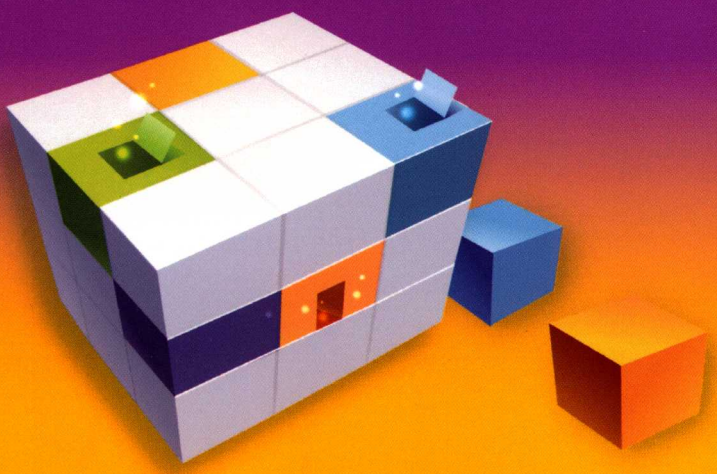
MEASUREMENT AND
ASSESSMENT IN EDUCATION

教育测量与评估

（原书第二版）

〔美〕Cecil R. Reynolds, Ronald B. Livingston, Victor Willson 著

霍 黎 霍 舟 译



科学出版社

教育测量与评估

（原书第二版）

〔美〕Cecil R. Reynolds
Texas A&M University

〔美〕Ronald B. Livingston
University of Texas at Tyler

〔美〕Victor Willson
Texas A&M University

霍　黎　霍　舟　译

科学出版社
北　京

图字:01-2015-2064

内 容 简 介

本书的主要内容包括:(1)在对学生以专业的方式进行评估的时候,教师所应该掌握的必要知识与技能;(2)教育评估研究的意义。本书介绍了在评估中经常采用的必要的基本数学概念和知识;扩展了传统的教育评估的内容,介绍了近几年使用比较广泛的表现性评估和成长记录袋评估;完整介绍了对残疾学生评估所必须进行的评估调整;还讨论了教育评估中的最佳实践。

本书主要是针对授课教师和准备从事教学工作的相关专业的大学生或教学管理人员编写的,从事教育工作的其他人员以及学生家长,在本书中也能找到对他们有所裨益的内容。

图书在版编目(CIP)数据

教育测量与评估:第2版/(美)雷诺兹(Reynilds, C. R.)等著;霍黎,霍舟译.
—北京:科学出版社,2015

书名原文:Measurement and Assessment in Education

ISBN 978-7-03-045991-6

Ⅰ. ①教… Ⅱ. ①雷… ②霍… ③霍… Ⅲ. ①教育评估-研究
Ⅳ. ①G40-058. 1

中国版本图书馆 CIP 数据核字(2015)第 245365 号

责任编辑:李静科 / 责任校对:郭瑞芝
责任印制:徐晓晨/ 封面设计:耕者工作室

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码:100717
http://www. sciencep. com
北京教图印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2015 年 11 月第 一 版 开本:B5(720×1000)
2015 年 11 月第一次印刷 印张:34
字数:658 000
定价:158. 00 元
(如有印装质量问题,我社负责调换〈印科〉)

献给 Julia，她为我和我的工作做出了许多牺牲。

——Cecil

献给我的密友 Kyle——你给我的生活带来极大欢乐！

——Ron

献给在测量和统计专业毕业了 34 年的学生。

——Vic

译 者 序

从我上学开始，就对为什么 60 分是及格线感兴趣。在满分为 100 分的考试中为什么 60 分就算及格？为什么只有各门课程的成绩都在 60 分以上，才能拿到毕业证书。直到我读大学的时候，很多课程还是 60 分及格。当然，有一些课程成绩采用了等级评定，有些任意选修课也采用了“通过”和“不通过”的成绩记录方法，可当时老师仍然告诉我们，通过也是过了 60 分才行。

60 分这个及格线使我对学习成绩的评定产生了浓厚的兴趣，尤其是在进入大学工作以后，发现多数老师仍然采用 60 分作为评定学生学习成绩合格的标准。那么，到底为什么是 60 分及格？有没有其他的成绩评定方法？对于类似的问题，使我越发感觉到了评价学生学习成绩这个问题的重要性，从而激发了自己想要对这一领域进行更加深入学习与研究的决心。

在寻找、查阅和学习相关资料的过程中发现，在这个学科领域，国内和国外的研究视角存在相当大的差异。关于国外在这方面近一段时期的研究成果，尤其是系统介绍这方面研究成果的中文资料并不多，也就是在这个时候，恰巧得到了由 Cecil R. Reynolds，Ronald B. Livingston 和 Victor Willson 合著的 *Measurement and Assessment in Education*，于是就有了将其翻译成中文的想法。

20 世纪以来，教育科学研究发展迅速，形成了一个庞大的教育科学体系。在这个教育科学体系中，有许多已经形成研究对象相对明确、研究内容相对独立、研究成果相对完整的学科分支。在当今世界许多发达国家，教育基本理论研究、教育测量与评价研究以及教育发展理论研究已成为现代教育科学研究的三大领域。

教育测量与评价是教育科学体系中极其重要的分支之一。这是因为，教育测量与评价理论不仅在教育教学及教学管理等实际工作中具有重要的应用价值，而且在社会各个领域的人才选拔与评价过程中也有广泛而重要的应用。所以，了解国外有关教育测量与评价的学科地位和作用，对我国教育测量与评价的建设和发展是非常有必要的。

Measurement and Assessment in Education 一书在美国教育测量与评估领域广受同行赞誉。作者之一的 Cecil R. Reynolds 博士是美国心理学教授，由于在心理测验和评估方面的出色工作而成为心理学界最著名的教授之一。他的研究获得过多项奖项，出版过多部著作和教材。我们希望这本书的翻译出版，能成为国内教育科学的重要参考书，能为我国教育测量与评估工作带来一些观念和方法上的启

示,为我国的教育科学的教学和研究提供一些借鉴。

历时两年多时间,终于将这本书翻译完了。为了准确起见,在翻译过程中,我们查阅了大量资料。但即便如此,有些术语可能还是第一次翻译成中文,经过仔细斟酌和向专家请教,才用了我们认为贴切的术语,并在后面加上了原来的英文术语。

本书的前十一章由霍黎翻译,后六章和附录由霍舟翻译,最后的整理由霍黎完成。

由于我们知识水平有限,虽然已很尽力,但翻译过程中难免还会有一些误解及不准确,敬请读者批评指正。

霍 黎 霍 舟
西安财经学院 西安科技大学
2015 年 4 月

原书第二版前言

我们很高兴有机会准备本书的第二版！这个版本保持第一版的组织结构，但也有一些实质性的改变。这一修订版的一个主要焦点是更新和增加联邦法律的覆盖范围，以及它们如何影响教育评估。这样做的目的是想强调这些法律如何影响教师的日常教学。我们的指导原则是为教师和读者服务——保留他们喜欢的，增加他们需要的。

致谢

我们要感谢编辑 Arms Burvikovs，正是他的领导帮助我们完成了这个版本。在获得许多高品质的评论和指导我们如何最好地实施建议上，他的工作对我们的写作有巨大的帮助。我们感谢以下审稿人：要塞军事学院（The Citadel）的 Nick Elksnin，东新墨西哥大学（Eastern New Mexico University）的 Kathy Peca 和布法罗大学堪利斯学院（University at Buffalo，Canisius College）的 Dan Smith。

对于我们家人给予我们工作上的热情支持、鼓励以及为保障我们的工作所付出的努力，我们心存感激。我们希望，本书能够对那些在课堂上辛勤工作的教师提供帮助，希望能帮助他们对学生进行更好的评价，促进更好的教学。

原书前言

当初次遇到某人时，都不可避免地要做出某种形式的评价。滑稽、有风度、聪明、诙谐、傲慢还是粗鲁，所有这些都是在初次遇到某人时能够联想到的。对班级中的学生也是这样，大学教授同其他教师一样，每年都会遇到一些新生，并且从相互交流中对他们形成一些印象。这些印象来源于对观察到的特征的评价或评估，或者来源于与这些新生的交流。我们每个人都在做这件事，但所做的都是非正式的。有时当我们对某人有了更多的了解后会发现，先前的评价是错误的。有的时候，评价必须是更正式和更精确的。这本书就是为了解决这些问题，并且能够使评估更精确，更有意义。

例如，我们必须给学生评定成绩来决定学生是否适合升级。心理学家需要通过某些适当形式的心理学诊断来获得如智力迟钝、学习障碍、精神分裂症、抑郁症、焦虑症等症状的准确判断。这些评价最好是通过比非正式的互动更严格的一些正式的测量方法来更好地完成。就像木匠可以估计木板的长度一样，我们也可以估计学生的特点——但对最后木匠的施工或我们的决策，这两种估计都不是令人满意的。它们都必须通过测量才能决定。

教育与心理测验正是解决这些问题的测量工具。这些问题包括：掌握一个科目的熟练程度、教育目标的实现、学生参加测验时显示的焦虑程度，甚至学生在课堂上的注意力。有些测验比其他测验更正式，并且测量技术的正式程度是在一个连续体上变化的：从教师自编的对某一特定任务进行的典型测验到用于标准化环境的、准备充分的、具有全国性代表参考样本的商业化标准测验。

本书的目的是给读者介绍不同的方法，用这些方法可以测量学校内感兴趣的结构，以及如何确保在设计自己的课堂评估时尽可能做到最好。我们还提供了学校中其他专业人员使用各种评估方法的详细资料，如学校心理学家，所以读者可以与这些专业人士更加智能化地进行交流，并且使用用于学校的很多评估结果，将与学生有关的工作做得更好。

本书不仅详细介绍课堂评估过程，同时也介绍各种不同的标准化测验。普通的或一般的课堂教育是重点，同时也注意并阐述对残疾学生的评价和测量过程。只要有可能，我们总是试图将这些原则应用到学校的日常问题中。虽然以应用在课堂上的问题为重点，通过集成的过程来描述和解释测验和测量的原则，但我们也希望能为读者准备好在学校中所面对的不断变化的评估和评价。基本原理可能变化不大，但在学校中的实际应用肯定会有变化。

本书的编写主要针对准备从事教学工作的教师或类似的学校管理人员。在教育环境中从事教育工作的其他人也可以找到翔实的内容,我们时刻希望这本书是实用的。在编写这本书的时候,我们反复问自己两个问题。第一个问题是,教师真的需要知道如何履行自己的本职工作吗?我们认为,大多数教师并不渴望成为评估专家,因此试图把重点放在基本知识和技能上,避免难懂的知识。第二个问题是,实证研究能告诉我们教育评估与测量是什么吗?有时,更容易跟随教育潮流和流行趋势,而忽视多年的研究。虽然这可能是诱人的,但却是不可接受的!对我们的读者,欠缺的是基于现有科学知识的最准确的有用信息。在你的职业生涯中评估的许多学生也欠缺这些信息。

作者已经开发了两个不可或缺的补充材料用以增加本书的效果。一个是对学生复习和掌握本书材料特别有用的、由 Victor Willson 博士主讲的 PowerPoint™ 课件;另一个是对教师有用的试题库。

目　　录

译者序
原书第二版前言
原书前言
第 1 章　教育评估简介 …… 1
1.1　评估语言 …… 2
1. 测验、测量和评估 …… 2
2. 测验类型 …… 4
3. 分数解释的类型 …… 8
1.2　教育评估的假设 …… 9
1. 心理与教育结构是存在的 …… 9
2. 心理与教育结构是可以测量的 …… 9
3. 虽然可以测量结构，但测量并不完美 …… 10
4. 存在不同的方法来测量任何给定的结构 …… 10
5. 所有评估方法都有其自身的优势和弱点 …… 10
6. 信息的多种来源应该是评估过程的组成部分 …… 10
7. 测验中的表现可以推广到非测验行为 …… 11
8. 评估可以提供信息用来帮助教育工作者制定更好的教育决策 …… 11
9. 可以用公平的方式进行评估 …… 11
10. 测验和评估可以使教育机构乃至整个社会受益 …… 12
1.3　评估过程中的参与者 …… 13
1. 开发测验的人 …… 13
2. 使用测验的人 …… 14
3. 参加测验的人 …… 14
4. 评估过程中的其他参与人员 …… 15
1.4　教育评估与相关法律 …… 15
1. 不让一个孩子掉队法案(NCLB,2001) …… 15
2. 残疾人教育改进法案 2004(IDEA,2004) …… 16
3. 1973 年《康复法案》的第 504 条款(504 条款) …… 17
4. 保护学生权利法案(PPRA) …… 18
5. 家庭教育权利和隐私权法案(FERPA) …… 19

1.5 教育评估的常见应用 …… 19
1. 学生评价 …… 19
2. 教学决策 …… 20
3. 选拔、安置和分类决策 …… 20
4. 政策决策 …… 21
5. 咨询和指导决策 …… 21
1.6 关于评估,教师需要了解什么 …… 21
1. 教师应该有能力选择适合做教学决策的、专业开发的评估方法 …… 22
2. 教师应该有能力开发适合做教学决策的评估方法 …… 22
3. 教师应该有能力管理、评阅和解释专业开发的和自己编制的评估方法 …… 22
4. 在做教育决策时,教师应该有能力使用评估结果 …… 23
5. 教师应该有能力开发包含评估信息的有效的评分方法 …… 23
6. 教师应该有能力交流评估结果 …… 23
7. 教师应该有能力识别不道德、非法和其他不恰当使用评估的方法或信息 …… 23
1.7 21 世纪的教育评估 …… 24
1. 计算机自适应测验(CAT)和其他技术进步 …… 24
2. "真实的"或复杂的表现性评估 …… 25
3. 教育问责和高风险测验 …… 26
4. 对残疾学生评估的趋势 …… 27
1.8 总结 …… 28
1.9 关键术语和概念 …… 30
1.10 推荐阅读 …… 31
1.11 感兴趣的互联网网站 …… 32
第 2 章 测量中的数学基础 …… 33
2.1 数学在评估中的作用 …… 33
2.2 测量量表 …… 34
1. 什么是测量? …… 34
2. 称名量表 …… 34
3. 顺序量表 …… 35
4. 等距量表 …… 35
5. 比率量表 …… 36
2.3 测验成绩的描述 …… 39
1. 分布 …… 39
2. 集中趋势测量 …… 42
3. 变异性测量 …… 46

2.4　相关系数 …… 49
1. 散点图 …… 50
2. 相关和预测 …… 52
3. 相关系数的类型 …… 52
4. 相关性与因果性 …… 54
2.5　总结 …… 55
2.6　关键术语和概念 …… 56
2.7　推荐读物 …… 57
2.8　感兴趣的互联网网站 …… 57
2.9　练习题 …… 58
第 3 章　测验得分的意义 …… 60
3.1　常模参照和标准参照得分的解释 …… 61
1. 常模参照解释 …… 62
2. 用于常模参照解释的派生分数 …… 67
3. 标准参照解释 …… 76
3.2　常模参照，参照标准，或两者的结合 …… 80
3.3　得分的定性描述 …… 82
3.4　总结 …… 82
3.5　关键术语和概念 …… 84
3.6　推荐读物 …… 85
3.7　感兴趣的网站 …… 85
3.8　练习题 …… 86
第 4 章　教师的信度 …… 87
4.1　测量误差 …… 88
1. 测量误差的来源 …… 90
4.2　估计信度的方法 …… 92
1. 重测信度 …… 93
2. 复本信度 …… 94
3. 内部一致性信度 …… 95
4. 评分者之间信度 …… 98
5. 总评成绩的信度 …… 99
6. 选择信度系数 …… 100
7. 评价信度系数 …… 102
8. 如何提高信度 …… 104
9. 估计信度的特殊问题 …… 105
4.3　测量的标准误 …… 107

1. 评价测量的标准误 …… 108
4.4 信度:教师的实践策略 …… 110
4.5 总结 …… 113
4.6 关键术语和概念 …… 114
4.7 推荐读物 …… 115
4.8 练习题 …… 115
第 5 章 教师的效度 …… 117
5.1 效度威胁 …… 118
5.2 信度和效度 …… 119
5.3 “效度类型”与“效度证据类型” …… 120
5.4 效度证据类型 …… 122
1. 基于测验内容的证据 …… 122
2. 基于与其他变量之间关系的效度证据 …… 125
3. 基于内部结构的证据 …… 132
4. 基于反应过程的证据 …… 133
5. 基于测验后果的证据 …… 133
6. 整合效度证据 …… 134
5.5 效度:教师的实践策略 …… 135
5.6 总结 …… 137
5.7 关键术语和概念 …… 138
5.8 推荐读物 …… 139
第 6 章 教师的试题分析 …… 141
6.1 试题难度指标(或试题难度水平) …… 142
1. 特殊评估情况和试题难度 …… 144
6.2 试题区分度 …… 145
1. 区分度指标 …… 145
2. 试题-整体测验相关系数 …… 148
3. 掌握测验的试题区分度 …… 149
4. 速度测验的试题分析 …… 150
6.3 干扰项分析 …… 150
1. 干扰项如何影响试题难度和区分度 …… 152
6.4 试题分析:教师的实践策略 …… 153
6.5 使用试题分析来改善试题 …… 154
6.6 表现性评估的试题分析 …… 157
6.7 定性试题分析 …… 158
6.8 使用试题分析改进课堂教学 …… 160

6.9　总结 …… 160
6.10　关键术语和概念 …… 161
6.11　推荐读物 …… 162
第 7 章　开发课堂测验的基本步骤 …… 163
7.1　教育目标的特点 …… 164
1. 范围 …… 164
7.2　教育目标的分类 …… 165
1. 认知领域 …… 166
2. 情感领域 …… 168
3. 动作技能领域 …… 169
7.3　行为与非行为教育目标 …… 169
7.4　编写教育目标 …… 170
7.5　开发测验提纲(或测验蓝图) …… 172
7.6　按照测验提纲来开发测验 …… 173
1. 常模参照和标准参照得分的解释 …… 169
7.7　在全州范围内开发课堂测验 …… 174
1. 选择使用哪种类型的试题 …… 174
2. 装配评估 …… 178
7.8　让学生为评估做准备和管理评估 …… 180
7.9　总结 …… 183
7.10　关键术语和概念 …… 184
7.11　推荐读物 …… 185
第 8 章　选择类试题的开发和使用 …… 186
8.1　选择题 …… 187
1. 开发选择题的准则 …… 188
2. 选择题的优势 …… 198
3. 选择题的弱点 …… 201
8.2　判断题 …… 202
1. 开发判断题的准则 …… 203
2. 判断题的优势 …… 205
3. 判断题的弱点 …… 205
8.3　匹配题 …… 206
1. 开发匹配题的准则 …… 207
2. 匹配题的优势 …… 209
3. 匹配题的弱点 …… 209
8.4　总结 …… 210

8.5 关键术语和概念 …… 211
8.6 推荐读物 …… 212
第 9 章 构造类试题的开发和使用 …… 213
9.1 口试:作为构造类试题先驱的口头论述 …… 214
9.2 论述题 …… 215
1. 论述题测验的目的 …… 215
2. 不同复杂程度的论述题 …… 216
3. 限制型论述题与扩展型论述题 …… 218
4. 开发论述题的准则 …… 219
5. 论述题的优势 …… 220
6. 论述题的弱点 …… 221
7. 评分论述题的准则 …… 223
9.3 简答题 …… 226
1. 开发简答题的准则 …… 228
2. 简答题的优势 …… 229
3. 简答题的弱点 …… 230
9.4 最后注意:构造类试题与选择类试题 …… 231
9.5 总结 …… 231
9.6 关键术语和概念 …… 232
9.7 推荐读物 …… 233
第 10 章 表现性评估和成长记录袋 …… 234
10.1 什么是表现性评估? …… 235
10.2 开发有效表现性评估的准则 …… 240
1. 选择合适的表现性任务 …… 240
2. 开发测验说明 …… 243
3. 开发评分答案的办法 …… 244
4. 减少评分误差的实施步骤 …… 248
5. 表现性评估的优势 …… 253
6. 表现性评估的弱点 …… 254
10.3 成长记录袋 …… 256
1. 开发成长记录袋评估的准则 …… 256
2. 成长记录袋评估的优势 …… 258
3. 成长记录袋评估的弱点 …… 258
10.4 总结 …… 259
10.5 关键术语和概念 …… 262

10.6　推荐读物 …… 263
10.7　感兴趣的网站 …… 263
第 11 章　基于课堂评估来评定成绩 …… 264
11.1　反馈与评价 …… 265
1. 正式和非正式评价 …… 267
2. 在终结性评价中使用形成性评价 …… 268
11.2　报告学生的进步:使用什么符号 …… 269
11.3　评定成绩的基础 …… 271
11.4　参考框架 …… 272
1. 常模参照评分(相对评分) …… 272
2. 标准参照评分(绝对评分) …… 274
3. 成就与改善或努力的关系 …… 275
4. 成就与能力的关系 …… 275
5. 建议 …… 276
11.5　将各类得分合并成总评成绩 …… 276
11.6　告知学生评分系统和获得的成绩 …… 281
11.7　家长会 …… 283
11.8　总结 …… 283
11.9　关键术语和概念 …… 284
11.10　推荐读物 …… 285
第 12 章　高风险评估时代的标准化成就测验 …… 286
12.1　高风险评估时代 …… 288
12.2　集体成就测验 …… 290
1. 商业开发的集体成就测验 …… 291
2. 各州开发的成就测验 …… 296
3. 增值评估:一个教育问责的新方法 …… 302
4. 在学校中使用标准化成就测验的最佳实践 …… 303
12.3　个体成就测验 …… 308
12.4　选择成就测验套装 …… 311
12.5　总结 …… 312
12.6　关键术语和概念 …… 313
12.7　推荐读物 …… 313
第 13 章　在学校中使用资质测验 …… 314
13.1　智力测验的简要历史 …… 317
13.2　在学校中使用的资质和智力测验 …… 319

1. 资质-成就的差异 …… 321
13.3 特殊学习障碍的一个新的评估策略:干预反应(RTI) …… 323
13.4 主要的资质/智力测验 …… 324
1. 集体资质/智力测验 …… 324
2. 个体资质/智力测验 …… 330
3. 选择资质/智力测验 …… 335
4. 理解智力评估报告 …… 336
13.5 大学入学考试 …… 350
13.6 总结 …… 351
13.7 关键术语和概念 …… 352
13.8 推荐读物 …… 353
第 14 章 行为和人格评估 …… 354
14.1 评估行为和人格 …… 355
1. 反应定势 …… 356
2. 在学校中的行为和人格评估 …… 358
14.2 行为评定量表 …… 359
1. 儿童行为评估系统-第二版——教师和家长评定量表(TRS 和 PRS) …… 360
2. Conners 评定量表-修订版(CRS-R) …… 365
3. 儿童行为检核表和教师报告表(CBCL 和 TRF) …… 366
14.3 自陈测量 …… 367
1. 儿童行为评估系统-第二版——人格自陈(SRP) …… 368
2. 青少年自陈量表(YSR) …… 372
14.4 投射技术 …… 372
1. 投射画 …… 374
2. 完成语句测验 …… 375
3. 统觉测验 …… 375
4. 墨渍技术 …… 376
14.5 总结 …… 377
14.6 关键术语和概念 …… 378
14.7 推荐读物 …… 379
第 15 章 评估调整 …… 380
15.1 影响残疾学生评估的重大立法 …… 381
15.2 残疾人教育法案(IDEA) …… 382