心 理 学 报 2008,**40**(6):737 ~ 747
Acta Psychologica Sinica

允许检查并修改答案的计算机化自适应测验*

陈 平[1,2] 丁树良[2]

([1]北京师范大学发展心理研究所,北京 100875) ([2]江西师范大学计算机信息工程学院,南昌 330027)

摘 要 采用计算机模拟程序对允许检查并修改答案的计算机化自适应测验(CAT)进行研究,并采用新的评分方式对付 Wainer 策略。结果表明:综合考虑被试的两次作答信息可以得到更精确的能力估计值。大部分被试进行了修改,只有少部分答案被修改,在被修改的答案中大部分是由错误改为正确;综合 Wainer 策略 CAT 的后验分布期望值(EAP)和极大似然估计值(MLE)可以"粗糙"对付 Wainer 策略。

关键词 计算机化自适应测验,项目检查,Wainer 策略,蒙特卡洛模拟。

分类号 B841.7

1 引言

1.1 问题提出

计算机化自适应测验(Computerized Adaptive Testing,CAT)依靠大型题库,采用项目反应理论(IRT),自行去适应被试水平,灵活施测难度最恰当而且性能最优的项目,从而实现对被试的高效测量[1]。目前,CAT 在心理教育测量领域已经得到非常广泛地应用。然而在 CAT 的所有应用中,只有一项允许被试检查并修改答案,即由美国临床病理学家协会的注册委员会(Board of Registry,简称 BOR)开发的资格考试[2,3],大多数 CAT 都不允许被试检查并修改答案。

检查并修改答案的问题,对参加测验的被试及组织测验的考试机构来说都非常重要。首先从被试的角度来看,大多数 CAT 不允许被试返回检查并修改答案,被试在纸笔(paper & pencil, P&P)测验中常用的答题策略(比如,有些被试从头答到尾,答完之后再检查,发现错误就修改;其他的一些被试,不能确定答案的项目先放着,答完其它项目之后再返回作答;等等)就不能应用到 CAT 中,这样会给被试带来压力。另外,对于不允许修改答案的 CAT,若某个被试完全有能力答对某个项目但笔误了,由于不允许修改答案,他/她的能力会被低估。相反,若某个被试没有能力答对某个项目却猜对了,又不允许修改,他/她的能力会被高估。这时,不允许修改答案的 CAT 是有失公平的。如果允许修改,修改的结果可以更真实地反映被试能力水平。而且,CAT 的自适应算法本身就决定了:被试只能答对一半左右的项目,这同样会增加被试的焦虑水平并影响被试的测验表现[3]。所以在参加 CAT 时,被试都希望得到检查并修改答案的机会;其次从考试机构的角度来看,大多数 CAT 不允许被试检查并修改答案有很多原因[3,4],但主要是担心:一些"聪明"的被试或"聪明"的备考机构所指导的被试通过使用 Wainer "作弊"策略获得能力高估值,影响测验的公平性、公正性和准确性。Wainer 策略由 Wainer 于 1993 年提出[5],是被试在 CAT 测验过程中可能采用的一种"作弊"策略:被试在自适应阶段故意错误作答所有项目,根据得分向量计算的能力估计值会越来越低,选题策略选出的项目会越来越容易,整个测验就会相对容易。之后在检查修改阶段,他/她尽全力作答。与传统的 CAT 相比,该被试很可能答对更多的项目。而且,如果被试的真实能力水平充分大于初始项目的难度水平时,他/她就很有可能答对所有的项目。

显然,CAT 中不允许检查并修改答案相对于 P&P 来讲是一种缺陷,而允许修改答案的 CAT 又有

收稿日期:2007 - 08 - 27

* 国家自然科学基金项目(60263005)、江西省教育厅科技项目(GJJ08154)、江西省高校人文社科研究项目(JY06201)、全国教育考试科研规划项目(2006JKS3063)和卫生部课题(KY200704,JM20060070)资助。

通讯作者:丁树良,E-mail:ding06026@163.com,电话:0791 - 8120410

可能为使用 Wainer 策略的被试提供可乘之机，那么是否存在某种解决方案，既使 CAT 允许检查并修改答案，又让那些使用 Wainer 策略的被试不致于“占便宜”——得到不合理的能力高估值呢？我们认为，这对 CAT 的发展是一个有意义的论题，理应成为学者们关注的焦点。目前关于这方面的研究，国内刊物未见报导，国外研究也还不多。已有研究总结如下：BOR 深入研究检查并修改答案的问题，在其 17 种资格测验中增设了检查修改机制[2]。BOR 为了防止被试使用 Wainer 策略，被试每作答完一个项目，测验就会计算他/她的正确作答率。当被试的正确作答率较低（比如说小于 30%）时，测验就自动根据标准参照点（pass point）选择最具信息量的项目，而不是根据当前的能力估计值选题，有效对付了 Wainer 策略[3]；Stocking（1997）采用模拟数据和 CAT 版本的 GRE 数据评价 Wainer 策略的有效性，并与其它三种允许或部分允许检查修改的备选模型进行了比较，发现通过限制修改机会可以有效对付 Wainer 策略[6]；Papanastasiou（2002）在研究中尝试着提供一种新的方法——“重新安排”方法，即为了更好地估计被试的能力重新安排并跳过一些项目，以防止被试使用 Wainer 策略。结果表明“重新安排”方法有是有效的，它能够降低贝叶斯估计的标准误，并且能够提高能力估计的可靠性[7]。另外，Gershon 和 Bergstrom（1995）、Vispoel（1999）等人就什么情况下 Wainer 策略会获得成功、成功使用 Wainer 策略对能力估计值及能力估计标准误的影响如何、如何判断被试使用了 Wainer 策略等等问题进行了深度的研究[2,8]。

然而，纵观已有研究，我们发现研究者在估计被试能力时，仅仅记录了修改后的答案并仅根据该记录评分，而对修改前的原始作答信息不予考虑。我们认为：其一，对于允许修改答案的 CAT，被试有可能第一次反应正确，修改后变得不正确。也有可能第一次反应错误，修改后变正确了。这与修改前后都正确或错误的那些被试的能力理应有所不同，所以对允许修改答案 CAT 的作答信息的收集不能只看结果而不看过程；其二，计算机化测验相对于 P&P 最大的优点之一就是，计算机能够记录被试的所有作答信息，不仅能够记录被试在自适应阶段的作答信息，而且还能够记录被试在检查修改阶段的修改信息。如果将被试尽可能多的信息（修改前后作答的信息）综合起来考虑，是否会获得更加精确的能力估计值呢？针对这个问题，本文提出新的评分方式及其模型，并根据相应估计试图对付 Wainer 策略。

1.2 模型提出

Javier Revuelta 提出三种项目修改的方案：作答完所有项目之后进行修改（RE）；每作答完一个项目块（比如说五个项目）进行一次修改（RB）；每作答完一个项目就进行一次修改（RI）[9]。本文只考虑 RE 修改方案，即被试作答完所有的项目之后，才允许其检查并修改答案。

被试作答完所有的项目之后再进行修改，可以理解为每个被试对每个项目作答两次。被试在自适应阶段的作答记为第一次作答，在检查修改阶段的作答记为第二次作答。如果没有进行修改，即确认第一次作答的答案为第二次作答的答案。本文只讨论长度为 m 的定长 CAT，只考虑 0 - 1 评分的情况，并选用 3 参数逻辑斯蒂克模型（3PLM），其项目特征函数表示如下：

$$P_{\alpha j} = c_j + \frac{1 - c_j}{1 + \exp[\,-D \times a_j \times (\theta_\alpha - b_j)\,]} \qquad (1)$$

（1）式中 $P_{\alpha j}$ 表示被试 α 答对第 j 个项目的概率，θ_α 表示被试 α 的能力值，a_j、b_j 和 c_j 分别表示第 j 个项目的区分度、难度和猜测度。本研究中量表因子 D 取 1.70。

我们假设第一次作答正确给 Beta 分（Beta $\in$ [0,1]），第二次作答正确给（1 - Beta）分。也即对第一次作答反应赋权值为 Beta，对第二次作答反应赋权值为（1 - Beta）。特别地，当 Beta = 0 时，只考虑第二次作答结果；当 Beta = 1 时，只考虑第一次作答结果。本文用 $U_{\alpha j}$ 表示被试 α 在第 j 个项目上的第一次作答反应，用 $U_{\alpha,j+m}$ 表示被试 α 在第 j 个项目上的第二次作答反应（$U_{\alpha j} \in \{0,1\}$，$U_{\alpha,j+m} \in \{0,1\}$，$j = 1 \ldots m$）。于是本文提出的新的评分方式可以记为：

$$V_{\alpha j} = Beta \times U_{\alpha j} + (1 - Beta) \times U_{\alpha,j+m} \qquad (2)$$

（2）式中 $V_{\alpha j}$ 表示被试 α 在第 j 个项目上两次作答的综合反应（$V_{\alpha j} \in \{0, Beta, 1 - Beta, 1\}$）。类似于 Lord 针对参数估计过程中的缺失数据而提出的新函数[10]，本文将综合得分矩阵 $V = [V_{\alpha j}]$ 的“似然函数”记为：

$$L = \prod_{\alpha=1}^{N} \prod_{j=1}^{m} P_{\alpha j}^{V_{\alpha j}} Q_{\alpha j}^{1 - V_{\alpha j}} \qquad (3)$$

（3）式中 N 为被试数，$Q_{\alpha j}$ 表示被试 α 答错第 j 个项目的概率。与（3）式对应的对数似然函数可记为：

$$\ln L = \sum_{\alpha=1}^{N}\sum_{j=1}^{m}(V_{\alpha j}\times \ln P_{\alpha j} + (1 - V_{\alpha j})\times \ln Q_{\alpha j}) \quad (4)$$

特别值得注意的是:传统的评分方式下,被试的项目得分要么为 0 要么为 1。而在这里,当 Beta 取 0 到 1 之间的小数时,被试的综合得分 $V_{\alpha j}$ 就有可能取 0 到 1 之间的小数(Beta 或 1 - Beta),似然函数 L 公式中的指数就有可能会出现 0 到 1 之间的小数(Beta 或 1 - Beta)。由此得到的能力估计值是否会出现异常呢? 如果能够得到正常值,那么 Beta 取什么值时,能力估计最精确呢?

另外,如果我们将对同一批项目的两次作答看成是作答了两批相同的项目,那么相应的似然函数可以表示如下(假设能力给定的条件下,对同一个项目的两次作答是局部独立的):

$$L = \prod_{\alpha=1}^{N}\prod_{j=1}^{2\times m} P_{\alpha j}^{U_{\alpha j}} Q_{\alpha j}^{1-U_{\alpha j}} \quad (5)$$

(5)式中 $U_{\alpha j}(j = 1,\ldots,m)$ 表示被试的第一次作答反应, $U_{\alpha j}(j = m+1,\ldots,2\times m)$ 表示被试的第二次作答反应。$P_{\alpha j}$ 和 $Q_{\alpha j}$ 中的 a_j、b_j 和 $c_j(j = 1,\ldots,m)$ 与 a_j、b_j、$c_j(j = m+1,\ldots,2\times m)$ 分别对应相等。

为了方便,本文记(3)为模型 1,记(5)为模型 2。模型 1 和模型 2 估计的未知参数,哪个更精确呢? 采用模型 1 和模型 2 估计被试能力,能否对付 Wainer 策略呢?

针对以上疑问,本研究进行蒙特卡洛(Monte Carlo)模拟实验。评价指标是对未知参数的修复(Recovery)能力,修复能力越强的方法或者模型就越好。而衡量能力估计的修复能力, 通常用绝对偏差平均 *ABS* 和偏移均方根 *RMSD* 两个指标。

绝对偏差平均 *ABS* 的计算公式如下:

$$ABS = \sum_{j=1}^{C}(\sum_{i=1}^{N}|\theta_i - \hat{\theta}_{ij}|/N)/C$$

式中 N 表示被试人数,C 表示实验独立重复模拟的次数。θ_i 为被试 i 的能力真值,$\hat{\theta}_{ij}$ 代表被试 i 在第 j 次模拟中得到的能力估计值。因此, *ABS* 指标反映了能力真值与能力估计值的绝对偏差的平均。*ABS* 值越小,能力估计准确性越高。

偏移均方根 *RMSD* 的计算公式如下:

$$RMSD = \sqrt{\sum_{j=1}^{C}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_{ij})^2/(C\times N)}$$

式中 N、C、θ_i 以及 $\hat{\theta}_{ij}$ 的定义同上。RMSD 值也是越小越好。

2　实验一

本实验采用 Monte Carlo 方法模拟允许检查并修改答案的 CAT,以考察模型 1 中不同的 Beta 值(本文取 Beta = 0,0.10,0.20,0.30,0.382,0.40,0.50,0.60,0.618,0.70,0.80,0.90,1)对能力估计的影响,并比较模型 1 和模型 2 的能力估计精确性。本实验假设被试在作答过程中没有使用 Wainer 策略,并记 $N(x,y)$ 表示期望为 x、方差为 y 的正态分布。

2.1　方法

2.1.1　被试和题库　计算机首先从标准正态分布中抽取 1000 名被试的能力真值($\theta \sim N(0,1)$)。然后模拟生成 1000 个项目形成题库,项目区分度 a 和难度 b 的分布分别为对数标准正态分布($lna \sim N(0,1)$)和标准正态分布($b \sim N(0,1)$),所有项目的猜测度 c 都定为 0.15。能力真值 θ 和 b 均介于 -3 至 3 之间,a 介于 0.20 至 2.50 之间。

2.1.2　测验长度和终止规则　为了方便,本实验只考虑长度为 30 的定长 CAT。也即测验长度达到 30,结束测验。

2.1.3　模拟过程　本实验固定被试和题库,将允许检查并修改答案的 CAT 全过程独立重复模拟 30 次,以尽可能的减小随机误差。由于本实验是考察模型 1 中 13 个不同的 Beta 值对能力估计的影响,所以每个被试在每次实验中都有 13 组综合得分反应向量(每组得分反应向量对应于一个 Beta 值),也就有 13 个能力估计值。实验结束后,相应地就有 13 组评价指标值(ABS 和 RMSD)。通过判断这 13 组指标值的大小,我们就可以判断模型 1 中 Beta 取何值时能力估计的修复能力最强。另外,采用模型 2 估计被试能力也可以得到且只能得到 1 组评价指标值,因为模型 2 中不包括 Beta 值。

允许检查并修改答案的 CAT 模拟全过程分为三个阶段:探测性阶段、正式测试阶段和检查修改阶段,相对于检查修改阶段,前两个阶段可以统称为自适应阶段。在探测性阶段,由于对被试的能力一无所知,所以对每个被试赋能力初值大约为均值水平(零附近),并根据选题策略选择最适合的项目给被试进行模拟作答。并采用固定步长法[11](假设步长为 0.20,如果正确作答,新的能力估计值为原值加上 0.20;如果错误作答,新的能力估计值为原值减去 0.20)估计被试能力,直到作答项目数不少于 3 且作答总分既不为 0 也不为满分时,结束探测性阶段。根据被试的作答得分向量,调用能力估计子程序求得更加精确的能力初估值,进入正式测试阶段。在正式测试阶段,由选题策略调用难度值与被试当

前能力估计值相匹配的项目施测被试。被试每作答完一个项目,就更新一次能力值,直到测验长度达到预设长度,结束正式测试阶段[12]。最后是检查修改阶段,被试对在自适应阶段已经作答的项目进行修改,并且只允许修改一次。

自适应阶段被试作答模拟称为第一次作答模拟。模拟方法为:根据被试 α 的能力真值 $\theta\alpha$ 和当前项目 j 的项目参数,计算被试 α 答对该项目的概率 P,再生成(0,1) 区间上的随机数 r。若 $P \geq r$,则认为被试 α 在该项目上得分 $U_{\alpha j}=1$;否则,认为被试 α 在该项目上得分 $U_{\alpha j}=0$。

检查修改阶段被试作答模拟称为第二次作答模拟。模拟方法为:根据被试的能力真值 $\theta\alpha$ 和当前项目 j 的项目参数,计算被试 α 答对该项目的概率 P。若 $P \geq 0.67$ 且 $U_{\alpha j}=0$,我们认为被 α 在检查修改阶段有能力把第 j 个项目的答案由错误改为正确(即 $U_{\alpha,j+m}=1$);若 $P<0.33$ 且 $U_{\alpha j}=1$,我们认为被试 α 在检查修改阶段会把第 j 个项目的答案由正确改为错误(即 $U_{\alpha,j+m}=0$);其他的情况下,答案保持不变($U_{\alpha,j+m}=U_{\alpha j}$)。由于模拟被试修改作答比较复杂、困难,本实验只考虑了将答案由错误改为正确或是将答案由正确改为错误的情况,并没有考虑将答案由错误改为错误的情况。

有了以上两次模拟作答,我们就可以利用(2)式求得被试的综合反应向量 V。

2.1.4 能力估计方法 探测性阶段采用固定步长法估计被试能力,步长设为 0.20。之后采用 EAP 或 MLE 方法估计被试能力。在实现 EAP 方法时,假设能力先验分布为标准正态分布。当测验长度不大于 20 时,积分结点数设为 13;测验长度大于 20 时,积分结点数设为 16。这样假设主要是为了方便运算。在实现 MLE 方法时,模拟程序始终将能力迭代值控制在[-3,3]之间,并且规定每次能力估计时迭代循环次数不超过 20 次,能力估计迭代精度定为 0.01。

2.1.5 选题策略 从题库(或剩余题库)中选择难度值与被试当前能力估计值最接近的项目。

2.2 结果

模拟过程中采用 EAP 方法估计被试能力时,模型 1 中不同的 Beta 值所对应的 *ABS* 和 *RMSD* 结果如图 1 所示。我们将模型 2 得出的结果也列入图 1,以资比较。

从图 1 中的变化趋势,可以发现当 Beta = 0.50 时,ABS 和 RMSD 值都最小,能力估计的修复能力最强。也就是说,既考虑被试第一次作答的作答信息又考虑被试第二次作答的作答信息,而且两者都赋以 0.50 的权值加权综合,能够对被试进行最精确地评价。当 0 ≤ Beta < 0.50 时,ABS 和 RMSD 值都随 Beta 值的增大而减小。特别地,当 Beta = 0 时,ABS 及 RMSD 值都达到最大。也即如果完全不考虑被试第一次作答的作答信息,而只考虑被试第二次作答信息的话,对被试能力估计的精确性最差。而当 0.50 < Beta ≤ 1 时,ABS 和 RMSD 值都随 Beta 值的增大而增大。特别地,当 Beta = 1 时,ABS 及 RMSD 值都达到次大值。说明只考虑被试第一次作答的作答信息,而完全不考虑被试第二次作答信息的话,对被试的能力估计也不精确。尽管由模型 2 得出的 ABS 和 RMSD 值比较小,但还是不如模型 1 中 Beta = 0.50 时的结果精确。

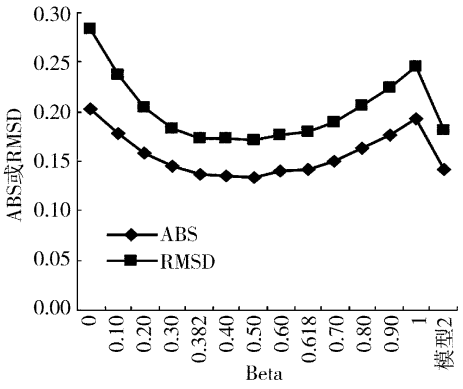

图 1 采用 EAP 方法时,模型 1 中不同的 Beta 值对应的 ABS 和 RMSD

记录被试的所有作答信息并进行统计,我们还可以得到其它的一些结果:(1)66.80% 的被试进行了修改;(2)大约 6.40% 的答案被修改;(3)在被修改的答案中,有 75% 的答案是由错误改为正确,25% 的答案是由正确改为错误。

另外,当模拟过程中采用 MLE 方法估计被试能力时,可以得到类似 EAP 方法的结果。限于篇幅,这里就不一一阐述。

2.3 结论

本实验中,模型 1 中 13 个不同的 Beta 值会对能力估计产生不同的影响。相对于只考虑被试第一次作答的作答信息(Beta = 1)或只考虑被试第二次作答的作答信息(Beta = 0),综合考虑被试两次作答的作答信息(Beta = 0.10,0.20,0.30,0.382,0.40,0.50,0.60,0.618,0.70,0.80,0.90)能够产生更精确的能力估计值,以 Beta = 0.50 时能力估计最为精确。在允许检查并修改答案的 CAT 中,传统的评分

方式是仅记录第二次作答的作答信息并根据该记录评分,而对第一次作答的信息不作考虑(也即 Beta =0 的情况)。本实验认为:按传统方式评分并进行能力估计的精确性最差,这似乎是对传统评分方式的一种挑战。而且,模型 2 得出的实验结果并不比模型 1 的实验结果好。另外,对记录的所有作答信息进行统计得出:大部分被试进行修改;只有少部分答案被修改;在被修改的答案中,大部分答案由错误改为正确,少部分答案由正确改为错误。这与前人的研究结果也是基本一致的[2~4,13]。

实验一通过模拟研究发现,既考虑被试自适应阶段的作答信息又考虑被试在检查修改阶段的作答信息,而且两者都以 0.50 的权值加权综合,能够对被试进行最为精确地评价。接下来在实验二中,我们希望使用这种评分方式及其模型对付 Wainer 策略。

3　实验二

本文将运用了 Wainer 策略的 CAT 记为 Wainer 策略 CAT。

本实验的目的是在新的评分方式下,通过模拟数据试图找到解决 Wainer 策略的办法。本实验主要讨论 Wainer 策略 CAT 得到的能力估计值,通过模拟分析对这些能力值的偏差和估计标准误进行计算。在模拟 Wainer 策略 CAT 的过程中,我们有一个基本假设:所有的被试在自适应阶段故意错误作答所有项目,而在修改阶段尽全力作答。为了验证 Wainer 策略的使用对能力估计的影响,我们将被试在 Wainer 策略 CAT 中的表现与他/她在传统 CAT 上的表现进行了比较。整个实验参考了 Vispoel 在文献[8]中使用的模拟方法。

3.1　方法

3.1.1　被试和题库　假设被试的真实能力 θ 范围是[-3.20,3.20],步长是 0.40,那么共有 17 个能力水平。在每个能力水平下,我们模拟了 100 名被试在 30 个项目上的作答情况(即共模拟 1700 名被试对 30 个项目作答)。采用 Monte Carlo 方法模拟容量为 1000 的题库,且 $\ln\alpha \sim N(0,1)$、$b \sim N(0,1)$,所有项目的猜测度 c 都定为 0.15。b 介于 -4 至 4 之间(有别于实验一),a 介于 0.20 至 2.50 之间。

3.1.2　测验长度和终止规则　本实验考虑长度为 30 的定长传统 CAT 和定长 Wainer 策略 CAT。

3.1.3　模拟过程　为了比较,对每个被试,首先模拟他/她在传统 CAT 上的作答,然后模拟他/她在 Wainer 策略 CAT 上的作答。本文实验一中介绍了允许检查并修改答案的 CAT 模拟过程(包括自适应阶段和检查修改阶段),其中的自适应阶段即本实验所指的传统 CAT,所以本实验中的传统 CAT 模拟过程以及被试作答模拟方法见实验一中所述。另外,由于 Wainer 策略的使用是建立在允许检查并修改答案基础之上,所以 Wainer 策略 CAT 的模拟过程实际上本身就是一种允许检查并修改答案的 CAT 模拟过程。于是,本实验中 Wainer 策略 CAT 的模拟过程与实验一中允许检查并修改答案的 CAT 模拟过程相同,不同之处在于对被试的模拟作答。Wainer 策略 CAT 在自适应阶段模拟被试故意错误作答所有项目(即得分向量为零向量)。而在检查修改阶段模拟被试尽全力作答,方法同实验一中自适应阶段被试作答模拟方法。

考虑到被试在 Wainer 策略 CAT 的自适应阶段故意错误作答所有项目,所以在 Wainer 策略 CAT 的自适应阶段不能采用 MLE 方法估计被试的能力值。为了统一处理,传统的 CAT 和 Wainer 策略 CAT 的自适应阶段都根据 EAP 能力估计值来选题。测验结束后,对传统 CAT 而言,被试的 EAP 能力估计值已经获得,再根据被试的反应向量用 MLE 方法又可以得到他/她的 MLE 能力估计值;而对 Wainer 策略 CAT 而言,根据被试的综合反应向量 V,分别采用 EAP 和 MLE 方法可以得到他/她的 EAP 和 MLE 能力估计值。接着分别计算相应的估计标准误(对 EAP 方法来说,标准误是后验分布的标准差;而对 MLE 方法来说,标准误是测验信息量的倒数的平方根)。于是,模拟结束后每个被试都有 8 个值,其中 4 个来自传统 CAT 的模拟结果,另外 4 个来自 Wainer 策略 CAT 的模拟结果。4 个值中一个是 EAP 能力估计值,一个是 MLE 能力估计值,一个是 EAP 估计标准误,还有一个是 MLE 估计标准误。在模拟 Wainer 策略 CAT 的过程中,被试有两次作答,因此也存在着如何评分的问题。模型 1 中 Beta = 0 是传统的评分方式,而在实验一中发现 Beta = 0.50 时能力估计精确度最高。因此本实验分 Beta = 0 和 Beta = 0.50 两种情况讨论。

3.1.4　能力估计方法　采用 EAP 和 MLE 方法估计被试能力。在实现 EAP 方法时,假设能力先验分布为标准正态分布。当测验长度不大于 20 时,积分结点数设为 17;测验长度大于 20 时,积分结点数设为 21。这样假设主要是为了方便运算。在实现 MLE 方法时,模拟程序始终将能力迭代值控制在

[-4,4]之间,并且规定每次能力估计时迭代循环次数不超过 20 次,能力估计迭代精度定为 0.01。

3.1.5 选题策略 从题库(或剩余题库)中选择难度值与被试当前能力估计值最接近的项目。

3.2 Beta =0 时的实验结果及分析

首先我们来看 Beta = 0 的情况,也即评分时只考虑被试在检查修改阶段的作答信息,不考虑被试在自适应阶段的作答信息(Beta 取何值主要是针对 Wainer 策略 CAT 而言,与传统 CAT 无关)。

3.2.1 EAP 结果及分析 Beta = 0 且采用 EAP 方法时,能力估计的偏差及相应的估计标准误见图 2 所示。

(1)偏差

图 2 中任何一个数据都是某个能力水平下 100 名被试的平均值。比如说,图 2(a)中的能力偏差值是某个能力水平下 100 名被试的能力估计值与真实能力值的差的平均值;图 2(b)中的能力估计标准误也是某个能力水平下 100 名被试的估计标准误的平均值。图 2 中"菱形"图例代表传统 CAT 的模拟结果,它起着重要的参照作用;而"正方形"图例代表的是 Wainer 策略 CAT 的模拟结果。从图 2(a)中可以看到,EAP 方法在能力分布两端存在着偏差。这是由能力估计值关于先验分布均值的回归效应所造成的。由于先验均值的影响,EAP 方法有一个向内的偏差模式,即低估高能力水平被试,高估低能力水平被试。当被试能力真值与施测项目难度愈不匹配时,偏差会愈显著,这在图 2(a)的 Wainer 策略 CAT 模拟结果中表现得非常明显。另外,在[-1.60, 0.40]的真实能力范围内,"聪明"的被试通过使用 Wainer 策略能够得到正偏差(相对于传统 CAT)。而且这个偏差范围非常小,从 0.07 到 0.26。也就是说,对于中低能力水平被试来说,即使他/她成功使用 Wainer 策略,能力高估程度也不大。而在[0.80,3.20]的真实能力范围内,"聪明"的被试通过使用 Wainer 策略却得到了严重的低估,最严重的低估出现在 $\theta = 3.20$ 处,低估值达到了 -2.26。

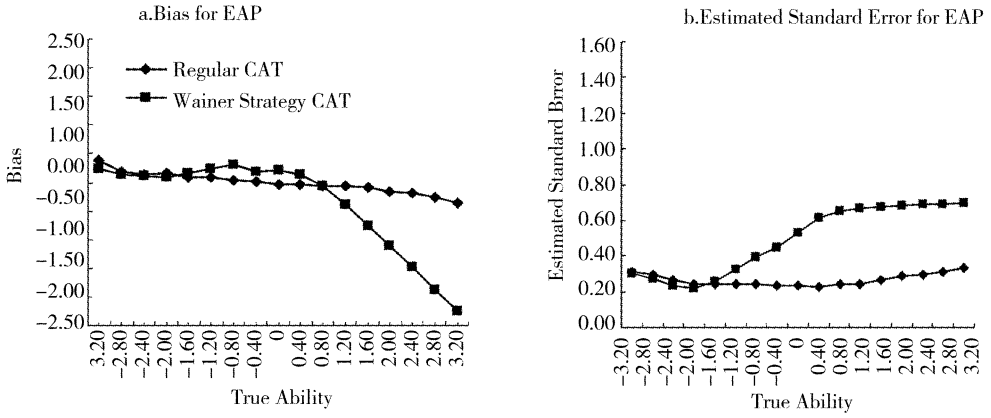

图 2 Beta = 0、采用 EAP 方法时,能力估计的偏差及相应的估计标准误

(2)标准误

图 2(b)显示,当被试真实能力增加时,由 Wainer 策略选择的项目就越来越不合适,估计标准误也就越来越大。从能力水平 -1.60 开始,Wainer 策略 CAT 产生的测验标准误就偏离于传统 CAT 的标准误。换句话说,对任何能力水平不太低的被试来说,测验一开始故意答错所有的项目会影响能力估计的精度。另外,在非常低的能力水平(比如说低于 -2),Wainer 策略 CAT 产生的标准误可能会低于传统的 CAT。产生这个结果的原因可能是:低能力水平被试在传统 CAT 中偶尔会答对项目,这样的话,低能力水平被试就有可能作答一些与他们能力不匹配的项目;而在 Wainer 策略 CAT 中,被试故意错误作答所有项目,作答项目难度较低,与被试的低能力比较相符。

3.2.2 MLE 结果及分析 Beta = 0 且采用 MLE 方法时,能力估计的偏差及相应的估计标准误见图 3 所示。

(1)偏差

图 3(c)显示:传统 CAT 的模拟结果在整个能力范围内的偏差不大。当 $\theta \in [-1.20, -0.40]$时,Wainer 策略 CAT 得到的 MLE 值只是稍高于传统 CAT 得到的 MLE 值。在这个能力范围内,即使使用 Wainer 策略,能力估计值相对于传统 CAT 估计值来说差别不是太大。当 $\theta \in [0, 0.80]$时,Wainer 策略 CAT 的能力高估现象愈发明显,最大高估值达到了

1.26。当真实能力大于0.80的时候,能力高估程度又逐渐下降。这是因为大部分能力大于0.80的被试在Wainer策略CAT上表现都非常出色,他们的能力估计值几乎都会达到最大值4。于是随着能力真值的增加,它们与最大值4之间的偏差就越来越小。

(2)标准误

图3(d)中反映的一个重要趋势就是当能力水平高于 -1.60的时候,Wainer策略CAT与传统CAT产生的标准误差异程度越来越大(这主要是因为被试能力与项目难度水平相差愈来愈远),这一点与图2(b)非常相似。另外,在非常低的真实能力水平上(比如[-2.80, -1.60]),Wainer策略CAT产生的标准误会低于传统的CAT,产生这个结果的原因同上。

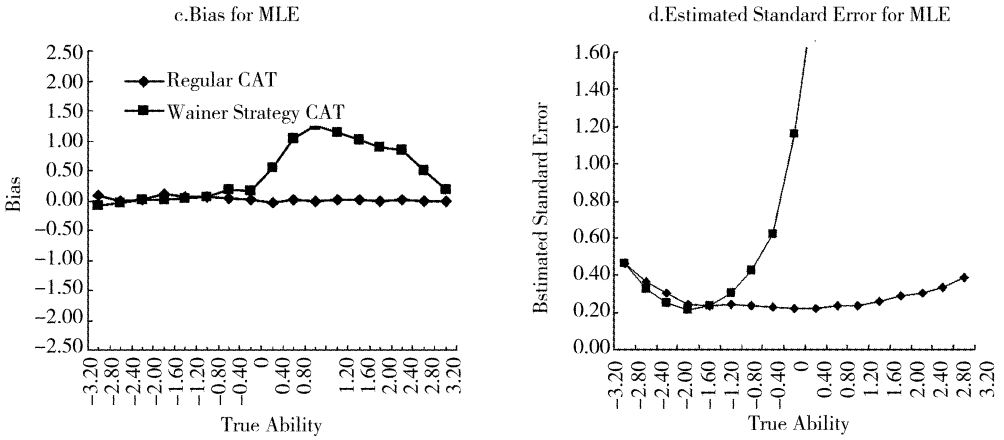

图3　Beta=0、采用MLE方法时,能力估计的偏差及相应的估计标准误

我们发现:采用EAP方法会严重低估参加CAT使用Wainer策略的中高能力水平被试;而采用MLE方法则会严重高估参加CAT使用Wainer策略的中高能力水平被试。低能力水平被试是否运用Wainer策略对能力估计影响不大。这与Vispoel得出的结论是基本一致的[8]。

3.3　Beta=0.50时的实验结果及分析

接着我们来看Beta=0.50的情况,也即评分时以1:1的比例综合考虑被试在自适应阶段和检查修改阶段的作答信息。

3.3.1　EAP结果及分析　Beta=0.50、采用EAP方法时,能力估计的偏差及相应的估计标准误见图4所示。

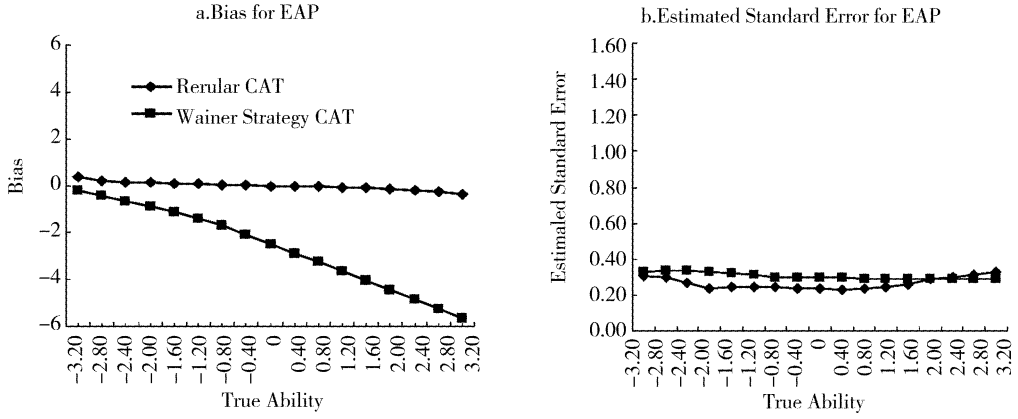

图4　Beta=0.50、采用EAP方法时,能力估计的偏差及相应的估计标准误

从图4(a)中可以发现,Wainer策略CAT产生的EAP值与能力真值存在着非常大的偏差,而且偏差随着能力真值的增大而增大。图4(b)显示:在整个能力范围内,Wainer策略CAT的估计标准误与传

统 CAT 的估计标准误的差异并不大,而且并不随着能力真值的增大而增大,这显然不合常理。在实验一(被试在作答过程中没有使用 Wainer 策略),模型 1 中 Beta 取 0.50 时不仅可以得到正常的能力估计值和标准误,而且能力估计最精确。而在这里,模型 1 中 Beta 取 0.50 时的结果却出现异常,偏差较大。产生这种结果的原因可能是:参加 CAT 使用 Wainer 策略的被试作答的项目难度太低了,导致能力估计出现异常。

3.3.2 MLE 结果及分析 Beta = 0.50、采用 MLE 方法时,能力估计的偏差及相应的估计标准误见图 5 所示。

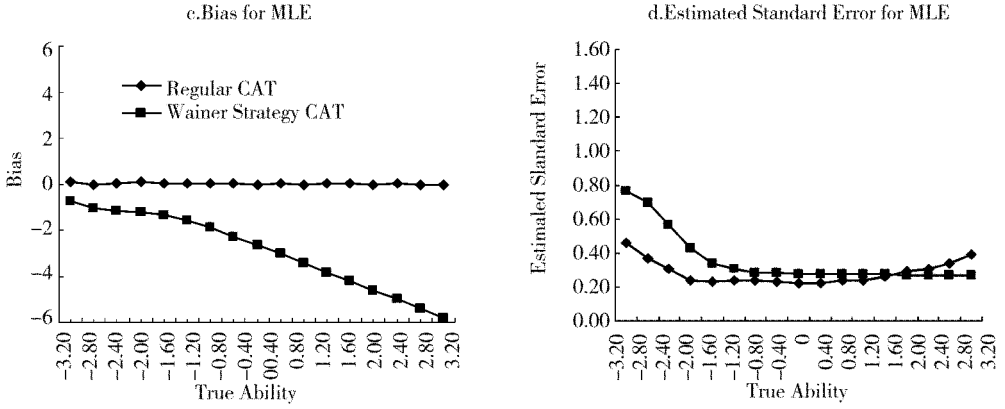

图 5 Beta = 0.50、采用 MLE 方法时,能力估计的偏差及相应的估计标准误

从图 5(c)中同样也可以发现,Wainer 策略 CAT 产生的 MLE 值与能力真值存在着非常大的偏差,而且偏差随能力真值的增加而增加。图 5(d)显示,Wainer 策略 CAT 产生的估计标准误竟然随着能力真值的增加而减小,这显然不符合常理。产生这种结果的可能原因同上。

另外,我们对 Beta 取其它 11 种值的情况(Beta = 0.10, 0.20, 0.30, 0.382, 0.40, 0.60, 0.618, 0.70, 0.80, 0.90, 1)也进行了大量的模拟实验,得出了类似于 Beta = 0.50 的实验结果:Wainer 策略 CAT 产生的能力估计值与能力真值存在着非常大的偏差,而且偏差随能力真值的增加而增加;Wainer 策略 CAT 产生的估计标准误并不随着能力真值的增大而增大。所以,本研究通过改变评分方式的方法并没能成功地对付 Wainer 策略。

3.4 对付 Wainer 策略的"粗糙"方案

我们从图 2(a)和图 3(c)中可以发现:当 Beta = 0 时,Wainer 策略的使用对低能力水平被试影响不大,对中高能力水平被试影响较大。而且对中高能力水平被试的影响又因能力估计方法的不同而不同。EAP 方法严重的低估中高能力水平被试,而 MLE 方法则高估中高能力水平被试。我们设想:EAP 方法低估,MLE 方法高估,如果能够将被试的这两个能力估计值进行综合,是否会得到更加满意的结果呢? 也就是说,即使一个中高能力水平被试使用 Wainer 策略,我们综合考虑该被试的 EAP 值和 MLE 值,是不是在一定程度上也能对付 Wainer 策略呢? 对此,我们对 Wainer 策略 CAT 的 EAP 值和 MLE 值进行整体加权综合(在整个真实能力范围内考虑),发现将 EAP 值和 MLE 值分别赋以 0.40 和 0.60 的权值加权求和时,得到的被试能力估计值与能力真值的绝对偏差最小。加权综合后的结果见图 6 中"三角形"图例所示。

从图 6(a)中可以看到,综合结果比 EAP 结果更令人满意。首先,综合结果较 EAP 结果更接近能力真值。虽然对中等能力水平被试(能力真值介于 0 到 2 之间)会产生高估,但高估程度不大。其次,对高能力水平被试(能力真值大于 2),综合结果会轻度低估他们的能力水平,起到惩罚的作用,这是我们所期望的。因为,成功运用 Wainer 策略获得正偏能力估计值的往往是高能力水平被试。另外,从图 6(b)中同样也可以看到,综合结果比 MLE 估计结果也更令人满意。首先,相对于 MLE 估计结果,综合结果也更接近能力真值。其次,对高能力水平被试(能力水平大于 2)来说,综合结果也会轻度低估他们的能力水平。

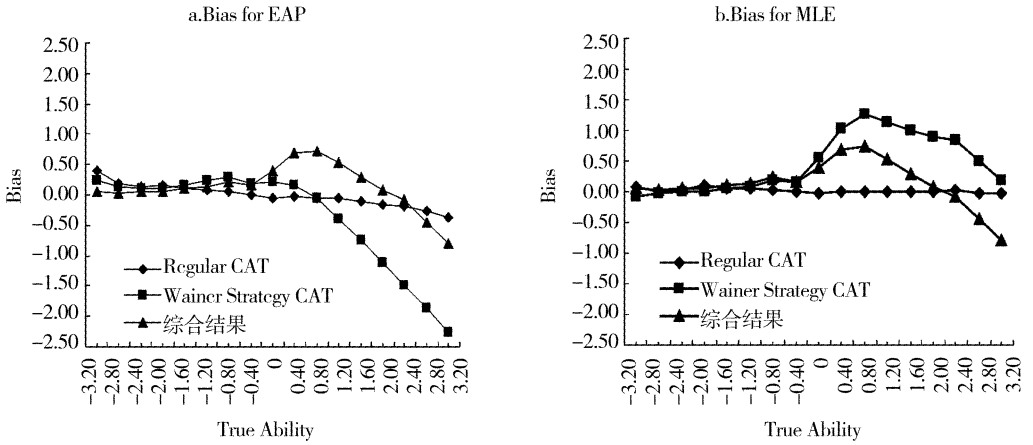

图 6　综合结果与原始结果比较图

3.5　结论

虽然本文提出的新评分方式并不能有效地对付 Wainer 策略,但是我们通过模拟研究发现当 Beta = 0 时,综合 Wainer 策略 CAT 的 EAP 值和 MLE 值可以比较“粗糙”地对付 Wainer 策略。“粗糙”对付 Wainer 策略的方法可以描述如下:首先根据答案被修改的项目数、修改前后能力估计值的差异以及能力估计的标准误,判断被试是否使用 Wainer 策略。如果被试修改了大量的项目或修改前后能力估计值差异很大或者说能力估计的标准误很大,我们就认为该被试使用了 Wainer 策略。对使用 Wainer 策略的被试,将他/她的 EAP 值和 MLE 能力估计值分别赋以 0.40 和 0.60 的权值加权求和,得到该被试的最终能力估计值。对于正常作答的被试,可以将他/她的 EAP 值或 MLE 值作为最终的能力估计值。这种方法相对“粗糙”,对使用 Wainer 策略的中等能力水平被试会产生一定程度的高估,对高能力水平被试会产生轻度的低估,但是比单独使用 EAP 值或 MLE 值作为最终能力估计值效果要好。

4　综合讨论

尽管已有研究提出能够有效对付 Wainer 策略的方法,但是我们发现相关研究者在估计被试能力时,仅仅记录了修改后的答案并仅根据该记录评分,而对修改前的原始作答信息不予考虑。我们认为考虑被试尽可能多的信息,很有可能会获得更加精确的能力估计值。基于这种想法,本文提出新的评分方式——“综合考虑被试在自适应阶段的作答信息和被试在检查修改阶段的修改信息”,试图从另一个角度出发对付 Wainer 策略。

本文实验一(被试在作答过程中没有使用 Wainer 策略)考察了不同的 Beta 值对能力估计的影响,发现 Beta = 0.50 时,模型 1 对能力估计的修复能力最强。可以理解为:如果两次作答都正确或都错误,将两者等量加权得到的综合结果保持不变;如果两次作答不同(一次正确、另一次错误),将两者等量加权得到的综合得分(0.50 分)最能够反应该被试的能力水平。也就是说,把两次作答看成同等重要、等量齐观,可以对被试进行最精确地评价。对于这个结果,如何从理论上给予有效的解释,还需要进一步的研究。另外,模型 1 得出的实验结果比模型 2 的实验结果好。本文实验二试图采用新的评分方式对付 Wainer 策略 CAT 中的 Wainer 策略,但发现能力估计值出现异常,新提出的评分方式并不能对付 Wainer 策略。于是回到 Beta = 0 的情况(只考察修改后的答案),综合 Wainer 策略 CAT 的 EAP 值和 MLE 值,给出对付 Wainer 策略的“粗糙”方案。

本研究同以前的研究一样[3],我们也假设在修改模型下,项目参数并不改变。然而,修改的情境会使得被试的参数变得不合理。因为修改使得被试花双倍的时间去作答一个项目,这意味着项目的难度会降低。另外,修改的情境改变了能力的定义,因为项目的属性不同于非修改的测验情境。有必要进一步地研究清楚修改会不会改变项目参数,会不会改变被试能力,还是两者都会改变。

多级评分模型下的 CAT 是目前研究的趋势。本研究只考虑了 0—1 评分的情况,对于允许检查并修改答案会给多级评分模型下的 CAT 带来什么影响,这还有待进一步的研究。本文也只讨论了定长 CAT 的情况,对于允许检查并修改答案会对变化长度的 CAT 产生什么影响,也将是我们下一步要研究的问题。而且在模拟允许检查并修改答案的 CAT

时,本文为了研究方便只是选用了简单的选题策略——“从题库(或剩余题库)中选择难度值与被试当前能力估计值最接近的项目”。在模拟允许检查并修改答案的 CAT 时采用 Sympson 和 Hetter 的 SH 方法以及张华华的 a - stratified 方法等有代表性的策略进行选题,是下一步需要进行的工作。另外,本文只讨论了 3PLM 情形,今后我们打算对 2PLM 情形也进行研究。

参 考 文 献

1 Qi S Q, Dai H Q, Ding S L. Principles of modern educational and psychological measurement (in Chinese). Beijing: Higher Education Press, 2002
(漆书青, 戴海琦, 丁树良. 现代教育与心理测量学原理. 北京: 高等教育出版社, 2002)

2 Gershon R, Bergstrom B. Does cheating on CAT pay: Not. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1995

3 Olea J, Revuelta J, Ximenez M C, et al. Psychometric and psychological effects of review on computerized fixed and adaptive tests. Psicologica, 2000, 21: 157 ~ 173

4 Wise S L. A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, April, 1996

5 Wainer H. Some practical considerations when converting a linearly administered test to an adaptive format. Educational Measurement: Issues and Practice, 1993, 12(1): 15 ~ 20

6 Stocking M L. Revising item responses in computerized adaptive tests: A comparison of three models. Applied Psychological Measurement, 1997, 21(2): 129 ~ 142

7 Papanastasiou E C. A ‘rearrangement procedure’ for scoring adaptive tests with review options. Paper presented at the National Council of Measurement in Education, New Orleans, April, 2002

8 Vispoel W P, Rocklin T R, Wang T, et al. Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? Journal of Educational Measurement, 1999, 36(2): 141 ~ 157

9 Revuelta J, Ximenez M C, Olea J. Psychometric and psychological effects of item selection and review on computerized testing. Educational and Psychological Measurement, 2003, 63(5): 791 ~ 808

10 Lord F M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39: 247 ~ 264

11 Barbara G D, De Ayala R J, William R K. Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 1995, 19(1): 5 ~ 22

12 Chen P, Ding S L, Lin H J, et al. Item selection strategies of computerized adaptive testing based on graded response model (in Chinese). Acta Psychologica Sinica, 2006, 38(3): 461 ~ 467
(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. 心理学报, 2006, 38(3): 461 ~ 467)

13 Vispoel W P. Reviewing and changing answers on computerized fixed - item vocabulary tests. Educational and Psychological Measurement, 2000, 60(3): 371 ~ 384

Research on Computerized Adaptive Testing that Allows Reviewing and Changing Answers

CHEN Ping[1,2], DING Shu-Liang[2]

([1] *Developmental Psychology Institute, Beijing Normal University, Beijing* 100875, *China*)

([2] *Computer Information Engineering College, Jiangxi Normal University, Nanchang* 330027, *China*)

Abstract

In the past decade, some paper - and - pencil (P&P) tests have been replaced by computerized adaptive testing (CAT) within many large - scale standardized testing programs. However, many researches and applications on CAT had limitations because most of the CAT did not allow examinees to review and change their answers. Among the operative CAT applications, there was only one that incorporated item review. Not allowing examinees to review and change their answers would result in test pressure and affect their performances. Moreover, a majority of examinees manifested a clear preference for item review because they believed that the inclusion of item review made the test fairer and considered it to be a disadvantage if review was disallowed. The CAT test organizers did not allow examinees to review and change answers mainly because they were apprehensive that examinees would use the deceptive Wainer strategy in the review stage to obtain positively biased ability estimates, consequently affecting the fairness and precision of the test. If we could provide a solution that not only allowed examinees to review and change answers but that was also able to deal

with the Wainer strategy, the meaning would be great for the development of CAT. Until now, there have been few relevant studies on this topic worldwide. Moreover, the previous studies had a nonnegligible disadvantage, in that the researchers only recorded the answers of the review stage and used them as a basis for scoring, without considering the answers of the adaptive stage. We assumed that comprehensively considering the answers before and after review could produce a more accurate ability estimation. Therefore, this paper employed a new scoring method and attempted to deal with the Wainer strategy:

$$V_{\alpha j} = \text{Beta} \times U_{\alpha j} + (1 - \text{Beta}) \times U_{\alpha, j+m}$$

This study involved two experiments. Experiment 1 used the Monte Carlo method to simulate the entire process of CAT that allows the reviewing and changing of answers, with the aim of investigating the influence of different beta values on ability estimation. Experiment 2 used simulation data generated by the Monte Carlo method to evaluate the effectiveness of the Wainer strategy and attempted to deal with the strategy by using a new scoring method.

The simulation results of Experiment 1 indicated the following. First, comprehensively considering the answers before and after review did produce a more accurate ability estimation, and the most accurate estimates occurred when beta = 0.50. Second, the share of examinees who changed their answers was 66.80%; further, 6.40% of the answers were changed, and 75% of the modified answers represented changes from incorrect to correct answers. Experiment 2 indicated the following: When using the new scoring method, the ability estimates generated by the CAT involving the use of the Wainer strategy obviously diverged from the true ability values. Moreover, the bias increased as the true ability value increased.

The new scoring method employed in this study was not able to effectively deal with the Wainer strategy because of the abnormal ability estimates and abnormal estimated standard error. However, through a simulation experiment, we found the following: When beta = 0, comprehensively considering the expected a posteriori (EAP) and maximum likelihood estimation (MLE) ability estimates of CAT involving the use of the Wainer strategy succeeded in roughly dealing with the Wainer strategy. Our future task involves developing a more accurate method to deal with the Wainer Strategy.

Key words　computerized adaptive testing, item review, Wainer strategy, Monte Carlo simulation.