教育方案评价丛书

如何测量成绩

[美] 林恩·L·莫里斯
卡罗尔·T·菲茨-吉本 著

蒋一驭 张九超 译
赵永年 校

上海翻译出版公司

Lynn L · Morris Carol T · Fitz-Gibbon
HOW TO MEASURE ACHIEVEMENT
Sage Publications, Inc. 1978
本书根据塞奇出版公司1978年版译出

教育方案评价丛书
如何测量成绩
[美] 林恩·L·莫里斯 卡罗尔·T·菲茨-吉本 著
蒋一飒 张九超 译
赵永年 校
上海翻译出版公司
（上海复兴中路597号）
新华书店上海发行所发行 上海东方印刷厂印刷
开本850×1156 1/32 印张4.25 字数110000
1988年9月第1版 1988年9月第1次印刷
印数1—4,000
ISBN7-80514-146-0/G·104 定价：1.85元

中译本序言

这套《教育方案评价丛书》系根据美国加州大学洛杉矶分校评价研究中心主编的评价丛书译成。原文分为八册，译著亦同。各分册书名是：

- 《评价人员手册》
- 《如何处理评价目标》
- 《如何设计方案评价》
- 《如何测量方案实施》
- 《如何测量成绩》
- 《如何测量态度》
- 《如何进行统计分析》
- 《如何写评价报告》

丛书基本体现了评价工作过程的全貌，同时也大致反映了目前常见的评价方法的轮廓。

半个世纪以来，人们在世界各地对社会上从生产到研究的各种活动方案，广泛地进行了评价。用于评价的方法和技术也因此取得了较显著的进步。这种进展主要是从对教育方案——除了学校教育以外，还包括各种专业性的培训活动——积极进行评价的工作经验中产生的。在这段时期内，评价对象种类很多，但以各种教学改革方案和战时人员培训方案的目标及其实现程度为主要组成部分。

在这段发展过程中产生了一些重要经验，其中一条是：应当努力做到使评价有助于取得更多、更好的具有社会价值的效果，即通过评价提高活动方案的社会功效。——一句话，评价是为了提高。——这条经验主要来自教育评价。教育，作为嬗传人类文明并促

使其加速进步的精神生产，必须经常进行适应于物质文明和精神文明发展需要的改革，才能充分发挥提高人民素质，多出合格人才的社会功能。

在贯彻中共中央关于教育体制改革的决定过程中，全国正在积极开展教育评价工作。如：工科、师范高等院校等正在分别筹划对专业教学方案和学校管理水平进行评价的工作。从更广范围来说，全国正处在举世瞩目的改革过程中。改革的具体建议来自对过程和现况的评价。要做好评价工作，就必然会愈益讲究评价方法。

半世纪来，世界各地在评价工作发展过程中积累了不少关于适应评价对象选用恰当的观察、分析和判断方式的经验，也就是选用评价模型的经验。他山之石，可以攻玉。国外经验中有不少可供参考借鉴之处。不过，当务之急，首先在于早让从事评价工作的广大干部迅速掌握基本方法。就这点来说，这套丛书的实效会是很显著的，值得向大家推荐。

这套丛书的译者多数是青壮年研究工作者，这是十分可喜的事。希望这些出版物的出现能引起更多人对于评价这门学科的重视，来共同推动有关工作和研究的开展，从而有益于我国社会主义改革与建设的进程。

邱　渊　1986年7月

目　录

第一章

为方案评价测量成绩

初步考虑

本书的目的是帮助确定正在评价的方案在多大程度上已经达到了成绩目标(通常就是方案正式介绍中提到的那些目标)，但可能还有一些其他的目标因种种原因(注1)被补充进测量计划中了。本书拟就以下两个方面帮助实现上述目的：(1)提出建议、程序和经验方法，以开展与测量方案评价的成绩有关的评价工作；(2)介绍作为编写和选择成绩测验以及解释那些测验结果的某些基础理论。

本书的内容是以下面三点作为根据的：(1)洛杉矶加利福尼亚大学评价研究中心的评价者们的经验；(2)教育测量领域中专家们的意见；(3)在学校、学区和州范围内曾使用过实地测验卷子的人的评论。可能的话，所有的程序都以步进方式提出。这样的叙述将给评价者以最大的实际帮助，而理论上的混乱将减到最低限度。不过请记住，本书提议的许多程序都是在最有利的条件下进行选择成绩测量、实施成绩测量和解释成绩测量的方法。但是，由于评价的情况很少会与这里所设想的完全吻合，因此，不应该指望可以

注1：无论哪种情况都可能提醒评价者去测量那些未为正式方案目标所提到的技能成就。例如，评价者也许想了解一下附带成果，即与方案有关，但未被方案包括在内的技能成绩。也可能会决定检查在未为方案所强调的某课程范围内，学生参与了方案之后是否成绩下降了。

丝毫不差地照搬本书提出的建议。希望先检查本书提出的原则和实例，并且使那些原则和实例适用于评价者本人的时间限制、评价范围和数据要求等方面的条件。

本书的一个根本目的是要通过评价者自己的工作提供帮助，目的在于改善这块新兴的、日益发展的评价领域的实践情况。尽管多数评价都不能就方案或方案的组成部分的价值提出绝对清楚的信息，然而任何评价者的任务还是必须提供尽可能理想的信息。这就是说，一方面要在所处的环境条件容许范围内收集最可靠的信息；另方面要用一种每个评价听取人(注2)都可以用的方式向他们提供那些信息。关于方案成绩目标的任务就是为每个听取人介绍该方案似乎已经产生的成绩大小，那些介绍且由以下事实予以证明：已经使用的成绩测量手段对于方案的目标是很敏感的，它们似乎会给出正确的信息。

本书各章分别提出了一个方面的实用意见。第一章内的决定方向的问题表将帮助决定该测量些什么以及在测量成绩上该花费多少时间和多少精力。本章和整本书都涉及到成绩测验。有的评价者也许会选择通过不太正统的手段——例如，通过给学生的设计进行评定的手段——来估计某方案的认知上的效果，然而教育评价方面的多数测量上的争论都是围绕正式编就的测验的。何况，本书所讨论的那些原则都是很一般的，它们几乎可以在收集和报告成绩信息的任何情景里都获得应用。

第二和第三章的讨论题是选择和获得已发表的测验。第二章介绍了几种也许已经可以弄到的成绩测验数据（根据课程内包含

注2：评价听取人是评价工作的一个重要概念。听取人是评价者的主人，评价者又是评价信息的收集者。如果评价者正在写的一份报告是要让人们读的，那么每个评价者至少要有一个听取人。许多评价都有好几个听取人。听取人是这样的一个人或一个组，他(或他们)因某个明确的目的需要从评价中获取信息。那些因为需要监视政治气候而想随时记录方案制定情况的行政管理人员是一个潜在的听取人；而那些想获得有关方案某特定组成部分正在产生多少成绩这一方面信息的课程编写者则是另一个潜在的听取人。每个听取人需要不同的信息；而且重要的是，每个听取人都有不同的标准来决定信息是否可信和可靠。

的测量工具和州或学区指定的测量手段所获得的数据)。该章讨论了评价可能采用的几种方法，还把能够买到或借到已经发表的测验的来源列制成表。

既然许多评价会要求采用一份已经发表的成绩测验，因此，在选择测验或确定评价要求采用的测验的合适性时都得有个依据。为了估计测验的合适性而提供一种部分定量法，第三章提出了一份方案测验比较表(TPTC)。如果循着步进程序完成该表，便可以算出对于正在评价的那个方案的测验相对合适性指数。那些指数显示了测验与方案的最主要目标之间的相配情况、该测验所包含的那些目标的比例、以及与方案相关联的测验项目的比例。

第四和第五章分别论述了测验的技术质量和应用。第五章讨论了成绩测验中信度和效度的问题。第六章简单讨论了测验的解释与评分报告的问题。

评价者的职责

《教育方案评价丛书》(本书即为其中的一本)主要是为从事方案评价工作的人而写的。作为一名方案评价者，花在选择、编制、实施和评定成绩测量工具上的精力，和需要收集有关每个工具测量精确性的信息量，在很大程度上都将取决于你在被评价的那个方案里所任的职务。根据被指定要完成的任务，可采取下面两种形式中的一种来开展评价工作。在某些情况下也许会发现，评价者必须同时承担每种职责的某些部分：

1. 可能需要评价者负责写出一份关于方案有效性的终结性报告。在这种情况下，评价者也许会写一份报告给方案的拨款机构、政府办公室、或者方案赞助者的另一位代表。他们也许希望在报告中介绍该方案、对其目标作出陈述、估计这些目标已经达到的程度、记录方案似乎已经取得的意料之外的结果，以及还可能要把该方案与备择方案作一比较。如果上述描述

与评价者的工作相符，那么这就是一名终结性评价者。

2. 也许指定评价者承担方案设计者和制定者的助手或顾问的任务。那么可能会被召去检查方案制订情况，找出方案中尚需完善的地方、查明其潜在的问题、为方案工作人员介绍其活动特性，也许还要定期测验方案参与者在达到认知上或态度上的目标方面的进展情况。如是这样，评价者就有许多工作要做，所有那些工作都要求与方案工作人员合作，以帮助他们制定出尽可能是最佳的方案。也许他们要评价者写一份报告，也许不要。如果这样大致限定的工作职责看来较接近评价者的工作，那么这就是一名形成性评价者。

本书中有关成绩测量的设计、实施和解释等方面的信息都是专门供形成性评价者和终结性评价者共同使用的。但是，评价者在成绩测量方面未来的工作可能会变，这要看他在这两种评价者之间任什么角色。

终结性评价者的未来工作

终结性评价者主要关心的是找出并应用工具，以测量方案是否达到了它的总体目标。为此，终结性评价者必须密切注意方案所宣布的成绩目标和显而易见的成绩目标。他的兴趣在于设计或购买对于确定那些目标完成状况最为敏感的测量工具。终结性评价者关于方案对成绩的全面影响的兴趣还应扩大，他还应关心测量那些似乎正在出现，然而方案设计者却未在他们制定的目标中提及到的认知上的收益。

终结性评价者可能还会有别的原因要测量成绩。如果他正采用一种评价设计来确定只可能是由于该方案的缘故才获得的良好结果，那么他也许希望采用某种成绩测验来给方案组和比较组分配学生、班级或者学校（注3）。这种对于成绩信息的普通应用称之

注3：可以在大多数论述教育研究的教科书中找到评价设计的讨论。特别请参阅《如何设计方案评价》。

为分组抽样或分层抽样，它保证了产生的小组一开始在初始成绩方面就已经尽可能相象的了。

终结性评价者可能不仅会被要求在最后的报告中写进方案本身目标的达到情况，还要求表明当时方案参与者的成绩同其他学校或学区的人的比较情况。在上述这些情况下，评价者需要选择、实施和报告能够提供这些常模数据的成绩测验的结果。

由于终结性评价者提出的报告可能会影响有关方案未来的重大决策，因此他必须确保他编制的或者购买的测量工具具有高信度。换言之，他必须采用的是评价听取人认为是合适的和准确的测量工具。

形成性评价者的未来工作

恰恰相反，形成性评价者所以要测量成绩的理由是不太正式的。一般说来，形成性评价者成绩测量的主要职责有两点，一是在方案的全过程中进行进展情况核查；二是保证学生们的确是在学习方案要他们学习的内容，并且保持着预期的进度。这种信息的基本听取人就是方案的工作人员与设计人员。由于他们与方案关系密切，所以通常他们并不一定要测量工具技术优点的论证——尽管他们当然也希望能够相信测量工具提供给他们的信息。

形成性评价者另一个任务可能是开展简短实验以试验方案的组成部分，或解决设计者们关于采用哪些方法实施方案效果最佳的争论。在这些情况下，形成性评价者要选择或编制测验以仔细测量那些组成部分所特有的成绩目标。

由于对形成性评价者收集数据的条件不太苛刻，所以他们在挑选成绩测量工具时要比终结性评价者来得灵活。例如，对于终结性评价来说，被列入课程材料中的单元测验如果作为测量方案成绩的工具，人们通常对它们持怀疑态度，因为那些测验范围可能过于狭窄，只集中于被使用的特定方案材料之上。终结性信息的需要者通常总是想知道方案是否已经教授了其他非方案组可能会成功运用的技能。但是，对于形成性评价者而言，单元测验则是收

集数据的理想途径，因为那些数据至少告诉工作人员学生是否正在学习方案材料所包括的内容。一般说来，形成性评价者可以采用她和直接听取人都认为能够提供可靠证据的任何办法来估价成绩。当然，假使形成性评价者发现自己提出的测量工具是为了解决工作人员之间关于实施方案的若干种备择方法的争论，那么就必须重新物色技术上站得住脚的测量手段。

如果终结性评价牵涉到方案的未来，那么形成性评价者也许还得增加一项目标：找出或者编写具有广泛可靠性的成绩测量工具，而其同时却又能颇为忠实地描述该方案的效果。为此，形成性评价者可以收集单元测验题、教师编制的测验题以及从有关课程中得到的测量工具，在方案制定期间对上述这些工具进行试验，以找出对该特定方案目标的成绩具有敏感性的测验。

形成性评价者可能想收集一些鼓励方案工作人员和设计人员展开讨论的数据。在完成方案目标过程中监视学生的进展情况可以促使工作人员就学生接触方案期间，方案能够完成哪些内容一事作出比较符合情理的估计。如果对未列入方案已述目标中的技能的成就进行测量，形成性评价者便能指出方案在哪些地方正在使学生学到意料之外的知识，或者在哪些地方方案正在把人们的视线从可能正在丢失成绩的另一个地方转移开去。形成性评价者对于方案的最大贡献最后可能是这样的：他发现了方案取得成功必不可少的重要技能——方案过去没有予以重视的技能；或者他辨认出了的确是由方案产生的，但过去却未被方案辨认为目标的一些能力。

不管在阅读此书时想的是形成性评价还是终结性评价，或者是为了增长专业知识，《如何测量成绩》将帮助评价者去熟悉围绕设计和应用成绩测量工具以评价方案的主要的论点和任务。如果想更加深入探索成绩测量，可在多数章节的末尾参阅参考文献。

测验的类型

如果有人不知道标准参照测验与常模参照测验之间的区别，在评价领域内就迈不了几步。事实上，适用于开展评价工作的成绩测验至少有两类。为了知道为什么会产生不同类型的测验，我们有必要知道一些历史背景。

常模参照测验和标准参照测验是如何形成的

最早获得大规模使用，并且有系统发展的测验是常模参照测验。用这种测验、将人分类，例如分成若干智商组，和陆军军官预备学校的候选人。为了讨论起见，其评分办法（评分是相对的）被视为它们的基本特点。既然测验的目的在于人与人之间的比较，所以仅仅知道新兵伯恩斯做对32道题是没什么用的。于是，测验便规范化了，也就是说，把这些测验给几大群人去做，他们的得分分布情况用图表示。测验的编制者知道了标准组的评分情况后便能够把"32题正确"转换成一个标准的、比较的分数。例如，假设标准组里有75%的人分数低于32，那么伯恩斯的百分位数就定为百分之七十五。对于选择人们和把人们分类，这个分数是更加有用的信息。

常模参照测验(NRT)的编制方法是由其用途确定下来的，即需要把人们中的差别区分开来。通常，编制常模参照测验的过程过去是(现在仍是)与下面的论述差不多：

1. 选择人类活动的一个广泛而又普通的领域作为测验的中心题目。一般说来，这种中心题目的领域或者"建构"在测验的大标题内反映出来：学术成就、学习能力、现代语言的精通、手工技巧等。
2. 测验的编制者为了保证测验把建构的主要表现包罗无遗，或者为了保证测验反映出获得普遍应用的课程，他们调查了学

科的内容，并且获得了能够被编成试题的一连串认知上的或感情上的行为。人们经常采用一种内容/过程矩阵来组织调查一切可能获得应用的项目，并且用此矩阵把人们的注意力集中于应该分配给该试卷中各个子范围多少项目（由此可获多少权）。表1列示了内容/过程矩阵的一个实例，它被用于设计一份西班牙语的单元测验。图表告诉我们该测验所包含的各种题目以及每种题目获得的相对强调程度。一名当地的测验编制者希望他的测验能够反映出某特定教师或特定方案所强调的目标；而商品化的NRT的编制者却试图采用能够反映通用的各种课本和各门课程的内容或过程矩阵，从而达到内容范围上的协调。

3. 对于每一种被要求放进设计中的项目，都要记下大量备择题。
4. 在一个适当的受试者小组中试验那些项目，并且要作项目统计分析。最后的试卷形式中包含的各个项目要尽可能产生对受试者进行优劣分类的分数。通常采用项目分析统计法来舍弃那些在分类上降低测验有效性的项目。每个项目计算所得的难度指数（注4）使测验编制者得以把那些过多的人都通过或未通过的项目删去，而保留那些接近半数受试者通过的项目。常模参照测验的编制者希望他们的测量工具具有的另一特性就是良好的项目区分度（注5）。项目区分度表示对于某

注4：如果所有的受试者把该项目全做错了，项目难度指数即为0；如果所有的受试者该项目全做对了，项目难度指数便为1。反之，项目难度指数从.01到.99不等，它反映了试验小组对于该项目的成功率。

注5：的确存在着若干种项目区分度指数。但是它们都能够得到类似的解释。跟相关系数一样，一个项目区分度指数所取的值也可以从－1起然后到0最后一直到＋1。事实上，指数的含义与相关性的含义很相象，因为某一个项目负的区分度指数指出该项目的结果与整个试卷的情况成负相关。具体地说，那些在整个测验中情况良好的人在那个项目中却往往做错了；而那些在整个测验中情况不妙的人却往往做对了这个项目。可以想象出，具有负的区分度指数的某个项目对于出试题的人来说是件令人头痛的事，因为这样的项目显然是在测量与整个试卷不同的东西。等于0的区分度指数同样不是件好事情，因为这意味着那些测验不及格的人在那个项目中的情况与得高分的那些人完全一样。在多数情况下，一个项目必须具有正的区分度指数，这样，该项目才会对于必须把受试者区分开来的测验有用。

个已知项目，该测验的高分受试者比低分受试者好的程度。为了确定多项选择题的弊端，测验编制者可以检查一下，看看大部分或全体受试者所排除的是哪些不正确的备择答案项目。

随着六十年代个别教学的崛起，课程设计者发现他们需要测验学生的进步状况。教育者为了让学生在某个教学方案期间取得进展，便想正确了解在各个关键时刻学生懂得了什么，以及学生是否懂得了足够的东西因而能够继续学习下去。

不少的人——特别是格拉泽(注6)——开始明白，常模参照测验的编制技术不适宜这个目的。其理由之一是，设计这种测验并非用来检查已经懂得多少学科内容；编制这种测验的意图是让它们来决定一个学生的一般学业成绩与其他学生的一般学业成绩的优劣比较。常模参照测验首先是以标准组作参照的，其次才以学科范围为参照。

由于这种情况，常模参照测验又产生了两个具体问题：

(1) 具有很强区分能力的项目的存在（尽管这一点被视为常模参照测验的精华），可能使有的学生倒了霉，那些学生在课程上学得很好，可以进一步学下去，但是在识破常模参照测验的某些项目的花招方面却还学得不够。事实上，某些教育测量专家认为，常模参照测验的以区分度指数为依据的选题程序把人们本来希望单元测验会包含的那种直接了当的项目都给摒弃了。

(2) 设计常模参照测验时采用的结构不太严格的内容/过程矩阵难以确切讲出一个成绩良好的人实际上懂得了多少。这个人是什么都懂一点，但是他通过（例如说是西班牙语吧）熟练测验一事是否就意味着他能够正确讲出该语言的最常用不规则动词的变化形式了呢？常模参照测验的分数不宜用来结论性地确定学生是否已经掌握了具体的技能。

注6，R·格拉泽美国心理学家，著《教学技术和学习结果的测量》一书。

六十年代后期，测验的设计者开始编制目的在于帮助教育者发现学生在特定学科范围内已经掌握的内容的测验。由于那种测验是被用于确定学生是否已经达到了教学目标，所以人们称它们为标准参照测验。学生如果达到或者超过预先规定的足够知识或技能的标准，他们便通过了测验。

的确，好长时间以来，上课教师一直非正式地同时对标准参照测验和常模参照测验进行解释。例如，一名教师可能出了20道拼写测验项目。如果决定班里最好的10%学生可以得A，那么所采用的就是一种常模参照解释。但如果决定凡拼写错误少于3个单词的学生应得A，那么这种解释就是以表现标准为依据的了。近十五年里关于标准参照测验(CRT)运动值得注意的一点是，为了设计并且更加有效地大规模应用标准参照测验，以及为了(诸如全州范围内的竞争性考试)胜过个别课堂教学，编制这种测验的技术获得了发展。

人们要求制定某些特定技能范围内的熟练标准，于是编制测验的新的途径便应运而生。如果测验的分数是为了具体表示学生已经学到的东西，那么测验发表者就需要一些比内容/过程矩阵远为详细的设计测验的方法。这样就使测量专家提出了知识领域参照测验的概念。知识领域参照测验是一种以一个指定的办法，从一套用以测量某个教学目的的，规定得很清楚的项目中进行抽样而汇集成的测验。采用这样的测验的一个优点是：它提供了一种清晰的手段来概括一个受试者所懂的东西。如果阿丽莎在一次知识领域参照测验里四个项目中做对了三个，我们便有充分理由认为，对于在同样的行为领域内的所有项目，也就是说，如果那个测验里有着如知识领域说明所介绍那样的全部项目，阿丽莎可能也会在每四个项目中大约做对三个。

一份知识领域说明规定了作为测验备择对象的正确特性。这个规定包括以下几个内容：

- 对于受试者的指导和围绕实施测验的其他条件
- 项目组成部分可以变动的程度，其细节内容一直到可容许的

表 1　西班牙语测验的内容/过程矩阵

内容 \ 过程	听力		阅读			总计
	翻译成口头英语	用西班牙语回答	翻译成书面西班牙语	用书面西班牙语回答	翻译成书面英语	
数量表示	1 *B-1*	2 *B-4, B-5	2 *A-1, A-2		1 *A-22	6
一天的时间	1 *B-2	2 *B-6, B-7	2 *A-3, A-4		1 *A-23	6
日期	1 *B-3	2 *B-8, B-9	2 *A-5, A-6		1 *A-24	6
por/para 的用法			2 *A-7, A-8			2
ser/estar 的用法		2 *B-10, B-11	3 *A-9, A-10, A-11	2 *A-16 A-17		7
数和人称 gustar		1 *B-12	1 *A-12	1 *A-18	1 *A-25	4
带有hacer tener, poner的习惯用语			3 *A-13, A-14, A-15	3 *A-19, A-20, A-21	3 *A-26, *A-27, *A-28	9
总　计	3	9	15	6	7	40

• 表内各格表示针对测验测量的每个内容范围的项目数量与项目编号

词汇、格式、和多项选择题的迷惑特点等。

• 目标评分程序