DOI:CNKI:50-1165/N.20120704.1115.023

一种选题策略在自适应的计算机等级考试中的应用*

秦春影，喻晓锋，仝海燕
(亳州师范高等专科 学校计算机系，安徽 亳州 236800)

中图分类号:TP391 **文献标志码**:A **文章编号**:1672-6693(2012)04-0127-04

全国计算机等级考试由教育部考试中心主办，是一种面向社会的计算机应用能力水平考试，为社会提供了一个相对统一、公正和客观的计算机应用能力考核标准[1]。这种考试深受社会各行各业的欢迎和认同，和英语四、六级一样，已成为大学生求职，晋升的重要考核凭证。截至 2009 年，全国计算机等级考试考生人数已经突破 800 万，获得证书的人数已超过 350 万。传统的基于计算机形式的考试形式在很多方面的表现不尽人意，比如测验项目没有针对性，不能做到“因材施测”，测验对于每位考生的测试效力是不相同的，测验的有效性和公平性受到质疑；测验长度偏长，测验效率不高。因此，选用更适合计算机等级考试的考试形式则是关键的一步。已有研究表明，选用自适应形式无论是在测验效率，测验精度，测验的有效性和公平性上都要优于传统形式的计算机等级考试[2-3]。

1 自适应形式的计算机等级考试介绍

自适应形式的计算机等级考试(Computer rank examination with an adaptive form，CRE_AF)也就是采用计算机自适应测验(Computerized adaptive test，CAT)[2-3]形式的计算机等级考试，这是和传统的采用基于计算机的计算机等级考试相对而言的。在传统的计算机等级考试中，计算机在整个考试中所起的作用仅仅局限于题目的呈现和成绩的统计，起到考试“执行者”的作用。由于考试并没有因材施测，考试不能很好地考察学生的真实水平，并生考生抽取试题时存在“运气”，考试的合理性和公平性受到很大质疑。自适应形式的计算机等级考试可以解决传统计算机等级考试固有的弊病，可以实现“因材施测”，提高测验精度，降低测验长度，并且考试的实施和组织更加经济。

2 选题策略

选题策略是 CAT 的重要环节之一，关系到测量准确性、测验安全和测验信、效度[4]。在 CAT 中比较常用的选题策略主要有两种，一种是信息函数最大化策略[2]，另一种是加权离差模型(Weighted deviation model，WDM)[5]。随着 CAT 的不断发展，这两种选题策略表现出了它们的局限性，主要表现在信息函数最大化策略虽然具有较的测验精度和测验效率，但是高区分度项目容易被过度使用，对题库的安全产生威胁；而加权离差模型在实际过程中权值容易受到主观因素的影响导致选题效果不理想，正因为如此，Chang 和 Ying[6-8]提出多阶段 a 分层的方法(Multistage a-stratified method，STR)的选题策略，并在此基础上 Chang 和 Ying 进行了改进，提出了 b 分区 a 分层的多阶段 CAT 选题策略(a-stratified multistage item selection with b-blocking，BAS)[7]，为节省篇幅，本文只对使用到的多阶段 a 分层的选题策略和 b 分区 a 分层的多阶段 CAT 选题策略进行介绍。

1)多阶段 a-分层的选题策略。多阶段的 a-分层的选题策略基本思想是在测验早期使用低区分度

* **收稿日期**:2011-11-10 **修回日期**:2012-01-04 **网络出版时间**:2012-07-04 11:15:00
资助项目:国家自然科学基金(No.31160203；No.31100756)；2010 安徽省自然科学基金(No.KJ2010B123)
作者简介:秦春影，女，讲师，硕士，研究方向为人工智能。
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20120704.1115.201204.127_023.html

的项目,在测验的后期使用更高区分度的项目,这主要是考虑项目的区分度在测验的各个阶段对自适应测验的不同影响。在测验初期,由于对于被试的信息了解很少,能力估计值不精度,此时选用区分度较低的项目施测,有利于控制整个题库中曝光率,并且可以提高测验结果的精度,因为区分度较低的项目的信息函数比较平坦,可以为跨度较宽的能力区间提供差别不大的项目信息量,而区分度较高的项目的信息量函数比较陡峭[9],如果能力值与项目难度值相差较大,则项目信息量急剧下降。整个测验中项目的选择分为 K 个阶段。

具体的操作过程如下:

第一步,按照项目的区分度的大小把题库中所有的项目分为 K 层,第一层和最后一层分别包含区分度最小和最大的项目;

第二步,与题库分为 K 层相对应,整个测验的选题过程也分为 K 个阶段;

第三步,在第 k 个阶段,从题库的第 K 层选择 n_k 个项目。在测验中,根据项目与被试能力估计值最接近原则选取下一个要施测的项目;

第四步,重复第三步,直到达到预定的测验精度。

2)b 分区 a 分层的多阶段选题策略。多阶段的 b 分区 a 分层选题策略是对 a-分层的多阶段选题策略的改进,其不同之处在于题库的每一层都有一个难度参数 b 的平衡分布来保证去匹配不能被试的能力值 θ[7-9]。

多阶段的 a 分层选题策略是仅考虑区分度参数 a 的分层方法,其中,项目的区分度参数 a 和项目的难度参数 b 之间存在相关性使得在单纯的多阶段 a 分层的选题策略中的 b 值分布不均匀;在项目区分度相同的情况下,项目的难度与被试的能力之间越接近,项目信息量越大,越有利于测出学生的能力值,b 分区 a 分层的多阶段选题策略正是基于这个思想而提出来的。相对于多阶段的 a 分层选题策略,b 分区 a 分层的多阶段选题策略有更高的测验精度和测验效率,后面的实验也验证了这一点。具体操作过程如下:

第一步,根据项目难度值的大小,把整个题库划分为 M 个区(Blocking),可以使每个区都含有相同数量的项目,也可以使每个区含有不同数量的项目。其中第一区包含最低难度的项目,第 M 区包含最高难度的项目。

第二步,根据项目的区分度的大小,把每一个区都分别划分为 K 层。这样,对于第 M 区来说,第一层包含区分度最小的项目,第 K 层包含区分度最高的项目。

第三步,对于 $k=1,2,\cdots,K$,在 M 个区之间把第 k 层的项目重新组合成一个单独的层。现在一共有 K 层。

第四步,把测验分为 K 个阶段。

第五步,在第 k 个阶段,从 k 层中选择与被试的能力估计值最接近的项目进行施测。

第六步,当 $k=1,2,\cdots,K$ 时,重复步骤五,直到达到预定的测验精度。

3 CRE_AF 中选题过程的模拟

由于在实际应用过程中,信息函数最大化策略和加权离差模型选题策略均表现出了局限性,自适应形式的计算机等级考试中采用 b 分区 a 分层的多阶段选题策略。

为了将注意力集中到新形式的计算机等级考试的测验过程中来,并且在测验中被试的能力真值和项目参数真值是不可能知道的,因此使用 Monte Carlo 模拟的方法来进行实验。模拟的过程如下。

3.1 模拟被试和题库

1)产生被试(即产生一批被试参数),假定被试的能力参数服从标准正态分布,即 $\theta\sim N(0,1)$。模拟 500 个被试。

2)产生项目(即产生一批项目参数,这里对项目只考虑难度和区分度,暂不考虑猜测度,并且全部是 0-1 计分的项目),假定项目区分度参数服从对数正态分布,项目难度参数服从标准正态分布,即 $\ln a\sim N(0,1)$,$b\sim N(0,1)$,模拟 1 000 个项目。

3.2 施测过程

由于被试和题库均是模拟产生的,并且参数完全服从相应的分布,是属于理想的情况,题库的质量可以得到保证,接下来可以安排被试进行 CRE_AF 测试。选题策略采用 b 分区 a 分层的多阶段选题策略。

首先进行 b 分区,将整个题库根据难度参数平均划分为 5 个区,每个区中的项目数为 200,其中第一区中项目的难度最低,第五区中项目的难度最高。

第二步进行 a 分层,在每一区中,将所有的项目按区分度分度,这里分为 5 层,其中第一层的区分度最小,第四层的区分度最大。

第三步分阶段,将五个区中各自的五层项目分别组合成 5 个阶段,即第一、二、三、四、五区中的第一层组合起来成为测验第一阶段所用的题库,第二层组合起来成为测验第二阶段所用的题库,依次得到测验所有 5 个阶段所用的题库。

第四步分阶段测验,测验中依次使用 5 个子题库(每个子题库有相应的结束条件)进行测验,直到满足整个测验结束的条件,结束测验。

3.3 评价指标

这里采用能力估计的返真性、能力估计的标准差、测验的平均长度、项目调用的均匀性、测验效率、统一量纲作为评价指标[6]。为了使实验结果更稳定,实验重复执行 R 次(本文取 20)。

1)能力估计的返真性。用返真性指标来衡量测量的准确性。返真性指估计的能力值与模拟的真值的绝对差的平均。值越小,测量越准确。计算公式为

$$\text{返真性}=\frac{1}{R}\sum_{i=1}^{R}\left(\frac{1}{N}\sum_{j=1}^{N}\text{abs}(\hat{\theta}_{ij}-\theta_j)\right)$$

其中 θ_j 和$\hat{\theta}_{ij}$分别表示第 j 个被试的能力真值和在 R 次实验中的估计平均值。

2)能力估计的标准差的计算公式为

$$\text{能力估计的标准差}=\frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{\sum_{j=1}^{R}(\theta_i-\hat{\theta}_{ij})^2/R}\right)$$

其中 M 为题库中的项目数,N 为被试个数。

3)测验的平均长度的计算公式为

$$\text{测验的平均长度}=\frac{1}{R}\sum_{i=1}^{R}\left(\sum_{j=1}^{N}r_{ij}/N\right)$$

其中 r_{ij} 是被试 j 在第 i 次测验中作答的项目个数。

4)项目调用的均匀性计算公式为

$$\text{项目调用的平均次数}=\frac{1}{R}\sum_{i=1}^{R}\sqrt{\sum_{j=1}^{M}(m_{ij}-\overline{m}_i)^2/M}$$

其中 M 为题库中总的项目个数,m_{ij} 表示第 j 个项目在第 i 次测验中被使用的次数,$\overline{m}_i=\sum_{j=1}^{M}m_{ij}/M$ 表示项目 j 在 i 次模拟实验中被使用的平均次数。

5)测验的效率计算公式为

$$\text{测验的效率}=\sum_{i=1}^{N}Infor_i/\sum_{i=1}^{N}L_i$$

其中 $Infor_i$ 表示被试 i 在测验中的信息总量,L_i 为被试 i 做答的项目总数。

6)统一量纲。由于前面有 5 个指标,为了更直观地反映实验整体效果,将这些指标进行综合统计,采用统一量纲的方法[10]。具体作法是,对于值越大越好的指标,将该指标上的最大值作分母,把其他选题策略在该指标上的值作分子,求出比值;对于值越小越好的指标,将该指标上的最小作分子,把其他选题策略在该指标上的值作分母,求出比值。将各选题策略指标进行加权求和得到统一量纲后的结果。这个值可以作为该种选题策略的综合表现评分。

4 实验结果及分析

通常情况下,最大信息量选题策略可以达到最佳的测验效率,随机选题策略可以使题库使用最均匀,为了和实验结果进行对比,模拟传统形式的计算机等级考试时分别使用最大信息量选题策略(Maximum information criteria,MIC)、随机选题策略(Stochastic criteria,SC)和多阶段 a 分层的选题策略(STR)将这 3 种选题策略的结果与 b 分区 a 分层的多阶段选题策略(BAS)进行比较。

表 1　CAT_AF 下不同选题策略模拟实验结果比较(实验重复 20 次)

选题策略	能力估计的返真性	能力估计的标准差	项目调用的平均次数	测验的平均长度	测验的效率	统一量纲
MIC	0.199 2	0.223 7	45.36	11.6	1.416	4.163
SC	0.224 7	0.265 9	7.428	46	0.228	3.148
STR	0.201 5	0.247 8	14.47	15.2	1.342	4.212
BAS	0.200 7	0.226 6	12.24	16	1.356	4.329

采用BAS选题策略,能力估计的返真性与MIC法相当接近,表明BAS选题策略会使能力估计的精确性很高,接近MIC法,并且能力估计的标准差也有类似的表现;在项目的使用率上,BAS选题策略提高了低区分度项目的使用率,降低了高区分度项目的使用率,从而使整个题库的使用更加均匀,对项目的曝光率有较好的控制,明显优于MIC法;BAS选题策略在能力估计的返真性,能力估计的标准差,项目调用平均次数,测验长度,测验效率上的表现均好于STR法。BAS选题策略使测验的效率更高,优于STR和SC法。从综合表现上来看,BAS法的统一量纲值最大,也印证了这一点。选择BAS选题策略更加符合CAT的测试过程,值得在实际应用中去验证和使用。

5 展望

采用BAS选题策略,在选题策略的各项指标上表现均比较好,使得测验效率提高的同时,提高了测验的精度,降低了测验的成本,使计算机等级考试更能发挥自身的社会作用,更能体现测验的公平性和合理性。但是,由于研究是在模拟的条件下进行的,各种参数都严格满足相应的假设条件(如能力满足正态分布、项目参数分布等),但是当在实际应用中,这样较强的假设条件不满足时,选题策略的表现如何值得去研究和证实。此外,BAS的具体操作应当如何进行、多级记分下的表现是否也如0-1记分下的表现一致等问题需要进一步探索。

参考文献:

[1] 郭道江.大学生如何应对全国计算机等级考试[J].安徽工业大学学报:社会科学版,2007,24(3):128-129.

[2] 漆书青,戴海崎,丁树良.现代教育与心理测量学原理[M],北京:高等教育出版社,2002:8.

[3] 戴海崎,丁树良,罗照盛.心理与教育统计测量专题研究文集(1995—2004)[M].南昌:江西科学技术出版社,2005:3.

[4] 毛秀珍,辛涛.计算机化自适应测验选题策略述评[J].心理科学进展,2011,19(10):1552-1562.

[5] 张忠华,谢小庆.国外计算机自适应测验选题策略的研究[J].中国考试,2004(7):18-21.

[6] Chang H H,Ying Z L. A-stratified multistage computerized adaptive testing[J]. Applied Psychological Measurement, 1999,23(3):211-222.

[7] Chang H H,Qian J H,Ying Z L. A-stratified multistage computerized adaptive testing with b blocking[J]. Applied Psychological Measurement,2001,25(4):333-341.

[8] Stocking M L,Swanson L. A method for severely constrained item selection in adaptive testing [J]. Applied Psychological Measurement,1993:23,277-292.

[9] 鹿士义,张坚.题目曝光控制的动态a分层方法[J].中国考试,2011(9):3-9.

[10] 陈平,丁树良,林海菁,等.等级反应模型下计算机化自适应测验选题策略[J].心理学报, 2006,38(3):461-467.

(责任编辑 方 兴)