

G740-61/6

东西方教育丛书

简明国际教育百科全书

教育测量与评价

许建钺 赵世诚 杜智敏 郑晓齐 编译



教育科学出版社



21250924

图书在版编目(CIP)数据

教育测量与评价:(瑞典)胡森(Husen, T.),
(德)特尔威斯特(Thwaite, N. P.)主编;许建铎等编译.
-北京:教育科学出版社,1992. 5(1997 重印)
(简明国际教育百科全书)
ISBN 7-5041-0731-X

I. 教… I. ①胡… ②特… ③许… III. ①教育-测验
②教育评估 IV. G449. 1

中国版本图书馆 CIP 数据核字(96)第 07144 号

TORSTEN HUSEN T. NEVILLE POSTLETHWAITE ed.
THE
INTERNATIONAL ENCYCLOPEDIA
OF
EDUCATION
Research and Studies

1985

责任编辑 许佩云

教育科学出版社出版

(北京·北太平庄·北三环中路 46 号)

新华书店北京发行所发行

河北省唐山市胶印厂印装

开本:850 毫米×1168 毫米 1/32 印张:14.75 字数:495,000

1992 年 5 月第 1 版 1997 年 8 月第 3 次印刷

印数:14 001—19 000 册

ISBN 7-5041-0731-X/G · 693 定价:18.00 元

《东西方教育丛书》总序

在人类文明史上，东西方文化教育的交流源远流长。早在古希腊时期，雅典的学校模式就曾影响过小亚细亚和阿拉伯，而埃及、印度、中国等古老文明发祥地也将其灿烂的文化传到了西方。近现代的史册上，记载着四海五洲教育思想传播交流相互作用的无数事实。一方面，西学东渐，西欧、北美的教育制度和科学技术广泛地影响了东方世界；另一方面，亚、非、拉大批新兴国家的崛起，历史悠久的东方文化重新焕发出令全球瞩目的光彩。在当代，东西方在文化教育领域的交流日益频繁，已成为推动教育革新和社会发展的重要因素。在相互依存中求发展，已成为势不可挡的潮流，任何人为的壁垒也无法阻遏。

历史的经验清楚地昭示我们：一切文明的成果，都是人类共同的财富。巍巍高山，不辞土石而成其高；茫茫大洋，不拒洪流而成其大。只有在继承和发扬本民族优秀文化的同时，汲取世界其他民族的进步文化，才能建设社会主义的物质文明和精神文明。过去在对待外国文化上走过的曲折道路又告诉我们：盲目排外固然要落伍吃亏，全盘照搬也会贻误后代。只有深深植根于本民族的土壤中，才能使嫁接的枝叶长出新芽蓓蕾。任何简单的移植都不会成功。无论是批判还是借鉴，都须以客观的了解和

科学的研究为前提条件。

正是在东西方交流的广阔背景上，在发展和改革本国教育的严重任务面前，比较教育作为一个一个研究领域获得了它的意义和生命力。

《东西方教育丛书》实质上是一套比较教育丛书，它以“面向世界、面向未来、面向现代化”为指针，有重点地介绍东西方较有影响的教育著作，旨在促进东西方教育学术交流，推动我国教育科学和教育事业的发展。编者希望它能成为观察世界教育动态的一个窗口，成为联接东西方教育思想的一座桥梁，成为使我们的教育发展建立在世界文明成果基础之上的一块基石。希望它能在神州大地建设以马克思主义为指导、立足本国、面向世界的社会主义精神文明事业中，发挥一分积极作用。

《丛书》选题范围广泛。它将介绍国际上优秀教育著作的同时，将着重介绍西方有代表性、有较高学术价值的教育论著，努力译介长期以来未受到应有重视的发展中国的教育研究成果。我们尤以殷切企盼的心情期待中国学者推出从国际比较角度探讨中国教育现实问题和发展战略的力作，使我“物华天宝，人杰地灵”的中华民族能在教育研究领域作出无愧于时代的贡献。

《丛书》译、著并重。编者希望在我国比较教育研究日渐繁荣的基础之上能有更多的著作问世，而逐步减少译作的比重。

《丛书》以学术质量为其生命，编者将本“选材严、出书精”的原则，注重科学性和理论的现实性，为不断提高学术水平而努力。

真理的探讨和学术的发展有赖于自由的思想交流。编者将身体力行，努力贯彻“百家争鸣”的方针，不以主观好恶决定选题的取舍，不以门户之见决定著译者的选择。如能得海内外学者惠赐书稿，将是编者最大的乐事。愿姹紫嫣红的百花在丛书的园

地开放，愿各有所见的百家在丛书的论坛争鸣。

《丛书》里献给读者的每一本书都不是为了猎奇，而是为了探求教育过程的规律，也都不是为了罗列描述性的事实，而是为了从纷繁的社会现象中寻找教育发展的内在逻辑。它们的着眼点不是外国，而是中国教育改革和发展实践的今天和明天。也许，这些书籍不能为您提供解决复杂的教育问题的现成答案，但我们希望，它们能以提出问题和理论探讨的方式促进您对教育问题的思考，能为您提供有助于解决问题的多种方案。愿《丛书》的每一本都能成为您学习、研究、实践过程中有益的工具和长久的伴侣，而不会因时间的流逝便丧失其价值。

《东西方教育丛书》编委会

1987年12月

《东西方教育丛书》编委会名单

一、顾问（按姓氏笔划为序）

刘承沛 苏 蓟 杜殿坤 汪永铨

金世柏 顾明远 曹青阳 滕大春

二、主编

周南照 杨典求

三、编委（按姓氏笔划为序）

王英杰 曲 程 成有信 许佩云

许建钺 苏渭昌 杨典求 杨春发

张学忠 周南照 蔡克勇 蔡荣生

编译说明

瑞典斯德哥尔摩大学国际教育研究所荣誉教授托斯顿胡森 (Torsten Husen) 和德国汉堡大学比较教育系教授纳维尔波斯特尔斯威特 (T. Neville Postlethwaite) 主编的《国际教育百科全书》，是现代教育科学研究领域具有深远意义的一件大事。

《国际教育百科全书》堪称第一部真正国际性的、全面反映当代教育现状和最新研究成果的大型教育辞书。全书收词目 45000 余条，由 100 多个国家的 1300 多位学者撰写。全书有 1500 多条、字数在数千乃至上万以上的长条目概述了教育科学各主要学科领域迄今为止取得的进展。无论是要了解应用性强的比较教育和国际教育，还是想探索学理深奥的皮亚杰发生认识论，都可由全书循径而入，得窥全貌。

我国生动而丰富的教育发展及改革的实践正在召唤科学的理论，而理论的发展要求我们以开放的襟怀汲取一切先进的人类文化成果。《国际教育百科全书》无疑为我们提供了有高度学术价值和实际意义的综合性工具书。这是我们组织编译《简明国际教育百科全书》的主要缘由。

《国际教育百科全书》原著共 10 卷，6000 余页，500 万词，内容按 17 个教育专业领域分类编纂，按词头英文字母的顺序排列。由于目前各种条件的限制，我们难以全部移译这部卷帙浩繁的巨著，而只能根据我国教育事业和教育科学研究的急需，采取简译本形式，按教育领域和学科分类出版若干册，以飨读者。在

中文简编本的编译过程中，我们约请有关方面专家学者分别担任各分册的主编工作，以利编选、翻译工作的进行。

《简明国际教育百科全书》，不仅可以作为教育科学研究人员和各级各类学校教师的重要工具，而且是为教育行政人员、规划人员、技术人员和各级教育管理部门提供的必备参考书。现代教育已远远超出传统的“学校教育”范围，而广泛涉及整个社会生活和人的终生。因此，一切关心自身和国家教育问题的读者都可以从这部以科学研究为基础的辞书中获得有益的信息、知识和见解。

参加全书各分册翻译工作的百多位译者来自许多不同的单位，我们衷心地感谢他们鼎力相助和通诚合作，没有他们在各自领域的专门知识和语言功夫，本书的出版无疑是不可能的。

各分册在词目编排上参考了原书索引分类标题，按内容分成几大部分，列出分类词目表，同时在词条先后顺序上适当作些变动。译文中外国人名后面的年份，一般指其著作的出版年份；凡用一符号连接两个年份的，则指其生卒年份。限于篇幅，原书词条后的“参考文献”部分一概从略，有志于作深入研究的读者可查阅原著。

《教育测量与评价》分册有103个条目，约48万字，由许建铨，赵世诚，杜智敏，郑晓齐负责主持词目的编选，审校和最终统稿工作。编译者对本分册的原来目录作了一些变动；对部分条目的部分内容，在不影响全文理解的前提下作了少量删减；对个别难解、生疏的专业术语都进行了查证。在此基础上由统稿人写出注释，以期能帮助读者有更深入的理解。

俞志圭，浦薇，顾婷等同志参加了本分册编译的组织及事务工作。

编译者

1991年1月

总 目

《东西方教育丛书》总序	(1—3)
《东西方教育丛书》编委名单	(4)
编译说明	(5—6)
目录	(7—11)
正文	(1—450)
英文条目索引	(451—454)

目 录

教育评价的一般理论 (General Theory on Educational Evaluation)

对手式评价 (Adversary Evaluation)	1
决策定向评价 (Decision-oriented Evaluation)	3
诊断性评定法 (Diagnostic Assessment Procedures)	9
评价、评定和测量 (Evaluation, Assessment and Measurement) ...	14
效用评价 (Evaluation for Utilization)	15
成人教育评价 (Evaluation in Adult Education)	18
发展中国的成人教育评价 (Evaluation in Adult Education : Deve- loping Countries)	23
评价模式和它的开发 (Evaluation Models : Development)	28
评价标准 (Evaluation Standards)	37
评价的影响 (Evaluation Studies, Impact of)	41
评价人员的作用 (Evaluator, Role of)	45
形成性与总结性评价 (Formative and Summative Evaluation)	48
目的游离评价 (Goal-free Evaluation)	51
启发性评价 (Illuminative Evaluation)	53
实质性评价 (Intrinsic Evaluation)	60
司法式评价 (Judicial Evaluation)	62
元评价 (Meta-evaluation)	65
自然主义评价 (Naturalistic Evaluation)	67
成人教育中的需要评估 (Needs Assessment in Adult Education)	71
需要评估研究 (Needs Assessment Study)	76
常模参照评估 (Norm-referenced Assessment)	80

计划评价 (Program Evaluation)	84
应答式评价 (Responsive Evaluation)	93
交互式评价 (Transactional Evaluation)	95

教育评价的量表与统计方法

(Scales and Statistical Methodology of Educational Evaluation)

相关的弱化 (Attenuation)	99
考虑自信的评分 (Confidence Marking)	100
德尔斐法 (Delphi Technique)	103
领域分数 (Domain Scores)	104
推断理论 (Generalizability Theory)	105
格特曼量表 (Guttman Scales)	112
试题分析 (Item Analysis)	114
试题反应理论 (Item Response Theory)	127
潜在特性测量模型 (Latent Trait Measurement Models)	132
利克特量表 (Likert Scales)	139
教育测量的充分统计量 (Measurement, Educational : Sufficient Statistics)	141
拉希测量模型 (Rasch Measurement Models)	143
等级量表分析 (Rating Scale Analysis)	150
测验的标准、等价和标度 (Test : Calibrating, Equating, and Scaling)	160
瑟斯顿量表 (Thurstone Scales)	161
真实分数模型 (True-Score Models)	167
个体测验反应模式的确认 (Validation of Individual Test Respon- se Patterns)	170

测验的类型

(Types of Tests and Examinations)

成就测验 (Achievement tests)	176
适应测验 (Adaptive Testing)	182
态度及其测量 (Attitudes and Their Measurement)	188
自动评分系统 (Automated Scoring Systems)	198

封闭式测验和保密性 (Closed Tests and Security)	203
填空测验 (Cloze Tests)	204
考前辅导 (Coaching for Tests and Examinations)	208
认知偏好测验 (Cognitive Preference Tests)	210
认知风格 (Cognitive Style)	211
计算机管理测验 (Computer-managed Testing)	214
文章的计算机评分 (Computer Scoring of Essays)	215
标准参照测量 (Criterion-referenced Measurement)	218
标准参照测验 (Criterion-referenced Tests)	226
成套鉴别测验和多项混合测验 (Differential Test Batteries and Omnibus Tests)	234
领域参照测验 (Domain-referenced Tests)	236
外部考试 (Examinations, External)	239
语言能力测验 (Foreign Language Performance Tests)	243
课堂教学中的外语测验 (Foreign Language Testing in the Classroom)	246
外语口语测验水平 (Foreign Language : Testing Speaking Proficiency)	249
在大学中取得成功的预测 (Higher Education : Prediction of Success)	251
招生面试 (Interviews for Selection)	255
试题编写技术 (Item Writing Techniques)	258
潜像显示测验 (Latent Image Testing)	265
学习能力测验 (Learning Ability Tests)	268
字体的清晰性 (Legibility)	272
智力测验的历史 (Mental Testing, History of)	273
非言语测验 (Nonverbal Tests)	283
客观测验 (Objective Tests)	287
口试 (Oral Examinations)	292
操作测验 (Performance Tests)	295
个性调查表 (Personality Inventories)	297
难度测验与速度测验 (Power and Speed Tests)	303
实践教学考试 (Practical Examinations)	304

教育领域中的个人能力剖析图 (Profiles, Educational)	312
投射测验技术 (Projective Testing Techniques)	319
可读性与阅读水平 (Readability and Reading Levels)	325
测验结果的可靠性 (信度) (Reliability of Test Results)	329
效度 (Validity)	338
自我报告 (Self-report)	349
填句测验 (Sentence Completion Tests)	351
标准化测验 (Standardized Tests)	352
测验的实施 (Test Administration)	359
测验偏向 (Test Bias)	365
书面模拟测验 (Testing by Written Simulation)	374
小组测验与个别测验 (Testing, Group and Individual)	378
测验的常模 (Test Norms)	379
参试的焦虑 (Test-taking Anxiety)	386
参试技巧 (Test wiseness)	389
语言推理测验 (Verbal Reasoning Tests)	392
词汇测量 (Vocabulary Measures)	395
职业能力倾向 (Vocational Aptitude)	399
职业兴趣测量 (Vocational Interest Measures)	404

评价机制及其它

(Evaluation Mechanism and Others)

职业证书的颁发 (Credentialing)	410
国际教育评价协会 (International Association for Educational Assessment, IAEA)	413
国际教育成就评价协会 (International Association for the Evalua- tion of Educational Achievement, IEA)	414
全国教育进展评议中心 (National Assessment of Educational Pro- gress, NAEP)	417
学校视导员和指导员 (School Inspectors and Advisers)	420
学校记录和成绩报告单 (School Records and Reports)	425
高等教育入学的选拔机制 (Selection Mechanisms for Entry to Higher Education)	430

中等教育入学的选拔机制 (Selection Mechanisms for Entry to Secondary Education)	436
教育质量标准的国家监控 (Standards, National : Monitoring)	441
测验信息的提供 (Tests, Sources of)	448

教育评价的一般理论

(General Theory on Educational Evaluation)

对手式评价 (Adversary Evaluation)

依照英美两国的法律，陪审团和法官在听取对立的各方证词后才能作出判决。争执者或他们的代理人（律师）自己去收集证据，他们选择、访问及准备证人以便确定证据的要点。其它形式的证据则由证人的证词提供。辩证程序一般由赞成者和反对者的律师所作的开始陈述和结束陈述所组成，证人的证词也是由这些律师的引发而用类似于叙述的方式表达出来的。证词的目的旨在证实案件的真实情况，说服法官和陪审团。对手之间广泛地互相出示作证书和文件，互相进行有力的诘问，其目的是使证人作出有损对方论点的承认和动摇法官或陪审团对对方主要证词的信赖。在辩论过程中，评价者可以独立地为他们的诉讼作准备和辩护，可以凭借他们的技能和专长进行反驳。

在法律范围内，人们可以根据明

确的法律标准得出结论，证据的确切性和可采纳性则由法官依照大家已接受的法规严格确定，这个法规也保护证人免受对方不公平的对待。

社会研究的范例规定了什么材料可以被认为是事实和这些材料如何采集和报告。科学研究的定量模型用实验或准实验的方法去测量所选变量之间的假设关系，并借助于统计技术以数字形式报告结果。定性研究模型基于个案研究、深入调查、现场观察和人类学等技术，以便从被调查者的语言中得出解释或“现象的了解”，而不受研究者原有期望的约束。

在复杂的社会情境下作决策要依靠最突出的有关变量和它们内在联系的综合信息。定量和定性材料相结合的评价能够最好地满足这种需要。一些评价模型，如斯塔弗尔比姆 (Stufflebeam) 的决策模型和普罗沃斯 (Provus) 的差异模型 (discrepancy model)，都可以协调这两种材料。由欧文斯 (Owens, 1973) 和沃尔夫 (Wolf, 1975) 提出的对手式方法也

允许这种综合的应用，并且由于采纳了有计划的对抗，这种方法相当于提供两种独立的评价。

对手式评价采用了“斗争理论”，靠相互对立的评价者通过出示更好的、也就是更有说服力的证据，或者用能导致优势的法律或辩论技术（或其它技术）去努力获胜。这个所谓的对手系统的主要优点是它的互相诘问过程（cross-examination）。人们可以相信这样的过程能使定量和定性的材料真实可靠。这种主张的基础是：人类智力是唯一能够感知关系复杂的社会变量和把它们转换为语言陈述的手段。这种语言就是任意找 12 个平常无知的人也能理解的，因为它是社会的真正剖视。

至少有两种观点或评价并存对减少“唯唯诺诺”综合症发挥了作用。这种综合症使专家在处理他所尊重的当事人意见时受到影响。人们假设评价者的辩驳本能足以揭露任何经验材料和逻辑上的假象。对手式评价固有的争论不休和针锋相对状态，并不符合沃尔夫的司法式评价模型（judicial evaluation model）的要求，但对它们进行权衡来消除偏见，从而使正反两种意见的距离更为明显。因此，这种方法抓住了决策范围的两端，而不太考虑其中间领域。适合于对手式评价的行为，在整个过程中都应当是有疑问的。这种过程倾向于劝说而不是判定，它能容忍材料处理的疏漏，

但不允许违法处理。

对手式评价的基础是人类动力学（human dynamics），它是（也许是）不能定量的，也不能检测。对一种情况的取舍取决于人的多方面的特性因素，如个性、偏见、权势、恐惧等等。把定量数据和定性材料转换成符合某些特定评价听取人需要的证词时，二者的相对份量要视情况而定。

对有关人员的利益至关重要的那些问题的处理，取决于对手们作为教育评价者正直的职业道德。即使过程本身会招致某些人的压力，他们也能公正处理。对手辩论过程中提供信息的可能要和它产生变化（特别重要的是对于产生正确的变化）的可能严格分开。这一情况促使非教育评价专家得以参与，并可以不根据证据所支持的理由作出结论。时间、人力、财力在提供证据上的限制，以及数据分析工作的限制都会严重地影响诉讼陈述的质量和裁决的性质。另一方面，有关基层人员评价的参与，以及在“开庭日”争辩过程的互相诘问中展示的才能所给予人们的概念，将使决策增加一个大众因素，这是传统的评价方法做不到的。

在对手式评价中提交出来的证据取决于证人是否愿意在公共场合陈述他们认为是私人的或敏感的事情，和那些公众无权了解的、或危及证人的事情。对手式评价过程有可能提高结果的合理性，使它非常符合现实世界。

通过互相诘问, 结果可以得到平衡和描述。这样一来, 任何对控制舆论的指控就可以消除。这种元评价 (metaevaluation, 指对评价的再评价——编者) 的形式能使对手式评价过程作为一个很有用的方法来鉴别和报告某些计划或准备采取的行动的优缺点。

评判员不作一般的辩论, 而集中注意于很可能需要确定事实的特定案件上。公众可能缺乏对“发现”事实真相和陈述不同观点作出区分所需要的经验。

对手模型对怎样选择和确定问题, 或选择和确定什么问题, 或如何建立和提出诉讼等都不提供任何指导。每一方都可申述自己的理由, 把最重要的事实提给对方。这个过程没有对时间分配或人员安排提出要求, 参与这个过程的人数以及耗费的时间或金钱都是无关紧要的。它只能是一种价值判断, 这种判断决定包含或不包含什么样的证据或哪些证人, 以及采取和放弃什么样的相互诘问。对手式评价的难点是争执, 而评判员或决策人可能不考虑(确切地讲是蔑视)那些曾告知他们的任何材料。对手们只是提出自己的诉讼, 他们无需写下结论, 也不需要统计方法计算其重要性。

据记载, 至少已经出现了九种对手式评价, 各种评价都采用不同的方法(沃思恩 Worthen 和欧文思, 1978)。评价可能会揭示出比最初的想

象更为复杂的问题。结果形成的方式不是确定的, 因为在社会科学领域中许多问题并无确切的标准。简单的肯定-否定模式会使许多合理的选择含糊不清。

S·克莱因 (Clyne)

(马陆亭 译)

决策定向评价 (Decision-oriented Evaluation)

决策定向评价(也可译为“面向决策的评价”——编者)是为了在几个备择的行动方向中作出选择而提供信息的过程。一个评价如果具备了以下条件, 那它就是决策定向的: (1) 是为决策服务的; (2) 有几个备择方案可供选择; (3) 为下一阶段的另一决策提供信息。决策定向评价指把评价的概念引入到广义的决策要求中去的一种观点, 它并非只是进行决策时所采用的特定方法与技术。

1. 决策定向的定义和概念

菲·德尔塔·卡帕国家评价研究委员会(The Phi Delta Kappa National Study Committee on Evaluation) (斯塔弗尔比姆 Stufflebeam等, 1971) 是首先提出从决策的角度来考虑评价这种意图的代表。他们将教育评价定义为“为判断各种备择的决策方案而进行的描述、获取