语言技术与教学研究

用现代化手段辅助语言测试

吴会芹
(浙江大学 宁波理工学院 外国语分院,浙江宁波 315100)

摘 要: 本文简要回顾了机辅语言测试的历史发展过程。从试题库建设、试题结构设计、命题设计等方面描述了机辅语言测试的准备过程,并分别从单机测试和网上测试两个方面描述了机辅语言测试的实施过程。机辅语言测试的现代化不仅包含测试过程的现代化,还包括阅卷、评分等方面的现代化。实现语言测试的现代化,不仅能提高工作效率、降低成本,而且对整个语言教学将会产生深远影响。

关键词: 试题库;机辅语言测试;机辅语言阅卷
中图分类号: H319.3 **文献标识码:** A **文章编号:** 1001-5795(2006)03-0049-0005

语言测试研究主要涉及两门学科内容:一是研究语言测试内容的学科,一是研究语言测试手段的学科。语言测试手段主要指计算机技术应用于语言测试,该技术的发展主要涉及电脑应用、多媒体技术、人工智能应用、自然语言处理(natural language processing)等方面,这些技术的发展将为实现语言测试现代化提供技术方面的支持。本文将着重探讨如何运上述技术手段辅助、优化语言测试,尤其是英语语言测试之问题。

1 机辅语言测试的回顾

计算机辅助语言测试走过了一段曲折的路程。语言学家李筱菊(2001:441-447)将其发展进程概括为五代:

第一代称为计算机化(computerized)语言测试。这一代语言测试在测试标准上有一定进步,受试者通过直接键入答案,减少了输入和计算分数过程的测量误差(measurement errors);试题及答案均存于机内,加强了试题的保密性(test security),提高了受试者应试时间的灵活性。不过,第一代语言测试有很多局限性,难以做到试卷的情景化、互动化、智能化、个性化。

第二代为调适性语言测试。这一代语言测试在个性化方面取得一定的进步,但人工智能开发欠佳,在智能化及互动化开发方面未能取得令人满意的成果,而且尚无法处理情景化的多维考试项目。

第三代语言测试是多媒体语言测试。多媒体语言测试实际上是多种技术的综合运用,它不但具有比上一代更成熟的调适性能,对人工智能也有一定成效的运用,增进了受试与测试之间的互动化和智能化,为测试建立多维度的情景,达到了语言的真实性、情景化的效果。

第四代语言测试实现了从静态到动态的转变,在把测试的焦点从成果转移至过程方面带来了根本突破。但是由于认知科学的应用和人工智能的开发尚不充分,故对整个做题过程不能用认知科学做出深刻的解释;虽然对受试者的学习情况和进程能做出动态的描写,但缺乏智慧的分析,因而对学习者尚不能提出指导性建议。

第五代语言测试,称为智能化测试。它是"人(教师)和机器(计算机)的结合,智慧(intelligence)与能量(power)的结合"(李筱菊,2001:446)。它"把最优秀的教师的智慧(知识、经验、理解、推理、分析、判断、预见等等)集中起来,用符号的形式变为计算机的知识库(knowledge bases)。以知识库为基础的专家系统,使得计算机能做有智慧的教师所能做的很多的事,而且做起来速度增加无数倍,幅度扩大无数倍"(李筱菊,2001:446)。

语言测试进入第五代以来,逐步开始走向个性化、情景化、互动化、智能化。从此,大量纷繁复杂的统计

作者简介:吴会芹(1963-)女,硕士,副教授。研究方向:机辅语言教学,词汇学,英汉语言文化对比。
收稿日期:2005-11-25

工作逐步由智能计算机完成，语言测试正逐渐朝现代化迈进。

2 机辅语言测试准备

2.1 试题库建设及理论依据

试题库是70年代教育测量领域中的一项重要技术成果，它能高效地、科学地、反复地生产信度高的、等值的试题，从而解决了考试从不科学到科学，从不稳定到稳定，进而达到标准化的一系列大问题。起初，试题库的建立主要依据经典测试理论（classical test theory），该理论建立了表示难度的值（p-value）、表示区分度（d-value）的项目与测试相关系数（coefficient）、标准差（standard diviation，简称SD）及由此推断出来的信度（reliability）公式等。但是，由于依据这一理论所进行的测试对考生情况估计不足，致使考试结果难以做出相应比较。后经统计学家Lord等人的努力，又推出了比较完善的项目反应理论（item response theory 简称IRT）。IRT认为，通过考生对具有一定难度和区分度等特征的项目的反映可以确定考生的潜能特征和倾向。它把考生能力值和项目难度值以统一的计量单位置于一个量表中，来研究考生分数与其实际能力或倾向之间的函数关系，具有项目难度不受样本影响（sample-free）、能力不受项目影响（item-free）的优点。IRT通过适宜性检验了解试题的效度，且可使各种试题项目入库及调整项目的难度。由于IRT具有比传统测试理论无法比拟的优越性和实用性，该理论曾在80年代几乎独霸整个欧美测试领域。

试题库的建设是实现机辅测试的基础，有了试题库不仅能产生不同目的的试卷，更重要的是，它充分利用计算机速度快、准确率高、容量大等优势，全面合理高效地对测试数据进行统计分析和处理。

2.2 机辅试题结构设计

机辅试题结构的设计主要应从以下三方面考虑：首先，测试内容要围绕测试大纲进行测试范围的划分和测试内容的确定。其次，均衡测试内容的比重。因为，考试内容的比重往往通过其反拨作用直接左右着教学的发展方向。第三，要充分考虑测试的题量、难易度及信度，有效把握主客观题型的比值。虽然大量的客观题型对于能否测试出应试者的语言应用能力的真实水平越来越受到语言教育工作者的质疑，但是，由于测试技术的局限性，开展大规模的测试，尤其是机辅测试还需依靠大比例的客观题型。

2.3 机辅命题设计

客观题命题一般有六个步骤：确定考点；编写题干和答案项；编写干扰项；修整选择项和题干；排列选择项；全卷调整答案分布。人工命题是个很复杂的过程，需要命题人员艰辛的劳动。不过，现在的电脑命题系统还只是比较机械的试题。这些系统大都利用专家系统（expert systems），靠的是人工智能的开发。可喜的是，试题库、资料库的建立使命题变得易于操作。资料库的建立带动了命题中现代化科学技术手段的渗透，它使得大规模的语言测试在短时间内即出结果成为可能。

3 机辅语言测试的实施

机辅语言测试主要有两种形式，单机测试和网上测试（包括局域网测试）。

3.1 单机测试

单机测试是由教师通过软件制作试题，然后交给学生一张软盘。学生通过软盘在电脑上进行考试后，将存盘交还教师。教师通过计算机进行阅卷。这种测试适合较小规模的阶段性测试或提交作业等。

例如，Dae Dae Enterprises软件公司推出的TrueTest测试软件，就是采用单机测试形式。测试时考生只需得到一张存有truetest.exe，truetest.hlp和default.tst三个文件的磁盘，就可以参加考试。考试进行中，界面下方有一计时窗口，动态表示考生已用去的时间。最后一分钟，时间表呈鲜红色，提醒考生时间即将结束。单机测试能在上述内容的基础上提高教师的工作效率，节省劳务开支，从而优化教学，并可通过反馈信息及时调整教学。

时下许多英语指法软件、单词学习软件的单机版即是该种类型。这种系统（软件）不仅能实现自动评分，而且还有自动纠错并给出解析之功能。

单机测试可以发生在没有网络支持的环境，或在受试者缺乏技术操作能力的情况下。单机测试有很大局限性，其安全性也相对较弱，因此，只能用于不太重要的测试中，如作业等。

单机测试实现了用电脑取代纸张的无纸化测试，而网上测试则使大规模的考试不受时间、空间、人力、物力等的限制成为可能。

3.2 网上测试（On-Line Test）

网上测试也叫联机测试或无纸化测试，它指的是不用纸张而直接利用计算机网络完成测试的全过程。网上测试需要配备高性能服务器，由测试人员将出好的试题变成数据库放入数据服务器中。考试开始前需

要考生输入个人资料，然后进行答题。时间一到，考试自动结束。考试中的客观题能马上显示考生的得分情况，并根据需要打印报表及相关合格证书。

网上测试实现了考务工作的全自动化管理，其考试模式灵活，试题类型多样化，更重要的是它数千倍地缩短了命题、测试、阅卷、评分及数据分析的周期。

我院为英语专业教育方向开设的教学技能课程，是一门以教授学生如何运用各种软件制作教学课件为主的课程，该课程的测试完全使用无纸化测试。命题教师选择内容相当的英语文章，由考生任选一篇，通过借助网络资源，在指定的考场（机房），在规定的时间内，运用指定软件完成多媒体课件制作全过程。然后交卷（存盘）到指定的 ftp 地址。监考只须将 ftp 地址上的所有试卷（课件）收齐（拷贝）即可。阅卷（对课件的评定）也是在电脑上进行。所有试卷打分完毕后，登陆教务处网站进行网上成绩报送。当然，该测试非纯粹的语言测试，其阅卷打分等过程的效率优势并未充分体现出来；但是，它的收发试卷过程却显出明显的优势，既避免了有纸化收发试卷的麻烦，又避免了丢失试卷的危险。由于试卷的文件名是考生的学号和姓名，因此，在试卷排序上也易于操作。

机辅测试不仅大大提高了阅卷工作的效率，而且节约了纸张，避免印刷、分发试卷等工作，节省了大量人力、物力、财力。笔者做过一番调查，现以每年两次的四、六级测试为例，如果每次考生人数按四百万人次计算，每份试卷需要使用 6 张正反两面的 B4 纸张，如果使用 70 克纸胶印，纸张价格按 0.07 元/张计算，胶印费按 0.12 元/张计算，那么，仅纸张和印刷费就是 6,480,000 元人民币，那么一年两次的相应费用则是 12,960,000 元，这是实行机辅测试可以节省的第一笔费用。网上测试的答题、阅卷、评分、统计的全自动化过程免去试卷的收发等工作，并且可节省由人工阅卷及统计所产生的一切费用，如果人工批阅一份试卷按一元计算，其产生费用就是 4,000,000 元，这是可节省的第二笔费用。这样，一年可节省的两项总开支为 16,960,000 元人民币。

机辅语言测试采用考生直接输入答案形式，与人工阅卷相比，减少了由于人为错误导致在计算分数过程中所产生的测量误差。计算机超凡的记忆力和准确率使阅卷记分等误差降到零，从很大程度上保证了测试的公正性、客观性。

正是机辅测试的上述优越性，使其越来越被语言测试领域看中，如 GRE 自 1998 年以来由传统的纸笔测试全面改为机辅测试；东软在线凯思考试（Computerized Assessment System for English Communication）提供的 CALT 系统已经广泛应用于企业、学校和政府，成为评价国际职业英语能力的标准；一些有影响的考试如 GMAT，TOEFL 等也已采用计算机测试方式；我国上海市民通用外语考试也在逐渐采用网上测试形式。TSE（Test of Spoken English）是美国教育考试中心（ETS）为母语为非英语的学生提供的英语口语水平考试。近年来，很多北美的高等院校将 TSE 考试成绩作为选拔教学助手（ITA）的依据。他们依据这一成绩对北美及世界范围英语教师进行资格认证。

可以说，网上测试是语言测试现代化的重要标志，它不仅是语言测试的发展方向，也是其它领域测试的发展方向。随着语言测试理论和计算机技术的不断发展，机辅测试形式将更先进、更科学，人们将会通过网络足不出屋完成报名、答题、查分、打印证书等全过程。

诚然，作为一个新生事物，机辅语言测试必然有许多不尽人意的地方；但随着人工智能的开发，随着语音识别技术的不断进步，这方面的工作必将日益完善。

3.3 计算机调试性语言测试

计算机调试性语言测试（computerized adaptive testing，简称 CAT）是现代语言测试理论与当代多媒体技术相结合的产物。它以 IRT 为理论基础，并依此为数学模型建立试题库，在测试过程中可以根据受试对象的答题情况，从试题库中选取符合受试者语言水平的题目进行测试，在最短的时间内准确评估受试者的语言能力，直至达到预定的测试目的。CAT 的优势在于具有明显的渐进性和个体性，即受试者即将面对的试题是以其前面答题情况而定，其题目或难或易，因人而易，循序渐进，具有很强的科学性和准确性。CAT 不但以信息函数这一综合质量指标为科学的选题标准，而且题目函数估计准确，能最大限度地测算受试者的实际语言水平，并且具有稳定性（Stability）、等值性（Equivalence）与客观性（Objectivity）。CAT 安全指数高，几乎不可能猜题、压题、舞弊。CAT 由于题量灵活，时间可长可短，测试后能即时记分并报告成绩，具有经济性和实效性。

4 机辅语言阅卷/评分

有了试题库就能快速自动地生成各种格式化试卷，有了测试软件，就可以实现无纸化测试。但测试之后另一项单调而繁重的工作便是阅卷和试题分析。计算机不仅可以用来储存材料、建立试题库、编制试卷

等，而且还可以用来阅卷评分、做试题分析和模拟录取等工作。

实现机辅阅卷主要有两种途径：一是在网上测试结束后，利用网络测试软件实现网上阅卷、评分并打印证书等；二是测试时采用答题卡形式，考试结束后由考务人员收取答题卡，并利用光标阅读器（阅卷机）实现阅卷、登录成绩、测试分析全过程。

目前，高考网上阅卷已陆续在我国部分省市展开，如《网络阅卷管理系统》是基于 WINDOWS NT /WINDOWS 9X 操作系统所开发的阅卷管理系统。该系统可支持多达 22 台计算机同时阅卷，具有安全性高、速度快、准确性高之特点。

我国部分省市地区还实现了中考网上阅卷。如，“ZK99”系统是专为中考而开发的通用阅卷及数据处理软件，它同时可进行主观题和客观题的阅卷，使用方便快捷，大大提高了工作效率，减轻了劳动强度。

5 机辅语言测试的展望

随着现代计算机技术的发展和语言测试理论的不断完善，测试软件如雨后春笋展现出勃勃生机。在我国，多种语言测试系统正在或已经被开发出来。如杰佛公司推出的面向企业的网上考试软件 WebExam；深圳市新为软件有限公司推出的 SmartExam 在线考试系统（http://www. pcdog. com/soft/31528. htm）；四川省干部英语考试系统（http://www. scrtvu. net/thesis/files/yjlc/lc200242. html）；由 http://www. superdown. com/soft/1306. htm 网站推出的 E-Exam 英语考试自测系统 1.1；成都祥和源科技发展有限公司推出的 E 灵通考试系统（NeoExam）；东软在线教育（http://www. neusoftonline. com/casec/info. jsp? id = 03）提供的凯思英语考试系统等等以及许多软件公司、科研机构利用隐马尔科夫模型（HMM）开发并推出的语音识别引擎，如微软的 Whisper，Scansoft 公司的 Dragon Naturally Speaking，飞利浦的 FreeSpeech，IBM 的 ViaVoice 等等。基于 IRT，20 世纪 80 年代已有 BICAL 软件、LOGIST 软件，90 年代以来，新软件层出不穷，由芝加哥大学心理学实验室设计的 BIGSTEPS 软件，由 Kansas 大学 David Thissen 教授设计的 MULTILOG 软件，有美国教育测试中心 Eiji Muraki 教授设计的 PARSCALE 软件及该中心的 Robert，J. Mislery 和芝加哥大学 Bock 共同开发设计的 BILOG 软件等。此外还有 PARELLA、ASCAL、RASCAL、XCALIBRE 及 WINMIRA 等软件，这些软件都能一次运行处理数千个考生和项目的数据，是当前较好的 IRT 软件。

语言测试智能化及互动化方面的工作已经取得丰硕成果。上个世纪 80 年代，美国 Clolrado State University、加拿大 Canadore College 及香港、台湾等地的大学就尝试借助计算机辅助写作教学。现在，美国 University of California，Berkeley，Brandeis University he Kickinson College 等大学全部采用了电子评分方法。学生用电子邮件将自己写的论文发给指导教师，教师利用 Word 的一些功能，如字体颜色、批注、插入——甚至可以在文本中插入语音评语来指导或评价学生的写作。美国的 ETS（Educational Testing Service for Institutions of Higher Education）机构还成功开发了电子作文打分系统（Electronic Easy Rater，简称 e-rater）。该系统根据评分专家事先设定的评估作文成绩的标准，可自动分析考生作文的特征，并与专家设定的特征相对照，给出等级。与人工阅卷相比，其准确率已达到 87% ~ 94%。

在许多语言测试软件中，一个最引人注目的测试系统当推 PhonePass 全自动英语考试系统。该系统集语音识别、测试学理论、心理学、数字分析、统计学模型理论等尖端技术于一身，通过数字化技术，将受试者的语音传输到中央处理服务器上进行数据分析和处理，通过受试者在交谈中使用的词汇、句子、短语及其语速、流利程度和发音情况，对受试者进行综合评估。

据了解，该系统曾被美国教育考试处 ETS 选用并作为 TSE 英语口语考试的测试平台，该技术平台还被用于新托福口语测试。PhonePass 现在已经被美国、加拿大、德国、韩国、日本、香港等国家和地区广泛认可，并被广泛用于全自动英语口语测试。据说，在汉城 2002 年世界杯足球赛期间，组委会就曾应用自动化测试对数万名志愿者进行英语口语的评估。在我国，一向锐意改革的北大在全国高校中首用英语口语考试新方法，率先采用了由联晨教育引进的 PhonePass 全自动英语考试系统。在最近结束的北大英语四级考试中，共有 1808 名在校本科生参加了口语考试。考试完毕，考生只需要拨通指定的电话号码，与全自动测试系统进行 10 分钟的互动交流即可完成英语口语考试。几分种后，老师就可以在网络上查询到成绩报告。

据介绍，该系统将发布 IDT 网络版本，届时，受试者无论在哪里，只要有可以上网的电脑，耳机和麦克，就可以轻松地完成测试。（http://www. neusoftonline. com/casec/info. jsp? id = 03）

如果将该系统引入到四六级口语考试，必将大大

推动我国英语口语教学，必将激励无数名莘莘学子张开难启的嘴巴，困扰英语教育工作者多年的“哑巴口语”的难题有望迎刃而解。

测试是教育中需要频繁进行的重要活动，也是教学工作的一个重要组成部分。应用计算机来完成外语教学测试可以完善教师的教学过程，对教学的评估、学生的评定提供充分的依据。CALT 是今后语言测试发展的方向。随着计算机人工智能、光学字符识别（Optical Character Recognition，OCR）、语音识别等技术的提高，在不远的将来，我们将能够使计算机通过控制软件与考生直接交互、识别和判断实际的书面甚至口头语言进行评估，高效率地进行测试。由此可见，实现语言测试的现代化，与世界接轨，是大势所趋，也是我们今后改革外语测试、提高外语教学质量的必备条件。□

（本文系 2006 年度浙江大学宁波理工学院教育与科技发展研究课题部分内容）

参 考 文 献

[1] 章国英. 计算机辅助外语教学[M]. 上海：上海外语教育出版社，1995.

[2] 李筱菊. 语言测试科学与艺术[M]. 长沙：湖南教育出版社，2001.

[3] 庄智象，等. 全国高校“新理念”大学英语网络教学试点方案[C]. 上海：上海外语教育出版社，2004.

[4] 沙国全. 计算机辅助语音训练与测试：问题与思考[J]. 外语电化教学，2005，(2)：.

[5] 孔 文. 大规模语言测试的方向：计算机适应性语言测试[J]. 外语界，2002，(2)：.

[6] 冯慎宇. 利用计算机辅助语言测试——兼介绍 TrueTest 测试软件[J]. 外语界，2000，(1)：.

[7] 蔺长旺. 教育测试新理论——IRT 的研究与应用[J]. 外语教学与研究，2000，(5)：.

[8] Bachman Lyle F. & Palmer Adrian S. Language Testing in Practice[M]. Oxford University Press 1996, Shanghai Foreign Languages Education Press 1999.

[9] http://www.wuyou.com.cn/Article_Print.asp? ArticleID=690.

[10] http://www.pcdog.com/soft/31528.htm.

[11] http://www.scrtvu.net/thesis/files/yjlc/lc200242.html.

[12] http://www.superdown.com/soft/1306.htm.

[13] http://www.neusoftonline.com/casec/info.jsp? id=03.

Computer-Assisted Language Testing

Wu Hui-qin

(Ningbo Institute of Technology , Zhejiang University, Ningbo, Zhejiang 315000, China)

Abstract: After looking back at the five generations of CALT and its application process, this thesis introduces us the preparing process of the CALT, in which it mainly involves Item Banking and Test Designing. The thesis fully describes its operating process, including On-Line Testing and Individual Machine Testing, and briefly points out its potential advantages and its increasing impact on the whole language teaching. CALT is the trend of language testing. Looking forward to CALT, the highly developed technology and highly advanced technical devices will not only lead to higher efficiency in language teaching and learning, but also bring forward to the strictness and higher reliability of language testing.

Key words: Item Banking; Computer-Assisted Language Testing (CALT); Computer-Assisted Language Rating