

基于 IRT 的计算机化自适应测试系统研究^{* 1}

杨跃诚, 钟汝能, 孙瑜, 肖梦雄
(云南师范大学 信息学院, 云南 昆明 650500)

摘要:计算机化自适应测试(Computerized Adaptive Test, CAT)系统是一种基于项目反应理论(Item Response Theory, IRT)的智能测试系统。它不但可以提高测试效率,而且能提高对测试者能力评价的精确度。作者介绍了IRT的详细内容,论述了CAT的工作原理和实施过程及其系统设计的相关知识,提出一种初始能力值估计的解决办法并对测试终止条件进行了优化处理。为个性化学习提供一种理想的测试方法。

关键词:经典测验理论;项目反应理论;计算机化自适应测试

中图分类号:TP 399 **文献标识码:**A **文章编号:**0258-7971(2011)S2-0294-05

建立在经典测试理论^[1](Classical Test Theory, CTT)下的传统测试是以真分数理论为基础,所有被试尽管水平差异可能很大,但他们都被安排去进行同一套试卷测试。最终导致低水平的被试不能完成高难度的项目;高水平的被试能轻松对低难度的项目作答,这样不但测试不出被试的真实能力水平而且测试出来的不同能力缺乏科学的比较,结果必然失去测试的意义。为了克服传统测试中的不足,采用基于项目反应理论的计算机化自适应测试来更准确地检测被试的真实能力,并将其用来对不同被试之间的能力进行比较。

1 IRT 的基本假设及其优点

项目反应理论^[2]早期也称潜在特质理论(Latent Trait Theory, LTT)。它是针对经典测试理论的不足提出的一种新的测试理论,其最大的优点是项目参数的估计与被测试的样本无关,即测试中的各种项目参数与被测试对象无关。各种测试理论都有其建立的基本假设,与 CTT 相比 IRT 是建立在强假设的基础上的,它的 4 条基本假设为:① 测验的潜在空间(Latent Space)单维性。通常被试对测验的全部项目的反应会涉及到他的 n 项潜在特质或能力构成了 n 维的潜在空间。IRT 假设该测验只测试被试的一项特质能力的特性称为单维性。换句话说,在所有影响被试反应的因素($\theta_1, \theta_2, \dots, \theta_n$)中,仅有该测验所要测量的能力或特质一个因子占据着主导地位,决定着该测试。② 测验项目间的局部独立性^[3]。主要指某一确定的相同能力或特质水平的被试对不同测试项目的反应在统计上是独立的,他们对一个测验项目的反应不受其它测验项目反应情况的影响。事实上,单维性假设和局部独立性假设是等价的。③ 项目特征曲线^[2](Item Characteristic Curve, ICC)模型假设。指被试对测验项目的正确反应概率与该项目所对应的能力水平之间的函数关系的具体形式的特定假设,通常称之为 IRT 模型。④ 测验的不限时性假设^[4]。就是在不限制时间的情况下进行的一种测试,假如被试有未作任何反应的测试项目就认为其能力不够并将该项目当作答错来处理。

项目反应理论在其建立之初就有许多优点,主要简述如下:首先、项目参数估计的不变性。也就是对测试项目参数的估计与被试样本没有相关性。其次、能力参数估计的确定性。即

* 收稿日期:2011-10-10

基金项目:国家自然科学基金资助项目(60903131);民族教育信息化教育部重点实验室项目;教育部科学技术研究重点项目(210210);云南省高校教育资源智能信息化科技创新团队项目;云南省学术技术后备人才项目。

作者简介:杨跃诚(1978-),男,云南人,硕士生,主要从事计算机远程教育方面的研究。

通讯作者:孙瑜(1974-),女,博士,教授,主要从事人工智能方面的研究。

对被试能力参数的估计和测验中采用的具体测试项目无关,各个测试项目对整个测验的贡献相互独立,不因测试项目的难易程度而改变。第三、提供了测验信息函数作为对被试能力水平参数估计量精度的客观指标。在项目反应理论中,一种测验分数的信息函数越大,则其对应的能力或特质水平的估计值就越精确。测验信息函数替代了 CTT 中的“信度”概念,在 CTT 中,测验信度一般无法精确估计。第四、建立了非线性模型。IRT 将被试对项目的反应与其潜在特质能力之间用统计概率的非线性模型表示。最后、对测验项目的分析与设计,能力参数估计及其测验等值和误差分析等问题给出了系统化的解决办法^[5]。

2 IRT 模型及其参数意义

根据所处理的测验数据类型通常将 IRT 模型分为 3 类。分别为二级评分 IRT 模型、多级评分 IRT 模型和连续型 IRT 模型。由于二级评分模型的评分方式相对简单(反应错误记 0 分,反应正确记 1 分),实用性较强,所以目前较为普遍使用。二级评分模型主要适用于对客观项目的测试评分,而多级评分模型则更适用于对主观项目的测验评分。目前二级评分的 IRT 模型也有多种,其中以 1957 年伯恩鲍姆提出的逻辑斯蒂(Logistic)模型应用最为广泛。

项目反应理论将被试对测试项目的反应(应答),用测试项目特性的项目参数和被试者能力的能力参数及其组合的统计概率模型表示^[6]。由于逻辑斯蒂模型与洛德(Lord)1952 年提出的正态曲线模型(Normal ogive model)极为近似,因此也有单参数,双参数和三参数 3 种模型,它们对应着不同的项目特征曲线。各种模型的特征函数为:

- (1) 单参数模型: $P(\theta) = 1/(1 + e^{-D(b-\theta)})$;
- (2) 双参数模型: $P(\theta) = 1/(1 + e^{-Da(b-\theta)})$;
- (3) 三参数模型: $P(\theta) = c + (1 - c)/(1 + e^{-Da(b-\theta)})$.

其中: $D = 1.702$ (常数)。 θ 为被测者能力值或潜在特质水平(一般取值在 $-4 \sim 4$ 之间)。 a 为项目的区分度参数,它反映测试项目对被试应答能力的区分程度,也即拐点处特征曲线的斜率,其值的大小反映了应答概率变化的快慢,它正常的取值范围在 $0 \sim 2$ 之间。 b 为测试项目的难度系数。 c 为项目的猜测系数,表示由于某种推论、猜测等偶然因素而对该测试项目做出正确应答的概率,对应特征曲线在 $P(\theta)$ 上的截距值($0 \leq c \leq 1$)。值越大,则表明即便是能力水平很低的被试也都有可能答对该项目。 $P(\theta)$ 为能力为 θ 的被试答对某项目的概率,称为项目反应函数(Item Response Function, IRF)。

三参数的逻辑斯蒂(Logistic)模型适用于对选择题和是非题之类需要考虑猜测系数的测验,而对于形如简答题和计算题这类的主观题测验,需要应用多级评分 IRT 模型,由于其评分和计算统计相对复杂,所以本论文的论述暂不涉及。三参数模型的项目特征曲线如图 1 所示。

很明显,在图 1 所示的三参数模型中,由于加入了猜测系数 c ,使得项目特征曲线的下限值不等于零而是该项目的猜测系数,项目特征曲线的上限值将随着能力值的不断增强而接近于 1。同样,根据其模型的特征函数计算可知,当能力参数 θ 等于该项目的难度参数 b 时,被试对该项目的应答概率为 $(1+c)/2$,对应于曲线上坐标点为 $(b, (1+c)/2)$ 称为项目特征曲线的拐点。如果忽略猜测因素的影响,则被试答对项目和答错项目的概率 $P(\theta)$ 均等为 0.5。图 1 中可以清楚地看到,被试对测试项目的应答概率 $P(\theta)$ 随着被试能力 θ 的增加而成单调递增的趋势。项目特征曲线形状是一条以拐点为中心的 S 形曲线^[7]。

因为伯恩鲍姆提出的 Logistic 模型形式简洁,更具数学模型的优点;同时对模型参数和能力参数的估计比较方便。因此,在实际应用中大多采用逻辑斯蒂模型。

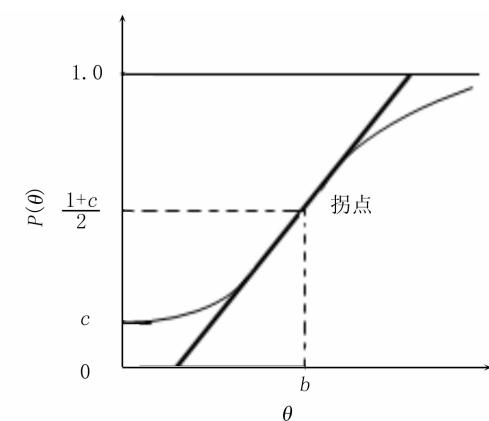


图 1 三参数模型项目特征曲线

Fig. 1 Item characteristic curve of three parameter models

3 CAT 工作原理与测试过程

计算机技术和网络技术的发展,使得应用基于IRT的计算机化自适应测试系统构建成为现实。自适应测试的目的就是通过应用测验分数来对被试的真实能力或特质水平作出有效估计。

整个CAT的工作原理和测试过程都建立在IRT基础上,包括系统的题库建设、测验所用的项目参数和能力参数估计、挑选项目和对被试反应的评分,这一切都在IRT理论的协调指挥下进行。其设计思想可简述为:首先,系统对被试初始能力值作一个粗略的估计(θ_0),然后以此估计值来从题库中挑选难度合适的项目供被试作答。其次,根据被试的反应(应答)情况,系统应用恰当的能力估计方法来估算被试的能力水平(θ_0)。再次,系统又依据最近估计的能力值来应用恰当的选题策略继续从题库中挑选与此能力匹配的项目供被试作答。接下来便是不断的考生能力估计和选题作答过程的反复,直到考生能力值的精度达到测试终止的条件时,此时的能力估计值($\Delta\theta$)即为此被试的真实能力值。实施CAT系统的整个测试过程需要完成2个方面的重要任务:

第一,对被试初始能力值的估计。通常在被试没有接受过任何测试之前,一般地,认为其能力为中等进而从题库中选择中等难度的试题作为测试的起点。反之,则利用以前测试的记录作为本次测试的开始点。在本文中,采用先让被试做一次经典测试理论下的计算机无纸化测试,得到被试一个初步能力值记为 b_s ,再将此初值(b_s)和项目库中的中等难度值(b_m)求它们的平均值 $\bar{\theta}$,将此平均值作为自适应测试的初始能力值。这样做的目的主要是尽量减小其初始值估计的误差。因为初始值估计是整个CAT测试过程中一个重要环节,其初值设计的好坏将直接影响对被试能力参数估计的精确程度。

第二,逐步修正被试的能力值。应用恰当的选题策略,根据当前的能力估计值,去题库中选择最合适的项目来进行测试,并不断估计能力和选取项目,直到对被试能力作出合理的估计。通常利用极大似然估计法或贝叶斯方法来估计项目的特征参数和被试者能力值,而应用极大信息量方法来作为CAT的选题算法。因为只有项目的信息量达到最大时,项目的难度才跟被试者的能力相匹配,能力估计的精度最好。

第三,终止条件的选择。终止条件的选择方法有多种,本文采用将固定测试长度法和比较被试能力的2次连续估计值^[8]结合起来。首先比较2次连续估计值,如果它们的差值小于某个预先指定的 ε 值时测试就终止;否则当测试达到一定的长度时强行终止。这样做的好处是既保证了测试能力的相对精度,又避免了测试过程的永无止境。

CAT对被试能力参数的估计:正确估计考生能力是CAT得以顺利进行的前提,通常用极大似然函数来估计能力参数值。首先需要构造似然函数:假设能力为 θ 的被试对项目*i*的反应为一随机变量 u_i (正答 $u_i = 1$,误答 $u_i = 0$),则矩阵 $U(u_1, u_2, \dots, u_n)$ 表示被试对这*n*个测试项目的反应向量, P_i 为正答概率, Q_i ($Q_i = 1 - P_i$)为误答概率,那么被试对测试项目*i*的反应为 U_i 的概率可用 $L(U_i | \theta, a, b, c)$ 来表示。根据局部独立性假设,得联合概率为^[2]:

$$L(U_i | \theta, a, b, c) = \prod_{i=1}^n P_i^{u_i} Q_i^{1-u_i}, \quad (1)$$

通常称(1)式为似然函数,求得似然函数达到最大值时的 θ 值作为被试能力参数的估计值。因为 L 和 $\ln L$ 有单调联系性,所以它们将在同一 θ 点上同时达到极值。为简化计算,将似然函数转化为(2)式对数似然函数^[6]并求其极值。

$$\ln L = \sum_{i=1}^n \{ u_i \ln(P_i(\theta)) + (1 - u_i) \ln(Q_i(\theta)) \}, \quad (2)$$

求对数似然函数 $\ln L$ 在参数 θ 上的一阶导数,使得导数值为零时的能力参数值即为最大似然估计值 θ_{\max} ,如方程(3)所示:

$$f(\theta) = \frac{\partial \ln L}{\partial \theta} = 0, \quad (3)$$

运用Newton-Raphson迭代(*N-R*)公式求解非线性方程(3)的 θ 值,一般靠计算机来完成。*N-R*迭

代公式为(4) 式:

$$\theta_{k+1} = \theta_k - \frac{f(\theta_k)}{f'(\theta_k)}, \quad (k = 0, 1, 2, 3, \dots). \quad (4)$$

首先确定一个初值 θ_0 , 然后根据迭代方程逐个计算 $\theta_1, \theta_2, \dots, \theta_n$. 直到 $|\theta_n - \theta_{n-1}| < \varepsilon$ 时迭代结束, θ_n 即为所求方程 $f(\theta) = 0$ 的解. 其中, ε 是预先指定的一个正数.

CAT 的选题策略: 根据 CAT 所估计的能力水平, 在对项目库中的测试项目进行检索时, 应基于项目的信息函数 $I_i(\theta)$ 进行选择. 一般总是选择那些 $I_i(\theta)$ 最大的项目, 因为它对被试有最好的分辨能力. 项目的信息函数^[6] $I_i(\theta)$ 与被试能力的关系如(5) 式.

$$I_i(\theta) = \frac{P_i'^2(\theta)}{P_i(\theta) \cdot Q_i(\theta)} = \frac{D^2 a_i^2 (1 - c_i)}{\left[c + e^{Da_i(\theta - b_i)} \right] \cdot \left[1 + e^{-Da_i(\theta - b_i)} \right]^2}. \quad (5)$$

由于项目信息量具有可加性, 所以由个项目组成的测验信息函数^[6] 为各个项目的信息函数之和记为如(6)式. 测试信息函数表示了对各种不同能力的被试者, 测试整体(不是其中某个项目)的测定精度.

$$I(\theta) = \sum_{i=1}^n \frac{P_i'^2(\theta)}{P_i(\theta) \cdot Q_i(\theta)}. \quad (6)$$

根据 CAT 的工作原理, 可画出其测试过程的流程图如图 2 所示.

测试开始, 系统弹出“是否首次测试?”询问窗口, 要求被试选择测试入口, 假如被试接受初次测试, 则先完成一阶段的计算机无纸化测试求其初步能力 b_s 值, 接着系统计算被试初始能力值 $\bar{\theta}$. 随后系统抽取难度为 $\bar{\theta}$ 的项目供被试作答并根据答题情况进行能力评估, 在结束条件的控制下对能力进行循环测试. 最终得到一个满意的被试能力估计值($\Delta\theta$).

4 CAT 系统设计

系统采用浏览器/服务器(Browser/Server)结构的方法实现 Web 环境下的计算机自适应测试系统(Computerized Adaptive Test System, CATS)的设计. CATS 要完成题库管理和考试管理以及对系统的后台管理三部分系统功能. 服务器上系统界面采用.NET 环境下的 C#语言编程(如: Visual Studio 2008), 并用 SQL Server2005 实现数据库搭建. 数据库主要完成学生基本信息表、管理员和教师基本信息表、试题信息表、考试情况记录表、成绩记录表和考试记录表的量化设计工作. 题库管理包括题库的建设和动态维护两部分功能. 对于题库的建设主要完成试题开发、试题的试测、试题的参数估计及其组织方式等四部分主要内容. 在题库开发时要注意题量和题型的选取以及试题与知识点的分布情况, 应尽量组织专家进行命题或到一些试题库中选取较为优秀的试题. 系统运行过程中要设有能不断进行题库扩充、删除和更新等动态维护的功能. 考试子系统提供用户通过浏览器完成在线测试工作. 设有基础练习、模拟测试、试题评分以及试题的导出功能. 系统管理完成对用户的增加和删除以及其身份验证等管理工作.

5 结束语

Web 环境下的计算机自适应测试(CAT)系统, 在未来尤其是在我国推行素质教育的背景下, 它将成为我们对学习程度和教学效果的重要检测系统. 经过多年的发展和实际应用证明了它比经典的测试效果要优越. 它完全根据测试者的能力进行有效测试, 不受测试时间上的限制, 这样既能提高被试者自主学习

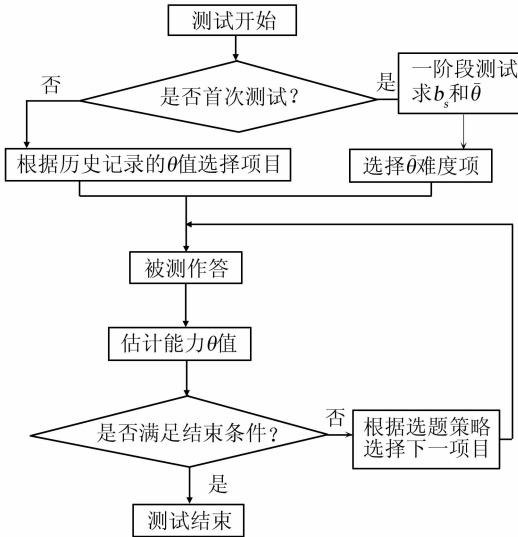


图 2 CAT 测试过程流程图

Fig. 2 CAT test procedure flow chart

的积极性,又不挫伤其接受测试的信心,是个性化学习下的一种理想测试系统。但是CAT系统在项目初始参数的确定,被测者初始能力值的估计和项目选题策略上均有待进一步发展完善,这也是我下一步所要研究的重点。

参考文献:

- [1] 王守佳.移动学习中考试系统的研究[D].长春:东北师范大学,2009.
- [2] 许祖慰.项目反应理论及其在测验中的应用[M].上海:华东师范大学出版社,1992.
- [3] 辛涛.项目反应理论研究的新进展[J].中国考试:研究版,2005:18-21.
- [4] 庄晓.项目反应理论在各类考试系统中的应用策略研究[J].西北医学教育,2006,14(1):30-31.
- [5] 郭庆科,房洁.经典测验理论与项目反应理论的对比研究[J].山东大学学报:自然科学版,2000,15(3):264-266.
- [6] 傅德荣,章慧敏.教育信息处理[M].北京:北京师范大学出版社,2009.
- [7] 赵秋.项目反应理论的发展综述及其在教育测量学中的应用[D].长春:东北师范大学,2008,29(1):84-88.
- [8] 高怀勇,金桂林.项目反应理论及其在计算机自适应测试中的应用[J].西华师范大学学报:自然科学版,2008,29(1):84-88.

Research of computerized adaptive test system based on item response theory

YANG Yue-cheng, ZHONG Ru-neng, SUN Yu, XIAO Meng-xiong
(School of Information, Yunnan Normal University, Kunming 650500, China)

Abstract: Computerized adaptive test (CAT) system is an intelligence test system based on Item Response Theory (IRT). It can not only improve test efficiency but also enhance the accuracy of ability evaluation to tester. This article describes the details of IRT, discusses the operational principle and implementation process of CAT as well as the correlative knowledge of system design, proposed a solution to the estimation of initial ability value and optimized the termination conditions of the test. Finally we provided an ideal test method for the personalized learning.

Key words: classical test theory; item response theory; computerized adaptive test