

丛书主编 廖平胜

新世纪考试科学丛书

现代测量理论在

考试

漆书青 著

中的应用

华中师范大学出版社

国家“十五”重点图书

新世纪考试科学丛书/廖平胜 主编

现代测量理论 在考试中的应用

漆书青 著

华中师范大学出版社

2003年·武汉

(鄂) 新登字 11 号

图书在版编目 (CIP) 数据

现代测量理论在考试中的应用/漆书青 著.

—武汉: 华中师范大学出版社, 2003.8

(新世纪考试科学丛书/廖平胜主编)

ISBN 7-5622-2773-X/G·1445

I. 现… II. 漆… III. 测量学—应用—考试

IV. G424.74

中国版本图书馆 CIP 数据核字 (2002) 第 005200 号

现代测量理论在考试中的应用

© 漆书青 著

华中师范大学出版社出版发行

(武昌桂子山 邮政编码: 430079 邮购电话: 027-67861321)

(网址: <http://press.ccnu.edu.cn> 电子信箱: hscbs@public.wh.hb.cn)

新华书店湖北发行所经销

华中科技大学印刷厂印刷

责任编辑: 严定友

封面设计: 新视点

责任校对: 张 钟

督 印: 姜勇华

开本: 850mm×1168mm 1/32

印张: 14.875 字数: 325 千字

版次: 2003 年 8 月第 1 版

2003 年 8 月第 1 次印刷

印数: 1-2000

定价: 30.00 元

本书如有印装质量问题, 可向承印厂调换。

总 序

——中国需要“考试学”

考试，特别是高考，起码在教育领域内，是评价极为矛盾的典型。早在20世纪60年代，毛主席重提阶级斗争，考试就沦为“培养精神贵族”、“复辟资本主义”的“万恶之源”。1966年，废除高考成了教育领域中开展“文化大革命”的突破口。^①1973年把张铁生当成“石头”抛出来“反回潮”，1975年拿周荣鑫开刀“反击右倾翻案风”，高考都是“突破口”。1977年，邓小平再次重新工作，恢复高考，又成了推翻“两个凡是”^②、打破“两个估计”^③的“突破口”，成了在十一届三中全会之前思想解放的先声。有趣的是，恢复高考20年后，1997年，一方面，原国家教委及各级考试机构纷纷举行纪念活动，《人民日报》、《中国教育报》开展纪念征文，《人民日报》用《伟大的转折》为题综述了征文活动，并用一位教育家的话作为结尾：“恢复高考，挽救

①《人民日报》1966.6.18社论。

②“凡是毛主席作出的决策，我们都坚决拥护，凡是毛主席的指示，我们都始终不渝地遵循。”

③“文化大革命前十七年教育战线是资产阶级专无产阶级的政，知识分子大多数世界观基本上是资产阶级的。”

了我们的民族和国家。”^①另一方面，把“应试教育”归罪于高考的言论也屡见不鲜，以至说高考导致“教无宁日”、“国无宁日”，更有甚者，《羊城晚报》采访1998年全国“两会”的记者报道：“激烈的‘素质派’提出取消高考。”^②

同对现实生活中考试的评价一脉相承，对中国古代科举考试的评价也是极其矛盾的，所不同的是，否定的意见居多。应当明确，上世纪初，废止科举，批判八股，是中国进步所绝对必要的，是完全正确的。但近来（包括现在）一些文章、论点，把科举考试说得一无是处；有的人抨击现代高考，联系到古代科举考试，似乎自古以来考试就是万恶之源，则使人感到缺乏根据。我接触过美、英、加等国考试机构的总裁、专家，他们到中国来介绍各自的考试，第一句话就是：“考试，是我们从中国学到的。”科举考试是在总结了荐举制的弊端，经过几百年的探索才逐步建立起来的。科举考试的建立，为选拔官员特别是文职人员，为缓和统治阶级内部、统治阶级与人民群众的矛盾，为传播儒家思想，为发展教育，为各民族的融和、祖国的统一，都发挥了重要作用。科举成为历史前进的障碍，是明末清初以后的事情，其原因主要是考试内容的陈腐及考试形式的僵化，而不在考试制度本身。恰在此时，欧洲传教士来到中国，看到科举考试制度对克服吏治腐败的巨大作用，盛赞不已。正是借鉴了中国科举考试制度的经验，16世纪首先在英国建立了文官考试制度，后逐步传遍西方。鸦片战争以后百余年来，中国古代的各项制度几乎淘汰殆尽，惟有考试，我们又从西方学回来，逐步建立了现代的教育考试制度，以及公务员、会计师、律师、医师等各项考试制度，还

①《人民日报》1997.11.21。

②《羊城晚报》1998.3.14。

有各种劳动技能鉴定制度。到2000年，废除科举已近百年，我们现在完全有条件有可能历史地、公正地评价科举考试。全面肯定当然是不对的，全面否定也肯定是不对的。厦门大学高教研究所所长刘海峰教授提出建立“科举学”，我认为完全是必要的、正确的、有远见的。

现在的世纪之交是一个重要的历史转变时期。我国将从自然经济、计划经济转向有中国特色的社会主义市场经济。在劳动力市场上，各行各业，无论对脑力劳动者还是体力劳动者，对他们的能力，都需要一个科学的测量或评价，以便于交流或交换，这就是考试。有如现代生活需要驾驶汽车。你具有驾驶汽车的能力吗？去考试。通过了，发给你驾驶证，你才有权驾车上路。我们正开始向知识经济转变，创新能力将成为生产力的首要因素。因此考试内容必须改革，必须注重考能力，特别是创新能力。考试内容改革的要求，再加上科学技术（尤其是计算机信息技术）的突飞猛进，考试形式、技术也必须改革。在知识经济时代，教育或学习将成为终身的事情。对一个人来说，考试在一生中将不仅不是小、中、大学的入学与毕业考试、校内考试，作为对学习成绩和所具有能力的评价，各种各样的考试将伴随一个人的终生。总之，未来的考试与今日的考试，将不可同日而语。

古代中国，在许多领域曾长期处于领先地位，包括考试。然而，明末清初，当欧洲从农业经济、自然经济向工业经济、市场经济转变的时候，中国朝廷却在那里妄自尊大，错过了历史的机遇，导致了后来在列强坚船利炮之下中华民族的百年屈辱。应当严肃指出，科举考试在这个时期起了阻碍历史前进的恶劣作用。现在，我们又面临这样一次历史的转变。中华民族再也不能错过这次历史机遇了。我们必须直面市场经济、知识经济的挑战。党中央制订了“科教兴国”的战略，江泽民为此奔走呼号。为了实

现历史的转变，教育必须改革，考试必须改革。要创造适应未来时代要求的考试吗？我们必须正确对待历史的考试，改革现在的考试。而要总结历史的经验，改革现在的考试，创造未来的考试，最关键的是必须有正确的指导思想，有科学的评价标准和方法，这就是关于考试的理论，或称为“考试学”。

中国受考试之苦是深重的，包括有考试但属于“不好的考试”和没有考试这两方面的苦难。而苦难的根源，部分的是因为没有或缺乏科学的考试理论。1963年提出克服“片面追求升学率”，1966年取消高考，1977年恢复高考不久再次提出克服“片面追求升学率”，前几年又上纲为“应试教育”。中小学生学习过重的学业负担是应当减轻的。问题是应如何减轻？嫁祸于高考，甚至取消高考，并不能解决问题。建国50年来，围绕高考，各种论调花样翻新，实践中却莫衷一是。直到今年1月，新任教育部长陈至立同志才把“在确保‘两基’的前提下，调整现有教育体系的结构，扩大高中阶段教育和高等教育的规模，拓宽人才成长渠道，减缓升学压力”，作为减轻中小学生学习过重负担的第一项根本措施。^①中国发明了考试，却没有建立起全面系统的考试理论。在科举考试一千三百年悠久的历史中，中国积累了丰富的经验，有完备严密的规则，这是我们祖先的荣耀。在这漫长的日子里，在朝廷这最高一级的水平上，围绕科举考试曾爆发过多次激烈的争论，宋、明、清各朝都曾多次废除科举考试，却又都因没有更好的办法，而无奈恢复。科举考试在古代中国几乎“无书无之”。但翻遍各种文献资料，却找不到关于考试的一本论著，更不必说全面系统的考试理论。没有理论足以统一人们的思想，没有理论指导可以逐步改革这种多次存废的全国性实践活动。欧洲传教士

^①《中国教育报》2000.1.13。

在明末清初看到的正是处于阻碍历史前进时期的科举考试。但是，西方引进了科举的“合理的内核”——考试制度，不久又创造了考试的理论，即教育测量学、教育统计学，以及认知心理学等。在这些理论的指导下，西方的考试不断改革；考试的丰富实践，又进一步推动了理论的发展。正当中国废除科举的时候，考试却在西方获得了长足发展，并一直处于世界领先地位。十一届三中全会以后，首先是这些理论恢复了在大学的地位，随之而来的是恢复高考入学的第一批学子们，用西方的理论，批判中国传统考试的不足。这种批判的直接结果就是始于1985年的高考标准化。可悲的是，一些“文化水平”很高的人，不屑于这些考试理论，却不能容忍考试的标准化，于是大发雷霆，说一些驴唇不对马嘴的外行话。这起最近的小插曲再一次说明，考试理论对于考试工作的重要性，以及缺乏理论在实践中的危险性。

我理解，考试既是对人们心理因素的测量，更是一种社会活动。我们研究考试，既要研究考试内部诸因素的关系，如命题、考试实施、评卷、分数制度、统计分析等，我认为，这属于考试认识论的范畴；同时，我们也要研究考试与其外部诸因素的关系，如考试与教育、教学的关系，与人事、劳动制度的关系，与社会、经济发展的关系等，我认为，这属于考试社会学的范畴。在中国长期的考试历史中，考试认识论的水平，决定了考试质量的高低，而考试社会学的是非，却决定着考试的存废。在西方的考试理论中，考试认识论的著作比较多，而考试社会学的论述很少见。我们应当在考试社会学方面作出自己的贡献。

华中师范大学廖平胜教授早就看到了考试中的这些问题，长期以来呕心沥血潜心研究，积劳成疾而不顾，其精神令人敬佩。是他首先提出“考试学”这个命题，并立有专著。在他的倡导下，集国内许多专家与考试工作者，共同编写《新世纪考试科学

丛书》。借此机会，我首先向丛书的各位作者致谢，向华中师范大学出版社致谢。同时我也知道，考试是十分复杂的问题，而考试理论的研究又很薄弱，因此，丛书必然存在许多缺陷。借此机会，我代表丛书各位作者，诚恳地欢迎各位读者给我们提出批评。在这里，我先道谢了！

杨学为

2002年元旦于北京

前 言

考试，是对社会个体掌握人类知识与发展相应能力的状态与水平进行测量的社会活动。个体所要掌握与发展的这类知识与能力，常常正是学校教学的传授对象。因此，考试天然地就是教学工作中的一个环节。由于古代掌握一定文化知识与相应能力是从事国家事务管理的一个必备条件，而近现代系统的科学文化知识与相应能力更是大规模商品经济发展的基础性决定条件，于是，考试也就成为现代人事管理的有力工具，并具有服务于现代经济的重要社会职能。人类已经迈进 21 世纪，现代考试必将在社会生活中发挥出更为重要的教育文化与经济政治作用。

测量就是要给所考察的对象在一定的数字系统或者说量尺（量表）上指定值。考试所要测量的对象正是人的内部的高级心理过程；这不但不同于物理对象，而且也不同于人的低级的心理—物理现象。因此，作为考试的测量就具有显著不同于物理测量和心理—物理测量的突出特点。考试是主试与应试间发生的主体间的社会交往活动，而非一般的人—物系统活动；考试所要考察的是知识、心智技能、抽象逻辑思维能力，以及创造想象、体验感悟能力，乃至非认知的情意品质与特性，它们跟外部可观察行为（含言语活动）变量间的关系，是间接的、非确定性的，宜采用非线性概率模型来描述。

作为考试活动规律性的理论表现形态之一的教育测量学，首先就应揭示考试中这种测量活动的本质特点。并在此基础上，对

考试所欲测量的人的内部心理特质结构和测量反应活动的动态过程作出深入分析。但测量的根本要求是将考察对象数量化，因此，在教育和心理测量学中，计量学理论当然应是重要的主干内容。20世纪初正式建立起来的教育和心理测量学，起初在很大程度上是依照着物理测量理论框架进行构建的，对高级心理活动测量的本质特点的认识尚属粗浅。所以，以真分数理论为代表的经典测量理论就不能不显出诸多局限性。随着现代教育学、心理学的发展，在现代统计学与计算机技术的支持下，更由于人们哲学观念的进一步革新转换，自20世纪中期以来，逐步产生和形成了现代测量理论。其中的主要代表是项目反应理论。

经典的真分数理论以人们的整体反应行为作为分析对象，它主要关心的是测验总分，遵循的是只管结果的机械的刺激—反应模式。这种心理计量学理论有数量的统计的探讨，但严重缺失活生生的心理学内容，特别是人的内部高级心理过程的内容。在计量模型应与现代认知心理学等实质心理理论相结合的新思潮的强劲推动下，以项目反应理论为代表的现代测量理论，就把计量分析的对象努力朝人的心理过程内部深化：它不再只简单关心测验总分，而更要考察作为总分基础的每个项目（试题）上的作答表现，考察这种作答表现跟被试心理特质水平和项目测量性能之间的并非线性的概率函数关系。而且，还正在进一步考察各测验项目上每一作答反应步骤以及解决问题的各心智加工成份，对最后的反应作答表现的制约影响作用。被试内部心理特质，正在被如实地承认为主动的、变化着的、可能是多维的、有内部结构和发展形成过程的高级精神性对象。

本书的目的正是要介绍和探讨现代教育测量理论在考试实践与研究中的应用。因此，首先就要对考试的测量特性作出应有的哲学、社会学、教育与心理学的分析。同时，还要对考试所欲测

察的心理特质的结构与测量反应过程作出较为深入的现代心理学分析。这些，就构成了本书第一章与第三章的主要内容，以及第十章的某些重要内容。但是，本书的主体部分却是心理计量学内容。第二章真分数理论与第四章概括化理论，主要都是针对测验总分作分析的计量理论。真分数理论在历史上有着重要地位，在现实实践中也仍在发挥基础作用；尤其是概括化理论，在优化测试条件控制测量误差上，更显现出极强的指导价值，人们还将其视为现代测量理论的一个分支，所以，本书仍要对它们予以认真概括评介。从第五章起，主要内容是论述项目反应理论的基本概念与方法，分析它超出真分数理论局限的优点，并着重对它在考试工作中的应用，包括估计被试水平、分析测验与项目质量、考察测验信息、控制测量误差、实现测验等值、避免测验偏差和正确实现测验目的要求等方面的技术与策略，进行多方面的探讨。

教育测量学是从技术的角度来研究考试的。在现代考试要尊重人的主体地位，要努力服务与促进人的全面发展，要成为现代教育与现代管理的有效工具的思想指导下，其考核标准就要允许多元化，施测策略就要突出灵活针对性，情境设置、资料传输处理更要力争信息化，从而不断推进技术的创新。经典测量理论只按总分排队，它的真分数也是依赖于测试项目组的，信度却建立在实践中很难做到的“平行测验”观念基础上，难于满足现代考试的要求。项目反应理论能根据作答反应资料估出不依赖于测试项目样本的被试特质水平值，可用特质分数来统率与解释真分数与观察分数，特别是提出与定义了经典理论中所没有的项目与测验信息函数概念。这样，项目反应理论就能实现难度有异的测验的“垂直”等值，建立起项目参数在统一量尺上取值的大型题库，做到有预控地来编制达到指定考核要求的试卷；尤其是，可以编制出按“因人施测”原则进行的计算机化自适应测验，从而

使测量理论真正和信息技术融合在一起，确实为现代考试技术的发展开辟出了新天地。因此，本书特设第六、七、八章来突出讨论这些内容。

当前，现代教育和心理测量理论正在迅猛发展，本书所涉及的内容，当然只属基本部分。就计量学理论来说，当前国内流行的主要仍是真分数理论。这跟国际测量学界的发展态势相比，自是显得滞后。更加上我国已经加入 WTO，教育、人事、职业等考试正面临新的机遇和挑战，加强现代教育和心理测量理论的应用与研究，特别显得紧迫。本书的撰写，正是为了能对此贡献一份力量，以协助促进现代测量理论在国内的普及与应用。

华中师范大学出版社高瞻远瞩，主持出版《新世纪考试科学丛书》。承蒙出版社和丛书组织者杨学为研究员，以及华中师范大学廖平胜教授热诚相邀，要我负责撰写《现代测量理论在考试中的应用》一书。虽实感力所不逮，但为促进考试事业发展计，仍只好奋勉遵行。本书大部分章节由我执笔。第六章则由江西师范大学计算机学院丁树良教授撰写。丁教授从 20 世纪 80 年代中期起就开始研究 IRT 的参数估计问题。本章概括了当前国际上这一方面的最新成果，以及他本人的研究心得，实为本书增色不少。第九章由江西师范大学教育科学学院青年教师董圣鸿撰写。测验偏差与项目功能差异问题，尚未引起国内测量学界足够关注，董老师已就此作过某些专门研究，本章内容也属新颖充实。全书由我策划、统稿，不当之处敬希读者赐教指正。写作过程中得到了华中师范大学出版社张红梅女士的热情支持与指导，也得到责任编辑严定友先生的鼎力襄助，对此特致谢忱！

漆书青

2003 年 3 月

目 录

总 序	(1)
前 言	(1)
第一章 考试与现代测量理论	(1)
第一节 考试的发展和性质	(2)
一、中国古代科举制前的考试	(2)
二、中国的科举考试	(5)
三、西方考试与测验的发展	(9)
四、考试的根本性质	(15)
五、考试的主要种类	(24)
第二节 测量和测量理论	(28)
一、测量活动及其科学定义	(28)
二、测量规则及其执行	(32)
三、不同对象领域的测量特点	(37)
四、心理测量的测量特点	(41)
五、心理和教育测量理论	(46)
第二章 真分数理论概述	(56)
第一节 真分数理论模型的基本概念	(56)

一、经典真分数理论的建立背景	(56)
二、经典真分数理论的假设与模型	(60)
三、真分数理论假设的几个推论	(63)
第二节 真分数理论的信度概念	(67)
一、信度概念的理论分析	(67)
二、信度系数的定义	(69)
三、信度系数的估计	(73)
四、信度系数的应用	(78)
第三节 传统的项目分析技术	(86)
一、项目分析的意义	(86)
二、项目难度	(89)
三、项目区分度	(95)
四、项目与测验的关系	(101)
第三章 常模参照测验与标准参照测验	
.....	(104)
第一节 常模参照测验	(105)
一、常模参照测验的定义	(105)
二、发展常模	(108)
三、组内常模	(112)
四、标准分数常模	(114)
五、常模参照测验的编制	(117)
第二节 标准参照测验	(121)
一、标准参照测验的定义	(121)
二、行为领域的界定和项目的编制	(124)
三、标准的建立和确定	(127)
四、项目与测验质量的统计分析	(131)

第三节	促进形成的教学考试	(135)
一、	教学中考试的形成性	(135)
二、	常模参照测验与标准参照测验的教学应用	(137)
三、	教学考试的综合多样化	(140)
第四章	概括化理论	(143)
第一节	概括化理论的基本概念	(143)
一、	概括化理论对经典理论的发展	(143)
二、	真方差与误差方差	(147)
三、	随机侧面与固定侧面	(149)
四、	交叉设计与嵌套设计	(151)
第二节	概括化理论分析的统计方法	(153)
一、	G 研究与 D 研究中的方差分析	(153)
二、	单侧面交叉设计	(156)
三、	双侧面交叉设计	(161)
第三节	概括化理论的应用	(165)
一、	交叉设计的应用实例	(165)
二、	嵌套设计的概括化理论分析	(171)
三、	标准参照解释分数时的概括化理论分析	(175)
四、	效度的概括化理论看法	(178)
第五章	项目反应理论的基本概念	(184)
第一节	项目反应模型	(184)
一、	经典真分数理论的局限	(184)
二、	项目特性曲线与项目特性函数	(187)
三、	项目反应模型	(191)

第二节	能力分数与被试水平估计	(196)
一、	能力分数与能力量表	(196)
二、	能力参数的条件估计	(202)
三、	联合最大似然估计	(208)
四、	能力的贝叶斯估计	(213)
第三节	项目与测验信息函数	(217)
一、	信息函数的概念	(217)
二、	项目反应理论的信度观	(223)
三、	项目反应理论的显著优点	(228)
第六章	参数估计	(233)
第一节	IRT 中参数估计概述	(233)
一、	参数估计及其困难	(233)
二、	参数估计的几种主要方案	(235)
三、	参数估计中的几个有关问题	(238)
第二节	条件似然估计、联合似然估计、 边际似然估计的若干讨论	(240)
一、	关于条件似然估计的讨论	(240)
二、	关于联合似然估计的讨论	(244)
三、	关于边际似然估计的讨论	(247)
第三节	通过 EM 算法实施 MMLE 方案	(249)
一、	MMLE/EM 方案的推导过程	(250)
二、	MMLE/EM 方案的实施步骤	(257)
第四节	贝叶斯估计方案	(259)
一、	采用贝叶斯估计方案的必要性	(259)
二、	后验期望估计和最大后验密度估计	(261)