计算机化考试的研究和发展①

李　虹　车宏生

中国科学院心理研究所（北京 100101）

[摘要]计算机化自适应考试是现代测验研究中的一个新领域。它对于提高测验效率与质量有着重要意义。文章主要介绍了计算机化自适应考试的发展历史及其与传统纸笔测验的关系，尚待解决的问题，并着重阐述和说明了计算机化自适应考试的原理。

关键词　计算机化自适应考试，项目反应理论，题库，信息函数

分类号　59.805_4

1　什么是计算机化自适应考试

计算机化自适应考试(Computer Adaptive Test,简称 CAT)是近年来测验研究中引人注目并取得了重大发展的领域。CAT 是由最早的适应性测验(Tailoring Test)发展而来的。William W.Turn Bull 于 1951 年最早提出适应性测验这一概念，当时适应性测验是指针对被试先前经验选取适合被试能力的题目进行施测,作答完立即评分,并以上一题的作答情况决定下一道测题,直至测验结束为止。也就是在题库中选取符合被试能力水平的题目进行测试。与自适应测验相近的概念有：分支测验（branched testing）、个别化测验（individualized testing）、程式测验（programmed testing）、连续项目测验（sequential item test）、反应权宜测验（response contingent testing）等[1]。名称虽多，意义大同小异，都不外乎具有下列一些特征[2]：（1）有一个由一定量精选而来的测题所组成的题库作为支持；(2) 按照一定的策略进行选题并根据被试的作答情况不断地调整测题；（3）按一定的规则终止测验，评分在施测的过程中进行。

因此，从这一意义上来讲，适应性测验最早应溯源于 Binet 智力测验量表，因为 Binet 量表也是根据被试先前的反应来决定以后测试项目的选择与分层，也是施测与评分同时进行，也是不同的被试可能接受长度不同的测试。

随着测验理论以及测验技术的发展，适应性测试的思想也在测验中逐步地得以体现与完善，尤其是计算机科学的迅速发展，给测验带来了全面而深刻的冲击，CAT 的出现便是

① 本文于 1998-09-28 收到。

这一冲击下的产物。从测验呈现方式、测验编制、施测过程，到评分规则,CAT 与传统的纸笔测试相比都迥然不同[3]。如今在美国，CAT 已在教育测验、职业测量、人事测评等领域中大显身手，如美国研究生入学考试（Graduate Record Examination）、工商管理类研究生入学考试（Graduate for Management and Administration Test）以及全美护士国家委员会资格考试（Nurse National Committee License Test）等都已采取了 CAT 的方式。不难看出 CAT 代表了今后教育、心理测验发展的方向与重点[4]。

CAT 最早是由 Lord 于 1971 年首先提出的[5]。它的出现不仅打破了两千多年沿袭下来的以纸和笔作为作答工具的方式而改为计算机显示屏呈现、键盘与鼠标进行作答的方式，而且与传统的测验相比，测验思想也发生了巨大的变革：它通过给每一个被试建立一个个人化的测验来达到更为准确的测量，因为项目的选择是根据被试的能力水平定身度量而成的，因而被试所做的每一个题目的难度都是与其能力相匹配的。也就是说，水平高的被试能够避免做到相当简单的题目，而能力低的被试能够避免做到超出其能力范围之外的题目。

从其测验思路中我们不难看出 CAT 考试的众多优点[2]：（1）测验效率高：实践表明，CAT 测验只需相当于纸笔测验一半的施测时间和约为 40%的长度，便可达到与纸笔测验同样的精度。因为在测验实施过程的同时进行了测验的评分，两个过程合二为一，且测验长度减少了，因此节约了时间；（2）可比性强：由于特定的实施方式与记分方式，CAT 最终的分数可转化为可相互比较以及解释的量尺，因此可同时应用于常模参照测验和标准参照测验。

CAT 测验是建构在现代测验理论 —— 项目反应理论（Item Response Theory IRT）基础之上的，从题库的建设到测题的选择再到最后的评分，都是以 IRT 为指导进行的。CAT 也被认为是现代测验理论对测验的最大贡献[5]。Hambleton 和 Swaminathan 对项目反应理论作了如下定义[6]：在测验情景中，通过定义被试的特征，即特质或能力、估计被试在这些特质上的得分（称作能力分数）、并运用这些分数来预测或解释项目以及答题情况，来解释和预测被试的作答。特质（trait or ability）和项目（item）是项目反映理论的核心概念。两者之间的关系是 IRT 的主要内涵。

2 计算机自适应考试的原理

2.1 题库的建设（construction of Item Pool）

题目的编制、筛选、组合成题库是进行 CAT 测试的首要步骤。题库并非是一些项目的简单集合，而是一种有机的组合。题库的最重要的特征是其中的项目不仅是测验题目，而且都有必要的参数，而这些参数指标必须满足一定的条件、符合一定的模式。项目是用来反映被试特性的，一个项目的质量直接影响到它对被试能力特性的反应质量。一个项目的性质往往是由项目的参数来进行刻划的。从认知心理学的观点来看，给予被试一个刺激（项目），被试通过加工（以其能力为中介）做出一定的反应，IRT 在已建立起的项目参数、答对概率、被试能力之间的关系模型基础之上，通过外显的、可测的被试的答题反应测量出被试的能力特性θ。因此参数估计决定了测验准确性的基础。在 IRT 理论中，项目的参数

有：（1）鉴别力（discriminative powder index）；（2）难度（difficulty index）；（3）猜测系数（guessing index）。采用不同的参数，就构成了能力θ与参数间不同的数学模型。常用的数学模型有：单参数模型、双参数模型和三参数模型等。

(1)单参数模型　单参数模型是指项目难度与被试能力θ之间的关系数学模型。有单参数常态钟型模型、单参数对数模型以及 Rasch 模型。其中以 Rasch 模型影响最大。 Rasch 模型为一种随机概率模型，数学公式的表达为：

$$p(\theta) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}$$

其中 θ 为能力估计值，b 为项目难度。

（2）双参数模型。如双参数对数模型：

$$p(\theta) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}$$

其中 a 为鉴别力指数。

（3）三参数模型。如三参数对数模型

$$p(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}$$

其中 c 为猜测系数。

IRT 的模型不下二十余种，如何选择恰当的模型进行参数估计是题库选题的关键。不同的模型具有不同的特点，适合于不同条件下的使用。就以上提及的三种模型而言：单参数模型比较简单，使用较为方便，但它对项目参数性质的要求较为苛刻，如在采用 Rasch 模式时，必须满足难度指数介于 0.85 和 1.25 之间，并且猜测系数很小等诸多条件，否则宜采用其他模式；双参数模型要求项目的猜测系数较小；三参数模型虽然具有涵盖较多项目信息的优点，但亦给参数估计带来更为复杂的工作。因此，虽然关于模型选择标准现在尚无定论，不过，可以从命题方式、记分方式、参数性质、样本人数、模型的强健性、假设的满足与否等方面得到一些选题的依据。

在作项目以及被试的取舍时，需要进行适合度的显著性检验以及模式假设检验。卡方检验是最常用的适合度的检验方法，其中以Wright 和 Panchapakesan[8]所提出的方法最受重视。其公式为：

$$\chi^2 = \sum_{i=1}^{n-1}\sum_{j=1}^{n}\left(\frac{f_{ij} - E(f_{ij})}{\sigma}\right)^2$$

其中 $\frac{f_{ij}-E(f_{ij})}{\sigma}$ 称为标准化残差，f_{ij} 为第 i 能力组答对第 j 题的人数，

$E(f_{ij})=r_i p_{ij}$，为第 i 能力组答对第 j 题的人数的期望值

$\sigma=r_i p_{ij}(1-p)$，为 f_{ij} 的标准差

在作项目适合度分析时，将所有被试的标准化残差平方相加，再除以自由度，所得的均方值符合 F 分布，如将之转化为近似的常态分布的 t 值，则可进一步解释项目的适合度；同理，在作被试的适合度分析时，以所有项目的标准化残差平方的总和求得均方值及 t 值。然而，Hambleton 和 Swaminathan 则认为标准化残差不符合常态分布，卡方检验受样本数的影响。只要样本数足够大，大部分的卡方检验都可以拒绝虚无假设，因此有必要进行模型假设的检验。现代测验理论的模型假设都属于强假设，一般的测验资料很难完全符合，但某些必要的检验不能省略。如单向性的检验适用于所有的模型；采用单参模型时要进行等鉴别力的检验；最小猜测系数检验适用于单、双参数模型。考验前提假设成立与否可作为确定所选取的模型与资料的符合程度的又一证据。当用校准样本进行完模型选择以及参数估计时，CAT 测验的后备库业已建立，在这样一个题库准备的支持下，便可进入 CAT 的施测阶段。

2.2 项目的选择

2.2.1 θ的估计

CAT 的诞生过程实质上是一个项目系列设计（item sequential design）的过程，概括地说，每一个项目的选取是根据被试先前的答题情况采用某种选择策略而进行的。具体步骤是根据对被试的能力θ进行估计，再挑选当能力值为θ时，具有最大信息值的项目。现代测验中最为普遍使用的方法是最大似然估计法以及 Bayesian 估计法。

以三参数对数模型为例：

（1）最大似然法（maximum likelihood estimation, MLE）

$$L(x_1,x_2,\cdots\cdots x_n\,|\,\theta)=\prod_{i=1}^{n}[P_i(\theta)^{x_i}Q_i(\theta)^{1-x_i}]$$

$$\ln L(x_1,x_2,\cdots\cdots x_n\,|\,\theta)=\sum_{i=1}^{n}[x_i\ln P_i(\theta)+(1-x_i)\ln Q_i(\theta)]$$

$$\ln(\theta_n)=\max\ln(\theta)$$

其中 L 表示概率，θ为能力估计值，Xi 为第 i 题的项目反应，θn 为最大估计值. 分别对 a, b, c, θ四个参数求偏导，得出由四个方程式组成的方程组：

$$\frac{\partial \ln L}{\partial a_i} = \frac{D}{1-c_i}\sum_{i=1}^{n}\frac{(\theta - b_i)(P_i - c_i)(x_i - P_i)}{P_i} = 0$$

$$\frac{\partial \ln L}{\partial b_i} = \frac{-D}{1-c_i}\sum_{i=1}^{n}\frac{(P_i - c_i)(x_i - P_i)}{P_i} = 0$$

$$\frac{\partial \ln L}{\partial c_i} = \frac{1}{1-c_i}\sum_{i=1}^{n}\frac{(x_i - P_i)}{P_i} = 0$$

$$\frac{\partial \ln L}{\partial \theta_i} = D\sum_{i=1}^{n}\frac{a_i(P_i - c_i)(x_i - P_i)}{(1-c_i)P_i} = 0$$

采用 Newton-Rapson 法，同时处理这四个方程式：a, b, c 三个参数的估计值在建立题库之初已经得到了，因此在用迭代法处理时，先以 a, b, c 的初始值当成已知，代入方程式估计能力参数θ，再将θ当成已知去估计项目参数，直到迭代程序完成。

（2）贝氏估计法

贝氏是逆概论理论的提出者。其理论的主要概念为先验概率与后验概率，因此贝氏估计法的实质是以后验概率作为项目选择的标准。数学公式表达为：

$$\ln f(\theta \mid x_1, x_2, \cdots\cdots x_n) = k + \ln(\theta \mid x_1, x_2, \cdots\cdots x_n) - \frac{1}{2\sum_{i=1}^{n}\theta i}$$

2. 2. 2　信息函数与项目选择

项目信息函数（item information function）是 IRT 的核心概念，这个基础性的概念对测验的应用领域起了诸多影响[9, 10]。项目信息在二维空间中被定义为 Fisher information，是能力θ的函数：

$$I_i(\theta) = \left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2 \Big/ P_i(\theta) Q_i(\theta)$$

项目信息函数的目的是为了获得最为精确的估计。目前，大多数的 CAT 的择题策略是选取当能力值等于上一题所得的估计值时，能使项目信息函数最大化的测题。然而，不难看出，这种选题策略存在一个明显的问题：如果所有估计的能力参数θ与真实的能力水平差距甚远，以这一估计值为依据所选择的项目未必是最适合被试能力水平的项目，从而影响整个测验的精确性。这种偏差在测验的初始阶段表现得最为明显，因为在初始阶段，对能力θ没有先前的估计。因此，当 CAT 的长度过短时，测验的有效性会受到很大的损害。所以，θ的一元函数是否能够提供出某一项目的完整信息内容，是很值得怀疑的。针对信息函数的这一缺陷，Chang 和 Ying 提出采取 global information 的方法进行项目选择[10]。Global information，即 Kollack-leible information，是用来测量θ_0（真实能力）与θ_1（能力估计值）两个概率分布的差异，定义为：

$$k(\theta \| \theta_0) = P_i(\theta_0)\ln\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + \left[1 - P_i(\theta_0)\right]\ln\left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)}\right]$$

K 信息的重要特征之一是：它是两个能力水平θ与θ_0的函数，因此它对于θ与θ_0的接近性并无要求，因而它所代表的信息内容所覆盖的范围较大，以 K 信息为择题标准可能产生的偏离θ_0的误差也就较小。

对被试施测一系列项目后，相应地得到一系列的θ估计值，如果序列{θ1，θ2，……，θn}最终收敛与某一值时，可以此作为测验的终点，收敛值θ则为能力的最终值。

3 小结

随着计算机技术的日益精进，CAT 已经取得了长足的发展，并不断得到完善。然而，CAT 在方法学上与理论上的发展还显得相当薄弱[12]。比如在测验起始点的选择、择题的策略、题库的安全与暴光率、模型的强健性等领域还存在许多亟待解决的技术问题。但无庸置疑，CAT 测验代表了今后测验发展的趋势和方向。在国内，已经出现了一些应用性的研究，自行编制了一些 CAT 测验，并取得了实际的运用效果[13, 14]，相信在未来的几年中，这一研究领域会倍受人们的关注，其研究也会不断深入下去，从而使 CAT 的理论与方法愈加成熟与完善。

参考文献

[1]王宝墉. 现代测验理论. 台湾：心理出版，1983.

[2]Weiss D J. New horizons in testing: Latent trait test theory and computerized adaptive testing. New York : Academic Press, 1983.

[3]韩布新，车宏生. 计算机施测对心理测验的影响研究述评. 见:心理与教育测量. 浙江：浙江教育出版社，1997. 441-449.

[4]Weiss D J. Improving measurement quality and efficiency with computerized adaptive testing . Applied Psychology Measurement, 1982, 6:473-492.

[5]Lord M F. Applications of item response theory to practical testing problems . Hillsdale NJ: Erlbaum, 1990.

[6]Wainer H. Computerized adaptive testing: A primer. Hillsdale NJ: Erlbraum , 1990.

[7]Hambletonm R K, Swaminathan H. Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing , 1985.

[8]Wright B B. Panchapakesan N A. procedure for sample free item analysis. Journal of Educational and psychological Measurement 29, 23-48.

[9]Cover T M, Thomas J A. Elements of information theory. New York: John Wiley & Sons, 1991.

[10]Chang H H, Ying Z l. A Global Information Approach to Computerized Adaptive Testing. Applied Psychological Measurement, 1996. (2) :213-219.

[11]Kullback S Information theory and statistics New York: Wiley, 1959.

[12]张厚粲. 心理测量学的研究与发展. 见:心理与教育测量. 浙江： 浙江教育出版社，1997. 1-6.

[13]孔祥斌，胡文斌. 电脑自适应测验、解题辅导题库. 见:心理与教育测量. 浙江：浙江教育出版社， 1997. 117-125

[14]淑慧，范光辉. 电脑适性测验应用于学业性向测验之成效评估. 见:心理与教育测量. 浙江：浙江教育出版社，1997. 459-472.