IT技能测评自动化

——理论·技术·应用

许 骏 柳泉波 著

科学出版社

IT 技能测评自动化

——理论·技术·应用

许　骏　柳泉波 著

科 学 出 版 社

2001

内 容 简 介

本书是作者根据近年来从事技能测评自动化研究所取得的成果，以及国内外最新文献资料撰写的一本学术专著。全书包括：绪论，主要介绍CAA的定义、分类、研究现状和发展趋势，提出了一种新的分类方法并以此为基础重构CAA研究的内容体系；理论篇，提出技能测评的信息模型，全面、深入地讨论了系统建模、形式化表示和知识推理等内容，建构了技能测评自动化的理论框架；技术篇，主要讨论IT技能测评自动化的关键技术——信息获取技术；应用篇，介绍应用系统的设计原理与实现技术，代表性成果是IT技能训练导师系统iTutor和IT技能测评系统iTAS；最后介绍测评自动化研究的新进展，包括两个内容，即程序自动测评系统和CAA应用引发的新课题——基于CAA数据库的知识发现。书后附录提供了与CAA研究相关的在线资源和参考文献。

本书可供高等学校计算机、自动化等相关专业研究生和高年级本科生阅读，对信息技术教育研究和教育软件研究开发人员也具有指导意义和参考价值。

图书在版编目（CIP）数据

IT技能测评自动化——理论·技术·应用/许骏，柳泉波著. -北京：科学出版社，2001

ISBN 7-03-009829-3

Ⅰ. I… Ⅱ. ①许… ②柳… Ⅲ. 计算机辅助教学 Ⅳ. G434

中国版本图书馆CIP数据核字（2001）第070378号

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

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2001年10月第 一 版 开本：850×1168 1/32

2001年10月第一次印刷 印张：8 1/2

印数：1—3 000 字数：226 800

定价：25.00元

（如有印装质量问题，我社负责调换〈新欣〉）

序

广东省科技厅于日前致函邀请我主持“IT技能测评自动化理论与技术”科技成果鉴定会。这是由本书作者及其领导的课题组完成的一项科研成果，其中包括这本即将由科学出版社出版的专著。

随着计算机技术的飞速发展和社会信息化水平的不断提高，将计算机应用于教育测量与评价的全过程，即计算机辅助评价（Computer-Assisted Assessment，简称CAA）已是大势所趋。CAA国际学术会议至今已经开过五届了，在国外，CAA已成为计算机教育应用研究的热点之一，但国内有关CAA研究的成果尚不多见。

CAA的应用引发了评价内容、方法和形式的深刻变革。从测评的内容和目标看，CAA大体上可分为知识测评和技能测评两大类，它们共同构成能力素质测评的基础。但目前国内外CAA研究主要集中在知识维度的客观题测试，技能测评研究还相当薄弱，尤其缺乏一般性的理论与方法。技能包括操作技能和心智技能，通常采用非客观题——操作题或主观题——的形式进行测评。因此，技能测评和非客观题自动阅卷是CAA研究的重点和难点，从而形成了CAA研究的新领域——技能测评自动化，显然，这是一个多学科交叉的研究课题。

我与作者在1999年初就认识，当时他主持完成的一个科研课题在北京召开科技成果鉴定会，我是鉴定委员会主任。该成果由于在CAA研究开发方面的创新性和领先性，被专家评价为计算机教育应用的开拓性工作。鉴定会上，专家们建议将技能测评自动化作为后续研究的重点。时隔2年半，作者及其领导的课题组对技能测评自动化理论与技术进行了系统深入的研究，在理论

探索—技术攻关—软件开发三个层面上均取得重要进展。我为作者及其同事取得的新成果感到由衷的高兴。

作者以自己近年来的研究成果为基础，同时借鉴吸收国内外同类研究的最新成果，写成了这本专著，其成功之处在于建立了技能测评自动化的理论、技术和应用体系，并提出了 CAA 应用的新课题——基于 CAA 数据库的知识发现。理论性与应用性相结合是本书的特色，作者不仅提出了技能测评的一般性理论与方法，应用系统设计也有所创新，以 IT 技能训练导师系统iTutor为例，它突破了 ITS（Intelligent Tutoring System）的两个关键问题——交互测评和适应性决策，为多年来徘徊不前的 ITS 研究开发提供了一种新的技术方案和一个成功的范例。

值得一提的是，这本著作是国内该研究领域迄今为止公开出版的第一本专著，我对本书的出版表示祝贺。作者在一个新研究领域进行开拓性探索，这种努力是可贵的，当然，它的稚嫩也是难以避免的，这并不影响此工作的意义。相信这本著作能对国内 CAA 研究起到推动作用。

中国工程院院士

2001 年 9 月 29 日

前　言

CAA（Computer-Assisted Assessment，计算机辅助评价）是指将计算机应用于教学评价过程，它引发了评价内容、方法和形式的深刻变革。从测评内容和目标分类，CAA 大体上可分为知识测评和技能测评两大类，它们共同构成能力素质测评的基础。从国内外研究的现状看，CAA 研究主要集中在知识维度的客观题测试，相应的理论、方法和技术都比较成熟，自动阅卷问题也容易解决。相比之下，技能测评研究还相当薄弱。

技能包括操作技能和心智技能。技能测评通常采用非客观题的形式，因此，当前 CAA 研究的重点和难点在技能测评和非客观题自动阅卷。我们的研究工作主要集中在技能测评自动化以及 CAA 应用引发的新课题——基于 CAA 数据库的知识发现。

本课题研究的理论创新在于：

①提出 CAA 分类的新方法并以此为基础重新建构 CAA 研究的内容体系，开拓了 CAA 研究的新领域——技能测评自动化；

②提出技能测评的信息模型，建构了技能测评自动化的理论框架，包括系统建模、信息获取、形式化描述、知识推理以及综合评价等内容；

③对 CAA 应用引发的新课题——基于 CAA 数据库的知识发现进行了初步探索，提出了一种新的分类算法。

本课题研究在 IT 技能测评自动化的关键技术——信息获取技术取得突破，提出三种信息获取方法：

①基于文件解析器的信息获取方法；

②基于自动化技术的信息获取方法；

③过程信息获取方法。

在上述理论与技术成果的基础上，研究开发了两类教育软件——IT 技能训练导师系统 iTutor 和 IT 技能测评系统 iTAS。其中 iTutor 突破了 ITS（Intelligent Tutoring System）的两个瓶颈问题——交互测评和适应性决策，为多年来徘徊不前的 ITS 研究提供了一种新的技术方案和一个成功的范例；iTAS 解决了长期困扰计算机信息技术教育的技能考核问题，具有重要的实践意义。

主观题的最重要特征是题目的解答通常要通过语言表述来完成，因此，主观题的自动测评必然涉及到自然语言理解问题。限于目前的技术水平，完全解决主观题的自动测评问题是不现实的。程序设计（编程）能力主要属于心智技能测评的范畴，试题形式也主要采用主观题。但由于程序设计语言与一般的自然语言相比，具有严格得多的约束和限制。因此，我们选取程序自动测评作为主观题自动阅卷研究的突破点，目前已取得一些重要进展。

本书介绍作者及其领导的课题组近年来在该课题研究所取得的主要成果，写作过程中参阅了大量国内外最新文献资料。本书是迄今为止该研究领域第一本公开出版的专著，它不仅凝聚了课题组全体同志的集体智慧和多年来辛勤工作的成果，也汇集了所引用文献中国内外同行专家的真知灼见。我们期望本书的出版能为扩大 CAA 在国内的影响起到促进作用，更热切期待有更多的同志加入到 CAA 研究的行列，为建构 CAA 理论、方法、技术和应用新体系而共同努力。

由于技能测评自动化在理论、技术和实践上还远未达到成熟的阶段，我们所做的工作也只是初步的，本书具有明显的探索性和试验性，而且写作时间又比较仓促，因此，书中的疏漏和不当之处在所难免，殷切期待专家和读者批评指正。

本课题研究得到广东省科学技术厅、广东省高等教育厅、广东广播电视大学、深圳市教育局、中央广播电视大学、北京师范大学现代教育技术研究所和科学出版社等单位的关心、支持和帮

助。课题研究和本书的写作是在我们的导师何克抗教授的具体指导下完成的，区益善教授、胡晓峰教授等专家学者给予我们许多鼓励和指导。中国工程院院士、中国科学院计算技术研究所倪光南研究员主持了该课题的科技成果鉴定会并为本书作序。在此一并表示诚挚的谢意。

在本书即将出版之际，我们再次向所有关心、支持本课题研究的领导、专家和朋友们表示衷心的感谢！

许　骏　柳泉波

2001.8.18

目　录

应　用　篇

测评自动化研究新进展

绪 论

技能测评自动化是计算机辅助测评（Computer-Assisted Assessment，简写为 CAA）研究的一个新领域。本章首先对 CAA 的定义及其分类做简单的介绍，接着对 CAA 理论和实践两方面的研究现状和关键技术进行综述和讨论，试图勾画出目前 CAA 研究的脉络、重点及其发展趋势。在此基础上，我们提出 CAA 分类的新方法，并对本书的体系结构作了简要介绍。

一、CAA 的定义、分类及研究重点

随着社会信息化水平的不断提高，越来越多的教育机构使用计算机代替纸和笔实施评价。考生在计算机屏幕前操作键盘和鼠标答题，计算机自动阅卷并立即给出测评结果。事实上，计算机是教学评价的一种理想工具。为说明这一问题，我们先看评价的一些特点：

①评价是重复性工作；

②评价可以被精确定义；

③速度（提供快速反馈）是评价的重要指标；

④人通常不是好的评价者，因为人工评分容易产生错误，而且不同的阅卷人对错误（特别是复杂错误）的解释不一样，这必然导致评价标准的不一致。

显然，评价的这些特点也正是计算机的长处所在。实践表明，计算机参与评价过程，可提供快速、准确和一致的评价，并能自动对结果进行统计分析，大大促进了教育评价的量化研究。

为适应知识经济时代的需要，终身学习体系正在形成，学习型社会已初见端倪，远程教育学习的人数急剧增长。在传统的远程教育中，对学生进行评价是一件非常困难的事，要耗费大量的

人力物力，往往只能进行次数很少的总结性评价。计算机进入评价领域，使这种状况大为改观，例如通过网络题库和远程自动测评系统，学员可以及时得到关于自己学习的准确评价，获取有针对性的反馈信息，这对师生处于分离状态的远程教育至关重要。

1．CAA 的定义

CAA 是指在评价学习者的知识、技能和能力的过程中引入计算机作为工具或手段[1]，它引发了评价内容、方法和形式的深刻变革。含义大致与 CAA 相同的术语还有：计算机辅助评价（Computer-Aided Assessment）、计算机化评价（Computerised Assessment）、基于计算机的评价（Computer-Based Assessment，简写为 CBA）和基于计算机的测试（Computer-Based Testing）。（注意：本书后续内容将用“测评”一词统一翻译“testing”和“assessment”，包含了测试和评价两方面的意思）。在评价过程中采用计算机手段是指：

① 在测试的过程中，传递材料、评分以及对测试结果进行分析。

② 比较并分析通过试卷搜集到的数据。

③ 记录并分析学习者的成就水平，生成测评报告。

④ 通过网络比较、分析和传送测评信息。

根据英国在 1995 年和 1999 年进行的全国调查发现，几乎所有的学科都有应用 CAA 的案例。CAA 支持的评价类型主要包括下面几种：

① 诊断性。目的是判断学习者对某个主题的预备知识的掌握程度。

② 自测。学习者检查自己对某个概念或术语的理解程度。

③ 形成性。目的是提供反馈来指导学习者的学习，这些反馈表明了学习者对某个主题的知识、技能理解和掌握的程度。在 CAA 中，形成性评价通常采取客观题的形式，在测评的过程中或者结束后立即给出提示和反馈。

④ 总结性。总结性评价能够给出量化分数，并对学习者在某个领域的成绩做出判断。总结性评价通常是正式的、有组织且有监考的评价形式。

此外还要区分“标准（criteria）参照”的评价和“常模（norm）参照”的评价。标准参照的评价是指评价有明确的标准和目标。在这种类型的评价中，过关标准是预先定义好的，只要学习者达到标准即可通过。常模参照的评价是对一组学习者的成绩或者态度进行分级，采用的分级标准是通过控制每个级别的学习者占学习者总数的百分比获得的，这种评价具有明显的选拔性。

2. CAA 的分类

从现有的文献资料看，对 CAA 的分类比较混乱。究其原因，主要是大部分学者在对 CAA 分类时所采用的分类标准前后不一致所造成的。为了避免分类的混乱现象，我们建议采用如图 0.0.1 的 CAA 分类方法。

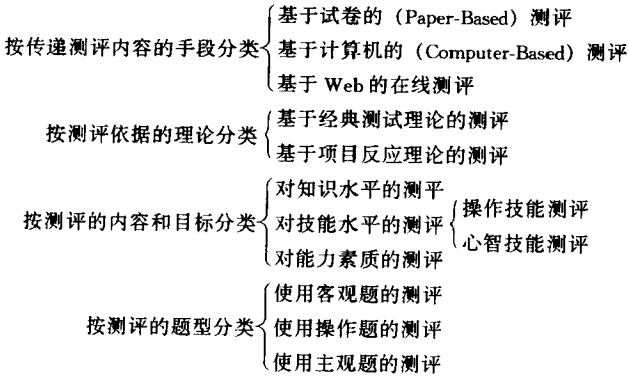

图 0.0.1　CAA 的分类

讨论 CAA 分类，主要是为了明确每一种类型的 CAA 能够做什么，要创建该类型的测评系统需要解决哪些问题，不同的 CAA 类型各有什么优点和不足之处等。对于上述这些问题，可以从教育的角度或技术的角度作出回答。我们认为，教育专家和技术专家对 CAA 的研究工作是相辅相成、互为补充的，共同构筑起 CAA 研究的大厦。虽然将 CAA 研究中教育层次的问题和技术层次的问题截然分开是不太可能的，但是作为计算机科学技术工作者，我们主要从技术的角度关注 CAA 研究的新课题及其解决方法。

（1）按传递测评内容的手段进行分类

第一种类型是基于试卷的测评，这是一种比较传统的测评形式。考生坐在教室里，监考者会给每个考生一份试卷、一张答题纸，并指导考生如何填写答题纸、涂写答题卡等。这种测评形式在我国已经取得了比较广泛的应用，例如高考试卷中客观题的评阅。考生的答题纸将被搜集起来，借助光标阅读器进行评分（Optical Mark Reader，简称 OMR）。OMR 软件对每份试卷的每个问题进行阅读和评分，并自动生成详细的统计分析报告，因此能够对学习者理解知识的程度进行分析。这种类型测评的主要特点是：一是只能处理客观性试题，应用范围受到很大的限制；二是应用的主要技术手段是 OMR，相关的硬件和软件都比较成熟。当然，这种类型的测评也可能有计算机的参与，例如用计算机对测评结果进行分析，并生成各种统计报表。

第二种类型是基于计算机的测评，试题可以通过软盘、硬盘或者局域网提供。考生坐在 PC 机旁，题目呈现在屏幕上，通过操作键盘和鼠标进行答题。考生可以在完成每个题目后立即得到反馈（主要适用于形成性评价），也可以在整个测评完成后得到最终的反馈（主要适用于总结性评价）。这种测评类型的一个主要优点是可以采用多媒体形式呈现试题，除了改善试题的表达，更重要的是有助于生成测评高级技能的题目。此外，由于测评过程是在计算机上进行的，通常具有自动阅卷的功能，因此测评结

束后可以马上得到结果。

第三种类型是基于 Web 的在线测评，这种类型的测评具有基于计算机的测评所具备的所有优点，而且允许考生从连接到 Internet 上的任何 PC 机上获得试题，这非常适合于网络学习。对于形成性评价，基于 Web 的在线测评允许学生在他们方便的任何时候任何地点参加测评，即学生可以不受时间和空间的约束，根据自身知识、能力水平和时间情况申请测评，并得到及时的评价和反馈，从而实现真正意义上的开放教育。对于总结性评价，基于 Web 的在线测评允许世界各地的考生在同一时间实施同样的测评，其显著特点是跨地域、范围广和规模大。此外，这种测评形式能够连结到 Internet 上所有可利用的资源，这对于形成性评价尤其具有重要的意义。在线测评的主要缺点是缺乏安全性，对被试的身份确认也是一件困难的事情。

(2) 按测评依据的理论模型进行分类

传统的测评都是基于经典测试理论，采用真分数模型[2]。所谓真分数，数学上定义为：测评中被试的真分数是观测分数(或者说测评实得分数)的期望值，表示为：$T = E(X)$。其中 T 代表真分数，X 代表被试在测评中的实际得分，E 代表数学上的期望。显然，在此意义上的真分数是无法得到的，因为定义中真分数 T 是被试在无数多次独立重复测评中获得的平均观测分数。从信息论的角度看，任何一组信息都含有无关的信息（噪音），测评的任务是获得被试的真实信息，排除无关信息，前者称作真分数，后者称为误差。经典测试理论由基本假设、信度、效度等概念组成，其方法体系主要包括项目分析和标准化。以经典测试理论为基础的综合测评，主要包括客观性测试和主观性测试两种形式。

另一种广泛应用的测试理论是项目反应理论（Item Response Theory，IRT)[3,4]。简单地说，项目反应理论以两个基本假设(一维性假设和局部独立性假设）为基础，将被试答对某一道题目的概率看作被试能力的函数，其图像称为题目特征曲线。项目

反应理论的方法体系包括参数估计（题目参数估计和能力参数估计）、信息估计和误差分析，这涉及到较多的统计学或数学方法问题。与经典测试理论相比，项目反应理论具有以下优点：

①题目参数估计更准确；

②根据信息函数可从题库中抽取符合被试能力的题目进行测试。

详细内容将在第三条中介绍。虽然从理论上讲，项目反应理论中涉及到的试题既可以是客观性的也可以是主观性的，但目前主要采用客观题的形式。建立在项目反应理论基础上的自适应测试是我们关注的重点。

（3）按测评内容和目标对 CAA 进行分类

知识、技能和能力是评价的三个主要方面，因此可以将 CAA 分为知识测评、技能测评和能力测评。现行教学评价的主要缺陷之一就是测评内容片面，即测评内容基本上只局限于知识维度，忽视了技能维度的测评。由于能力包含了对知识运用和技能的掌握。因此，知识测评和技能测评是能力测评的两个基石。

（4）按照测评的题型对 CAA 进行分类

客观题、操作题和主观题是测评中最常用到的三种题型。客观题和主观题的概念大家都很熟悉了，这里不做讨论。操作题是介于客观题和主观题之间的一种题型，主要考核被试的操作过程和方法，通常在真实或者模拟的环境中进行。人们一般将主要使用客观题型的测试称之为客观性测试，将主要使用主观题型的测试称之为主观性测试。注意它们与客观评价、主观评价的概念是有区别的。以测评被试提交的 Word 文档为例，文档的格式设置情况以及是否按要求插入图片或添加艺术字等都有客观的评判标准，因而属于客观评价。如果要对文档的外观效果进行评价，则通常采用专家打分的方式，评判标准往往具有主观随意性，因而属于主观评价。

在详细讨论各类 CAA 的特点之前，先对技能测评做简单介绍。

3. 技能测评

在《美国传统辞典》中，对“技能”（skill）是这样解释的：“Proficiency, facility, or dexterity that is acquired or developed through training or experience”，翻译过来，就是“通过训练或经验而得到或发展起来的熟练性、能力或灵巧度”。从定义可见，技能与实践是密切相关的。技能分为操作技能以及心智技能，前者是控制操作活动动作的执行经验，其动作是通过外显的机体运动来实现的，动作的对象是物质性的客体。后者则是控制心智活动动作的执行经验，其动作常借助于内潜的头脑内部语言来实现，动作的对象为事物的信息。从目前 CAA 研究的现状看，技能测评研究的报导甚少。

技能测评的一般方法是：通过对事实的收集和分析，确定被试的技能水平，测评过程通常可分为四个步骤：

① 定义评价目标和要求，设定目标技能及其评价标准；

② 收集能够体现被试技能水平的事实（信息）；

③ 将事实和测评目标相匹配，最简单的方法是将被试的答案与标准答案进行对比，找出差异；复杂一些的问题往往要应用知识推理的方法；

④ 根据测评标准，将这些差异与被试所表现出的技能水平联系起来，判定被试是否已具备相应的技能认证资格。测评标准由一整套规则构成，这些规则将事实中错误的存在或不存在与特定技能的缺乏或拥有联系起来。

测评自动化是指事实的收集和分析能自动进行。以 IT 技能测评为例，可供分析的事实（证据）主要有两类：

① 用户提交的文档，它通常是以某种格式存储的文件，通过构造文件解析器可获取相关的事实；

② 被试完成技能任务过程中的动作序列——事件流。

为保证测评精度，同时收集结果信息和事件流进行综合评价是必要的，因为事件流可以提供更多的过程信息。以字处理中的字符串删除为例，如果只靠结果信息作评价，可能会出现不正确