创刊三十周年特约稿-外语测试学研究专辑

计算机化考试的设计模型

曾用强
(广东外语艺术职业学院,广东 510640)

摘 要:计算机化考试的发展取决于两个重要的因素:计算机技术和测试理论。计算机化考试从测试理论获得考试的内容,从计算机技术获得考试的形式,内容和形式有机地结合,形成一种新型的考试模式。本文重点讨论这种新型考试模式的四个基本构成及其设计:题型设计、试题组织、能力估算和成绩报告。计算机化考试的主要优势特征也集中体现在这四个方面:可以应用创新型试题;可以实现适应性测试;实现了多维的能力估算;可以向考生提供即时的诊断性信息。在实际应用中,计算机化考试的设计就是优化设计这四个基本构成,使计算机的优势得到最大限度的发挥,满足测试的目标要求。

关键词:计算机化考试;创新题型;适应性测试;诊断性信息
中图分类号:H319.3 **文献标识码**:A **文章编号**:1001-5795(2012)01-0022-0006

1 计算机化考试的发展

未来的考试如何发展? Hambleton(2004)认为,要预测教育心理测试的未来发展方向并非易事,因为测试方法和计算机技术的发展十分迅速。但是毫无疑问,未来 20 年的最大变化就是更多的考试将通过计算机来实现。计算机技术的应用促使考试在诸多方面发生了变化,比如:考试开始从线性(所有的考生完成相同的试题)到非线性(测试项目的难度适应于考生的能力水平)转变;题型设计从传统的选择型试题(如单项选择等)向构建型试题发展,因为主观题的自动化评分技术得到迅速的发展;考试内容向更高层次的认知能力方向发展;考试的评价开始引入过程参数,比如:答题的反应时、答题顺序及修改次数等;考试的组织更加灵活。

计算机化考试研究始于上世纪 70 年代,至 90 年代才开始走向成熟、并真正开始应用于考试的实践中,比如:

· 1996 年欧洲几所大学联合开发了欧洲诊断语言评价系统——DIALANG,免费向学习者开放,它涵盖十四种欧洲语言。

· 美国 ETS1998 年推出托福机考 TOEFL_cBT,2005 年又发展成为托福网考 TOEFL_iBT。

· GMAT 和 GRE 分别于 1998 年和 1999 年开始施行计算机化考试。

· 从 2011 年 8 月起,托业 TOEIC 开始施行计算机化考试。

· 2004 年广东省推出高考英语口语考试机考,每年考生约 20 万。

· 2009 年大学英语四六级考试机考开始在全国部分高校试点。

· 2011 年广东省进行高考英语听说考试机考的改革,每年考生超过 60 万。

· 2011 年全国成人高等教育申请学士学位英语考试机考在广东试行,考生约 1 万。

计算机化考试由于其具备许多传统纸笔考试所无法比拟的优势,而成为未来考试发展的趋势。这种发展趋势的最终结果是,语言教学和语言测试之间的界限越来越模糊,直至最终完全一体化,满足学习者的终生学习需求。

2 计算机化考试的基本构成

计算机化考试的发展取决于两个重要的因素:计算机技术和测试理论。计算机化考试不是简单地把考

作者简介:曾用强:博士,教授,博士生导师。研究方向:语言测试。
收稿日期:2011-11-20

试试题移植到计算机，而是把测试理论和计算机技术有机地结合起来：从测试理论获得考试的内容，从计算机技术获得考试的形式，内容和形式有机地结合，形成一种新型的考试模式。与传统的纸笔考试相比，这种新型的考试模式的优势特征主要体现在四个基本构成的设计上：题型设计、试题组织、能力估算以及成绩报告。

题型设计：设计能够满足考试目标需要，适用于计算机化考试的题型。计算机化考试题型设计有两种基本思路：沿用传统纸笔考试的题型，或设计创新型的考试题型。

试题组织：按照特定的规则确定题量、试题内容以及试题的编排。计算机化考试通常有两种试题组织方式：线性和非线性。线性的考试是预先定义好考试内容和题量，所有的考生都回答相同的试题，只是有时为了防止舞弊，考生的答题顺序或单选题的选项顺序可能不完全一样。非线性的考试也称适应性测试（adaptive testing），考生只需回答与其能力水平相适应的试题，每个考生所需回答的试题数量也不一定相同。非线性考试需要建设一个带参数的试题库。

能力估算：根据考生的考试表现估算其能力值。计算机化考试的能力估算主要有两类：基于结果的能力估算和基于过程的能力估算。前者就是根据考生的答对情况对其能力进行评估；而后者则是通过采集考试的过程数据，对考生的能力做出更加客观和科学的评估。

成绩报告：采用特定的形式向考生报告能力估算结果。计算机化考试可以向考生反馈单一的成绩报告单，也可以提供详细的能力诊断报告。

题型设计、试题组织、能力估算和成绩报告是设计计算机化考试时必须考虑的四个基本组成部分，也是体现计算机化考试优势特征的四个主要方面：在题型设计方面，计算机化考试可以应用创新型试题；在试题组织方面，计算机化考试可以实现适应性测试；在能力估算方面，计算机化考试实现了多维的能力估算；在成绩报告方面，计算机化考试可以向考生提供即时的诊断性信息。

3 计算机化考试的设计

3.1 创新题型的设计

开发创新试题是计算机化考试领域中最具吸引力的环节，它不同于传统的、单一考点的、基于文本的试题类型。正是因为计算机化考试具有创新性的试题类型，才得以大大改进了测量质量。创新试题类型是指那些试题包含有只在计算机上运行实施才能体现的特征。这些创新特征主要体现在以下5个方面：试题格式（item format）、反应行为（response action）、媒体引入（media inclusion）、交互层次（level of interactivity）和评分方法（scoring method）。

试题格式　选择题是考试中最常见的试题格式之一。大家熟悉的纸笔考试的选择题要求考生从2～5个备选项中选择最佳答案。但是，在计算机化考试中，这种格式的试题可以增加备选项或提供更直接的测试方法，以减低猜测度。比如，要求考生点击阅读短文中的某个句子或选择复杂图形中的某一部分。再如，在测试写作技能时，可以要求考生面对一篇包含一些语法和文体错误的文章，但不标示错误的位置。考生用移动键指向他/她认为需要修正的错误位置，然后计算机显示用于改写的备选项，考生从中选择其中的一个选项或放弃选择。

还有其他的选择题格式，这些格式在传统纸笔考试中使用过，也很容易移植到计算机上实现。其中一个示例就是允许考生多次选择答案，每次选择后即时反馈是否正确，最后的得分是基于直到选对为止总共选择的次数。

反应行为　反应行为是指考生回答试题时所作出的物理方面的行为。最常见的反应行为就是传统纸笔考试中使用铅笔在相应的位置做出选择或涂个椭圆形等。

在计算机化考试中，反应行为通常是靠鼠标或键盘来完成。考生的反应通过键盘输入，或点击鼠标等。比如，选择题的答案选择可以使用键盘、或输入相应的字母或用鼠标点击来完成。考生可以使用鼠标点击文本文章或图形中的某位置、或加亮文本或图形、或拖拉文本，数字或图标到指定的位置或顺序等。

除了键盘和鼠标还可以有其他的输入设备，比如，触摸屏、光笔、操纵杆、轨迹球和麦克风等。比如，使用麦克风收集口头反应等。

媒体引入　媒体引入为测量带来的最主要好处也许就是扩展了测试的内容和认知技能的覆盖面。增加非文本媒体可以提高任务的情景性和考试的效度，减少对阅读技能的不适当的依赖。

电脑内置的功能可以把非文本媒体引入到计算机化的考试中。计算机可以用来显示图形、播放声音、运行动画和视频等。包括这些非文本媒体的创新试题类型经常是把这些媒体包含在题干中，然后再与传统的

考试题型配合使用。

除了单一使用这些非文本媒体外，创新试题类型还可以同时包括多种形式的媒体，比如在听力考试中，可以应用图片或两人会谈的视频，同时播放他们会谈的录音。

交互层次 大多数的计算机化考试试题都是非交互性的，考生完成动作后（如点击鼠标作出选择）就完成了解题过程。但是，也有一些计算机化考试试题应用了有限的交互（高度交互的试题较少）。比如，在要求考生编辑文本的试题中，编辑后的新文本可以显示在原来的文章内，让考生再读这篇修改后的文章。在排序的试题中，一旦考生排完序后，他们能够看到新排序的项目。

还有一种交互性结合了两步或多步骤分支功能。比如，某车间发生冲突，考生的任务是解决这场冲突。在这个应用中，试题首先显示车间冲突的视频场景，然后是涉及到如何解决冲突的多项选择问题。考生一旦选择了一个选项，就显示与该选项有关的第二个视频场景。这种评估使用了两个阶段交互式的分支层面。这种交互式的题型可以应用于计算机化英语口语考试中。

评分方法 传统的标准化纸笔考试以及计算机化考试中的传统题型都采用计算机自动评分。自动评分还应用于一些创新性的题型。如果试题不包括交互性、只需要考生做出选择判断，自动评分就是一件很容易的事。但是对于其他类型的试题，自动评分就是一个比较复杂的过程，比如作文评分。

至今已经有一些现成的自动评分软件，如：PEG，E-rater，Intellimetric Engineer，Intelligent Essay Assessor，InQuizit 等。这些软件的评分标准存在很大的差别，其中有些系统只考虑表层特征，而有些系统应用了高级计算语言学的理论和方法。即使存在这么大的差别，有关这些系统的研究已经表明，它们的评分信度可以接近人的评分信度，可以充当第二个评分员的角色。

3.2 适应性考试的设计

计算机化适应性考试（Computerized Adaptive Tests，CAT）属于非线性考试，其考试的流程不是预定义的，而是根据考生在考试进程中的不同表现确定测试项目的内容和题量。其理论依据和实现方法源自项目反应理论。

计算机化适应性测试的发布方式就是，测试项目的选择决定于考生前面的项目反应。这种考试的目的是，根据潜在的评估标准对每个考生的水平提供精确的评估。测试项目的数量、具体的测试项目以及测试项目的出现顺序都可能因不同的考生而不同。每个考生的评估都基于项目反应理论潜在能力估算的同一量表上，可实现即时的成绩反馈。

适应性测试由四个基本部分组成：起始项目、项目选择规则、能力估算和终止准则。计算机化适应性测试主要有以下几种常见的设计模型：双阶测试、误差控制测试、多层次分支测试和阶梯式测试。

双阶测试（two-stage tests）分两个阶段进行：全体考生首先参加第一阶段的常规测试，然后是第二阶段的优选测试。优选测试将每一考生在第一阶段测试中获得的能力估计值作为起始水平，计算机据之直接为每个考生选出一套最优化的测试项目。项目选择的规则是，以最少的项目提供最大的测试信息。最大测试信息量同时也将作为双阶测试的终止准则，能力估计的方法是平均难度评分法（Weiss，1974）。在这个评分方法中，每个考生的能力就是他正确回答的所有项目的平均难度。

误差控制测试（error-controlled tests）从中等难度的项目（难度 = 0.00）或与考生能力相适应的项目开始。如果考生正确回答了初始项目，下一个项目就更难些（难度增值一般为 0.5），反之，难度就减值 0.5。当考生至少出现一个正确的和一个错误的项目反应时，将获得一个能力估算值，测试可以反复进行，终止准则是，测量标准误差是否低于预先确定的精度水平（比如，预设的标准误差 = 0.30）。

在误差控制测试模式中，由于变换了其中的初始项目以及项目选择的方法，在实际运用中，这种模式存在三种变体形式，三种变体形式的基本步骤是完全一致的。

多层分支测试（variable-branching tests）始于一个中等难度的项目，项目选择的一般规则是，如果回答正确就进入更难的项目，否则进入更容易的项目。当考生答完全部项目后，使用极大似然值估算方法估算出考生的能力值。

阶梯式测试（step-ladder tests）题库中的全部项目首先按不同的难度划分成几个不同的区域（如，将难度从 -3 到 +3 的项目分成 12 个不同的区域）。但是，每个区域必须包含大致相同数量的测试项目。在测试过程中，首先从中间区域随机选择一个起始项目，如果考生反应正确，计算机就往上一个区域随机选择一个新项目，否则就往下一个区域随机选择一个新项目。终止准则是（1）考生的项目反应在某两个区域中出现正

表1 CAT模型之对比

模型	起始项目	项目选择规则	能力估算方法	终止准则
双阶测试	每个考生在常规测试中的能力估计	最大测试信息函数	平均难度评分法	测试信息函数
阶梯式测试	中间区域中的随机项目	根据项目反应在各个区域上下移动	极大似然值计算	答对某区域中的一定数量的项目
多层次分支测试	中等难度的项目	按常数进行增值或减值		固定数量
误差控制测试(Ⅰ)		项目难度与考生能力相匹配		测量标准误差
误差控制测试(Ⅱ)	与(已经获得的)考生能力相适应的项目	最大项目信息函数		
误差控制测试(Ⅲ)				

误徘徊n次后(比如,预设n=5)或(2)测量标准误差低于预先确定的精度水平(比如,预设的标准误差=0.30),测试就终止。考生的能力根据其在当前区域以及前后一个区域中的全部项目反应进行估计,也采用极大似然值计算方法。

以上介绍的四种基本模型及其变体都通过运用不同的方法来完成CAT中的四个基本部分。表1归纳了这六种CAT模式的异同点。

3.3 基于过程的能力估算

使用计算机组织考试可以更好、更有效地获取考生更有意义的考试行为抽样,详细记录考生的考试过程信息,比如答题反应时、解题顺序、修改次数以及是否求助帮助信息等,这些信息在一定程度上反映了考生对所测知识点的把握程度。此外,还可以利用计算机技术,开发一些多维能力的测试题型或测试系统。

迄今,几乎所有的考试,包括计算机化考试,都只能依据考生的项目反应(答对或答错)计算成绩或估算能力值。在经典真分模式中,考生的能力估算只是简单地计算其回答正确的测试项目数量(即答对率);项目反应理论则是根据考生对不同难度项目的反应推导出其能力值。不论是经典模式,还是项目反应理论,目前采用的能力评估模式都是建立在单维能力假设的基础上,即:影响考生测试行为的因素只是考生某一特定的单一语言能力(如:词汇能力或阅读理解能力等)。但是,我们都知道,在实际的语言测试中,影响考生测试行为的因素除了考生的语言能力外,还有许多其他因素,如:测试条件、测试焦虑以及其他个性特征等。这些因素在单维能力假设中被认为是构成测试误差的主要来源。单维能力假设的最大缺陷就是,在估算能力真分的过程中,它不考虑影响考生测试行为的诸多因素,因此无法对考生的能力真分作出客观、准确的评估。随着语言测试理论的发展和测试手段的现代化,语言测试已经开始朝着多维能力评估的方向发展,即,考生的测试行为被认为是多个能力共同作用的结果。考生语言能力的评估不仅仅考虑测试结果,而且开始考虑测试过程,即:考生在答题过程应用了哪些策略和心理思维过程等,比如考生的答题用时、修改答案次数以及正确回答测试项目的条件(如:是否获得有关的帮助等)。基于过程的能力估算是计算机化考试的最大优势之一。所以,应该采集哪些过程数据以及它们如何参与能力估算是计算机化考试的重要研究课题。

3.4 诊断式成绩报告单

成绩报告包括单一分数(single score report)和多分数报告(multi-scores report)两种。计算机化考试除了根据考生的项目反应对其能力水平做出评价外,还可以收集考试过程的信息,对考生的考试过程进行全面、客观和科学的分析。因此计算机化考试的成绩报告单应该包含分数以外的更多信息,对考生的能力做出更加详细的、多维度的分析和评估,通过各种数据给考生一个知识、能力和潜质的报告。考生从成绩报告单中应该了解到自己哪些方面是强项,哪些方面是弱项,明确自己今后努力的方向,而不仅仅是知道自己的成绩或是否通过考试。诊断式成绩报告单除了必须的总成绩以外,还包括成绩分布(各单项成绩)、成绩分析(比如:考生的成绩与全体考生的平均分、最高分和最低分之比较等)、答题异常(比如:答错相对其他考生容易的试题、答对相对其他考生较难的试题)、过程分析(比如:答题用时、答案回溯、答题顺序与修改次数等)和成绩评语等。而且诊断式成绩报告单还可以使用图表更加直观表示这些信息。托福网考的成绩报告单就是诊断式的,它为考生提供了成绩分布信息以及能力诊断分析报告。

4 计算机化考试的应用

在实际应用中,计算机化考试的设计必须充分考虑理想目标与现实情况,优化组合题型设计、试题组织、能力评估和成绩报告的不同形式,使计算机的优势得到最大限度的发挥,满足测试的目标要求。比如,基于已有纸笔考试的计算机化考试改革最好先沿用传统

的题型设计，实现平稳过渡后逐步引入创新题型，这样有利于考生的备考，提高考生的接受度。大规模高风险考试通常不采用非线性的设计，因为建设题库的成本太高，尤其是建设一个符合适应性测试标准的题库实在太难了，甚至有可能题库一旦建成，考试已经落后了。美国 ETS 的托福考试的发展就是一个例子。

托福经历了三大主要发展阶段：pBT（纸考），cBT（机考）和 iBT（网考）。从 pBT 到 cBT，托福的最大变化体现在，它从一个线性考试发展到一个非线性考试，实现了计算机化适应性测试，即考试的题量和选题不是预先定义的，而是根据考生的测试表现不断估算能力值，并据此选择合适难度的测试项目。适应性考试对试题库的量和质都有非常高的要求，要满足大规模，尤其是高风险考试的要求决不是一件容易的事情。美国 ETS 花费了大量的人力和财力，用了几十年的时间建立了题库，但是最终还是不能满足考试不断发展的需求，比如，试题的安全性和考试的高效运行等。题库建设是一项耗时、耗力和耗财工程，建成一个题库可能需要几年，甚至几十年。但是题库一旦建成，准备投入使用时，试题库的试卷结构和试题内容可能已经远远落后于考试的发展。

2005 年托福放弃了 cBT，选择了 iBT。从 cBT 到 iBT，托福的最大变化体现在，它从非线性考试又回到了线性考试，但是在题型设计上有很大的突破，引入了多项创新型的考试题型，大大增强了考试的真实性和科学性。新托福在成绩报告方面也有了很好的发展。

从托福的发展，我们可以看出，在实际应用中计算机化考试的设计应遵循以下几条基本原则：

（1）效度和信度优先原则：不论选择哪一种的设计方法（包括题型设计和能力估算等），它首先必须能够满足考试的目标要求，对考生的能力做出准确有效的评估。

（2）可行性是必须条件：计算机化考试的设计必须具有可操作性，也就是，它的运行条件必须能够得到充分的满足，比如，硬件的要求、考场的设置标准、考务工作以及人力和财力的投入等。

（3）先进性是必要条件：计算机化考试的设计需要考虑如何通过发挥技术的优势，在题型设计、试题组织、能力估算或成绩报告方面有所突破，体现出它的先进性，从而提高考试的测量质量。

（4）发展是检验的重要标准：推行计算机化考试之后，与传统的纸笔考试相比，在哪些方面得到了发展以及这些方面是否体现考试的发展趋势？这是在设计计算机化考试时首先必须要回答的问题。

计算机化考试作为正在兴起与发展中的一种考试形式，从考试设计到考试实施的每个阶段它都要经历从开发到改进完善的过程。计算机化考试已经逐步发展成为本世纪语言测试发展的主要趋势，这种发展趋势要逐步模糊语言教学与语言测试之间的界限，最终实现教学与测试的一体化，满足人的语言学习的终生需求。但是在我国，计算机化考试还处在一个发展的初级阶段。不论在理论，还是在实践方面，我们都需要不断的努力，逐步构建适合我国实际的计算机化考试模型。 □

参 考 文 献

[1] Chapelle, C. A., J. Jamieson & V. Hegelheimer. Validation of a web-based ESL test[J]. *Language Testing*, 2003, 20(4):409 - 439.

[2] Hambleton, R. K. Theory, methods, and practices in testing for the 21st century[J]. *Psicothema*, 2004, 16(4):696 - 701

[3] Luecht, R. M. and Clauser, B. E. Test models for complex computer-based testing[A]. In C. N. Mills, M. T. Potenza, J. J. Fremer and W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* [C]. Hillsdale, NJ: Lawrence Erlbaum Associates, 2002.

[4] Mills, C. N., Potenza, M. T., Fremer, J. J. & Ward, W. C. Computer-Based Testing, *Building the foundation for future assessment*[M]. Mahwah, NJ: Lawrence Erlbaum Associates, 2002.

[5] Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davery, T. *Practical Considerations in Computer-based Testing* [M]. Springer-Verlag New York, Inc, 2002.

[6] Valenti, S., Neri, F., & Cucchiarelli, A. An overview of current research on automated essay grading[J]. *Journal of Information Technology Education*, 2003(2).

[7] Weir, C. J. *Language Testing and Validation*[M]. Palgrave: Macmillan, 2005.

[8] Weiss, D. J. *Strategies of adaptive measurement, Research Report* 74 - 5[M]. Minneapolis: University of Minnesota, Psychometric Methods Orogram, Department of Psychology, 1974.

[9] Weiss, D. J., and Kingsburry, G. G. Application of computerized adaptive testing to educational problems[J]. *Journal of Educational Measurement*, 1984, 4:361 - 375.

[10] 李筱菊. 语言测试科学与艺术[M]. 湖南教育出版社，1997.

[11] 曾用强. 对计算机化考试的思考[J]. 外语电化教学，

2010(1).
[12] 曾用强. 自信心与语言测试行为[J]. 现代外语,2002(2).
[13] 曾用强. 个性化自适应测试初探[J]. 外语教学与研究,2002(3).
[14] 曾用强. 电脑顺应性测试模式的设计[J]. 外语教学与研究,1992(2):19-22.

Design Models of Computer-Based Testing

ZENG Yong-qiang

(Guangdong Vocational College of Foreign Languages and Arts, Guangdong 510640, China)

Abstract: Development of Computer-Based Testing(CBT) depends on two factors: computer technology and test theory. CBT is a new type of test model developed from the combined implementation of computer technology and test theory. The paper will deal with its four basic design elements: test format, presentation of test items, ability estimation and score report. These are also four advantages of CBT over the conventional paper-and-pencil tests: use of innovative test items, adaptive testing, process-oriented ability estimation, and diagnostic information about test performance. In the practical application, design of CBT is to find the best solution to such issues as item format, item presentation, ability estimation and score report, which should be in accordance with the test purpose.

Key words: Computer-Based Testing; Innovative Test Items; Adaptive Testing; Diagnostic Information