教育测量新理论在医学考试中的应用报告

医学计算机化题库系统的开发与应用研究

医学计算机化题库系统的开发与应用研究课题组

第二军医大学出版社

医学计算机化

题库系统的开发与应用研究

Development and Application of CoMTIB

医学计算机化题库系统的开发与应用研究课题组

课题组顾问 张华华　应子良

课题组组长 朱　樑

课题组主要成员 梅长林　张红武　姚定康

赵铮民　李兆申　吴　正

吕一刚　雷新勇　徐丽萍

蔡瑞宝　徐晓璐　周　全

王占齐　刘昌堃　张颖秋

蒋小龙　蔡剑飞　姜　平

内 容 提 要

本书首先介绍了国内外医学计算机化题库及教育测量和评价的现状，然后详细介绍了医学计算机化题库系统(CoMTIB)研究的指导原则，题库系统的建立及组卷、考试、阅卷评分和考后分析的全过程，最后评述诊断性评价在医学教育评估中的应用，另外还对项目反应理论及概化理论在医学教育中的应用进行了介绍。

图书在版编目(CIP)数据

医学计算机化题库系统的开发与应用研究/朱樑，梅长林，张红武等编著. —上海：第二军医大学出版社，2006.6

ISBN 7-81060-531-3

Ⅰ.医… Ⅱ.①朱… ②梅… ③张… Ⅲ.计算机应用—医药学—考试—诊断性评价—测量理论—研究 Ⅳ.R-39

中国版本图书馆CIP数据核字(2006)第069316号

医学计算机化题库系统的开发与应用研究

主　　编　朱　樑　梅长林　张红武

责任编辑　王　楠

第二军医大学出版社出版发行

上海市翔殷路800号　邮政编码：200433

发行科电话/传真：021—65493093

全国各地新华书店经销

第二军医大学印刷厂印刷

开本：787×1092　1/16　印张：8　字数：90 千字

2006年 6 月第1版　　2006年 6 月第1次印刷

ISBN 7-81060-531-3/R.472

定价：20.00元

特别说明 —— 国家教育科学“十五”规划课题

课题组由下列单位的研究人员组成

（以姓氏笔划排序）

王占齐	第二军医大学教务处	博士
吕一刚	第二军医大学长征临床医学院	教授
朱　樑	第二军医大学长征临床医学院	教授
刘昌堃	上海教育考试院	教授
李兆申	第二军医大学长海临床医学院	教授
吴　正	复旦大学	教授
应子良	美国哥伦比亚大学	教授
张华华	美国德克萨斯大学	教授
张红武	第二军医大学长征临床医学院	副教授
张颖秋	第二军医大学长征临床医学院	讲师
周　全	第二军医大学长征临床医学院	讲师
赵铮民	第二军医大学长征临床医学院	教授
姚定康	第二军医大学长征临床医学院	副教授
徐丽萍	第二军医大学长征临床医学院	副教授
徐晓璐	第二军医大学长海临床医学院	教授
梅长林	第二军医大学长征临床医学院	教授
蒋小龙	第二军医大学教务处	教授
雷新勇	上海教育考试院	教授
蔡瑞宝	第二军医大学长征临床医学院	教授

目　　录

一、项目概况

1999年，国家实行国家执业医师、医师职称资格认定等考试，对医学院校的考试提出了新的要求。为了进一步加强对医学生临床思维能力和临床实践能力的培养，我校从2000年开始筹备，在2001年正式成立临床医学院，改原有传统的教学模式(4+1，即4年大学基础与临床课程学习，1年医院实习)为临床医学院新教学模式(3+2，即3年大学医学基础课程学习，2年医院临床学习和实习)。医学考试也随之有大的改革以与之相适应。课题组于2001年向上海市教委提出了《新型医学考试题库的建立与应用》的项目申请，并于同年获得上海市教委批准及投资1万元，第二军医大学及长征临床医学院也配套1万元支持本项目的实施。为深化研究，2002年立项为国家教育科学十五规划课题。

本项目已进行了整整三年，现已完成预定目标，且有多方面的深化与拓展，开发了医学计算机化题库系统(Computerized Medicine Test Item Bank，CoMTIB)，进行应用研究，并予以推广。

(一) 研究人员概况

本课题组成员承担的主要研究工作有：课题设计，题库软件编制，组织相关考试并对结果进行分析，进而优化题库，提出并实施课题深化与拓展方案等。

除课题组成员外，我们还组织了校内外30余名有临床医学教学经验的教授、讲师和管理干部参与试题编写、筛选、审定及有关的组织管理工作。组织了在长海临床医学院、长征临床医学院、南总临床医学院、济总临床医学院、海总临床医学院、北总临床医学院、利群临床医学院和南京军医学院及长宁区中心医院、黄浦区中心医院、浦东区中心医院、杨浦区中心医院、吴淞人民医院、东方医院、411医院和上海市第二人民医院等近20个单位学习、实习的两届本科毕业学员、四届实习学员(五年制、七年制)及研究生共1 000余名学生及同济大学、

复旦大学、上海交通大学医学院部分学生参与考试研究的实践。

为了使课题得到深化与拓展，尤其是国内外先进学术思想得以应用，课题组又邀请了上海教育考试院、复旦大学、美国哥伦比亚大学、美国德克萨斯大学等单位的有关专家学者加盟。他们参与了课题的部分研究工作。

（二）总体目标及要解决的问题

1. 总体目标

本课题总体目标是：根据国内外医学教育、教育测量学、教育心理学等领域的最新发展趋势，开发针对新世纪医学考试实际需要的新型题库及软件并进行实际应用，提出并解决与题库组成、组卷原则、评分标准、诊断性评价等有关的一系列理论应用问题。本题库主要用于水平测试，以医学本科生、研究生内科学学业测试为主，进而拓展到临床医学的各种考试，也适合于执业医师模拟考试用。

2. 要解决的问题

（1）确定符合医学教育要求，又适宜计算机组卷、分析以及计算机环境下考试的题库内容。还应解决已经编辑完成的资料与计算机程序之间的连接问题。

（2）对题库中的试题进行科学合理的分类及分层，为组卷提供依据。

（3）应用项目反应理论对试题进行参数估计与等值化，提高试题质量；并应用多元概化理论进行试卷分析，不断优化题库。

（4）多维度诊断性指标及诊断性评价在试卷及临床技能考试分析中的应用。

（三）研究过程与方法

第一阶段（2001.01～2001.08） 运用文献法、比较研究法等进行项目调查研究和文献检索；收集与编写试题，进行初步属性标定，同时完成题库软件程序编写。

第二阶段(2001.09～2002.08) 完成题库所需试题的编写和筛选工作(共7 000余题),并运用项目反应理论及统计分析、数据处理方法完成试题属性标定与试题参数等值化分析,进行尝试性考试,并对软件的实用性、安全性进行改进。

第三阶段(2002.09～2003.08) 组织多次实际考试,运用多元概化理论等方法对考试结果及试卷、临床技能评价指标体系进行定量化评价,进行必要的修正,并引入多维度诊断性评价系统,同时进行推广应用。

第四阶段(2003.09～2003.10) 综合运用实验法、调查法、评估法对课题进行过程中积累的全部资料进行汇总总结,完成本报告。

二、医学考试的国内外现状与立项依据

(一) 医学考试的需求

随着国家执业医师考试、医师职称资格认定考试等一系列新的医学考试的推出,迫切需要建立一种高质量、多功能、检测标准和考试形式都与国际先进水平接近的新型医学考试题库,使其既适合于高等医学院校临床医学专业理论课程考试,又能适用于各类医师资格考试,同时还可以为检测不同等级医学院校的教学质量提供公共平台。

在从应试教育转变为素质教育的过程中,对医学考试也提出了如何更有利于素质教育和能力培养的新要求。

2000年,我校进行教学改革。2001年正式将原有传统教学模式(4+1)改为新教学模式(3+2)。医学生在医院的2年内,先后在内科、外科及专科范围内的各临床科室轮转学习。在每个临床科室学习结束时进行小出科考试,测试学生掌握该学科理论及临床能力的水平;内、外、专三大科结束时的大出科考试则测试其各大科理论及临床思维能力和临床实践能力的水平。

新教学模式对医学考试的方式和内容提出了更高的要求，这是进行本课题研究最直接的推动力。

（二）国内外现状

目前，国内外已经建成了许多考试用智能题库系统，许多计算机公司也开发了基于 Web 的网络考试系统，特别是在计算机认证考试、外语考试方面，研究较为深入。在医学考试方面，目前国内研究不多，尚未见有系统、新颖的大型医学考试题库的报道。而在国外，计算机化的医学题库应用比较广泛，如美国的执业医师资格考试、执业护士资格考试等。但美国医学院入学考试仍是纸笔考试，未实现计算机化，更无自适应考试。

最近，在第六届全国医学考试与教育评价学术研讨会上，国家医学考试中心领导明确提出，近几年内要在全国逐步推广计算机化医学考试，今后医师执照考试将采用计算机题库上机考试方式。

医学考试题库历经了卡片式、档案式到图书馆式等多种形式。国内比较成熟的有国家教委委托中国医科大学完成的基于 DOS 系统的《临床医学综合考试国家题库计算机软件系统》和天津市卫生局建立的《实践技能计算机医学考试题库》，它们在实际应用中发挥了一定的作用，但由于计算机操作系统较旧，目前已经难以被推广应用。

目前，我国的医学考试仍然普遍采用笔纸考试形式，少数单位在住院医师考试中对于影像诊断学（B 超、X 线）等少数内容采用计算机考试。但是，迄今为止的所有医学考试并不重视测量理论的指导，题库建设基本以传统测量理论为基础，还没有见到用现代测量理论如项目反应理论（IRT）或概化理论（GT）来指导考试的报道。

美国医学考试中运用信息技术和心理测量理论是世界上最先进的。美国全

国医学考试委员会(NBME)十分重视心理测量理论与技术在执业医师资格考试(USLME)中的应用,除了对考试中获得的大量数据进行系统的处理与分析外,还注重用测量学理论指导有关考试工作。目前,以项目反应理论为代表的现代测量理论与技术已经开始应用于 USMLE 的统计分析、题库建设以及考试设计。项目反应理论的测量模型包括能力的估计、试题、人或题的特征(包括能力、难度、区分度、信度等)、模型拟合等要素。NBME 现在使用的 IRT 模型是单参数模型,即拉希模型(Rash)。

美国医师执照考试一般根据 IRT 理论中信息函数相等原则,事先根据考试委员会要求组织 50 套试题(从题库中选出,均为选择题,每套试卷的信息函数相同,题目均经过等值处理,保证在质量、内容覆盖面、难度以及答题速度等方面的一致性)存于计算机题库中,每位考生随机从中选择一套题目进行考试。因此,美国执业医师考试并没有采用计算机自适应考试(CAT),它采用的是 IRT 理论指导下的计算机化考试。它能通过较少的考题就对被试者的水平(θ)做出更加有效的测度,这是它较之传统纸笔测验的最大优势。在美国的教育测试中,CAT 已经得到了普遍应用,如美国研究生院入学测验(GRE)、美国商学院研究生招生测验(GMAT)、美国护士执照测验、美国军事职业性向测试(ASVAB)等。CAT 的发展走过弯路,也存在着一些问题。但是总体看来,CAT 仍然有着巨大的发展潜力和广阔的应用前景,因此也是研究的热点。目前国内尚无医学考试采用 CAT 的报道,本课题的研究也是实现 CAT 的前期工作。

根据文献检索,美国的医学考试分析还没有引入诊断性评价的概念,也尚无采用概化理论指导的报道。尽管美国于 2001 年颁布的名为"一个也不能掉队"的教育法案(No Child Left Behind Act of 2001)中提出,要尽快在教育评估标准化考试中加入诊断性评价。因此,为大规模的标准化考试做出诊断性评价,为考

生提供更全面更有价值的信息，成为当前考试研究的一个热点。

（三）原有的工作基础

长征临床医学院在2000年曾组织临床各科专家编著了一套主要以国家医师资格考试为目的的临床医学习题集，因此就有了大量经过多年实践应用的优秀试题，为进一步研究开发新型医学考试题库奠定了重要基础。同时，已经公开发表的中国医科大学的《临床医学综合考试国家题库计算机软件系统》和天津市卫生局的《实践技能计算机医学考试题库》，以及其他兄弟单位在实践中形成的各种医学考试试卷，也为本课题研究提供了研究参考资料。

长征临床医学院及本校其他临床医学院有一批长期从事医学考试的医学教育专家，他们有丰富的医学考试及评价的理论与实践经验。

美国哥伦比亚大学应志良、美国德克萨斯大学张华华是美国医学考试及相关心理测量研究领域的知名学者，他们在项目反应理论、概化理论等现代教育测量理论和应用方面有许多开创性的工作。上海教育考试院刘昌堃、雷新勇、复旦大学吴正等也都在教育测量理论与计算机化考试的领域里有长期工作经验和多种理论成果。他们的加盟加强了本课题的理论基础。

根据社会的需求、国内外现状和原有的工作基础，本课题提出以现代测量理论为理论基础，针对医学教育的特点，开发实用、科学、有效的新型医学考试题库系统，进而研究考试的诊断评价和教学反馈等问题，在理论水平上是先进的，在应用前景上是明确的，在工作基础上也是可靠的。

三、研究指导原则

（一）素质教育导向原则

考试的直接目的是为了检测教学效果，而终极目的是全面提高学生的素质。

因而，考试只是一种重要的教育手段，而全面提高学生的素质才是教育的根本目标。

临床能力，包括临床思维能力和临床实践能力，是医学专业技术人员专业能力的主要组成部分和专业技术水平的直接反映，也是其综合素质的根本体现。国内外医学教育都非常重视临床能力的培养，临床能力的评价方法也在不断得到改进。医学考试不仅应该能够具备评估考生的作用，更应当成为学生自我检测和提高的一个有力途径，因此，诊断性评价是素质教育宗旨下的医学考试体系应当具备的一个重要环节。

在本课题的研究中，我们始终十分重视题库及其应用对学生临床能力的评价作用。长征临床医学院所在的第二军医大学从 2001 年开始的新教学模式更强调临床能力的培养，这一点在我们题库系统的诊断性评价功能及应用中得到了充分体现。

（二）评价的科学性原则

以项目反应理论、概化理论等为代表的现代教育测量理论，对于促进考试和测评的科学化、建设大型题库和进行大规模测验，都有十分重要的指导作用。我们在医学考试题库的建设中根据项目反应理论的原理，对试题标定了难度、区分度、猜测度；并按临床医学专业要求确定系统疾病分类、知识结构分层和认知层次等一系列参数。因而能够据此对被测试者做出诊断性评价。然后，再运用概化理论对已经使用过的试卷和测试结果进行可信性和有效性分析，提出对题库的评价和改进、优化意见，因此，整个试题库的开发和应用建立在较强的科学性分析水平基础上。

（三）评价的诊断性原则

认知诊断是目前国际教育评价中十分推崇的理念之一。这一理念的核心是

通过对学生的测评，发现学生个体认知的优势方面和不足方面，并对不足方面提出改进和提高的建议。医学教育的目标是使医学生获得知识，培养分析和处理临床问题的能力。如何透过学生的分数，关注其背后所包含的实质，即认知、专业素质及临床能力的差别？我们根据临床医学教育的特点，创造性地提出多维度诊断性指标及诊断性评价系统，并成功地应用于题库软件，进而对考生的试卷进行多方面的分析、评价，并提出针对性建议，帮助教师深入了解学生掌握知识和能力的程度，也帮助学生自我认识，发扬优势，找出差距，从教、学双方面提出针对性的改进措施。

四、题库的建立

（一）建库原则

1. 学术要求：科学性、规范性、严肃性

题库建设是一项系统工程，在管理上应保证科学性、规范性和严肃性。在长征临床医学院有关领导的支持下，成立了题库建设指导委员会，负责题库建设的策划、论证、设计、督查和评估；又成立了专家组，负责题库建设的命题、审核、修订和增补。以教研室作为命题单元，负责题库的编写、初审等工作。这是题库建设的组织保证。

2. 内容要求：先进性、兼容性、实用性

题库内容以新版全国统编教材为准，充分考虑专业特点，从知识点着手，尽可能涵盖全部教学内容。在专业技术上应体现先进性、兼容性和实用性。

3. 题型要求：多样性、层次性、准确性

医学考试内容多，综合性强，要求题量大，覆盖面广，并且必须体现考生临床能力水平。所以，题型有客观题和主观题，试题往往文字叙述较多，并辅有图片、

动画等多媒体形式。在操作上应着重协调性、层次性和准确性。

（二）试题来源

本课题形成的题库以长征临床医学院经多年实践检验证明为较高质量的试卷题目为基础，参考兄弟单位和国外最新试题集，特别是美国国家医师资格考试中的 Step3（临床知识及能力考试）和 CCS 试题（计算机模拟病例分析），并结合本院的教学实际，增加大量新编试题。具体做法是：

1. 组织专家命题

为了本课题命题及分类需要，课题组编写了《内科学考试题库命题要求》（见附件一）和《内科学疾病的编号》（见附件二），对试题的属性进行了科学的界定。同时组织第二军医大学各临床医学院的内科教研室部分教员参加编写综合内科学考试试题。然后，请有关教员根据《内科学考试题库命题要求》，结合自己的教学经验及学生学习情况编辑试题。

2. 收集国内外新的试题

通过翻译国内外公开发表的题库著作、试题库等收集新的试题，并经校内专家审定，进行筛选改进，再按题库要求进行属性的确定。

3. 原有考试试题的再开发

课题组修改、审定了近 10 年我校各类考试及同济大学医学院考试的共 40 余份试卷的试题，同样按题库要求进行属性的确定。

（三）试题参数确定

1. 属性分类

为了便于参数的估计，我们根据科学性和可操作性原则，将试题属性分为以下几类：

（1）题型编号

题型区分为：A 型单选题、B 型单选题、X 型多选题、填充题、名词解释、简答题、论述题及病例分析等八大类。

(2) 系统疾病

对呼吸、循环、消化、泌尿、血液、内分泌、风湿等 7 个系统及感染性疾病、理化因素所致疾病共 102 种主要疾病均予以分别编号。

(3) 知识分层(知识结构)

将知识分层为：健康维持、疾病机制、诊断、治疗及医患关系等 5 项。

(4) 难度、区别度和猜测系数

难度、区别度和猜测系数是根据项目反应理论组卷要求的参数。

(5) 认知水平

认知水平包括记忆、理解、应用三个层次。

(6) 大纲要求

大纲要求分为掌握、熟悉、了解三级。

2. 难度、区分度和猜测系数的确定

以项目反应理论为基础建库，要求获得的参数与考试群体无关，因而其应用中最大的困难是试测群体要足够大，需经若干届学生考试才能获得可供分析的足量试题。

我们采用项目反应理论(IRT)的参数估计分析软件，对近 10 年我校各类考试共 40 份试卷的近 100 万条数据进行了参数估计(见图 1)。

IRT 是一种以试题参数为前提条件的理论，每道试题都有其特征曲线(即参数估计曲线，见图 2、图 3)，从此曲线可以得到该试题的难度、区别度及猜测系数的估计值，从而判断试题的质量。

IRT 参数估计具有以下优点：①试题参数确定更为准确；②能全面地解决测