握情况。因而，强迫选择题和任务经常用于一类任务。在表 4.7 中，传统选择题的例子涉及对“region”这个术语，填空题的例子涉及对“selectively permeable”这个术语的理解。“事实”是表 4.7 中两个多项选答题例子的核心。如配对题的第一个例子所示，强迫选择题可以用来检查心理程序。在这种情况下，学生必须要操作计算程序才能选择出正确的答案来。

表 4.7 强迫选择题的类型

1. 传统式单项选择 (Traditional Multiple-Choice) —— 提供题干和选项，其中有些选项是干扰项，只有一个选项是正确的。
 (题干) “Region” 一词的最佳定义是_____。
 A. An area of land between two bodies of water (干扰项)
 B. An area of land that has common topographical or political features (正确答案)
 C. An area of land that is a specific size (干扰项)
 D. An area of land that has a specific shape (干扰项)
2. 配对 (Matching) —— 提供多个题干和选项。
 (题干) (选项)

A. 3×5	= _____	1. 28	6. 15
B. $7/6$	= _____	2. 1.05	7. 1.28
C. 12×13	= _____	3. 120	8. 114
D. 7×6	= _____	4. 156	9. 42
		5. 22	10. 1.17

 扩展式

人物	活动	时间
A. Kennedy	1. Led U. S. forces in Europe (WWII)	6. about 1790
B. Jefferson	2. Was elected first Roman Catholic president	7. about 1980
C. Reagan	3. Was elected first U. S. president	8. about 1800
	4. Purchased Louisiana Territory	9. about 1860
	5. Hostages released at start of presidency	10. about 1960
3. 二项选择 (Alternative Choice) —— 提供题干和两个相似的选项。
 传统式
 (题干) 用来连接两个子句的词是_____。
 A. a preposition
 B. a conjunction
 变式
 (A. An architect, B. A draftsman) is an engineer who designs buildings.
4. 真/假判断 (True/False) —— 提供需要判断真假的语句。
 如果语句为假，则在语句前面标上 F；如果语句为真，则在语句前面标上 T。
 _____ 1. The first thing to do with an automobile that does not start is to check the battery.
 _____ 2. A cause of premature tire wear is improper tire pressure.

续表

_____ 3. The automobile's onboard computer should be replaced if the automobile drives poorly.

_____ 4. Under harsh driving conditions, an automobile's oil should be changed every 3 months/3 000 miles, whichever comes first.

5. 填空 (Fill-in-the-Blank) ——提供题干, 每空只能填入一个正确答案。

As it relates to the cell membrane, the term selectively permeable means that it allows in _____, but keeps out _____.

6. 多项选答 (Multiple-Response) ——考虑二个或二个以上的正确答案。

传统式

下列哪一个选项可以用作句子结束的标点符号。

1. A period

2. A dash

3. A exclamation point

4. A question mark

A. 1 and 2

B. 2, 3 and 4

C. 1, 3 and 4

D. 2 and 3

变式

请在下列列举的发生在盖茨堡战役的事件前面填上 Y。

_____ 1. Pickett's Charge

_____ 2. the end of the Civil War

_____ 3. Confederate soldiers occupied Culps Hill

_____ 4. Meade's Maneuver

_____ 5. 15 citizens of Gettysburg killed at Devil's Den

书面简答

按照定义, 书面简答是一种“答案构建”题或任务, 与强迫选择题或任务相对。顾名思义, 答案构建题和任务要求学生构建出一个正确的答案而不是辨别一个答案的对与错。例如, 下面就是一个可以用于经济学课堂的书面简答题:

解释汽油价格大幅上涨为什么是供求原理的一个例子。

要回答这个问题，学生必须检索有关供求关系的知识，按照逻辑组织信息，然后解释他们的逻辑。正如本例所示，简短答案构建题和任务常用于信息的二类题目的设计中。本例中的题目要求学生呈现对具体原理的理解。

书面简答也是用来设计信息的三类任务的有效题型。前面谈到过信息的三类任务通常要求学生进行比较、分类、创造比喻、创设类比或进行错误分析，所有这些都容易设计成书面简答题。例如，下面就是一个涉及供求原理的书面简答题：

将供求原理和生产过剩原理进行比较，描述造成两者相似与不同的具体方式，然后描述你通过分析得出的一、二个结论并加以解释。

用书面简答题型来设计心理程序的一、二、三类任务也是很有效的。现在让我们举一个看城市地图的心理程序的例子。教师给学生一张地图，让他们做一系列书面简答题，题目如下：

1. 说出图例中每个符号的意义。
2. 将市区购物商场圈起来，并描绘出它在地图中的方位以及它所处方位的具体地点。
3. 判断下列三者相互之间哪两个离得最远？购物商场，棒球场，还是工厂？解释你是怎样计算出它们的距离的。
4. 描绘出从购物商场到棒球场的最短路线，只用单行线描绘。

书面简答题型也是用来设计心理程序的二类任务的有效题型。例如，下列就是一组可以用来检测实验调查的心理程序的书面简答题：

课堂上，我们讨论了“人们认为天气和气候影响人的心情和个性”这一观点。

1. 描写你所注意到的天气/气候与心情/个性的关系。
2. 根据你所注意到的有关这一关系的情况做一个推论。

3. 解释你将怎样用实验来验证你的推论。
4. 解释你需要什么样的结果来说明你的推论是正确的。

最后，书面简答题型还可以用来设计心理程序的三类题目和任务。下面就是一个看城市地图的心理程序的三类题目：

假设你要重新设计你手里的地图。解释你将会对图例和地图本身所做的改动。对每一处改动，说明它为什么会改进当前版本的地图。

书面问答

依据马克·杜姆（Mark Durm, 1993）的说法，书面问答是最早用于公共教育的评价方式之一。书面问答题和书面简答题很相似，因为学生必须要以书面形式作答。然而，问答题比起简答题，给学生提供更多的句子结构，对所给的测量主题提供更多的信息。现在让我们看下面这个例子：

1858年，亚伯拉罕·林肯代表伊利诺伊州参与竞选美国参议院席位，与参议员斯蒂芬·道格拉斯展开了一场辩论。你们已经拿到两位辩论者的部分发言，阅读道格拉斯和林肯的开场白，然后回答下列问题：

1. 道格拉斯提到了林肯在斯普林菲尔德做的演讲。道格拉斯提到的是哪篇演讲稿？
2. 道格拉斯说这句话“联邦政府自1789年成立至今，一直都存在蓄奴州和自由州”是什么意思？
3. 课堂上，我们讨论了这个观点：“每一个可辩解的提案都有一个可以辩解的反提案”。解释林肯—道格拉斯辩论是如何为这个观点提供例子的？
4. 找出一个令你想起林肯—道格拉斯辩论的现代情景。说明该辩论与这一情景的相似之处与不同之处。

书面问答题通常给学生提供启动反应的信息。跟这个问题一起

提供给学生的还有林肯—道格拉斯辩论的节选，如表 4.8 所示。

表 4.8 林肯—道格拉斯辩论节选

斯蒂芬·A. 道格拉斯 在共和党代表大会提名林肯先生作为他们的候选人竞选美国参议员之前，林肯先生在斯普林菲尔德发表了一篇演说。在演说中，他说道： “分裂之家无可持存。” “我相信本政府不会让半奴隶制半自由的分裂状态长存下去。” “我不希望联邦分崩离析，我不希望这个大家庭灭顶倾覆，而我的确希望它停止分裂。” “它要不全是这样，要不全是那样。” 那就是他在这场竞选运动中提出的基本观点。唔，一旦你们仔细分析，看到它的后果，我想你一个字都不会相信。联邦政府自 1789 年成立至今，一直都存在蓄奴州和自由州，……然而，我们却被告诫：政府将不能容忍这种状态的持续存在，除非它们全都变成自由州或蓄奴州。因此，他说…… 亚伯拉罕·林肯 道格拉斯法官转述了我最近在斯普林菲尔德演说中的两个观点。他说它们是这次竞选运动的焦点问题。他转述我在斯普林菲尔德讲话的第一点，我相信我还记得很清楚。我当时说道：“自从制定了方针，目的明确，公开宣布要结束对奴隶制的鼓吹，五个年头早已过去了。但在此方针的执行期间，对奴隶制的鼓吹不但没有结束反而不断升级。”“我认为，对奴隶制的鼓吹不到危机降临并四处蔓延之时，不会结束。‘分裂之家无可持存。’我相信，政府不会让这种半奴隶制半自由的分裂状态长存下去。”“我不希望联邦分崩离析”——我在引用我演说中的话——“我不希望这个大家庭灭顶倾覆，而我的确希望它停止分裂。它要不全是这样，要不全是那样。它将朝着两个截然相反的方向发展：不是奴隶制的反对者遏止其进一步蔓延，让公众之心得以平静，相信奴隶制正走向灭亡；就是奴隶制的鼓吹者将其传播四方，直至奴隶制从北到南，在所有的州都获得合法地位……” 资料来源：Political Debates Between Abraham and Stephen A. Douglas, by Cleveland (1902) in Baker, Aschbacher, Niemi, and Sato (1992, pp. 43 -47).
--

书面问答题很有效的一个方面就是它们能够将一、二、三类题目的要素全部放入同一个问答题中。该例中的第 1、2 题实质上是一类题目，因为它们处理的是事实性信息；第 3 题本质上是一个二类

题目，因为它涉及一个规律；第4题实际上是一个三类题目，因为它要求学生运用课外知识。

如林肯一道格拉斯任务所阐明的，问答题对于检测信息性知识很有用。有人会辩解说，跟书面简答题一样，问答题用来检测心理程序也是很有用的。这可能是事实，但书面问答题需要高水平的写作能力。假定使用书面简答题足以检测心理程序，而且不那么依赖于写作能力，那么建议用书面简答题（而不是问答题）来检测心理程序。

口头回答与口头报告

从某种意义上说，口头回答与口头报告可当做是口头形式的书面简答和问答。事实上，只要要求学生口头回答问题，林肯一道格拉斯辩论的问答题示例就很容易被转换成口头报告。如果重点是要学生呈现知识而不是表述能力，那么，口头回答和口头报告类似于书面简答。因此，它们是设计信息和心理程序的一、二、三类任务的好方式。实际上，上面书面简答题中提供的每个例子都可以转换成口头回答或口头报告。

在课堂评价中未得到充分运用的口头回答的一种形式很可能就是师生之间的即兴讨论。具体地说，只要时间和情况允许，教师与学生临时讨论一个具体的测量主题就行了。例如，学生在做课堂作业时，教师在教室里踱步，这时有一个学生想提问。觉察到该生在思考一个特殊问题，该教师就及时发起一场讨论。

这种互动方式是潜在的最有效的课堂评价方式之一。为了说明这个问题，研究人员希拉·瓦伦西亚（Sheila Valencia）与她的同事对44名小学生和31名初中生进行了调查（Valencia, Stallman, Commeyras, Pearson, Hartman, 1991）。他们用四种方式来检查学生在各个主题方面的知识：结构式讨论，填空测试，简答测试和书面问答。在结构式讨论中，教师简单地向某一个学生提出与主题知识相关的一些问题，学生给予口头回答。结构式讨论题提供的信息比其他三种评价方式提供的信息多很多。具体地说，其他三种方式提

供的信息只有结构式讨论提供的信息的 34%。正如研究人员所说：“平均起来，学生在访谈期间提供的特别相关的观点，66% 是其他方法没有检验到的”（p. 226）。瓦伦西亚的结论之一就是：“要全面检查一个人的主题知识需要采用多种方法，每种方法都会给这个图景提供特有的信息”（p. 230）。

演示与表演

顾名思义，演示和表演就是要求学生“做事”，用技术术语来说，演示和表演要求学生操作程序。就这点而论，它们是用来设计心理性肌肉运动程序的一、二、三类任务的最佳方式。举个例子说，在一对一的情形下，防守一个篮球选手的一类任务就是要求学生完成整个防守程序。在五对五的情形下，打防守的二类任务就是要求学生在比赛中展示他们的能力。打防守的三类任务就是要求学生做课堂上没有教过的防御进攻的动作。总之，演示和表演是评价心理性肌肉运动程序的自然方法。

题型与知识类型的匹配

用于课堂评价设计的五种题型和任务类型中的任何一种——强迫选择、书面简答、书面问答、口头回答和报告，演示和表演都适合用来检测一定类型的知识。表 4.9 表示各种知识类型跟题型和任务类型之间的关系。

再论课堂评价的设计与评分

课堂评价设计的直觉方法在本章的开头讨论过。除直觉方法可行之外，牢固掌握这三类知识——信息，心理程序和心理性肌肉运动程序，以及它们跟一、二、三类题目的关系可以给教师在进行课堂评价设计时提供很多选择，还可以增进教师对课堂评价设计的理解。为清楚起见，让我们回顾一下本章开头关于课堂评价设计的直觉方法的讨论，其中谈到对繁殖与遗传这个主题的评价设计。通过

表 4.9 知识型、题型和任务型及评价表格

	强迫选择	书面简答	书面问答	口头回答和报告	演示和表演
信息					
一类	×	×	×	×	
二类		×	×	×	
三类		×	×	×	
心理程序					
一类	×	×		×	
二类		×		×	
三类		×		×	
心理性肌肉运动程序					
一类					×
二类					×
三类					×

回顾，我们发现该主题的内容本质上是信息性的。同时我们还意识到该课堂评价设计特别依赖于书面简答题。完全掌握了我们前面所讨论的题型知识的教师可能会使用如下的配对题来设计一类任务：

将每个术语与最能描写它的答案配对，并将正确答案前面的字母写在相应的术语的边上。答案比术语的个数多，所以不是每个答案都可以找到一个相匹配的术语。

Vocabulary term (术语)

Answer (答案)

- | | |
|------------------------|--|
| • heredity | (a) the process that results in an offspring that is an exact copy of the one parent |
| • offspring | |
| • sexual reproduction | (b) the product of the reproductive process of an animal or plant |
| • asexual reproduction | |
| • gene | (c) the process of passing instructions for specifying traits from one generation to another |
| | (d) the process that involves a male and fe- |

male parent

- (e) the product of asexual reproduction
- (f) the element of a cell that carries a single unit of information
- (g) the product of a cell dividing
- (h) the element of a cell that allows the cell to split
- (i) the contribution of the male in the reproductive process
- (j) the part of the cell that houses the chromosomes
- (k) the type of reproduction used by all animals and all forms of bacteria

对于二类题目的设计，教师可能会混合使用传统的多项选择题和书面简答题：

下列哪一个句子最能解释这种现象：如果在田野间生活的昆虫和鸟儿突然大量死去，而又没有别的昆虫和鸟儿来补充，长满开花植物的田野会发生什么？

(a) 所有的植物都会死去，因为鸟儿和昆虫留下肥料使植物生长。

(b) 植物不会受到很大的影响，因为没有鸟儿和昆虫它们同样也能够生存。

(c) 所有的植物都会死去，因为昆虫和鸟儿有助于植物进行有性繁殖。

(d) 所有的植物都会死去，因为昆虫和鸟儿有助于植物进行无性繁殖。

(e) 一部分而不是所有的植物会死去，因为昆虫和鸟儿有助于植物进行有性繁殖。

(f) 一部分而不是所有的植物会死去，因为昆虫和鸟儿有助于植物进行无性繁殖。

最后，对于三类题目的设计，教师可能会决定采用前面谈到的书面简答形式：

解释遗传特征与环境特征之间的差异，然后列举一些你自己拥有的遗传特征与环境特征，最后解释为什么你的行为更受遗传特征的影响或更受环境特征的影响。

如第三章所述，为了进行评价打分，教师会给学生做的每一道题标上“+”（完全正确）、“0”（错/漏做）或“part”（部分正确）。通过检查学生对这三类题目的反应模式，教师可以给学生评出最合适的分数。

再论不合逻辑的题目应答模式问题

如第三章所提到的，有时候学生的题目应答模式似乎没有什么逻辑。例如，某一学生做错了一道一类题目（强迫选择题），但却答对了所有的二类题目。造成这种异常的应答模式的一个可能原因就是某一道题没有涉及到一类知识——较容易的知识，心理程序或心理性肌肉运动程序。记住这一点很重要：具体的信息，心理程序或者心理性肌肉运动程序本身并不存在难易问题，尽管有些知识比另一些知识更复杂些是事实（例如：规律比事实复杂；一个宏程序比一个步骤复杂）。知识在课堂上讲解的明白程度与学生感到难或易有很大的关系。大多数学生不能正确回答某一题时就传达出这样的信号：有关知识在课堂上讲解得不够清楚，不够透彻。为了弥补不足，教师必须将该题重新归为二类题目。

另一种可能就是某一个学生回答问题或是理解内容时很粗心。处理这个问题的最好的方法就是将试卷还给这个学生，并附上书面（或口头）批语，如“如果你答对了第5题，我会给你的成绩评3.0分而不是1.5分，因为做错第5题表明你对一些基本事实的理解有困难”。另外，教师还可以要求学生完成教材中的一类练习、教师提供的练习或学生自己设计的练习来证明他理解了题中呈现出来的基本信息。在任何情况下，教师都要让学生有责任、有机会证明做错

较容易的题是异常现象。这不仅是有效的课堂评价实践，而且还是有效的教学实践。

刚才举的例子涉及书面作答。现在让我们讨论一个用演示或表演进行的一次评价。这个例子再一次涉及篮球比赛中打防守这一动作。学生表演时，体育教师在一旁观察每个学生的表现。首先，教师观察学生是怎样表现出一类协调技巧（component skill）的，然后给每个学生记录分数。跟其他类型的评价打分一样，教师使用符号“+”、“0”或“part（部分）”来评价学生协调技能的表现。下一步，教师观察学生整个防守过程，以相同的方式给学生打分。如果某个学生展示出更复杂的心理性肌肉运动程序的能力（如全面防守的能力），但在恰当体位协调技巧方面犯错，篮球比赛之后，教师可以要该生演示一下正确的体位来纠正这个不规范的动作。

对生活技能主题的评价

第二章谈到课堂活动的参与、作业的完成、行为与协调合作这些生活技能的重要性。这些生活技能在评价打分阶段也都需要进行评价。这样做需要设计一个评分标准表来评价地区或学校已经确定的每一个生活技能主题。表 4.10 列出了参与活动、完成作业、行为、小组活动这些生活技能主题的一般评分标准。需要注意的一点是表 4.10 中的一般评分标准十分笼统。如第二章所讨论的，地区或学校应该替每个年级制定生活技能主题的具体行为标准。例如，对于中学一级的课堂活动的参与，学区或学校可制定以下行为标准：

- 试图回答教师提出的问题。
- 主动说出自己的观点，不用教师点名。
- 注意表述方式。

这些行为标准会写入每个评分标准表里。例如，活动的参与技能达到分值 3.0 的标准应包括下列要素：

学生的课堂活动参与技能符合如下的行为标准：

- 试图回答教师提出的问题。
- 主动说出自己的观点，不用教师点名。
- 注意表述方式。

表 4.10 生活技能主题的一般评分量表

	参与活动	完成作业	行 为	小组活动
4.0	除了达到分值 3.0 要求的表现外，学生参与没有料想到的课堂活动。	除了达到分值 3.0 要求的表现外，学生作业完成出色，超过一般要求。	除了达到分值 3.0 要求的表现外，学生遵守没有特别要求的规则和秩序。	除了达到分值 3.0 要求的表现外，学生表现出以前没有得到强调的维持小组活动的能力和处理人际关系的能力。
3.0	学生参与活动的情况满足课堂期待。	学生按时交作业，符合所有的要求。	学生遵守所有的课堂规则和秩序。	在小组活动中，学生表现出已经得到强调的维持小组活动的能力和处理人际关系的能力。
2.0	除一些特殊情况外，学生参与活动的情况满足课堂期待。	除一些特殊情况外，学生按时交作业，符合要求。	除一些特殊情况外，学生遵守课堂规则和秩序。	除一些特殊情况外，学生表现出已经得到强调的维持小组活动的能力和处理人际关系的能力。
1.0	在教师的帮助或鼓励下，学生参与活动的情况满足课堂期待。	在教师的帮助或鼓励下，学生按时交作业，符合要求。	在教师的帮助或鼓励下，学生遵守课堂规则和秩序。	在教师的帮助或鼓励下，学生表现出已经得到强调的维持小组活动的能力和处理人际关系的能力。
0.0	甚至在教师的帮助或鼓励下，学生参与活动的情况也不能满足课堂期待。	甚至在教师的帮助或鼓励下，学生也不按时交作业，或不符合要求。	甚至在教师的帮助或鼓励下，学生也不遵守课堂规则和秩序。	甚至在教师的帮助或鼓励下，学生也没有表现出已经得到强调的小组活动能力和处理人际关系的能力。

有关一般评分标准需要注意的另一点是它们只涉及整数分值。如第三章所讨论的，对于有些测量主题来说，使用简化版评分量表评分比完整版评分标准表更合适。生活技能主题通常更适合用简化版评分标准表来进行评价。

不像对学科测量主题的评价，生活技能主题是通过教师的观察来评价的。奥德丽·克莱恩萨瑟（Audrey Kleinsasser, 1991）解释说：“教师观察涉及与学生的非正式谈话和教师全天对学生的观察”（p. 9）。里克·斯蒂金斯（Rick Stiggins, 1994）提供了下列教师观察学生的社交技能的例子。

例如，小学教师会观察某一学生跟其他同学的交往，推测该生的社交能力的发展水平。如果该生的社交能力的发展水平能用语言清楚地界定，观察者就能轻松地做出解释。那么，教师通过仔细观察后获得的信息将有助于进一步设计促进学生社交能力发展的策略。因而，这不像判断对/错答案的测试，而像问答测试。我们依赖教师的判断将学生的行为表现定格在由低到高的成绩等级连续体系里的某一处。（p. 160）

观察生活技能主题的一个做法就是每周挑选具体的一天给学生评分。例如，一位中学社会学科教师会利用每周星期五课堂的最后10分钟给学生一周的表现评分——评价他们在参与课堂活动、作业的完成、行为举止、团队合作这几个方面的表现。这样，经过为期9周的评价打分，教师就会在所有的生活技能主题上给每个学生记录9个分数。

对于生活技能主题的有趣的做法就是让学生在同一时间进行自评。学生的自评分可以与教师的评分进行比较。师生的评分差异都能成为师生讨论的话题。还有一种做法就是观察学生时，教师给学生的生活技能表现记分。例如，教师注意到某学生某天在小组活动中表现得特别好，教师会给该生在小组活动方面的生活技能表现记录一个分数。这种“情景分”可以添加到每周星期五课堂结束时记录的分数里去。

小结和归纳

知识可以分为三类：信息，心理程序，心理性肌肉运动程序。课堂评价题目和任务可以划分为三类：第一类，涉及基本细节和技能；第二类，涉及更为复杂的思想内容和程序；第三类，要求学生超出课堂所教授的知识进行推理和运用。课堂评价题还可以用5种形式进行设计：强迫选择，书面简答，书面问答，口头回答与报告，演示与表演。总之，任课教师在设计形成性课堂评价时有很多选择。

第五章 促进学习的评价方法

在第一章所讨论的四个有效的课堂评价原则中，第二个原则，即鼓励学生进步，实施起来也许是最具挑战性的。如第一章所述，反馈对学生的学会产生各种不同的效果。如果反馈处理不当，则会影响学生的学习。第一章中的表 1.2 表明：简单地告诉学生答案的对与错对学生的学会产生负面效应。反馈的正面效应不会自动产生。本章将讨论促进学习的三个方法。

跟踪学生的进步

教师能给学生提供的最直接、最有效的反馈方法就是让学生跟踪自己在主题学习上所取得的进步。有个简单的方法就是发表格（如表 5.1 所示）给学生填，一张表填写一个主题或评价打分阶段选学的主题。该折线图各栏呈现出对学生学习测量主题——概率的不同评价。第一栏呈现该生第一次评价的得分，第二栏呈现第二次评价的得分，如此类推。这一方法可以给学生提供一个可视的成绩表述，让学生看见自己的进步。它还可以给学生提供一个确定自己的学习目标，不用在班上进行成绩排名就可以判断自己的学习是否成功的手段。如第一章所讨论的，动机心理学家马丁·科温顿（Martin Covington, 1992）认为，观点上的这一简单转变有助于促进学生的学习。依据动机学家的说法，在整个评价打分阶段（the whole grading period）让学生看见他们“学习知识的收获”，可以激发他们的“内在（intrinsic）”动机。

表 5.1 学生进步图表

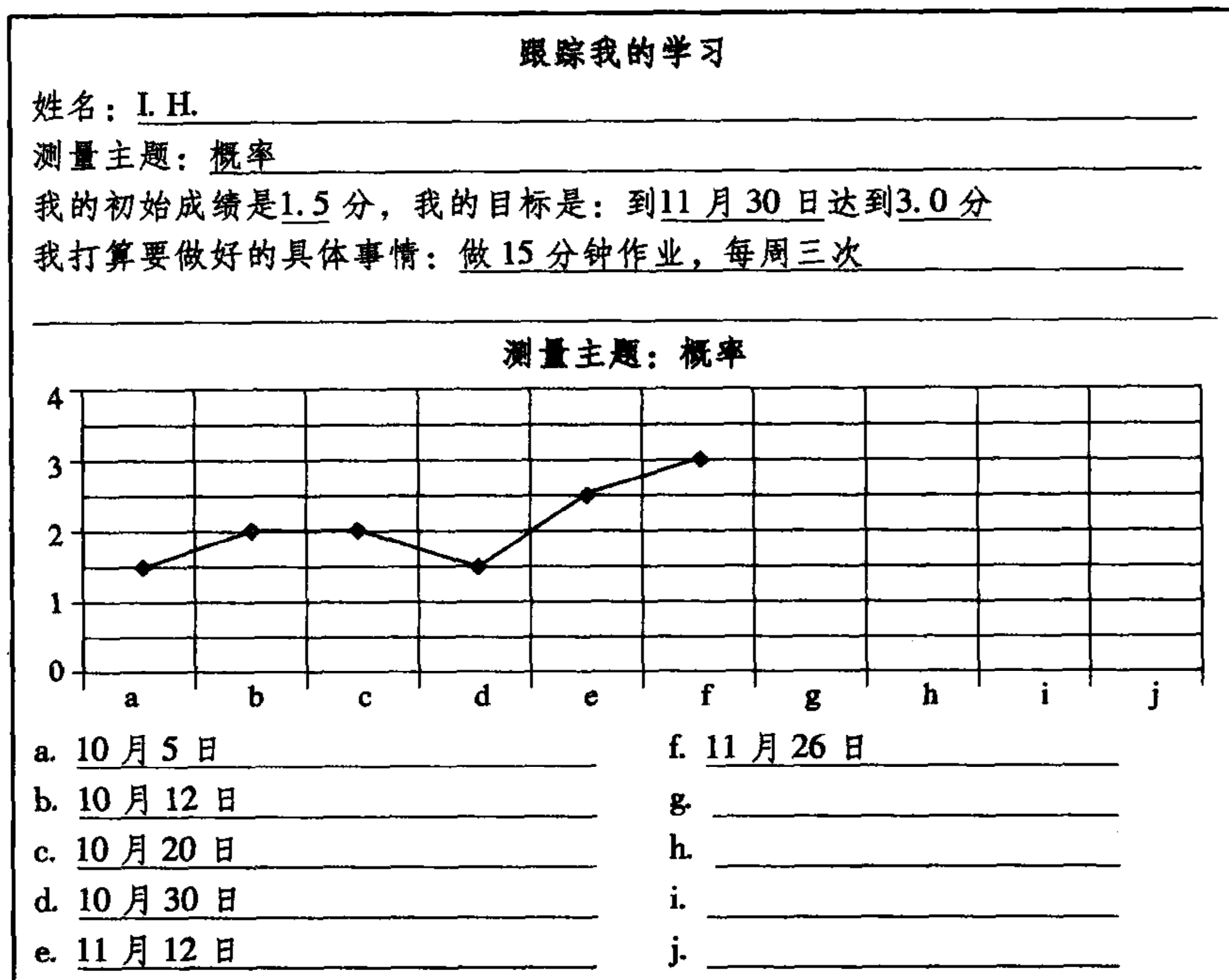
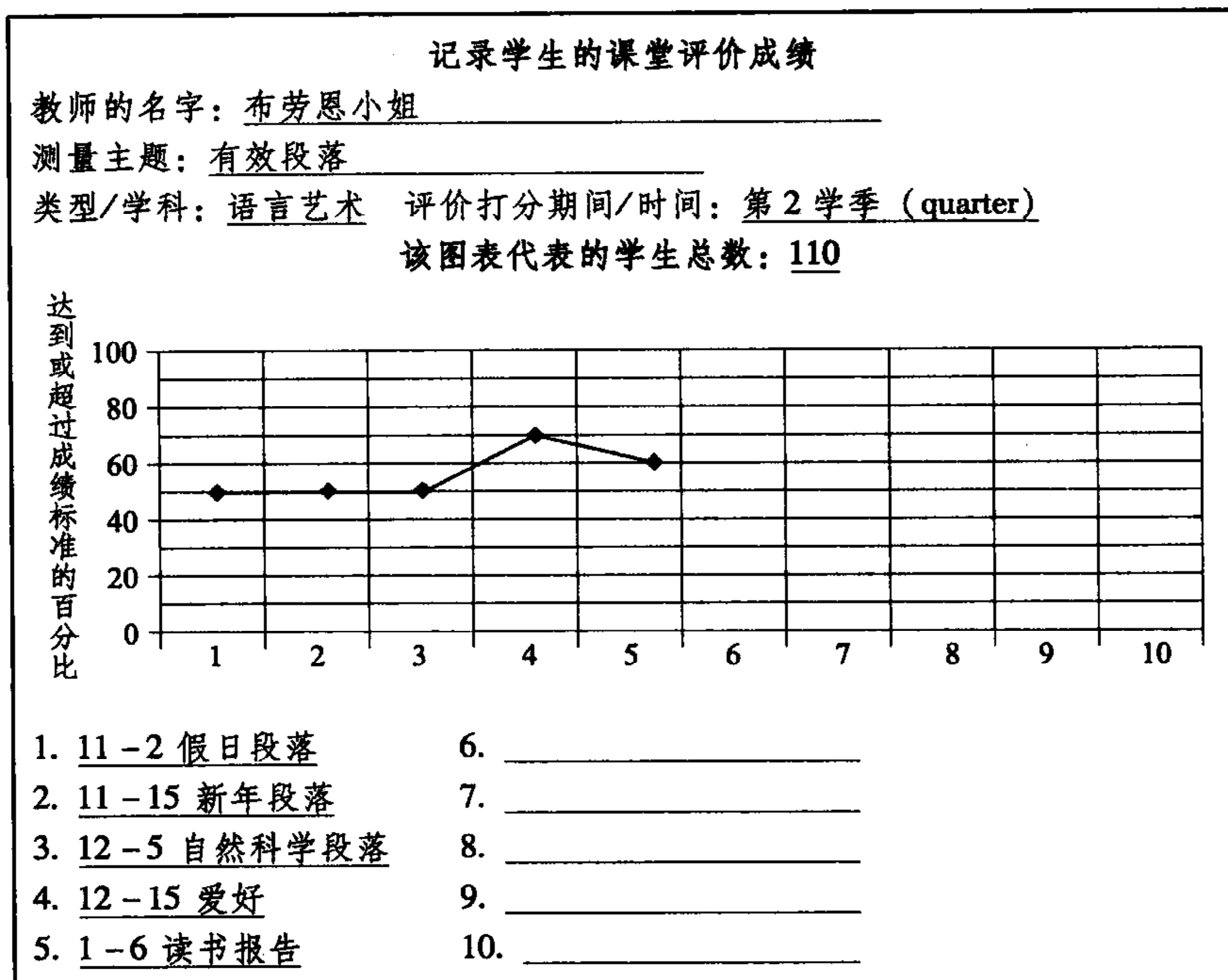


表 5.2 说明一位教师是如何跟踪她负责的 4 个语言艺术班级的进步的。此表与众不同的是, 它用百分比来表示在有效段落这个测量主题上的得分超过某一分数或“具体操作能力标准”的学生数量。第三章讨论了具体操作能力标准的概念。简而言之, 具体操作能力标准指的是评分量表上的某一级分数 (在此情况下, 指的是完整的九级评分量表), 是所有的学生要达到的成绩水准或理解水准。

在表 5.2 中, 11 月 2 日, 在布劳恩小姐所教的班级中, 50% 的学生达到或超过了具体操作能力标准 (performance standard), 这跟下两次评价的结果一样。到 12 月 15 日, 70% 的学生达到或超过了成绩标准。

这类总数据可以为教师与管理人员提供整个年级甚至整个学校的学生进步的快相。教师个人或教师组可以使用此类总数据来确定

表 5.2 班级图表



将来的教学重点。如果总数据表明某个年级达到或超过规定的具体操作能力标准的学生百分比不足,该年级的教师应该齐心协力提高学生对该测量主题的成绩。

促进学生进行反思的方法

促进学生学习的另一方法就是确保学生有机会利用来自于课堂评价的信息对学习进行反思。这样做至少有两种方法可行。

第一种促进学生进行反思的方法就是让学生进行自评。在课堂评价的文献中常常提及学生自评(参见 Stiggins, Arter, Chappuis, Chappuis, 2004)的方法。越来越多的事实证明它可以对学生的学习产生积极的影响(Andrade Boulay, 2003; Butler, Winne, 1995;

Ross, HogaboamGray, Rolheiser, 2002)。在本书的情形中,学生自评指的是学生自己给每次课堂评价打分。例如,在表 5.1 中,学生将教师给他的一系列课堂评价打的分数记录了下来。对于这样的课堂评价,可以要求学生给自己打分。

为了减少学生自评的困难,教师可以给学生提供一个简化版评分量表。表 5.3 呈现了一个简化版五级评分标准表与完整版九级评分标准表。

学生自评的基本用途之一就是给教师的评价提供必要的参照。具体地说,教师会将他在某一具体的评价中给学生打的分数与学生自己打的分数进行比较。打分差异会给师生提供互动的机会。如果某个学生给自己打的分数高于教师打的分数,教师会在该生得到观察分数之前,指出他需要改进的地方。如果学生给自己打的分数低于教师打的分数,教师会指出该生可能没有意识到的自己的长处或强项。

促进学生进行反思的另一个方法则是让学生发表对学习的看法。K. 帕特里夏·克罗斯 (K. Patricia Cross, 1998) 提出了达到这一目的的许多手段。例如,她将“一分钟检查”作为进行自我反思的手段:

在课堂结束前,教师要求学生简要地写出这两个问题的答案:在今天的课堂上,你学到的最重要的东西是什么?
今天课堂上你留下的未回答的主要问题是什么?(p. 6)

一分钟检查的目的就是检查学生“感到最混乱不清的地方 (muddiest point)”。就是让学生说出他们在课堂上感觉最迷惑的东西。教师了解到每个学生最模糊不清之处,依据得到的这些信息进一步策划教学,而后将学生分成小组进行学习。

表 5.3 所示的学生版评分量表可以用来帮助学生确定最模糊不清的地方。为清楚起见,以简化版和完整版评分标准表中的分值 2.0 为例进行说明。给自己评 2.0 分的学生承认自己对有些内容模糊不清。如果再让学生详细说出来,他们会确定他们最模糊不清的内容。

表 5.3 学生版评分量表

简化版评分量表	完整版评分量表
4.0 我熟知它，能将它与未教的内容联系起来。	4.0 我熟知它，能够将它与未教的内容联系起来，而且是对的。 3.5 我熟知它，能够将它与未教的内容联系起来，但有时出错。
3.0 我理解（能理解）所教的内容，没有出错。	3.0 我理解（能理解）所教的内容（包括容易的和较难的部分），没有出错。 2.5 我理解（能理解）所有容易的内容和一些（不是所有）较难的内容。
2.0 我理解（能理解）所有容易的内容，但不理解（不能理解）较难的内容。	2.0 我理解（能理解）所有容易的内容，但不理解（不能理解）较难的内容。 1.5 我理解（能理解）一些比较容易的内容，但出了些错误。
1.0 在老师的帮助下，我明白（能明白）一些教学内容。	1.0 在老师的帮助下，我明白（能明白）一些较难的内容和一些较易的内容。 0.5 在老师的帮助下，我明白（能明白）一些较易的内容，但不明白较难的内容。
0.0 我什么都不懂，什么也不能做。	0.0 我什么都不懂，什么也不能做。

克罗斯（Cross, 1998）认为最为复杂的反思形式是诊断性的学习日志，这涉及学生对四个问题的回答：

1. 简要描述你刚才完成的作业。你认为此作业的目的是什么？

2. 为你做得最成功的一、两个答案提供一个例子。解释你是怎样成功的？
3. 举一个答案做错或答案不完整的例子。为什么这些题做得不对或欠佳？
4. 准备下周的作业时，能有什么改进吗？(p. 9)

克罗斯建议，教师可以通过这些答案，寻找能够为将来设计班级互动、小组互动和学生互动活动提供依据的模式。

这些例子说明了反思的本质——也就是，学生对他们参与课堂活动的表现和对课堂任务的理解进行评价。这样的行为就是德博拉·巴特勒和菲利普·温纳（Deborah Butler, Philip Winne, 1995）所提及的“自我调节学习（self-regulated learning）”。

注重对学生在评价打分阶段末期学习的评价

对学生在测量主题上的学习进行评价的最终目的就是要对学生在评价打分阶段末期的学习进行评价。为清楚起见，参见表 5.4。该表呈现了一个学生在 9 周的时间里学习概率这个测量主题所得的五次测试分数。该生在第一、二次测试中得 1 分，第三次得 2.5 分，等等。在评价打分阶段结束时，教师会计算出该生学习这一测量主题的最后综合成绩。计算最后综合成绩的最普通的办法就是求出平均值。实际上，有人会说基础教育偏爱平均值。许多有关课堂评价的教材或明或暗地推荐平均值（参见 Airasian, 1994; Haladyna, 1999）。如下一章将讨论的，有些情况下计算平均值是有意义的。但是，学生在评价打分期间所得的形成性评价分数不宜计算平均值。表 5.5 有助于说明情况为什么会这样。如前所述，表中的柱形条分别表示学生在五次测试中所得的分数。在此表中，补充了平均值 2.0 分，用虚线表示。要了解用平均值 2.0 分作为学生的最后综合成绩的意义，请参见第三章有关真实分数的概念的讨论。学生每次测试得到的分数由两个部分构成——真实分数与误差。理论上，学生得到的测试分数（指观察分数）大部分由学生的真实分数构成。然而，

学生得分中的误差部分会大大地改变观察分数。例如，一个学生在一次测试中得了 2.5 分，但实际上应得 3 分。那 0.5 分的误差是由于学生看错题或曲解了一些题所致。相反地，一个学生得了 2.5 分，但实际上应得 2 分，因为她猜对了一些题。

表 5.4 一段时间内一个学生在一个主题上的得分柱形图

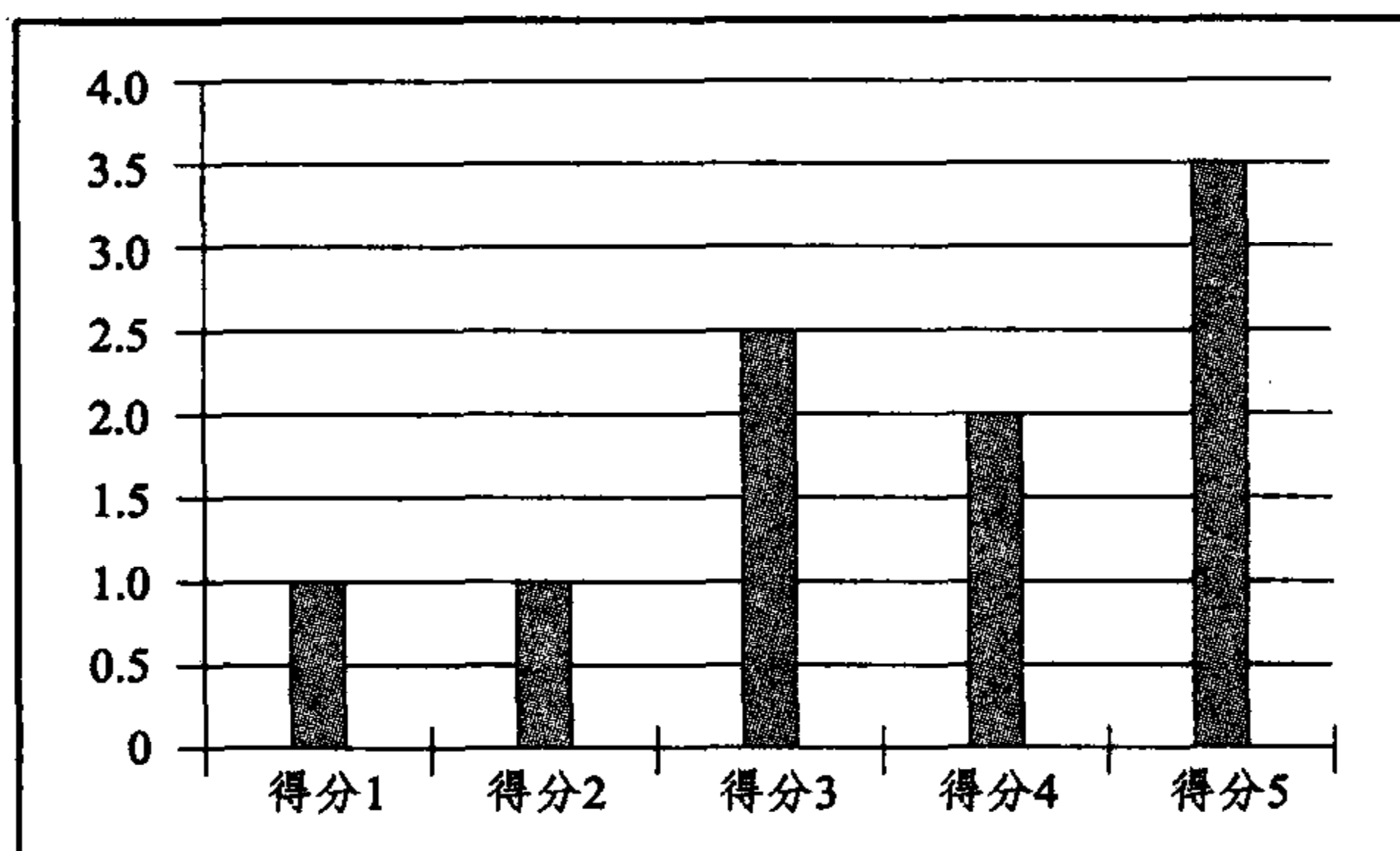
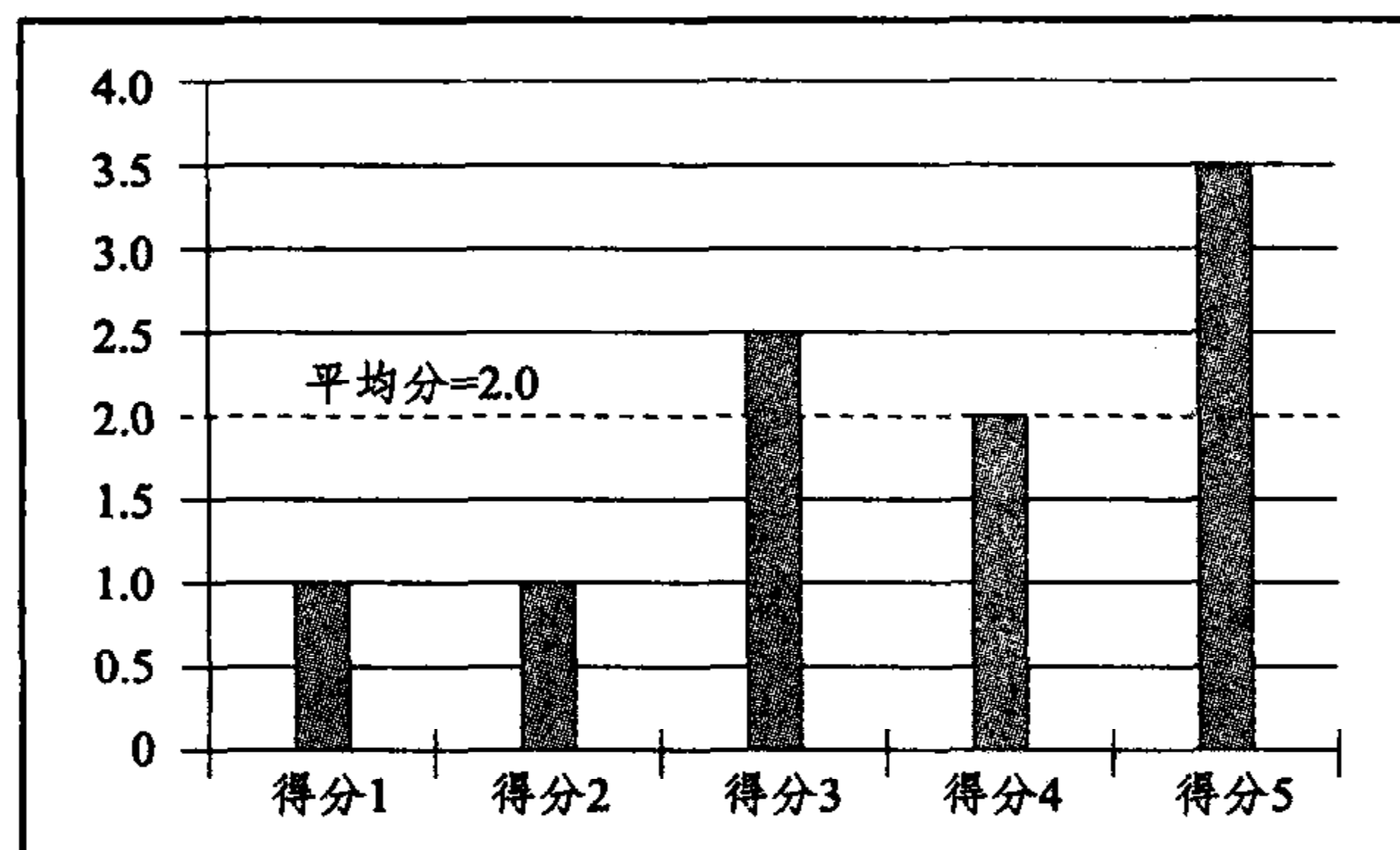


表 5.5 含有平均分数线的得分柱形图



一个学生学习某一测量主题所得的最后综合分数最好视作该生学习该主题所得的最后真实分数。回到表 5.5，如果将该生的平均分视作她在评价打分阶段末期得到的真实分数，我们就会得出这个结论：她得到的真实分数是 2.0 分。这就意味着她已经掌握了一些简

单细节与程序，但对比较复杂的思想内容与程序几乎是无知的。然而，当我们审视整个评价打分阶段所有的分数时发现这种解释是没有意义的。在前两次的测试中，该生的回答表明若不借助任何帮助，她什么也不能做。但是，从第三次开始，她得到的分数从没降到 2.0 分以下，这表明她掌握比较简单的细节与程序是不成问题的。实际上，在第三次测试中，该生对比较复杂的信息和程序表现出一定的理解；而在第五次测试中，该生表现出有一定的能力掌握课堂教学外的知识。很显然，在该例中，平均分 2.0 并不能代表该生在评价打分阶段末期学习这一主题所得的真实分数。

若将学生在形成性评价中各次所得的分数进行平均的话，存在的主要问题是假定平均值在任何一次测验中都没有发生。这个观念是古典测试理论所固有的。实际上，测试理论家常常依据某一个学生的情况来解释真实分数。如建筑师弗雷德里克·洛德（Frederic Lord, 1959），他早期的许多思想都涉及古典测试理论与题项应答理论，他解释说：如果在测验过程中受试者没有发生任何变化，真实分数“常常被定义为受试者参加所有可能的平行测验所得到的分数的平均值”（p. 473）。在这种情形中，平行测验可当做是这样的测验——在这些测验中学生可能得到不同的观察分数，但得到了相同的真实分数。因此，当教师将某一学生的所有考试成绩平均时，她就是在默认该生每次考试的真分数是一样的。换句话说，使用平均值是设想每两次测验所得的观察分数之间的差距是“随机误差（random error）”造成的结果，而求取平均值的行为是想“抵消（cancel out）”每次测验的随机误差（Magnusson, 1966, p. 64）。

遗憾的是，学生每次测验得到的真实分数都相同这一观点跟我们所了解到的学习和为跟踪学习而设计的形成性评价是相抵触的。学习理论与常识告诉我们，学生在评价打分阶段初期也许对某一主题知之甚少或完全不知，但到评价打分阶段末期则知之甚多。学习理论家（为讨论起见，参见 Anderson, 1995）对此现象已做过详尽的描述。特别是认知心理学研究中最普遍的结果表明学习与表 5.6 所示的曲线相似。如该表所示，被考查的学生起初对这一主题毫不

了解，处于零知识状态。尽管这种情况可能并不常见，或者极其少见，但它为我们了解学习的本质提供了有用的观察视角。有趣的是，学习曲线表示从一段时间到另一段时间，学习量起初是很大的，但在学习一段时间之后，学习量从零增加到 20% 多，然后逐渐减少。在认知心理学中，这一学习趋势（根据 Newell, Rosenbloom, 1981）被称之为学习的幂函数，因为描绘表 5.6 中的柱形条的数学函数可用幂函数来计算。

表 5.6 对学习的幂函数的描写

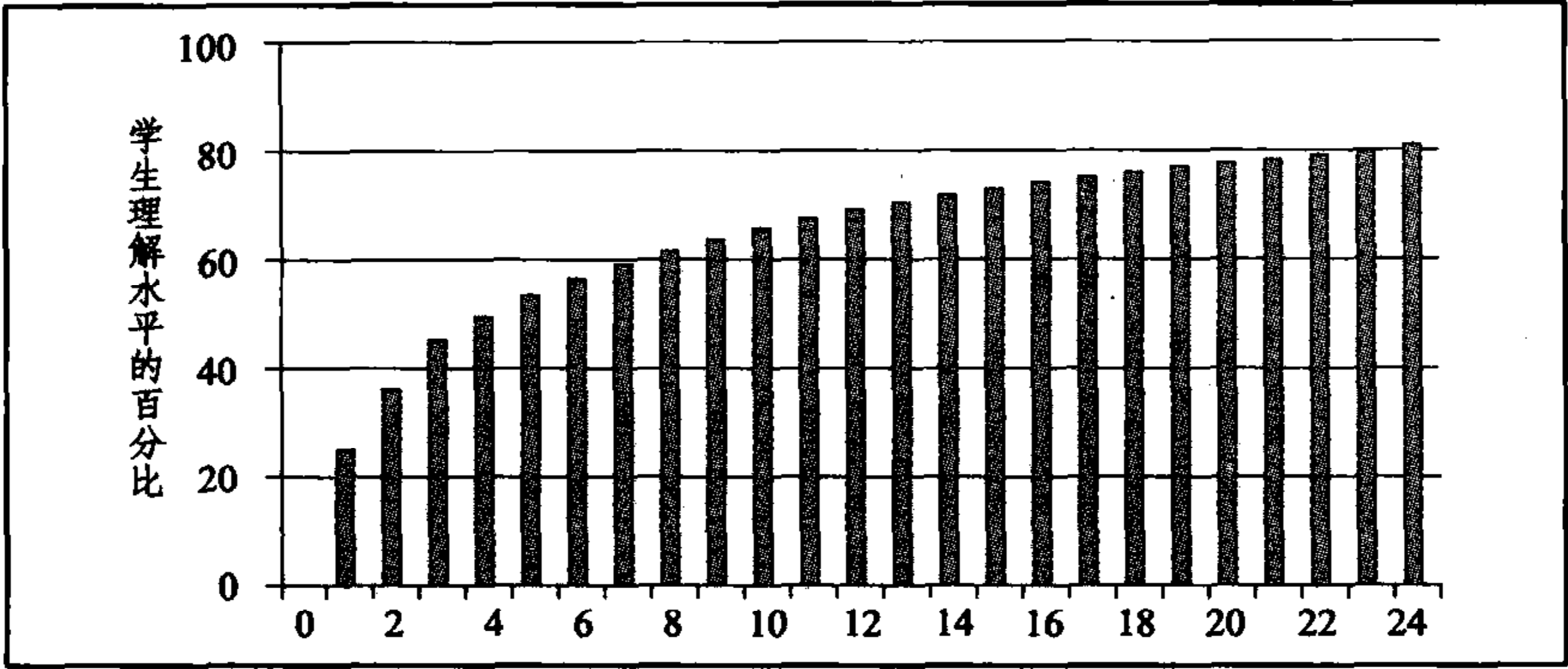


表 5.7 包含幂函数分的柱形图

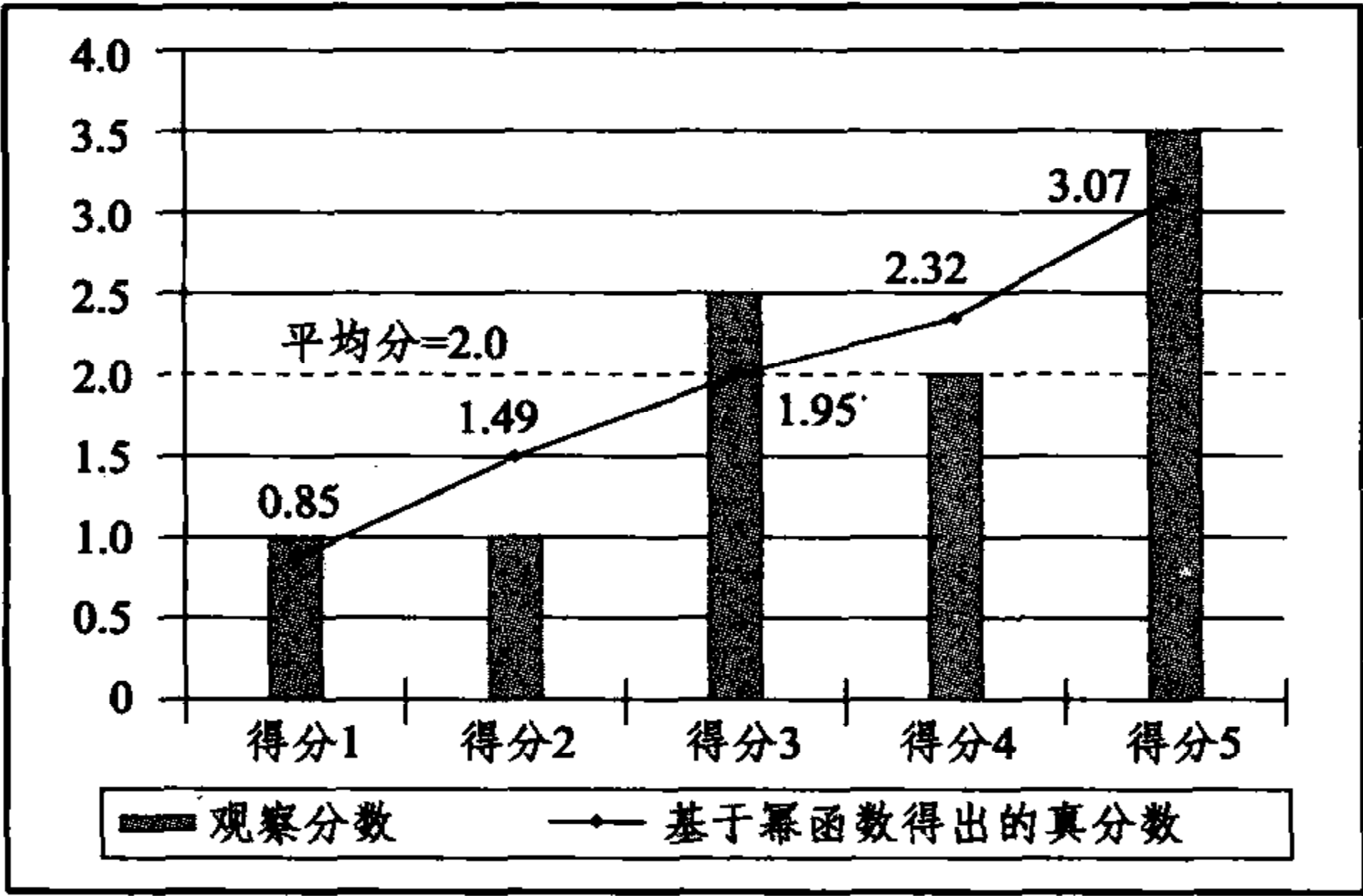


表 5.8 观察分数、平均分数与幂函数算得分数之比较

评 价	1	2	3	4	5	总 差
观察分数	1.00	1.00	2.50	2.00	3.50	n/a
平均分数	2.00	2.00	2.00	2.00	2.00	n/a
幂函数算得分数	0.85	1.49	1.95	2.32	3.07	n/a
观察分数与平均 分数之差	1.00	1.00	0.50	0.00	1.50	4.00
观察分数与幂函 数算得分数之差	0.15	0.49	0.55	0.32	0.43	1.94

技术术语注释 5.1 对幂函数进行了详尽的讨论。扼要地说，幂函数已经被用来描绘大量不同情景中的学习。研究员约翰·安德森 (John Anderson, 1995) 解释说：“自从幂函数得到纽厄尔和罗森布卢姆 (Newell, Rosenbloom) 的证实，它在心理学界引起了极大的关注，研究人员试图弄明白为什么学习在所有的实验中呈现出相同的形式” (p. 196)。根据幂函数在形成性评价中的运用情况，学习的幂函数充分说明运用幂函数能对某一学生在评价打分阶段末期的真实分数做出最恰当的评估。显然，学习的幂函数支持早期探讨的观点：平均分的使用很可能无法对学生在评价打分阶段末期在某一测量主题上的得分提供恰当的评估。实际上，用平均分去评价学生的学习就等于对学生说：“我认为你在这个学习阶段根本没有学习。你在这个主题上每次测试得到的分数高低不同是由测量误差所致。”

学习的幂函数也间接地表明还有一种方法可用来评定学生评价打分阶段末期时所得的真分数。参见表 5.7，该表描绘了用幂函数计算学生每次测验所得的分数。也就是，该生第一次测验得到的观察分数为 1.0，但用幂函数计算出的真实分数为 0.85。该生的第二次得到的观察分数为 1.0，而用幂函数计算出的真实分数为 1.49，等等。在评价打分阶段末期，用幂函数计算出该生的真实分数为 3.07，比平均分 2.0 高出很多。幂函数做出这些估算是通过审查评价打分期间 5 个观察分数呈现出的模式而得到的（参见技术术语注解 5.1 的讨论）。有了这个模式，（数理上）人们有理由认为第二次观察分数为 1.0 有错，人为地减少了观察分数；第三次的观察分数由于人

为地增加观察分数而有错。

对真实分数的计算也仅仅是一种估计而已，这一点需要注意。实际上，测评理论专家告诉我们，学生在某一次测验中所得的真实分数是无法直接观察到的，我们一直在试图估计它（参见 Gulliksen, 1950; Lord, Novick, 1968; Magnusson, 1966）。然而，在测量主题中，幂函数对学生的最后真实分数的计算几乎总是优于对真实分数的平均计算。为明白起见，参见表 5.8。该表生动地展示出幂函数对学生的真实分数的计算优于对真实分数的平均计算。通过对用两种方法（平均法与幂函数法）计算出的真实分数与观察分数之间的差异进行对比，用幂函数计算学生的真实分数的优越性就大大地凸显出来。具体地说，第一次测验的观察分数为 1.0，通过平均计算出的真实分数为 2.0，但通过幂函数计算出的真实分数为 0.85。结果平均分数与观察分数相差 1 个单位，而幂函数算得分数与观察分数相差 0.15 个单位。第二次测验的观察分数为 1.0，通过平均计算出的真实分数为 2.0（平均法会给每次测验的观察分数计算出相同的真实分数），而通过幂函数计算出的真实分数为 1.49。结果平均分数与观察分数相差 1 个单位，而幂函数算得分数与观察分数相差 0.49 个单位。在表 5.8 的最后一栏，我们会发现这 5 次测验的观察分数与平均分数之间的总差为 4.0，而观察分数与幂函数算得分数之间的总差为 1.94。总之，通过幂函数计算出的真实分数比平均计算出的真实分数更接近观察分数，通过幂函数计算出的真实分数比平均计算出的真实分数更符合观察数据。我们将在第六章继续讨论“最合适 (best fit)”这一概念。

因此，讨论的结果是强烈要求用幂函数计算每个学生在评价打分阶段末期在每一个测量主题上所得的真实分数。很显然，教师不必自己进行这些计算。第六章将讨论使用技术方法来解决此问题，也就是使用能进行自动计算的计算机软件。但是，教师也可以用不需要很特别的计算机软件的低科技的解决办法，我将这种解决办法称为“证据收集法 (the method of mounting evidence)”。

证据收集法

证据收集法是一种非常直观与直接的方法。使用这种方法，教师需要使用一个如表 5.9 所示的成绩簿。这种成绩簿的不同之处就是它的每页只能登记 5 个同学的成绩（为方便讨论，表 5.9 只展示了一个同学的成绩）。不像一般的成绩登记簿，一页就可以登记一个班 30 人的成绩，这种成绩簿需要 6 页才能登记完一个班 30 人的成绩。一位教 5 个班级的中学教师（每班 30 人，学生总数为 150 人），共需要 30 页的成绩簿，每班 6 页。尽管这种成绩簿比传统的成绩簿需要的纸张更多，但并不因此而紊乱。使用标准的文字处理软件，很容易创建空白表格。只想试用这种系统的教师记住下面这一点很重要：使用这种成绩簿，只需要考虑中间的步骤。一旦教师相信这个系统将成为保持记录的长久办法，就可购买适当的计算机软件，如第六章所讨论的。

表 5.9 供证据收集法使用的成绩簿

	物质与 能量	力与 运动	繁殖与 遗传	地球的演 变过程	适应性	行 为	完成 作业	参与课堂 活动
阿伊达	1.5 2.5	2.0	1.0 2.5	2.0 3.0	3.0	3.0 3.5	4.0 3.5	2.5 3.5
	2.0	2.0	1.5	2.0	3.0	3.0 3.0	4.0 4.0	3.0 3.5
	2.0	2.0	2.0	2.0	3.5	2.5 3.5	4.0 3.5	3.5 3.5
	2.0	2.5	2.5	2.5	3.5	3.0 3.5	3.5 4.0	3.0 4.0
	2.5	2.5	2.5	3.0	3.5	3.5	4.0	3.5
注：圆圈表示老师已经给该生机会提高前一次测试的成绩。方框表示该生的成绩被判定已经超过框中的那一个分数级别。								

表 5.9 中的第一行呈现出教师在评价打分阶段所教的各个测量主题。在该例中，教师教了 5 个科学主题：物质与能量，力与运动，繁殖与遗传，地球的演变过程，适应性。该教师还跟踪了生活技能主题，如行为、完成作业、参与活动。首先让我们讨论学术主题。

要知道成绩簿是如何使用的，请参看阿伊达在物质与能量这一主题上的得分。在成绩簿的每一列里，先从左上至下，后从右上至下的顺序填写分数。于是，在物质与能量这个主题上，阿伊达得到的6个分数按顺序分别为：1.5，2.0，2.0，2.0，2.5，2.5。注意第二个分数是用一个圆圈圈起来的。这表示教师给了机会让阿伊达提高前一次测试的成绩。这就是证据收集法的活力所在。阿伊达在这个主题上的第一次测试得分为1.5。阿伊达正确回答了一些一类题目，而另一些则答错了，这说明她对该主题比较简单的内容没有完全掌握。但是，教师把试卷还给阿伊达后，跟她谈了考试的情况，指出她做一类题目所犯的错误，解释她的试卷为什么判了1.5分，同时，还给机会让她证明她在测试中做错一类题目不是她对该主题理解的真实表现。换言之，教师给她一个机会证明1.5分没有准确地反映出她应得的真实分数。教师也许让她做完教材中与该主题相关章节后面的一些练习，或编一些阿伊达能够完成的练习给她做，或让她想办法展示真知。

当学生在某一主题的单独测验中所得的分数与他们在课堂上的行为不一致时，教师就应该这样主动跟学生进行沟通。例如，也许在课堂上讨论物质与能量的问题时，阿伊达表现出了对基本细节与过程的理解，这就说明她应该得2.0分。第一次测验的结果似乎与教师平时在课堂上随意获得的有关阿伊达的信息不相吻合。教师用她早期对阿伊达的了解来指导她对阿伊达学习这个主题的评价。基于对阿伊达以前的了解，教师决定要收集更多有关阿伊达对这个主题的理解水平和技能水平的证据。注意教师不是简单地修改测验分数，而是给阿伊达一个机会让她对这个测量主题提供更多的信息。如果阿伊达提供的新信息证实了教师的看法：阿伊达达到了分值2.0的要求，教师就会修改成绩簿上的分数并圈上，表示这个分数是依据补充信息判定的。

表5.9中还有一个值得注意的做法就是将有些分数是用方框框起来，如阿伊达在物质与能量主题上的第4次得分就是加了框的。当教师这样做时，这就意味着她已有足够的证据可以得出这样的结

论：该生已经达到评分标准表上某一个分数级别。到教师给阿伊达输入第4个分数时，她相信阿伊达已经达到了分值2.0所要求的水平。从那次测验起，教师审查阿伊达的答案以寻找她已经超过这个分数的证据。这就是说，从那时起，老师检查阿伊达的试卷以找出阿伊达应得2.0分的证据。这并不意味着阿伊达可以跳过一类题目的测试。实际上，凡是一类题目没有回答正确的试卷都要交还给阿伊达，并告诉她必须改正错误以达到得2.0分要求的准确度。然而，教师认为这些错误是因为努力不够，或推理失误，或二者皆有的原因所致，而不认为阿伊达的真实分数低于2.0分。

证据收集法的灵活之处就是一旦学生提供了足够的证据让教师得出某一学生已经达到一定分数级别的结论，那么那个分数就是该生那段时间的那一时刻学习该主题所得的真实分数。以此为基础，教师寻找下一个更高一级的分数的证据。一旦收集到足够的证据，教师可以得出这样的结论：这就是代表真实分数的下一个分数级别。以此类推，直到评价打分阶段结束。这样，证据收集法为确定学生达到一定的理解水平或技能水平提供依据。

这种方法逻辑性强，可以得到各种研究和理论的支持。第一，第一章中的表1.2说明获得20个百分点是与这种做法联系在一起的：要学生重复做某一个动作直到做好为止。证据收集法也具有一定这种“求精法（mastery-oriented）”的特点。实际上，本杰明·布鲁姆（Benjamin Bloom, 1968, 1976, 1984）和汤姆·古斯基（Tom Guskey, 1980, 1985, 1987, 1996a）早期的研究就是基于类似的方法。证据收集法还可以得到一种称之为“贝氏推理（Bayesian inference）”的统计推理的支持。欲要更深入地了解“贝氏推理”，请参见术语注释5.2。简而言之，贝氏推理认为，对于一个学生在一定时间内所得真实分数的最佳评定必须要考虑我们从该生过去的经历中所了解到的情况。每一次评价都不是依据孤立的信息而进行的，而是依据我们所了解到的有关该生与某一具体的测量主题相关的情况进行评价的。在某种意义上，贝氏推理提出了这样的问题：“了解到学生学习这一测量主题的情况，那么在这次评价中，给她的真实分

数评多少才是最合适的呢？”证据收集法是一种再生形式（generative form）的评价方式，当教师在某一次评价打分中，对一个具体分数的确定没有把握时，他就要利用这种方法去收集更多的信息。

生活技能主题

生活技能主题可以从证据收集法的角度进行探讨，但情况稍有变化。现在让我们查看一下表 5.9 中阿伊达的生活技能得分。这些分数与具体的测验无关。第四章提到，教师可能会利用每周星期五课堂的最后几分钟，给学生在这三个主题上的表现进行评价，一周评一次。教师在行为一栏里记录了 9 个分数，在评价打分期内，一周记录一个。再者，分数是从左上至下，然后又从右上至下进行登记。这样阿伊达的分数按照规定顺序为：3.0，3.0，2.5，3.0，3.5，3.5，3.0，3.5，3.5。注意这些分数当中有许多被框了起来。另外，方框表示教师认为它是学生在一定时间里某一时刻得到的真实分数。因此，阿伊达的第二个分数 3.0 用方框框了起来，表示该分数就是教师给阿伊达那一时刻评定的行为表现的真实分数。注意下一个分数是 2.5，比教师上周的评分低 0.5 分（假定生活技能分是每周星期五记录的）。由于阿伊达的表现分下降了，教师去见了阿伊达，告诉她下周必须将分数恢复到 3.0 分。既然这样，阿伊达只得照做了。然后教师将那个分数框起来，再次申明 3.0 分实际上是阿伊达得到的真实分数。

小结和归纳

有效的形成性评价应该是促进学生学习的。有三个方法有助于这个目标的实现。第一个是让学生填表跟踪自己学习某一具体测量主题的进步；第二个是要学生对在测量主题学习上取得的进步进行不同方式的反思；第三个是评定学生在评价打分阶段末期的真实分数。尤其要注意的是，在形成性评价中，要对学生的最后综合成绩进行有效评定不宜将学生在形成性评价中所得的分数进行平均。这

里还有两种可供选择的成绩评定方法：一个是使用幂函数来评定学生的最后综合成绩；第二个是使用证据收集法来评定学生的最后综合成绩。

第六章 最后综合分数与等级

可以说，在美国教育中最不容易改变的传统就是总成绩。十多年前，林·奥尔森（Lynn Olson, 1995）说：“成绩是美国教育中最神圣的传统之一……事实上，成绩在美国的学校中如同宗教崇拜一样重要。成绩是向家长报告孩子在学校里学习情况的最基本的快捷工具（p. 24）。近来，测评专家苏珊·布鲁克哈特（Susan Brookhart, 2004）表示出相同的看法。她说：“在一个完美的世界里，将不会需要当今学校里所使用的那种成绩……（但是）无论如何，成绩不会很快从学校消失（p. 4）。在这一章里，我们将探讨计算总成绩的最有效的方法。首先，让我们讨论有效的形成性评价系统所需要的一种计算机软件。

在第五章里，我提到了用“高科技（high-tech）”的方法来记录学生在测量主题上的得分。很显然，本书描绘的课堂评价系统需要的是一种能够有效记录与加工学生在形成性评价中获得的各个测量主题分数的计算机软件。有许多软件包可供教师输入学生的一些传统项目上的得分，如课堂小测验、家庭作业、期中考试，等等。但这些软件包一般不可用来跟踪或加工具体的测量主题分数，尤其是形成性评价中的主题分数。本章将描述适用于本书所介绍的形成性评价的软件系统的三个特点：

- 软件允许教师依据自己的设计创建多个测量主题，并轻松输入相应的主题分数。
- 在评价打分阶段的末期，软件应该最准确地呈现学生在每一个测量主题上所得的分数。

- 软件应该提供反映学生成绩的图表与表格。

由求精软件（Excelsior Software）公司开发的品尼高加法系统（Pinnacle Plus System）是一个具有这些特点的软件系统。^① 如果还有别的软件程序能有本章所描述和前面各章所提及的功用，那我就不清楚了。

多个测量主题与轻松输入成绩

对适用于形成性评价的软件系统的最首要的要求就是允许教师创建多个测量主题并能轻松输入学生在每个主题上的得分。我们注意到，有的程序设有现成的项目，如家庭作业、课堂小测验、期中考试，等等。尽管这些项目可以重新标上标签以表示本书所描述的测量主题，但软件如果专门设有此功能是最好的。

使用品尼高加法系统时，教师要做的一件事情就是确定他们将在整个评价打分阶段要跟踪的学科测量主题和生活技能测量主题。本系统还允许教师用自己使用或学校使用的术语来提及这些主题。例如，教师可以将他们的主题说成功率标准，评论性学习标准，基本技能，学习成果，等等。

本系统还允许教师将测量主题划分成更大的范畴。例如，一位语言艺术教师要持续跟踪六个测量主题。除了跟踪六个主题的形成性评价的得分外，教师还希望将其中三个主题合并成一个范畴称为阅读。其余三个主题合并成一个范畴称为写作。

确定了测量主题并将它们划分成更大的范畴之后，教师能够轻易地将评出的分数输入成绩簿。举个例子说，一位五年级的教师跟踪六个数学测量主题（概率，数据分析与分布，集中趋势与离差，测量法，解题，图形与函数）和三个生活技能主题（完成作业，参

^① 编者按：读者须知，美国视导与课程发展学会（ASCD）认可品尼高加法系统，并相信它是当前本书编撰时唯一可用的软件。该软件可用来实施作者所推荐的数据分析方法。然而，本章的技术术语注释则是为那些想自己对数据分析进行编程的人提供的。

与活动，行为举止)，一共九个主题。教师设计了一次测试并进行了评分，内容涉及两个主题：(1) 概率；(2) 图形与函数。教师给每个学生的试卷评了两个分数，一个主题一个分数。现在，教师将学生的成绩成套输入系统。教师首先解释这次测试涉及两个测量主题。当教师准备好给学生输入测试成绩时，一个如表 6.1 所示的屏幕就会出现。在该表中，第一列中的每一格显示一个学生的名字，第一行的每一格显示一个测量主题。第三列、第四列显示教师刚刚评出的测试分数（包含两个主题），第二列显示第一次测试的分数（测试 1），这次测试只涉及一个主题：概率。教师可以用他希望的方式给测试命名，还可将测量主题与任何一次测试链接起来。

表 6.1 供输入测试成绩的品尼高加法系统表

学 生 姓 名	测试 1 概率	测试 2 概率	测试 2 图形与函数
阿伊达 (Aida)	3.0	3.5	3.0
艾什莉 (Ashley)	2.0	2.5	2.5
卡门 (Carmen)	2.5	3.0	3.0
塞西莉亚 (Cecilia)	3.0	3.5	3.0
克里斯蒂娜 (Christina)	1.5	2.5	2.0
珍娜 (Jana)	1.0	1.5	2.5
朱莉娅 (Julia)	3.0	3.0	3.0
马克 (Mark)	2.5	3.0	3.5
托德 (Todd)	0.5	1.0	2.5
惠特尼 (Whitney)	3.0	2.5	3.5

如表 6.1 所示的全班成绩登记表允许教师一次性输入全班所有学生的成绩（品尼高加法系统表有足够行数输入一个班 50 人的成绩）。一个设有完整九级评分标准表所需要的整数分与半分的下拉菜单允许教师给学生在每个主题上的得分选择恰当的分数，不需要教师将这些数字与小数点敲打出来。表 6.1 中所绘制的表格一直都可以使用，因此教师随时都可以查看任何一个学生在评价打分期间所有的测试成绩。

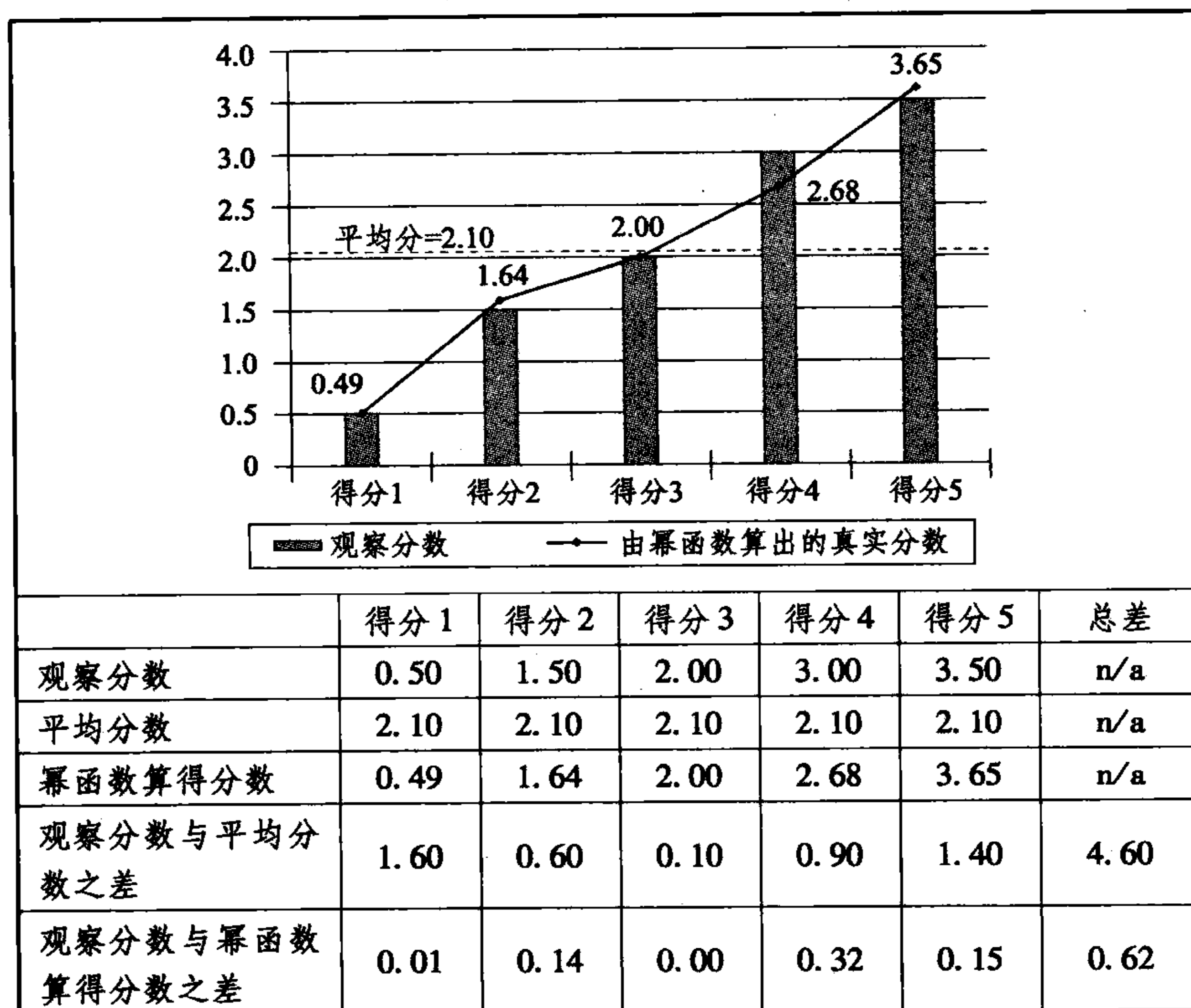
准确呈现评价打分阶段末期的分数

在第五章，我们了解到在评价打分阶段末期，某一测量主题的平均分数不一定是对学生真实分数的最佳呈现。我们还了解到由我们称之为“学习的幂函数”计算出的真实分数通常是比较合适的分数。有效的软件应该给教师提供一个利用幂函数计算出来的学生的最后真实分数和一个平均计算出来的真实分数，还应该提供一种判定用哪种方法计算出的分数与有关数据最相符的方法。

现在让我们以表 6.2 中的一组得分为例。如第五章所述，通过平均计算出的学生在每次测试中所得的真实分数都是一样的。在表 6.2 中，通过平均计算，学生在第一次、第二次及后面的测试中所得的真实分数为 2.10。而由幂函数计算出的真实分数是不同的。幂函数计算出第一次测试的真实分数为 0.49，第二次为 1.64，第五次测试的真实分数高达 3.65。正如我们在第五章所看到的，将由平均计算出的分数与由幂函数计算出的分数进行对比，可以看出它们与实际数据相符的程度。这是通过计算每一个预测得到的真实分数和观察分数之间的差，然后将这些差加起来而得到的。这些数字显示在表 6.2 的下半部分。如前所述，由幂函数计算出的分数更为接近观察分数。具体地说，平均分数与观察分数总共相差了 4.60；而幂函数算得的分数与观察分数总共只相差了 0.62。所以说，幂函数分更“符合”观察数据。软件程序应该为每个测量主题计算这些数据并利用这些数据来确定用哪种方法计算出的分数更合适——通过平均计算出的分数，还是通过幂函数计算出的分数。品尼高加法系统对这些数据进行运算，并选用最合适的真实分数，除非教师对这种真实分数的精确度的选择不做考虑。

也许有人会理智地提出这样一个问题：在什么情况下基于平均计算出来的真实分数会比基于幂函数所计算出来的分数更符合实际呢？当学习过程发生时，难道平均分数总会低估学生在某一测量主题上得到的真实分数吗？答案是肯定的，但是在某些情况下平均分数

表 6.2 最合适的分数



更加合适。为清楚起见，请参见下列这个分数分布模式：2.0，3.0，1.5，2.5 和 2.0。依据学习观，这种分数分布模式是没有什么道理的。该生得到的第一个分数为 2.0，这表明该生能够处理比较简单的细节和程序，但在处理比较复杂的细节与程序时出现大的错误和遗漏。第二个分数为 3.0，这表明该生既能够处理比较简单的细节和程序，又能够处理比较复杂的思想内容和程序。这两次测试得分给人的印象是该生很快就掌握了与测量主题有关的知识。该生的学习趋势是向上的。但是接下来的第三个分数打破了这种模式。第三个分数为 1.5，这表明该生部分掌握了比较简单的思想内容和程序，而在处理比较复杂的思想内容与程序时出现大的理解错误。而后面接下来的两个分数分别为 2.5 和 2.0，继续了这种不平衡的模式。这种不

平衡的模式是怎样发生的呢？至少有两个原因能解释这种不平衡模式的出现。

第一个原因是测量误差。回顾第四章和第五章的讨论，测量误差可能是由多方面的原因引起的。比如学生在考试时误解了试题或者试题本身有歧义。测量误差产生的另一个原因是教师评卷时看错或误解了学生的答案。还有一个原因则是记分错误。例如，教师本来要在某学生的试卷上记3.0，但最后却写成了2.0。最后，测量误差的出现可能是由于学生本身缺乏努力。如果一个学生试图在某一次考试好好努力，而下次考试则不努力，那么第二次考试成绩就很可能无法准确反映该生在这个测量主题上的表现。

某一测量主题的分数分布出现这种不平衡模式的另一个原因可能跟测量主题本身有关。第二章谈到的一个问题就是，测量主题包含的各要素应该是一维的或是紧密相关的，这样处理一个要素的能力就会与处理另一个要素的能力紧密相连。我们把这种特征称作协变性（covariance）。处理某一要素的能力有所提高，则处理另一要素的能力也随之提高。当学区或学校设计测量主题时，其目的应该是确保同一主题中的各要素之间呈现出这种关系——牢固的协变关系。但有时候这个目标达不到。当这种情况出现时，主题评价的精度将受到影响。为清楚起见，我们来看一下第二章中介绍的“要旨阅读”这一测量主题。某学区确定了以下要素为将六年级的阅读水平评为3.0分的指标：

阅读与年级水平相当的材料时，学生能够识别并描述：

1. 明显的与暗含的复杂的因果关系。
2. 明显的与暗含的基本论点。
3. 带有明显的和暗含的单个故事情节的结构。
4. 基于神话的术语。
5. 技术术语。

学生没有出现大的错误或遗漏（omissions）。

要素1、2和3之间是有联系的，它们涉及认知和描述组织模式

的能力。如果你能做其中一个，就很可能也能做其余的。然而，要素4和5之间并没有联系。它们要求理解的是不同类型的知识。因此，学生有可能能做好要素4和5，但却做不好要素1、2和3。用第二章的术语说，这个测量主题是多维的，而这些维度是非协变的。这会造成不平衡的分数模式，如同前面所提到的2.0，3.0，1.5，2.5和2.0的分数模式一样。第一次测试是有关要素1的，第二次测试是有关要素2的。而要素2与要素1之间是有联系的。但是，第四次测试涉及要素4，而要素4跟要素1和要素2并没有联系。所以，该生这次的测试分数很低，这并不是因为该生失去了认知组织模式的能力，而是因为第四次测试所涉及的是跟前两次测试不相关的新维度。

当测量误差或主题的多维性造成了不平衡的分数分布模式时，平均分数是数据分布模式的最佳呈现，而不是幂函数分数。所以，评分软件需要同时提供平均分数和幂函数分数，而且它还应该确定最符合实际数据的那个分数。

学生的成绩图表及表格

在第五章我们探讨了这种方法，那就是学生通过创建自己的成绩图表，跟踪他们自己在各个测量主题学习上取得的进步。我们同时还讨论了这样做的好处。除了学生自己绘制的这些图表外，评分软件在任何时候都应该快速和有效地提供有关各个测量主题的电子图表。如表6.3提供了由品尼高加法系统生成的有关某学生在某个测量主题上的得分情况的图表。与表6.2一样，表6.3提供了一个表示每个测量主题的观察分数的柱状表，由幂函数计算出的分数以及代表平均分数的横线。像这样的表可以制作并打印发给每个学生。教师可以利用这种表在学生—教师会议中加强对话。也许可以给学生打印一份硬拷贝，还可以发给学生家长一份。

表 6.3 品尼高加法系统分数报告单

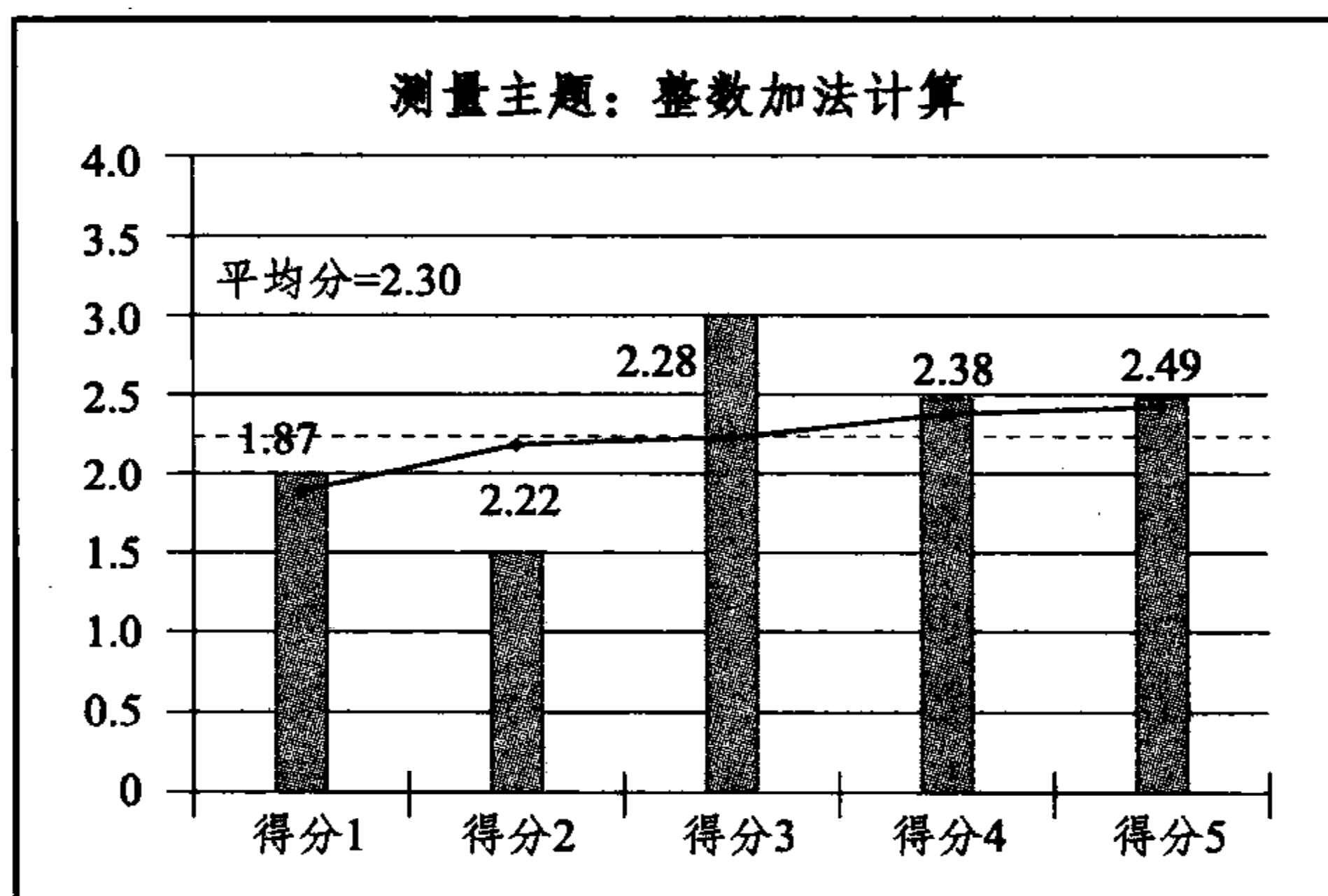


表 6.4 呈现了软件系统应该生成的另一种分数报告单。这种分数报告单是给学生个人的，而且其中只包含学术主题。该软件还可以提供生活技能主题分数报告单，以及两者都包括的报告单。

该报告单既没有反映学生在各个测量主题上的得分情况的柱状图，也没有代表平均分数的横线，但却提供了所有的测量主题和分数。从某种意义上说，该报告单是一幅反映某个学生在所有测量主题上评价得分情况的全景图。报告单顶部的“评价提示”栏，显示了各个主题评价进行的时间。注意教师把一些评价称为“作业”，而另一些称为“测试”。正如第五章所说的，评价有很多多种方式，包括传统的纸笔考试和在课堂上布置而在家里完成的作业。表 6.4 的报告单还提供了各个主题的最后综合分数。正如前面所讨论的，最后综合分数可以是幂函数分数，也可以是平均分数——最后综合分数不管是哪种分数，只要最符合实际数据就行。

如表 6.3 和表 6.4 所描绘的分数报告单可以用做促进教师、学生和家长之间对话的工具。当这些报告单用于这种目的时，它们将成为评价过程中有价值的一部分。

表 6.4 测量主题一览表

学生：简娜						
评价提示						
1. 9月10日：作业	6. 9月24日：作业	11. 10月13日：作业				
2. 9月13日：测试	7. 9月29日：作业	12. 10月15日：测试				
3. 9月17日：作业	8. 10月2日：测试	13. 10月17日：作业				
4. 9月20日：测试	9. 10月5日：测试	14. 10月26日：期末考试#1				
5. 9月22日：测试	10. 10月10日：作业	15. 11月1日：期末考试#2				
	概率	数据分析 与分布	集中趋势 与离差	测量	解题	图形 与函数
1	0.5	1.0				
2	1.5	0.5				
3			2.0	1.0		
4	2.0	2.5				
5			1.5	1.0		
6					1.0	0.5
7			3.0	2.0		
8	3.0	1.5				
9						1.5
10			2.5	2.0		
11					1.5	
12					2.0	
13						2.0
14	3.5	2.0			2.5	2.5
15			2.5	3.0	3.0	3.5
最后综合分数	3.65	1.71	2.49	2.41	2.80	3.65

常见的问题

多年来，本书介绍的系统一直在完善，但问题也在不断地出现，而常见的问题有四个：测试的比重问题，评价的次数问题，如何处

理课堂小测验的问题以及评分系统的精度问题。

关于比重问题

当介绍该系统的时候，有些教师问，为什么在这个系统中一次测试所占的比重不会比另一次测试所占的比重大？然而，难道期末考试不应该比课堂小测验更重要吗？在该系统中，权重是没有意义的。事实上，如果测试要增加权重的话，就无法采用幂函数来计算分数。但这并不是说所有的测试所占的比重都一样。当然，教师通常会使得一次测试所占比重比另一次测试的比重大，其原因在本系统是很容易得到解释的。

占比重较大的测试通常都涉及多个主题。例如，课堂小测验所占的比重小是因为它只涉及一个主题。一位五年级的数学教师可能会进行4次课堂小测验，每次只涉及“计算”这个主题。每次小测验只占10分。但是，期末考试则涵盖了一个学季所教的所有的5个主题的内容。在传统的记分系统中，教师分配给期末考试的分数很可能比小测验多。

在本书介绍的系统中，期末考试会被自动地增加权重，因为它包含了5个分数——每个主题一个。每个分数都会输到相应的主题之下。结果，期末考试所占的比重就比一次课堂小测验所占的比重大。例如，表6.4输入了两次期末考试的成绩——10月26日的期末考试和11月1日的期末考试。该教师举行了两次期末考试：一次是10月26日举行的，包含了4个主题；另一次是11月1日举行的，包含了4个主题。事实上，这两次考试所占的比重是其他任何一次测试的两倍。

应该进行多少次评价？

另外一个经常被提到的问题就是，需要进行多少次评价才能保证在评价打分阶段末期学生的真分数能够得到准确的评估。技术术语注释6.1提供了有关该问题的更为详细的讨论，包括在整个评价打分期间应该隔多久进行一次评价的讨论。简要地说，虽然从数学

的角度来说，最少要进行3次评价才能评出幂函数分数，但是从测试的角度来看，我通常建议每个评价打分阶段每个主题至少应该进行4次评价，最理想的是5次。正如我们在第四章所谈到的，这并不意味着教师要为每个主题进行5次正式的笔头测试。事实上，教师可以利用非正式的评价手段，例如探究性讨论，甚或选做一些家庭作业来获得学生学习各个测量主题的多样成绩。

一个要遵循的总规则是：当你越不确定一个学生在给定的主题上的真实分数时，你要收集的评价信息就应该越多。因此，某个学生只要经过3次测试后，教师就能自信地评出这个学生在给定的主题上的得分是4.0。而对于另一个学生，在经过3次相同的测试后，教师依然不能确定学生的真实分数。这种不确定性可能是由于学生的题目反应模式极其不平衡造成的。如果是这样的话，教师就应该让这位学生完成更多的练习或者是参加探究性讨论来获得更多的信息。教师应该把这些信息当做是附加的评价，输入成绩簿。

与应该评价多少次有关的一个问题就是各次评价打分之间的时间间隔问题。要精确地计算出如表6.3和表6.4所表示的幂函数分数，就必须考虑每次评价打分的时间。举个例子说，假定某教师在新的评价打分阶段的第一天就给每个学生在某一测量主题上的学习进行了评价打分，在第一个、第三周、第七周、第九周又分别进行了一次评价打分。这样，该教师就获得了该测量主题的5个分数，但它们是学习过程中学生得到的不同的分数。数学上，这些分数差异可以通过幂函数计算出的真实分数得到解释。此外，术语注释6.1解释了这是怎样被诠释的。此处的讨论表明计算机软件应该进行这种调整。事实上，在计算幂函数分数时，如果对各评价之间的时间差不进行调整的话，那么对学生的真实分数的评定就会出现偏差。

另外一个与应该评价多少次有关的问题是如何对待缺考的学生以及那些未能交作业（而这些作业又作为一次评价）的学生。在本书介绍的系统中，有一个绝对原则，那就是学生不能因为缺考、晚交或不交作业而得零分。托马斯·古斯基、简·贝利和道格拉斯·里夫斯（Thomas Guskey, Jane Bailey, 2001; Douglas Reeves, 2004）

对此问题所导致的消极影响进行了全面的讨论。简要地说，给学生打零分是不合理的评价，而且还会导致无论是平均还是幂函数计算出的分数都变得毫无意义，但这并不是说教师应该忽视学生不愿意交作业的现象。当然，如果学区和学校规定了将完成作业作为一个生活技能主题，那么学生履行完成作业的职责的程度就可以得到正规的记录了。

为清楚起见，请参照表 6.5 的评分量表，这跟第四章介绍的有些差别。如果每次作业的评分都是按照表 6.5 的评分标准来进行的话，教师将会掌握每个学生完成作业这一行为表现的详细情况。在评价打分阶段的末期，教师可以计算这一生活技能的最后综合分数。像这种倾向于行为的主题，我通常建议通过平均来计算学生的真实分数而不是通过幂函数来计算，因为诸如像按时交作业这样的行为更多的是出于责任感而不是一种技能。在评价打分期间不同的时段里，学生表现出的责任心有大有小。换句话说，他们以前做作业的表现并不能预示他们下次作业的表现。在这种情况下，平均分数作为学生在整个评价打分期间的行为表现方面的最后综合成绩会更合适一些。不过，如果教师认为幂函数分数能更好地反映学生的行为表现，也可以用幂函数分数。

表 6.5 生活技能主题作业完成情况的评分量表

4.0	除了有分值 3.0 要求的表现外，学生的作业完成情况超过一般要求。
3.0	学生按时交作业，符合所有的要求。
2.0	除一些特殊情况外，学生按时交作业，符合要求。
1.0	借助于帮助和鼓励，学生按时交作业，符合要求。
0.0	甚至借助于帮助和鼓励，学生也不按时交作业或不符合要求。

如前所述，我建议每个主题至少要进行 4 次评价，最好是 5 次。如果某学生缺评了好几次，我建议教师把做出补偿性评价的任务转移给学生。也就是说，教师在决定评出最后综合分数之前，告诉该生，他需要更多地理解和掌握有关某个主题的内容。教师可以指导学生做一些书本上的附加练习或者询问学生这种附加的评价信息应

该如何提供。

关于课堂小测验的问题

如本书所描绘的系统在使用时，一个经常出现的问题就是怎样处理无法让学生展示完整版评分量表上各级水平的课堂小测验的问题。如前所述，要想得到 2.0 分，学生必须完全掌握比较简单的细节和程序的全部知识。要想得到 2.5 分，学生必须完全掌握较简单的细节和程序，部分掌握比较复杂的思想内容和程序。因此，如果课堂小测验只涉及比较简单的细节和程序（也就是，该小测验只涉及一类题目），根据完整版评分表中的标准，学生能够得到的最高分只是 2.0 分。这里至少有三个方法可以解决这个问题。

最直接的一个方法就是每次小测验都包含几道二类题目和三类题目。这个做法会让学生在每次小测验中展示分值 2.5 或 2.5 以上级别所要求的知识，尽管每次小测验的重点是比较简单的细节和内容。教师必须将这一点向学生解释清楚：测试中的二类题目和三类题目只当做附加题，所以只有在做完涉及基本细节和程序的题之后才去做它们。简而言之，给每次小测验补充一些二类题目和三类题目好让学生在考试中完全、彻底地展示他们的理解水平和技能水平，以此解决小测验的最高分值为 2.0 这个问题。

如果教师不想在每次小测验中添加一些二类题目和三类题目，那么设定最高分为 2.0 的小测验只能在评价打分阶段的初期进行。例如，某教师计划在评价打分阶段的初期，就某一测量主题进行 3 次小测验，而后又在评价打分阶段的末期进行 3 次测试，后 3 次测试含有二类题目和三类题目，允许最高分值达到 4.0。在前 3 次小测验中，学生能够得到的最高分是 2.0，但在后 3 次的测试中，学生能够得到的最高分为 4.0。假如幂函数程序不是那么特别处罚学习初期缺乏理解和技能的学生的话，那么给学生计算出的最后真实分数就不会受到最高分为 2.0 的小测验的分数的影响。

假定一个学生在初期 3 次小测验中得到 2.0 分，而在评价打分阶段末期的 3 次测试中得到 4.0 分。又假定这些分数是均衡地分布

在评价打分期间，那么幂函数会将该生在评价打分阶段末期的真实分数计算为 3.57 分。这个由幂函数计算得出的分数实际上低估了该生的真实分数，但如果教师意识到该生在各次测试中都获得满分，就会将幂函数计算得出的最后分数改为 4.0 分。

这个例子强调这样一个事实：通过幂函数和平均进行的计算是以一定的假设为基础的。一旦意识到某些假设没有得到证实，教师只需要上下调整最后评定的真实分数来解决幂函数分数和平均分所依赖的数学模型与实际数据不完全相符的问题。

表 6.6 课堂小测验的评分量表

4.0	学生答对所有的题。
3.0	学生答题出现小错误，或漏做的题很少。
2.0	学生答题出现大错误或漏做许多题。
1.0	借助于帮助，学生答对一些题。
0.0	借助于帮助，学生也没有答对一道题。

解决课堂小测验问题的第三个方法（最不可取的方法）就是将小测验的所有得分进行单独处理，也就是为小测验单独设计测量主题。我认为这种方法是最不可取的，因为这类小测验的测量主题的设计不是以共变要素为基础的。实际上，这类小测验是奉行陈旧的（我认为）评价体系而衍生出来的东西。尽管如此，我建议教师使用表 6.6 中的评分量表，或者它的改编版本来给所有小测验评分。表 6.6 中的评分量表很明显是改编了第三章所介绍的简化评分量表。它允许小测验用最大分值 4.0 来评分。分值 0.0 和 1.0 基本上跟以前的要求相似。分值 4.0 表示学生答对所有的题，3.0 表示答题中有小错误，2.0 表示答题中有大错误。如前所述，评分标准要求教师对学生的理解水平和技能水平进行判断。教师不是简单地把测试分数加起来。如我们在前面的章节里所谈到的，这样的判断是评价的关键，是一个有效的形成性评价系统必不可少的。

该评分系统有多精确？

一个非常理智的问题是，该评分系统有多精确？精度这个问题

已经在第三章讨论信度的时候提到过。现在让我们深入探讨这个问题。在第三章我们提到，研究表明由某位教师使用本书介绍的评分方法评出的分数，其信度为 0.719；而用传统的评分法评出的分数，其信度为 0.294。另外，两位教师使用该评分方法分别给每份试卷评分后，再将两者评出的分数综合起来，所得分数的信度为 0.882。如果 4 位教师分别给每份试卷评分后，再将四位评出的分数综合起来，所得分数的信度为 0.901。

这个结果意味着学校或学区可以采用教师设计的课堂评价来对学生评价打分，所评出的分数的精度可以比得上标准化考试和国家级考试的精度。例如，学区和学校每学期会设计并进行一次评价，评价的内容覆盖某一科目所涉及的所有主题。如果两位教师各自评分，那么评出的分数的信度会很高。教师可以将这些测试分数用作附加信息来评估学生学习各测量主题的最后综合成绩。学区或学校也会通过这些分数来获得各个年级或学习小组等的总数据。学区或学校也可以依据这个总数据来设定全区的，或全校的，或各班的教学目标。

信度应该考虑的另一个方面就是对学生在某一主题上的最后真实分数评定的信度。在测评文献（参见 Collins, Sayer, 2001）中，这个问题是人们经常讨论的主题。术语注释 6.2 比较详细地解释了这个问题。简而言之，用幂函数计算学生的真实分数的信度可以通过考查评价打分期间每次幂函数计算出的分数与观察分数之差来计算。之前我们在讨论用平均和幂函数计算最后真实分数谁合适时，曾提到这个问题。在此，观察分数与幂函数算得分数之间的差距有着不同的意义。简单地说，每个学生得到的幂函数分数与观察分数之间的差距越小，则评出来的最后真实分数的信度就越高。在由我主持的一次研究（Marzano, 2006）中，评定了一组 25 名学生的最后真实分数，发现其信度为 0.67。跟其他的评分方法的评分信度相比，这一信度是比较高的了。而且，检查学生最后真实分数评定的信度为考查学生评价的总精度提供了另一条途径。也就是，我们通常只考虑每次评价结果的信度。之前很少有人注重对某一个学生的一系

列分数构成的总数据的信度进行讨论。本书所描述的形成性评价可以让我们从两个角度去分析信度：每次评价的结果以及最后真实分数的评定。

最后综合成绩

一旦每个测量主题的最后综合分数被计算出来，教师就可以计算总成绩（an overall grade）了。在第七章，我们提供了计算总成绩的一些可供选择的方法。事实上，从评价的角度来说，总成绩是相对无意义的。但是，总成绩大概会在未来的一段时间内依然是大多数学校衡量教学质量的标准。在本章开头，我们引用了林·奥尔森（1995）的话：“成绩在美国的各学校获得了如同宗教崇拜般的重要地位（p. 24）。”因此，如果学区或学校希望评出总成绩，那么它所使用的软件系统会将这件事做得又快又好。

如表 6.7 所示，评定总成绩的最直接的方法就是给各个测量主题的最后综合分数分派权数。该表表明教师在整个评价打分期间跟踪记录了 9 个主题——6 个数学主题和 3 个生活技能主题。除概率、集中趋势与离差和解题这三个主题之外，教师给其他主题各分配了 1 个指标的权数，而上述提到的三个主题所获得的权数是其他主题的 2 倍。

表 6.7 分配给每个主题的权数

测量主题	权 数
概率	2
数据分析与分布	1
集中趋势与离差	2
测量	1
解题	2
图形与函数	1
完成作业	1
参与活动	1
行为	1

表 6.8 加权平均值

测量主题	主题的最后综合分数	权 数	质量点数
概率	3.5	2	7.0
数据分析与分布	2.5	1	2.5
集中趋势与离差	2.5	2	5.0
测量	1.5	1	1.5
解题	3.0	2	6.0
图形与函数	2.0	1	2.0
完成作业	2.5	1	2.5
参与活动	2.5	1	2.5
行为	<u>3.5</u>	<u>1</u>	<u>3.5</u>
总计	n/a	12	32.5

通过这个加权方案，教师可以计算出总加权分，如表 6.8 所示。在该表中测量主题的最后综合分数是以 0.5 为递进单位来记录的，这与完整版评分量表中的分值是一致的。我建议用这种方式记录分数，不要记录由幂函数或平均计算出来的精确分数。为清楚起见，假定由幂函数计算得出某个主题的最后综合分数为 2.65，那么可以确定这个分数比通过平均计算得出的分数更符合实际数据。虽然 2.65 这个分值要比半分（half point）分值 2.5 或 3.0 更加精确，但记住这一点很重要：总目标就是尽量准确地、精确地评估学生的真实分数。尽管像 2.65、2.66 或 2.67 这样的分数看起来更加精确，因为它们有更多小数位（事实上，无穷小数位都可能存在），但在这里可以肯定地说，半分分值更加准确。这是因为分值 2.5（或者 1.0、1.5，等等）可依据理解和技能水平来解释：分值 2.5 表示学生掌握了基础知识，部分了解某一主题比较复杂的思想内容。但是，要解释 2.65 所表示的理解与技能水平就比较困难了（2.65 比 2.6 多 0.05，这意味着得 2.65 的学生掌握的知识更多吗？）。所以，我建议教师用幂函数或平均计算主题的最后综合分数，并将其转换成半分分值，这可能是学生所得真分数的最佳表示。例如，一个主题的最后综合分数为 2.65，通过幂函数计算，四舍五入为 2.5。然而，将

分数进行四舍五入时要多加考虑。例如，2.74 不应随意地四舍五入为 2.5，而 2.76 也不应随意地四舍五入为 3.0。对一个主题的最后综合分数进行四舍五入时，应该以教师的最佳判断为依据，尽可能利用可用的信息。

表 6.8 中还有一点值得注意。每个主题的最后综合分数与权数相乘得到的乘积被称之为“质量点数 (quality points)”。总质量点数相加再除以总权数（此例中，总权数为 12）得出加权平均值为 2.71。表面上看，计算加权平均值的建议是与前面几章里有关基于幂函数计算出的真实分数优于基于平均计算出的真实分数的论述相矛盾的。差别就在于前面的讨论假定了一个测量主题中的所有分数表示相同的纬度或共变的维度。在这样的情形中，从一次又一次的评价中得到的分数是相互关联的，其意义是从一次次的评价中，学生会提高在该主题上学习的真实分数。至于说，学生在不同的测量主题上获得的最后综合分数，它们是各自独立无关的。学生在“行为”这个测量主题上的得分与他在“概率”这个测量主题上的得分没有任何关系。每一个主题的最后综合分数代表不同维度的真实分数的评定。当将不同主题的分数合并时，唯一有效的数学计算方法就是以加权或未加权的方式算出这些分数的平均值。

平均分的使用似乎与前面有关学生的得分应该四舍五入到整分或半分的近似值的讨论是相矛盾的。在这里，情况有所不同，因为我们是在将多个主题的分数合并起来，而不是计算单个主题的最后综合分数。这种情况好比是要概括出 9 个人的高度（9 个主题），这通常是通过平均计算而成。凡是对 9 人的高度的分布感兴趣的人都想知道平均值以外的东西。下一章将讨论如何在成绩单上提供更多的信息。

将分数转换为传统的字母等级

一组学术主题和非学术主题的平均分数能够作为一个评价打分期间所有主题的最后综合分数。在这样的系统中，某学生得到的最后综合分数就是该学生学习该课程中的测量主题得到的加权或未加

权平均分。在表 6.7 和表 6.8 的例子中，该生得到的最后综合分数是 2.71。但是，希望沿用传统的 A、B、C、D 和 F 等级评分方式的学校和学区需要进行如下的分数等级转换：

3.00 ~ 4.00 = A

2.50 ~ 2.99 = B

2.00 ~ 2.49 = C

1.50 ~ 1.99 = D

低于 1.50 = F

使用这个转换方案时，我们例子中的这个学生会得到 B 级。当然，这只是一个例子。学校和学区应该建立起自己的转换机制以便能够最好地体现评级的意义。但是不管学校或学区采用什么样的转换机制，他们都应该意识到完全用字母评级是人为的行为。因为在字母评级中，各级分数的划分是任意的。这一特点是使用字母评级的最大的弱点之一。古斯基（Guskey, 1996b）解释说：等级划分的任意性的本质是这种评分体系固有的弱点。

等级范围之间的分界线总是任意而又难以公正的。假如 B 级的范围是从 80 分到 90 分，得 89 分的学生跟得 80 分的学生得到同样的等级，虽然他们的分数有 9 分之差。而得 79 分的学生却会因 1 分之差而只拿到 C 级，因为 B 级的分界线是 80 分。

采用联合法

如前面所描述的，给主题的最后分数加权并计算总成绩的方法是一种补偿性方法（Kifer, 1994），因为使用这种方法计算总分，一个测量主题分数可以“补偿”另一个测量主题的分数。为清楚起见，假定有一个学生获得以下 8 个测量主题的分数：

主题 1 = 1.5

主题 2 = 2.5

主题 3 = 2.5

主题 4 = 4.0

主题 5 = 1.5

主题 6 = 1.5

主题 7 = 1.5

主题 8 = 4.0

这一组分数的（未加权的）平均分是 2.38。根据完整版评分量表中所描述的半分分值标准，这个平均分表示该学生掌握了基础知识——也就是较简单的细节和程序，部分掌握了较复杂的细节和程序。然而，该生在这 8 个主题上的得分有 4 个是 1.5 分。这表明该生对于这 4 个主题，只表现出部分掌握，甚至是部分掌握较简单的细节和程序。显然，另外两个高分 4.0 拉升了学生的一般成绩（typical performance）。这个精心设计的但却简单的例子说明了使用能够生成加权或未加权平均值的补偿法的不足。有时候平均值并不能代表一组分数中的一般分数，因为这组分数之间的差距很大。

可以取代补偿法的另一方法就是联合法（the conjunctive approach）（Plake, Hambleton, Jaeger, 1995）。如果采用联合法来评定总成绩的等级，一个分数就不会拉升另一个分数。更确切地说，总成绩的等级评定是依据测量主题分数模式（score pattern）来确定的。表 6.9 描述了众多选择中的两个联合评级系统。第一个系统具有严密的逻辑。要得到 A 级，所有的主题分数必须达到 3.0 分或 3.0 分以上；要得到 B 级，所有的主题分数必须达到 2.5 分或 2.5 分以上，等等。遗憾的是，该系统在某些情况下非常有惩罚性。举例来说，假定一个学生除了一个主题只得了 1.5 分外，其余的都得 4.0 分，那么使用第一个系统评级，该学生就只能拿 D 级。第二个系统虽然不那么苛刻，但仍然是有要求的。如果使用第二个系统评级，该学生可拿到 C 级。

创建第二个系统的各种变化形式以使特低分数的影响最小化是可能的。但是，通过最后的分析，发现总成绩的评定并不能以完全公正和合理的方式进行。在第七章我们将讨论总成绩评定的其他方法。

表 6.9 两个联合评级系统

系统 1	
等级	分数模式 (Score Pattern)
A	没有主题分数低于 3.0
B	没有主题分数低于 2.5
C	没有主题分数低于 2.0
D	没有主题分数低于 1.5
F	有些主题分数低于 1.5
系统 2	
等级	分数模式 (Score Pattern)
A	没有主题分数低于 2.5, 且多数在 3.0 或 3.0 以上
B	没有主题分数低于 2.0, 且多数在 2.5 或 2.5 以上
C	没有主题分数低于 1.5, 且多数在 2.0 或 2.0 以上
D	没有主题分数低于 1.0, 且多数在 1.5 或 1.5 以上
F	有些主题分数低于 1.0, 或多数不超过 1.5

小结和归纳

我们可以利用各种方法来计算主题的最后综合分数并将这些分数转换成各个等级。适合本书描述的系统所使用的电脑软件具有 3 个特征。第一, 软件应该允许教师能轻松地输入测量主题的多个测试分数。第二, 软件应该提供对学生在每个主题上得到的最后综合分数的最恰当的评估。第三, 软件应该能够提供描绘学生进步的图表。补偿法和联合法是把主题的最后综合分数综合起来进行总成绩计算的两个方法。

第七章 成绩单与基于标准或主题的教育前景

第六章中所传达的信息之一就是单独用字母等级来反映学生的学习成绩并不是反映学生进步的最佳方法。显然，依据以标准为基础的教育观，单独用字母等级表示总成绩（或单独用百分数，甚至是用平均标准分数表示）是远远不够的，因为它们无法提供促进学生所需要的详细的反馈信息。许多评估专家已经认识到总成绩与总分，或综合成绩与综合分数固有的不足并对此进行了讨论（参见 Guskey, 1996b; Marzano, 2000; Stiggins, 1997）。本章的基本观点是，改变成绩单的形式有可能使美国的基础教育发生史无前例的巨大变化。我相信总成绩不再成为衡量教学质量的标准时，美国教育将会发生巨大的变化。但是，考虑到用字母等级表示总成绩的观念在美国社会根深蒂固，学校会先考虑使用既提供这种具有“文化偶像”（cultural icon）意义的总成绩，又提供测量主题的详细信息的成绩单。

含有字母等级总成绩和主题分数的成绩单

改变成绩单最合理的做法就是制作一张既含有传统的字母等级成绩，又含有评价打分阶段各个测量主题的最后综合分数的成绩单。这种成绩单跟我们在第六章提到的关于一组人的平均高度的类比举例是一致的。平均高度跟总成绩相似，但是为了能得到一幅完整的高度分布图，我们必须知道每个人的高度。而个人高度就如同单个

的主题分数。

表 7.1 展示了一张包含了字母等级总成绩和测量主题分数的成绩单。这是一张五年级的成绩单样本，但这种成绩报告形式能够轻易地用于所有的年级——从幼儿园一直到十二年级。为了方便讨论，我们假定学校已经进行了教学任务分派，每门学科都由不同的教师负责。在这个评价打分阶段，对 5 门学科进行了评价，即语言艺术、数学、科学、社会学科和艺术。每位教师都使用第五章和第六章介绍的方法来计算主题的最后综合分数。各学科中的主题分数都用长条图形来表示。注意每一门学科都包含了学术主题和生活技能主题（参与活动，完成作业，行为，小组活动）。

在成绩单的顶部，是传统的 A、B、C 字母等级，用以报告学生每一学科的学习成绩。这些最后的综合成绩都是用第六章所说的补偿法或联合法计算出来的。实际上，成绩单的上半部分是相当传统的，但是由学科中的主题分数提供的详细情况并不传统。主题分数给学生和家长提供了一个便捷的途径来了解学生学习所有主题的表现，而这种学习表现对成绩的计算具有重要的影响。

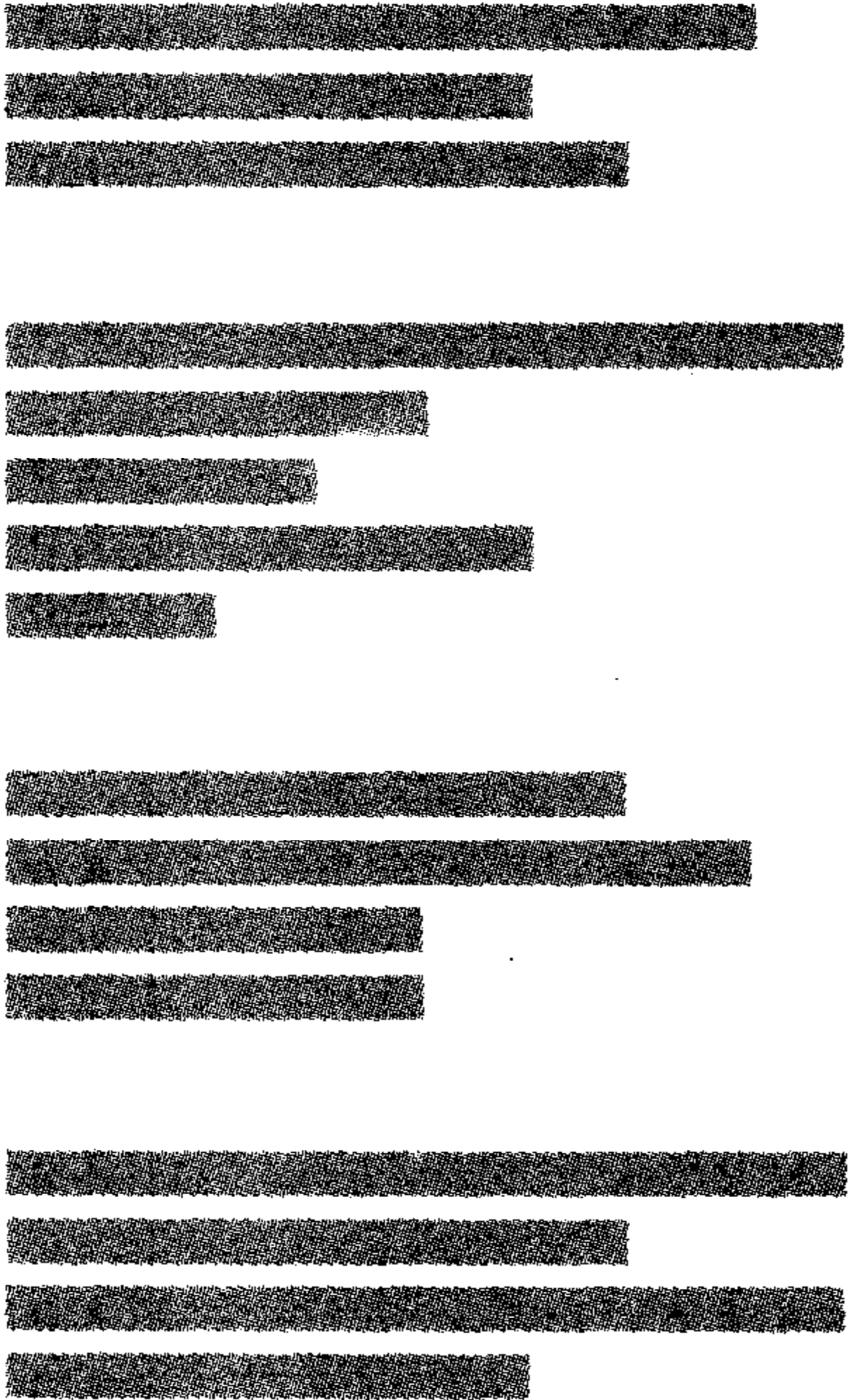
像这样的成绩单还可以附上一张传统的总成绩单，上面列有已修课程、获得的学分（在中学阶段）以及总平均成绩（GPA）。尽管这样的成绩报告方式并不理想，但它的确有两个优点。首先，它将人们熟悉的东西联系了起来，因为它报告总成绩和传统的总平均成绩。其次，它报告评价打分期间所有测量主题的最后综合分数。因此，相比起现在使用的传统版本的成绩单，它提供了更多的信息。此外，它还让我们看到更有效的成绩单会是怎么样的。

不含字母等级总成绩的成绩单

不含字母等级总成绩的成绩单将主要报告测量主题的分數。为清楚起见，请参见表 7.2。表 7.2 也是一张五年级的成绩单样本。这张成绩单虽然没有 A、B、C 字母等级总成绩，但它依然是根据学科和主题来记分的。例如，在这个评价打分阶段，对数学中的这五个

学术主题进行了评价：数进制，预算，加法/减法，乘法/除法和比率/比例/百分比。这五个主题的平均分是 2.80，记在成绩单顶部的数学学科后面，也没有转换成字母等级。

表 7.1 含有等级成绩和主题分数的成绩单

姓名：阿伊达·海斯蒂德 地址：某街 123 号 城市：某镇，CO 80000 年级：五年级 指导教师：贝克小姐		学科： 语言艺术 B 数学 B 科学 D 社会学科 A 艺术 B
语言艺术 阅读 识字与词汇 要旨阅读 文学分析 写作 语言规约 组织与中心 研究与技巧 评价与修改 写作技巧应用 听与说 理解 组织与演说 口头媒介的分析与评价 说话技巧应用 生活技能 参与活动 完成作业 行为 小组活动	3.5 2.5 3.0 4.0 2.0 1.5 2.5 1.0 3.0 3.5 2.0 2.0 4.0 3.0 4.0 2.5	

续表

数学		
数进制	4.0	
预算	2.5	
加法/减法	3.5	
乘法/除法	3.5	
比率/比例/百分比	0.5	
生活技能		
参与活动	4.0	
完成作业	2.5	
行为	3.0	
小组活动	2.0	
科学		
物质与能量	1.5	
自然力	2.5	
生物的多样性	1.0	
人类的特性	2.0	
生物的相互依存	0.5	
生活技能		
参与活动	1.5	
完成作业	2.0	
行为	3.0	
小组活动	1.0	
社会学科		
文化的影响	3.0	
时事	3.5	
个人职责	4.0	
政府代表制	4.0	
人权与民权	3.0	
生活技能		
参与活动	4.0	
完成作业	4.0	
行为	4.0	
小组活动	3.5	

续表

艺术		
艺术宗旨	1.5	
艺术技能	2.5	
艺术与文化	3.0	
生活技能		
参与活动	2.5	
完成作业	2.0	
行为	4.0	
小组活动	3.0	
资料来源：What Works in Schools: Trans lating Research into Action by Robert J. Marzano, 2003, pp. 41 -42. 版权© 2003 ASCD, 经同意改编。		

表 7.2 不含字母等级成绩的成绩单

姓名：阿伊达·海斯蒂德 地址：某街 123 号 城市：某镇，CO 80000 年级：五年级 指导教师：贝克小姐		学科与主题： 语言艺术 2.54 参与活动 3.20 数学 2.80 完成作业 2.70 自然学科 1.50 行为 3.60 社会学科 3.50 小组活动 2.40 艺术 2.33	
语言艺术			
阅读			
识字与词汇	3.5		
要旨阅读	2.5		
文学分析	3.0		
写作			
语言规约	4.0		
组织与中心	2.0		
研究与技巧	1.5		
评价与修改	2.5		
写作技巧应用	1.0		

续表

听与说		
理解	3.0	
组织与演说	3.5	
口头媒介的分析与评价	2.0	
说话技巧应用	2.0	
生活技能		
参与活动	4.0	
完成作业	3.0	
行为	4.0	
小组活动	2.5	
语言艺术的平均分数	2.54	
数学		
数进制	4.0	
预算	2.5	
加法/减法	3.5	
乘法/除法	3.5	
比率/比例/百分比	0.5	
生活技能		
参与活动	4.0	
完成作业	2.5	
行为	3.0	
小组活动	2.0	
数学的平均分数	2.80	
科学		
物质与能量	1.5	
自然力	2.5	
生物的多样性	1.0	
人类的特性	2.0	
生物的相互依存	0.5	
生活技能		
参与活动	1.5	
完成作业	2.0	
行为	3.0	
小组活动	1.0	
科学的平均分数	1.50	

续表

社会学科		
文化的影响	3.0	
时事	3.5	
个人职责	4.0	
政府代表制	4.0	
人权与民权	3.0	
生活技能		
参与活动	4.0	
完成作业	4.0	
行为	4.0	
小组活动	3.5	
社会学科的平均分数	3.50	
艺术		
艺术宗旨	1.5	
艺术技能	2.5	
艺术与文化	3.0	
生活技能		
参与活动	2.5	
完成作业	2.0	
行为	4.0	
小组活动	3.0	
艺术的平均分数	2.33	
资料来源：What Works in Schools: Trans lating Research into Action by Robert J. Marzano, 2003, pp. 41 -42. 版权© 2003 ASCD, 经同意改编。		

如前所述，每位教师都要给各学科中的生活技能主题（参与活动，完成作业，行为，小组活动）评出最后综合分数。这5门学科中的生活技能主题的平均分记在成绩单的顶部。这样，在此例中，学术主题分数与生活技能主题分数根本就没有综合在一起。这是一个很重要的区别。表7.1中的成绩单比较传统，报告了单个的学术主题和生活技能主题分数，但它将同一个学科中的所有主题分数综合成总成绩。这种做法把具体的学术主题跟生活技能主题混合起来。虽然这两类主题都很重要，但是把它们混合起来没有意义。

表 7.3 传统科目和相关的跨课程主题

传统科目	跨课程测量主题
阅读	理解
写作	段落结构
数学	解题
科学	提出假设与验证

解决这个问题的变通方式就是确定一些跨学科测量主题。例如，假定一个学校和学区已经确定表 7.3 中的四个测量主题为多个学科所涉及的主题。每个主题都属于一门传统的学科。每个主题都代表了一个非常重要的信息和技能领域，即便是在其他学科领域这些主题也需要巩固加强。但这并不意味着每堂课、每个学季都要涉及这些主题。准确地说，学校和学区可能会让非数学类的选修课涉及“解题”这个主题，而非科学类的选修课程涉及“提出假设与验证”等主题。将这些跨学科测量主题的最后分数综合起来，以表 7.2 所示的生活技能相同的方式记分，这类主题似乎可以取名为信息与推理。

当然，像表 7.2 这样的成绩单需要附带一张反映学生在一定时间内不断取得进步的成绩单。表 7.4 就展示了这类成绩单的一个版本，为简单起见，这里只列出了语言艺术中的各个主题。其他学科都可以采用相同的形式。该表中的栏目表示某一学年内的各个学季（quarter）或评价打分阶段（第一学季、第二学季等）。注意并不是所有的测量主题都会在每一个学季中（这在第二章讨论过）进行评价。对于那些经常进行评价的主题来说，基于幂函数计算出来的真实分数可能是最合适的分数。对于那些一个学年内只进行一两次评价的主题，平均分则是最合适的分数。为了遵守原则，表 7.4 中的最后两列列出了平均分和幂函数分。

表 7.4 成绩单样本

语言艺术	第一学季	第二学季	第三学季	第四学季	平均分	幂函数分
阅读						
识字与词汇	3.5	3.0	3.5	4.0	3.50	3.67
要旨阅读	2.5	2.5	3.0	3.5	2.88	3.27
文学分析	3.0	3.0	3.5	3.5	3.25	3.49
写作						
语言规约	4.0	4.0			4.00	
组织与中心	2.0	2.0	2.5	2.5	2.25	2.49
研究与技巧	1.5		2.0		1.75	
评价与修改	2.5	2.0	2.5	3.5	2.63	2.90
写作技巧应用	1.0	1.5	2.0	2.0	1.63	2.15
听与说						
理解	3.0	3.0	3.5	3.0	3.13	3.12
组织与演说	3.5			3.5	3.50	
口头媒介的 分析与评价	2.0			2.0	2.00	
说话技巧应用	2.0		2.5	3.0	2.50	2.86

转变学区和学校文化

最后，要在成绩单上实现重大的改变，必须要转变学校或学区的文化：使其从教师各自形成自己独特的评分方法的文化转变到教师之间以及学科之间都进行统一评分和成绩报告的文化。正如我们从第二章到第六章所讨论的，即使使用本书描述的系统，教师在怎样或在什么时间对学生进行评价这些方面依然有很大的自由。但是，他们所要报告的主题分数和呈现有关学生在那些主题学习上表现出的信息的方式必须是统一的。然而，这种统一不是一时能完成的。所以我建议，将以下几个阶段或事件的进展顺序作为将现有的评价和分数报告系统转变成本书所推荐的系统的方法。

第一阶段

让一个小组的教师带头以某个主题为基础进行评价试验并做记

录。

改变一个学校或学区的文化的的第一步就是要召集一小批教师自愿进行以某个主题为基础的评价试验并跟踪记录。这一步需要花费大约一个学期的时间。在此期间不需要做出重大改变，不需要修改成绩单，不需要设计全区性和全校性的测量主题，不需要进行统一评价设计。或者说，带头小组的每位教师只需要设计他们要在特定的评价打分阶段进行评价的学术主题和生活技能主题。他们使用第三章提到的简化版或完整版评分量表的模板来为各个主题制定评分标准。他们围绕主题进行课堂评价设计，并在适当的时候进行评价。为了跟踪记录学生主题学习的表现，教师可能会用到第五章提到的证据收集法。这样做，他们就要制作一个跟表 5.9 一样的书面填写的成绩簿。此外，带头教师还可以选择使用含有成绩簿的廉价软件。为此，我一般推荐求精软件公司生产的品尼高加法系统的单机用户版。该成绩簿对那些想要试用本书描述的系统的个体用户来说是免费的。有关单机用户版的信息可登录网站 [http://www. excelsiorsoftware. com](http://www.excelsiorsoftware.com) 查询。

为了说明第一步是如何开展的，我们假定一个教九年级社会学科的教师志愿在学年的第二个学季做一名带头小组的成员。该教师会通过考查一个学季内要教的典型内容来设计一套测量主题，然后如第二章所描述的一样将那个内容编入各个测量主题。一般情况下该教师不应该设法跟踪、记录太多的学术主题（我推荐每个评价打分阶段大概 5~7 个学术主题），否则她会发现她无法跟踪、记录要教学的内容。也许她还会发现有些内容通过组织能很好地编入测量主题，而有些则不能。结果，她会决定使用第二章提到的教师选择类的内容。这类内容包含了所有那些她认为很重要的、但又不很适合编入任何一个主题的内容。除了学术主题外，她还需要设计几个她认为重要的生活技能主题。

随着个人主题的确立，该教师将会用第三章中表 3.11 中的普通模板来为每个主题制作一个评分量表。根据第三、四章的建议，她将要进行课堂评价设计与课堂评价打分。根据第五、六章的建议，

她还要计算主题的最后综合分数。

成立带头小组的目的就是要让志愿教师获得更多有关新的评价评分系统的经验。我的体会是建立带头小组是改变学校和学区文化的关键一步。由于这个新的评价评分系统的可行性，带头小组的教师将会赢得同事们的信任。我还发现，在带头小组的教师中，绝大多数都是这个新系统的大力提倡者。他们看到了传统系统固有的不足，同时也看到了新系统的潜力。

第二阶段

设计整个学区或学校使用的测量主题，确定跟踪学生主题学习的表现所使用的软件与将要使用的评分系统。

如果带头小组的经历是一个好的开始，那么下一步就要设计测量主题。比较理想的是依据学区级水平来设计测量主题，从第一阶段的末期开始进行，这应该需要大约一年至一年半的时间。在第二阶段期间，主题专家小组需要频繁会面以草拟各个年级各学科的测量主题设计方案的初稿。依据第二章所讨论的，我建议要为幼儿园到九年级甚至十年级的学生设计测量主题——不管哪个年级的学生达到了应该掌握的教学内容的终极水平，也不管学生高中毕业后的计划如何。这些测量主题应该被看成是一般的读写能力主题（general literacy topics），因为它们蕴涵着人们成功地参与整个社会活动所需要的重要的信息和技能。在九年级（或十年级）之后，就应该为专业课程和高级课程设计不同的测量主题，也就是要设计超出基本测量主题的测量主题。

在任何情况下，一年内，专家在特定的学科中设定的测量主题不应该超过20个。测量主题设定的理想数目是15个左右。在这些主题内，要设定相对少的共变要素——每个主题不超过5个，最理想的，大约是3个。除了学术主题外，还需要设计一组生活技能主题。我建议不超过5个。尽管高年级的每个主题中的具体的共变要素与低年级的是不一样的，但这些主题可能从幼儿园到十二年级都是一样的。对所有的学术主题的规定都应该以完整的9级评分量表

(最早出现在第三章,表 3.11) 中的标准为依据(还可参见表 3.14 的例子)。而对生活技能主题的规定可以以简化的 5 级评分量表(参见表 4.10 的例子)中的标准为依据。一旦完成了学术主题与生活技能主题设计方案的初稿,就应该让那些没有参与设计过程的教师和项目相关人员进行审查。从这些审查中得到的信息应该在主题设计方案的二稿中加以采用。

在第二阶段进行的同时,来自于该学区的一些知识渊博的学者应该选定用来跟踪学生主题学习表现的计算机软件。理想的是,在第一阶段,带头教师就有机会试用该软件的廉价版本。选定的软件应该能够完成第六章所描述的如下任务:

- 基于幂函数或平均计算最后的真实分数。
- 在数理上决定哪种方法计算出的真实分数最符合实际数据。
- 对具有不同权重的主题进行处理。
- 用多种方法将数个主题合并成更高一级的范畴。

第二阶段的最后一步就是决定所要使用的评分系统。本章介绍了两种系统:一种包含字母等级总成绩、学术主题分数和生活技能主题分数;另一种不包含字母等级总成绩,但按学科和主题分类记录测试主题的分数。也许还有很多其他方法可以对测量主题的分数进行综合记录。具体使用哪一种记分报分系统应该在第二阶段做出选择,所选择的系统应该允许相关项目人员进行适当的检查和信息的输入。

第三阶段

分步实施该系统。

一旦第一阶段和第二阶段结束了,就应该分步实施已经设计好的基于主题的记分报分系统。如前所述,第一阶段和第二阶段需要花一年半到两年的时间才能完成。因此,新的记分报分系统大约要在第三年才能真正开始实施。这让教师们有足够的时间去识别和修复漏洞,也让教师和项目相关人员有时间投入该系统的设计。

系统一旦设计好，就应该在全区实施。要做出这样重大的改变向来不容易。因此，尽量一次性把它做好。但是，该系统也可以分步实施。在基础教育体制中，对新的记分报分系统抵触最小的应该是小学。在小学实施一年之后，该系统便可以在初中实施；再一年之后，该系统便可以在高中实施。在阶段性实施该系统的过程中，需要做一些调整和补充。例如，在第三阶段，学区可能开发出季末评价或统一评价供教师结合形成性课堂评价使用。总之，在第一阶段和第二阶段结束之后，整个系统可以采用分步的办法在三年内分步进行实施。

从改变文化到改变教育

如果该系统得到正确实施，这三个阶段将极大地改变基础教育的文化，使其从没有什么形成性数据去确定学生优、缺点的教育转变成形成性数据丰富的教育。因此，教师可以在学习的早期发现那些在特别的测量主题方面学习有困难的学生，并在需要时给予帮助。一个数据丰富的系统将会大大地降低学生在升级后主要课程基础十分薄弱的可能性。这样的成就是引人注目的。

然而，一个严格实行以主题为基础的记分报分系统具有引起更为巨大变化的潜力。也就是说，它能将教育转化成一个系统。在该系统中，学习是一个常量，教学时间是一个变量；而在现有的系统中，学习是一个变量，教学时间是一个常量。这种革命性的概念很多人都提倡过，讨论过，该主题也变更过很多次（例如 Bloom, 1976; Boyer, 1983, 1995; Goodlad, 1984; Guesky, 1980, 1985, 1987; Spady, 1988, 1992, 1994, 1995）。根据 J. 罗纳德和詹姆斯·拉利（J. Ronald Gentile, James Lalley, 2003），这个丰富多样的概念表现形式来源于约翰·卡罗尔（John Carroll）的著作（1963, 1983），其中提到卡罗尔模式（Carroll model）。卡罗尔模式可以用下列公式表示：

$$\text{学习量} = \frac{\text{实际花费的时间}}{\text{学习所需要的时间}}$$

根据这个公式，学生学习某一测量主题的学习量是该生实际学习该主题所花费时间和学习该主题所需时间的函数。如果一个学生在一个需要花 10 小时学习的主题上只花了 5 个小时，那么就还没有把这个主题学好。这个情况对于第二章有关学习国立和州立标准规定的内容估计所需时间的讨论具有重要的意义。估计现在需要额外 71% 还多的时间来教国立和州立标准所规定的内容，所以要美国学生能学好标准规定的内容几乎是不可能的。因此，现在需要大量地删减规定的教学内容。

卡罗尔公式揭示了另一个有趣的问题——不同的学生需要不同的时间去学习同一内容。具体地说，大量的研究表明，学生入学时，他们所拥有的学科背景知识量是大不相同的（供讨论，参见 Marzano, 2004b）。那些在某一学科领域拥有大量学科背景知识的学生学习起来要快一些，而那些没有很多背景知识的学生则需要花更多的时间学习。

卡罗尔模式从根本上表明一个最好的教育体制应该能让学生根据需要花费时间的多少去学习测量主题的内容。但是，在现行的教育体制中至少有两个根深蒂固的习惯做法阻碍着卡罗尔模式的实现——那就是年级制和学分制。

年级的概念是戴维·泰亚克和威廉·托宾（David Tyack, William Tobin, 1994）所提及的美国教育的“基本原理（grammar）”中不可缺少的一部分——学校应该按年级组织教学的这种观念在美国文化中根深蒂固，因此很少受到挑战。可以说，根据学生的年龄划分年级进行教学的方法在 20 世纪初就深入美国教育体制中，其部分原因是这种方法在公共教育迅速发展期间，以一种直接、直观而又吸引人的方式将学生进行分类。正如戴安娜·拉维茨（Diane Ravitch, 1983）所说：

近 10 年来随着时间的推移，美国青年上学的时间越来越

越长：一年中上学的时间长了，一生中上学的时间也长了。从1870年到1940年，人口增长了三倍，学校的入学率飙升。中学生人数差不多增长了90倍，从1870年的8万人到1940年的700万人……接受教育的人数的增加归因于经济因素和社会因素。1900年，大多数男性工作者不是农民就是体力劳动者。随着本世纪的发展，越来越少的人从事这些职业，越来越多的人从事白领工作或需要技能的蓝领工作（pp. 9-10）。

一旦学校建立起年级制，整个社会便把这种做法看成是有效教育的基本要素。然而，我们可以说，年级制阻碍了学生根据自己的学习进度学习教学内容。所有学生（少数几个除外），无论对某个年级的教学内容理解和掌握与否，都必须以相同的步子往上升级。学校的教学时间是相同的，但学习效果却是大不相同的。

阻碍卡罗尔模式实现的另一个习惯做法就是中学把学分当成是学生在某一课程学习中进步的指标。为了获得某一课程的学分，学生必须花一定的时间去学习该课程。实际上，学生只要求表现出最小的能力便能获得学分（D级）。在此我们又可以说，时间是不变的，但学习效果却大不相同。学分制的起源可以追溯到大约100年前，也就是1906年。亨利·史密斯（Henry Smith），卡内基教学促进基金会的会长，把一个学分定义为一学年每周上5节课（Tyack, Tobin, 1994）。在《高中：关于美国中等教育的报告》这本书中，欧内斯特·博耶（Ernest Boyer, 1983）解释说，学分制实际上已经成了一个障碍，阻碍着学生根据自己的学习进度进行课程学习（p. 237）。

也许这还不是一件显而易见的事，使用以课堂为基础的形成性评价，围绕测量主题组织教学的地区有可能打破传统的年级制和中学实行学分制的常规。事实上，一个基于主题的教学体系使年级和学分的不合理性凸显出来。举个例子说，一个七年级学生在一所使用以测量主题为主的教学方法的地区性学校学习。在数学方面，该生有能力学习八年级的主题，而强迫该生学习已经掌握的七年级的

主题很可能会使她感到厌倦。但是，七年级的语言艺术主题可能超出了她现有的能力水平，在这一学科上，她最好去学习六年级的语言艺术主题。学习七年级的语言艺术主题很可能会让她倍感沮丧，因为以她的水平来说，这些内容太复杂了。一个如书中所描述的基于主题的系统，由于对进度表进行调整，对资源进行重组，会减少这些问题。更准确地说，一个使用卡罗尔模式的学校必须要灵活。

在一个十分灵活的系统中，学生可以学习任何学科中任何水平的内容。该系统是完全灵活的，因为（理论上来讲）学生可以学习一门学科中二年级的内容，也可以学习另一门学科中十年的内容。事实上，在这样的系统里，“年级（grade level）”这个词语将很可能简化成更受欢迎的普通词语“级别（level）”。虽然这样的系统是理想化的，但如果要在大多数学区实施该系统，那也是非常困难的，因为这些学区刚进入设计。而且通常高中校园跟初中校园相距很远，初中校园又跟小学校园相距很远。事实上，大多数学区的后勤操作并不考虑一个完全灵活的基础教育方法。但是，我们还可以有别的选择。

在《一个称之为学校的地方》这本书中，约翰·古德拉德（John Goodlad, 1984）提议学校应该按照3个层级组织教学，每个层级由4个年级构成。我在别的著作中（Marzano, Kendall, 1996）建议分为4个层级，即：K至二年级，三至五年级，六至八年级，九至十二年级。每一层级都有其灵活性。也就是说，一个五年级的学生可以很轻易地向上学习六年级的数学主题和向下学习三年级的语言艺术主题。同一层级的教师负责教学不同年级的主题。例如，在三至五年级这一层级，有些教师负责教学三年级的数学主题，有些负责教学四年级的主题，还有一些负责教学五年级的主题。当然，同一个层级的数学教学应安排在同一时间里，以便学生按需要自由选择高一年级或低一年级的主题学习。对其他学科来说，情况也是这样。一旦学生掌握了一个年级所要求学习的主题，便可以继续下一个年级的主题的学习。这样，学生就不需要在一个年级花一整年的时间学习。那些花6个月时间便掌握了某个年级所要求学习的主

题的学生可以继续学习下一个年级的主题。同样地，那些需要花一年多的时间学习才能掌握某一个年级的主题的学生将继续留在该年级学习直到掌握为止。

有个逻辑问题是：第2层级（三至五年级）的学生，需要学习某学科中二年级的主题或者六年级的主题，该怎么办呢？通过把那些低于或高于某层级期望的学生分开组成同类小组，可以很容易地解决这个问题。下层小组的学生将学习那些他们没有学好的主题，目的是尽快将他们的能力提升到常规层级所要求的最低水平。将某一课程的学习超过某层级期望的学生组合在一起，这样他们就可以学习高级主题了。当这些学生进入下一个层级时，就已经超过了该层级对学生最低水平的期望了。实际上，直到第3层级，每个层级都应该采取适当的手段帮助那些进入某层级时水平低于某学科所要求的基础水平的学生，同时也要帮助那些水平超前的学生。

高中生需要用到初级证书和高级证书这两个概念。这两个概念在第二章提到过。例如，一张数学初级证书表明该生已经掌握了对该生走入社会起重要作用的普通数学内容。一个希望获得数学高级证书的学生需要学习高级数学主题的课程。根据学生的兴趣和动机，他们高中毕业时会获得一些学科的初级证书和一些学科的高级证书。要想毕业，所有的学生都必须获得某些核心课程的初级证书，如数学、阅读、写作和科学（或者任何地区设置的核心课程）。与核心课程的要求不同，学生可以自由地选择他们想要进一步发展的学科领域，通过学习获得高级证书。

虽然这里提到的系统不是没有操作和资源问题，但我坚信在现有的系统和该系统可用的资源内，这些问题都是可以解决的。换句话说，假如有学校已经做好了测量主题和形成性评价的必备工作，书中描述的系统可以马上实施。

教育改革在哪里进行？

很多时候，种种方案如这里所讨论的一样仅仅是设想而已。它

们给我们以启示，但同时也使我们对其实用性产生质疑：这些方案描绘出的美好设想在近期内是可望而不可即的，目前根本没有人运用这个系统的事实似乎也证明了这一点。尽管如此，目前出现了一个比较令人鼓舞的情况。一个很小但是正在不断壮大的地区团体已经开始实践这个设想。这其中就有许多地区参加了学校教育改革联盟（Reinventing Schools Coalition, RISC），并在运用学校教育创新模式。正如里克·施赖伯和温迪·巴蒂诺（Rick Schreiber, Wendy Battino, 2002）在《学校教育改革指南》（A Guide to Reinventing Schools）中描述的，位于阿拉斯加州中南部的巧克奇学区（Chugach School District）是学校教育改革模式的起源地。巧克奇学区覆盖了大部分威廉·桑德王子海岸线与岛屿，其学生分布在方圆 22 000 平方米的校区内。由于巧克奇的教师努力创新，负责人理查德·德洛伦佐（Richard DeLorenzo）领导有方，巧克奇成为了第一个获得新美国中学奖的学区，也是第一批获得著名的马尔科姆·波多里奇优秀奖（Malcolm Baldrige Excellence Award）的三个教育机构之一。

1994 年，通过征求所有的项目相关人（包括家长、学生、教师、行政官员、商业团体的代表和整个社区）的意见后，巧克奇开始进行全面的改革。基于这些意见，巧克奇设置了 10 门标准课程，并制定了每门课程具体的理解标准和成绩标准。这 10 门课程包括（1）数学，（2）科学，（3）科技，（4）阅读，（5）写作，（6）社会研究，（7）服务学习，（8）职业发展，（9）文化意识，（10）个人/社会/健康发展。项目相关人理解并接受这些课程，因为他们参与了设计和设置。正如施赖伯和巴蒂诺（2002）写道：“教师、家长、学生和社区成员都了解学生教育的目标，因为他们帮助制定了这些标准（p. ix）。”在巧克奇的教育体系中，学生提升的是各种水平而不是年级。学生是否达到某一级水平的能力是通过教师设计的课堂评价和学区设计的统一评价来验证的。

正如人们所期待的，巧克奇获得了令人瞩目的波多里奇奖（Baldrige Award），这就已经证明其教育体系对学生的学业成绩产生了积极影响。为清楚起见，请参见表 7.5。它呈现了巧克奇的学生

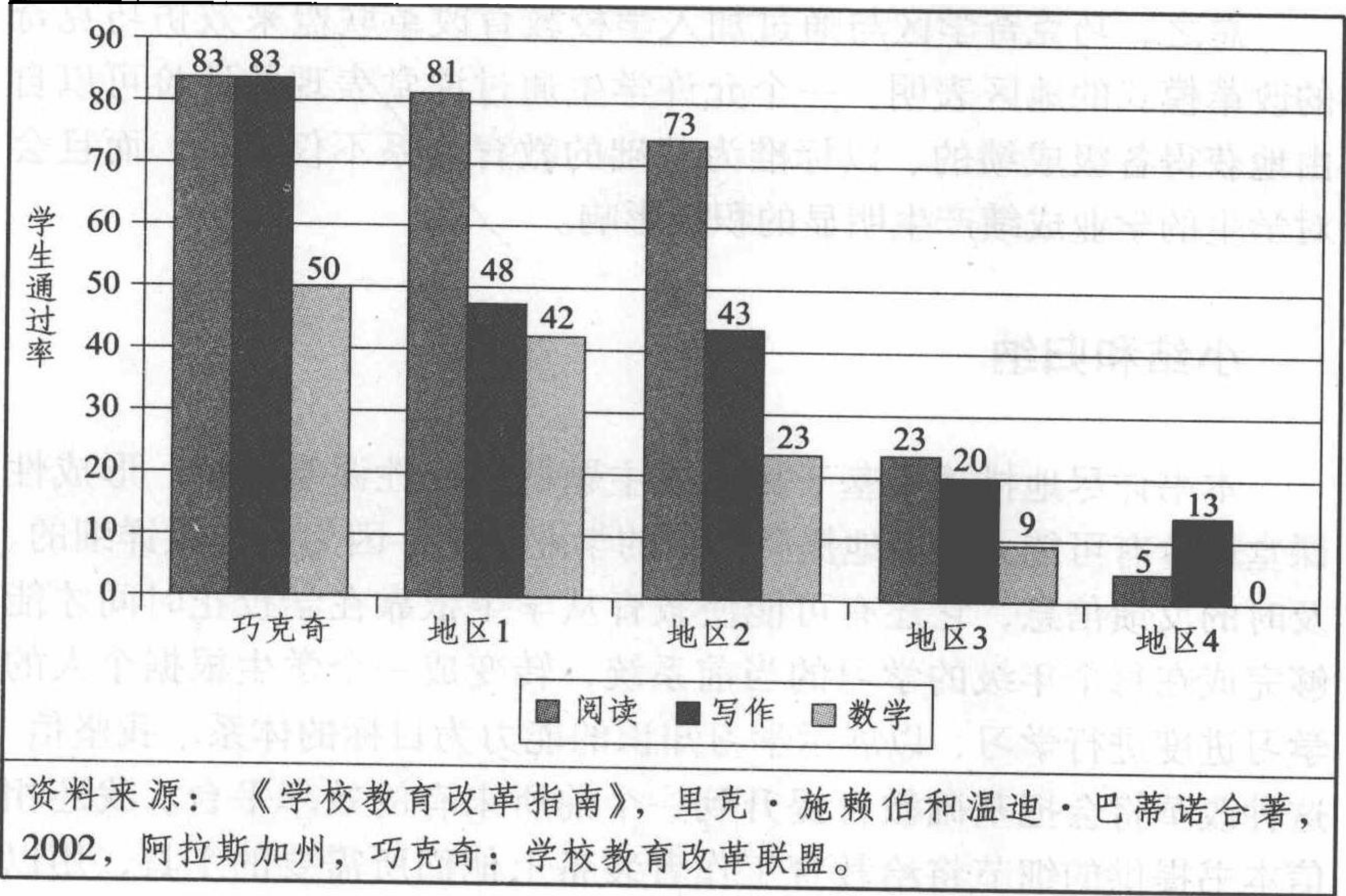
在加利福尼亚州成绩测验的全国百分制分数。加利福尼亚州成绩测验始于1994—1995年期间，该系统也是在这个时候启动的。如表7.5所示，巧克奇的学生5年来在所有4门课程中的百分制分数不断提高。施赖伯和巴蒂诺（2002）说道：“当前的水平和趋势表明巧克奇的学生比其竞争地区的学生提高得快些”（p. 320）。

表 7.5 巧克奇的学生获得的全国百分位分

学 年	阅读总分	语言总分	数学总分	拼写总分
1994—1995	28	26	36	22
1995—1996	43	44	54	32
1996—1997	56	50	58	35
1997—1998	63	60	66	46
1998—1999	71	72	78	65

资料来源：《学校教育改革指南》，里克·施赖伯和温迪·巴蒂诺合著，2002，阿拉斯加州，巧克奇：学校教育改革联盟。

表 7.6 阿拉斯加州中学质量和基准考试通过率



的确，巧克奇与阿拉斯加州其他地区相比是相当有优势的，参看表 7.6。该表呈现了巧克奇和其他 4 个地区的学生通过 2000 阿拉斯加州中学质量和基准考试的百分比。学生参加了阅读、写作和数学三门课程的考试。尽管其中有两个比较地区的学生在社会经济学结构方面占有极大的优势，但是巧克奇的学生在所有三门课程中的考试通过率却更高些。

也许巧克奇可以提供的最有说服力的成绩数据就是第三方对其他学区效仿巧克奇模式所取得的成绩的评价（参见 Coladarci, Smith, Whiteley, 2005）。具体地说，15 个学区进行了教育体系改革，即运用学校教育改革联盟（RISC）的方法改革以标准为基础的教育。一项研究调查了 15 个地区对学校教育改革联盟（RISC）改革方案的实施程度与其学生在阿拉斯加州中学资格考试和基准考试中取得的成绩之间的关系。调查发现学校教育改革联盟（RISC）改革方案的实施与阿拉斯加州州级测试成绩之间的密切关系有利于阅读、写作和数学成绩的提高（分别为 0.57, 0.33, 0.54）。

总之，巧克奇学区与通过加入学校教育改革联盟来效仿巧克奇的改革模式的地区表明：一个允许学生通过课堂表现的评价可以自由地获得各级成绩的、以标准为基础的教育体系不仅可行，而且会对学生的学业成绩产生明显的积极影响。

小结和归纳

本书详尽地描述了基于标准或主题的形成性课堂评价。形成性课堂评价有可能大幅度地提高学生的学业成绩，因为它提供详细的、及时的反馈信息，它还有可能使教育从学生依靠在学校花时间才能够完成在每个年级的学习的当前系统，转变成一个学生根据个人的学习进度进行学习、以展示学习知识的能力为目标的体系。我坚信，这种改革将会把基础教育提升到一个前所未有的效率平台。我还相信本书提供的细节将给教育工作者装备上他们所需要的工具，用以促进一个更有效的教育体系的发展。

术语注释

术语注释 1.1

判断“教学效果最好 (most effective)”与“教学效果最差 (least effective)”的教师可通过排列学生的成绩,然后按排列顺序分成五个级别。“教学效果最好”的教师是那些排在最高一级 (Quintile 1) 的教师,“教学效果最差”的教师被排在最低一级 (Quintile 5) 中 (进一步的技术讨论请参见 Haycock, 1998; Sanders, Horn, 1994; Wright, Horn, Sanders, 1997)。

术语注释 1.2

术语“效应规模 (effect size)”指如下一些指数 (indices),即能用来表示自变量 (independent variable)、因变量 (dependent variable) 或预测变量 (predictor variable) 和被预测变量 (predicted variable) 之间的关系。这些指数包括 r (双变量关联系数, the bivariate correlation coefficient), R (多变量关联系数, the multiple correlation coefficient) 以及 PV (方差百分比, the percentage of variance)。如本书所使用的,“效应规模”的总称指格拉斯 (Glass, 1976) 提出的最常见的效应规模: 标准化均数差 (standardized mean difference)。实验平均数和控制平均数之差,除以总体标准差的估计数——称为

“标准化均数差”。

$$\text{标准化均数差效应规模} = \frac{\text{实验小组平均数减去控制组平均数}}{\text{总体标准差的估计数}}$$

举例说明标准化均数差（这里指效应规模）怎样计算和阐释。假设一所具有特色的学校测验成绩的平均分是 90，而另一所没有特色的学校平均分是 80，再假设总体标准差是 10。那么，效应规模是

$$\text{效应规模} = \frac{90 - 80}{10} = 1.0$$

该效应规模可解释为：特色小组（实验组）的平均分比无特色小组（控制组）的平均分大一个标准差（standard deviation）。假设所谈的特色不是简单的可能性函数（function of chance），而且除了该特色，两个小组的其他所有特征都是相同的，可以推测，实验组所拥有的特色使学生的平均分提高了一个标准差。换句话说，效应规模可用“z 值（z score）”的形式表示，还能用百分数值增加（percentile gain）的术语解释。

百分数值增加指的是，与控制组的第 50 百分位数值的学生（percentile）相比，在实验组的第 50 百分位数值的学生估计增长（或减少）的百分位数值点。假设效应规模是 1.0，可以推测，实验组中获得平均分的学生在控制组中大约位于第 84 百分位。这是因为分布理论（distribution theory）告诉我们，z 值 1.0 在标准正态分布中是第 84.134 百分位。因此，实验组中居第 50 百分位的学生的平均分比控制组大一个标准差（standard deviation）。

术语注释 1.3

为了建构表 1.1，布莱克和威廉（Black, William, 1998）报告的效应规模 0.70 被译成用公式 $r = \text{效应规模} / (\text{效应规模}^2 + 4)^{0.5}$ 来表达双变量关联系数。结果是 $r = 0.33$ 。这使人们能预测因变量的 z 值，若自变量的 z 值提高很快。这个预测的等式是

$$z'_y = r_{xy} z_{xy}$$

这个等式用文字表述就是，变量 y 中所预测的 z 值或标准分（用“'”表示）等于 x 和 y 之间的相关系数 x 上的 z 值相乘。如马格努森（Magnusson, 1996）所说：

当我们了解一个个体（individual）在 x (z_x) 观察到的标准分和 x - 分布与 y - 分布的分数之间关系的相关系数后，我们通过用相关系数乘以 z_x ，能够获得 y 上个体标准分的最好的可能性预测。（p. 39）

要计算学生成绩上的进步，并与表 1.1 上部分显示的形成性评价里教师教学效果联系起来，假设形成性评价的效果是自变量，学生成绩是因变量。同时还假设，若有名学生在成绩上位于第 50 百分位，有位教师在形成性评价的效果中也是第 50 百分位。因此，自变量和因变量的 z 值都是 0。然而，假如教师在形成性效果中提高一个标准差，她的 z 值就会提高到 1.0，她在使用形成性评价时被分布的位置是第 84 百分位。用以上的公式，可以预测学生的 z 值将提高到 0.33 [即 $z'_y = r_{xy} z_{xy} = (0.33) \times (1.0) = 0.33$]。参考标准正态分布，当 z 值 0.33 转化为（在该背景中）百分位数值，就会增加 13 点。计算表 1.1 下部分所显示的学生成绩进步，可做同样的假设。但是，在这个例子中，假设教师把形成性评价中的效果提高 2.33 个标准差，她就会在第 99 百分位上。用以上的公式，可以预测学生的 z 值提高到 0.769，即 $z'_y = r_{xy} z_{xy} = [0.33] [2.33] = 0.769$ 。参考标准正态分布表，当 z 值 0.769 转化为百分位数值时，就会增加 28 点。

术语注释 1.4

罗伯特·班格特德劳恩斯、陈林·库利克和詹姆斯·库利克（1991）的元分析推测以 15 周为一周期的形成性评价效应规模，并假设学生成绩和所进行的评价数量之间是非线性关系。他们提供了

一个图表来表达他们的非线性推测。表 1.3 描述的数值被输入他们的图表中。

术语注释 2.1

哈迪尔 (Hattie, 1984, 1985; Hattie et al., 1996) 在一系列文章中阐明了关于建构一维性 (construct of unidimensionality) 的许多错误概念。他解释道:

测试理论 (test theory) 的基本假设是, 分数只有在有一套试题仅测量一个方面或一个难度时才有意义。假如测量工具包括测量不同方面的题项, 很难从一套题目中阐释总分, 在相关的变量中很难产生心理学意义 (psychological sense), 很难解释个体差异 (individual differences)。尽管这个假设对所有的测试模式都很重要, 但在调查这一假设的系统化努力上却极为匮乏, 并且, 直到最近, 仍无法提供有力的程序来评价一维性的观点。(Hattie et al., 1996, p. 1)

在 1984 年的研究中, 哈迪尔检测了 87 个一维性指数 (indices of unidimensionality), 得出的结论是“几乎所有的指数……都是无效的”(p. 75)。作为一维性指数研究的结果, 哈迪尔质疑把一维性作为测试发展的理想目标 (ideal), “最终, 人们会认为寻找一维性指数或一套套一维性题目是不现实的”(1985, p. 159)。哈迪尔说, 基于题目应答本身的指数和题项不可能确定一套测试项是否是一维性的, 因为每个被测者都会以不同的方法做题。“许多最新的研究显示, 同一套测试题目会被人们用不同的认知策略来应对”(1985, p. 159)。最后, 哈迪尔认可对测试的一维性必须进行裁定:

可能所要求的是判断的行为而不是指数。凯利 (Kelly, 1942) 论证说, 隐含在这些一维性概念中的是调查者的一个观念或观点, 例如, 当研究者确认题目测验的是相同内

容时要求有个裁定。因此，与另一份试题比较时，不仅可能要通过检验来认定测试是否一致，而且即使有指数，也要进行裁定（1985，p. 159）。

虽然哈迪尔没有强调第二章阐述的测量主题概念，但他的论述很有意义，因为题目的一维性或协变性（covariance，又称协方差）在某种程度上起到了解释在教学上怎样处理内容的作用。因此，同一个学区的课程专家必须参与到一维性的建构或协变因素建构（constructs with covarying elements）的鉴别与阐释中来。

术语注释 2.2

正如第二章讨论中提到的，协方差指的是某种关系。其中，某一特征或某一维度的能力与另一特征或另一维度的能力是相关的，当一种能力提高时另一种能力也会提高。在更加强调技术化和统计学知识的层面上，协方差被定义为比率，即，两组变量的离差之积的总和除以变量的组数：

$$\text{协方差} = \frac{\sum (y_i - \text{平均分 } y) (x_i - \text{平均分 } x)}{n}$$

y_i 指的是学生 i 在 y 变量中的分数（或因变量 dependent variable）， x_i 指的是学生 i 在 x 变量中的分数（或自变量 independent variable）。分子括号里的术语指的是离差（deviation scores），或者说是观察分数和观察分布的平均数之间的距离。当一套分数的协方差除以自变量和因变量的标准差的积时，得出双变量关联系数（Cohen, Cohen, West, Aiken, 2003, p. 29）。即，

$$r_{yx} = \frac{\text{协方差}_{yx}}{(sd_y)(sd_x)}$$

术语注释 3.1

真实分数 (true score, 又称 T 值) 的概念基于古典测试理论 (CTT), 与题项应答理论 (IRT) 相悖。汉布尔顿、斯瓦米纳丹和罗杰斯 (Hambleton, Swaminathan, Rogers, 1991) 是这样解释 T 值的: “能力指什么? 在古典测试理论框架中, 能力的概念用 T 值来表达, 被称为 ‘在兴趣测试中对被观察到的绩效的期望值’ ” (p. 2)。洛德 (Lord, 1959) 解释说, T 值 “经常被定义为平均值, 假如被测者在测试过程中不改变的话, 他在所有可能的平行测试中都将获得这个平均分” (p. 473)。格利克森 (Gulliksen, 1950) 定义为, 对学生来说, T 值 “是他在大量测试方法中的平均分的终值, 在类似测试数量……无限提高的基础上” (p. 28)。最后, 马格努森 (Magnusson, 1966) 是这样描述 T 值的: “完全肯定, 来源于潜在的连续体 (latent continuum) 预测的 T 值对每个个体的所有平行测试都是相同的” (p. 63)。

这个定义的含义是, T 值与某一具体测试联系起来。每个定义都在某个具体测试或平行测试的情景中描述 T 值。这局限了对 T 值的解释。如汉布尔顿 (Hambleton) 及其同事 (1991) 所解释的, “被测者的能力仅在特定测试的意义上被确定。测试 ‘难’ 时, 被测者表现出有能力; 测试 ‘易’ 时, 被测者表现的能力似乎更强” (p. 2)。IRT 模式解释潜在特性 (latent traits) 的术语正好与 T 值相反。潜在特性连续体并不由具体测试确定, 而是以分值的假设分布来确定, 这些分值指向由测试所测量的基本维度。

术语注释 3.2

虽然有许多 IRT 模式, 但这些模式都有相同的特点。如恩布莱特森和赖泽 (Embretson, Reise, 2000) 所解释的, “所有基于 IRT 的评分策略都试图评估每个被测者在潜在特性连续体上 (latent-trait

continuum) 的位置, 他们通过运用被测者的项目应答模式并结合估计性的项目参数来进行评估”(p. 158)。最早期和最基本的IRT 模式之一是拉希模式, 以创始人乔治·拉希 (Georg Rasch, 1960) 的名字命名。拉希提出了许多模式, 但与他的名字联系最密切的一个模式是应用个人的特征分数 (trait scores) (T) 和题目的难度 (D) 之间的简单差 (simple difference):

$$p(X_{is} = 1 | T_s, D_i) = \frac{\exp(T_s - D_i)}{1 + \exp(T_s - D_i)}$$

- T_s 指个人 s 的特征分数。
- D_i 指题目 i 的难度。
- $p(X_{is} = 1 | T_s, D_i)$ 指个人 s 在题目 i 的应答 (X) 是正确的 (1) 可能性, 若个人 s 的具体特征分数 (T) 和题目 i 的难度 D 确定。
- $p(X_{is} = 1 | T_s, D_i)$ 指使用个人的特征分数和题目难度之差的自然反对数 (antilog)。该数也可写成 $\varepsilon(T_s - D_i)$ (参见 Embretson, Reise, 2000, p. 50)。

用文字表述的话, 这个公式意指, 带有特征分数的人能正确回答给定难度的具体题目的概率等于 ε (自然对数的底, 大约为 2.7183) 乘以个人的特征分数的幂 (power) 减去题目难度, 除以 1 加上 ε , 乘以个人的特性分数的幂减去题目难度。对测试中的每道题来说, 特征分数全距 (range) 的题目应答曲线 (item response curves) 用类似的模式公式来计算。题目应答曲线显示正确应答题目的可能性, 或确认每个潜在特征分数的题目。通常, 特征分数和题目难度用一个量表 (scale) 表示, 从 -3.0 到 +3.0。尽管这些米制 (metric) 与 z 值相似, 但它们不是相同的。总体说来, 要计算个人在测试上的特征分数, 先用这些题目应答曲线计算特征分数全距的可能性, 得出个人的应答模式。带有最高计算可能性的特征分数被看成是通过测试证明的个人特征分数。这些特征分数通常被转换成另一种更易阐释的米制 (比如, 从 0 到 1000 不等)。

术语注释 3.3

正态分布 (normal distribution) 的概念在教育实践上有着深远的影响。正态分布的数学方程 (mathematical equation) 早在 1733 年就由数学家亚伯拉罕·棣莫弗 (Abraham de Moivre, 1667—1754) 列成公式了; 然而, 它的重要性是由数学家皮埃尔·西蒙·拉普拉斯 (Pierre Simon de Laplace, 1749—1827) 和卡尔·弗里德里希·高斯 (Carl Friedrich Gauss, 1777—1855) 提出的。高斯关于正态分布的文章震撼世界, 因而今天他被普遍认为是正态分布的先驱。事实上, 正态分布有时被称为“高斯分布”。

正态分布所指的实际上是两个参数 (parameters) 作用的分布群, 分布的平均数和方差。带有相对较大方差的分布在形状上是扁平的, 带有相对较小方差的分布呈高尖状。如果不考虑这些参数, 正态分布群中的所有分布都有相同的特点, 最突出的特点是中间有最高点、平均数周围形成对称起伏的曲线, 如表 3.4 显示。正态分布最常指的是用 z 值形式表达的原始数值 (raw scores) 的分布群成员。这个分布被称为单元正态分布 (unit normal distribution)。大多数统计课本运用这种单元正态分布报告了会在平均数和 z 值形式的分数全距之间下降的数值比例表。

正态分布的一个最突出的方面是它能很好地描述许多特点。如, 乔弗瑞·哈里森、约瑟夫·威纳、詹姆斯·塔纳和 N. A. 巴尼科特 (Geoffrey Harrison, Joseph Weiner, James Tanner, N. A. Barnicot, 1964) 报告说, 1939 年招募英国青年到军队服役时, 青年们的身高服从正态分布。同样, 西里尔·伯特 (Cyril Burt, 1957) 报告说, 从伦敦学校随意挑选的 2 835 个学生的 IQ 值也满足正态分布。简言之, 人们通常认为, 根据任何特点随意挑选的数值大约都满足正态分布。从 20 世纪初期到目前的测量理论都使用单元正态分布作为参考的关键点。

术语注释 3.4

戴维·弗里斯比 (David Frisbie, 2005) 在《测量 101: 再论基本要素》一文中论证了关于测量理论的本质和应用的一系列错误概念。最重要的是, 他阐明了测量理论基础中的两个概念: 效度 (validity) 和信度 (reliability)。关于效度, 他说, 术语文献清楚地阐述“效度本身不是工具, 而是对数值的阐释和运用” (p. 22), 因而, 谈论一次测试本身的效度是不准确的。这个事实站不住脚, 他列举了大量错误应用效度概念的事例:

以下来自各方的事例说明了对效度的误解:

1. “……‘甄别阅读评价 (screening reading assessment)’ 这个术语意指有效的评价……, 基于科学的基础阅读研究;” …… (2002 年的 NCLB 法案, 公众法律, 107 - 110)
2. “……你能确保测试对所有的学生都是有效的和公正的。” (来自面对全国的评价项目的考官手册, 2005)
3. “测试效度的证明……应该对公众公开。” (来自一个著名的测试机构的一本重要出版物, 2002)
4. “在评价范围里, 它指的是测试的效度。” (来自一本评价入门指南课本, 2005)
5. “[测试名称] 50 多年来在使用中已证明它本身是一种……有效的测试……。” (来自一个著名测试类出版社的网站, 2005)
6. “这种努力代表了测试效度的奠基石。” (来自一本著名的成绩测试术语手册, 2003, p. 22)

弗里斯比 (Frisbie, 2005) 提供了关于信度概念的共同基本结论: “看起来需要做一个关于信度的相似陈述, 要同样严谨和重视。即, 信度不是工具本身……信度是一套分值的一个特性, 而不是对产生分值的评价” (p. 25)。他还提供了有关明显误解信度概念的

例子：

1. “这些评价将用于某些目的，为此，这些评价是……可信的……”（2002 年的 NCLB 法案，公众法律，107-110）
2. “这些是用于研究……评价的信度。”（标准和评价同等检验指南，NCLB，2004）
3. “订约者将进行心理测试分析来监控测试的内容、结构效度和信度。”（来自国家第三教育部门的 RFP，2003）
4. “卖方必须用文件证明所建议的测试是可信的、有效的、不带偏见的……通常，对学生个体在某个学科领域的测试分数要有大约 0.85 或更高的测试信度系数。”（来自一份国家第四教育部门的 RFP，2003）
5. “由于测试信度受到测试中题量的影响很大……”（来自一本著名的成绩测试术语手册，2003，pp. 25-26）

弗里斯比关于效度和信度的论点直接关联着第三章关于为课堂评价打分所建议的测量量表。这个量表强调教师对学生应答的阐释，而不是一种孤立的评价。

术语注释 5.1

在整个篇章中，幂函数（power function）或“幂定律（power law）”作为数学选择模式用于反映学习过程。幂函数的基本公式是 $y = at^b$ ， y 是一个具体评价的分值（score）， t 是进行评价的时间（time）， a 和 b 是常数（constants）。如第三、五和六章所述，许多心理学家认为，幂定律是普遍存在的（Newell, Rosenbloom, 1981），因此是一个模拟真实得分变化的有用函数模型。

幂函数虽然是该书用于描述学习的模式，但它不是真实分数发展的唯一可能的模式。事实上，还潜有一套函数。假设 y 代表具体的主题或特点的个人分数， t 代表获得成绩的时间。假设学习不断进

行，下面都是可用于模拟学习的函数：（1）线性函数（linear function）；（2）指数函数（exponential function）；（3）对数函数（logarithmic function）；（4）二次函数（quadratic function）；（5）“近似-幂函数（pseudo-power function）”——是通过把自变量和因变量转换成自然对数（natural log），计算线性函数，然后把预测的分数转变回原有的米制而生成的。可用表 TN5.1A 检查这些函数的可行性。第一行包括了一套原来所观察的分值。第二行包括预测分值，这些分值是通过应用幂函数计算所观察的数据而得出的。其余几行包括应用各自的函数而得出的预测分值。表 TN5.1A 表明，预测分值各不相同，从原始分值到更大和更小的范围。

表 TN 5.1 A 模拟学习的函数

原分数	1	1.50	1.50	2.00	2.50
幂函数	0.984	1.413	1.747	2.031	2.282
线性函数	1.00	1.350	1.700	2.050	2.400
指数函数	1.062	1.313	1.623	2.006	2.480
对数函数	0.903	1.480	1.817	2.057	2.243
二次函数	1.071	1.315	1.629	2.014	2.471
近似-幂函数	0.984	1.413	1.747	2.031	2.282

判断用于模拟真实分数发展的各种函数的相对有效性的方法之一是检验方差的百分比，如表 TN5.1B 所示。除了对数函数，其他函数在所观察的分值中，方差都在 90% 以上，其中二次函数的方差最大。

表 TN 5.1 B 各种函数的方差百分比

函 数	方差百分比
幂函数	92.5
线性函数	94.2
指数函数	94.2
对数函数	86.1
二次函数	95.6
近似-幂函数	92.5

简言之，虽然幂函数在该章节中被用于模拟学习，但还有大量的函数可供使用。读者也要注意第六章描述的品尼高加法（Pinnacle Plus）软件使用了近似 - 幂函数，因为它容易被镶嵌到已存的软件代码中。

术语注释 5.2

增强证明的方法是松散地建立以一个英国牧师命名的贝氏推理模式（Bayesian model of inference），他早前研究过概率（probability）和决策（decision）理论。最简明的贝氏理论表明，要证明 A 和 B 两个事件，必须具备以下的关系：在给定事件 B 的条件下，事件 A 发生的概率等于事件 A 和 B 相交的概率（即，两个事件同时发生的概率）除以事件 B 的概率。贝氏理论转为计算式是：

$$p(A|B) = \frac{p(B|A) p(A)}{p(B|A) p(A) + p(B|\bar{A}) p(\bar{A})}$$

$\bar{A}$ 表示 A 的补集，或事件 A 没有发生的概率。

Hays (1973) 列举了以下情景：半夜，一个人起床要吃一片安眠药。他走到冲凉房，在黑暗中打开小药箱，抓起三个药瓶中的其中一瓶，吃了瓶中的一片药丸。回到床上，他开始感觉不舒服。他突然想起小药箱里除了有两瓶安眠药，还有另一瓶装着重药。他手头正好有一本医学手册，他查看了与服用那种重药相关的症状，发现其中有 80% 服用重药的人会出现与他现在一样的征兆，而 5% 服用安眠药的人也会有相同的症状。假设 B 代表他目前的症状， A 代表服用了重药，那么

$$\begin{aligned} p(B|A) &= 0.80 \\ p(B|\bar{A}) &= 0.05 \end{aligned}$$

假设每瓶药在黑暗中被拿到的概率是相同的，那么

$$\begin{aligned} p(A) &= 0.33 \\ p(\bar{A}) &= 0.67 \end{aligned}$$

把这些数据套进上面的公式，得出：

$$p(A|B) = \frac{(0.80)(0.33)}{(0.80)(0.33) + (0.05)(0.67)} = 0.89$$

因此，根据贝氏理论，此人服用了毒药的概率是0.89。

虽然增强证明的方法并不完全运用贝氏理论本身，它运用了基本的启发式（heuristic），但学生在原先的评价上对所给主题的表现程度应该能使教师判断学生在当前评价的表现程度。

术语注释 6.1

威列特（Willett, 1988）曾在他的关于测量学习变化的讨论中论述过恰当的评价次数的问题。很明显，评价次数越多越好，数据与曲线也越相符。但威列特仅用4个数据点就例证了测量知识变化的方法（用他的术语是4个数据“波”）。

看似合理的是，在评分期间，教师可能会与测试主题的4次评价的最小值相符，特别是用单独的测试、测验等方法评价多个主题时尤其如此。如第六章所述，每个主题我都推荐5次评价。所涉及的问题是，若只有5个数据点可用时，如何准确评估学生真实分数的提高？要回答这个问题，就要假设选择幂函数作为评判真实分数提高的模式。

在与学习相关的幂定律的心理学文献中，多数讨论都认为评价之间的间隔（interval）是相同的，或是基本相同的。要想在课堂上满足该条件的假设，将会有难度。教师难以在固定的时间（假设每隔一天）给予评价。同样，周末的间歇（hiatus）也增加了调节因素（mitigating factor）。

在论述评价之间的非相同间隔问题上至少有两种途径，表述在表TN 6.1中。表中第一栏代表学生在连续的20天中每天练习一种技能。第二栏的真实分数来自应用如下的幂函数： $y = 1.00 x^{0.45}$ 。换言之，第二栏假设的分数是20个相等练习间隔期的真实分数，这

些都是假设学习的进程符合幂函数，并且每人以 1.00 分开始。

表 TN 6.1 的第三栏代表教师跟踪学生成绩的一种方法，称为“基于顺序的计算 (order-based accounting)”。教师在 20 个间隔中分布了 5 次评价，仅连续计算评价，不考虑时间间隔之间的偏差。这对任课教师来说是轻而易举的。可以想象教师在课程计划允许的情况下进行评价，然后连续计算这些评价，不考虑评价之间的时间间隔。与之相关的问题是，使用这种基于顺序的绩效进行评价的话，20 天期限后对学生真实分数评估的准确性有多大？要检验这个问题，

表 TN 6.1 基于顺序和时间的绩效

次数	真实分数 $y = 1.00x^{0.45}$	基于顺序的 评价#	基于顺序的 预测	基于时间的 评价#	基于时间的 预测
1	1.000	1	1.104	1	1.000
2	1.366				
3	1.639				
4	1.866				
5	2.063				
6	2.239	2	1.977	6	2.239
7	2.400				
8	2.549				
9	2.688				
10	2.818				
11	2.940				
12	3.059	3	2.781	12	3.059
13	3.172				
14	3.279				
15	3.382				
16	3.482	4	3.542	16	3.482
17	3.579				
18	3.672				
19	3.762				
20	3.850	5	4.273	20	3.850

可将这5次评价（栏2）的真实分数换成顺序分布次数（栏3），预测的分数用幂函数计算。使用这种方法所预测的最后得分（如，第20个序列的预测分数）是4.273，很明显，对3.850这个最后的真实得分估计过高了。事实上，这个估计超过了该量表的上限（upper limit）（该量表分数不高于4.0）。

表TN 6.1的第5栏描述一个可选系统记录，称为“基于时间的计算”系统。教师安排了与学生的练习和复习天数相同的辨别数字。因此，第二次评价（按次序）的评价数字是6，因为是在教学/评价周期的第6天；第三次评价（按顺序）的评价数字是12，因为是在教学/评价周期的第12天，照此类推。在这个系统中，评价数字反映了教学/评价周期在时间上的实际数。当这5次评价的真实分数转换成基于时间的评价数字时，所预测的最后分数正好与真实分数相符。

这个模拟虽然是人为的，但却很重要。具体来说，基于时间的系统比基于顺序的计算为学生的真实分数提供了更准确的评估。事实上，这意味着即使当只使用四分之一的数据点（如20次中的5次）时，基于时间的系统也能为最后的真实分数提供非常准确的评估。

术语注释 6.2

最近这10年见证了人们对课堂评价理论基础的极大兴趣。事实上，2003年，一个整体性的论题“教育测量：问题和实践”（第22卷，第4）被加入到对课堂评价理论的现行地位的讨论中。实际上，每个研究这个问题的测试专家都明显或隐含地提到，古典的测试理论（CTT）和项目应答理论（IRT）都是为大规模的评价而设计的，难以很好地转换到课堂评价中。该卷的每一篇文章都建议用新的理论或新的方法去评估单独的课堂评价的信度（reliability）（或其他心理测试特点）。但对学生学业总体得分的心理测试特点没有讨论。第三、四、五、六章的讨论表明，使用形成性课堂评价使得对心理测

试特点的思考（特别是对信度）超出了单一的评价。弗里斯比（Frisbie, 2005）强调了考虑多种信度的需要。具体来说，就是假设真实分数随评价次数而改变时（形成性评价的必然例证），可以从两个方面获取信度。

第一个方面是来自一个特别评价的分数信度。通过计算某种基于一个或多个研究的信度系数或普遍性系数来具体强调，这在当时被认为代表了那个特殊评价的信度，虽然弗里斯比（2005）评论过这种阐释的非准确性（参见术语注释 3.4）。第三章展示了这种测量表，这种研究一直在进行中（参见 Marzano, 2002a）。

第二个方面是对每个学科评分的信度。更具体地说，如，用于描述测试次数之间得分的变化的数学模式信度。威列特（Willett, 1988）曾用丰富的史料证实了这个文献，但时至今日，它已不再应用到课堂评价中。如威列特（1985）所描述的，基本的测量公式是

$$X_{ip} = F_p(t_i) + e_{ip} \quad (1)$$

公式中的下标 i 表示测量的次数， t_i 表示测量次数 i^{th} 产生的时间，下标 p 表示被测试者， F_p 代表被测试者 p 的真实名次，括号里测量次数 i^{th} 的时间指 F_p 随时间而变化。因而， $F_p(t_i)$ 代表了描述个人 p 在各次测量中的真实名次； e_{ip} 表示与被测试者在时间 i 上测量误差。

将形成性测量公式与古典测试理论的终结性公式进行对比：

$$X_p = F_p + e_p \quad (2)$$

F_p 代表被测试者 p 的真实名次， (t_i) 的空缺表明古典测试公式假设了固定的名次。在等式（1）中，人们能推测构建一个可行的形成性测量公式的关键是证明最恰当的增长函数（growth function） $F_p(t_i)$ 。这两个增长函数是：差分函数（the difference score function）和多维线性函数（multiwave linear function）。

差分函数（The Difference Score Function）

作为课堂评价基础的增长函数（growth function）是起始名次与

最后名次之间的偏差。在这种情况下，任课教师会使用学生在学完一个教学单元所含的内容后的第一次成绩测试。假设要使第一次和最后一次测试平衡的话，每个学生的首次评价得分都会从期末测试分数中减除。对预测的敏感性研究表明，任课教师甚至会在头尾两次测试中使用同一份试题（Wilson, Putnam, 1982）。这个增长函数的数学公式是

$$D_p = X_{fp} - X_{1p} \quad (3)$$

公式中， X_{fp} 指被测者 p 的最后测试（如，教学单元的最后评价）， X_{1p} 指被测者 p 的第一次测试。观察差分（observed difference score）与产生于 t_f 和 t_1 之间的真实分数的变化之间存在一种简单的关系：

$$D_p = \Delta_p + e_p^* \quad (4)$$

$\Delta_p = F_p(t_f) - F_p(t_1)$ ， $e_p^* = e_{fp} - e_{1p}$ 。作为误差术语的独立假设的结果，人们归纳出 e^* 是与零平均数（zero mean）及方差（variance） $2\sigma_e$ 的正态分布。除此之外，还能看出，个人 p 的观察差分（difference score）是量（quantity） Δ_p 的公正评估（unbiased estimate）。罗格萨、布兰特和齐默斯基（Rogosa, Brandt, Zimowsky, 1982）强调的事实是，忽略测量误差（measurement error）大小的话，观察差分等于真实差分的公正评估。这是因为个人 p 的观察差分的预期值（expected value）是个人 p 的真实差分。威列特（1988）指出，尽管“这是个明显令人满意的统计结果”，对差分（difference score）的批评“多年来一直如此广泛，以至调查者使用它时总是小心翼翼”（p. 366），这主要由于它的信度低。尽管事实上差分常常表现出低信度，但这并不一定意味着它们在个体差分层面上欠精确。对差分信度的概念基础的检验给该问题提供了新的视野。

差分的总信度 $p(D)$ 被定义为总体中所有个体的方差 Δ_p 对方差 D_p 的比率。若 $p(D)$ 表达为观察偏差，威列特（1988, p. 368）提供了下面的公式，它需要没有简化的假设：

$$p(D) = \frac{\sigma_{x_1}^2 p(x_1 x_1) + \sigma_{x_f}^2 p(x_f x_f) - 2\sigma_{x_1} \sigma_{x_f} p(x_1 x_f)}{(\sigma_{x_1}^2) + (\sigma_{x_f}^2) - 2\sigma_{x_1} \sigma_{x_f} (p x_1 x_f)} \quad (5)$$

$p(x_1 x_1)$ 和 $p(x_f x_f)$ 是 x_1 和 x_f 的总信度 (population reliability), $p(x_1 x_f)$ 是起始分数 (x_1) 和最后名次分数 (x_f) 的总体相关关系 (population correlation)。包含起始和最后名次 (initial and final status) 的相关关系在分子和分母中被减除, 这能解释差分的低信度。如威列特 (1988) 所解释的那样:

由于分子前两项的方差乘以它们各自的信度, 因而它们比分母里相同的项小, 减除包含分子和分母相关的起伏关系的项能确保当 $p(x_1 x_f)$ 大并且为正数时, $p(D)$ 会降低。由于起始和最后名次的相关关系经常被误解为结构效度指数 (index of construct validity), 作者总是报告说差分无法同时作为信度和效度 (pp. 368 - 369)。

为了说明起始和最后名次之间密切的相关关系对差分信度的影响, 要考虑 $p(x_1 x_f)$ 是零的情况。分子和分母减除的项目为零, 等式 (5) 为

$$p(D) = \frac{\sigma_{x_1}^2 p(x_1 x_1) + \sigma_{x_f}^2 p(x_f x_f)}{\sigma_{x_1}^2 + \sigma_{x_f}^2} \quad (6)$$

在等式 (6) 中, 差分的信度减为 $p(x_1 x_1)$ 和 $p(x_f x_f)$ 的方差——加权平均数 (variance-weighted average)。假设 $\sigma^2(x_1) = 4$, $\sigma^2(x_f) = 9$, $p(x_1 x_1) = 0.90$, $p(x_f x_f) = 0.80$, 而 $p(D) = 0.83$ 。然而, 若假设起始和最后名次之间的相关关系为中等的 0.50, 那么 $P(D)$ 就会减少为 0.69 [运用等式 (5)]。

若起始和最后测试名次之间的相关关系通常很高, 实际就意味着所计算的差分信度总是很低。如费尔特和布伦曼 (Feldt, Brennan 1989) 指出的:

当人们查看标准成绩测试的报告时, 通常会发现五年级和六年级的名次分数信度大约是 0.88, 两年之间的相关

关系是 0.82。因而，增长量数（growth measures）的信度经常在 0.33 或更低中徘徊。然而，没有人会否认这些五年级学生在一年的时间中取得的巨大进步（p. 119）。

威列特（1988，p. 369）运用个体增长率（individual growth rate）的均匀（homogeneity）角度扩大了差分信度的视野。他提出了更能说明差分信度的公式：

$$p(D) = \frac{\sigma_{\Delta}^2}{\sigma_{\Delta}^2 + 2\sigma_e^2} \quad (7)$$

很明显，差分的信度随着个体中真实变化偏差的提高而提高。如威列特（1988）指出的：

真实增长中的个体偏差越大，差分的信度就越大。没有发现真实增长的个体偏差的话（如，当 σ_{Δ}^2 等于零时，所有个体的实得分数增长轨线是平行的），差分信度只能是零，不考虑测量时的精确度（p. 370）。

最后，对差分信度的理论基础的 analysis 使人们得出这样的结论，即这个结构无法胜任检验个体差分可行性（viability）的任务。在定义上，“信度是区分个体差别的尺度，仅能够就一个小组或总体进行定义”（Rogosa et al., 1982, p. 730）。若增长率的多相性（heterogeneity）是差分高信度的前提条件，均衡增长率的评价情景会产生低信度，并无法区别测量的准确性。如罗格萨及其同事们所指出的那样：“尽管增长中的个体偏差对高信度是必要的，欠缺这种偏差并不妨碍对个体变化的评价意义”（p. 731）。

总而言之，尽管与差分相关的信度特别低，起始到最后名次差分对形成性测量模式的增长函数是个可行的选择。具体而言，教师可以在一个单元的起始和终结时进行全面的评价，把差分作为学生成绩进步的可行评估。就是说，多点的（multipoint）或多维的（multiwave）方法比差分方法有更多的好处。

多维线性函数 (The Multiwave Linear Function)

虽然这种讨论支持把差分作为评估学生学习的手段，但在实际应用中还是有些问题。即使差分信度的评估不能很好地评价个体变化，它们仍应被评估。然而，这种评估需要的信息任课教师难以获取。如威列特（1988）指出的：

除了增长数据的两条曲线提供的样板方差和相关关系外，调查者还必须要有两条补充的信息：两次测试中各自的观察名次所评估的信度。两条补充信息必须从增长调查中客观地获取（例如，从手写的测试或以前经验的信度研究中），或观察名次的重复测试（duplicate measurement）必须在每个学科进行，其中两次分数符合需要信度的情景评估（situ estimation）（p. 371）。

很明显，这种补充信息的要求并不适合课堂评价情景。能够更容易获得对个体增长的评估更有意义的信度指数评估的方法是使用多点的或多维的数据。而且，多维数据能提供比差分更好的真实增长（true growth）评估：

对两次的个体排名拍张快照（snapshot）并不能使调查者看清个体增长原因的复杂性……实际上，要准确地测试个体增长，需要更多多维数据形式的增长信息。若多维数据在样本中的每门学科都能获得，调查者就能检验详细的经验式增长（empirical growth）——轨线图（trajectory plots），从而归纳出随时间变化的个体观察增长（observed growth）（Willet, 1988, pp. 384 - 385）。

在课堂情景下，教师可以通过设计多项评价来收集多维数据，用第三、四章描述的方式评分。

当只能得到两个数据分数时，把多维数据的优势概念化的方法是把差分作为代表个体增长的线性函数回归权重（regression weight）

的评估。事实上，当只能得到两个数据分数时，个体学生的直线函数（straight-line function）回归权重是

$$\frac{x_f - x_1}{t_f - t_1} \quad (8)$$

这直接与原差分成比例（参见 Willet, 1988, p. 385; Rogosa et al., 1982, p. 728）。

运用多维数据来模拟增长（modeling growth）所基于的假设是：随着时间的推进，学习遵循着一个可确定的数学函数发展，而且不断进行的观察测试是对增长函数不准确的评估。

当个体增长时，仿佛基本的真实增长随着时间不断顺畅并且难以观察，但调查者会周期性地看到增长中容易产生错误的（fallible）测量工具。这样的话，个体的观察增长记录被收集起来，它包括了根据时间顺序的系列分散测试，每次都是真实名次和测量误差的无名组合（unknown combination）。当然，对调查者来说，最大的兴趣是基本的、连续的真实增长轨线；观察增长记录中的多元项目（multiple entries）仅是易错的镜头（fallible lens），真实的个体增长是通过这镜头观察的。（Willet, 1988, pp. 386 - 387）

R. 米德和 D. J. 派克（R. Mead, D. J. Pike, 1975）列举了各种代数函数（algebraic functions），能用于模拟个体真实增长，就像罗格萨及其同事们做的那样（1982）。这些在术语注释 5.1 中讨论过。这个讨论要运用线性或直线函数，虽然这里说的逻辑应用于其他函数（Seigel, 1975）。特别是，当近似 - 幂函数被用做幂函数的可行评估时，直线函数逻辑应用到自然对数的转换中。

在这个情景中，模拟这些数据的直线函数可表示为

$$F_p(t) = F_p(t^*) + B_p(t - t^*) \quad (9)$$

$F_p(t^*)$ 代表真实的首次名次， $(t - t^*)$ 代表测试 t 和次数 t^*

之间的区别， B 是代表次数之间（常数）变化比率的回归系数（regression coefficient）（Willet, 1988, p. 390）。

若使用线性函数，基本的测量公式变成

$$X_{ip} = F_p(t^*) + B_p(t - t^*) + e_{ip} \quad (10)$$

为了评估个体增长，关键参数是 B_p ——测量次数变化中观察分数的回归系数，或者是在测量次数中单元变化而产生的观察分数变化比率（假设各单元之间的间隔是相同的）。这个评估的准确性类似于某个时间点上个体真实名次评估的准确性。

多项数据分能够评估准确性和信度。威列特（Willet, 1988, p. 402）解释说，增长比率的总体方差（ σ_B^2 ）和样本方差（ σ_B^2 ）的关系是：

$$\sigma_B^2 = \sigma_B^2 + \frac{\sigma_e^2}{SST} \quad (11)$$

SST 是观察次数的平方和——观察次数与它们平均数（ $\sum [ti - \bar{t}]$ ）的平方差。这是测量次数分布的测量。由测量误差引起的方差用 σ_e^2 表示。等式（11）表明增长率的样本方差会过高地估计增长率的真实方差。

若多维数据可行，每个人的测量误差方差都能计算出来，结果也可被应用在增长率的总体方差的评估中。特别是，若符合数据的增长模式是正确的话，观察分数和预测分数之差能就评估每次测量中的测量误差（ MSE ）。如威列特（1988）指出的：

测量误差的重要性能直接从这些残数（residuals）中看到。按照本文的假设，通过计算平方残数（squared residuals）能相当简单地估计出 σ_e^2 ……可以在 [跨越] 次数和人数的基础上，得出残数平方之和，然后除以总自由度（total degrees of freedom）（p. 403）。

通过代数处理，平方残数之和能够等于每人测量误差（ MSE ）的简单平均数。即：

$$\hat{\sigma}_e^2 = \frac{\sum_{p=1}^n MSE_p}{n} \quad (12)$$

评估增长率的信度

从语言学的角度来说，增长率的信度指的是真实增长率的方差与观察增长率的方差之间的比率。因为观察增长率的方差可以类似真实增长率方差的可行评估那样计算，增长率的总信度能通过以下的公式进行评估：

$$p(\hat{B}) = \frac{\sigma_B^2}{\sigma_B^2 + \frac{\sigma_e^2}{SST}} \quad (13)$$

因而，在含有大量真实增长交叉轨迹的总体上，得出可信的信度在实际中是可能的。另一方面，假如在真实增长率（ $\sigma_B^2 = 0$ ）中没有个体之差的话，所有的真实增长轨迹就是平行的，增长信度只能是零，不考虑所得的测量精确度（Willet, 1988, p. 404）。

简言之，在课堂上不断使用形成性评价不仅能评估学习终结时学生个体的真实分数，还能评估学生的成长轨迹，对每个学生的测量误差以及作为一个整体的班级的成长轨迹的信度。

最后，罗格萨及其同事们给个人变化的测试提出了许多“格言（mottos）”，对我们所讨论的课堂评价很有用：

- 两维数据比一维数据要好些，但好不了很多。两个数据分对个体变化提供的信息不足，因此，测量变化经常会比传统的前后数据（pre-post data）要求更多。
- 当只有两维数据时，差分是个体变化的一个自然而有用的评估。
- 不只有一个方法来判断对变化的测量。在变化的测量中，信度不是“从头到尾都是好的”（be all and end all）。统计数据很重要。

- 信度低并不意味着不准确。
- 两个易错的测量之差可能像测量本身那样可信。

(p. 744)

若运用完善的变化测量理论，这些格言可用于构成形成性课堂评价模式的基础。

附录 A：半分分值与整数分值的对照

如下表所示，充分理解完整版评分量表的一个方法就是将半分分值与它们相邻的整数分值进行对照。

半分分值	半分分值与相邻的整数分值的描述
0.5	分值 0.0 表示…… - 借助于帮助，学生对比较复杂的思想内容和程序（二类）或者比较简单的细节和程序（一类）表现出不理解。 分值 1.0 表示…… - 借助于帮助，学生对比较复杂的思想内容和程序（二类）或者比较简单的细节和程序（一类）表现出部分理解。 分值 0.5 表示…… - 借助于帮助，学生对比较简单的细节和程序（一类）表现出部分理解，但对比较复杂的思想内容和程序（二类）表现出不理解。
1.5	分值 1.0 表示…… - 借助于帮助，学生对比较复杂的思想内容和程序（二类）或者比较简单的细节和程序（一类）表现出部分理解。 分值 2.0 表示…… - 学生处理比较复杂的思想内容和程序（二类）时出错或出现遗漏，但处理比较简单的细节和程序（一类）时无错误或遗漏。 分值 1.5 表示…… - 学生处理比较复杂的内容和程序（二类）时出现严重错误或遗漏，但对比较简单的细节和程序（一类）表现出部分掌握，尽管仍然有错。

续表

2.5	分值 2.0 表示…… • 学生处理比较复杂的思想内容和程序（二类）时出错或出现遗漏，但处理比较简单的细节和程序（一类）时无错误或遗漏。 分值 3.0 表示…… • 学生处理比较复杂的思想内容和程序（二类）和比较简单的细节和程序（一类）时无错误或遗漏。 分值 2.5 表示…… • 学生处理比较简单的细节和程序（一类）时无错误或遗漏，而对比较复杂的思想内容和程序（二类）表现出部分掌握。
3.5	分值 3.0 表示…… • 学生处理比较复杂的思想内容和程序（二类）和比较简单的细节和程序（一类）时无错误或遗漏。 分值 4.0 表示…… • 学生水平超出课堂所教，表现出推理能力和运用知识的能力（三类）。 分值 3.5 表示…… • 学生水平超出课堂所教，在一定程度上能进行推理和对知识的运用。

附录 B：语言艺术测量主题样本的评分细则

	评分细则	
	学科：语言艺术	
	测量主题：要旨阅读	
	九至十年级（低分区）	
分值 4.0	除了有分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了有分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 明显的和暗含的复杂因果关系； · 带有明显的和暗含的复杂支持体系的论点； · 带有明显的和暗含的复杂的解决办法的问题； · 带有明显的和暗含的多重故事线索的复杂情节。 学生没有出现错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现错误或遗漏，如确认…… · 明显的复杂因果关系； · 带有明显的复杂支持体系的论点； · 带有明显的复杂解决办法的问题； · 带有明显的多重故事线索的复杂情节。 但是，学生在确认比较复杂的思想内容和程序方面出现重大错误或遗漏。	

续表

	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0		借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0		甚至借助于帮助，学生也表现出不理解或无技能。

	学科：语言艺术	
	测量主题：要旨阅读	
	八年级	
分值 4.0		除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… - 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； - 解释并对有关内容的推理进行辩解。
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，并在一定程度上进行推理和知识的应用。
分值 3.0		在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… - 明显的和暗含的复杂因果关系； - 明显的和暗含的基本论点； - 带有明显的和暗含的复杂的解决办法的问题； - 带有明显的和暗含的多重故事线索的复杂情节。 学生没有出现重大错误或遗漏。
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。

续表

分值 2.0	对比较简单的细节和程序的处理, 学生没有出现重大错误或遗漏, 如识别…… · 明显的复杂因果关系; · 明显的基本论点; · 带有明显的复杂解决办法的问题; · 带有明显的多重故事线索的复杂情节。 但是, 学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序, 但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助, 学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助, 学生能部分理解一些比较简单的细节和程序, 但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助, 学生也表现出不理解或无技能。	

	学科: 语言艺术	
	测量主题: 要旨阅读	
	七年级	
分值 4.0	除了分值 3.0 要求的表现外, 学生的所学超出课堂所教, 通过…… · 解释课文中哪些部分的内容是明显的, 哪些部分的内容需要推理; · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外, 学生能够超出课堂所教的内容, 在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时, 学生能够确定课文的要旨, 如…… · 明显的和暗含的复杂因果关系; · 明显的和暗含的基本论点; · 带有明显的和暗含的复杂解决办法的问题; · 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	

续表

	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0		对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 明显复杂的因果关系； · 明显的基本论点； · 带有明显复杂的解决办法的问题； · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0		借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0		甚至借助于帮助，学生也表现出不理解或无技能。

	学科：语言艺术	
	测量主题：要旨阅读	
	六年级	
分值 4.0		除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。

续表

分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… • 明显的和暗含的复杂因果关系； • 明显的和暗含的基本论点； • 明显的和暗含的复杂时间序列； • 带有明显的和暗含的基本解决办法的问题； • 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… • 明显复杂的因果关系； • 明显的基本论点； • 明显的复杂时间序列； • 明显的基本解决办法的问题； • 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

学科：语言艺术		
测量主题：要旨阅读		
五年级		
分值 4.0	除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 明显和暗含的复杂因果关系； · 带有明显和暗含的复杂时间序列； · 带有明显和暗含的基本解决办法的问题； · 带有明显和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 明显复杂的因果关系； · 明显的复杂时间序列； · 带有明显的基本解决办法的问题； · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

		学科：语言艺术
		测量主题：要旨阅读
		四年级
分值 4.0	除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 明显的和暗含的基本因果； · 明显的和暗含的简单时间序列； · 带有明显的和暗含的基本解决办法的问题； · 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 明显的基本因果； · 明显的简单时间序列； · 带有明显的基本解决办法的问题； · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

学科：语言艺术		
测量主题：要旨阅读		
三年级		
分值 4.0	除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 明显的和暗含的基本因果关系； · 带有明显的和暗含的简单的时间序列； · 带有明显的和暗含的基本解决办法的问题； · 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 明显的基本因果； · 明显的简单时间序列； · 带有明显的基本解决办法的问题； · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

		学科：语言艺术
		测量主题：要旨阅读
		二年级
分值 4.0	除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 明显的和暗含的基本因果； · 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 明显的基本因果； · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

	学科：语言艺术	
	测量主题：要旨阅读	
	一年级	
分值 4.0	除了分值 3.0 要求的表现外，学生的所学超出课堂所教，通过…… · 解释课文中哪些部分的内容是明显的，哪些部分的内容需要推理； · 解释并对有关内容的推理进行辩解。	
	3.5 分	除了分值 3.0 要求的表现外，学生能够超出课堂所教的内容，在一定程度上进行推理和知识的应用。
分值 3.0	在阅读适合本年级的材料时，学生能够确定课文的要旨，如…… · 带有明显的和暗含的单一故事线索的情节。 学生没有出现重大错误或遗漏。	
	2.5 分	处理比较简单的细节和程序时，学生没有出现重大错误或遗漏；部分掌握比较复杂的思想内容和程序。
分值 2.0	对比较简单的细节和程序的处理，学生没有出现重大错误或遗漏，如识别…… · 带有明显的单一故事线索的情节。 但是，学生在识别比较复杂的思想内容和程序方面出现重大错误或遗漏。	
	1.5 分	学生部分掌握比较简单的细节和程序，但在处理比较复杂的思想内容和程序时出现重大错误或遗漏。
分值 1.0	借助于帮助，学生能部分理解一些比较简单的细节和程序以及一些比较复杂的思想内容和程序。	
	0.5 分	借助于帮助，学生能部分理解一些比较简单的细节和程序，但不能理解比较复杂的思想内容和程序。
分值 0.0	甚至借助于帮助，学生也表现出不理解或无技能。	

参考文献

Ainsworth, L. (2003a). Power standards: Identifying the standards that matter most. Denver, CO: Advance Learning Press.

Ainsworth, L. (2003b). Unwrapping the standards: A simple process to make standards manageable. Denver, CO: Advance Learning Press.

Ainsworth, L. & Viegut, D. (2006). Common formative assessments. Thousand Oaks, CA: Corwin Press.

Airasian, P. W. (1994). Classroom assessment (2nd ed.). New York: McGraw-Hill.

American Association for the Advancement of Science. (2001). Atlas of Science Literacy. Washington, DC: American Association for the Advancement of Science and the National Science Teachers Association.

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1995). Learning and memory: An integrated approach. New York: Wiley.

Anderson, J. R. , Greeno, J. G. , Reder, L. M. , & Simon, H. A. (2000). Perspectives on learning, thinking and activity. Educational Researcher, 29 (4), 11 – 13.

Anderson, J. R. , Reder, L. M. , & Simon, H. A. (1995). Applications and misapplications of cognitive psychology to mathematics education. Unpublished paper, Carnegie Mellon University, Department of

Psychology, Pittsburgh, PA. Available: <http://act.psy.cmu.edu/personal/ja/misapplied.html>

Anderson, J. R. , Reder, L. M. , & Simon, H. A. (1996). Situated learning and education. *Educational Researcher*, 25 (4), 5 – 11.

Andrade, H. G. , & Boulay, B. A. (2003). Role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research*, 97 (1), 21 – 34.

Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychological Review*, 64, 359 – 372.

Atkinson, J. W. (1964). An introduction to motivation. Princeton, NJ: Van Nostrand.

Atkinson, J. W. (1987). Michigan studies of fear of failure. In F. Halisch & J. Kuhl (Eds.), *Motivation, intention and volition* (pp. 47 – 60), Berlin: Springer.

Atkinson, J. W. , & Raynor, J. O. (1974) . *Motivation and achievement*. New York: Wiley.

Baker, E. L. , Aschbacher, P. R. , Niemi, D. , & Sato, E. (1992). CRESST performance assessment models: Assessing content area explanations. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Bangert-Drowns, R. L. , Kulik, C. C. , Kulik, J. A. , & Morgan, M. T. (1991). The instructional effect of feedback on test-like events. *Review of Educational Research*, 61 (2), 213 – 238.

Bangert-Drowns, R. L. , Kulik, J. A. , & Kulik, C. C. (1991). Effects of classroom testing. *Journal of Educational Research*, 85 (2), 89 – 99.

Birnbaum, A. (1957). Efficient design and use of tests of mental ability for various decision-making problems. Series Report No. 58 – 16. Project No. 7755 – 23, USAF School of Aviation Medicine. Randolph Air Force Base, TX.

Birnbaum, A. (1958a). Further considerations of efficiency in tests of mental ability. Series Report No. 17. Project No. 7755 - 23, USAF School of Aviation Medicine. Randolph Air Force Base, TX.

Birnbaum, A. (1958b). On the estimation of mental ability. Series Report No. 15. Project No. 7755 - 23, USAF School of Aviation Medicine. Randolph Air Force Base, TX.

Black, P. , & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5 (1), 7 - 75.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1 (2), 1 - 12.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership*, 42 (3), 4 - 17.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement, Issue and Practice*, 16 (4), 21 - 33.

Boyer, E. L. , (1983). High school: A report on secondary education in America. New York: Harper & Row.

Boyer, E. L. , (1995). The basic school: A community for learning. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Brookhart, S. M. , (1994). Teachers' grading: Practices and theory. *Applied Measurement in Education*, 7 (4), 279 - 301.

Brookhart, S. M. , (2004). Grading. Columbus, OH: Pearson.

Burt, C. (1957). The distribution of intelligence. *British Journal of Psychology*, 48, 161 - 175.

Butler, D. L. , & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65 (3), 245 - 281.

Cahen, S. , & Davis, D. (1987). A between-grade levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal*, 24, 1 – 2.

Carnevale, A. P. , Gainer, L. J. , & Meltzer, A. S. (1990). *Workplace basics: The essential skills employers want*. San Francisco: Jossey-Bass.

Carroll, J. B. (1963). A Model of school learning. *Teachers College Record*, 64, 723 – 733.

Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, 8 (1), 26 – 31.

Cohen, J. C. , Cohen, P. , West, S. G. , & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Coladarci, T. , Smith, L. , & Whiteley, G. (2005). *The re-inventing schools implementation monitoring survey, Alaska benchmark/high school graduation qualifying examination data, and relationships between the two*. Anchorage, AK: Reinventing Schools Coalition.

Collins, L. M. , & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. New York: Cambridge University Press.

Covington, M. V. , Omelich, C. L. , & Schwarzer, R. (1986). Anxiety, aspirations, and self-concept in the achievement process: A longitudinal model with latent variables. *Motivation and Emotion*, 10, 71 – 88.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58 (4), 438 – 481.

Cross, K. P. (1998). *Classroom research: Implementing the*

scholarship of teaching. In T. Angelo (Ed.), Classroom assessment and research: An update on uses, approaches, and research findings (pp. 5 – 12). San Francisco: Jossey-Bass.

Durm, M. W. (1993, Spring). An A is not an A is not an A: A history of grading. *The Educational Forum*, 57, 294 – 297.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

English, F. W. (2000). Deciding what to teach and test: Developing, aligning, and auditing the curriculum. Thousand Oaks, CA: Corwin Press.

Farkas, S., Friedman, W., Boese, J., & Shaw, G. (1994). First thing first: What Americans expect from public schools. New York: Public Agenda.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105 – 146). New York: Macmillan.

Fitts, P. M., & Posner, M. I. (1967). Human performance. Belmont, CA: Brooks/Cole.

Flicek, M. (2005a). Consistency of rubric scoring for common assessments for math that are a part of NCSD body of evidence (BoE) for high school graduation. NCSD Assessment and Research Brief, 8, 1 – 21.

Flicek, M. (2005b). Moving toward a valuable and reliable teacher judgment of student performance on standards. NCSD Assessment and Research Brief, 4, 1 – 3.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, 24 (3), 21 – 29.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta analysis. *Exceptional Children*, 53 (3), 199 –

208.

Gagne, R. M. (1977). *The conditions of learning* (3rd ed.). New York: Holt, Rinehart & Winston.

Gagne, R. M. (1989). *Studies of learning: 50 years of research*. Tallahassee: Florida State University, Learning Systems Institute.

Gentile, J. R., & Lalley, J. P. (2003). *Standards and mastery learning*. Thousand Oaks, CA: Corwin Press.

Glaser, R., & Linn, R. (1993). Foreword. In L. Shepard, *Setting performance standards for student achievement* (pp. xiii-xiv). Stanford, CA: National Academy of Education, Stanford University.

Glass, G. V. (1976). Primary, secondary, and meta-analyses. *Educational Researcher*, 5, 3-8.

Goodlad, J. I. (1984). *A place called school: Prospects for the future*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guskey, T. R. (1980). What is mastery learning? *Instructor*, 90(3), 80-86.

Guskey, T. R. (1985). *Implementing mastery learning*. Belmont, CA: Wadsworth.

Guskey, T. R. (1987). Rethinking mastery learning reconsidered. *Review of Educational Research*, 57, 225-229.

Guskey, T. R. (Ed.). (1996a). *Communicating student learning* (1996 ASCD Yearbook). Alexandria, VA: Association for Supervision and Curriculum Development.

Guskey, T. R. (1996b). Reporting on student learning: Lessons from the past—Prescriptions for the future. In T. R. Guskey (Ed.), *Communicating student learning* (1996 ASCD Yearbook, pp. 13-24). Alexandria, VA: Association for Supervision and Curriculum Development.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin

Press.

Haladyna, T. M. (1999). A complete guide to student grading. Boston: Allyn & Bacon.

Hambleton, R. K. , Swaminathan, H. , & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE Publications.

Harrison, G. A. , Weiner, J. S. , Tanner, J. M. , & Barnicot, N. A. (1964). Human biology: An introduction to human evolution, variation, and growth. London: Oxford University Press.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 19, 49 – 78.

Hattie, J. (1985). Methodology review: Assessing the unidimensionality of tests and items. *Applied Psychological Measurement*, 9 (2), 139 – 164.

Hattie, J. A. (1992). Measuring the effects of schooling. *Australian Journal of Education*, 36 (1), 5 – 13.

Hattie, J. , Krakowski, K. , Rogers, H. J. , & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20 (1), 1 – 14.

Haycock, K. (1998). Good teaching matters... a lot. *Thinking K – 16*, 3 (2), 1 – 14.

Hays, W. L. (1973). Statistics for the social sciences. (2nd ed.). New York: Holt, Rinehart, & Winston.

Kendall, J. (2000). Topics: A roadmap to standards. *NASSP Bulletin*, 84 (620), 37 – 44.

Kendall, J. S. , & Marzano, R. J. (2000). Content knowledge: A compendium of standards and benchmarks for K – 12 education (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Kifer, E. (1994). Development of the Kentucky instructional re-

sults information system (KIRIS). In T. R. Guskey (Ed.). High stakes performance assessment: Perspective on Kentucky's educational reform (pp. 7 – 18). Thousand Oaks, CA: Corwin Press.

Kleinsasser, A. (1991, September). Rethinking assessment: Who's the expert? Paper presented at the Casper Outcomes Conference, Casper, WY.

Kluger, A. N. , & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary intervention theory. *Psychological Bulletin*, 119 (2), 254 – 284.

LaBerge, D. , & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading comprehension. *Cognitive Psychology*, 6, 293 – 323.

Lauer, P. A. , Snow, D. , Martin-Glenn, M. , Van Buhler, R. , Stoutemyer, K. , & Snow-Renner, R. (2005). The influence of standards on K – 12 teaching and student learning: A research synthesis. Aurora, CO: Mid-continent Research for Education and Learning.

Lord, F. M. (1959, June). Problems in mental test theory arising from errors of measurement. *Journal of the American Statistical Association*, 54 (286), 472 – 479.

Lord, F. M. , & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lou, Y. , Abrami, P. C. , Spence, J. C. , Poulsen, C. , Chambers, B. , & d'Appollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66 (4), 423 – 458.

Magnusson, D. (1966). Test theory. Reading, MA: Addison-Wesley.

Marzano, R. J. (2000). Transforming classroom grading. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2001). Designing a new taxonomy of educational

objectives. Thousand Oaks, CA: Corwin Press.

Marzano, R. J. (2002a). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, 15 (3), 249 – 268.

Marzano, R. J. (2002b). Identifying the primary instructional concepts in mathematics: A linguistic approach. Centennial, CO: Marzano & Associates.

Marzano, R. J. (2003). What works in school: Translating research into action. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2004a). Applying the theory on measurement of change to formative classroom assessments. Centennial, CO: Marzano & Associates.

Marzano, R. J. (2004b). Building background knowledge for academic achievement. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2004c). Workshop materials. Centennial, CO: Marzano & Associates.

Marzano, R. J. (2006). Unpublished data.

Marzano, R. J. , & Haystead, M. (in press). Making standards useful to classroom teachers. Centennial, CO: Marzano & Associates.

Marzano, R. J. , & Kendall, J. S. (1996). A comprehensive guide to designing standards-based districts, schools, and classrooms. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. , & Kendall, J. S. (in press) . The new taxonomy of educational objectives. Thousand Oaks, CA: Corwin Press.

Marzano, R. J. , Kendall, J. S. , & Cicchinelli, L. F. (1998). What Americans believe students should know: A survey of U. S. adults. Aurora, CO: Mid-continent Research for Education and

Learning.

Marzano, R. J. , Kendall, J. S. , & Gaddy, B. B. (1999). *Essential knowledge: The debate over what American students should know*. Aurora, CO: Mid-continent Research for Education and Learning.

Marzano, R. J. , Marzano, J. S. , & Pickering, D. J. (2003). *Classroom management that works: Research-based strategies for every teacher*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. , Pickering, D. J. , & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

McMillan, J. H. (1997). *Classroom assessment: Principles and practices for effective instruction*. Needham Heights, MA: Allyn & Bacon.

McMillan, J. H. (2000). *Basic assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Press.

Mead, R. , & Pike, D. J. (1975). A review of response surface methodology from a biometric viewpoint. *Biometrics*, 31, 803 – 851.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Education Goals Panel. (1991). *The National Education Goals report: Building a community of learners*. Washington, DC: Author.

National Education Goals Panel. (1993, November). *Promises to keep: Creating high standards for American students. A report on the review of education standards from the Goal 3 and 4 Technical Planning Group to the National Education Goals Panel*. Washington, DC: Author.

National Governors Association. (1996, March). *1996 National Education Summit policy statement*. Washington, DC: Author.

National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22 (2), 155 – 175.

Newell, A. , & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.). *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Nunally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Nye, B. , Konstantopoulos, S. , & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*. 26 (3), 237 – 257.

O'Connor, K. (1995). Guidelines for grading that support learning and success. *NASSP Bulletin*, 79 (571), 91 – 101.

Ohio Department of Education. (2001). *Academic content standards: K – 12 mathematics*. Columbus, OH: Author.

Olson, L. (1995, June 14). Cards on the table. *Education Week*, 23 – 28.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research and Evaluation*, 8 (11) [Online]. Retrieved December 29, 2003, from <http://PAREonline.net/getvn.asp?v=8&n=11>

Plake, B. S. , Hambleton, R. K. , & Jaeger, R. M. (1995). A new standard-setting method for performance assessments: The dominant profile judgement method and some field test results. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision

and Curriculum Development.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Ravitch, D. (1983). The troubled crusade: American education 1945 – 1980. New York: Basic Books.

Reeves, D. B. (2002). Holistic accountability: Serving students, schools, and community. Thousand Oaks, CA: Corwin Press.

Reeves, D. B. (2004, December). The case against the zero. *Phi Delta Kappan*, 86 (4), 324 – 325.

Rogosa, D. R. , Brandt, D. , & Zimowsky, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 90, 726 – 748.

Ross, J. A. , Hogaboam-Gray, A. , & Rolheiser, C. (2002). Student self-evaluation in grade 5 – 6 mathematics: Effects on problem-solving achievement. *Educational Assessment*, 8 (1), 43 – 59.

Rothblum, E. D. , Solomon, L. J. , & Murakami, J. (1986). Affective, cognitive, and behavioral differences between high and low procrastinators. *Journal of Counseling Psychology*, 33, 387 – 394.

Rothman, R. (1995). The certificate of initial mastery. *Educational Leadership*, 52 (8), 41 – 45.

Sanders, W. L. , & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS); Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8, 299 – 311.

Schmidt, W. H. , McKnight, C. C. , & Raizen, S. A. (1996). Splintered vision: An investigation of U. S. science and mathematics education: Executive summary. Lansing, MI: U. S. National Research Center for the Third International Mathematics and Science Study. Michigan State University.

Schreiber, R. , & Battino, W. (2002). A guide to reinventing

schools. Chugach, AK: Reinventing Schools Coalition.

Scriven, M. (1967). The methodology of evaluation. In R. F. Stake (Ed.), Curriculum evaluation: American Educational Research Association monograph series on evaluation, No. 1 (pp. 39 – 83). Chicago: Rand McNally.

Secretary's Commission on Achieving Necessary Skills. (1991). What work requires of schools: A SCANS report for America 2000. Washington, DC: U. S. Department of Labor.

Seigel, D. G. (1975). Several approaches for measuring average rate of change for a second degree polynomial. The American Statistician, 29, 36 – 37.

Seligman, M. E. P. (1975). Helplessness: On depression, development, and death. San Francisco: Freeman.

Seligman, M. E. P., Maier, S. F., & Greer, J. (1968). The alleviation of learned helplessness in the dog. Journal of Abnormal Psychology, 73, 256 – 262.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for general causal inference. Boston: Houghton Mifflin.

Smith, J. K., Smith, L. F., & DeLisi, R. (2001). Natural classroom assessment. Thousand Oaks, CA: Corwin Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263 – 331). New York: American Council on Education and Macmillan.

Snyder, C. R. (1984, September). Excuses, excuses: They sometimes actually work to relieve the burden of blame. Psychology Today, 18, 50 – 55.

Solomon, L. J., & Rothblum, E. D. (1984). Academic procrastination: Frequency and cognitive-behavior correlates. Journal of Counse-

ling Psychology, 31, 503 – 509.

Spady, W. G. (1988). Organizing for results: The basics of authentic restructuring and reform. *Educational Leadership*, 46 (2), 4 – 8.

Spady, W. G. (1992). It's time to take a clear look at outcome-based education. *Outcome*, 11 (2), 6 – 13.

Spady, W. G. (1994). Choosing outcomes of significance. *Educational Leadership*, 51 (6), 18 – 22.

Spady, W. G. (1995). Outcome-based education: From instructional reform to paradigm restructuring. In H. H. Block, S. T. Everson, & T. Guskey (Eds.), *School improvement programs* (pp. 367 – 398). New York: Scholastic.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Merrill.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. (2nd ed.). Columbus, OH: Merrill.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning*. Portland, OR: Assessment Training Institute.

Terwilliger, J. S. (1989, June). Classroom standard setting and grading practice. *Educational Measurement: Issues and Practice*, 8 (2), 15 – 19.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurement*. New York: Teachers College Press.

Tyack, D. K., & Tobin, W. (1994). The “grammar” of schooling: Why has it been so hard to change? *American Educational Research Journal*, 31 (3), 453 – 479.

Valencia, S. W., Stallman, A. C., Commeyras, M., Pearson, P. D., & Hartman, D. K. (1991). Four measures of topical knowledge: A study of construct validity. *Reading Research Quarterly*, 26

(3), 204 – 233.

Walberg, H. J. (1997). Uncompetitive American schools: Causes and cures. In *Brookings Papers on educational policy*, 1997. Washington, DC: Brookings Institution.

Weiner, B. (1972). *Theories of motivation: From mechanism to cognition*. Chicago: Markham.

Weiner, B. (1974). *Achievement motivation and attribution theory*. Morristown, NJ: General Learning Press.

Weiner, B. , Frieze, L. , Kukla, A. , Reed, L. , Rest, S. , & Rosenbaum, R. (1971). Perceiving the causes of success and failure. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 95 – 121), Morristown, NJ: General Learning Press.

Wiggins, G. , & McTighe, J. (2005). *Understanding by design* (expanded 2nd ed). Alexandria, VA: Association for Supervision and Curriculum Development.

Willett, J. B. (1985). *Investigating systematic individual difference in academic growth*. Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345 – 422. Washington, DC: American Educational Research Association.

Wilson, V. L. , & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects on experimental design. *American Educational Research Journal*, 19 (2), 249 – 258.

Wright, S. P. , Horn, S. P. , & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher education. *Journal of Personnel Evaluation in Education*, 11, 57 – 67.

作者简介

作者罗伯特·J. 马扎诺是科罗拉多州奥罗拉市的中部北美大陆教育学习研究所 (McREL) 的高级学者、威斯康星州大学的副教授、马扎诺协会主席。他开发了用于基础教育课堂实践和评价的研究项目，把当前的认知研究理论转化为有效的教学方法。他还是一位国际知名的教师教育培训者、演讲者和多产作者，在阅读和写作教学、思维技能、学校效应、认知和标准的实施，特别是课堂评价等方面撰写了 24 本著作和 150 多篇文章。近几年的作品有：《学校领导的作用：从研究到成果》(ASCD, 2005)；《为学术成就构建背景知识：对学校作用的研究》(ASCD, 2004)；《课堂管理的作用：基于研究的教师策略》(ASCD, 2003)；《学校的作用：把研究转化为行动》(ASCD, 2003)；《课堂教学的作用：提高学生成就的研究策略》(ASCD, 2001)；《为教育目标设计新的分类学》(Corwin, 2001)。马扎诺在纽约的爱奥纳学院获得英语专业的学士学位，在西雅图大学获得了语言艺术教育专业的硕士学位，在华盛顿大学获得了课程和教学论的博士学位。他的电子邮箱是：robertjmarzano@aol.com。

译 后 记

在今年南方罕见的雪灾冰冻中，在农历新年的爆竹声中，我和彭春艳老师一直在各自寒冷的书斋里伏案查译，驰骋在罗伯特·J. 马扎诺的《有效的课堂评价手册》中，赶在新学期开学前夕完成了译稿。收笔之余，仍感到意犹未尽。因为在翻译的过程中，我被马扎诺先生扎实的教育理论功底、严谨的数理逻辑思维、丰富的教育教学实践经验、对课堂评价的独特视角、精确的教育测量和统计范例所深深吸引……

《有效的课堂评价手册》著于2006年。这正是美国政府实施各学科教育标准的关键时期，也是世界各国纷纷研究、制定和实施教育标准或课程标准，并对学科教育和课程教学开始进行评价的关键时期。在《有效的课堂评价手册》一书中，马扎诺通过运用教育学理论、测试学理论、教育测量和统计的方法，研究分析了大量的测试分数、测量数据和评价案例，列举了许多测量样本、图表和模式，深刻地论述了形成性课堂评价的意义和原则，详尽地阐释了设计有效的课堂评价综合体系的模式和方法。该书顺应了当前教育评价的国际发展趋势，对如何通过课堂教学来落实和达到各学科的教育标准和课程标准、科学评价学生的学业成就、提高课堂教学效果，具有深刻的现实意义和很强的指导作用。可以说，这本手册对于正在从事或有志于从事教育教学工作的人们来说，都是一本必备读物和启蒙开道的案头指南。

教育改革从哪里进行？马扎诺说，“即使是最好的改革也无法体

现出在教育体制上对学生成绩影响最大的重要变量——任课教师”。

该书引用美国研究员保罗·怀特和桑得拉·鸿等人经过分析6万多名小学三年级至五年级学生的测试分数，得出的观察意见是：影响学生学习的最重要因素是教师。一个教学效果好的教师比学校教育的其他任何可控因素更能强化学生的学习。因此，必须“通过提高教师的教学效果来改进教育”。

教师教学效果的提高涉及许多方面的因素，一个世纪以来，无数的教育专家和学者都从不同的角度对此进行了研究和论述。而在《有效的课堂评价手册》中，马扎诺仅把关注点放在常被教师忽视的如何进行有效的课堂评价上。正如他开章明义所说的：该书只强调教学中常常在讨论如何提高学生成绩的途径中被忽视的一个方面：课堂评价（classroom assessment）。他和同事们综合了250多项研究表明，课堂评价是教师手中最强有力的武器之一。若全美的数学教师都能通过课堂评价获得0.7的效应规模（effect size），那么，数学成绩“中等”的美国，其“数学成绩分数能提高到继太平洋圈内的新加坡、韩国、日本和中国香港之后的‘前五名’（top five）”。可见，课堂评价确实能促进学生学习，成绩提高的幅度还相当大。

与其他论述课堂评价的篇著不同的是，马扎诺主要阐述怎样的课堂评价才能让学生真正了解自己，从而促进学生的自主学习，以及怎样的反馈才能给学生提供有效的指导。

教师能给学生提供的最直接、最有效的反馈方法就是让学生跟踪自己在学科学习上所取得的进步。但长期以来在所使用的成绩报告单中，无论是用百分数、五分制还是A、B、C、D字母等级表示都是远远不够的，因为它们无法提供促进学生学习所需要的详细的反馈信息。因此，马扎诺说，“改变成绩单的形式有可能使美国的基础教育发生史无前例的巨大变化。”他在书中用文字和图表的形式详细介绍了任课教师应该如何改进以往单一的成绩报告形式，在整个评价打分阶段（the whole grading period）与学生共同记录成绩量表（scoring scale），给学生提供一个可视的成绩表述。既有教师和学生的估计成绩，也有学生本人第一次、第二次、第三次……评价的得

分，还有不同教师对该生的评价打分，使学生不用在班上进行成绩排名就可以判断自己的学习是否成功。让学生看见他们“学习知识的收获”，看见自己的进步，激发他们的“内在”动机，促进他们的自主学习。这种在课堂教学过程中师生双方不断形成和积累的学生个体成绩量表，能提供详细的、及时的反馈信息，还能给学生提供一个确定自己的学习目标的杆尺，测量学生对超出课堂所教的知识进行推理和运用的能力。使学习从变量转为常量，使教学时间从常量转为变量。这就是书中所倡导的以课堂为基础的形成性评价的目的。

该书的特点是用大量的案例和精细的图表来介绍形成性课堂评价系统的设计。包括：

一、阐述如何把学科内容细化分成测量主题系列，让教师在不同的年级、在规定的时间内完成。

二、详细介绍了简化版和完整版测量主题的标准量表，从而能清晰地了解学生对每个主题的掌握情况。

三、让教师学会设计适用于形成性评价的软件系统，使教师能创建多个测量主题并能轻松输入学生在每个主题上的得分。

四、用幂函数式来计算和统计学生的真实得分，并能观测学生的成长轨迹，提供有效的测量信度。

可以说，21 世纪迎来了评估、评价的时代。人类的各种工作和活动都要接受评估或评价，通俗地讲，就是人们要按照各自不同的标准来进行检验检查。对教育界而言，近年来，国际上对教育评价的重视程度越来越高，研究者们不断推出评价项目。如国际教育成就评价协会（IEA）主持的国际数学和科学研究（TIMSS）、经济合作与发展组织（OECD）主持的国际学生评估项目（PISA），都是国际上很有影响的学业成就评价项目，向世人展示了近年来国际学生评价项目的改革发展趋势。从微观来说，对教师而言，课堂应涵盖多少知识内容？应该实施什么样的教学方法？学生是否掌握了所要求的知识和技能？学生的测试成绩反映了他们的真实水平吗？如何

提高测试的效度和信度？能否对学生学习的进步及提高做出清晰的、鼓励性的指导？怎样的成绩反馈才能鼓励而不是打击学生的进步？凡此种种，都与课堂评价有关。

在我国中小学的教学研究中，课堂评价还是一块待开垦的处女地，很多教师不了解教育测量理论和测试统计方法。教师们缺乏对课堂评价技能的必要培训，相关的基础知识还不够普及。随着我国素质教育的进一步深化、随着我国基础教育新课标的颁布和实施、随着教师教育发展的专业化程度不断提升，教师们都面临着重新学习和终身学习的挑战，才能在以建构标准和评价体系为新的突破口的国际教育变革中把握方向。因此，掌握有效的课堂评价方法就显得更为必要和迫切。

该书的目的就是“给教育工作者装备上他们这个所需要的工具”。通过真实的成绩反馈有效地大幅度提高学生的学业成绩。这种改革将会把基础教育提升到一个前所未有的效率平台，并促进一个更有效的教育体系的发展。我们希望通过引进该书关于课堂评价的一些最新观点、方法和手段，拓展我国中小学一线教师的视野，使他们了解具有国际视野的学生学业成就评价模式，掌握和探究有效的课堂评价策略，促进我国的基础教育评价体系与国际接轨，提高我国基础教育的教学质量。

本书译者的具体分工是：目录、表目录、第一章、第二章、第三章、术语注释和作者简介由邓妍妍译；第四章、第五章、第六章、第七章及附录由彭春艳译。程可拉教授和吕林海博士指导、校稿、修改和整理。

该译稿化为墨迹之时，特别想感谢程可拉博士和胡庆芳博士。是他俩精心策划和组织我们一批高校英语教师组成翻译团队，带领我们在教学之余欣然从译。他俩不仅在原著的挑选、译稿的校正等方面做了大量的工作，不仅在我们翻译过程中给予学术上的指导和思想上的鼓励，更重要的是，让我们更关注国外的教育教学研究的发展趋势，思考我国教育教学存在的问题，尽快把外国教育专家学

者的最新研究成果介绍给我国广大的一线教师、应用到我国的教育
教学实践中。无论是程博、胡博，还是我们这个团队的教师，我们
所做的一切，全因心中那份不舍的教育情结、对我国教育事业发展的
美好心愿，以及作为教育工作者的那份责任。

湛江师范学院外国语学院
邓妍妍

2008 年早春二月于寸金湖畔

摆渡者教师书架(现已出版部分)

丛书名称	主编或作者	书 名	定价(元)
大师背影书系	张圣华	《陶行知教育名篇》	24.90
		《陶行知名篇精选》(教师版)	16.80
		《朱自清语文教学经验》	15.80
		《夏丏尊教育名篇》	16.00
		《作文入门》	11.80
		《文章作法》	11.80
		《蔡元培教育名篇》	19.80
		《叶圣陶教育名篇》	17.80
教育寻根丛书	张圣华	《中国人的教育智慧·经典家训版》	49.80
		《过去的教师》	32.80
		《追寻近代教育大师》	29.80
		《中国大教育家》	22.80
杜威教育丛书	单中惠	《杜威教育名篇》	19.80
		《杜威学校》	25.80
		《杜威在华教育讲演》	29.80
班主任工作创新丛书	杨九俊	《班集体问题诊断与建设方略》	19.80
		《班主任教育艺术》	22.80
		《班级活动设计与组织实施》	23.80
新课程教学问题与解决丛书	杨九俊	《新课程教学组织策略与技术》	16.80
		《新课程教学现场与教学细节》	15.00
		《新课程备课新思维》	16.80
		《新课程教学评价方法与设计》	16.80
		《新课程说课、听课与评课》	16.80
新课程课堂诊断丛书	杨九俊	《小学语文课堂诊断》(修订版)	18.60
		《小学数学课堂诊断》(修订版)	18.60
		《小学综合实践活动课堂诊断》	23.60
		《小学品德与生活(品德与社会)课堂诊断》	22.80
名师经验丛书	肖 川	《名师备课经验》(语文卷)	25.80
		《名师备课经验》(数学卷)	25.60
		《名师作业设计经验》(语文卷)	25.00
		《名师作业设计经验》(数学卷)	25.00
个性化经验丛书	华应龙	《个性化作业设计经验》(数学卷)	19.80
		《个性化备课经验》(数学卷)	23.80
	于永正	《个性化作业设计经验》(语文卷)	20.60
		《个性化备课经验》(语文卷)	23.00

续表

丛书名称	主编或作者	书 名	定价(元)
深度课堂丛书	《人民教育》 编辑部	《小学语文模块备课》	18.00
		《小学数学创新性备课》	18.60
课堂新技巧丛书	郑金洲	《课堂掌控艺术》	17.80
课改新发现丛书	郑金洲	《课改新课型》	19.80
		《学习中的创造》	19.80
		《多彩的学生评价》	26.00
教师成长锦囊丛书	郑金洲	《教师反思的方法》	15.80
校本教研亮点丛书	胡庆芳	《捕捉教师智慧——教师成长档案袋》	19.80
		《校本教研实践创新》	16.80
		《校本教研制度创新》	19.80
		《精彩课堂的预设与生成》	18.00
美国教育新干线丛书	胡庆芳	《美国学生课外作业集锦》	35.80
美国中小学读写教学 指导译丛	胡庆芳	《教会学生记忆》	22.50
		《教会学生写作》	22.50
		《教会学生阅读:方法篇》	25.00
		《教会学生阅读:策略篇》	24.80
提升教师专业实践力 译丛	胡庆芳 程可拉	《创造有活力的学校》	22.50
		《有效的课堂管理手册》	24.00
		《有效的课堂教学手册》	32.80
		《有效的课堂指导手册》	24.80
		《有效的教师领导手册》	25.80
		《提升专业实践力:教学的框架》	30.80
		《优化测试,优化教学》	22.50
		《有效的课堂评价手册》	26.80
教育心理	俞国良 宋振韶	《现代教师心理健康教育》	25.80
教师在研训中成长丛书	胡庆芳 林相标	《校本培训创新:青年教师的视角》	21.80
		《教师专业发展:专长的视野》	21.60
		《听诊英语课堂:教学改进的范例》	31.60
其他单行本	胡庆芳	《美国教育360度》	15.80
	徐建敏 管锡基	《教师科研有问必答》	19.80
	杨桂青	《英美精彩课堂》	17.80
	陶继新	《教育先锋者档案》(教师版)	16.80
	单中惠	《西方教育思想史》	59.80
	孙汉洲	《孔子教做人》	27.90
	丰子恺	《教师日记》	24.80
	陶 林	《家有小豆豆》	27.00
	徐 洁	《教师的心灵温度》	26.50

“新课程教学问题与解决丛书”荣获第七届全国高校出版社优秀畅销书一等奖!

《陶行知教育名篇》荣获第八届全国高校出版社优秀畅销书一等奖!

“大师背影书系”荣获第八届全国高校出版社优秀畅销书二等奖!

《名师作业设计经验》(语文卷)、《名师作业设计经验》(数学卷)、《名师备课经验》(语文卷)荣获第17届上海市中小学幼儿园优秀图书三等奖!

《西方教育思想史》荣获全国第二届教育科学优秀成果二等奖(1999)!

在2006年全国教师教育优秀课程资源评审中,“新课程教学问题与解决丛书”中的《新课程教学组织策略与技术》《新课程教学现场与教学细节》《新课程备课新思维》和《新课程说课、听课与评课》被认定为新课程通识课推荐使用课程资源,《陶行知教育名篇》被认定为新课程公共教育学推荐使用课程资源,《课改新课型》被认定为新课程通识课优秀课程资源,《小学语文课堂诊断》被认定为新课程语文课优秀课程资源,《小学数学课堂诊断》被认定为新课程数学课推荐使用课程资源!



提/升/教/师/专/业/实/践/力/译/丛

网站:<http://www.msyd.cn>

策划:蚂蚁王

有效的课堂 评价手册

罗伯特·J.马扎诺：华盛顿大学课程和教学论博士，科罗拉多州奥罗拉市中部大陆教育学习研究所的高级学者、威斯康星州卡迪纳尔·斯特里奇大学副教授。国际知名的教师教育培训者、演讲者。

责任编辑:王峥媚

封面设计:  翼之扬设计

上架建议 教育阅读
定价:26.80元

ISBN 978-7-5041-4311-2



9 787504 143112 >