

2018 届研究生硕士学位论文

分类号: _____

学校代码: 10269

密 级: _____

学 号: 51150104043



華東師範大學

East China Normal University

硕士学位论文

MASTER' S DISSERTATION

基于内容平衡的计算机自适应测试选题策略研究

院 系: 教育学部教育信息技术学系

专 业: 教育技术学

研 究 方 向: 计算机辅助教育

指 导 教 师: 孟玲玲 副教授

学位申请人: 刘梦娇

2018 年 3 月完成

Dissertation for master' s degree in 2018

University code:10269

Student ID:51150104043

East China Normal University

**Item Selection Strategy of Computer Adaptive
Testing based on Content Balancing**

Department: Educational Information and Technology

Major: Educational Technology

Research direction: Computer Based Education

Supervisor: Associate - Prof. Lingling Meng

Candidate: Mengjiao Liu

March , 2018

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于内容平衡的计算机自适应测试选题策略研究》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：刘哲齐

日期：2018 年 5 月 21 日

华东师范大学学位论文著作权使用声明

《基于内容平衡的计算机自适应测试选题策略研究》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

☐ 1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，于 年 月 日解密，解密后适用上述授权。

☒ 2. 不保密，适用上述授权。

导师签名：马永波

本人签名：刘哲齐

2018 年 5 月 21 日

* “涉密”学位论文应是经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

刘梦娇 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
吴永和	研究员	华东师范大学	主席
叶长青	副教授	华东师范大学	
郁晓华	副教授	华东师范大学	

摘 要

计算机自适应测试能够根据被试的答题情况，即时估计被试的当前能力水平，并从题库中选择出最符合当前被试能力水平的试题进行施测。计算机自适应测试的提出和发展为教育评价提供了新的手段，与传统的测试方式相比，自适应测试具有测试效率更高、结果更准确等优点。

由于计算机自适应测试的核心是选择最适合的试题让被试作答，因此如何在题库中选择试题，即选题策略，是计算机自适应测试研究领域的核心问题之一。目前虽已提出许多自适应测试的选题策略，但是对于测试中内容平衡约束条件的控制关注较少，而内容平衡对于测试准确性具有较大的影响。本论文以此为切入点，在现有选题策略的基础上进行改进，提出了 c-STR-ST 内容平衡选题策略，并采用 Monte Carlo 模拟实验的方式将 c-STR-ST 与现有的 CCAT、MMM 以及 STR-C 选题策略的测试效果进行对比。实验结果表明 c-STR-ST 选题策略在测试准确性、曝光率、测试重叠率等多项指标上综合表现优于其他三种选题策略。

此外，本论文基于 c-STR-ST 选题策略设计并开发了计算机自适应测试系统，并建设基于高一数学三个内容域的题库，将系统与题库应用到高中数学阶段性教学评价中，根据测试的结果对学生内容域总体掌握情况、学生的认知情况进行分析，为教师和学生提供更加详实、更有针对性的评价结果，促进个性化学习的展开。

关键词：计算机自适应测试；项目反应理论；选题策略；内容平衡；参数估计；认知诊断

Abstract

The computer adaptive testing is a testing method that can estimate the current ability level of the examinee in time according to the response of previous test items, and select the item that best meets the current ability level from the item bank. The development of computer adaptive testing provides a new method for educational evaluation. Compared with traditional testing methods, adaptive testing has the advantages of higher test efficiency and can get more accurate results.

Because the core of computer adaptive testing is to select the most suitable item according to the current ability estimation of the examinees, therefore, how to select the test item, that is, the item selection strategy is one of the core issues in the research area of computer adaptive testing. Although many strategies have been proposed, there is little concern to the control of contents balancing. However, contents balancing has a great impact on the test accuracy. Therefore, in this thesis, a contents balancing item selection strategy named c-STR-ST is proposed, and Monte Carlo simulation is used to verify the test results of c-STR-ST. Compared with the existing CCAT, MMM and STR-C strategies, the results of Monte Carlo show that the c-STR-ST strategy is superior to the other three strategies in terms of test accuracy, exposure rate, and test overlap.

In addition, this thesis designed and developed a computer adaptive testing system based on the c-STR-ST strategy, and built an item bank of mathematics in the first year of senior high school. This system and item bank are applied to the evaluation of mathematics for senior middle school students. The test results include students' mastery of knowledges, cognitive diagnosis of each student, which provide teachers and students

a more detailed and individualized evaluation results.

Key words: Computer Adaptive Testing; Item Response Theory; Item Selection Strategy; Content Balancing; Parameter Estimation; Cognitive Diagnosis

目录

摘要.....	I
Abstract.....	II
目录.....	IV
表目录.....	VI
图目录.....	VII
第 1 章 绪论.....	1
1.1 研究背景.....	1
1.2 研究现状.....	3
1.3 研究目标与研究内容.....	9
1.4 研究思路与方法.....	10
1.5 论文结构与各章概要.....	11
第 2 章 计算机自适应测试的理论基础.....	12
2.1 项目反应理论概述.....	12
2.2 项目反应理论的基本假设.....	14
2.3 项目反应理论的模型.....	14
2.4 项目反应理论的参数估计方法.....	16
第 3 章 c-STR-ST 选题策略提出与模拟实验	20
3.1 计算机自适应测试中的选题策略概述.....	20
3.2 c-STR-ST 选题策略的提出	28
3.3 内容平衡选题策略算法实验效果分析.....	31
3.4 内容平衡选题策略测试效果比较分析.....	43
3.5 内容平衡选题策略测试效果比较与讨论.....	48
3.6 总结.....	53

第 4 章	c-STR-ST 选题策略在高中数学测试中的应用	55
4.1	基于 c-STR-ST 选题策略的自适应测试系统设计.....	55
4.2	基于 c-STR-ST 选题策略的自适应测试系统开发.....	59
4.3	基于 c-STR-ST 选题策略的自适应测试系统在高中数学测试中的应用	64
4.4	高中数学自适应测试结果分析.....	70
第 5 章	总结与展望.....	77
5.1	研究总结.....	77
5.2	研究展望.....	77
参考文献	78
附录 1	83
附录 2	84
附录 3	85
附录 4	87
致谢	89
攻读硕士学位期间科研成果	90

表目录

表 3-1 题库 1 模拟实验结果.....	47
表 3-2 题库 2 模拟实验结果.....	47
表 3-3 题库 3 模拟实验结果.....	47
表 3-4 题库 4 模拟实验结果.....	48
表 4-1 试题信息表.....	58
表 4-2 考生信息表.....	58
表 4-3 测试用题表.....	58
表 4-4 答题情况表.....	58
表 4-5 测试结果表.....	58
表 4-6 题库数据表结构.....	62
表 4-7 题库中试题涉及的内容域及其描述	65
表 4-8 被试情况统计	71
表 4-9 内容域 1 学生掌握情况统计表	73
表 4-10 内容域 2 学生掌握情况统计表	73
表 4-11 内容域 3 学生掌握情况统计表	74

图目录

图 1-1 研究思路	10
图 2-1 项目特征曲线	13
图 2-2 三参数模型项目特征曲线.....	16
图 2-3 牛顿—拉夫逊迭代法流程图	19
图 3-1 最大信息法算法流程图	20
图 3-2 a 分层算法流程图	22
图 3-3 影子题库算法流程图.....	23
图 3-4 约束 CAT 法算法流程图	24
图 3-5 修正的多项模型算法流程图	25
图 3-6 c 分层法算法流程图.....	27
图 3-7 新选题策略算法流程图.....	29
图 3-8 题库中试题区分度 a 满足 $\ln a \sim N(0,1)$ 分布图.....	33
图 3-9 题库中试题区分度 a 满足 $a \sim U(0.2,2.5)$ 分布图	34
图 3-10 题库中试题难度系数 b 满足 $N(0,1)$ 分布图	35
图 3-11 题库中试题难度系数 b 满足 $U(-3,3)$ 分布图	35
图 3-12 题库中试题猜测系数 c 分布图	36
图 3-13 被试能力水平分布图	37
图 3-14 反应向量的生成方法	38
图 3-15 采用 CCAT 选题策略试题曝光曲线	49
图 3-16 采用 MMM 选题策略试题曝光曲线.....	49

图 3-17 采用 STR-C 选题策略试题曝光曲线	49
图 3-18 采用 c-STR-ST 选题策略试题曝光曲线	50
图 3-19 四种选题策略在题库 1 模拟实验中可量化指标雷达图	51
图 3-20 四种选题策略在题库 2 模拟实验中可量化指标雷达图	52
图 3-21 四种选题策略在题库 3 模拟实验中可量化指标雷达图	52
图 3-22 四种选题策略在题库 4 模拟实验中可量化指标雷达图	53
图 4-1 自适应测试用户需求分析	55
图 4-2 基于 c-STR-ST 选题策略的自适应测试系统功能结构	57
图 4-3 数据库 E-R 图	59
图 4-4 系统结构示意图	60
图 4-5 系统功能模块	60
图 4-6 登录界面	61
图 4-7 登录错误提示框	61
图 4-8 自适应测试界面	62
图 4-9 查看测试结果界面	64
图 4-10 学生答题情况统计	67
图 4-11 BILOG 3.0 参数估计指令	67
图 4-12 BILOG 3.0 参数估计结果	68
图 4-13 高一数学自适应测试题库中试题区分度 a 分布图	69
图 4-14 高一数学自适应测试题库中试题难度系数 b 分布图	69
图 4-15 高一数学自适应测试题库中试题猜测系数 c 分布图	70
图 4-16 学生的能力水平估计	71

图 4-17 学生的能力水平估计分布	72
图 4-18 学生 1 认知曲线	75
图 4-19 学生 9 认知曲线	75
图 4-20 学生 24 认知曲线	75

第1章 绪论

1.1 研究背景

教育评价是教育教学活动中的一项重要环节,能够对学生的学习成果以及教师的教学质量做出价值判断。现阶段的教育评价主要采用测试的形式,通过测试学生能够直观地反映出对知识的掌握情况,为下一阶段学习和教学活动的开展提供指导。

传统测试主要是纸笔测试,在此模式下,学生作答指定的试题,根据学生对所有试题的作答情况以及试题的分值给出最终的测试成绩。采用该模式进行测试虽然能够有效地反映出学生对于知识的掌握情况,但是这种“千人一卷”的测试并不适应于每一个学生,可能存在与学生的真实能力水平相比,难度过高或过低的试题,从而导致测试结果不准确的问题。此外,由于试题难度的不合理也可能导致学生在测试过程中的心理变化,对测试结果产生影响。

自适应测试(Adaptive Testing)的提出很好地解决了传统测试模式中存在的问题。自适应测试的主要特点是测试中不再是所有学生作答相同的试题,而是根据学生当前的能力估计值,选择与之相适应的试题施测,与传统的测试方式相比能够实现“因人施测”。在自适应测试中,前几道试题随机给出,根据学生的作答情况,对其能力水平做出估计,之后的测试中据此能力估计值,从本次测试的题库中按照一定的选题策略选出难度适合学生的试题进行施测,学生作答后根据当前的作答情况重新估计其能力水平并选择下一道试题,重复这一选题、作答的过程,直到满足预先设定的测试终止条件。

由于自适应测试的模式涉及到复杂的数学计算,且自适应出题的模式较为繁琐,若由人工完成出题和即时能力估计,则会导致测试过程不流畅,且对于人力、物力的消耗量大等问题。计算机技术的发展解决了自适应测试中测试成本高的问题,计算机自适应测试(Computerized Adaptive Testing, CAT)在技术的支持下能够迅速即时估计被试能力水平,并据此筛选出最适合的试题予以作答,能够在保证测试过程流畅性的同时,提高自适应测试模式的实用性和

便捷性。

与传统的测试模式相比，计算机自适应测试具有以下四个方面的优势：

1. 提高测试结果的准确性。首先，由于测试中每道试题都是根据学生能力水平估计值筛选出的，因此避免了在测试中出现难度过高或是难度过低的试题，从而影响测试结果的问题。此外，由于试题的难度是根据学生作答情况而变化的，因此对于学生在测试中的心理状态影响较小，使得学生在测试中能够保持平稳的心态，避免因心理变化而影响真实水平的发挥。最后，不同于传统测试中单纯以对错以及试题的分值计算学生的测试结果，自适应测试中的测试结果是基于相关测试理论和数学模型对学生能力水平的估计值，因此具有更高的科学性、准确性和参考价值。

2. 提高测试的效率。传统测试中由于试题是预先设定的，为了对能力水平不同的学生进行施测，除了常规难度的试题外，通常存在难度过高或过低的试题。这些试题对于大部分能力水平适中的学生而言，其对于太难或太简单试题的答题情况并不能反映对知识的掌握情况，出现在测试中不仅不能对测试结果产生积极的作用，反而降低了测试的效率。而在计算机自适应测试中，通过自适应的出题方式，学生的能力估计值将不断趋近于真值，测试的试题更有针对性，从而使得测试能够通过更少的试题得到更准确的结果¹。

3. 提高测试的安全性。由于计算机自适应测试中每个学生的试题都不同，且是随着考试的进行不断筛选出来的，可预测性较低，因此能够很大程度上减少测试中的舞弊现象。此外，由于自适应测试题库的设置和更新机制，加之在选题时的相关约束条件，试题的曝光率能够控制在合理的范围，避免因同一道试题施测多次而影响测试安全性的问题。

4. 能够促进个性化学习的开展。计算机自适应测试题库中的每一道试题都标定有详细的信息，包括与知识点之间的对应关系，根据这些试题信息在测试

¹ Weiss D J. Improving Measurement Quality and Efficiency with Adaptive Testing[J]. Applied Psychological Measurement, 1982, 6(4):473-492.

后能够分析出每个学生对于各个知识点的详细认知情况。因此，通过自适应测试，教师和学生能够获取到更多的评价信息，对于学生了解自己上一阶段学习情况、差缺补漏以及教师制定下一阶段教学计划、为学生提供个别化指导给予了支持。

1.2 研究现状

比奈测试是计算机自适应测试的前身。在比奈测试中采用了根据被试对上一个项目的反应情况选择下一个项目的基本思想，已经具备“自适应”的特点。但是由于技术条件的限制，比奈测试通常是通过人工来完成的，花费的人力成本和时间成本高。随着计算机技术的不断发展和测试理论的完善，美国科学家 Lord 首次提出了计算机自适应测试的概念²。由于计算机自适应测试所具有的优势，受到了专家学者以及教育工作者的关注，相关的理论研究不断推进，且逐步应用到了实际的测评中。

1.2.1 计算机自适应测试研究现状

1. 国外研究现状

在理论研究方面，计算机自适应测试领域的理论研究主要集中在对于其测试理论基础的研究。

自适应测试的理论基础为项目反应理论，该理论将被试的能力水平量化，并用数学关系表示被试正确作答试题的概率和被试能力水平、试题特征之间的关系。项目反应理论的模型在研究中不断优化，在其发展过程中具有代表性的研究成果有双参数正态卵形模型（Probit IRT Model）³、等级反应模型（Graded

² Lord, F. M. Applications of Item Response Theory To Practical Testing[M]. LAWRENCE ERLBAUM ASSCCIAATES, 1980.

³ Lord F. A theory of test scores[M]. Psychometric Monograph, 1952, 7.

Response Mode)⁴、逻辑斯蒂模型(Logistic Model)⁵等。其中,应用最广泛的为逻辑斯蒂模型。

计算机自适应测试的理论研究使得其理论模型能够满足于更多测试的需求,同时测试的效率、准确性也得到极大的提高,为自适应测试的发展和应用提供了理论依据和基础。

在应用研究方面,自适应测试被广泛地应用于教育评价、认知诊断、心理测量等领域。其中,在教育评价领域已有大量进入到实践应用的案例。国外对于自适应测试的应用起步较早,1984年第一个全国范围内的计算机自适应测试系统 CAST(Computerized Adaptive Screen Test)在美国军方实验室得以应用⁶。1991年 Novell 公司在其认证资格考试中采用了自适应测试的方法。1993年,一些大规模的测试也使用自适应的方式进行,包括美国研究生院入学测验(GRE)、美国商学院研究生招生考试(GMAT)等⁷。

2. 国内研究现状

国内关于计算机自适应测试的理论研究相对起步较晚,但是仍然展开了许多深入的研究,并有相关研究成果产出。在自适应测试的理论基础方面,余嘉元教授在《项目反应理论及其应用》中对项目反应理论进行了剖析和进一步研究⁸;路鹏在项目反应理论中经典模型逻辑斯蒂模型的基础上提出了Logistic-T模型

⁴ Samejima F. Estimation of latent ability using a response pattern of graded scores. [J]. Ets Research Report, 1969, 34(1):1-97.

⁵ Lord, F. M. Applications of Item Response Theory To Practical Testing[M]. LAWRENCE ERLBAUM ASSCIIAATES, 1980.

⁶ 秦珊珊. 面向高中英语的自适应测试系统中项目参数的实验研究[D]. 东北师范大学, 2013.

⁷ 张华华, 程莹. 计算机化自适应测验(CAT)的发展和前景展望[J]. 考试研究, 2005(2):14-26.

⁸ 余嘉元. 项目反应理论及其应用[M]. 江苏教育出版社, 1992.

⁹；多级评分计算机化自适应测验方面也取得颇多研究成果^{10、11}，为自适应测试由二计分制推广到多级评分制提供了理论基础；此外，国内学者在自适应测试题库建设优化方面也做出相关研究¹²。

在应用方面，国内有代表性的应用案例有大学英语四、六级考试（CET4/6）¹³以及中国汉语水平考试（HSK）¹⁴等。其中大学英语四、六级考试为了提高测试的信效度、使得测试更加客观，采用了自适应的方式对客观题进行测试，取得了较好的测试效果，并逐步形成较为成熟的自适应测试体系。中国汉语水平考试（HSK）经过不断的研究，形成了计算机自适应性的 HSK，测试的信度、效度较传统的测试方式都得到了提升¹⁵。

1.2.2 计算机自适应测试选题策略研究现状

计算机自适应测试的最大特点在于其选题模式，即根据学生的作答情况，动态地筛选出下一道试题。因此，如何选择适合被试能力水平的试题以达到较高的测量精度及测验目标，即选题策略（Item Selection Criteria, ISC）是自适应测试技术最为重要的部分¹⁶。选题策略的优劣直接影响了测试的效率和准确性，好的选题策略能够在最短的测试长度内得到最准确的测试结果。

现阶段的选题策略根据选题目的不同可以分为以下四类：

（1）以提高测试准确性为主要目的的选题策略

⁹ 路鹏. 计算机自适应测试若干关键技术研究[D]. 东北师范大学, 2012.

¹⁰ 罗芬, 丁树良, 王晓庆. 多级评分计算机化自适应测验动态综合选题策略[J]. 心理学报, 2012, 44(3):400-412.

¹¹ 王晓庆, 罗芬, 丁树良, 等. 多级评分计算机化自适应测验动态调和平均选题策略[J]. 心理学探新, 2016, 36(3):270-275.

¹² 游晓锋, 丁树良, 刘红云. 计算机化自适应测验中原始题项目参数的估计[J]. 心理学报, 2010, 42(7):813-820.

¹³ 李慧. 浅析计算机自适应考试系统在大学英语测试中的应用前景[J]. 中国现代教育装备, 2009(3):27-29.

¹⁴ 付聪. 计算机自适应测试研究进展[J]. 现代情报, 2005, 25(1):61-64.

¹⁵ 付聪. 计算机自适应测试研究进展[J]. 现代情报, 2005, 25(1):61-64.

¹⁶ 简小珠, 戴海琦, 张敏强, 等. CAT 选题策略分类概述[J]. 心理学探新, 2014, 34(5):446-451.

(2) 以控制试题曝光率为主要目的的选题策略

(3) 结合试题内容平衡的选题策略

(4) 综合型选题策略

测试的准确性是指测试能够直接反映被试特征的程度¹⁷。由于准确性对于测试的重要意义，许多计算机自适应测试选题策略将提高测试准确性作为主要目的。这类选题策略的主要思想是将被试的能力水平和试题的参数利用数学模型关联起来，选择出最适合被试当前能力估计的试题。这类选题策略目前有代表性的有：

(1) 最大信息法 (Maximum Fisher Information, MFI)¹⁸，该方法的核心思想是在测试中，根据被试当前的能力估计，利用项目信息函数 (Item Information Function, IIF) 计算试题的信息量，信息量越大的试题有越高的测试价值，因此选择信息量最大的试题进行测试。

(2) 最小期望后验标准差 (Minimum Expected Posterior Standard Deviation, MEPSD)¹⁹，该方法基于贝叶斯思想、三参数正态比肩曲线模型提出，根据被试能力后验期望估计筛选出难度最适合的试题。

(3) 最大全局信息量 (Maximum Global-Information, MGI)²⁰，该方法中引入了 K-L 信息量 (Kullback-Leibler information) 以标定全局信息量，选择具有最大全局信息量的试题进行施测。

试题曝光率是试题在测试中被使用的频率。自适应测试题库中的试题若是具有较高的曝光率，则表示该试题已被施测过多次，被试可能在测试前对这些

¹⁷ Mao X Z, Xin T. Item Selection Method in Computerized Adaptive Testing[J]. Advances in Psychological Science, 2011.

¹⁸ Lord, Frederic M. A Broad-Range Tailored Test of Verbal Ability[J]. Applied Psychological Measurement, 1977, 1(1):95-100.

¹⁹ Owen R J. A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing[J]. Journal of the American Statistical Association, 1975, 70(350):351-356.

²⁰ Chang Hua-Hua, Ying, Zhiliang. A Global Information Approach to Computerized Adaptive Testing. [J]. Applied Psychological Measurement, 1996, 20(3):213-229.

试题就已有了解和准备；而若是题库中的试题具有过低的曝光率，则说明该类试题很少被施测，造成试题的“浪费”。因此试题曝光率对于题库的建设和测试结果都会起到很大的影响。为了保证题库的安全、提高试题的利用率，学者们提出了很多以控制试题曝光率为主要目的的选题策略²¹，主要有：

(1) a 分层法 (a-Stratified, STR-a)²²，该方法将题库按照试题的区分度（即参数 a）分层，在测试的不同阶段从对应的层中选择试题施测，以保证不同区分度的试题都能均匀曝光。

(2) b 分层法 (a-Stratified with b-Blocking, STR-b)²³，该方法在 a 分层法的基础上将试题的难度系数（参数 b）纳入到分层策略的考量中，在保证曝光度的基础上提高了测试的准确性。

(3) S-H 法 (Simpson-Hetter method, S-H)²⁴，该方法将条件概率引入到选题策略中以达到控制试题曝光率的目的。

(4) 项目合格法 (Item Eligibility Method, IE)²⁵，该方法将被选出的概率高于阈值的试题认为是“合格”的，并从所有判定为合格的试题所组成的子题库中选择出最佳试题施测。

(5) 多重极大曝光率法 (Multiple Maximum Exposure Rate Method, MRM)²⁶，该方法的核心思想是，规定一个最大曝光率，并采用使被测试题的曝光率不

²¹ Ozturk, Nagihan Boztunc|Dogan, Nuri. Investigating Item Exposure Control Methods in Computerized Adaptive Testing. [J]. Kuram Ve Uygulamada Egitim Bilimleri, 2015, 15(1):85-98.

²² Chang H H, Ying Z. α -stratified multistage computerized adaptive testing[J]. Applied Psychological Measurement, 1999, 23(4):211 - 222.

²³ Chang H H, Qian J, Ying Z. a-stratified multistage computerized adaptive testing with b blocking. [J]. Applied Psychological Measurement, 2001, 25(4):333-341.

²⁴ Simpson J B, Hetter R D. Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association[C]. San Diego, CA: Navy Personnel Research and Developmen, 1985.

²⁵ Linden W J V D, Veldkamp B P. Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests[J]. Journal of Educational & Behavioral Statistics, 2004, 29(3):273-291.

²⁶ Barrada R, Bernard P. |Olea. Multiple Maximum Exposure Rates in Computerized Adaptive Testing. [J]. Applied Psychological Measurement, 2009, 33(1):58-73.

超过最大曝光率的方法实现曝光控制。

(6) 最大信息量组块分层法 (Maximum Information Stratified with Blocking, MIS-B)²⁷, 该方法结合了分层方法的特点并引入 c 参数对现有方法进行修改而提出。

在上述选题策略中, 试题的选择依据均是由被试能力、试题参数等筛选而出的, 而这种选题依据可能会导致试题测试内容的不平衡, 这种不平衡可能会造成各内容域之间难度差异过大的问题, 使得学生因在某个内容域内作答情况良好而获得比真实能力水平更高的测试结果, 或是学生因在某个内容域内作答情况不理想而获得比真实能力水平低的测试结果, 这些都会导致测试结果的不准确。然而内容平衡本身并不是自适应测试理论基础模型所考虑的, 因此为了保证在各个内容区域平衡的测试条件, 需要在自适应测试系统中加入内容平衡的过程²⁸。这类选题策略主要有:

(1) 约束 CAT 法 (Constrained CAT, CCAT)²⁹, 该方法通过设定期望内容比例, 从离期望最远的内容域中选择试题实现内容平衡。

(2) 修正的多项模型 (Modified Multinomial Model, MMM)³⁰, 该方法的核心思想是通过设定内容域累计分布, 并用产生的随机数与累计分布做比较选择出相应内容域的试题。

(3) c 分层法 (α -Stratified Method with Content-Blocking, STR-C)³¹

²⁷ Barrada J R, Mazuela P, Olea J. Maximum information stratification method for controlling item exposure in computerized adaptive testing. [J]. Psicothema, 2006, 18(1):156-159.

²⁸ Leung C K, Chang H H, Hau K T. Content Balancing in Stratified Computerized Adaptive Testing Designs. [J]. Adaptive Testing, 2000, 2000(1):20.

²⁹ G. Gage Kingsbury, Anthony R. Zara. Procedures for Selecting Items for Computerized Adaptive Tests[J]. Applied Measurement in Education, 1989, 2(4):359-375.

³⁰ Chen S Y, Ankenmann R D. Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing[J]. Journal of Educational Measurement, 2004, 41(2):149-174.

³¹ Yi Q, Chang H H. α -Stratified CAT design with content blocking[J]. British Journal of Mathematical & Statistical Psychology, 2003, 56(2):359-78.

该方法在 b 分层法的基础上，在分层时将内容域参数作为分层因素之一，从而达到试题内容平衡的目的。

计算机自适应测试中选题需要考虑多方面的约束，因此也有相关的选题策略针对不同约束的结合而提出。这类选题策略主要包括：

(1) 加权离差模型 (Weighted Deviation Model, WDM)³²，该策略将组合线性测验的离差模型运用到自适应测试中，实现试题筛选中的多因素控制。

(2) 影子测验 (Shadow Test, ST)³³，该方法的核心是在测试中不断形成满足当前多个约束条件的试题组成的影子题库，施测试题将从影子题库中依据其他选题方法选出，从而保证了备选试题满足各项约束条件。

1.3 研究目标与研究内容

作为自适应测试系统中的一个重要环节，选题策略对测试最终结果的信度、效度、测试效率、测试准确性以及测试安全性等都有极大的影响。试题的内容平衡作为选题策略中的一个重要因素，也应纳入到试题选择的约束条件中。目前有关内容平衡的选题策略对于其他选题因素（如：测试准确性、曝光率等）的考虑较少，因此，本论文将以此为切入点，在现有的控制曝光率和提高测试准确性的选题策略的基础上，添加内容平衡约束条件，对选题算法加以改进，从而实现测试试题在各内容域间的平衡。此外，由于现阶段自适应测试在中小学教育评价中的应用较少，本研究将基于改进后的选题策略，设计开发自适应测试系统，并将该系统运用到中小学的教学评价中，验证其在实际应用中的效果，并为师生提供更加详细、准确的评价信息。

本论文的研究内容总结如下：

(1) 在前人研究的基础上，改进计算机自适应测试的选题策略算法，在保

³² Swanson L, Stocking M L. A Model and Heuristic for Solving Very Large Item Selection Problems. [J]. Applied Psychological Measurement, 1993, 17(2):151-166.

³³ Linden V D, Wim J. | Reese, Lynda M. A Model for Optimal Constrained Adaptive Testing. [J]. Applied Psychological Measurement, 1997, 22(3):259-270.

证测试准确性、选题的难度、曝光率、利用率的基础上，增加内容平衡的约束条件。

（2）通过模拟实验验证改进后的选题策略算法的效果，并与现有选题策略对比。

（3）基于改进的内容平衡选题策略设计和开发自适应测试系统，将改进后的选题策略应用在此系统中。

（4）在实际的中小学教育评价中验证系统的可行性和有效性，基于测试信息对学生进行认知诊断，为教师和学生提供更加详细、准确的测试信息。

1.4 研究思路与方法

本论文的研究思路如图 1-1 所示：

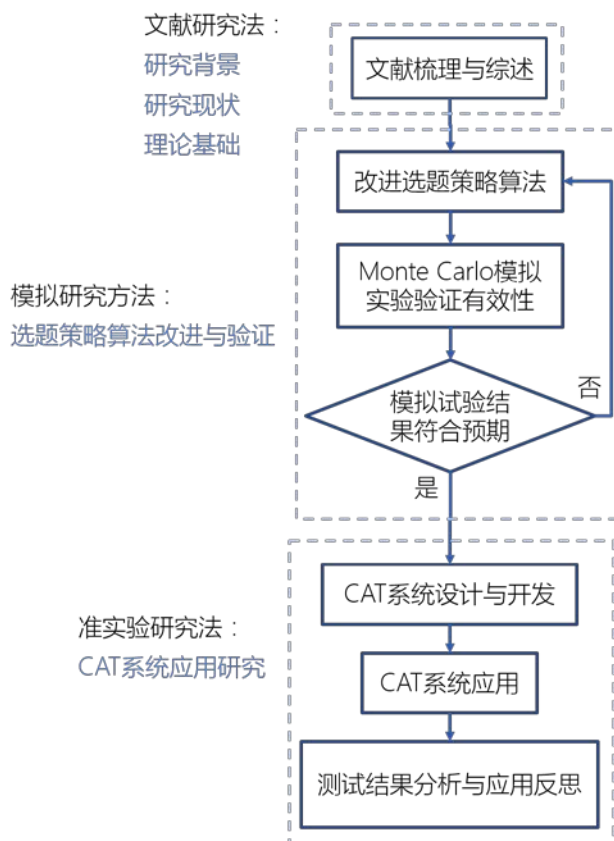


图 1-1 研究思路

本论文采用的研究方法主要有：

（1）文献研究法：通过文献资料深入了解自适应测试的基本测试方法和过

程,研究其理论基础的相关内容;通过文献的搜集和整理,梳理计算机自适应测试的研究现状、应用现状,以及现阶段提出的选题策略及其特点与不足,奠定研究的理论基础。

(2) 模拟研究方法:以概率和统计理论方法为基础,利用 Monte Carlo 随机模拟方法,验证改进后的选题策略在能力估计、测试效率、项目调用均匀性和曝光率方面的表现。

(3) 准实验研究法:根据改进后的选题策略,开发计算机自适应测试系统,并将此系统应用到中小学的实际教学评价中,验证系统的可行性和有效性。

1.5 论文结构与各章概要

论文共分为五个章节,各章节的主要内容如下:

第一章绪论,主要介绍相关的研究背景、意义,提出了论文的研究目标和主要内容,确定研究思路和方法。

第二章计算机自适应测试的理论基础,主要介绍了计算机自适应测试系统的理论基础,包括经典测试理论和项目反应理论。

第三章 c-STR-ST 选题策略提出与模拟实验,主要介绍了研究中提出的基于内容平衡的选题策略 c-STR-ST,并用 Monte Carlo 模拟实验的方法将 c-STR-ST 与现有的内容平衡选题策略 CCAT、MMM 以及 STR-C 的测试效果进行对比,规定测试精准度、曝光度等多项指标,对四种方法的测试结果进行综合分析。

第四章 c-STR-ST 选题策略在高中数学测试中的应用,主要介绍了基于 c-STR-ST 选题策略的自适应测试系统的设计以及开发、高一数学自适应测试题库的建设,并将该系统与题库应用到实际的高中数学教学评价中,并对测试结果进行分析,为教师和学生的教学活动提供参考依据。

第五章总结与展望,对研究进行了总结,分析了研究中存在的不足,并对之后的研究方向做出分析与展望。

第2章 计算机自适应测试的理论基础

2.1 项目反应理论概述

2.1.1 经典测试理论及其局限

传统测试主要以经典测试理论(Classical Test Theory, CTT)为理论基础,该理论出现于20世纪初,经过在理论和实践中的不断地发展和完善逐步形成了完整的体系,并总结归纳出三个基本假设³⁴:

1. 真分数能够反应真实能力,因此认为是恒定不变的。
2. 测试误差具有随机性,误差的期望为0。
3. 误差与真分数独立。

测试的结果,即观测分数,是误差与真分数的线性和,可用公式 2-1 表示:

$$X = T + e \quad \text{公式 2-1}$$

其中 X 表示观测分数, T 表示真分数, e 表示误差。

虽然经典测试理论已是一个完善的理论系统³⁵,且在传统测试中得到了广泛的应用,但仍然存在以下局限³⁶:

1. 试题的难度估算有赖于所选被试样本的能力和作答情况。在经典测试中,试题的难度由正确作答人数与被试总人数决定,因此被试的样本会对试题的参数产生影响,导致试题参数估计不够客观的问题。
2. 平行测试由于客观条件的限制难以实施,因此对于测试的信度考察较为困难。
3. 缺乏预测力,测试前不能预先估计出被试正确作答试题的概率。

2.1.2 项目反应理论

由于上述经典测试理论中所存在的局限,20世纪50年代提出了项目反应理

³⁴ Gulliksen H. Theory of mental tests [M]. New York: Wiley, 1950.

³⁵ Lord F M, Novick M R, Birnbaum A. Statistical Theories of Mental Test Scores[M]. Statistical theories of mental test scores. UT Back-in-Print Service, 1968.

³⁶ 李伟明, 陈富国. 经典测验理论和项目反应理论对题目分析的对比研究[J]. 心理学报, 1987(3):312-318.

论 (Item Response Theory, IRT)。项目反应理论认为被试对于项目的作答反应是由其某种心理“特质”所决定的, 由于这些特质难以直接观测, 因此可以称为“潜在特质”或“能力”³⁷。被试正确作答项目的概率和其能力水平之间的数学关系可以用项目特征函数 (item characteristic function) 来表示。项目特征函数图像的基本形式如图 2-1 所示:

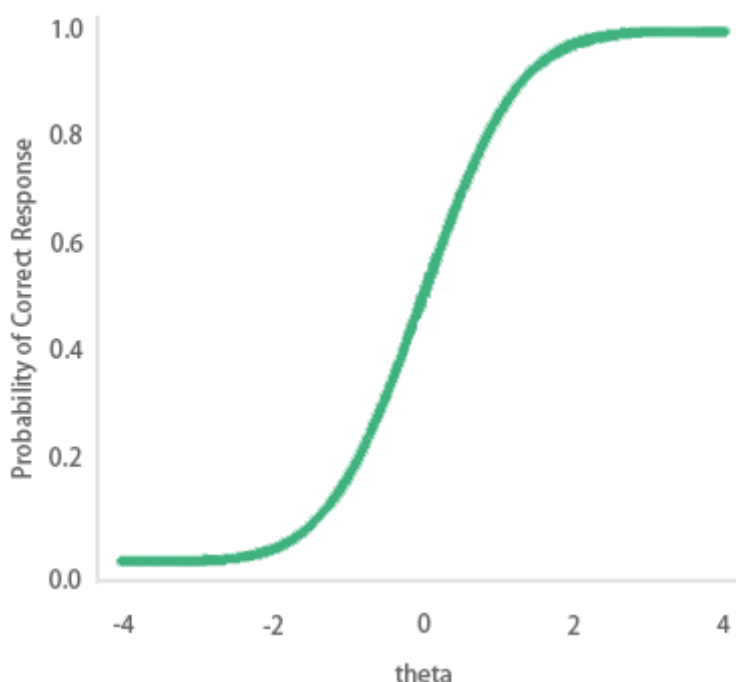


图 2-1 项目特征曲线

该图像称为项目特征曲线 (item characteristic curve, ICC)。图中, 横轴为被试能力水平, 纵轴为做出正确反应的概率。从图像中可以看出曲线单调上升, 被试的能力水平与正确反应的概率呈正相关, 即具有越高能力水平, 能够做出正确反应的概率越大, 而能力水平越低, 做出正确反应的概率越小。由于不同项目的参数不同, 因此对应了不同的项目特征函数和项目特征曲线³⁸。

³⁷ Birnbaum A. Some latent trait models and their use in inferring an examinee's ability[J]. Statistical Theories of Mental Test Scores, 1968:395-479.

³⁸ 金瑜. 心理测量. 第2版[M]. 华东师范大学出版社, 2005.

2.2 项目反应理论的基本假设

项目反应理论主要基于以下三个基本假设：

1. 能力的单维性 (Unidimensionality)

该假设认为项目只针对被试某一种能力或特质进行测试，而其他客观特征，包括心理状态、动机等因素，不会对测试结果产生影响。

2. 局部独立性假设 (Local Independence)

局部独立性假设认为对于某一项目而言，被试的能力水平是唯一决定其反应情况的因素，而对于不同项目之间的反应是相互独立的。该假设可由公式 2-2 表示：

$$\text{Prob}(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta)P(U_2 | \theta) \dots P(U_n | \theta) \quad \text{公式 2-2}$$

其中， $P(U_i | \theta)$ 表示能力为 θ 的被试在第 i 个项目上做出正确反应的概率， $\text{Prob}(U_1, U_2, \dots, U_n | \theta)$ 表示能力值为 θ 的被试在第 1 至 n 个项目上做出正确反应的概率。

3. 单调性假设 (Monotonicity)

单调性假设认为对于任一项目而言，被试的能力与其做出正确反应的概率具有正相关关系。

2.3 项目反应理论的模型

由于被试“特质”或“能力”是潜在的、难以观测，若要建立能力与被试正确反应概率之间的数学关系就需要建立数学模型。目前应用较为广泛的是逻辑斯蒂模型 (Logistic Model)。根据模型所涉及到的参数个数的不同，可以分为三类：单参数逻辑斯蒂模型 (One-Parameter Logistic Model)³⁹、双参数逻辑斯蒂模型 (Two-Parameter Logistic Model) 以及三参数逻辑斯蒂模型 (Three-Parameter Logistic Model)⁴⁰。三种模型的项目特征函数如公式 2-3、

³⁹ Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. [J]. Achievement Tests, 1961:199.

⁴⁰ Birnbaum A. Some latent trait models and their use in inferring an examinee's ability[M].

公式 2-4 以及公式 2-5 所示：

$$\text{单参数模型: } P_i(\theta) = \frac{e^{D(\theta-b_i)}}{1 + e^{D(\theta-b_i)}} \quad i = 1, 2, \dots, n \quad \text{公式 2-3}$$

$$\text{双参数模型: } P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \quad i = 1, 2, \dots, n \quad \text{公式 2-4}$$

$$\text{三参数模型: } P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \quad i = 1, 2, \dots, n \quad \text{公式 2-5}$$

其中， θ 表示被试能力水平； $P_i(\theta)$ 表示能力水平为 θ 的被试在项目 i 上做出正确反应的概率； b_i 表示项目 i 的难度系数； a_i 表示项目 i 的区分度； c_i 表示项目 i 的猜测系数； D 为常量，一般取 $D=1.7$ ； n 表示项目的总数。

单参数模型只涉及到项目的难度系数 b ，双参数模型涉及到项目的区分度 a 和难度 b ，三参数模型涉及到项目的区分度 a 、难度系数 b 和猜测系数 c 。其中，区分度 a 是项目鉴别被试水平高低特性的度量，难度系数 b 是被试对项目做出正确反应时所表现出来的困难程度的度量⁴¹，猜测系数 c 是被试真实能力水平不足以做出正确反应某项目，但在测试中却表现出正确反应的概率的量度。针对某一特定的项目而言，其三参数都是定值，因此被试的能力水平 θ 是唯一决定被试是否能够对项目做出正确的反应的因素。

由公式 2-3、公式 2-4 以及公式 2-5 可以看出，单参数模型是双参数模型当区分度 $a=1$ 时的特殊形式，双参数模型是三参数模型当猜测系数 $c=0$ 时的特殊形式，即可以将单参数模型和双参数模型看作是特殊的三参数模型。因此，本文主要介绍三参数模型的相关理论研究，其他两种模型可由此特殊情况推导而出。

三参数模型项目特征曲线如图 2-2 所示：

Statistical Theories of Mental Test Scores. 1968:395-479.

⁴¹ 冯艳宾，马洪超. 关于经典测量理论和项目反应理论中难度和区分度的探讨[J]. 中国考试, 2012(4):10-14.

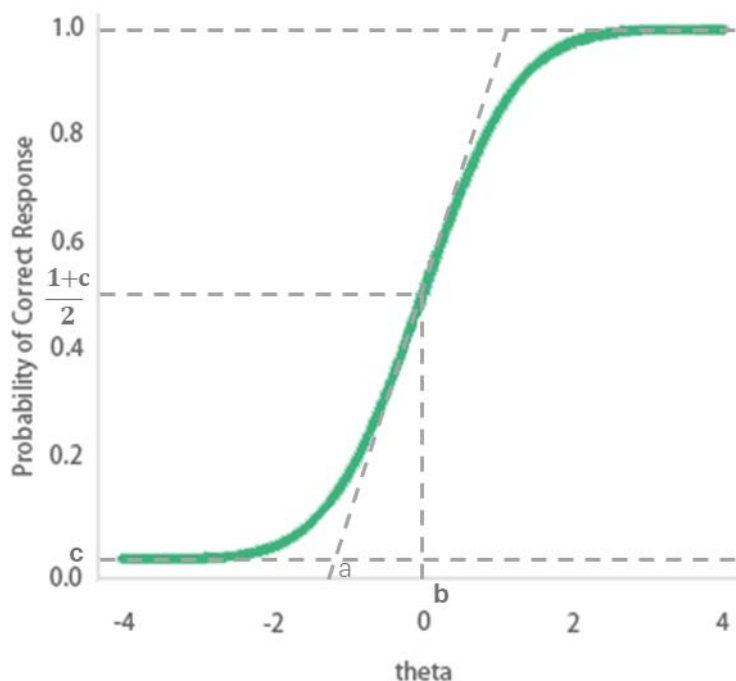


图 2-2 三参数模型项目特征曲线

在图像中，对于项目 i ，曲线与纵轴的截距即为猜测系数 c_i ，答对该项目的概率为 $\frac{1+c}{2}$ 的被测试者的能力值即为难度系数 b_i ，曲线在点 $(b_i, \frac{1+c_i}{2})$ 处的斜率即为其区分度 a_i 。

2.4 项目反应理论的参数估计方法

2.4.1 极大似然估计 (Maximum Likelihood Estimate, MLE)

为了确定项目的参数，需要获得若干被试对于项目的反应结果，据此估计出其参数。一般项目反应理论所讨论的情况为被试的反应是客观的，即被试的反应只分为正确和错误两种情况。因此可以用 1 代表被试做出正确反应的情况，用 0 代表被试做出错误反应的情况。若测试中存在 N 个被试和 n 个项目，则通过测试可以得到 $N \times n$ 个反应结果，并将这些结果列为大小为 $N \times n$ 的矩阵。若用 u_{ij} 表示第 j 个被试对第 i 个项目所做出的反应，则可以得到如下的反应矩阵：

$$\begin{bmatrix} U_{11} & U_{12} & \dots & U_{1j} & \dots & U_{1N} \\ U_{21} & U_{22} & \dots & U_{2j} & \dots & U_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ U_{i1} & U_{i2} & \dots & U_{ij} & \dots & U_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ U_{n1} & U_{n2} & \dots & U_{nj} & \dots & U_{nN} \end{bmatrix}$$

在该矩阵中, U_{ij} 只存在为 0 或为 1 两种情况。 $U_{ij} = 0$ 表示被试 j 在项目 i 上的反应为错误; $U_{ij} = 1$ 表示被试 j 在项目 i 上的反应为正确。

根据局部独立性假设, 对于某一被试, 公式 2-2 可以写为:

$$L(U_1, U_2, \dots, U_i, \dots, U_n | \theta, a, b, c) = \prod_{i=1}^n P_i^{U_i} Q_i^{1-U_i} \quad \text{公式 2-6}$$

其中 $Q_i = 1 - P_i$ 。考虑测试中 N 个被试的情况, 公式 2-6 可以写为:

$$L(U | \theta, a, b, c) = \prod_{i=1}^N \prod_{j=1}^n P_{ij}^{U_{ij}} Q_{ij}^{1-U_{ij}} \quad \text{公式 2-7}$$

根据极大似然估计的基本思想, 设事件 A 发生的概率与因素 x 有关, 若 A 发生了, 则可以认为 x 的取值是当 $P(A|x)$ 取得最大值的情况。因此, 需要估计参数 θ, a, b, c 使得 $L(U | \theta, a, b, c)$ 取得最大值。

由于 $L(U | \theta, a, b, c)$ 与 $\ln L(U | \theta, a, b, c)$ 具有相同的增减性, 因此可以将求 θ, a, b, c 使得 $L(U | \theta, a, b, c)$ 取得最大值的问题转化为求 θ, a, b, c 使得 $\ln L(U | \theta, a, b, c)$ 取得最大值。对公式 2-7 两边取对数, 得到公式 2-8:

$$\ln L(U | \theta, a, b, c) = \sum_{j=1}^N \sum_{i=1}^n [u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln Q_{ij}] \quad \text{公式 2-8}$$

对于公式 2-8, 分别对 θ, a, b, c 求导, 得到如下方程式:

$$\left\{ \begin{array}{l} \frac{\partial \ln L}{\partial \theta_j} = 0 \end{array} \right. \quad \text{公式 2-9}$$

$$\left\{ \begin{array}{l} \frac{\partial \ln L}{\partial a_j} = 0 \end{array} \right. \quad \text{公式 2-10}$$

$$\left\{ \begin{array}{l} \frac{\partial \ln L}{\partial b_j} = 0 \end{array} \right. \quad \text{公式 2-11}$$

$$\left\{ \begin{array}{l} \frac{\partial \ln L}{\partial c_j} = 0 \end{array} \right. \quad \text{公式 2-12}$$

该方程组共有 $N + 3n$ 个方程式, 求解即可得到参数 θ, a, b, c 。

由于计算量过大，采用传统的求解方法过程复杂，且耗时长，因此求解此类方程式通常使用牛顿-拉夫逊迭代法 (Newton-Raphson method)。牛顿-拉夫逊迭代法的具体求解方法是：

设需要求解的方程为 $f(x) = 0$ ，方程的根为 x ，其中一个近似根为 x_0 ，近似根的误差为 dx ，即有 $x = x_0 + dx$ ，则方程 $f(x) = 0$ 可以表示为：

$$f(x_0 + dx) = 0 \quad \text{公式 2-13}$$

将公式 2-13 在 x_0 处泰勒展开，得到：

$$f(x_0 + dx) = f(x_0) + dx f'(x_0) + \frac{(dx)^2}{2} f''(x_0 + \theta dx) = 0 \quad \text{公式 2-14}$$

由于误差 dx 较小，其高次项可以忽略不计，因此可以将公式 2-14 化简为：

$$f(x_0 + dx) \approx f(x_0) + dx f'(x_0) \approx 0 \quad \text{公式 2-15}$$

得误差项：

$$dx = -\frac{f(x_0)}{f'(x_0)} \quad \text{公式 2-16}$$

因此， $f(x) = 0$ 的第一次迭代近似根为 $x_1 = x_0 + dx = x_0 - \frac{f(x_0)}{f'(x_0)}$ 。

将 x_1 看作是初始值，带入所求方程 $f(x) = 0$ 中，进入下一次的迭代计算并得到下一次计算的近似根。重复执行这一过程，直到误差小于设定的值。

牛顿-拉夫逊迭代法的步骤如图 2-3 所示：

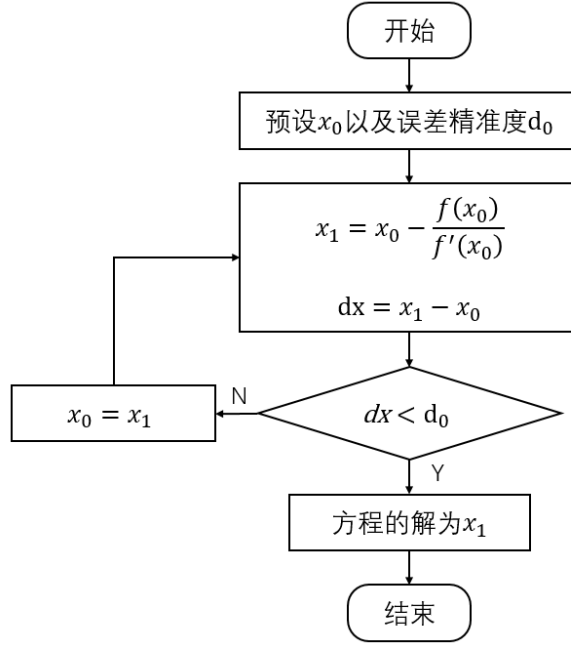


图 2-3 牛顿—拉夫逊迭代法流程图

在利用牛顿—拉夫逊迭代法解决极大似然估计中的方程时，将公式 2-9 至公式 2-12 的方程组看作 $f(x)$ ，按步骤求解即可。

2.4.2 边际极大似然估计

在计算机自适应测试的初始题库建设时，由于被试的能力水平 θ 是未知的，为了估计项目的 a, b, c 参数，通常使用边际极大似然估计的方法消除 θ 参数进行计算。

边际极大似然估计由博克（Bock）和利伯曼（Lieberman）提出，该方法利用概率密度函数进行积分转换⁴²，将等式中的能力参数 θ 消除，得到不含 θ 的似然函数如下：

$$\ln L = c + \sum_{i=1}^{\infty} r_i \cdot \ln(\pi_i) \quad \text{公式 2-17}$$

其中， c 表示常数， r_i 表示项目反应是 $u = \{u_1, u_2, \dots, u_n\}$ 的被试的数量， π_i 表示 $u = \{u_1, u_2, \dots, u_n\}$ 的边际概率。得到该似然函数后，利用 2.4.1 中介绍的方法可以计算出初始题库中项目的参数。

⁴² 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

第3章 c-STR-ST 选题策略提出与模拟实验

3.1 计算机自适应测试中的选题策略概述

3.1.1 最大信息法 (Maximum Fisher Information, MFI)

最大信息法的核心思想是在被试每次作答试题后，根据当前能力估计值，利用项目信息函数 (Item Information Function, IIF) 计算题库中尚未施测的试题所具有的信息量，并从中选择出信息量最大的试题进行测试⁴³。项目信息函数可以表示为：

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad \text{公式 3-1}$$

其中， $I_i(\theta)$ 表示对于能力估计为 θ 的被试试题 i 所具有的信息量； $P_i(\theta)$ 表示能力估计为 θ 的被试对试题 i 做出正确回答的概率； $P'_i(\theta)$ 表示 $P_i(\theta)$ 对 θ 的一阶偏导数； $Q_i(\theta) = 1 - P_i(\theta)$ 。

最大信息法的具体算法如图 3-1 所示：

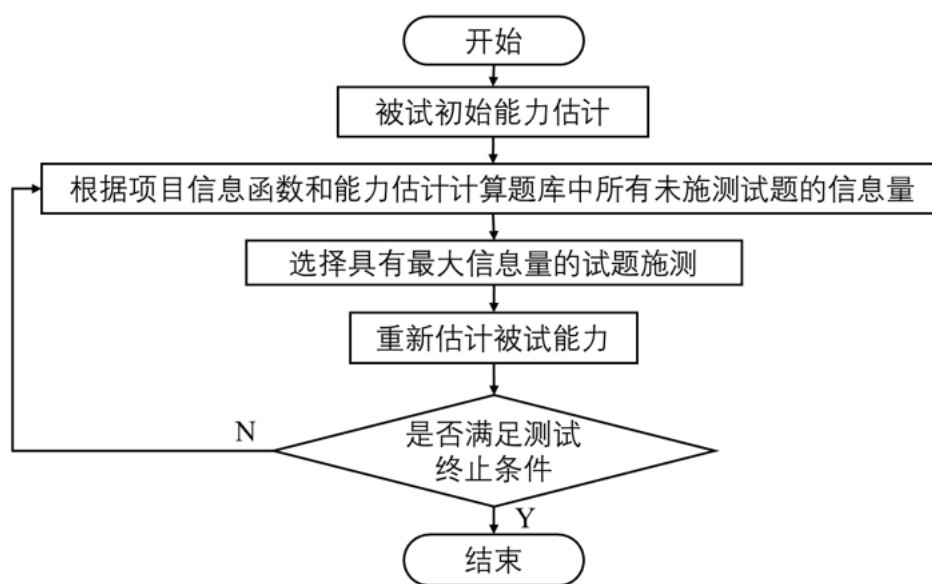


图 3-1 最大信息法算法流程图

从流程图中可以看出，最大信息法选题策略主要包含以下几个步骤：

⁴³ Wainer H. Computerized adaptive testing : a primer[M]. L. Erlbaum Associates, 2000.

- (1) 被试初始能力值估计;
- (2) 利用项目信息函数, 结合试题的参数、能力估计, 计算题库中所有未施测试题具有的信息量;
- (3) 比较所有试题的信息量, 选择最大信息量所对应的试题进行施测;
- (4) 被试作答后, 根据答题情况, 重新对其能力水平进行估计;
- (5) 重复第 2 至第 4 步, 直到满足测试终止条件。

最大信息法是现阶段计算机自适应测试中应用最为广泛的选题策略之一, 该方法虽然提高了测试的准确性, 但是仍然存在一些不足, 如对双参数和三参数逻辑斯蒂模型而言, 该方法倾向于选择高区分度的试题, 从而导致项目曝光不均匀的问题⁴⁴; 另外, 由于试题的信息量是基于被试能力估计计算而来的, 因此最大信息法的结果准确性有赖于能力的真值和估计值之间的差异⁴⁵。

3.1.2 α 分层法 (α -Stratified, STR- α)

α 分层法将测试分阶段的思想引入到自适应测试的选题策略中, 利用试题的区分度参数 α 控制曝光率⁴⁶。该方法的具体步骤如图 3-2 所示:

⁴⁴ 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.

⁴⁵ Chang H H, Qian J, Ying Z. α -stratified multistage computerized adaptive testing with b blocking. [J]. Applied Psychological Measurement, 1999, 23(4):211-222.

⁴⁶ Chang, H & Ying, Z. (1999). α -stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23, 211 - 222

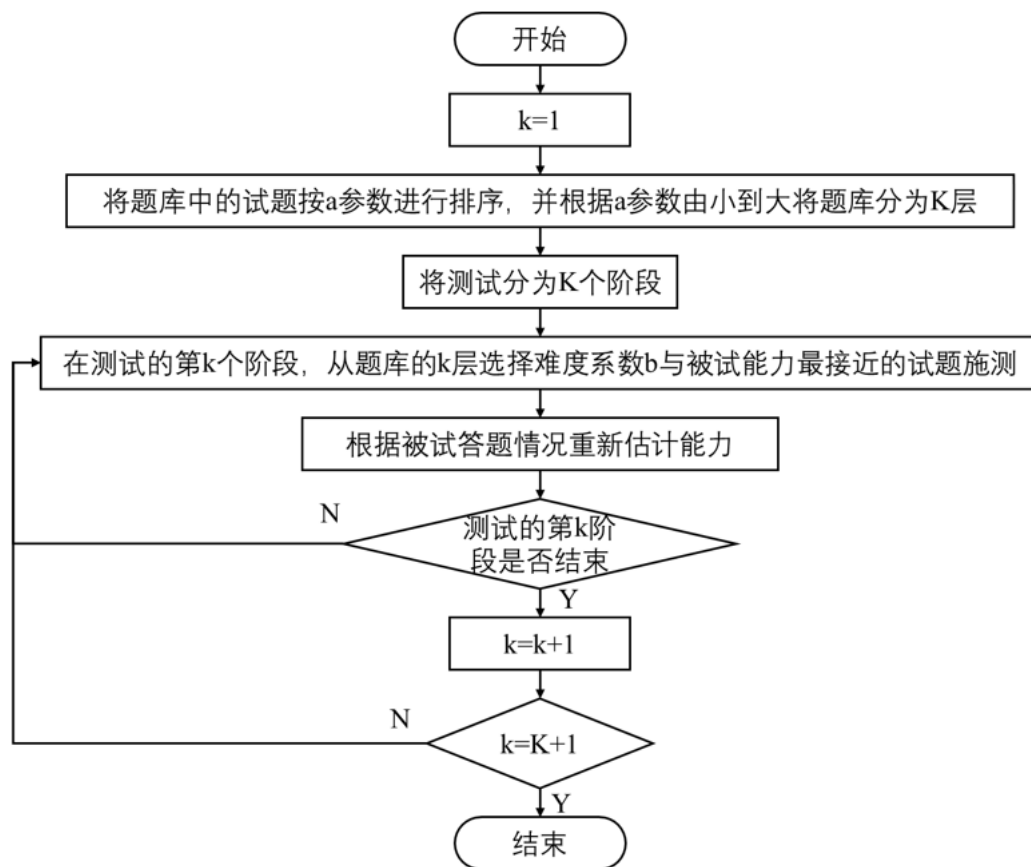


图 3-2 a 分层算法流程图

该方法可概述为:

- (1) 将题库中的试题按参数 a 进行排序, 并根据参数 a 由小到大将题库分为 K 层;
- (2) 将测试分为 K 个阶段;
- (3) 当测试进行到第 k 个阶段时, 从题库所对应的第 k 层中, 选择难度系数 b 与被试能力估计差值最小的试题施测;
- (4) 根据被试答题情况重新估计能力;
- (5) 若第 k 阶段结束则进入第 $k+1$ 阶段;
- (6) 重复第 3 至第 5 步直到完成所有测试阶段。

通常对于区分度低的试题而言被筛选出并施测的频率较低, 采用 a 分层法能够均衡不同区分度的试题的利用率, 因而达到了平衡总体曝光率的目的。但

是 a 分层法仍然存在忽略了试题区分度和难度之间存在的关系的不足⁴⁷。

3.1.3 影子题库 (Shadow Test , ST)

1998 年 van der Linden 和 Reese 提出影子题库的选题策略⁴⁸。影子题库的核心思想是在进行试题选择前,将满足所有约束条件的试题选出并组成影子题库,下一个即将施测的试题将从影子题库中选出,选择方法为最大信息量法。影子题库的算法可以用图 3-3 表示:

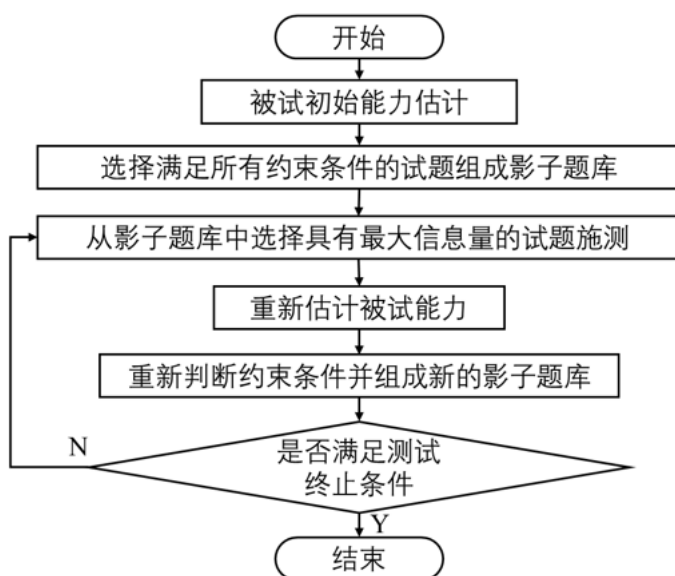


图 3-3 影子题库算法流程图

由流程图可以看出,该选题策略主要包括以下的步骤:

- (1) 被试初始能力估计;
- (2) 根据所有预先设定的约束条件,选择满足所有要求的试题,将这些试题组成影子题库;
- (3) 计算影子题库中的所有试题所对应的信息量,并选择具有最大信息量的进行施测;
- (4) 重新估计被试的能力水平;

⁴⁷ Wingersky, M. S., & Lord, F. M. (1983). An investigation of methods for reducing sampling error in certain irt procedures *. Ets Research Report, 1983(2), i - 52.

⁴⁸ Linden V D, Wim J. |Reese, Lynda M. A Model for Optimal Constrained Adaptive Testing. [J]. Applied Psychological Measurement, 1998, 22(3):259-270.

(5) 根据当前的能力估计, 重新判断题库中的试题是否满足所有约束条件, 并将满足所有约束条件的试题组成新的影子题库;

(6) 重复执行第 3 至 5 步, 直到满足测试终止条件。

影子题库的选题策略能够满足自适应测试中的多种条件约束, 并选择出最佳试题⁴⁹。

3.1.4 内容平衡选题策略

1. 约束 CAT 法 (Constrained CAT, CCAT)

为了实现自适应测试内容平衡约束, 约束 CAT 法将按照测试内容的不同将试题分类, 标定每道试题对应的内容域, 并在测试前设定不同内容域之间的期望比例, 在测试期间, 每次选题时都需要计算当前已测试试题的内容比例, 并与期望值作比较, 下一题将从离期望值最远的内容域中选出⁵⁰。约束 CAT 法的流程如图 3-4:

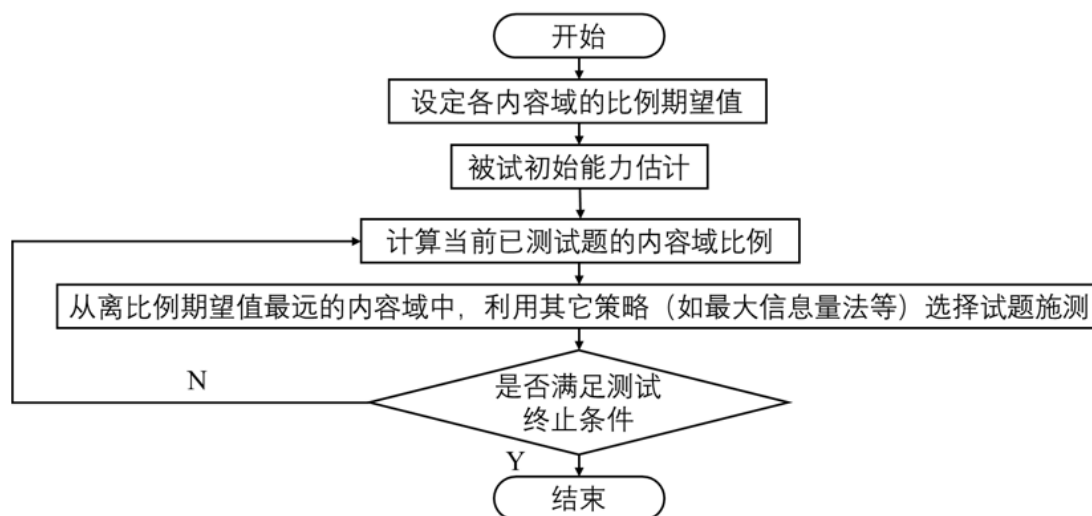


图 3-4 约束 CAT 法算法流程图

约束 CAT 法的选题步骤总结如下:

(1) 设定本次测试的内容域比例期望值;

⁴⁹ Linden V D, Wim J. | Reese, Lynda M. A Model for Optimal Constrained Adaptive Testing. [J]. Applied Psychological Measurement, 1998, 22(3):259-270.

⁵⁰ G. Gage Kingsbury, Anthony R. Zara. Procedures for Selecting Items for Computerized Adaptive Tests[J]. Applied Measurement in Education, 1989, 2(4):359-375.

- (2) 被试初始能力估计;
- (3) 计算当前已测试题的内容域比例;
- (4) 确定当前离期望值最远的内容域, 并利用其它选题策略(如最大信息量法等)选择试题施测;
- (5) 重复第 3 到第 4 步, 直到判定终止条件满足, 结束测试。

约束 CAT 法算法简单且易于执行, 在实现内容平衡的同时保证了测试的准确性和测试长度的合理性⁵¹。但是由于算法中内容域选择的机制使得下一道试题出自哪个内容域具有可预测性。

2. 修正的多项模型 (Modified Multinomial Model , MMM)

为了规避约束 CAT 法存在的可预测性的不足, Chen 和 Ankenman 提出了修正的多项模型选题策略⁵²。该策略在测试前设定内容域的累积分布, 通过随机数与累积分布的比较, 选择出对应的内容域。具体算法如图 3-5 流程图所示:

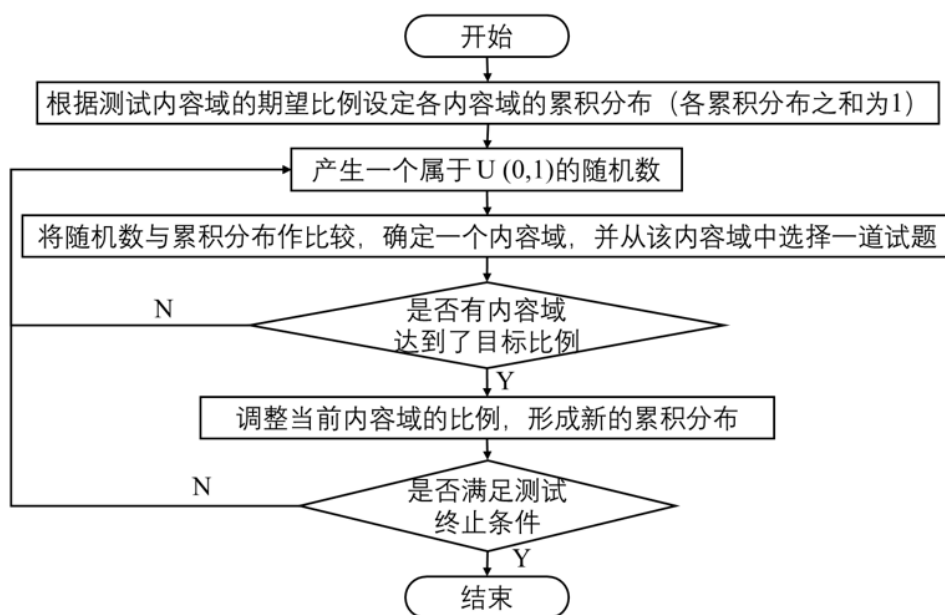


图 3-5 修正的多项模型算法流程图

⁵¹ G. Gage Kingsbury, Anthony R. Zara. Procedures for Selecting Items for Computerized Adaptive Tests[J]. Applied Measurement in Education, 1989, 2(4):359-375.

⁵² Chen S Y, Ankenman R D. Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing[J]. Journal of Educational Measurement, 2004, 41(2):149-174.

该方法的主要步骤有：

- (1) 根据本次测试的需求设定各内容域的期望比例，据此设定各内容域的累积分布，累积分布之和为 1；
- (2) 生成一个服从 $U(0, 1)$ 的随机数；
- (3) 将所产生的随机数与预先设定的累积分布进行比较，确定对应的内容域，并从中选择一道试题施测；
- (4) 重复第 2 至 3 步，直到某一内容域达到目标比例；
- (5) 调整当前的内容域期望比例，形成新的累积分布；
- (6) 重复第 2 至 5 步，直到满足测试终止条件。

修正的多项模型在保证内容平衡的基础上，消除了约束 CAT 所具有的可预测性的不足。

3. c 分层法 (a-Stratified Method with Content-Blocking, STR-C)

为了弥补 a 分层法存在的不足，Chang, Qian 和 Ying 在 a 分层法的基础上加以优化，将难度系数 b 纳入到分层的维度中，提出了 b 分层法⁵³。 b 分层法虽然较 a 分层法效果更佳，但是没有对内容平衡做出约束。因此在此基础上提出了 c 分层法⁵⁴，对试题内容加以约束。

c 分层法的算法流程图如图 3-6 所示；

⁵³ Chang H H, Qian J, Ying Z. a-stratified multistage computerized adaptive testing with b blocking. [J]. Applied Psychological Measurement, 1999, 23(4):211-222.

⁵⁴ Yi, Qing, Chang, et al. a-Stratified CAT design with content blocking[J]. British Journal of Mathematical & Statistical Psychology, 2003, 56(2):359-78.

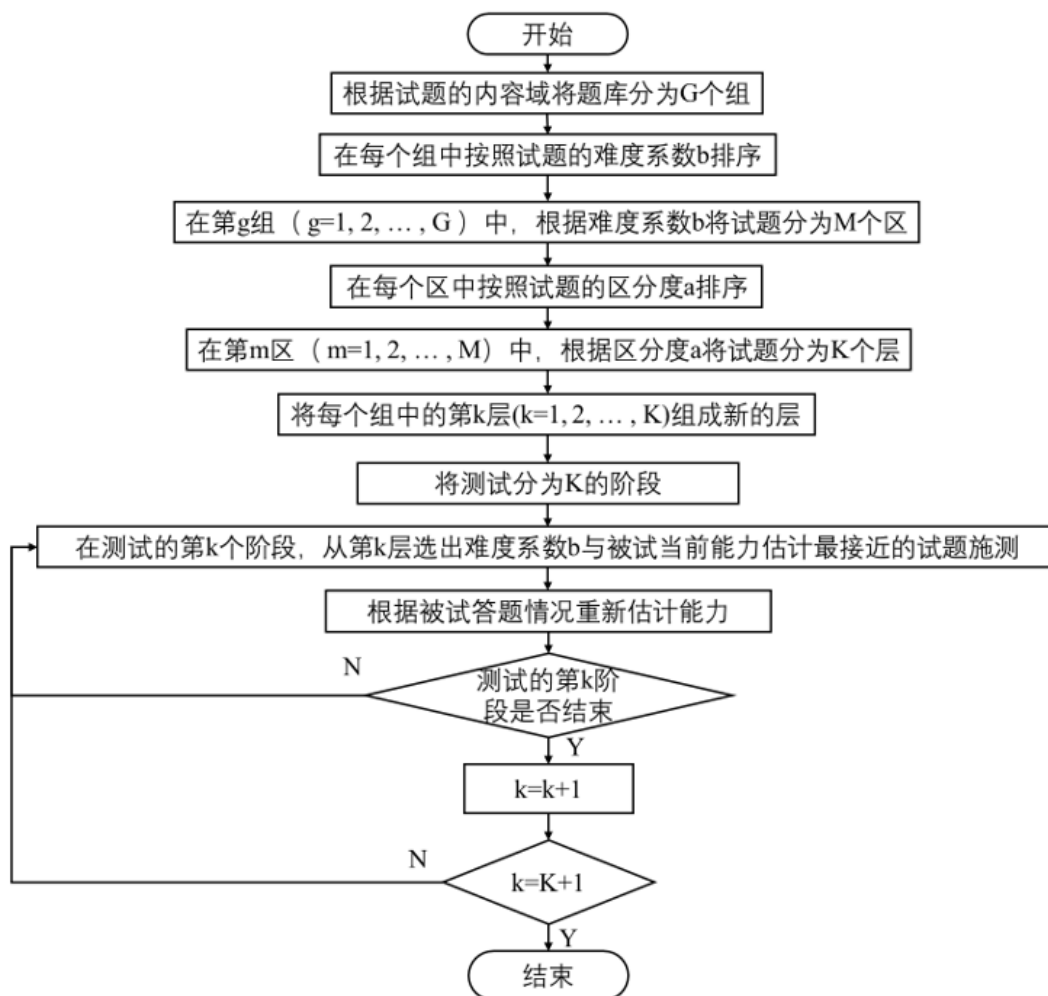


图 3-6 c 分层法算法流程图

c 分层法主要包括以下步骤:

- (1) 根据试题的内容域将题库分为 G 个组;
- (2) 在每个组中按照试题的难度系数 b 排序;
- (3) 在第 g 组 ($g=1, 2, \dots, G$) 中, 根据难度系数 b 将试题分为 M 个区;
- (4) 在每个区中按照试题的区分度 a 排序;
- (5) 在第 m 区 ($m=1, 2, \dots, M$) 中, 根据区分度 a 将试题分为 K 个层;
- (6) 将每个组中的第 k 层 ($k=1, 2, \dots, K$) 组成新的层;
- (7) 将测试分为 K 的阶段;
- (8) 当测试进行到第 k 个阶段时, 从题库所对应的第 k 层中, 选择难度系数 b 与被试能力估计差值最小的试题施测;

(9) 根据被试答题情况重新估计能力水平;

(10) 若第 k 阶段结束则进入第 $k+1$ 阶段;

(11) 重复第 8 至第 10 步直到完成所有测试阶段。

c 分层法在保留了 a 分层法和 b 分层法对于试题曝光度约束特点的基础上, 加上了内容平衡的约束, 提高了测试的精准度。

3.2 c-STR-ST 选题策略的提出

3.2.1 c-STR-ST 选题策略的提出原则

1. 当前内容平衡选题策略存在的问题

现阶段的内容平衡选题策略对于测试的准确性和试题曝光约束关注较少。现有的 CCAT、MMM 和 STR-C 选题策略中, 前两种策略利用最大信息量的方法能够保证测试的准确性, 但是对于试题的曝光率未做出约束; STR-C 策略虽然利用分层的方法保证了测试的安全性, 但是对于测试精准度未做出约束, 且测试的内容域预期比例不能够自定义。

2. 选题策略提出原则

基于上述当前内容平衡选题策略存在的问题, 本研究中提出一种新的选题策略, 在保证内容平衡的同时, 兼顾测试的准确性和试题曝光率。因此对于新选题策略的提出做出如下要求:

(1) 满足测试内容平衡要求。能够根据测试需求, 由施测者自定义测试的预期内容域比例, 各内容域测试用题量满足预期比例。

(2) 保证测试准确性。在自适应出题阶段的选题需要适合于被试的能力水平, 确保测试的自适应性以及结果的准确性。

(3) 控制试题曝光率。为保证测试安全性, 选题策略需要对试题的曝光率做出约束和控制, 避免出现大量曝光率过高或过低的试题。

3.2.2 c-STR-ST 选题策略的提出思想

本研究中将提出一种约束测试内容平衡的方法, 将提出的内容平衡约束方法与现有的最大信息量法、a 分层法以及影子题库的思想结合, 使得测试在达到内

容平衡的同时，确保测试的准确性以及将试题的曝光率控制在合理范围。

该方法主要通过测试前设定期望内容域的比例，在每一次选题时形成影子题库，使得影子题库中的试题内容域比例满足期望比例，再从影子题库中选择试题进行施测。为确保测试的准确性，从影子题库中选题时选用了最大信息量法，计算影子题库中所有试题的信息量，并选择具有最大信息量的试题施测。在控制曝光率方面，采用 α 分层法将题库和测试分层，确保具有不同区分度的试题能够均匀施测。

3.2.3 c-STR-ST 选题策略的主要步骤

基于上述思想，该选题策略的具体选题流程可由图 3-7 表示：

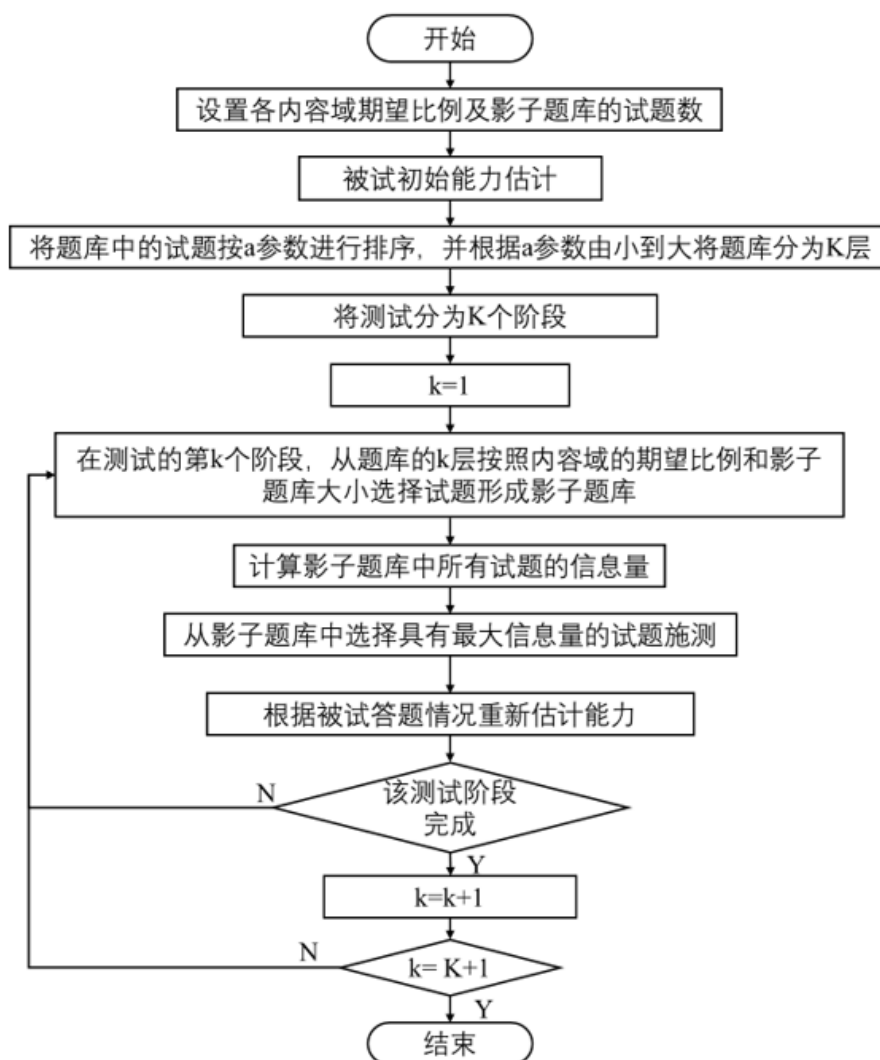


图 3-7 新选题策略算法流程图

新选题策略可描述为以下几个步骤：

(1) 根据本次测试的需要设定各内容域的预期比例，并根据题库大小和测试长度设定影子题库的试题数；

(2) 被试初始能力估计；

(3) 将题库中的试题按照区分度 a 参数进行排序，并将题库分为 K 个层；

(4) 将测试分为 K 个阶段；

(5) $k=1$ ；

(6) 在测试的第 k 个阶段，从题库的 k 层按照内容域的预期比例和影子题库的大小选择试题，组成影子题库；

(7) 计算影子题库中所有试题的信息量；

(8) 选择影子题库中具有最大信息量的试题施测；

(9) 根据被试的答题情况进行能力估计；

(9) 若该测试阶段完成，则 $k=k+1$ ；

(10) 重复执行第 6 至第 9 步，直到完成所有测试阶段。

该算法可描述如下：

```

Begin
输入 内容域预期比例  $q_1:q_2:q_3\cdots$ 
估计被试初始能力值  $\theta$ 
将题库中的试题按照区分度  $a$  参数进行排序，并将题库分为  $K$  个层
 $k=1$ 
for  $k=1$  to  $K$ 
{
    从第  $k$  层选择试题组成影子题库，试题总量为  $s$ 
    影子题库内每个内容域试题数量= $p$  ( $q_1:q_2:q_3\cdots$ )， $p$  为自然数
    for  $i=1$  to  $s$ 
    {

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$
 //计算每道试题的信息量
    }
    next_item=1
    for  $i=1$  to  $(s-1)$ 
    {
        if( $I_{i+1}(\theta) > I_i(\theta)$ )

```

```

        {
            next_item=i+1 //找到具有最大 Fisher 信息量的试题
        }
    }
    施测 next_item
    重新估计被试能力值 $\theta$ 
    if (该测试阶段完成)
    {
        k=k+1
    }
}
End

```

由于该内容平衡选题策略主要融合了分层思想和影子题库的思想，故将其命名为 c-STR-ST。

3.3 内容平衡选题策略算法实验效果分析

为验证 c-STR-ST 在自适应测试中的效果并与现有策略做比较，本研究中采用 Monto Carlo 模拟实验的方法对选题策略的效果进行检验。

Monto Carlo 模拟法是一种基于模拟随机数的统计抽样实验方法⁵⁵，而在自适应测试中，被试的能力水平、试题的参数都可以看作是满足一定的经验概率分布的，因此可以采用 Monto Carlo 模拟法随机产生测试中所需的已知参数，将 Monto Carlo 模拟法运用到自适应测试的模拟实验中⁵⁶。

利用 Monto Carlo 法模拟自适应测试的过程需要模拟的参数有：

1. 模拟题库。模拟题库主要是通过产生随机数的方式生成试题的三个参数，即区分度 a 、难度系数 b 和猜测系数 c 。试题的三个参数符合何种概率分布可以根据研究的需要设定。

2. 模拟被试。模拟被试主要是随机产生被试的能力值，同样可以根据实际研究的需要使得被试的能力满足某种分布。

⁵⁵ 余嘉元，汪存友．项目反应理论参数估计研究中的蒙特卡罗方法[J]．南京师大学报(社会科学版)，2007(1):87-91.

⁵⁶ 余嘉元．项目反应理论研究中的计算机模拟方法[J]．心理科学，1991(2):49-51.

3. 模拟被试反应。对于能力水平为 θ 的模拟被试,根据测试选择的模型不同,选择对应的项目特征函数,即公式 2-3、公式 2-4、公式 2-5,计算答对模拟试题 i 的概率 $P_i(\theta)$,并生成一个服从 $U(0,1)$ 的随机数,将随机数与 $P_i(\theta)$ 作比较,如果随机数小于或等于 $P_i(\theta)$,则认为该模拟被试答对了试题 i ,否则认为该模拟被试答错了试题 i 。

以上三个模拟步骤,结合选题策略、测试终止条件判定和被试能力估计,能够实现整个自适应测试的过程。因此本论文中将采用 Monte Carlo 模拟法,利用 R 语言模拟自适应测试,设定评估标准,对得到的测试结果进行分析,从而对不同选题策略的测试效果进行评估。

3.3.1 题库与被试的生成

1. 题库模拟

本次模拟的自适应测试选用三参数逻辑斯蒂模型,因此需要模拟的试题参数有区分度 a 、难度系数 b 、猜测系数 c 。由于选题策略涉及到测试的内容域,因此试题的参数还应包括试题的内容域。

为验证选题策略在不同结构的题库中的表现效果,在实验中模拟了 4 种题库,这 4 种题库中系数的分布情况为:

题库 1: $\ln a \sim N(0,1)$, $b \sim N(0,1)$, $c \sim \text{Beta}(5,17)$, 各内容域试题数量相等。

题库 2: $\ln a \sim N(0,1)$, $b \sim U(-3,3)$, $c \sim \text{Beta}(5,17)$, 各内容域试题数量相等。

题库 3: $a \sim U(0.2,2.5)$, $b \sim U(-3,3)$, $c \sim \text{Beta}(5,17)$, 各内容域试题数量相等。

题库 4: $a \sim U(0.2,2.5)$, $b \sim N(0,1)$, $c \sim \text{Beta}(5,17)$, 各内容域试题数量相等。

每个参数进行模拟后,根据项目反应理论中对于参数区间的要求进行筛选,即区分度 a 在 $(0.5, 2.5)$ 区间,难度系数 b 在 $(-3, 3)$ 区间,经过筛选后,每个题库中的试题数量不尽相同,但都在 450 至 500 道试题之间。

实验将分别采用上述 4 种结构的题库,对不同选题策略的测试结果进行对比分析,评价选题策略的效果。

(1) 区分度 a

本次模拟测试的题库区分度 a 包括两种结构,分别是满足 $\ln a \sim N(0,1)$ 和满足

$a \sim U(0.2, 2.5)$ 。

对于 $\ln a \sim N(0, 1)$ 分布，其模拟方法为：

- (a) 生成满足 $N(0, 1)$ 的随机数；
- (b) 取 e 的随机数次方，得到满足 $\ln a \sim N(0, 1)$ 条件的区分度 a ；
- (c) 由于自适应测试中试题的区分度在 $(0.5, 2.5)$ 区间为宜，去掉不在该区间的参数，得到题库所有试题的参数 a 。

通过该方法模拟出区分度参数 a ，并确定题库中总题量，用 n 表示。R 语言中 $\ln a \sim N(0, 1)$ 模拟方法如下：

```
x<-rnorm(n=800,mean=0,sd=1) #生成满足 N(0,1)的随机数
a<-exp(x)                      #取 e 的随机数次方
a<-a[a<2.5]
a<-a[a>0.5]                    #区分度在 (0.5,2.5) 区间
n<-length(a)                  #试题个数
```

题库中试题区分度 a 满足 $\ln a \sim N(0, 1)$ 的分布如图 3-8 所示：

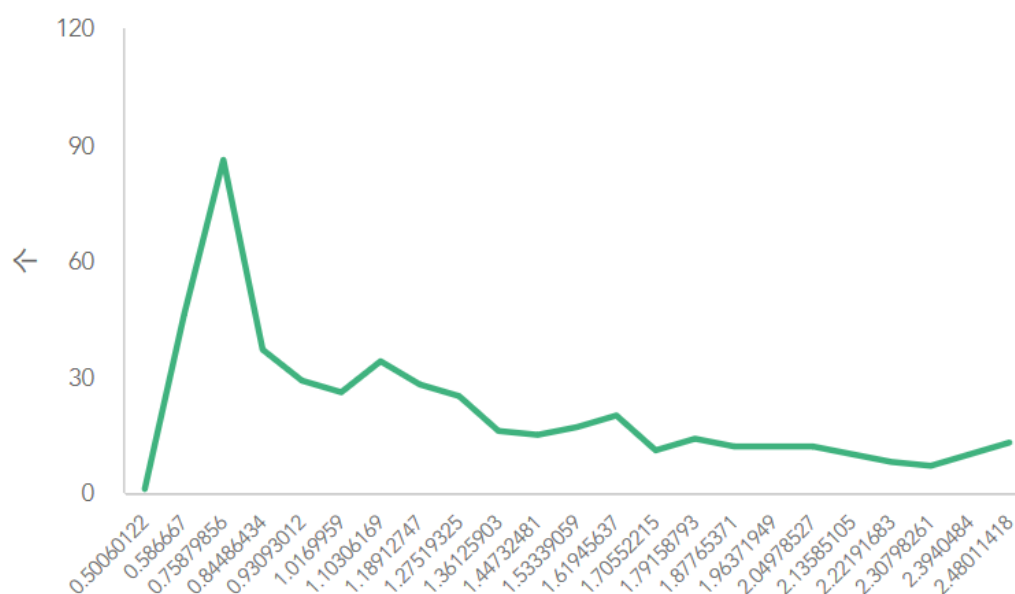


图 3-8 题库中试题区分度 a 满足 $\ln a \sim N(0, 1)$ 分布图

对于 $a \sim U(0.2, 2.5)$ 的情况，模拟方法为产生 0.2 至 2.5 之间的均匀分布随机数，具体为：

```
a<-runif(n,0.2,2.5) #产生 n 个 0.2 至 2.5 之间的均匀分布的随机数
```

题库中试题区分度 a 满足 $a \sim U(0.2, 2.5)$ 的分布如图 3-9 所示：

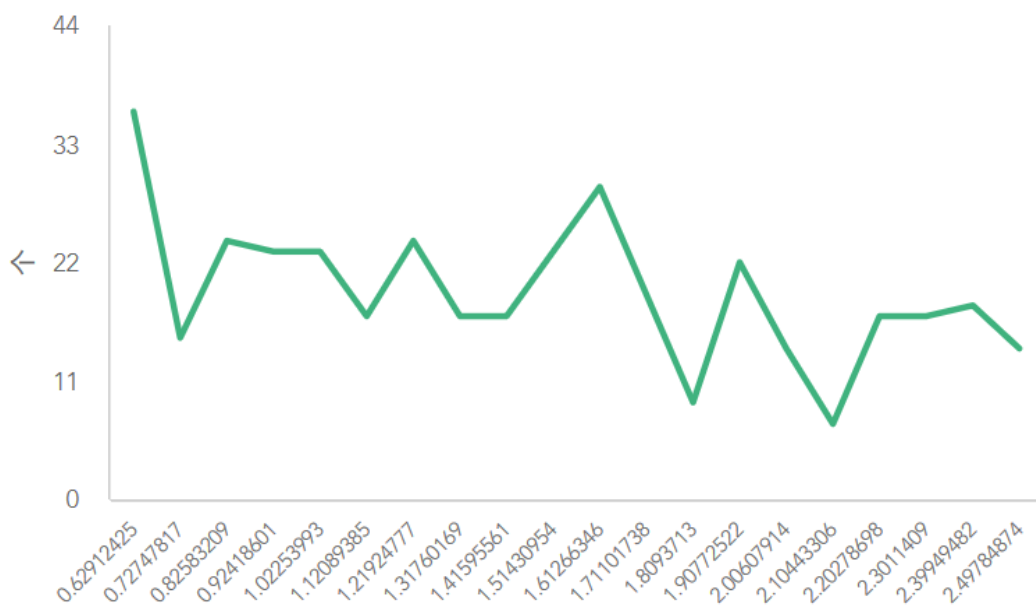


图 3-9 题库中试题区分度 a 满足 $a \sim U(0.2, 2.5)$ 分布图

(2) 难度系数 b

题库难度系数 b 分别满足 $b \sim N(0, 1)$ 和 $b \sim U(-3, 3)$ 两种结构。

$b \sim N(0, 1)$ 结构的模拟方法为：

(a) 生成满足 $N(0, 1)$ 分布的 n 个随机数；

(b) 由于试题的难度与被试的能力水平在数值上应该是统一的，被试的能力水平区间为 $(-3, 3)$ ，因此难度系数也应控制在此区间内。

R 语言具体模拟方法为：

```
b<-rnorm(n,mean=0,sd=1) #生成 n 个满足均值为 0，标准差为 1 的正态
分布随机数
b<-b[b>-3]
b<-b[b<3] #难度系数在 (-3,3) 区间
```

难度系数 b 满足 $N(0,1)$ 的分布如图 3-10 所示:

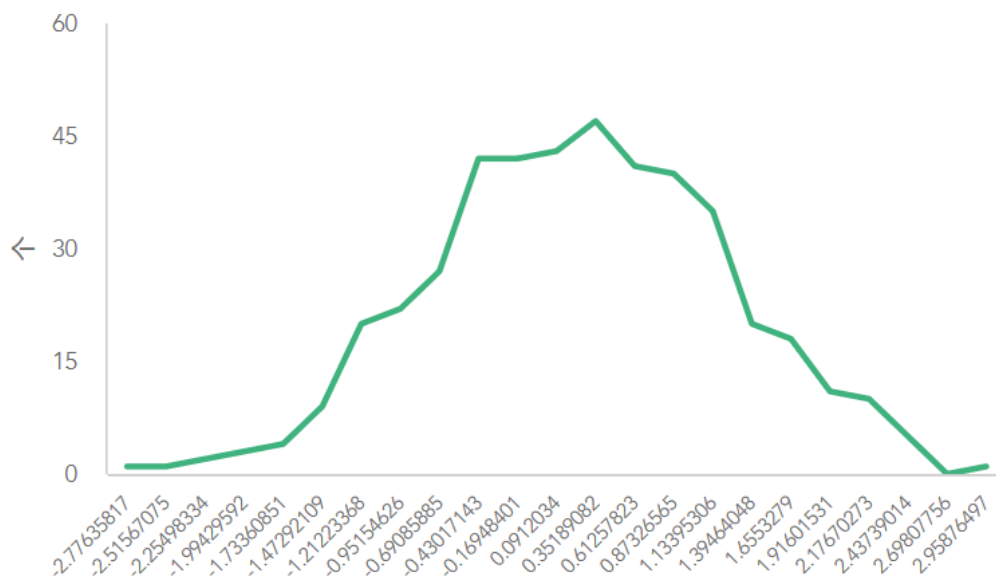


图 3-10 题库中试题难度系数 b 满足 $N(0,1)$ 分布图

$b \sim U(-3,3)$ 的模拟方法为产生 n 个在 $(-3, 3)$ 区间内满足均匀分布的随机数, 具体实现语句为:

```
b<-runif(n,-3,3) #产生 n 个-3 至 3 之间的均匀分布的随机数
```

该结构下难度系数的分布如图 3-11 所示:

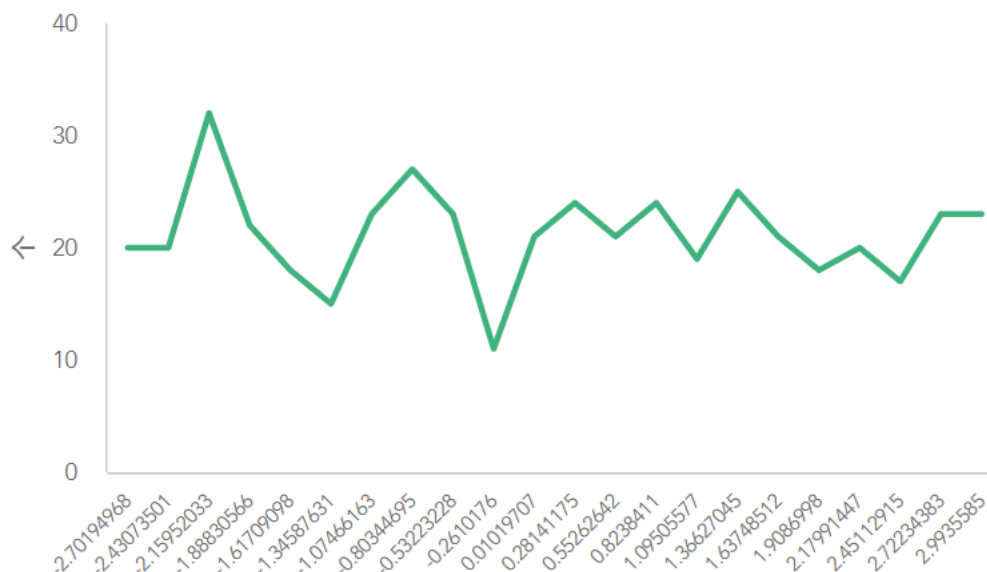


图 3-11 题库中试题难度系数 b 满足 $U(-3,3)$ 分布图

(3) 猜测系数 c

猜测系数 c 满足 $c \sim \text{Beta}(5, 17)$, 模拟方法是产生 n 个满足贝塔分布的随机数, 具体为:

```
c<-rbeta(n,5,17) #产生 n 个系数  $\alpha=5$ ,  $\beta=17$  的贝塔分布的随机数
```

c 参数的分布如图 3-12 所示:

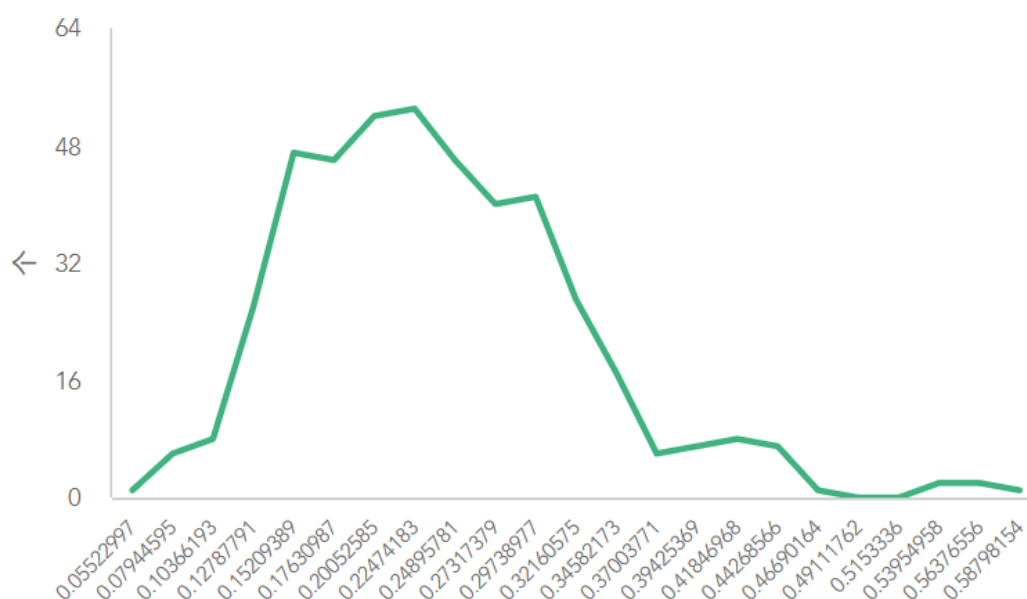


图 3-12 题库中试题猜测系数 c 分布图

(4) 内容域模拟

题库中的模拟试题共属于 5 个内容域, 分别用数字 1 至 5 表示, 且每个内容域占试题总数的 $\frac{1}{5}$ 。

模拟出试题的参数后, 组成一个 $n \times 4$ 的矩阵, 用于存放试题的 4 个参数, 完成题库的模拟。

2. 被试模拟

由于在实际测试中被试的能力水平通常是满足正态分布的, 因此在实验中将使得被试的能力也符合正态分布规律, 即 $\theta \sim N(0, 1)$, 且 $-3 < \theta < 3$ 。模拟实验中设定被试的个数为 1000 个, 用 m 表示, 即 $m=1000$ 。被试模拟方法如下:

```

theta<-rnorm(n=1000,mean=0,sd=1) #产生 1000 个满足正态分布的随机数
theta<-theta[theta>-3]
theta<-theta[theta<3]             #能力水平的区间为 (-3,3)
m<-length(theta)                 #被试个数为 m

```

模拟被试能力分布的分布如图 3-13 所示：

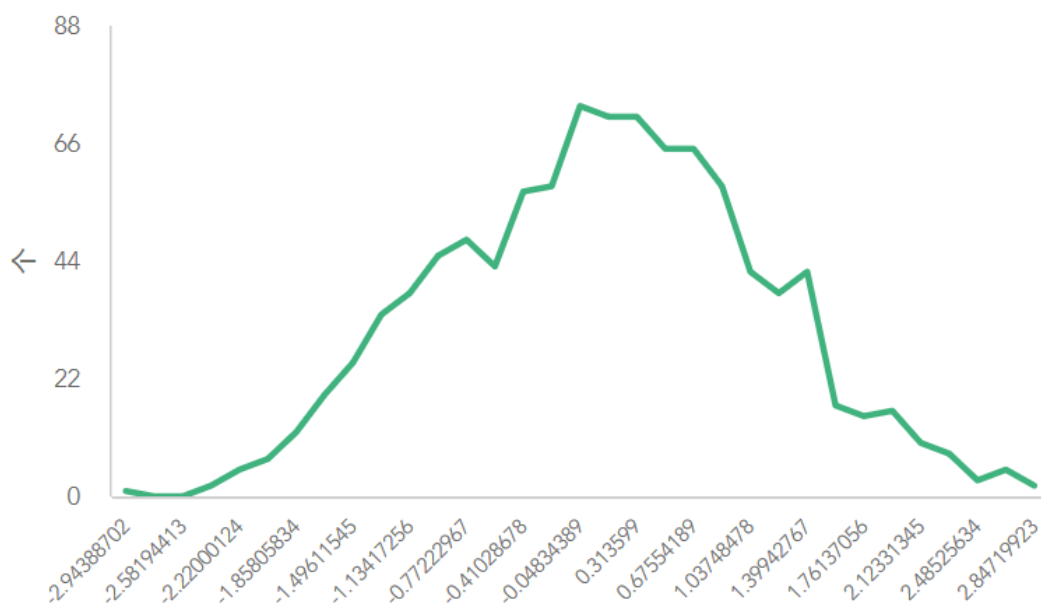


图 3-13 被试能力水平分布图

3.3.2 被试反应向量的生成

被试的反应向量即被试的作答情况，在测试中采用二值计分制，即答对计为 1，答错计为 0。用 U_{ij} 表示被试 i 对于项目 j 的反应，即：

$$\begin{cases} U_{ij} = 0, & \text{被试 } i \text{ 答错项目 } j \\ U_{ij} = 1, & \text{被试 } i \text{ 答对项目 } j \end{cases}$$

对于某一被试，在测试中共作答了 r 题，就有 r 个由 0 和 1 组成的答题情况，将答题情况组成一个具有 r 个元素的向量即为该被试的反应向量。

反应向量的生成有以下三个步骤：

(1) 根据被试的能力水平、试题的参数，计算被试答对该试题的概率。由于本次实验选择了三参数逻辑斯蒂模型，则能力水平为 θ 的被试答对试题 i 的概率，即 $P_i(\theta)$ ，可用公式 2-5 计算得出；

(2) 产生一个服从 $U(0,1)$ 的随机数；

(3) 将随机数与 $P_i(\theta)$ 做比较，若 $P_i(\theta)$ 大于等于随机数，则认为被试答对试题 i ，否则认为被试答错试题 i 。

反应向量的生成方法如图 3-14 所示：

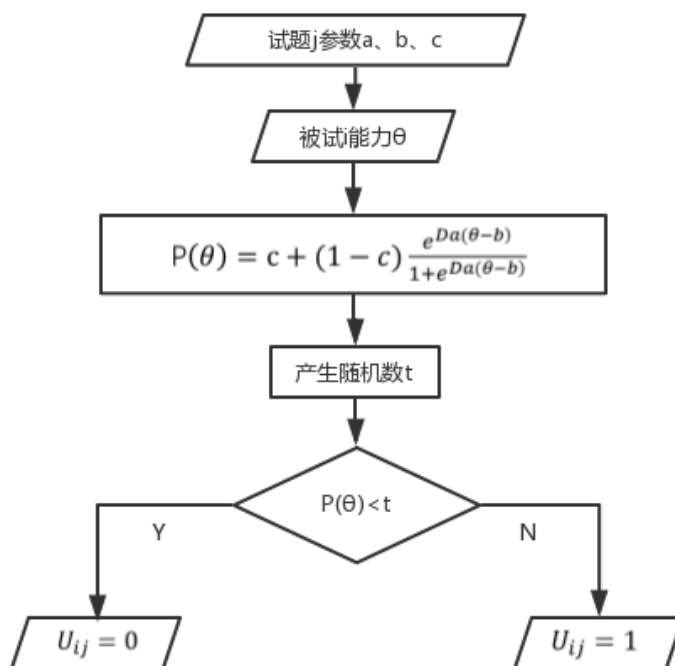


图 3-14 反应向量的生成方法

在 R 语言中被试反应向量的生成方法为：

```

for(i in 1:m)
{
  for(j in 1:n)
  {
    an=item[j,1]
    bn=item[j,2]
    cn=item[j,3]
    thetan=theta[i]
    p_theta[i,j]=cn+(1-cn)*exp(1.7*an*(thetan-bn))/(1+exp(1.7*an*(thetan-bn)))
    #计算 P_i (θ)
    j=j+1
  }
}
  
```

```

    }
    i=i+1
  }
  for(i in 1:m)
  {
    for(j in 1:n)
    {
      y=runif(1,0,1) #产生一个 0 至 1 之间的随机数
      if(p_theta[i,j]<y) #若 P_i (θ)小于随机数
        p_theta[i,j]=0 #认为被试答错
      else
        p_theta[i,j]=1 #否则认为被试答对
      j=j+1
    }
    i=i+1
  }
}

```

通过上述方法以及模拟出的试题参数、被试能力水平，得到所有被试对于所有试题的反应向量。

3.3.3 被试能力估计

1. 被试初始能力估计

由于在测试开始时被试的能力水平是未知的，为了能够利用尽可能少的试题估算被试能力水平的区间，需要对被试初始能力进行估计。现阶段常用的初始能力估计的方法是：第一道试题的选择为难度水平适中，若被试能够正确作答，则下一题稍提高难度，否则下一题稍降低难度，经过若干道题后（题量可自行设定），可由 $\theta = \frac{\text{答对试题数}}{\text{答错试题数}}$ 得到被试的初始能力⁵⁷。

基于上述方法，在本次模拟实验中，被试初始能力估计用到了 4 道试题。对于 CCAT 和 MMM，将采用的方法为：

（1）从题库中随机选择一道难度范围在 $(-1, 1)$ 区间的试题，该试题的难度为 b_1 ；

（2）根据被试第一题的答题情况，若答对，则第二题难度范围为 $(b_1, b_1+0.5)$ ，若答错，则第二题难度范围为 $(b_1-0.5, b_1)$ ，选出第二道题施测，其难度用 b_2

⁵⁷ 漆书青. 现代教育与心理测量学原理[M]. 江西教育出版社, 1998.

表示;

(3) 根据被试第二题的答题情况, 若答对第二题, 则第三题难度范围为 $(b_2, b_2+0.5)$, 若答错第二题, 则第三题难度范围为 $(b_2-0.5, b_2)$, 选出第三道题施测, 其难度用 b_3 表示;

(4) 根据被试第三题的答题情况, 若答对第三题, 则第四题难度范围为 $(b_3, b_3+0.5)$, 若答错第三题, 则第四题难度范围为 $(b_3-0.5, b_3)$;

(5) 根据对上述四道题的答题情况, 被试初始能力计算方法为 $\theta = \frac{\text{答对试题数}}{\text{答错试题数}}$ 。特别地, 若四题全部答对, 则 $\theta = 3$ 。

对于 STR-C 以及 c-STR-ST, 由于试题已进行分层, 被初始能力估计阶段具体方法如下:

(1) 将题库按照区分度由大到小分为四层;

(2) 第一道试题从第 1 层选出, 难度范围为 $(-1, 1)$, 该试题的难度用 b_1 表示;

(3) 第二道试题从第 2 层选出, 根据被试的答题情况, 若答对第一题, 则第二题难度范围为 $(b_1, b_1+0.5)$, 若答错第一题, 则第二题难度范围为 $(b_1-0.5, b_1)$, 选出第二道题施测, 其难度用 b_2 表示;

(3) 第三道试题从第 3 层选出, 根据被试的答题情况, 若答对第二题, 则第三题难度范围为 $(b_2, b_2+0.5)$, 若答错第二题, 则第三题难度范围为 $(b_2-0.5, b_2)$, 选出第三道题施测, 其难度用 b_3 表示;

(4) 第四道试题从第 4 层选出, 根据被试的答题情况, 若答对第三题, 则第四题难度范围为 $(b_3, b_3+0.5)$, 若答错第三题, 则第四题难度范围为 $(b_3-0.5, b_3)$;

(5) 根据四道题的答题情况, 初始能力计算方法为 $\theta = \frac{\text{答对试题数}}{\text{答错试题数}}$ 。特别地, 若四题全部答对, 则 $\theta = 3$ 。

R 语言中选择第一道难度范围为 $(-1, 1)$ 的试题的实现方法为:

```

x<-sample(1:n, size = 1)    #随机选择一道试题
while(1<item[x,2]||-1>item[x,2])
{
    x<-sample(1:n, size = 1) #若所选试题的难度不在（-1,1）范围则重新选择
}

```

第二至四题的选题方法为：

```

if(ans[length(ans)]==0) #判断上一题的答题情况，若为错误
{
    x<-sample(1:n, size = 1)    #随机选择试题
    while(b_l<item[x,2]||item[x,2]<(b_l-0.5))
    {
        x<-sample(1:n, size = 1)    #若所选试题的难度不在（上一题的难度-0.5，
上一题的难度）范围内则重新选择
    }
}
if(ans[length(ans)]==1) #判断上一题的答题情况，若为正确
{
    x<-sample(1:n, size = 1)    #随机选择试题
    while(b_l>item[x,2]||item[x,2]>(b_l+0.5))
    {
        x<-sample(1:n, size = 1) #若所选试题的难度不在（上一题的难度，上一题
的难度+0.5）范围内则重新选择
    }
}

```

初始能力的计算实现方法为：

```

if(sum(ans)==4)
{
  theta_e[i] <- -3 #若 4 题全部答对，初始能力为 3
}
else
{
  theta_e[i] <- sum(ans)/(4-sum(ans)) #计算初始能力
}

```

2. 测试中被试能力估计

基于被试的初始能力估计，模拟实验中将采用不同的选题策略进行若干次实验，被试每作答一道试题，根据作答情况，能力水平估计方法采用 2.4.1 节中介绍的极大似然估计法。

R 语言中能力估计的实现方法为：

```

if (sum(ans)==length(ans)) {
  theta_e <- -3 #若全部答对，能力估计为 3
}
else if (sum(ans)==0) {
  theta_e <- -3 #若全部答错，能力估计为 -3
}
else {

theta_e <- optimize(log_e, interval=c(-3,3), maximum=TRUE, item_par=item_par, d=d,
ans=ans)$max
#极大似然估计
}

log_e <- function
(

```

```
item_par,  
d,  
theta,  
ans  
{  
  p <- p_3(item_par,d,theta) #计算  $p_i(\theta)$   
  log_est <- sum(log(p)*ans+log(1-p)*(1-ans))  
  return(log_est)  
}
```

3.3.4 测试终止条件

现阶段常用的自适应测试终止条件有：

（1）固定长度。在测试前设定本次测试每个被试需要作答的试题数，当被试做完规定的题数后判定测试结束。

（2）比较连续两次能力估计的数值，若两值之差小于设定值，则结束测试。

（3）在每次能力估计后，计算其标准差，若标准差小于设定的阈值，则判定结束测试。

上述三种方法中，第 2 和第 3 种方法较第 1 种方法能够得到更高的测试精准度，但是存在测试长度可能过长的问题，在实际应用中可操作性较低。此外，阈值的设定需要在实践中多次检验⁵⁸，过程较为复杂。相较而言，第 1 种固定长度的方法实施简单、可行性高，考虑到在实际应用中也多采用固定长度的终止测试方式，因此在本次模拟实验中选择该方法判定测试的终止，且单次测试长度设定为 40 道试题。

3.4 内容平衡选题策略测试效果比较分析

本研究中提出的 c-STR-ST 选题策略主要针对自适应测试中的试题内容平衡

⁵⁸ 王勤云. 计算机自适应测验中选题策略的分析比较[D]. 山东师范大学, 2012.

约束, 因此在与新策略的效果做对比时, 选用现有的同样具有内容平衡约束的选题策略进行比较。常用的内容平衡选题策略在 3.1.4 节中做出了详细介绍。

本次对比实验将选择 CCAT、MMM、STR-C 三种内容平衡选题策略与 c-STR-ST 进行对比。四种选题策略的核心算法实现方法见附录。模拟实验中题库的结构、被试的能力分布相同, 试题的内容域总数均为 5 个。对于 CCAT、MMM 和 c-STR-ST, 由于测试的期望内容域比例可以由用户设置, 模拟实验中定为 1:2:2:2:3。c-STR-ST 的影子题库大小为 10。因此除选题策略外, 其余客观条件均认为是相同的。

在上述条件下, 评估三种现有选题策略与 c-STR-ST 在测试精确度、试题曝光率等方面的表现, 据此对 c-STR-ST 的效果做出评价。

3.4.1 内容平衡选题策略测试效果评价指标

选题策略作为自适应测试的核心, 对于测试的准确性、试题的曝光率、测试重叠率等都有重要的影响, 因此在对选题策略的效果做出评价时也应该从这几个方面出发。本实验中用于评价选题策略的指标有:

1. 测试准确性相关指标

测试的准确性即对被试能力估计的准确程度。常用的评价测试准确性的指标有:

(1) 能力估计平均偏差 (bias)

bias 用于评价测试对能力水平估计的精确程度, 计算公式为:

$$\text{bias} = \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \theta_i) \quad \text{公式 3-2}$$

其中 m 表示测试总人数, $\hat{\theta}_i$ 表示能力估计值, θ_i 表示能力真值。由 bias 的计算公式可以看出, bias 越接近 0, 则能力估计与真实能力越接近, 测试的准确性越高。

(2) 能力估计均方误差 (mse)

mse 是用于考察能力估计准确性的度量, 其计算方法为:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \theta_i)^2 \quad \text{公式 3-3}$$

mse 越小，则能力估计值越接近真实能力水平，测试准确性越高。

2. 试题曝光度相关指标

(1) 试题曝光度

试题曝光度是指试题在测试中使用的频数。其计算方法为：

$$A_j = \frac{t_j}{m} \quad \text{公式 3-4}$$

其中 A_j 表示试题 j 的曝光度， t_j 表示试题 j 被选择并施测的次数， m 表示被试总人数。

理论上认为所有试题的曝光度应该是一致的⁵⁹，但是在真实的测试环境下难以实现曝光度完全一致。因此测试后题库中试题的曝光度分布相对均匀的选题策略可以认为测试安全性更高。

(2) 未使用试题数量

对于测试而言，除了保证试题的曝光度基本一致外，为了防止题库中的试题因为从未使用过，而造成“浪费”的情况，对于选题策略的评价标准中还应包括对于未使用过的试题数量的统计。未使用试题的数量越少，代表测试中“浪费”的情况越少，选题策略越优。

(3) χ^2 检验统计量

由于可能出现试题曝光度差异较小的情况，将对曝光情况做进一步分析，引入 χ^2 检验统计量，其计算公式为：

$$\chi^2 = \sum_{j=1}^N \frac{(A_j - \bar{A}_j)^2}{\bar{A}_j} \quad \text{公式 3-5}$$

其中 N 表示题库中的总题数， \bar{A}_j 表示试题曝光度的平均值。

χ^2 检验统计量可以反映试题曝光率的均匀程度， χ^2 检验统计量越小，曝光率越均匀。

3. 测试重叠率指标

⁵⁹ 王勤云. 计算机自适应测验中选题策略的分析比较[D]. 山东师范大学, 2012.

测试重叠率 (TOR) 是随机选出的两个被试所作答的试题的重复比率⁶⁰。在传统的测试中, 由于所有被试作答同一份试题, 测试重叠率为 1, 可能导致测试结果不准确、舞弊等问题。自适应测试的一个重要优势在于弥补了传统测试中“千人一卷”的弊端, 因此, 在评价选题策略时, 也应对测试的重叠率进行比较和分析。

在计算 TOR 时, 需要计算任意两个被试在本次测试中使用的相同试题的个数。若测试中被试总人数为 m , 则共有 $\frac{m(m-1)}{2}$ 种组合, 因此需要计算 $\frac{m(m-1)}{2}$ 次两个被试使用相同的试题数。将这 $\frac{m(m-1)}{2}$ 个数相加, 除以总数 $\frac{m(m-1)}{2}$, 即为测试重叠率 TOR。TOR 的数学表达式为:

$$TOR = \frac{\sum_{p=1}^{m-1} \sum_{q=p+1}^m R_{pq}}{\frac{m(m-1)}{2}} \quad \text{公式 3-6}$$

其中, R_{pq} 表示第 p 个被试和第 q 个被试在测试中使用的相同试题数量。TOR 越小, 说明不同被试使用相同试题的情况越少, 测试的安全性也越好。

3.4.2 内容平衡选题策略测试效果实验结果

在其他测试条件相同的情况下, 更换模拟实验中使用的选题策略, 对四种题库结构下 c-STR-ST、CCAT、MMM 和 STR-C 各进行了 10 次测试, 对测试的以下几个方面进行比较, 评价不同选题策略的效果, 结果如表 3-1、表 3-2、表 3-3 以及表 3-4 所示:

⁶⁰ Way W D. Protecting the Integrity of Computerized Testing Item Pools[J]. Educational Measurement Issues & Practice, 1998, 17(4):17-27.

表 3-1 题库 1 模拟实验结果

评价指标	CCAT	MMM	STR-C	c-STR-ST
bias	0.00897	0.01136	0.01969	0.00778
mse	0.02801	0.029694	0.045199	0.044174
平均曝光率	0.087394	0.089608	0.087135	0.087038
未使用的试题数	0	0	0	0
χ^2 检验统计量	84.48118	80.95065	9.52575	22.67328
测验重叠率	0.268508	0.267695	0.106034	0.135733

表 3-2 题库 2 模拟实验结果

评价指标	CCAT	MMM	STR-C	c-STR-ST
bias	0.00829	0.01284	0.02691	0.00942
mse	0.029283	0.030838	0.049761	0.049284
平均曝光率	0.113722	0.107933	0.090498	0.0906
未使用的试题数	0	0	0	0
χ^2 检验统计量	95.73946	83.89912	30.8454	34.97332
测验重叠率	0.328167	0.295537	0.159284	0.168715

表 3-3 题库 3 模拟实验结果

评价指标	CCAT	MMM	STR-C	c-STR-ST
bias	0.01512	0.0079	0.0195	0.0072
mse	0.024268	0.02556	0.034116	0.03406
平均曝光率	0.123769	0.116764	0.103708	0.103951
未使用的试题数	0	0	0	0
χ^2 检验统计量	67.77292	64.74476	28.78927	36.90022
测验重叠率	0.296732	0.281949	0.177275	0.198497

表 3-4 题库 4 模拟实验结果

评价指标	CCAT	MMM	STR-C	c-STR-ST
bias	-0.002722	0.00188	-0.007316	-0.002742
mse	0.018401	0.017577	0.033884	0.031735
平均曝光率	0.10485	0.104548	0.103627	0.103896
未使用的试题数	0	0	0	0
χ^2 检验统计量	59.38077	50.40645	8.370708	26.05118
测验重叠率	0.25719	0.233674	0.12438	0.170381

3.5 内容平衡选题策略测试效果比较与讨论

3.5.1 测试准确性

1. 能力估计平均偏差 (bias)

根据 4 种题库的模拟结果, c-STR-ST 的 bias 在题库 1 和题库 3 中表现优于其他三种策略;在题库 2 中低于 CCAT, 优于其他两种方法;在题库 4 中低于 MMM, 与 CCAT 基本持平, 优于 STR-C。

由于 CCAT、MMM 与 c-STR-ST 融合了最大信息法的选题思想, 因此从 bias 的结果上可以看出这三种选题策略的准确性高于 STR-C, 且综合四种题库的结果来看, 与 CCAT 与 MMM 相比, c-STR-ST 的 bias 数值最接近 0。

2. 能力估计均方误差 (mse)

在模拟实验的结果中, 在 4 种题库结构下, 均是 CCAT 和 MMM 的 mse 值小于 c-STR-ST, 测试准确性最好, 而 STR-C 的 mse 值最大, 测试准确性与其他三种选题策略相比略差。

通过模拟实验结果可以看出, 在测试准确性方面, 比较 bias 和 mse 的计算方法可以看出, c-STR-ST 总体测试准确性较好。从 bias 指标来看 c-STR-ST 方法的测试结果最好, 而从 c-STR-ST 的 mse 指标较 CCAT 和 MMM 高可以看出, CCAT 和 MMM 选题策略中被试的能力水平估计高于真值和低于真值的情况较为平均, 而

c-STR-ST 能力水平估计高于真值的情况较多。

3.5.2 试题曝光度

模拟实验中用到的四种选题策略的试题曝光度均值差别不大, 选用 CCAT 和 MMM 策略时试题曝光度略高于 STR-C 和 c-STR-ST。此外, 四种情况下都不存在曝光率为 0 的试题。

四种选题策略的试题曝光曲线在不同题库结构下基本相同, 选择题库 1 中各选题策略下的一个曝光曲线为例, 如图 3-15、图 3-16、图 3-17 和图 3-18 所示:

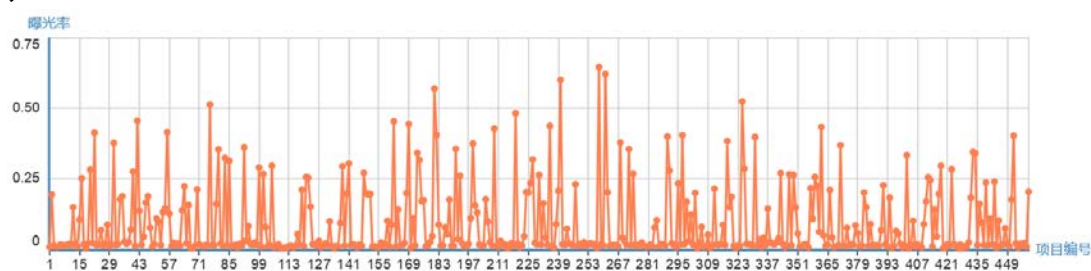


图 3-15 采用 CCAT 选题策略试题曝光曲线

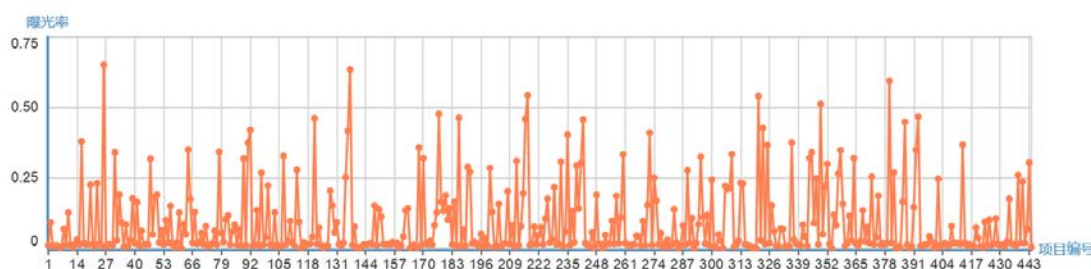


图 3-16 采用 MMM 选题策略试题曝光曲线

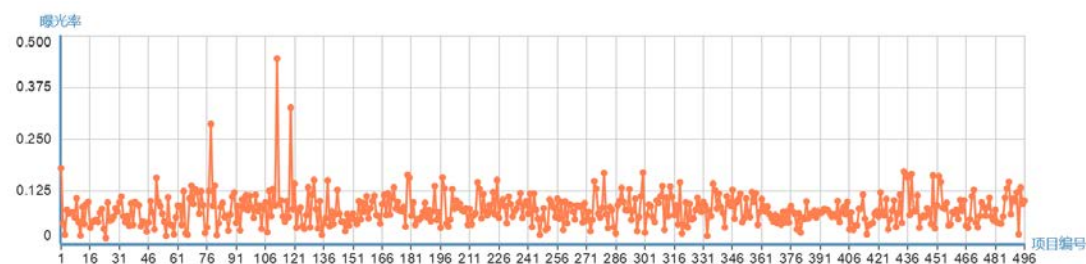


图 3-17 采用 STR-C 选题策略试题曝光曲线

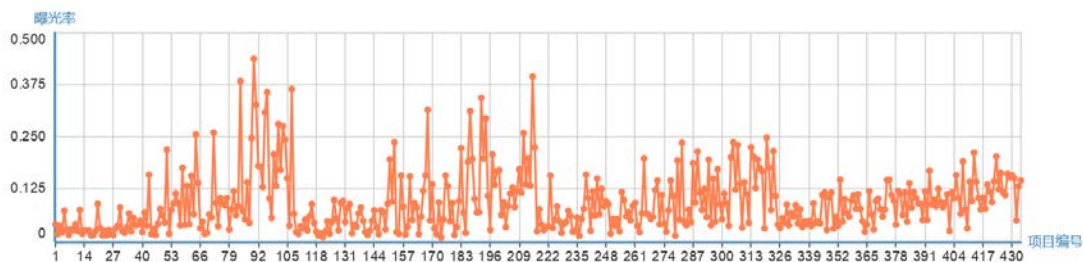


图 3-18 采用 c-STR-ST 选题策略试题曝光曲线

从曝光曲线可以看出, CCAT 和 MMM 的试题曝光率分布范围较大, 且存在曝光率大于 0.5 的试题; STR-C 的曝光率分布最均匀, 除个别试题曝光率较高外, 大多数试题的曝光率在 (0, 0.25) 的区间; c-STR-ST 的曝光率较均匀, 大多数试题的曝光率在 (0, 0.25) 的区间, 但是落在区间外的试题数量较 STR-C 多。

为进一步分析试题曝光情况, 评价中引入了 χ^2 检验统计量。由实验结果可以看出, CCAT 和 MMM 的 χ^2 检验统计量远大于 STR-C 和 c-STR-ST, 曝光率最不均匀。STR-C 曝光均匀性最好, 其次是 c-STR-ST。

在测试安全性方面, c-STR-ST 较 CCAT 和 MMM 明显偏好, 与 STR-C 相比, 在测试平均曝光率上基本持平, 虽然总体曝光率控制在合理的区间内, 但是仍然存在题库每个层中有个别试题曝光率过高的问题, 因此在其他安全性指标上较 STR-C 略有不足。

3.5.3 测试重叠率

在 4 种题库结构下的模拟实验中采用 STR-C 的测试重叠率最低、测试安全性最好, 其次是 c-STR-ST, CCAT 和 MMM 测试重叠率最高。

3.5.4 测试内容域

对于 CCAT、MMM 和 c-STR-ST 三种选题策略, 测试的内容域比例是可以由施测者自定义的, 在模拟实验中设定 5 个内容域的预期比例为 1:2:2:2:3。选用 CCAT 和 MMM 的模拟实验中, 10 次测验所用试题内容域的比例为 1:2:2:2:3。选择 c-STR-ST 的模拟实验中, 10 次测试所用试题的内容域比例为 1:1.9:2.0:2.0:2.8。而由于 STR-C 不可以自定义测试的内容域期望比例, 且题库中每个内容域占试题总数的 $\frac{1}{5}$, 因此 10 次测试所用试题的内容域比例约为 1:1:1:1:1。

在内容平衡方面, CCAT、MMM 和 c-STR-ST 都能够自定义测试内容域的比例,

且各内容域测试用题数量基本满足预期比例。而 STR-C 由于不能够自定义内容域预期比例，因此各内容域测试用题数量与题库中试题内容域比例相关性较大。

除上述定量指标外，在算法用时上 c-STR-ST 也有较大优势。由于组成了影子题库，且影子题库在保证了试题符合当前选题要求的情况下试题量较整个题库大大减少，因此在计算试题信息量时需要计算的试题数量少，因而整个测试过程中算法的总计算量较少，能够提高测试的流畅性。

3.5.5 综合评价

根据各选题策略在模拟实验中可量化指标进行分析并可视化如图 3-19、图 3-20、图 3-21、图 3-22 所示：

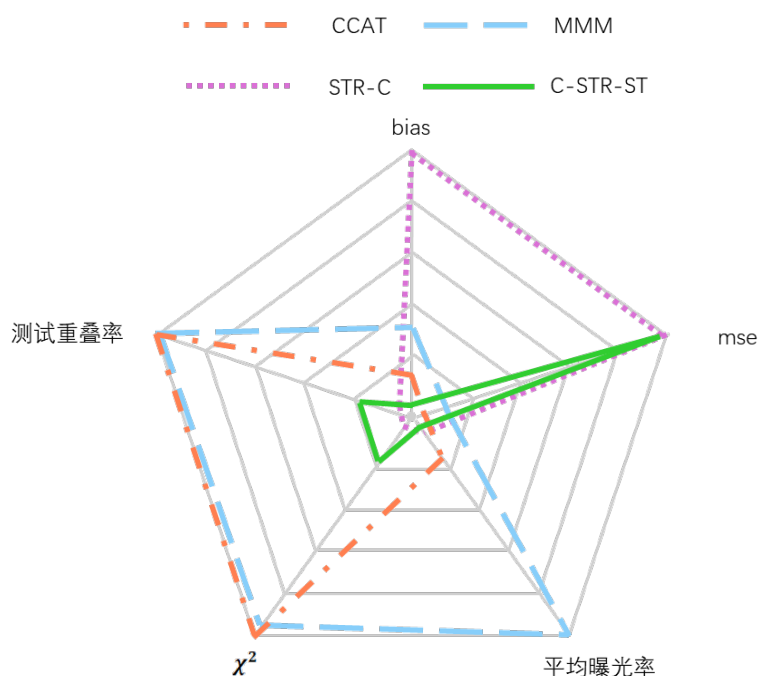


图 3-19 四种选题策略在题库 1 模拟实验中可量化指标雷达图

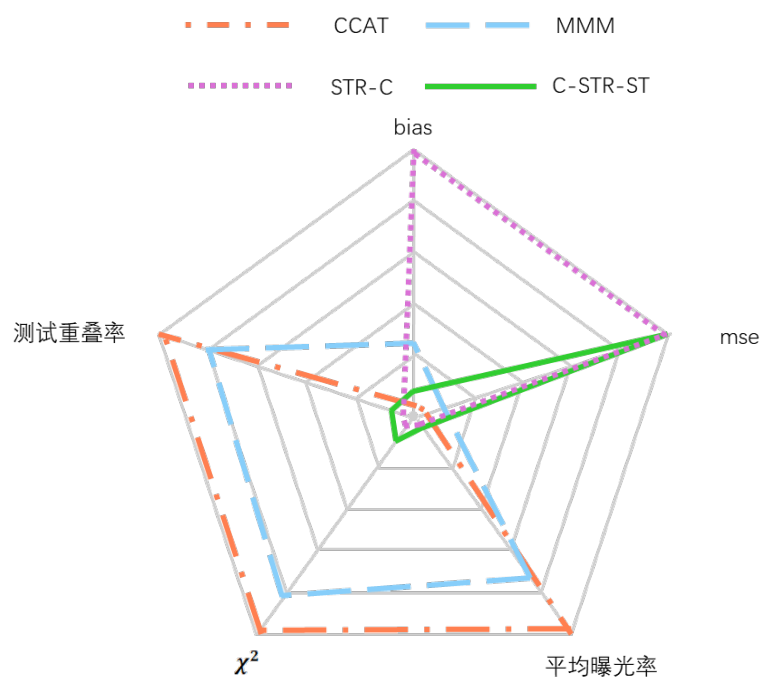


图 3-20 四种选题策略在题库 2 模拟实验中可量化指标雷达图

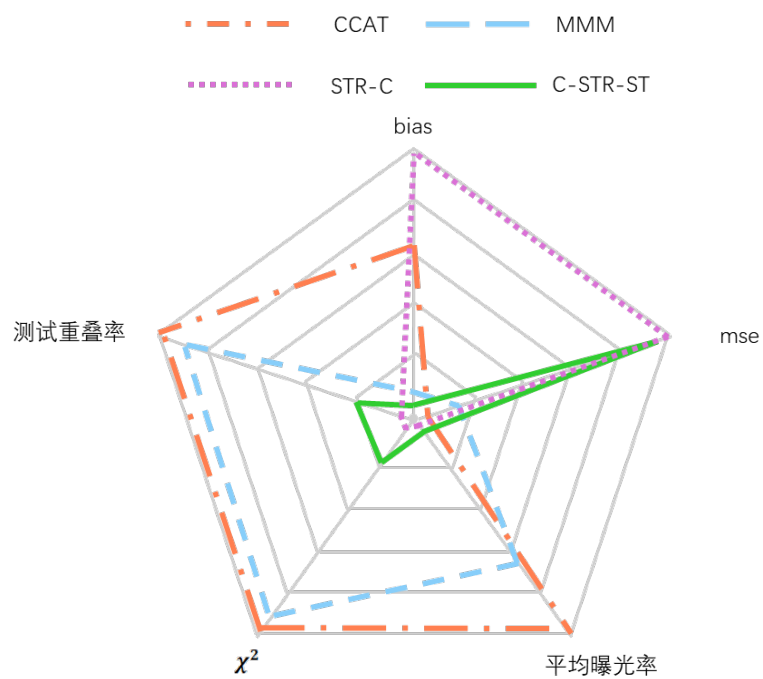


图 3-21 四种选题策略在题库 3 模拟实验中可量化指标雷达图

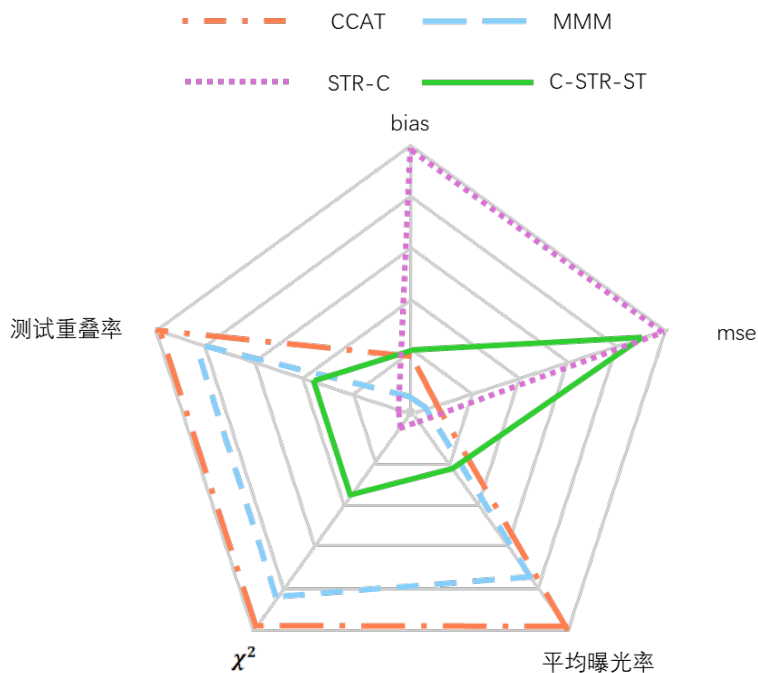


图 3-22 四种选题策略在题库 4 模拟实验中可量化指标雷达图

由于所有量化指标均是反向指标，即数值越小表示结果越好，从四个题库测试结果雷达图可以看出，c-STR-ST 所代表的线条围成的图形的面积均是最小的，因此综合五项量化指标，在 4 种题库结构的情况下综合测试结果最好。

结合雷达图所示与其他非量化指标（测试内容域控制、算法计算量等），c-STR-ST 与其他三种选题策略相比具有综合优势，在保证测试准确性、安全性的基础上，能够满足自定义的测试内容域比例。

3.6 总结

在本章中，基于现有内容平衡选题策略存在的不足之处，在现有的最大信息法、a 分层法以及影子题库选题策略的核心思想的基础上，提出了 c-STR-ST 内容平衡选题策略，并通过 Monte Carlo 实验的方式，模拟了四种不同参数分布的题库，将 c-STR-ST 与现有的内容平衡选题策略 CCAT、MMM、STR-C 进行对比分析，分析结果如下：

测试准确性方面，四种题库结构下 c-STR-ST 选题策略在 bias 指标上表现较其他选题策略更好，在 mse 指标上优于 STR-C 选题策略。

试题曝光率方面，四种选题策略平均试题曝光率差别不大，但是 c-STR-ST

选题策略在试题曝光评价指标 χ^2 检验统计量上优于 CCAT 和 MMM 选题策略。

测试重叠率方面, c-STR-ST 选题策略明显优于 CCAT 和 MMM。

内容平衡效果方面, c-STR-ST 选题策略能够实现用户自定义内容域期望比例, 且 Monte Carlo 模拟实验结果表明, 内容域实际比例能够满足用户期望。

算法计算量方面, 由于 c-STR-ST 选题策略采用了影子题库算法的思想, 从总题库中选出部分满足当前被试的试题组成影子题库, 并从影子题库中采用最大信息法选择最优试题, 与从总题库中采用最大信息法选择试题相比, 算法计算量大大减少。

综合上述实验结果与雷达图可以分析得出, c-STR-ST 选题策略在上述几个方面的综合表现优于其他三种选题策略, 能够实现用户自定义测试内容域之间的预期比例, 保证了测试准确性、安全性, 且有效地控制了测试重叠率。

第4章 c-STR-ST 选题策略在高中数学测试中的应用

4.1 基于 c-STR-ST 选题策略的自适应测试系统设计

4.1.1 基于 c-STR-ST 选题策略的自适应测试系统需求分析

1. 用户需求分析

基于 c-STR-ST 选题策略的自适应测试系统用户为学生和教师。

对于学生而言，学生能够应用该系统，采用自适应出题的方式获得试题并作答，在完成测试后得到个人能力水平估计的结果。

对于教师而言，教师能够查阅所有学生的测试结果，包括每个学生在测试中的系统选题情况、学生作答情况以及能力估计结果。

2. 功能需求分析

学生在进入测试前系统需要能够标识考生的用户信息；在进入测试后能够采用自适应测试的机制为考生出题、获取考生的答案并判断正误；在测试结束后能够给出测试的结果。

对于教师而言，系统需要能够导出考生的用户信息和测试结果，以便教师查阅分析。

用户需求分析如图 4-1 所示：

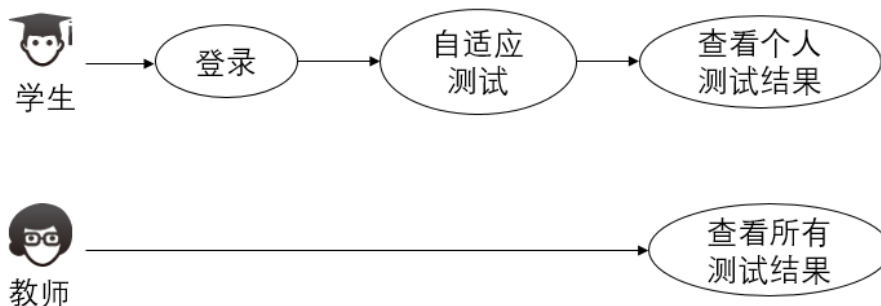


图 4-1 自适应测试用户需求分析

3. 非功能需求分析

结合对现有的自适应测试系统的分析，基于 c-STR-ST 选题策略的自适应测试系统设计主要应遵循以下几点原则：

（1）实用性。该系统能够实现自适应测试的功能，根据考生的作答情况和

选题策略给出试题以供学生作答，系统选题用时短，测试过程流畅。

(2) 易用性。该系统界面应简洁明了，试题呈现完整，作答方式简便，考生能够迅速掌握系统的使用方法，避免因系统操作问题影响测试。

(3) 开放性。该系统应能够面向多种类型的测试，改变试题信息数据库即可进行其他内容的测试。

4.1.2 基于 c-STR-ST 选题策略的自适应测试系统设计

1. 基于 c-STR-ST 选题策略的自适应测试系统主要功能

结合需求分析与设计原则，基于 c-STR-ST 选题策略的自适应测试系统的主要功能有：

(1) 用户管理

- a. 用户登录
- b. 用户身份记录

(2) 自适应测试

- a. 自动出题
- b. 即时判断考生作答正误
- c. 判断测试结束

(3) 测试结果管理

- a. 测试结束后考生查看本人测试结果
- b. 教师查看所有考生测试结果

2. 基于 c-STR-ST 选题策略的自适应测试系统功能结构

基于上述分析，本自适应测试系统的功能结构如图 4-2 所示：

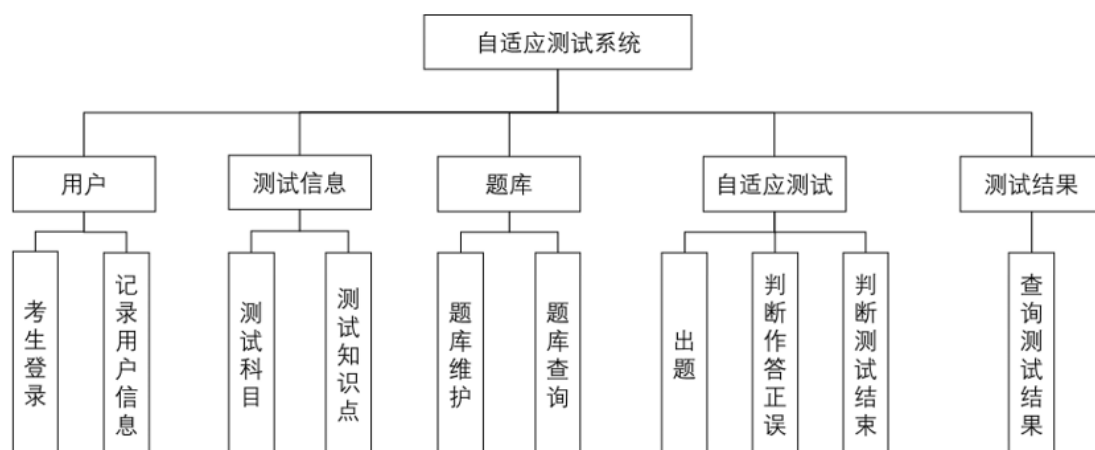


图 4-2 基于 c-STR-ST 选题策略的自适应测试系统功能结构

在该自适应测试系统中，测试用户即为考生，在进入测试时需要登录，系统会记录下考生的个人信息。

测试信息为本次考试的基本信息，包括科目和知识点等。

自适应测试的题库主要用于存储试题信息，教师能够查询和维护试题信息。

自适应测试功能主要是出题、判定考生作答的正误、判断测试结束条件。

测试结果主要供考生和教师查询。

3. 数据库设计

数据库是自适应测试系统的重要组成部分，用于存储用户信息、测试信息、试题信息以及测试结果等。在本自适应测试系统中选用关系型数据结构，用具有两个维度的表格存储数据，表格中的列代表信息字段，行代表信息，通过特定的行和列可以定位到某一具体数据。

（1）确定自适应测试系统需要的数据表

根据本系统中需要存储的信息，数据表应包括：

- a. 试题信息表
- b. 考生信息表
- c. 测试用题表
- d. 答题情况表
- e. 测试结果表

（2）确定数据表中的字段

各数据表中的信息字段如表 4-1、表 4-2、表 4-3、表 4-4 和表 4-5 所示：

表 4-1 试题信息表

字段	类型	主键
试题编号	char	是
区分度	int	否
难度	int	否
猜测系数	int	否
所属内容域	int	否
正确答案	char	否

表 4-2 考生信息表

字段	类型	主键
学号	char	是
姓名	char	否

表 4-3 测试用题表

字段	类型	主键
学号	char	是
试题编号	char	否

表 4-4 答题情况表

字段	类型	主键
学号	char	是
答题情况	char	否

表 4-5 测试结果表

字段	类型	主键
----	----	----

学号	char	是
能力估计值	int	否

各数据表之间的关系可以用图 4-3 所示的 E-R 图表示。

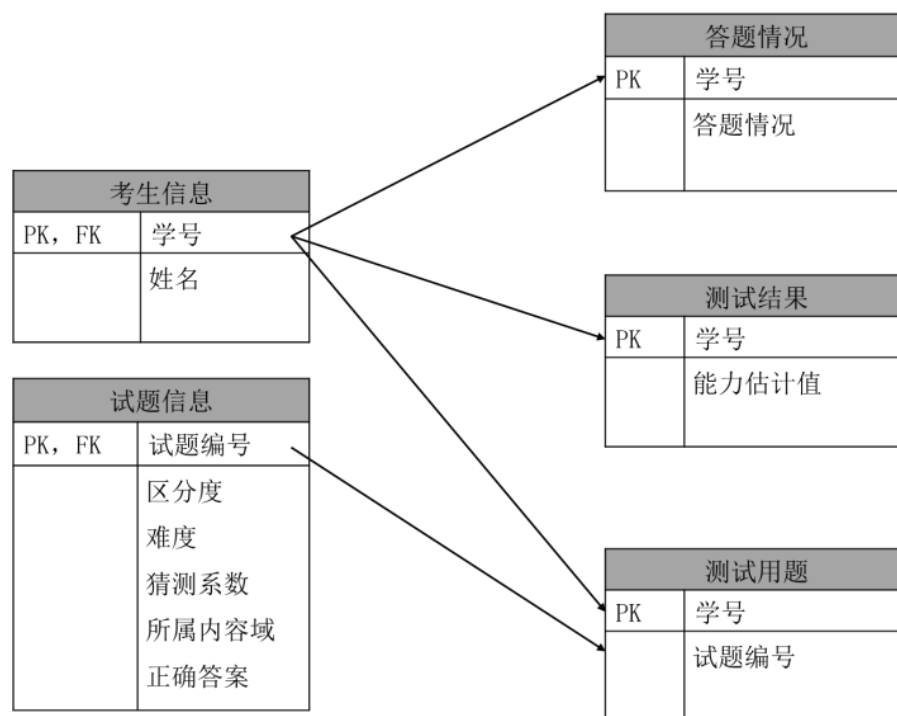


图 4-3 数据库 E-R 图

4.2 基于 c-STR-ST 选题策略的自适应测试系统开发

本自适应测试系统开发环境为 Windows 操作系统下的 Java (NetBeans IDE 7.4) 和 R (Rstudio), 数据存储与操作环境为 Microsoft Excel。

客户端采用 Java 开发, 学生通过客户端登录并进行测试, 答题情况数据将传至 R 中进行数据处理, 包括能力估计和根据选题策略选择下一道试题, 并将数据处理结果传回 Java 客户端, 呈现出试题题面并等待下一次作答。所有的答题情况和能力估计结果将写入 Excel 中, 并保存在由学生学号命名的文件夹中。整个系统的构成如图 4-4 所示:

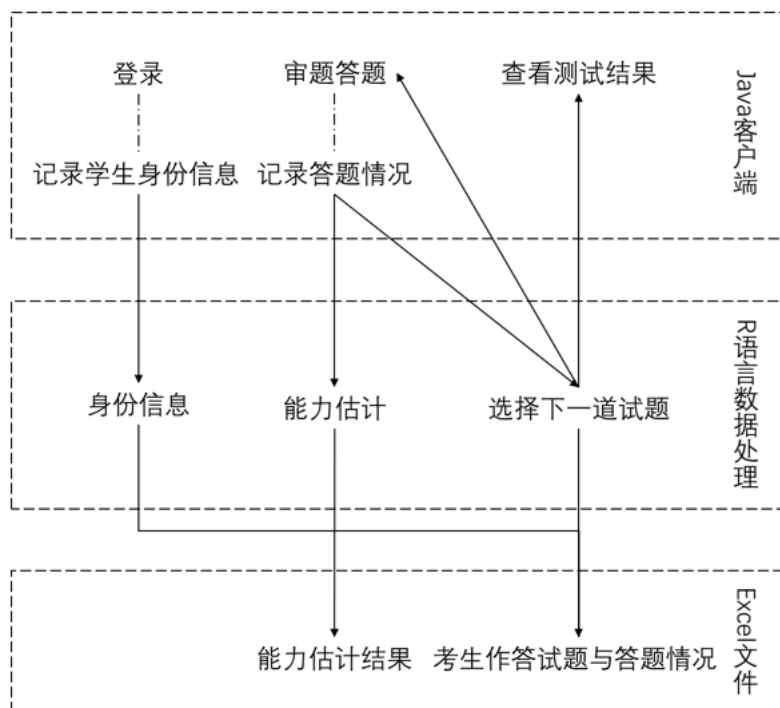


图 4-4 系统结构示意图

根据上述需求分析和系统分析，将系统分为三个功能模块，分别是用户登录模块、自适应测试模块和查看测试结果模块，如图 4-5 所示：

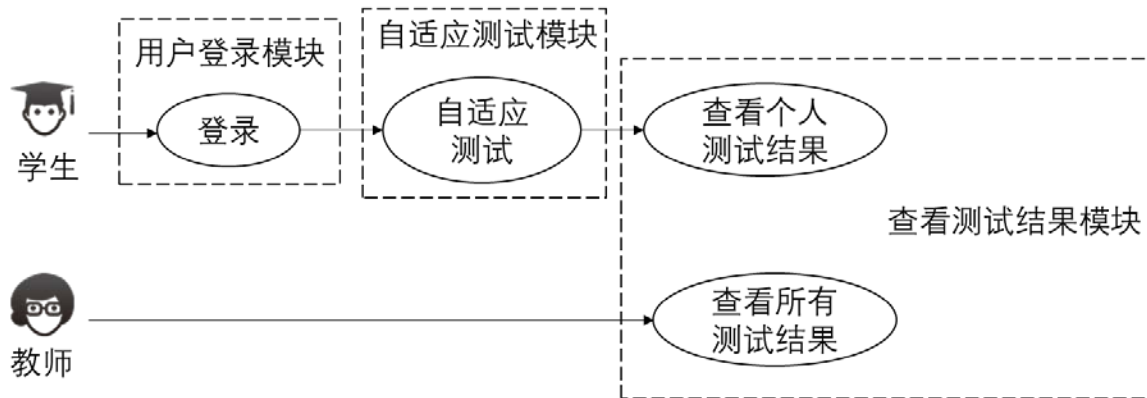


图 4-5 系统功能模块

本系统三个功能模块的具体设计与实现方法如下：

4.2.1 用户登录模块

在本系统中登录的用户为学生，登录时需要输入的信息包括姓名以及学号，登录界面如图 4-6 所示，点击“开始测试”即实现用户登录，记录下当前学生的关键用户信息，即学号，并开始测试。



图 4-6 登录界面

学号为标识学生身份的关键字段，不可为空，若未输入则在点击“开始测试”按钮后弹出提示并返回登录界面重新登录，提示界面如图 4-7 所示：

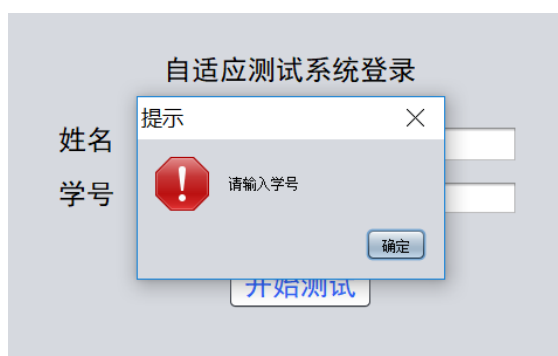


图 4-7 登录错误提示框

4.2.2 自适应测试模块

自适应测试模块主要功能是采用自适应出题的模式为学生提供测试，并即时判断学生作答情况，据此给出下一道试题。自适应测试模块的界面如图 4-8 所示：

科目：高一数学

考生姓名：张三

考生学号：10110123

6

集合 M 和 N 分别含有 10 个和 12 个元素，若集合 $M \cap N$ 有 5 个元素，则 $M \cup N$ 含有

() 个元素。

A. 10 B. 17 C. 22 D. 15

请输入你的答案: 提交

图 4-8 自适应测试界面

自适应测试采用 3.2 节中提出的 c-STR-ST 选题策略,终止规则为定长测试。题库中的试题在测试前按照 a 分层法进行分层,每道试题的信息包括题号、区分度 a、难度系数 b、猜测系数 c、内容域以及正确答案。题库的相关信息将保存在 Excel 文件中,存储结构如表 4-6 所示,存放在指定的路径下。该表格中 V1 列表示试题编号, V2 列表示区分度, V3 列表示难度系数, V4 列表示猜测系数, V5 列表示试题所属内容域, V6 列表示试题的正确答案。

表 4-6 题库数据表结构

V1	V2	V3	V4	V5	V6
1	0.476	-2.842	0.203	3	A
2	0.487	-2.982	0.206	3	C
3	0.497	-1.963	0.209	2	C
4	0.499	-2.388	0.204	1	A
5	0.5	-2.901	0.204	3	D
.....					

根据能力水平估计的不同,可以将计算机自适应测试的过程分为两个阶段。

第一个阶段为初始能力估计阶段，在每一层的题库中各选择 1 道试题进行测试。第二个阶段为自适应测试阶段，根据能力初始估计的结果和学生的作答情况，采用 c-STR-ST 选题策略选择试题，直到测试试题数量达到测试要求。

测试开始后该 Java 客户端将调用 R 进行选题和判断作答正误。具体调用方法为：

```
RConnection connection = null;

try {
    connection = new RConnection();
    connection.eval("source('D:/itemselect_new.R')"); //R 文件位置
    s=connection.eval("itemselect_new("+itemn+","+id+","+getans+","+userid
+"")").asInteger();    //调用 R 文件中 itemselect_new 函数并获得返回值
} catch (RserveException e) {
    e.printStackTrace();
} catch (REXPMismatchException ex) {
    Logger.getLogger(test.class.getName()).log(Level.SEVERE, null, ex);
}

connection.close();
```

在 R 的 `itemselect_new` 方法中，将通过 Excel 文件读取题库中的试题信息，并采用 3.2 中提出的 c-STR-ST 选题策略选择试题，并将所选试题的题号通过整型变量传回 Java 客户端，同时将测试信息保存于学号命名的文件夹下。Java 客户端接收到题号后，将选择对应的试题题面呈现以供学生作答。

考虑到实际测试中，试题的内容可能存在非文字的形式，如图片、公式等，因此在本系统中统一采用图片的形式存储试题的题面，试题图片的格式采用 png 格式，图片以试题编号命名，在调用 R 得到选题编号的返回值后，根据图片名称找到相应的题面，呈现给考生。实现方法如下：

```
String items=s+""; //将返回值即试题编号转换为字符型变量  
itemId.setText(items); //显示试题编号  
  
t="/newpackage/1-"+s+".png"; //定义试题题面图片路径及名称  
  
itemPic.setIcon(new javax.swing.ImageIcon(getClass().getResource(t)));  
  
//显示题面
```

4.2.3 查看测试结果模块

自适应测试结束后系统将通过 R 返回值得到考生能力水平估计结果并在系统中显示，学生能够了解自己的最终能力估计结果，如图 4-9 所示：

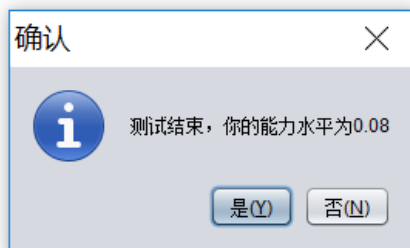


图 4-9 查看测试结果界面

教师可以查看更详细的测试结果，在测试结束后系统将生成一个由学生学号命名的文件夹，其中文件包括能力估计值、测试用题、用题的参数、学生小题答题情况。

4.3 基于 c-STR-ST 选题策略的自适应测试系统在高中数学测试中的应用

4.3.1 测试目的

为验证 c-STR-ST 选题策略及基于此开发的自适应测试系统在中小学实际教学评价中的作用，将基于 c-STR-ST 选题策略的自适应测试系统用于高中数学的测试中。

通过实际教学评价中的应用一方面验证该选题策略及系统的实用性，另一方

面能够为参与测试的学生和教师提供一种教学评价的新手段。由于自适应测试较传统测试所具有的优势，测试结果能够反映出更多测试信息，为总结上一阶段的学习和教学活动提供依据。

4.3.2 测试方法

测试采用本研究中设计和开发的基于 c-STR-ST 选题策略的自适应测试系统。

测试与上海市 H 高中合作完成，测试的科目选定为数学，年级选定为高一。

测试的试题由数学教师提供并组成自适应测试题库。

测试对象为 H 高中高一年级两个班级的学生，班级的选择机制为随机选择。

4.3.3 高中数学自适应测试题库建设

1. 试题来源

为建设自适应测试题库，首先要搜集一定数量的试题。本实验与上海市 H 高中合作完成，试题由学校提供。在获取试题时考虑的因素如下：

(1) 试题的测试内容域

在实际的测试中，测试的内容域越多，则题库的总量需求越大，所需要耗费的成本越高。因此在与学校相关老师沟通后，综合考虑测试的需要和现有条件，将测试的内容确定为高一上学期的 3 个内容域，包括集合、命题以及不等式（具体对于内容域的描述见表 4-7），并由学校提供了 2015 年至 2017 年高一上学期的阶段性测试试题共 88 道，涵盖了上述 3 个内容域，以及所有试题对应年份高一全年级学生的作答情况。

表 4-7 题库中试题涉及的内容域及其描述

编号	知识点	描述
内容域 1	集合	学生能理解集合的含义和性质，并能正确完成集合的基本运算。
内容域 2	命题	学生能理解命题的概念和命题的形式，并能判断命题的真假及命题之间的关系。
内容域 3	不等式	学生能理解不等式的基本性质，并熟练运用到计算中求解不等式。

（2）试题的形式

由于自适应测试的理论模型项目反应理论是基于二计分制的，即试题的作答评判只有正确和错误两种情况，因此试题的答案是唯一的，均为客观题。因此从上述三年试题中选择符合要求的 60 道选择题组成题库。

2. 试题参数估计

（1）试题区分度 a 、难度系数 b 、猜测系数 c

由于选题策略是基于三逻辑斯蒂模型的，因此题库中的试题也应包含三个参数，即试题区分度 a 、难度系数 b 、猜测系数 c 。在新题库建设时，由于参与测试的学生能力、试题参数都是未知的，基于 2.4 节中的理论分析，需要获得若干被试对试题的答题情况，并采用边际极大似然估计的方法进行参数估计。

边际极大似然估计中涉及到大量的计算过程，通常题库参数估计采用学术界认可的参数估计软件 BILOG⁶¹完成，该软件能够基于项目反应理论对数据进行处理，能够选择常用的逻辑斯蒂模型，对项目参数和能力进行估计。本文中将采用 BILOG 3.0 版本对题库的三参数进行估计。

采用 BILOG 3.0 估计试题参数首先需要整理历年学生答题情况数据。由于试题正确答案是唯一且确定的，因此是符合二计分制的。将获取的学生答题情况进行处理，用 1 表示回答正确、0 表示回答错误。以 2015 年的 16 道试题为例，2015 年高一年级 306 名学生在当年的测试中作答了这 16 道试题。对于每一名学生，将其 16 个作答情况用 16 个 0 或 1 的数字表示，得到长度为 16 的数字串，如图 4-10 所示，其中第一个数字表示学生的编号，编号后长度为 16 的数字表示该学生的作答情况。将该数据保存于 ans2015.dat 的文件中。

⁶¹ 张华华. 计算机化考试与中国教育评估[J]. 心理学探新, 2013, 33(5):387-391.

```

1 1111111111011111
2 1111111010011110
3 1111111101011111
4 1111111111011111
5 1011111010001110
6 1011101111001110
7 1111001110101110
8 1111101111011110
9 1100101110011100
10 1111111000011110

```

.....

```

298 1111101001001110
299 1111111110011101
300 1110111101000110
301 0001101010001100
302 1111101111011111
303 0111101100001110
304 1110001000011111
305 1101101001001010
306 1101110010011111

```

图 4-10 学生答题情况统计

在 BILOG 3.0 软件中新建项目，编写代码设定参数估计属性，如图 4-11 所示：

```

>GLOBAL DFName = 'C:\Users\mjiao\Desktop\ans2015.dat',
      NPArm = 3;
>LENGTH NItems = (16);
>INPUT NTotal = 16,
      NIDchar = 3;
>ITEMS ;
>TEST1 TName = 'TEST0001',
      INumber = (1(1)16);
(3A1,1X,16A1)
>CALIB ACCel = 1.0000;
>SCORE MThod = 1;

```

图 4-11 BILOG 3.0 参数估计指令

其中，DFName 表示答题情况数据 ans2015.dat 的存放路径；NPArm=3 表示采用三逻辑斯蒂模型；NItems 以及 NTotal 表示试题的数量，为 16；NIDchar 表示用于标识被试信息的字符在 ans2015.dat 所占的字符数，在这里学生编号占 3 个字符；TName 表示本次测试名称；INumber 表示答题情况数据格式；ACCel 表示算法加速；MThod=1 表示采用边际极大似然估计。

运行后得到参数估计结果如图 4-12 所示：

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
ITEM0001	0.816	0.408	-2.000	0.378	0.212
	0.142*	0.104*	0.589*	0.096*	0.093*
ITEM0002	0.572	0.529	-1.083	0.467	0.219
	0.153*	0.129*	0.391*	0.114*	0.094*
ITEM0003	-0.139	0.651	0.213	0.545	0.188
	0.194*	0.171*	0.279*	0.143*	0.079*
ITEM0004	0.704	0.407	-1.728	0.377	0.209
	0.142*	0.100*	0.535*	0.092*	0.092*
ITEM0005	0.477	0.547	-0.872	0.480	0.200
	0.150*	0.130*	0.348*	0.114*	0.088*
.....					
ITEM0012	-0.159	0.468	0.339	0.424	0.251
	0.219*	0.130*	0.437*	0.117*	0.095*
ITEM0013	1.194	0.465	-2.568	0.422	0.205
	0.155*	0.124*	0.683*	0.112*	0.091*
ITEM0014	0.341	0.406	-0.839	0.376	0.231
	0.159*	0.107*	0.471*	0.099*	0.097*
ITEM0015	-0.057	0.656	0.086	0.549	0.200
	0.192*	0.173*	0.284*	0.145*	0.083*
ITEM0016	-1.016	0.753	1.350	0.602	0.233
	0.410*	0.266*	0.349*	0.213*	0.065*

图 4-12 BILOG 3.0 参数估计结果

其中,SLOPE 列表示试题的区分度 a , THRESHOLD 列表示难度系数 b , ASYMPOTE 列表示猜测系数 c 。每个试题对应两行参数, 第一行的参数为列名对应的参数估计值, 第二行带有*符号的参数表示该估计的标准误。

参照此方法对 2015 至 2017 所有客观题进行三参数估计。由于项目反应理论中被试的能力水平在 $[-3, 3]$ 之间, 因此试题的难度系数也应落在该区间, 将 60 道选择题中不满足要求的 3 道试题删除, 余下 57 道试题组成题库。题库中试题的参数分布如图 4-13、图 4-14、图 4-15 所示:

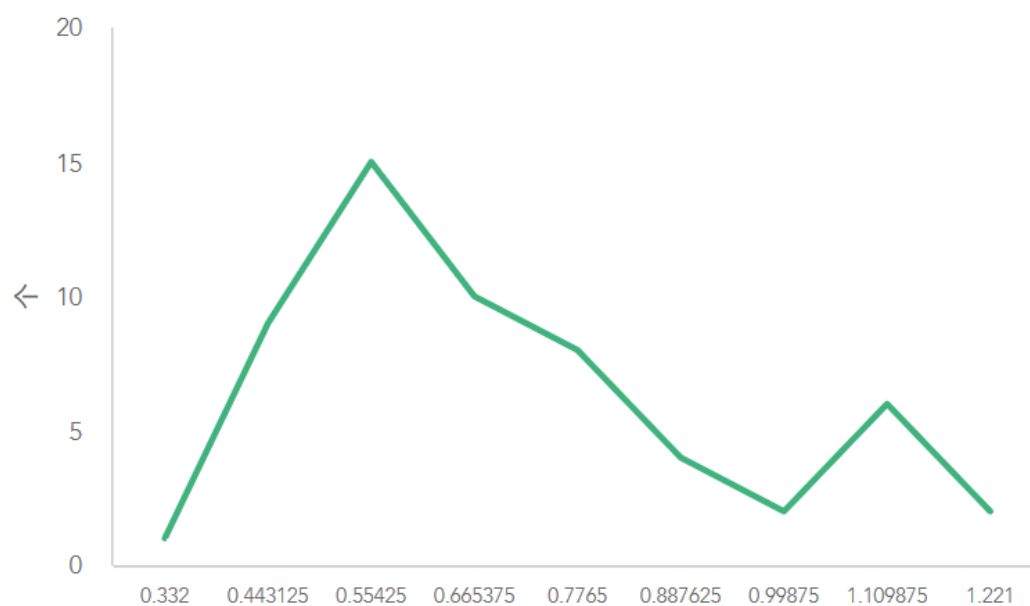


图 4-13 高一数学自适应测试题库中试题区分度 a 分布图

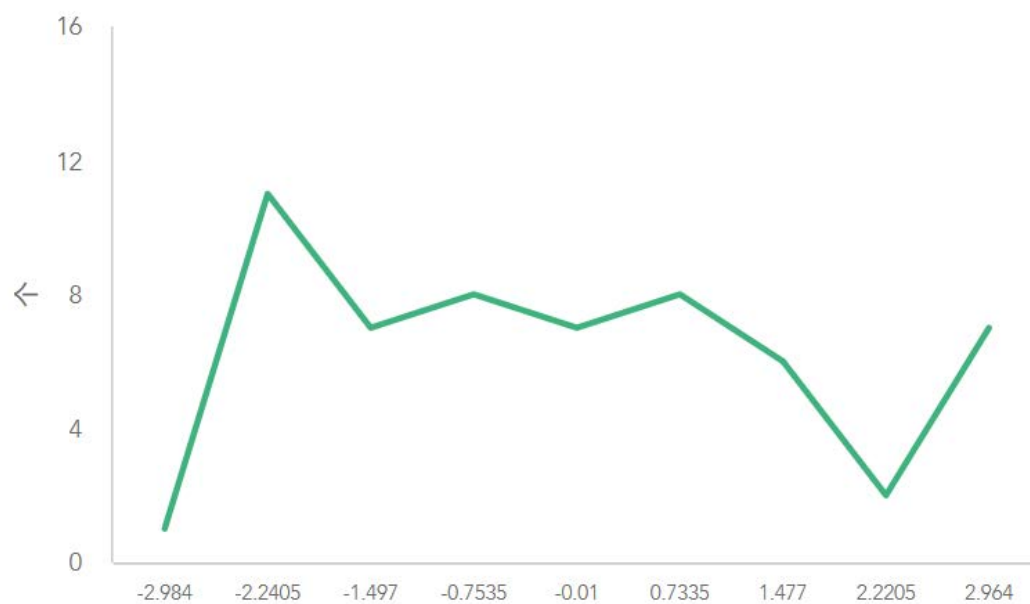


图 4-14 高一数学自适应测试题库中试题难度系数 b 分布图

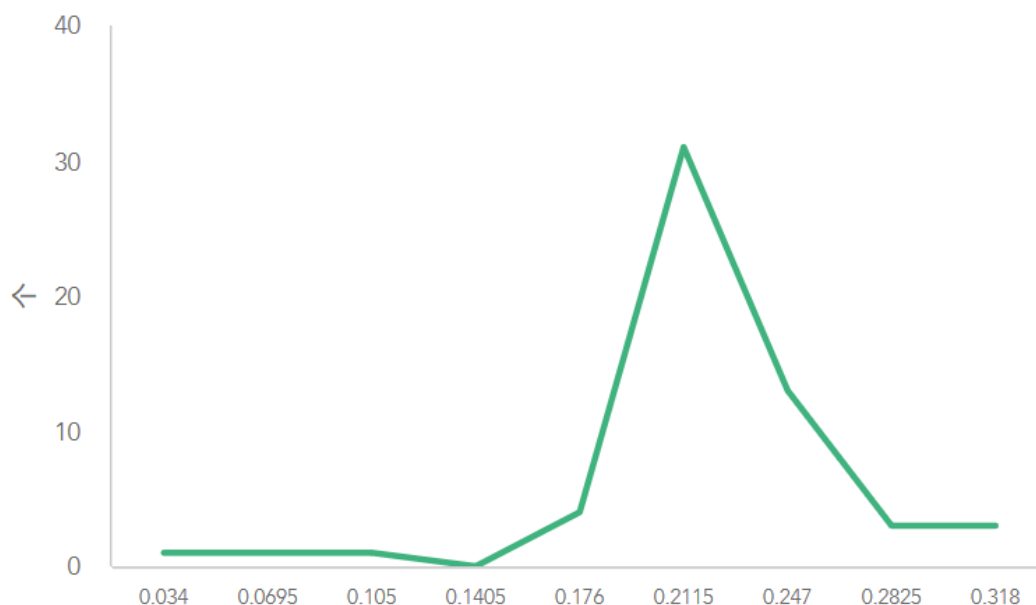


图 4-15 高一数学自适应测试题库中试题猜测系数 c 分布图

(2) 试题的内容域与正确答案

由于选题策略和对学生认知诊断的需要,由高中数学教师将每一道试题归纳到测试涉及的 3 个内容域之中。试题的正确答案也由数学教师提供。

综上,获得自适应测试系统题库中的试题以及试题的区分度 a 、难度系数 b 、猜测系数 c 、内容域以及正确答案。

4.4 高中数学自适应测试结果分析

结合上述基于 c -STR-ST 选题策略的自适应测试系统以及高一上学期数学题库,将其应用于高一数学实际教学活动中,对学生的阶段性学习成果做出评价,同时验证选题策略及其自适应测试系统在实际教学评价应用中的价值。

4.4.1 测试基本信息

1. 内容域及其期望比例

根据教师提供的题库,本次测试的内容域包括表 4-7 中的三个内容域,并由教师提供三个内容域的期望比例为 3:1:2。

2. 测试长度

测试为定长测试,即每个学生作答固定数量的试题后结束测试。根据教师的建议,测试长度设定为 20 道试题。

3. 被试样本

由于测试试题均包含在高一上学期数学的三个内容域中，因此被试的选择为上海市 H 高中随机选择两个班级。班级 A 人数为 42 人，班级 B 人数为 43 人，共 85 名被试参加本次测试，被试的具体情况如表 4-8 所示：

表 4-8 被试情况统计

班级	班级人数	男同学人数	女同学人数
A 班	42	18	24
B 班	43	20	23
合计	85	38	47

4.4.2 测试结果分析

1. 学生能力水平估计

根据每个学生在测试中的作答情况，系统能够对被试的能力水平做出估计。

测试得到 85 个学生的能力水平估计散点图如图 4-16 所示：

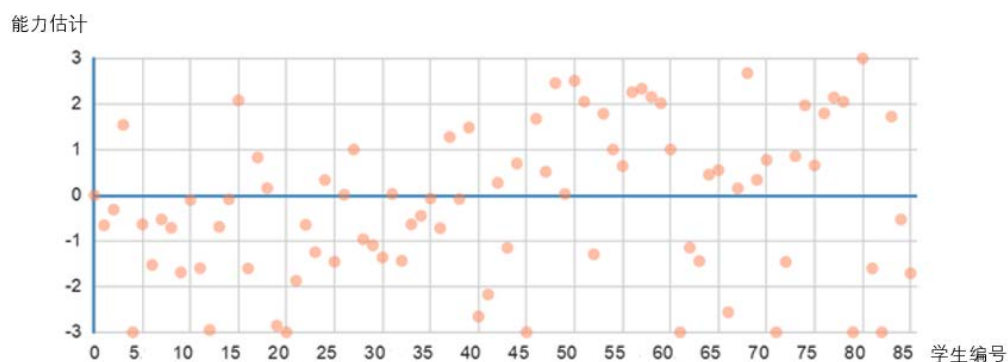


图 4-16 学生的能力水平估计

85 名学生的测试能力水平估计均值为-0.18047，最大值为 2.999947，最小值为-2.99995，极差为 5.999894，标准差为 1.6519。能力估计的分布情况如图 4-17 所示：

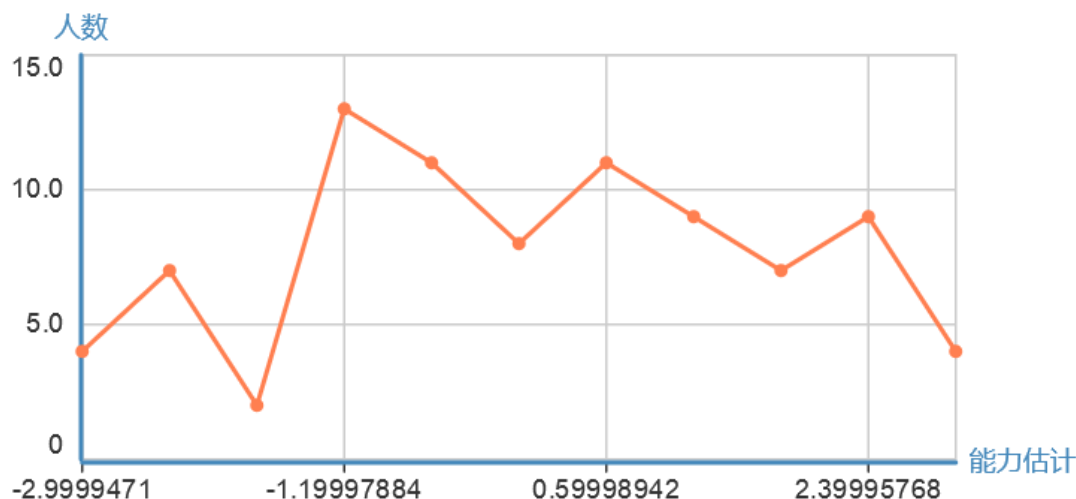


图 4-17 学生的能力水平估计分布

2. 知识点掌握情况

由于每道试题都标定了所测试的内容域,因此可以由学生的答题情况分析学生对于内容域的掌握情况。自适应测试中学生的总体能力估计是基于项目反应理论计算得出的,而对于一次测试中一个学生对某一内容域的试题选择和答题情况不能看作是完整的自适应测试,因此不能采用项目反应理论的方法估计学生对于某一内容域的能力水平。为了将学生对内容域的掌握情况进行分析和区分,对于某一内容域,统计每个学生所作答的属于该内容域的试题数量以及答对数量,计算学生答对试题的比例,即 $\frac{\text{答对属于内容域的试题的数量}}{\text{作答的属于内容域的试题数量}}$ 。将该比例划分为4个层级,分别为:

- (1) 属于 $[0, 0.25)$ 区间为掌握水平差;
- (2) 属于 $[0.25, 0.5)$ 区间为掌握水平较差;
- (3) 属于 $[0.5, 0.75)$ 区间为掌握水平较好;
- (4) 属于 $[0.75, 1]$ 区间为掌握水平好。

根据上述方法分析85名学生对于3个内容域的掌握情况分析如下。

对于内容域1,共有1名学生掌握情况差,11名学生掌握情况较差,44名学生掌握情况较好,29名学生掌握情况好。具体学生掌握情况和学生编号如表4-9所示:

表 4-9 内容域 1 学生掌握情况统计表

掌握情况	差	较差	较好	好
学生编号	44	48 32 9 22 43 5 23 2 26 49 7	4 6 8 46 74 3 29 13 24 28 47 25 40 51 71 12 21 52 76 84 16 19 45 1 10 15 20 35 36 39 55 58 67 73 17 18 38 61 62 33 41 69 78 37	11 34 59 66 42 57 63 68 79 81 14 27 70 30 56 80 83 50 85 60 75 31 53 54 64 65 72 77 82

对于内容域 2，共有 4 名学生掌握情况差，10 名学生掌握情况较差，26 名学生掌握情况较好，45 名学生掌握情况好。具体学生对于该内容域的掌握情况和学生编号如表 4-10 所示：

表 4-10 内容域 2 学生掌握情况统计表

掌握情况	差	较差	较好	好
学生编号	3 1 10 75	43 49 11 56 57 9 6 16 15 45	13 12 35 17 18 68 14 50 5 46 72 47 79 65 51 48 2 52 19 58 62 37 60 77 44 28	32 7 20 67 38 61 59 23 74 40 78 34 27 70 54 64 82 22 4 85 29 24 39 55 42 26 8 25 71 21 76 84 36 73 33 41 69 66 63 81 30 80 83 31 53

对于内容域 3，共有 6 名学生掌握情况差，11 名学生掌握情况较差，37 名学生掌握情况较好，31 名学生掌握情况好。具体学生对于该内容域的掌握情况和学生编号如表 4-11 所示：

表 4-11 内容域 3 学生掌握情况统计表

掌握情况	差	较差	较好	好
学生编号	11 29 1 50 14 46	12 18 52 2 48 8 30 3 7 66 53	10 43 57 15 17 68 47 19 62 44 38 27 55 25 21 54 6 31 75 16 51 32 65 36 45 13 5 37 61 74 63 80 56 60 20 39 82	35 59 78 4 24 42 26 41 9 77 70 71 79 23 64 22 33 83 58 34 69 72 81 73 40 84 76 49 28 67 85

对比 3 个内容域学生的掌握情况，内容域 1 学生的总体掌握情况较好，只有 1 名学生掌握情况为差，多数学生掌握情况较好。内容域 2 大多数学生能够在测试中表现掌握情况为好，而掌握情况差的学生人数较内容域 1 多。内容域 3 学生掌握水平在差和较差的人数最多，较内容域 1 和内容域 2 总体掌握情况欠佳。

结合上述分析，学生对于内容域 1 由于掌握情况总体较好，但优秀人数相对不多，因此在今后的教学活动中，教师可适当减少对于基础知识部分的讲解，而对该内容的重难点内容进行着重讲解。对于内容域 2，由于大多数学生能够在测试中表现出掌握情况好，因此教师可以对掌握情况欠佳的学生进行个别化指导，帮助他们赶上总体学习进度。由于内容域 3 学生的掌握情况较差，教师在教学活动中可以对内容域 3 给与更多的关注，加强这部分知识点的讲解和习题的练习，提高总体认知水平。

此外，测试的结果可以分析得出每个学生的认知曲线。以学生编号为 1、9、24 为例，三个学生的认知曲线如图 4-18、图 4-19 和图 4-20 所示：

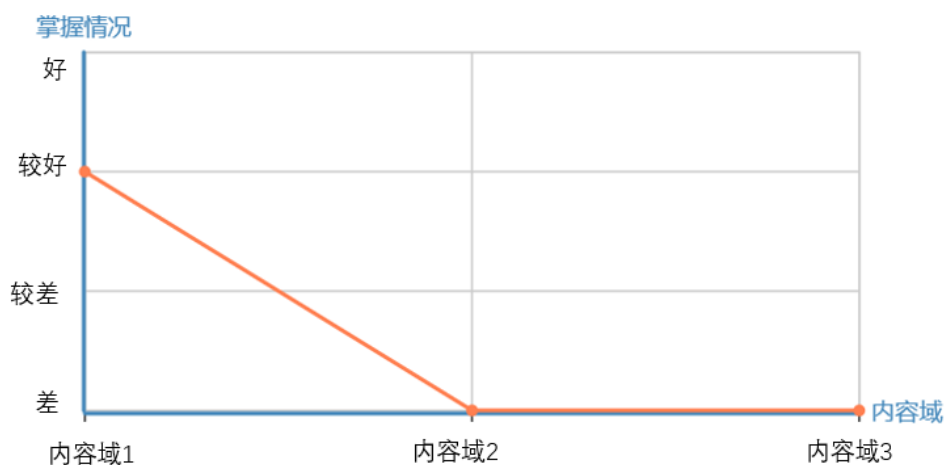


图 4-18 学生 1 认知曲线

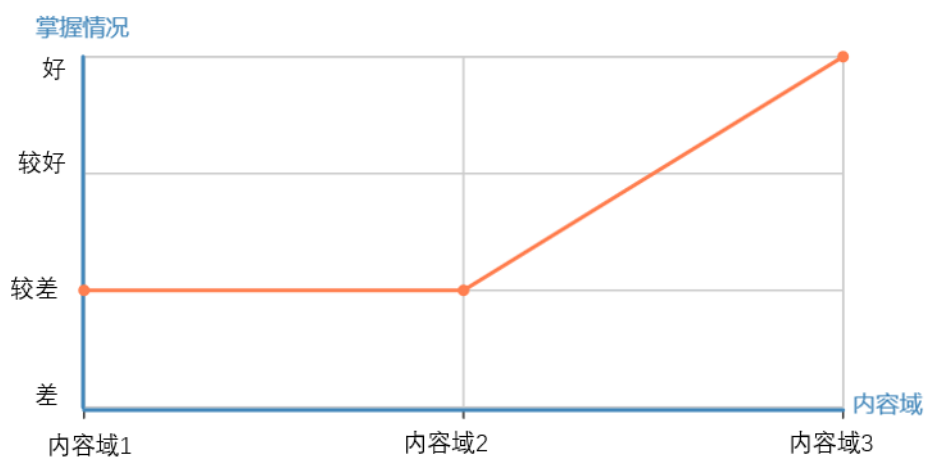


图 4-19 学生 9 认知曲线

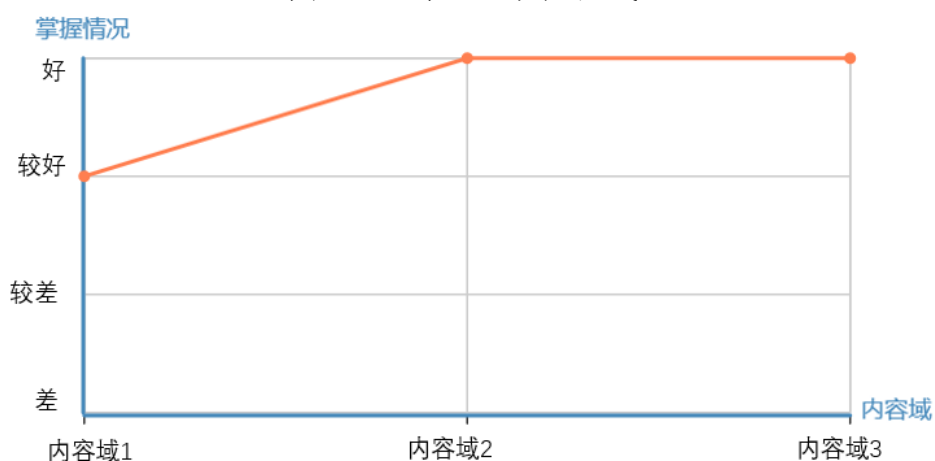


图 4-20 学生 24 认知曲线

学生 1 对于内容域 1 的掌握情况较好，而内容域 2、3 掌握情况差，因此在今后的学习中应加强对内容域 2、3 的学习和理解，在此基础上保持并提高对于内容域 1 的掌握水平。

学生 9 对于内容域 1、2 的掌握情况较差，应将更多的学习重点放在此内容域上，而内容域 3 掌握情况好，继续保持现有掌握情况。

学生 24 总体掌握情况良好，对于内容域 1 可以进行拔高练习，获得更深层次的知识理解。

根据学生的认知曲线，教师能够更好的掌握每个学生的认知情况，能够有针对性地对尚未达到学习要求的学生进行个性化的指导。而对于学生而言，能够对自己的阶段性学习成果有较为清晰的认识，了解对于所学知识点的情况和掌握程度，及时查缺补漏，能够更好地提高学习效率和学习效果。

第5章 总结与展望

5.1 研究总结

本论文以项目反应理论为理论基础,在现有的最大信息法、a 分层法选题策略以及影子题库算法思想的基础上,提出了基于内容平衡的选题策略 c-STR-ST,并采用 Monte Carlo 模拟实验的方法,规定了测试精准度、曝光度、重叠率等指标,模拟了 4 种不同结构的题库,将 c-STR-ST 选题策略与现有的内容平衡选题策略 CCAT、MMM 以及 STR-C 进行比较。根据 Monte Carlo 模拟实验的结果,c-STR-ST 选题策略在各项评价指标中综合表现优于其他三种选题策略,在保证测试准确性、控制测试曝光率和重叠率的基础上,能够实现用户自定义测试内容域期望比例,以达到测试选题的内容平衡。

基于 c-STR-ST 选题策略,本研究中根据实际需求设计和开发了计算机自适应测试系统。该系统能够实现考生的登录、自适应测试以及测试结果查看等功能。为验证基于 c-STR-ST 选题策略的自适应测试系统的实用性,本研究与高中数学实际教学评价相结合,建设基于高一数学内容域的题库,估计题库中试题的参数,并将题库与自适应测试系统结合,应用于实际的高中数学阶段性教学评价中,获得学生的测试结果。测试结果除了对于学生总体能力水平的评价外,还包括对每个内容域学生总体掌握情况的分析,以及对于学生认知情况的个性化分析。与传统的测试相比,通过基于 c-STR-ST 选题策略的自适应测试系统进行测试,能够为教师和学生提供更有针对性、更加个性化的教学评价,为之后的教学活动开展提供依据。

5.2 研究展望

在自适应测试算法改进方面,论文中提出的 c-STR-ST 选题策略虽然较 CCAT、MMM 以及 STR-C 综合表现更好,但是仍然存在题库分层后每个层中均有个别试题曝光率较大的问题。因此,在今后的研究中可以进一步控制该选题策略算法的试题曝光率,提高测试的安全性。

参考文献

- [1] Barrada J R, Mazuela P, Olea J. Maximum information stratification method for controlling item exposure in computerized adaptive testing[J]. *Psicothema*, 2006, 18(1):156-9.
- [2] Barrada J R, Olea J, Ponsoda V, et al. Item Selection Rules in Computerized Adaptive Testing: Accuracy and Security[J]. *Methodology European Journal of Research Methods for the Behavioral & Social Sciences*, 2009, 5(1):7-17.
- [3] Barrada R, Bernard P. |Olea. Multiple Maximum Exposure Rates in Computerized Adaptive Testing. [J]. *Applied Psychological Measurement*, 2009, 33(1):58-73.
- [4] Belov D I, Armstrong R D, Weissman A. A Monte Carlo Approach for Adaptive Testing with Content Constraints. [J]. *Applied Psychological Measurement*, 2008, 32(6):431-446.
- [5] Birnbaum A. Some latent trait models and their use in inferring an examinee's ability[J]. *Statistical Theories of Mental Test Scores*, 1968:395-479.
- [6] Chang H H, Qian J, Ying Z. α -stratified multistage computerized adaptive testing with b blocking. [J]. *Applied Psychological Measurement*, 1999, 23(4):211-222.
- [7] Chang H H, Ying Z. α -stratified multistage computerized adaptive testing[J]. *Applied Psychological Measurement*, 1999, 23(4):211 - 222.
- [8] Chang Hua-Hua, Ying, Zhiliang. A Global Information Approach to Computerized Adaptive Testing. [J]. *Applied Psychological Measurement*, 1996, 20(3):213-229.
- [9] Chang, H & Ying, Z. (1999). α -stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211 - 222
- [10] Chen S Y, Ankenmann R D. Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing[J]. *Journal of Educational Measurement*, 2004, 41(2):149-174.
- [11] Cheng Y, Chang H H, Yi Q. Two-Phase Item Selection Procedure for Flexible Content Balancing in CAT. [J]. *Applied Psychological Measurement*, 2007, 31(6):467-482.

- [12] Deng H, Ansley T, Chang H H. Stratified and Maximum Information Item Selection Procedures in Computer Adaptive Testing[J]. Journal of Educational Measurement, 2010, 47(2):202-226.
- [13] G. Gage Kingsbury, Anthony R. Zara. Procedures for Selecting Items for Computerized Adaptive Tests[J]. Applied Measurement in Education, 1989, 2(4):359-375.
- [14] Gulliksen H. Theory of mental tests [M]. New York: Wiley, 1950.
- [15] Leung C K, Chang H H, Hau K T. Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods[J]. Journal of Technology Learning & Assessment, 2003(5).
- [16] Leung C K, Chang H H, Hau K T. Content Balancing in Stratified Computerized Adaptive Testing Designs. [J]. Adaptive Testing, 2000, 2000(1):20.
- [17] Linden V D, Wim J. |Reese, Lynda M. A Model for Optimal Constrained Adaptive Testing. [J]. Applied Psychological Measurement, 1997, 22(3):259-270.
- [18] Linden W J V D, Veldkamp B P. Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests[J]. Journal of Educational & Behavioral Statistics, 2004, 29(3):273-291.
- [19] Lord F M, Novick M R, Birnbaum A. Statistical Theories of Mental Test Scores[M]. Statistical theories of mental test scores. UT Back-in-Print Service, 1968.
- [20] Lord F. A theory of test scores[M]. Psychometric Monograph, 1952, 7.
- [21] Lord, F. M. Applications of Item Response Theory To Practical Testing[M]. LAWRENCE ERLBAUM ASSCCIAATES, 1980.
- [22] Lord, Frederic M. A Broad-Range Tailored Test of Verbal Ability[J]. Applied Psychological Measurement, 1977, 1(1):95-100.
- [23] Lu P, Zhou D, Qin S, et al. The Study of Item Selection Method in CAT[J]. 2012:403-415.
- [24] Mao X Z, Xin T. Item Selection Method in Computerized Adaptive Testing[J]. Advances in Psychological Science, 2011.

- [25] Owen R J. A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing[J]. Journal of the American Statistical Association, 1975, 70(350):351-356.
- [26] Ozturk, Nagihan Boztunc|Dogan, Nuri. Investigating Item Exposure Control Methods in Computerized Adaptive Testing. [J]. Kuram Ve Uygulamada Egitim Bilimleri, 2015, 15(1):85-98.
- [27] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. [J]. Achievement Tests, 1961:199.
- [28] Revuelta J, Ponsoda V. A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing[J]. Journal of Educational Measurement, 2010, 35(4):311-327.
- [29] Samejima F. Estimation of latent ability using a response pattern of graded scores. [J]. Ets Research Report, 1969, 34(1):1-97.
- [30] Su Y H. A Comparison of Constrained Item Selection Methods in Multidimensional Computerized Adaptive Testing[J]. Applied Psychological Measurement, 2016, 40(5).
- [31] Swanson L, Stocking M L. A Model and Heuristic for Solving Very Large Item Selection Problems. [J]. Applied Psychological Measurement, 1993, 17(2):151-166.
- [32] Sympson J B, Hetter R D. Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association[C]. San Diego, CA: Navy Personnel Research and Development, 1985.
- [33] Tao Y H, Wu Y L, Chang H Y. A Practical Computer Adaptive Testing Model for Small-Scale Scenarios[J]. Journal of Educational Technology & Society, 2008, 11(3):259-274.
- [34] Wainer H. Computerized adaptive testing : a primer[M]. L. Erlbaum Associates, 2000.
- [35] Way W D. Protecting the Integrity of Computerized Testing Item Pools[J]. Educational Measurement Issues & Practice, 1998, 17(4):17 - 27.
- [36] Weiss D J. Improving Measurement Quality and Efficiency with Adaptive Testing[J]. Applied Psychological Measurement, 1982, 6(4):473-492.

- [37] Wingersky, M. S., & Lord, F. M. (1983). An investigation of methods for reducing sampling error in certain irt procedures *. Ets Research Report, 1983(2), i - 52.
- [38] Yi Q, Chang H H. a-Stratified CAT design with content blocking[J]. British Journal of Mathematical & Statistical Psychology, 2003, 56(2):359-78.
- [39] Zhao C H. The Design of IRT-Based Online Adaptive Testing System for College English Vocabulary[J]. Modern Educational Technology, 2008.
- [40] 冯艳宾, 马洪超. 关于经典测量理论和项目反应理论中难度和区分度的探讨[J]. 中国考试, 2012(4):10-14.
- [41] 付聪. 计算机自适应测试研究进展[J]. 现代情报, 2005, 25(1):61-64.
- [42] 简小珠, 戴海琦, 张敏强, 等. CAT 选题策略分类概述[J]. 心理学探新, 2014, 34(5):446-451.
- [43] 金瑜. 心理测量. 第2版[M]. 华东师范大学出版社, 2005.
- [44] 李慧. 浅析计算机自适应考试系统在大学英语测试中的应用前景[J]. 中国现代教育装备, 2009(3):27-29.
- [45] 李伟明, 陈富国. 经典测验理论和项目反应理论对题目分析的对比研究[J]. 心理学报, 1987(3):312-318.
- [46] 路鹏. 计算机自适应测试若干关键技术研究[D]. 东北师范大学, 2012.
- [47] 罗芬, 丁树良, 王晓庆. 多级评分计算机化自适应测验动态综合选题策略[J]. 心理学报, 2012, 44(3):400-412.
- [48] 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.
- [49] 漆书青. 现代教育与心理测量学原理[M]. 江西教育出版社, 1998.
- [50] 秦珊珊. 面向高中英语的自适应测试系统中项目参数的实验研究[D]. 东北师范大学, 2013.
- [51] 王勤云. 计算机自适应测验中选题策略的分析比较[D]. 山东师范大学, 2012.
- [52] 王晓庆, 罗芬, 丁树良, 等. 多级评分计算机化自适应测验动态调和平均选题策略[J]. 心理学探新, 2016, 36(3):270-275.
- [53] 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

- [54] 游晓锋, 丁树良, 刘红云. 计算机化自适应测验中原始题项目参数的估计[J]. 心理学报, 2010, 42(7):813-820.
- [55] 余嘉元, 汪存友. 项目反应理论参数估计研究中的蒙特卡罗方法[J]. 南京师大学报(社会科学版), 2007(1):87-91.
- [56] 余嘉元. 项目反应理论及其应用[M]. 江苏教育出版社, 1992.
- [57] 余嘉元. 项目反应理论研究中的计算机模拟方法[J]. 心理科学, 1991(2):49-51.
- [58] 俞晓琳. 项目反应理论与经典测验理论之比较[J]. 南京师大学报(社会科学版), 1998(4):79-82.
- [59] 张华华, 程莹. 计算机化自适应测验(CAT)的发展和前景展望[J]. 考试研究, 2005(2):14-26.
- [60] 张华华. 计算机化考试与中国教育评估[J]. 心理学探新, 2013, 33(5):387-391.
- [61] 张敏强, 刘晓瑜. 项目反应模型的应用问题研究[J]. 心理学报, 1998, 30(4):436-441.

附录 1 CCAT 选题策略算法实现代码

```
ccat<-function
(
  d,
  theta
){
  d=d
  theta=theta
  cr_t<-NULL
  content<-item_selected_par[,4]
  content=na.omit(content)
  for(i_1 in 1:5) {
    cr_t[i_1]<-length(content[content==i_1])
  }
  cr_t<-cr_t/40
  cr_t<-abs(cr_t-cr_e)
  if(sum(cr_t)==0) {
    f=f+1
  }
  cr_max<-which(cr_t==max(cr_t))
  if(length(cr_max)!=1) {
    cr_max<-sample(cr_max,size=1)
  }
  cr_max_selected<-Fl(cr_max,d,theta)
  return(cr_max_selected)
}
```


附录 2 MMM 选题策略算法实现代码

```
mmm<-function
(
  d,
  theta
)
{
  d=d
  theta=theta
  f=runif(1,0,1)
  content_selected=0
  if(f<=cr_f[1]) {
    content_selected=1
  }
  if(f>cr_f[1]&&f<=cr_f[2]) {
    content_selected=2
  }
  if(f>cr_f[2]&&f<=cr_f[3]) {
    content_selected=3
  }
  if(f>cr_f[3]&&f<=cr_f[4]) {
    content_selected=4
  }
  if(f>cr_f[4]&&f<=cr_f[5]) {
    content_selected=5
  }
  cr_max_selected<-FI(content_selected,d,theta)
  return(cr_max_selected)
}
```

附录 3 STR-C 选题策略算法实现代码

```
str_c <- function
(
  k,
  theta
)
{
  if(k==1) {
    b_dif<-abs(item1[,2]-theta)
    str_a_min=which(b_dif==min(b_dif[]))
    while(str_a_min %in% item_selected1[i,]) {
      b_dif[str_a_min]=9999
      str_a_min=which(b_dif==min(b_dif[]))
    }
  }
  if(k==2) {
    b_dif<-abs(item2[,2]-theta)
    str_a_min=which(b_dif==min(b_dif[]))
    while(str_a_min %in% item_selected2[i,]) {
      b_dif[str_a_min]=9999
      str_a_min=which(b_dif==min(b_dif[]))
    }
  }
  if(k==3) {
    b_dif<-abs(item3[,2]-theta)
    str_a_min=which(b_dif==min(b_dif[]))
    while(str_a_min %in% item_selected3[i,]) {
      b_dif[str_a_min]=9999
```

```
str_a_min=which(b_dif==min(b_dif[]))
    }
}
if(k==4) {
    b_dif<-abs(item4[,2]-theta)
    str_a_min=which(b_dif==min(b_dif[]))
    while(str_a_min %in% item_selected4[i,]) {
        b_dif[str_a_min]=9999
        str_a_min=which(b_dif==min(b_dif[]))
    }
}
return(str_a_min)
}
```

附录 4 c-STR-ST 选题策略算法实现代码

```
str_new <- function
(
  d,
  theta,
  item
)
{
  item_s<-matrix(nrow=1,ncol=4)
  s<-NULL
  for(i in 1:cr_e[1]) {
    s1<-sample(which(item[,4] %in% c(1)),size = 1)
    s<-c(s,s1)
  }
  for(i in 1:cr_e[2]) {
    s2<-sample(which(item[,4] %in% c(2)),size = 1)
    s<-c(s,s2)
  }
  for(i in 1:cr_e[3]) {
    s3<-sample(which(item[,4] %in% c(3)),size = 1)
    s<-c(s,s3)
  }
  for(i in 1:cr_e[4]) {
    s4<-sample(which(item[,4] %in% c(4)),size = 1)
    s<-c(s,s4)
  }
  for(i in 1:cr_e[5]) {
    s5<-sample(which(item[,4] %in% c(5)),size = 1)
```

```
s<-c(s,s5)
}
item_s<-cbind(item[s,],s)
FI<-NULL
for(i in 1:length(item_s)/5) {
  a<-item_s[i,1]
  b<-item_s[i,2]
  c<-item_s[i,3]
  FI[i] <- ((1-c)*d^2*a^2)/((c+exp(d*a*(theta-b)))*(1+exp(-d*a*(theta-b)))^2)
}
FI_max=item_s[which(FI==max(FI)),5]
return(FI_max[1])
}
```

致谢

当在论文的末尾敲下“致谢”两个字时，我才深切地体会到在华东师范大学7年的读书生涯步入尾声了。在这7年间，我从一个刚步入大学校园懵懂无知的女孩，成长为现在有担当的研究生，是学校给了我发展的平台和无限的可能，让我成就了现在的自己。

在完成硕士论文的期间，从开题到最终的定稿，是对我的一次综合性考验。在这一过程中，我碰到了许多困难，有时能够体会到解决问题后的成就感，有时也会陷入迷茫和无助。然而我的身边一直有支持我的老师、家人、同学和朋友，是他们对我的帮助，让我能够将所有困难一一克服，最终完成本篇论文。

首先要感谢的是我的导师孟玲玲老师，是孟老师带我认识了计算机自适应测试这一研究领域，引导、鼓励我在这个方向上不断进行研究。在论文开题时帮助我理清研究思路，在我遇到研究问题时用她丰富的研究经验引导我朝正确的方向思考。在孟老师的指导下不仅完成了论文，还学到了新的思考问题的思路。除了在学术方面的指导外，在生活上孟老师时常的关心让我感到，师门不仅仅是一个科研的团队，也是一个有爱的小团体。

其次是要感谢我的小伙伴们。在学校里，身边的同学是我最亲爱的人，是他们让我的研究生生活更加多姿多彩。在我的论文写作期间，当我想不到如何把数据可视化时，是室友集思广益和我一起讨论，让我打开了新的思路，实现了自己想要的效果；在我犯拖延症时，是小伙伴们拉着我一起去图书馆学习，让我的论文进度突飞猛进。还有本科的同学王怡，现在成为了一名“园丁”的小王老师在我的论文需要数据支持和实验时为我提供了帮助，让我的研究能够顺利进行下去。

研究生生涯的结束代表着另一端生活的开始，道阻且长，行则将至。

最后感谢所有给予过我帮助和支持的人，感谢评委专家给予的重要指导意见，感谢评审委员会的老师悉心指导，谢谢！

攻读硕士学位期间科研成果

1. 冷静, 刘梦娇. 未来学习技术及教育中的文化变迁——访著名学习技术专家斯蒂夫·哈蒙斯教授[J]. 开放教育研究, 2015(6):4-9.