2019届研究生硕士学位论文

分类号： 学校代码： 10269\_\_\_

密 级： 学 号： 51164102012



**East China Normal University**

**硕士学位论文**

**MASTER’S DISSERTATION**

**论文题目:语言模型在K-12英语题目质量分**

**析中的工程应用研究**

**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： 学习分析与数据挖掘**

**指 导 教 师： 顾小清教授**

**学位申请人 ： 吴秉聪**

**2019年 3月 10日**

Dissertation for master’s degree in 2019 Student ID: 51164102012

University code:10269

**East China Normal University**

**Title: The engineering application study of language model in K-12 English items quality analysis**

**Department:** Educational Information and Technology

**Major:** Educational Technology

**Research direction:** Learning Analytics and Data Mining

**Supervisor:** Professor Xiaoqing Gu

**Candidate:**  Wu Bingcong

**Mar ,2019**

**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《自适应测试思想在个性化练习中的应用研究》，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： 日期：** 年 月 日

**华东师范大学学位论文著作权使用声明**

《自适应测试思想在个性化练习中的应用研究》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 本人签名

年 月 日

\* “涉密”学位论文应是已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）。

**戴静 硕士学位论文答辩委员会成员名单**

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 职称 | 单位 | 备注 |
|  |  |  | 主席 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**摘 要**

《国家中长期教育改革和发展规划纲要（2010-2020）》中指出“要减轻中小学生课业负担，过重的课业负担严重损害青少年身心健康”，但为了追求成绩与升学率，“题海战术”在教学过程中仍然比较常见，中小学的作业情况存在量大、重复做和“一刀切”等实际问题，根据学生个体的水平做针对性的练习是解决以上问题的有效策略。可以说，为满足减轻学生负担的需求，在个性化学习的指导思想下，在教育信息化的发展浪潮下，个性化练习应运而生。

让学生做最少的题，最有效地帮助学生提高，这是个性化练习追求的目标。如何向学生推荐练习题才能达到这样的目标，是个性化练习的核心所在，也是个性化练习需要解决的关键问题。Lord在1970年提出，当某道题对学生来说既不是太难又不是太容易时，这道题对于该生来说就是好题，这就是自适应测试的核心思想，笔者认为这一思想在个性化练习中同样适用。

本文在分析自适应测试思想应用于个性化练习的可行性的基础上，从个性化练习的需求出发，对CAT和MST进行比较并确定将CAT的选题策略和MST的题组形式相结合作为个性化练习的选题机制，并增加对学习效果有强大影响力的反馈因素，以期帮助学生更好地消化个性化练习选择的题目。于是，形成了本文提出的应用自适应测试思想的个性化练习。

阐述应用自适应测试思想的个性化练习的组成要素、操作步骤及特点后，投入准实验研究，采用不相等实验组对照组前测后测设计，选取上海市某中学同一教师教的两个原始班级开展实验，以某中学数学学科初二年级“二次根式”章节为例进行效果检验，结果表明，应用自适应测试思想的个性化练习可以有效提高学生的能力水平和成绩。

**关键词**：语言模型；质量分析；工程应用研究；自动出题系统；决策辅助

**ABSTRACT**

“*The Medium and Long Term Reform and Development of Education Plan of China(2010-2020)*”pointed out that “To reduce the burden of primary and secondary school students, excessive schoolwork burden damage to the physical and mental health of young people.”. However, “questions tactics” is still relatively common in the teaching process for performance and enrollment rate. There have been some practical problems such as large quantity, repeating and "one size fits all" in the homework of middle and primary schools. And targeted exercises based on the level of individual students is an effective strategy to solve these problems. It can be said that, in order to meet the needs of reducing the burden of students, under the guidance of individualized learning, in the development of educational informationization, personalized exercise emerge consequentially.

Personalized exercise pursue the goal of doing the least questions and getting the most effective improvement. How to recommend exercises to students to achieve this goal is the core and key issues of personalized exercise. In 1970, Lord proposed that when a question is neither too difficult nor too easy for a student, this question is a good question for the student. This is the core idea of adaptive testing, and the author holds that it could also be applied in personalized exercise.

This paper analyzes the feasibility of applying adaptive testing to personalized exercise, compares CAT and MST according to the needs of personalized exercise, and determines the combination of item selection strategy of CAT and testlets form of MST as the item selection mechanism of personalized exercise. And increasing the feedback, which has strong influence on the learning effect, in order to help students better digest the items in personalized exercise. Consequently, the personalized exercise of applying adaptive testing has been proposed.

This paper expatiates on the constituent elements, operating steps and characteristics of the personalized exercise of applying adaptive testing, and then carrys out the quasi-experiment research. This research adopts non equivalent control group pretest posttest design, selects two classes of the second grade junior middle school students as the subjects,takes the chapter of quadratic radial as an example.The experimental results show that the personalized exercise of applying adaptive testing can effectively improve the students' ability level and performance.

**Keywords:** Personalized Exercise, Adaptive Testing, Application Study, Feedback

**目 录**

摘 要 I

目 录 IV

图目录 VII

表目录 VIII

第一章 绪论 1

1.1 研究背景 1

1.1.1 减轻作业负担的需求 1

1.1.2 个性化学习的趋势 1

1.1.3 教育信息化的发展 1

1.2 问题的提出 2

1.3 国内外研究现状 2

1.3.1 国内研究现状 2

1.3.2 国外研究现状 3

1.4 研究目的与意义 5

1.4.1 研究目标 5

1.4.2 研究意义 5

1.5 研究内容与问题 6

1.5.1 研究内容 6

1.5.2 研究问题 6

1.6 研究过程与方法 7

1.7 论文的创新与独特之处 7

1.8 论文的框架结构 7

第二章 理论基础 9

2.1 项目反应理论 9

2.1.1 项目反应理论概述 9

2.1.2 三参数Logistic模型（3PLM） 9

2.1.3 信息量与信息函数 10

2.2 反馈理论 11

2.2.1 反馈的基本介绍 11

2.2.2 反馈的影响力 12

2.2.3 反馈理论对实施个性化练习的启示 12

第三章 个性化练习 13

3.1 个性化练习的概念界定 13

3.2 个性化练习的特点 13

3.3 个性化练习的实施步骤与条件 14

3.4 个性化练习系统的推荐机制 14

第四章 基于计算机的自适应测试 16

4.1 计算机化自适应测试（CAT） 16

4.1.1 题库建设 16

4.1.2 选题策略 16

4.1.3 被试能力水平估计 18

4.1.4 终止规则 20

4.2 计算机化多阶段自适应测验（MST） 20

4.2.1 阶段（Stage）与模块（Module）数目 20

4.2.2 模块的长度 21

4.2.3 路径（Pathways） 21

4.2.4 路由规则（Routing Rules） 22

第五章 自适应测试思想应用于个性化练习 23

5.1 可行性分析 23

5.1.1 追求目标 23

5.1.2 实施对象 23

5.1.3 操作对象 23

5.1.4 推荐过程 24

5.2 应用于个性化练习的自适应测试选题机制的确定 24

5.2.1 CAT与MST的本质区别 24

5.2.2 CAT与MST的优缺点 25

5.2.3 应用于个性化练习的自适应测试选题机制 27

5.3 应用自适应测试思想的个性化练习 29

5.3.1 应用自适应测试思想的个性化练习组成要素 29

5.3.2 应用自适应测试思想的个性化练习的过程 30

5.3.3 应用自适应测试思想的个性化练习的特点 31

第六章 自适应测试思想应用于个性化练习的实践研究 34

6.1 实验目的与假设 34

6.1.1 实验目的 34

6.1.2 实验假设 34

6.2 实验设计 34

6.2.1 实验对象与学科 34

6.2.2 实验模式的设计 34

6.2.3 实验变量 35

6.2.4 实验安排 35

6.2.5 练习形式 36

6.2.6 辅助工具 37

6.3 题库的建设 37

6.3.1 题目的来源 37

6.3.2 获取学生的作答结果 37

6.3.3 模型拟合与参数估计 37

6.3.4 筛选题目入库 38

6.4 实验过程 38

6.5 实验结果与分析 40

6.5.1 学生能力水平值的整体变化情况 40

6.5.1.1 实验theta值的变化情况 40

6.5.1.2 实验与模拟theta值的对比变化情况 42

6.5.2 学生成绩的整体变化情况 45

6.5.2.1 实验班和对照班前测成绩差异比较 45

6.5.2.2 实验班、对照班各自前后测成绩差异比较 45

6.5.2.3 实验班和对照班后测与前测成绩差值差异比较 46

6.5.3 个性化练习对不同成绩段学生的影响的尝试性分析 46

6.5.3.1 个性化练习对各分数段学生theta值的影响 47

6.5.3.2 个性化练习对各分数段学生成绩的影响 48

第七章 总结与展望 50

7.1 研究总结 50

7.2 研究局限与不足 51

7.3 研究展望 52

参考文献 54

后记 58

**图目录**

图1 应用自适应测试思想的个性化练习 28

图2 实验安排 36

图3 题目难度参数的概率分布曲线 38

图4 实验过程 39

图5 两个班各theta值的概率密度函数图 40

图6 两个班各theta值的箱形图 41

图7 两个班theta均值的变化对比情况 42

图8 一班实验和模拟theta值的描述性统计指标对比 44

图9 四班实验和模拟theta值的描述性统计指标对比 44

图10 两个班各分数段学生的前后测分数分布情况 47

图11 两个班各分数段学生的theta均值变化情况 48

**表目录**

表1 题库中的参数分布 38

表2 两个班各theta值的描述性统计指标 41

表3 两个班前后测成绩的描述性统计指标 45

表4 实验班和对照班前测成绩独立样本t检验结果 45

表5 实验班和对照班前测、后测成绩配对样本t检验结果 46

表6 实验班和对照班的后测与前测成绩差值独立样本t检验结果 46

表7 两个班60-80分的学生后测与前测成绩差值独立样本t检验结果 48

表8 两个班80分及以上的学生后测与前测成绩差值独立样本t检验结果 49

**第一章 绪论**

## 1.1 研究背景

### 1.1.1 减轻作业负担的需求

《国家中长期教育改革和发展规划纲要（2010-2020）》中指出“要减轻中小学生课业负担，过重的课业负担严重损害青少年身心健康”[[1]](#footnote-1)。但是现在社会竞争非常激烈，为了追求成绩与升学率，“题海战术”在教学过程中仍然比较常见。作业量大，机械重复做题，直接导致了学生的作业负担。学生为了完成作业，可能出现不认真完成以致正确率不高、抄袭作业等低效作业的现象。

### 1.1.2 个性化学习的趋势

中小学作业除“题海战术”之外，还有“一刀切”的特点。由于中国学生人数多，小班教学比较少，教师没有时间和精力对学生一一辅导，普遍以班级的整体平均水平来布置作业。有些同学已经掌握了相关知识，却仍需一遍又一遍做自己已经会做的题。学习是一个循序渐进的过程，有些同学对概念还没有理解清楚，却要做对他来说比较难的题，不仅容易打击信心，对提高也无益。教师应该依据学生个体的水平，布置能够帮助学生提高的作业，这也符合了当今时代个性化学习的思想。

### 1.1.3 教育信息化的发展

信息技术推动教育的变革。手机、平板、电子白板等终端的研发，网络技术覆盖范围的扩大与传输速度的加快，各种学习软件与资源的开发等，都推动着教育信息化的发展。教育信息化从最初的计算机辅助学习，到数字化学习，再到现在的个性化学习，逐步提高了学习效率[[2]](#footnote-2)。

## 1.2 问题的提出

目前中小学的作业情况存在量大、重复做和“一刀切”等实际问题，根据学生个体的知识水平做针对性的练习是解决以上问题的有效策略，这一策略突出“个性化”，可称之为“个性化练习”。然而，在我国人口众多的基本国情下，让教师根据每个学生的知识水平，为其布置有效的作业，是不切实际的，但是计算机和网络为“个性化练习”的实现提供了可能。可以说，为满足减轻学生负担的需求，在个性化学习的指导思想下，在教育信息化的发展浪潮下，个性化练习应运而生。

如何让学生做尽可能少的题，最有效地提高学生的知识水平，这是个性化练习追求的目标。那么，如何向学生推荐练习题才能达到这样的目标，是个性化练习的核心所在，也是个性化练习需要解决的关键问题。而Lord认为，当某道题对学生来说既不是太难又不是太容易时，这道题对于该生来说就是好题，这是自适应测试的核心思想，且自适应测试的相关研究已较为成熟。因此，笔者对如何将自适应测试思想应用于个性化练习进行了探索。

## 1.3 国内外研究现状

### 1.3.1 国内研究现状

个性化练习作为个性化学习的重要组成部分，备受关注。二十世纪以来，越来越多的研究者和一线教师对个性化练习进行了探索和实践，国内关于个性化练习的研究呈逐渐上升趋势，且主要研究成果集中于近五年。以中国知网收录的文献数为例，截止2016年11月10日，以“个性化练习”或“个性化作业”为主题的文献共有467篇，其中2012年至今有366篇，占总数量的78%。

笔者认为，目前个性化练习的研究可分为两大类，绝大多数为第一类研究。此类研究是以一线教师为主力军，根据教学经验，结合学科特点提出个性化练习设计的原则，并针对具体课程设计出实践案例。这一类研究中的个性化练习以学生的个性差异、兴趣特长为主要关注点，提倡练习的内容、形式、评价方式的层次性、多样性和灵活性，给予学生自主发挥的空间。但从个性化练习的设计、实施到评价，需要耗费教师大量的时间和精力，而且对教师的专业能力也有较高的要求，使得此类研究进一步推广遭遇困境。

第二类研究较少，是关于个性化练习系统的研究，以学生的知识水平为主要关注点，如在线错题库系统。错题库系统可看作简单的个性化练习系统，该系统的主要功能是，记录学生在线答题时做错的题，为每个学生形成自己的错题库，学生可以重复练习错题库中的题。虽然错题库也是因人而异，但不能根据学生的知识水平出题，显然“个性化”的程度还不够。而且随着错题被练习次数的增加，学生记住答案的可能性也逐渐增大，错题库的价值将逐渐降低。故有学者在错题库的基础上进行完善，如蒋一君等[[3]](#footnote-3)提出当试题满足一定条件后则从错题库移除；郭辰[[4]](#footnote-4)与王文泉[[5]](#footnote-5)选择与错题类似的试题以供练习等。此外，吴皖赣[[6]](#footnote-6)从练习的评价环节入手，提出将知识地图应用于练习题的管理中，练习完成后根据作答情况生成个性化知识地图。

总的来说，国内目前关于个性化练习的研究还处于探索阶段，但成为研究热点的趋势已不可阻挡，尤其是信息技术与个性化练习相结合的研究，值得教育工作者进一步探索。

### 1.3.2 国外研究现状

国外关于个性化练习的研究不多，但关于测试的个性化研究成果丰硕，称之为计算机自适应测试，可为个性化练习的研究提供很多可借鉴之处。自适应测试的思想可追溯至1905年的比内智力测验，Alfred Binet提出不同年龄层次的被试做不同的测试题[[7]](#footnote-7)。1952年Lord提出双参数模型[[8]](#footnote-8)，项目反应理论正式创立，为计算机自适应测试提供了理论基础。再加上计算机的发展使得计算机自适应测试得以实现。1970年，Lord提出当测验题目对于被试既不是太难又不是太容易时测量效率最高[[9]](#footnote-9)，这是计算机自适应测试最早的选题策略思想。自此之后，越来越多的学者投入到计算机自适应测试的研究领域中,并取得了显著成果。

自适应测试的研究不仅面广且点深，从大的方面来说，有自适应测试系统的建构（例如[[10]](#footnote-10)Triantafillou, Georgiadou, & Economides,2008）、自适应测试与认知诊断的结合（例如[[11]](#footnote-11)Xu, Chang, & Douglas,2003）、不同测试形式的比较（例如[[12]](#footnote-12)Čisar, Čisar, & Pinter,2016）等；从小的方面来说，如选题策略（例如[[13]](#footnote-13)Chen, Ankenmann, & Chang,2000）、曝光率的控制(例如[[14]](#footnote-14)Stocking & Lewis,1998)、被试能力水平的估计（例如[[15]](#footnote-15)Samejima,1969）、作答时长（例如[[16]](#footnote-16)Hornke,2000）、测验情绪(例如[[17]](#footnote-17)Lu, Hu, & Gao,2016)、题库的建设（例如[[18]](#footnote-18)Yi, Zhang, & Chang,2005）等；更为细致的，如等值方法(例如[[19]](#footnote-19)Yao & Boughton,2009)、项目功能差异(例如[[20]](#footnote-20)Mantel & Haenszel,1959)、模型拟合(例如[[21]](#footnote-21)Orlando & Thissen,2000)等。

自适应测试理论成果丰硕的同时，实际应用也非常广泛。国外许多大型测试均采用计算机自适应测试的形式，如美国研究生入学考试(GRE)、美国注册护士执照考试（NCLEX-RN）、美国军队职业倾向测验(ASVAB)、美国工商管理类研究生入学考试(GMAT)、美国全国教育进展评估(NAEP)等[[22]](#footnote-22)-[[23]](#footnote-23)[[24]](#footnote-24)。教育、计算机、心理、医学等各个方面的专家都在关注计算机自适应测验，可见其具有良好的实用价值和广阔的发展前景。

## 1.4 研究目的与意义

### 1.4.1 研究目标

在学生作业负担重、个性化学习趋势、教育信息化发展的研究背景下，个性化练习应运而生。本文尝试将自适应测试的思想应用到个性化练习中，主要研究目标包括以下两个方面：

（1）探索如何将自适应测试思想应用于个性化练习，并总结应用自适应测试思想的个性化练习的组成要素、特点及操作过程，为个性化练习系统的建设提供理论指导。

（2）通过准实验研究，以某中学数学学科初二年级“二次根式”一章的复习为例，检验本文提出的应用自适应测试思想的个性化练习是否能够提高学生的能力水平及成绩。

### 1.4.2 研究意义

中小学生的身心健康与作业负担逐渐受到社会及学校的重视，目前中小学的作业情况存在题海战术、重复做和“一刀切”等实际问题，而个性化练习是解决这些问题的有效途径之一，随着教育信息化的发展，使得个性化练习在教学中的应用成为可能，探索如何有效实施个性化练习具有十分重要的意义。本文的研究意义在于：

（1）理论研究意义：一是，丰富了个性化练习的应用研究，促进了个性化练习理论研究的发展；二是，提出了应用自适应测试思想的个性化练习，并用实验研究验证应用自适应测试思想的个性化练习是否有助于提高学生的能力水平和成绩，为其他研究者提供借鉴。

（2）实践指导意义：一是，对应用自适应测试思想的个性化练习进行了实践检验，整个实验过程为其他研究者提供参考；二是，对建设应用自适应测试思想的个性化练习系统提供理论指导。

## 1.5 研究内容与问题

### 1.5.1 研究内容

基于以上分析，确定本文的主要研究内容为：

首先，梳理目前的个性化练习和自适应测试的相关研究，明确个性化练习的目标和需求，以及为何将自适应测试的思想应用于个性化练习；

第二，在自适应测试思想应用于个性化练习的可行性分析的基础上，根据个性化练习的需求，选择合适的自适应测试选题机制，作为本文个性化练习的推荐机制；

第三，对应用自适应测试思想的个性化练习的组成要素、特点、操作过程进行阐述；

第四，采用准实验研究，实验组采用个性化练习、对照组采用普通练习，验证应用自适应测试思想的个性化练习对提高学生的能力水平和成绩的有效性。

### 1.5.2 研究问题

本文拟解决的关键问题如下：

（1）自适应测试思想是否可以应用于个性化练习；

（2）基于计算机的自适应测试形式有CAT和MST两种，应该选择何种自适应测试选题机制应用于个性化练习；

（3）如何将自适应测试的思想应用于个性化练习中；

（4）检验应用自适应测试思想的个性化练习是否能够提高学生的能力水平和成绩的实验如何开展，题库是开展此实验的前提条件，如何建设可以使用的题库。

## 1.6 研究过程与方法

本文的研究过程主要包括两个阶段，第一阶段提出将自适应测试思想应用于个性化练习，对于应用何种自适应测试选题机制、如何应用以及应用自适应测试思想后个性化练习的特点等方面进行阐述；第二阶段是将自适应测试思想应用于个性化练习的实践研究，主要实验目的是检验本文提出的应用自适应测试思想的个性化练习是否能够有效提高学生的能力水平及成绩。

主要采用的研究方法是文献研究和准实验研究法。

（1）文献研究

本文主要通过文献研究，围绕当前个性化练习的发展现状和存在的问题，提出将自适应测试思想应用到个性化练习中。通过文献学习项目反应理论、基于计算机的自适应测试的两种形式及各自的优缺点、反馈理论，分析如何将自适应测试思想应用到个性化练习中。

（2）准实验研究

本文的准实验研究采用不相等实验组对照组前测后测设计，选取上海市某中学同一教师教的两个原始班级开展实验，以某中学数学学科初二年级“二次根式”章节为例，以验证本文提出的应用自适应测试思想的个性化练习是否能够有效提高学生的能力水平和成绩。

## 1.7 论文的创新与独特之处

本文的创新与独特之处主要表现在以下三个方面：第一，提出将自适应测试的思想应用到个性化练习中并进行实践研究；第二，将CAT和MST这两种基于计算机的自适应测试选题机制相结合应用于个性化练习中；第三，使用反馈理论帮助学生更好地消化自适应选择的题目，提高个性化练习对学生帮助的效果。

## 1.8 论文的框架结构

本文大致可分为四个部分，第一部分为研究的绪论和理论基础，包括第一章和第二章；第二部分为提出应用自适应测试思想的个性化练习，包括第三章、第四章和第五章；第三部分是将自适应测试思想应用于个性化练习的实践研究，位于第六章，第四部分是全文的总结与展望，位于第七章。各章节的具体内容介绍如下：

第一章为绪论，主要介绍本文的研究背景、问题的提出、研究现状、目的与意义、内容与问题、过程与方法、研究创新点。

第二章为理论基础，介绍了项目反应理论和反馈理论，并针对各理论阐述其在本研究中的应用或对本研究的启示。

第三章是对个性化练习的基本介绍。包括概念界定、特点、实施步骤与条件以及当前研究中个性化练习系统的推荐机制存在的不足。

第四章是对基于计算机的自适应测试的基本介绍。基于计算机的自适应测试可分为CAT和MST两种，介绍了CAT的题库建设、选题策略、被试能力估计、终止规则四个方面，以及MST的阶段与模块数目、模块长度、路径、路由规则四个方面。

第五章是在第三章和第四章的基础上提出应用于自适应测试思想的个性化练习。首先，从追求目标、实施对象、操作过程、推荐过程等角度分析了自适应测试思想应用于个性化练习的可行性；接着从个性化练习的需求出发对CAT和MST进行比较并确定将CAT的选题策略和MST的题组形式相结合作为个性化练习的选题机制；然后讨论了应用自适应测试思想的个性化练习的组成要素、过程及特点。

第六章为自适应测试思想应用于个性化练习的实践研究。从实验目的与假设、实验设计、题库建设、实验过程以及实验结果与分析五个方面对整个实验过程进行介绍。实验之前，需要收集题目与学生作答结果，再进行模型拟合与参数估计，最后选择题目进入题库。建设好题库后才能进入正式实验，以检验应用自适应测试思想的个性化练习对提高学生的能力水平和成绩的有效性。

第七章为总结与展望。从个性化练习的目标和需求、为何将自适应测试思想应用于个性化练习、如何应用和检验效果的实证研究这四个方面进行总结。并说明本研究实验过程中的许多不足之处及后续研究方向。

**第二章 理论基础**

## 2.1 项目反应理论

### 2.1.1 项目反应理论概述

项目反应理论（Item Response Theory，简称IRT）是为了突破经典测量理论的局限性而产生的一种新的测量理论，最初人们称之为潜在特质理论，后来称之为项目特征曲线理论，现在更多的人称之为项目反应理论。

IRT较之经典测量理论最大的特点是被试能力参数和题目参数的不变性以及将两者置于同一刻度。对于一种理论，了解其成立的前提假设是进行应用的基础。IRT有三条基本假设：第一，局部独立性，即被试答对某题的概率与答对另一题的概率相互无影响；第二，递增性，即对于某一题项，随着学生能力值的提高，答对该题项的概率就越高。第三，能力单维性，即测验中的所有题项均测量同一潜在特质。

### 2.1.2 三参数Logistic模型（3PLM）

项目反应理论模型根据评分方式可分为0-1评分模型和多级评分模型，根据被试参数可分为单维IRT模型和多维IRT模型，本文主要讨论0-1评分单维IRT模型中应用较多的logistic模型。logistic模型根据项目参数的数量可分为单参数Logistic模型（1PLM,即Rasch模型）[[25]](#footnote-25)、双参数Logistic模型（2PLM）[[26]](#footnote-26)、三参数Logistic模型（3PLM）26和四参数Logistic 模型（4PLM）[[27]](#footnote-27)。本研究中使用的是三参数Logistic模型，公式如下：

其中，i表示第i个题项；D为常数，通常取1.7；表示第i题的区分度；表示第i题的难度；表示第i题的猜测系数。当时，就是2PLM；当且时，就是Rasch模型；在3PLM的基础上再增加失误参数，就是4PLM。

### 2.1.3 信息量与信息函数

统计学中信息量是指信息的确定性程度，项目反应理论借用信息量的概念来表示项目或测验估计被试能力水平时所提供的信息的确定性大小[[28]](#footnote-28)。同一道题或同一测验对于不同的被试所提供的的信息量是不同，针对这一现象，项目反应理论提出了项目与测验信息函数的概念。项目信息函数是指项目在评价被试特质水平时贡献的信息量大小28，公式如下：

其中，是项目i的被试反应函数，是对的一阶导函数。由于项目反应理论的局部独立性假设，测验中各个项目的项目信息函数也是相互独立的。将本研究使用的三参数Logistic模型的公式带入，得3PLM项目信息函数如下：

对于三参数Logistic模型的项目信息函数，影响其大小的关键因素是该题项的区分度和猜测系数。区分度越大、猜测系数越小的题项，其项目信息函数的值越大。项目信息函数是关于被试能力水平的函数，其图像随被试能力水平的变化而呈钟形状态。项目信息函数值达到最大时，对应的被试能力水平为28：

从上式可以看出，使得项目信息函数达到最大值的值略大于题目的难度值。区分度越大、猜测系数越小时，越接近题目的难度值。

项目信息函数还具有可累加性，测验信息函数就是测验所含题项的项目信息函数的累加28，即：

容易发现，增加测验题目的数量可以提高测验信息量，尤其是增加区分度高、猜测系数小的题目，同时使测验题目的难度分布集中于被试的能力水平附近。

## 2.2 反馈理论

### 2.2.1 反馈的基本介绍

“反馈”是控制论的创始人维纳(Wiener)于1948年首次提出[[29]](#footnote-29)，后来被广泛应用于其他领域。在教育领域中，反馈可以帮助学习者修正自己的思想和行为以促进学习，可以帮助教师根据学习者的水平来调整教学计划以适应学习者[[30]](#footnote-30)，可以说，反馈对于提高教与学的效果有着非常重要的作用。

Hattie and Timperley认为反馈的目的就是缩短当前水平与目标水平之间的距离[[31]](#footnote-31)。以练习和测验中给予学习者的反馈为例，在答题时，学生可能答对也可能答错，教师、家长或计算机依据不同的表现给予学习者一定的反馈，学习者根据接收到的反馈调节自己的学习过程，如增加努力程度、对具体知识进行针对性地学习等，从而获得更好的学习结果，逐渐向目标水平靠近。

练习和测验中的反馈，从内容上来说，包括告知学生其作答正确与否、告知学生每一题的正确答案、让学生一直尝试直至答案正确、给予学生解题线索等；从时间上来说，包括作答一个题目后立即给予反馈、作答整个测验后立即给予反馈，又或者是作答一道题或整个测验延迟一段时间后再给予反馈；从效果上来说，分为正反馈和负反馈。Kulhavy and Stock(1989)[[32]](#footnote-32)认为有效的反馈不仅仅提供答案是否正确等确定性信息，还应对学习者有一定的指导作用。

### 2.2.2 反馈的影响力

反馈对学习效果有着强大的影响力，这一点已被许多研究者证实16-[[33]](#footnote-33)。不同的反馈内容、反馈时间、反馈方式等会对学习效果产生不同程度的影响，但具体的影响程度并没有达成共识。如Pashler等[[34]](#footnote-34)关于卢干达语-英语词对学习的实验结果表明，告知学习者正确答案比仅仅告知其答案对错学习效果更好；而Marsh等[[35]](#footnote-35)关于作答常识多项选择题的实验结果表明，告知学习者正确答案与仅仅告知其答案对错在答案正确率的保持上无显著差异。Corbett等[[36]](#footnote-36)关于ACT程序教学的实验结果表明，当学生出现错误时立即给予反馈并要求其改正的学习效果比其他方式更好；Metcalfe等[[37]](#footnote-37)关于六年级的学生学习词汇的实验结果表明，延迟反馈比立即反馈的学习效果更好；Clariana等[[38]](#footnote-38)关于高中生基于计算机阅读八篇文章并答题的实验结果发现，立即反馈和延迟反馈的效果因题目难度不同而不同。

### 2.2.3 反馈理论对实施个性化练习的启示

个性化练习也是练习的一种，反馈的影响力同样适用。本文的做法是在每次练习作答完成后给予学习者反馈，即作答结果反馈，让学习者及时了解自己的当前水平，以期对提高学习者的能力水平有催化和深入的作用。

值得一提的是，笔者认为个性化练习本身也是反馈理论的体现。本文的个性化练习是作答完成一道题后根据学习者的能力水平来选择下一道题，可以理解为，下一道题是计算机根据学习者的当前能力水平估计值所给予的一种反馈。

**第三章 个性化练习**

## 3.1 个性化练习的概念界定

个性化练习可以理解为练习题的个性化推荐。在线学习系统中的个性化推荐是指向学习者推荐与其需求最相符的学习材料[[39]](#footnote-39)，这里的需求主要从当前知识水平、学习兴趣、学习目标、学习者背景、学习风格、认知风格等[[40]](#footnote-40)方面考虑。本文的个性化练习是指，针对已学知识，根据学习者的当前知识水平来向学生推荐练习题，其追求目标是，用最少的题来最有效地提高学生的知识水平。

## 3.2 个性化练习的特点

个性化练习与普通练习相比，具有如下几个特点：

（1）针对性

不同的学生应该根据自己的水平有针对性地做题。每个学生的接受能力和理解能力是不同的，当新课学习完后每个学生对于所学知识的掌握水平是不同的，对于全班学生布置同样的作业显然是不合理的。但实际生活中，教师普遍以班级的整体平均水平来布置作业。个性化练习可以依据学生个体的当前知识水平，为其推荐适合的题目。

（2）高效性

个性化练习可以用较少的题有效提高学生的知识水平。普通练习中，有些题太简单无需重复做，有些题太难根本不会，过于简单和过于难的题目对于提高学生水平都没有太大的作用，可归为无效练习，故实际上普通练习中的有效练习题是较少的，只有足够多量的题才能“遇见”较多的有效练习题，从而达到提高学生水平的作用，也就是所谓的题海战术，显然普通练习的效率比较低。个性化练习根据学生的知识水平出题，提高练习的有效性，循序渐进地帮助学生提高。

## 3.3 个性化练习的实施步骤与条件

个性化练习的核心步骤主要有两个：其一，了解学生的知识水平；其二，根据学生的知识水平出题。

若由教师人工出题，则对教师来说是一项巨大的挑战。首先，教师需要通过上课时学生的表现和一些基础练习来了解并判断每个学生当前的知识水平，这对于经验丰富的教师来说可能相对容易，对新教师较难。其次，为每一个学生出题，这项任务的工作量是非常大的，教师在找题、针对性地出题以及最后改作业的过程中，都需要花费大量的时间和精力。最后，教师还需要掌握认知心理学等知识，以便更好地把握学生的个性需求等。总的来说，由于班级人数众多，教师的时间与精力有限，故由教师长期进行个性化练习出题是不现实的。

另一个实施途径就是建立个性化练习系统，由计算机来出题，在当今信息技术飞速发展的时代背景下，这一途径是可以实现的，那么个性化练习的两大核心步骤就需要个性化练习系统中的算法来实现，我们把根据学生的知识水平来推荐习题的算法称之为推荐机制。

## 3.4 个性化练习系统的推荐机制

推荐机制是个性化练习系统的核心与关键所在，推荐机制的目标就是针对已学知识向学生推荐有效帮助学生提高的练习题。目前这方面的研究还比较少。实践中已使用的错题库，只是满足了不同的学生针对自己做错的题目重复练习，并没有做到真正的个性化出题，而且容易引起作答次数过多记住答案的问题。蒋一君等3将学生要做的练习分为三种，即做错的题、未掌握的知识点以及已掌握的知识点，并根据不同的理论采取不同的推送机制。针对做错的题，满足一定条件则将该题移出错题库；对于未掌握的知识点，优先选择错题库中的题，若不够则根据知识点权重值来选题；对于已经掌握的知识点，根据艾宾浩斯遗忘曲线规律，在适当的时间进行巩固练习。这种机制分类细致、考虑全面，但没有关注到各知识点下的题还有不同难度层次之分，而且仍然无法摆脱做错题次数过多记住答案的问题。郭辰4通过建立知识体系进行诊断性练习，并为学习者推荐与错题具有相同知识点和题型的题目以供练习。这种机制避免了重复做错题，但仍没有考虑到题目的难度层次。王文泉5主张为学习者推荐与错题具有同等难度且同属一个知识点的高质量试题。这种推荐机制避免了重复做错题，也考虑到了题目的难度层次，但只分为三个难度等级且仍依赖错题库，若某些知识点没有练习到，那么推荐过程中也不会出现。

可见，目前的个性化练习推荐机制过于依赖错题，且很少考虑到题目的难度，即使考虑到也是简单地划分为几个等级，显然不足以很好地匹配学生的水平。

**第四章 基于计算机的自适应测试**

基于计算机的自适应测试主要分为计算机化自适应测试（computerized adaptive testing，简称CAT）和计算机化多阶段自适应测验（multistage adaptive testing，简称MST）。

## 4.1 计算机化自适应测试（CAT）

计算机自适应测试的过程包括题库的建设、选题策略、被试能力水平估计、终止规则等。由于选题策略和被试能力水平估计最能体现自适应测试的思想，故对这两节的介绍较为详细。

### 4.1.1 题库建设

题库是进行计算机化自适应测试的基础。本文是采用三参数Logistic模型，故题库中的每个题目都具有区分度、难度、猜测系数三个参数，区分度是指题目区分不同被试的能力，理论上取值范围为[-∞,+∞]，一般取值范围为[0,+3]；难度是指题目的难易程度，理论上取值范围为[-∞,+∞]，一般取值范围为[-3,+3]；猜测系数是指能力水平趋于-∞时答对该题目的概率，理论上取值范围为[0,1]，一般取值范围为[0,0.5]28。实际操作过程中，三参数的取值范围比以上的一般取值范围更小，具体取值范围需依据情况而定。

### 4.1.2 选题策略

选题通常分为两个阶段，第一阶段是学生能力水平未知时进行初始题项的选择，比较常见的方法有以下几种：从整个题库（或某个特定题目集）中随机选取[[41]](#footnote-41)（Cheng,2008）、从中等难度题目中随机选取[[42]](#footnote-42)（Embretson & Reise,2013）、根据被试的先验信息选取42（Embretson & Reise,2013）等。

第二阶段是根据被试在第一阶段的作答结果，计算出能力水平值作为初始值后，依据更新的值进行题目的选择。通常所说的选题策略主要是指第二阶段的选题方法，这一阶段是整个计算机化自适应测试的核心。最早的选题算法是b匹配法[[43]](#footnote-43)（Lord,1971）、贝叶斯选题法[[44]](#footnote-44)（Owen,1975）、最大Fisher信息量法[[45]](#footnote-45)（Lord,1977），后来因考虑曝光率和内容平衡等因素，在上述算题算法的基础上产生了S-H条件概率法[[46]](#footnote-46)（Sympson & Hetter,1985）、KL全域信息量法[[47]](#footnote-47)（Chang & Ying,1996）、a分层法[[48]](#footnote-48)（Chang & Ying,1999）、a分层b分块法[[49]](#footnote-49)（Chang, Qian, & Ying，2001）、a分层内容分块法[[50]](#footnote-50)（Yi & Chang，2003）等多种选题策略。每种选题策略都有各自的优缺点，选择时要依据具体测试条件和要求而定。以最大Fisher信息量法和a分层法为基础衍生出的一系列选题策略是最为流行的两大类[[51]](#footnote-51)，故这里主要介绍这两种基本算法。

（1）最大Fisher信息量选题策略（MFI）

最大Fisher信息量选题策略是Lord在1977年提出的45，基本原理是根据被试当前的能力水平估计值，对题库中该被试没有做过的所有题目，进行Fisher信息量的计算（3PLM项目信息函数的计算公式已在2.1.3中给出）并排序，选取Fisher信息量最大的那一道题，作为被试将要作答的下一道题。

在Logistic模型下，使用最大Fisher信息量选题策略选出的题目难度系数值与能力值匹配且区分度高、猜测系数小。这种方法测量效率高且测量误差低[[52]](#footnote-52)，是之前应用最广泛的选题策略。但在实际应用过程中，发现高区分度的题使用较多，不仅使得题库使用不均匀造成题库的浪费，还会引起题库的安全性问题[[53]](#footnote-53)。

（2）a分层选题策略

a分层法是Chang和Ying(1999)48针对最大Fisher信息量法的不足而提出的，基本原理是将题库按照题目区分度系数值从小到大分成K个层次，测验也分为K个阶段，测验的第K个阶段对应从题库的第K个层次中选择选题，选择难度与当前能力值相匹配的题目作为下一道题。简单来说，a分层法的思想就是在测试的初始阶段使用区分度较低的题，随着测试的进行，逐渐使用区分度越来越高的题。但是，题目的区分度和难度系数是很有可能存在正相关的[[54]](#footnote-54)，这样a分层法就会使得每一层题库中题目难度分布不均匀。

### 4.1.3 被试能力水平估计

被试能力水平估计也是CAT的重要环节，这里对于被试能力水平的估计均在已知题目参数的基础上。方法有极大似然估计（MLE）26（Birnbaum,1968）、贝叶斯极大后验估计（MAP）[[55]](#footnote-55)(Samejima,1969)、贝叶斯期望后验估计（EAP）[[56]](#footnote-56)(Bock & Mislevy,1982)、加权似然估计（WLE）[[57]](#footnote-57)（Warm,1989）等。其中，最常用的是方法是极大似然估计和贝叶斯期望后验估计。

（1）极大似然估计（Maximum Likelihood Estimation，简称MLE）

使用极大似然法进行参数估计的第一步是写出似然函数，0-1评分三参数Logistic模型的似然函数为：

其中，u为该被试的作答反应向量（如001100），为该被试在第i题的作答结果（0或1），为该被试的能力水平值，为三参数Logistic模型的公式即该被试在第i题上答对的概率。

我们可以这样理解，能力水平为的被试进行测验的最可能的作答结果为u，将看作已知变量，即算出该被试对试题进行作答的反应向量是u的概率；但实际上，上述变量中的u和题目的三参是已知变量，是未知变量，此时就是一个关于的函数，我们需要的是使得最大的值。通常我们会先将似然函数取对数，然后求对数似然函数最大时对应的值，原因是log函数是单调的，最优的似然函数对数值就是最优的似然函数值；另外似然函数是很多概率值相乘，得到的结果较小，取对数之后得到的值相对较大，便于解释。最后实际上是求对数似然函数对的偏导数等于零这一方程的解，一般使用数值迭代法如Newton-Raphson方法求解。

极大似然估计是无偏估计，但具有一定的局限性，不能处理极端情况，即被试所有题目作答结果为全对或全错。因此在使用极大似然估计时，需人为预先设定最大的θ值和最小的θ值，当全对（或全错）时直接给定最大的θ值（或最小的θ值），而不是使用MLE方法估计出。实际上，我们也想获得这些特殊被试的θ值，贝叶斯期望后验估计可以解决这个问题。

（2）贝叶斯期望后验估计（Expected a Posteriori Estimation, EAP）

贝叶斯期望后验估计的理论基础是贝叶斯定理，基本原理是将的后验分布的均值作为估计值。基于给定的作答反应向量u，关于的后验分布公式如下：

其中，是能力水平参数的先验分布，是似然函数（见公式6），是的后验分布，我们需要的是后验分布的期望值即均值，公式如下：

先验分布的引入使得EAP估计有偏，而先验分布的选择是影响EAP估计误差大小的重要因素，一般采用标准正态分布作为的先验分布，但并不一定是最佳选择[[58]](#footnote-58)。除此之外，贝叶斯期望后验估计拥有估计值稳定、用时少等许多能力估计方法的优点[[59]](#footnote-59)。

总的来说，MLE是无偏估计但不能处理极端情况，EAP可以解决MLE的问题但是有偏估计。故也有学者将这两种方法进行整合41，当被试作答都正确或都错误时，使用EAP方法；当被试作答既有答对也有答错时，使用MLE方法。

### 4.1.4 终止规则

终止规则是指CAT测验结束的条件，主要分为两种28：一种是固定测验的题目量，即通常所说的定长测验，这种方法在设定题目量时，需要考虑将被试能力水平估计误差控制在合理范围之内。另一种是固定测量误差，即固定信息量标准，虽然每个被试答题数量可能会不同，但这种方法在累积信息量足够高的前提下，可以保证每个被试有相同的信度，较为公平，是体现CAT核心思想的终止规则。

在测验效率和能力水平估计的精确度方面，第二种方法均优于第一种方法。但第二种方法在实际应用过程中，可能出现个别被试能力水平估计值不收敛而导致测验过长，此时需要为不定长测验设置题目量的上限值[[60]](#footnote-60)，因此常将两种终止规则结合使用。

## 4.2 计算机化多阶段自适应测验（MST）

计算机化多阶段自适应测试是以题组为单位的自适应测试，本节主要对其基本元素进行简单介绍。

### 4.2.1 阶段（Stage）与模块（Module）数目

MST由若干个阶段组成，每个阶段由若干个不同难度层次的模块组成。阶段数目和模块数目通常合并表示，如“1-2”、“1-3-3”，数字的个数表示阶段数目，用“-”隔开，数字的大小表示不同难度层次的具体模块数目。

早期的MST是纸笔测试，为了便于操作，多采用两阶段MST，使用计算机测试形式后，阶段数目和模块数目的选择相对灵活。在大多数学者的研究中，通常采用三至四个阶段，第一个阶段只包含一个中等难度水平的模块，后续阶段中包含三至四个模块[[61]](#footnote-61)。

### 4.2.2 模块的长度

模块的长度，即这一模块中包含的题目数量。虽然题目数量越多，对被试的测量精度越高，但过多显然会给被试带来负担。同一MST中，各个阶段的模块长度可以相同也可以不同。在固定每个被试测试总题数的前提下，起始阶段和后续阶段的模块分别使用多少道题依据具体需要而定，比如想减少建立MST所需准备的题目数量，可以使第一阶段的模块长一些，后续阶段的模块短一些，毕竟后续阶段的难度层次和模块数更多[[62]](#footnote-62)。

总的来说，阶段数目、模块数目、模块长度的具体数值，可在综合考量测量精度、组卷的复杂性、建立MST所需准备的总题数、学生做题负担等方面后，依据实际情况而定。

### 4.2.3 路径（Pathways）

在MST中，每个阶段都有若干个模块，每个被试在每个阶段只会做其中的一个模块，路径可以理解为被试在作答MST的整个过程中，相邻阶段之间模块的跳转。对于一个阶段数目、模块数目已确定的MST，所有可能的路径也是确定的。以1-3-3 MST为例，如图1(b)所示，所有可能的路径已在图中用箭头表示，可以发现，被试只允许从上一个阶段某难度层次的模块跳转到同等或相邻难度层次的模块，如可以从“2难”模块转到“3中”模块，但不可以从“2难”模块转到“3易”模块[[63]](#footnote-63)。

### 4.2.4 路由规则（Routing Rules）

路由规则可理解为替被试挑选路径的方法，是最能体现MST自适应性的元素，其主要目的是为被试选择适合其水平的难度层次的模块。仍以1-3-3 MST为例，如图1(b)所示，“2难”模块既可以转到“3难”模块，也可以转到“3中”模块，具体跳转到哪个模块就由路由规则而定。路由规则的一般思路是基于正确作答次数、被试能力水平或者模块信息量[[64]](#footnote-64)，具体有近似最大信息量法（AMI）62、定义人口间隔法（DPI）61等。

**第五章 自适应测试思想应用于个性化练习**

自适应测试思想是选出与学习者能力水平相匹配的题，即对学生来说既不会太难也不会太简单，笔者认为做这样的题有助于学生提高，故尝试将自适应测试思想应用于个性化练习中。本章将对自适应测试思想应用于个性化练习进行可行性分析，在此基础上探索如何借鉴自适应测试的思想，最后对应用自适应测试思想后的个性化练习进行阐述。

## 5.1 可行性分析

自适应测试与个性化练习的关系类似于传统教学过程中的考试与作业，两者目的不同，自适应测试是为了测出学生的真实水平，而个性化练习是为了帮助学生提高知识水平，除此之外，两者在其他方面都是类似的。

### 5.1.1 追求目标

自适应测试和个性化练习都追求高效率。自适应测试的目标是使用最少的题最精确地测量学生的真实水平，个性化练习的目标是使用最少的题最有效地帮助学生提高知识水平。

### 5.1.2 实施对象

自适应测试和个性化练习的实施对象都是学习者。尽管目前自适应测试在国内还没有得到广泛应用，但对于学习者来说，基于计算机的测试并不陌生，如计算机等级考试、托福考试等。虽然目前基于计算机的测试的实施对象以大学生为主，但在当今这一数字化的时代，中小学生的信息素养也较高，稍加指导，即可适应在计算机上做题。

### 5.1.3 操作对象

自适应测试和个性化练习的操作对象都是题目。自适应测试的题库建设以项目反应理论为指导，题目的难度参数与被试能力具有统一量表，且具有参数不变性的特点28，这都是项目反应理论带来的优点。个性化练习针对不同的知识内容也需要大量的题目，可借鉴自适应测试的题库建设方式。

### 5.1.4 推荐过程

自适应测试和个性化练习都属于“个性化”推荐，即根据学生的能力水平来推荐适合的题目。自适应测试推荐题目的基本思路是重复循环以下两个步骤：根据学习者的能力估计值出题；根据题目的作答结果再重新估计能力值。可见，自适应测试对于学生能力值的估计是实时估计、及时更新的过程。但由于自适应测试是在某一个时段中连续进行的，在此期间学生的真实能力水平几乎不会发生变化，故学习者的能力水平值会逐渐收敛，最终趋于稳定。

不同于自适应测试，个性化练习是时间上不连续的、阶段性的过程，学生的真实能力水平可能会随着作答个性化练习而提高，故在个性化练习的过程中，学生的真实能力水平是动态变化的，因此需要实时估计学生的能力水平，并根据最近的能力水平估计值来出题，恰好与自适应测试推荐题目的基本思路类似。

## 5.2 应用于个性化练习的自适应测试选题机制的确定

由可行性分析可知，个性化练习与自适应测试有许多相似之处，而且二者最实质的操作都是做题。借鉴自适应测试思想，实质上主要是借鉴自适应测试的选题机制，选题机制又离不开能力水平的估计，而这两个方面是自适应测试的主要构成方面，也是不同自适应测试之间的主要差别方面。

第四章中已经提到，基于计算机的自适应测试有两种：CAT和MST，这两种测试各有优缺点，个性化练习应该借鉴何种自适应测试是首先要思考的问题。本章借助图1来说明自适应测试的选题机制。

### 5.2.1 CAT与MST的本质区别

CAT是以单个题为单位进行的自适应测试，而MST是以一组题为单位进行的自适应测试。两者的本质区别主要表现在以下几个方面61：

（1）两次自适应选题之间的题目数量

如CAT和MST的示例图，即图1中(a)和(b)所示，CAT是做完一道题就进行能力值的估计并自适应选出下一道题。而MST是做完一组题才进行能力值的估计并自适应选出下一组题，一组题的具体题目数量依据实际情况而定。也正因此，对于同样长度的测试，CAT自适应选题次数比MST多。

（2）测试题目的预知性

CAT是采用选题算法直接从题库中选题，在学生作答之前无法预知下一道题。而MST中每个阶段有几个不同难度水平的模块是预先设计的，且模块都是预先组装好的，MST的自适应出题并不是直接从题库中选题，实际上是从下一阶段不同难度层次的模块中选择最适合学生的模块，故MST所有的路径是可以遍历的。

（3）自适应选题的灵活性

只要题库中的题目充足且参数分布合理，CAT可以针对能力参数范围内的任意值出题。而MST的模块是预先组装好的，难度水平是固定的，故无法准确地根据任意能力值出题。

### 5.2.2 CAT与MST的优缺点

在3.3中已经提到，个性化练习的两大核心步骤分别是了解学生的水平以及根据学生的水平出题，而检查作业并修改答案是一种常见的作业行为，允许线下做题也体现了作业形式的多样性。故本节从学生水平估计的准确性、自适应出题、检查题目与修改答案、允许线下做题这四个方面对CAT和MST进行比较。

（1）学生水平估计的准确性

自适应测试以能力值来表示学生的水平。早期的CAT在实践过程中产生了高估（或低估）学生能力水平的问题[[65]](#footnote-65)，这是由于学生在初始阶段的作答过程中超常（或失常）发挥造成能力水平估计过高（或过低）[[66]](#footnote-66)，在测验长度较短的的情况下，难以调回学生真实的能力水平53。MST可减轻这一问题，因为MST的初始阶段是作答一组题才进行初始能力水平的估计，即使超常（或失常）发挥，偏差相对较小。

（2）自适应出题

一方面，在5.2.1中也提到，CAT可针对任意能力水平值来出题，而MST的模块只有特定的几个难度层次，也就是说CAT选择的题与学生的能力水平更相近，尤其是对于能力水平非常高和非常低的学生。

另一方面，由于CAT是直接从题库中选题，只要建设好题库即可；而MST需要预先设计好阶段数、模块数并预先组装好模块以供测试过程中进行选择，不得不考虑MST组装试卷的复杂性。

（3）检查题目与修改答案

完成所有题目后对已作答的题目进行检查这一行为，在纸笔练习或考试中普遍存在。但目前已有的CAT实践是不允许学生返回检查题目并修改答案的，具体的原因有[[67]](#footnote-67)降低测验的效率、增加考试时长、增加软件技术的复杂性等，而且存在学生“操控”试卷题目的风险，比如学生故意答错题目以使后面选出来的题目难度逐渐降低[[68]](#footnote-68)，全部作答完成后再返回修改答案，尽可能地将错误的答案都修改正确，尽管这一行为并不一定会提高学生的分数。

MST是做完一组题才进行自适应选择下一组题，换句话说，在这一组题内是没有自适应的，学生可以在作答题组内的题目时往前或往后跳转来检查并修改答案[[69]](#footnote-69)，对测量的精度无明显的影响[[70]](#footnote-70)。需要注意的是，只允许学生在当前阶段内修改，当进入下一个阶段时，不能返回上一个阶段进行检查并修改。

（4）允许线下做题

允许线下做题是给予学生选择做题形式的空间，尤其是在学生还未完全适应线上做题形式时，允许线下做题是一大优势。CAT是做完一题选下一题，MST是做完一组题选下一组题，显然题组形式更适合线下做题。

总的来说，MST可以弥补CAT高估（或低估）能力水平、不允许检查与修改答案、不适合线下做题的问题，但MST组装试卷较为复杂且不能针对任意能力值来出题。

### 5.2.3 应用于个性化练习的自适应测试选题机制

综合以上分析，笔者认为CAT与MST结合的形式更适合应用于个性化练习，具体的结合方式如图1所示。单从图的表象来看，图(c)是将图(a)复制过来并将圆改成矩形，实质上是将CAT的选题算法和MST的题组形式相结合。Zheng & Chang[[71]](#footnote-71)在2015年从自适应测试的角度对这种结合的方式进行了理论论证，称之为OMST框架。

这种混合形式的基本思路是，先选一组题给学生做，估计学生的初始能力水平；然后使用CAT的选题策略自适应选择下一组题，做完再重新估计学生的能力水平，如此循环。它与CAT的本质区别在于以题组为单位而不是单个题为单位进行自适应选题，与MST的本质区别在于采用CAT选题策略从题库中选题而不是从预先设计好的模块中选择。

这种混合形式综合了CAT和MST的优点，相互抵消了不足。由于是题组形式，解决了CAT高估（或低估）学生能力水平、不能检查题目并修改答案两大问题；由于是CAT选题策略，解决了MST不能针对任意能力值出题及试卷组装复杂性的问题。而且，还拥有MST的另一个优点，就是线下做题，一组组题发给学生做比一道道题更有效率。

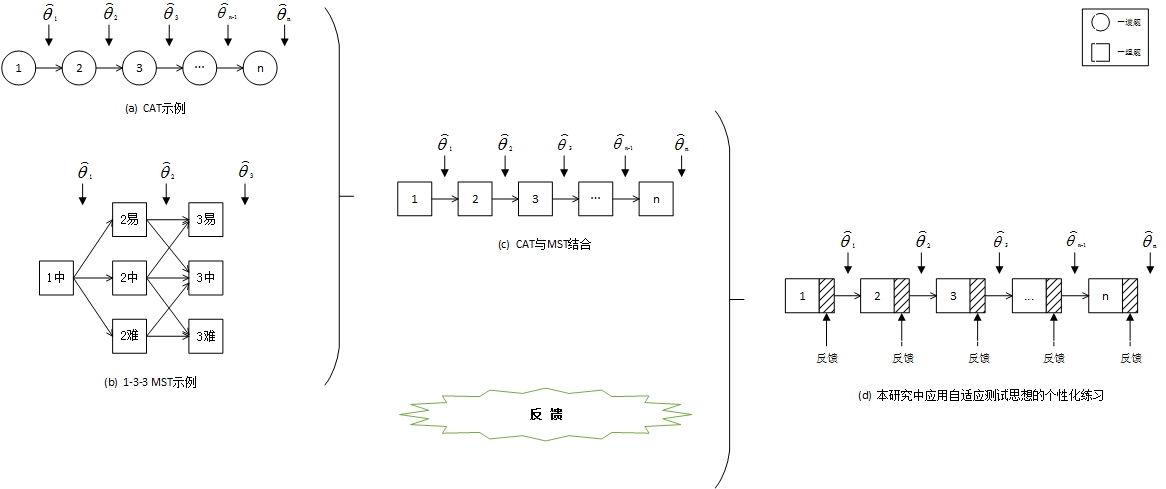


图1 应用自适应测试思想的个性化练习

## 5.3 应用自适应测试思想的个性化练习

上文已经讨论了自适应测试思想应用于个性化练习的可行性以及应该借鉴何种形式的自适应测试选题机制，本节将展开讨论应用自适应测试思想后个性化练习的组成要素、过程及特点。

### 5.3.1 应用自适应测试思想的个性化练习组成要素

应用了自适应测试思想的个性化练习如图1中（d）所示，从图中来看，是在自适应测试思想的基础上加上作答反馈。虽然应用自适应测试思想的个性化练习不是自适应测试思想与作答反馈的简单相加，但自适应测试的思想和作答反馈是构成本研究中个性化练习的两大组成要素。

（1）自适应测试思想

Lord在1970年提出观点“当一道题对于被试来说既不会太难，又不会太容易时，其测量效率是最高的”9，这就是自适应测试的核心思想。自适应测试是依据被试的能力水平进行的量体裁衣式测试[[72]](#footnote-72)，选出对被试来说既不是太简单又不是太难的题让被试作答，这样的题为可为估计被试能力水平提供最大的信息，是甄别学生能力水平最好的题。

笔者非常认同Lord的观点，当某道题对学生来说既不是太难又不是太容易时，这道题对于该生来说就是好题，故认为这一核心思想在个性化练习中同样适用。个性化练习需要选择对学生能力水平提高最有帮助的题，太简单的题重复做没有意义，太难的题不仅不会提高学生的能力水平且容易打击学生的信心，最有帮助的题就是让学生觉得恰到好处的题。

故笔者想借鉴自适应测试的思想来实施个性化练习，具体做法是借鉴CAT中的选题策略和能力水平估计方法，能力水平的估计实质上是作为选题算法下一次选题的依据。这里对选题策略思想上可能存在的一个误区进行解释，CAT的选题策略一直遵循的思想就是选出与学生能力水平最匹配的题，即难度系数等于学生能力水平的题，难度系数与学生能力水平相等并不代表这道题在学生的能力范围内，也不代表学生会做这道题。实质上，这样的题目对于学生来说，做对和做错的概率几乎是对半，Rasch模型和双参数Logistic模型是选出答对概率为0.5的题，三参数Logistic模型是选出答对概率为（1+c）/2的题，其中c指题目的猜测系数。

（2）反馈

为学生提供作答结果的反馈，是个性化练习区别于自适应测试的重要方面，也是二者目的不同所引起的不同需求。通过本文2.2节的阐述可知，反馈对学习效果有着强大的影响力。因此，在个性化练习的过程中，每一次练习作答完成后，都会告知学生作答正确与否以及正确答案，以期帮助学生更好地消化自适应选择的题目，增强个性化练习对提高学生能力水平的效果。当然，这建立在学习者接受反馈并调节自己学习过程的基础上，如增加努力程度、向教师询问不理解之处、对具体知识进行针对性地学习等。

### 5.3.2 应用自适应测试思想的个性化练习的过程

应用自适应测试思想的个性化练习的基本步骤为：首先，完成准备工作，包括建立题库、确定每一次练习的题目数量即题组长度；接着正式进入个性化练习阶段，包括根据能力水平使用选题策略选出一组题、根据作答结果进行能力水平的估计，循环执行这两个步骤，但相邻循环之间需要插入反馈，即学生接受上一组题的作答结果反馈以后，方可进入下一组题的作答；至于是否需要确立终止规则可依实际情况而定。对以上个性化练习步骤中涉及的几个主要方面进行介绍如下：

（1）题库的建设

题库的建设借鉴自适应测试的题库建设方法，以项目反应理论为基础建设即可，题库中的字段主要包括题目的参数和答案。题目的参数系数依据选用的项目反应理论模型而定，如本文实验中采用3PLM即包含区分度、难度、猜测系数三参数。

（2）题组长度的确定

这里的题组是指每一次练习的题目集合，题组长度即指每一次练习的题量。最简单的方式是设定每一次练习都具有相同的题目数量。

（3）题组的选择

对于题组的选择可以借鉴CAT的选题策略。对于初始题组的选择可以采用随机选取、从中等难度的题中选取等方法，也可由教师依据学生的平时表现进行选取。在初始能力水平估计完成之后，可采用最大Fisher信息量法、a分层法等，若有题目字数、图片数量等限制要求，可采用0-1规划方法、启发式方法等。本文实验中采用的是应用较为广泛的最大Fisher信息量法。

（4）能力水平的估计

能力水平的估计方法可直接借鉴CAT中被试能力水平的估计方法，最常用的是MLE和EAP两种方法，本文实验中采用的是MLE和EAP相结合的方法。

（5）作答结果反馈

作答结果反馈的作用是让学生了解自己的作答是否正确，帮助学生更好地消化自适应选择的题目，增强个性化练习对提高学生能力水平的效果。本文实验中反馈的具体做法是告知学生正确答案，让学生自己订正出解题过程再上交。这一过程可由家长或教师监督、检查。

（6）停止练习的规则

对于个性化练习可不制定终止规则，毕竟每一次练习完成均可看作一次暂停，几个小时或几天后再进行下一次练习都是可以的，随时可以停止练习。若要制定终止规则来判定学生何时不再需要继续练习，可以观察学生的能力水平，当练习过程中学生的能力水平不再提高或提高到某一个程度时，可提醒学生短时间内无需再继续相关知识的练习。

### 5.3.3 应用自适应测试思想的个性化练习的特点

从出题、提高能力水平的效果、题库等角度来看，应用自适应测试思想的个性化练习具有如下特点：

（1）针对学生的能力水平出题

自适应测试就是针对学生的能力水平来出题，应用自适应测试思想后的个性化练习也具有这一特点，对于每个学生产生适合他能力水平的一组题。这里的能力水平实质上是指学生对于所作答的相关知识的掌握水平，如本文是以二次根式这一章节为例进行实验，则实验中学生的能力水平实际上是指学生对二次根式这一知识的掌握水平，也是指以二次根式为限定范围的数学能力水平。

（2）动态调整学生的能力水平

最期望的情形是学生作答个性化练习后，能力水平会有不同程度的提高，此时学生的真实能力水平是动态变化的，故需要及时对学生的能力水平进行估计，持续更新每个学生的最近能力水平值，以确保选题时依据的theta值的实效性。

（3）提高能力水平的高效性

尽管自适应测试在国内还没有得到广泛应用，但自适应测试的理论已经非常成熟，CAT已经能够做到使用最少的题最精确地测量学生的能力水平，借鉴自适应测试思想的个性化练习也可以做到使用最少的题最有效地帮助学生提高能力和知识掌握水平。究其本质原因，是遵循自适应测试的核心思想，即Lord所说，选择对学生来说既不是特别难也不是特别容易的题，这样的题可以提供最大的信息量，是甄别学生能力水平最好的题，也是对学生最有帮助的题，因此对提高学生的能力水平效率较高。

（4）能力参数与题目难度参数具有同一量表及参数不变性

这是项目反应理论带来的优点，项目参数一旦估计放入题库后，一段时期内便不再更改。由于传统考试中同一学生做不同难度的考卷会得到不同的分数，相同的分数并不能代表学生的能力水平相同。故个性化练习中学生的能力水平不是用分数而是用theta值表示，学生的theta值估计不依赖于具体的题目。学生的能力水平和题目的难度参数共用同一个刻度，为应用和解释带来了很大的便利，如能力水平为0.6的学生答对难度为0.7的题的概率比答错的概率小，当该生通过个性化练习将自身的能力水平从0.6提高到0.8以后，其答对难度为0.7的题的概率比答错的概率大。

（5）间断性与连续性的结合

个性化练习与自适应测试不同，不是连续进行的过程，而是一次练习后，经过一段时间再做下一次练习，这期间的时间长度依据学生实际情况而定。但下一次练习的题目是依据上一次练习的作答结果估计出的能力水平值产生的，相邻两次个性化练习并不是没有关联，故个性化练习整体过程可看作时间上的间断性与能力水平上的连续性的结合。

（6）不是重复做相同的题

已有的个性化练习系统中，大多强调错题库的作用，即为学生保留每一次练习的错题，虽然每个学生做错的题也不一样，也可以称得上是个性化练习，但是重复做错题可能带来学生记住答案的问题，若如此，错题库的价值将大打折扣。应用自适应测试思想的个性化练习系统会自动记录该生已经做过的题，每一次练习都是在题库里未做过的题目中重新产生适合其能力水平的题，不会出现重复做题。

（7）线上和线下结合的方式

应用自适应测试思想的个性化练习需要借助软件技术来实现，最好的实现方案是建立个性化练习系统，学生拥有自己的账号，系统会记录每个学生的作答历史和每一次的能力水平值。若学生在线上作答，计算机自动给予作答反馈，学生可以线下订正及寻求教师的帮助；若学生的信息素养不够高或家长担心孩子借作答练习的名义玩电脑，则可将计算机为学生生成的题目打印出来以纸质形式给学生做，做完后由家长或教师向计算机输入答案，计算机给予反馈。这是题组形式带来的优势，既可以线上作答，也允许线下做题，但显然线下做题会给教师或家长增添工作量。

**第六章 自适应测试思想应用于个性化练习的实践研究**

## 6.1 实验目的与假设

### 6.1.1 实验目的

本次实验研究的目的在于，检验本文提出的“应用自适应测试思想的个性化练习”（本章中后文简称“个性化练习”）是否能够有效提高学生的知识水平，主要表现为是否有效提高学生的能力水平值即theta值和考试成绩。

### 6.1.2 实验假设

假设一：相比于普通练习，个性化练习可以有效提高学生的真实能力水平值。

假设二：相比于普通练习，个性化练习可以有效提高学生的考试成绩。

## 6.2 实验设计

### 6.2.1 实验对象与学科

（1）实验对象

上海市某中学同一数学教师教的初二（1）班和初二（4）班的学生，两个班各22人，共44人。

（2）学科内容

上海市初二数学：第一学期第一章“二次根式”。

### 6.2.2 实验模式的设计

本实验研究是准实验研究，采用不相等实验组对照组前测后测设计，具体的操作过程如下：

（1）教师针对“二次根式”一章内容出一份试卷对学生进行前测，以此掌握实验对象的原始情况。

（2）运用原始群体，以初二（4）班作为实验班进行个性化练习的训练，以初二（1）班作为对照班，进行同等练习量的普通练习的训练，并观察练习训练期间两个班级学生能力水平值的变化情况。

（3）教师针对“二次根式”一章内容再出一份试卷对学生进行后测，分析两个班测验成绩的变化。

（4）综合两个班级学生能力水平值和测验成绩的变化，判定个性化练习对提高学生知识水平的效果。

### 6.2.3 实验变量

（1）自变量

本次实验的自变量是不同种类的练习：个性化练习和普通练习。

实验班（四班）：个性化练习，采用自适应测试的思想出题。

对照班（一班）：普通练习，采用随机出题。

（2）因变量

本次实验的因变量是学生在练习训练期间的能力水平值和实验结束时的后测成绩。

（3）干扰变量

本实验的干扰变量有教师教授的课程内容、练习的题量与时间等，采用变变量为常量法来控制干扰变量。具体做法是：选择同一教师教授的两个班级，且练习训练时已不再学习第一章的内容（见图2实验安排）；实验班和对照班每次练习的题量、发练习及收练习的时间均相同。另外，由于实验班是每个人做不同题，为了避免这种“定制作业”让学习者产生新奇感而多付出努力，对照班也是每个人做不同题，对每一个学生进行随机出题。

### 6.2.4 实验安排

本次实验的时间安排如图2所示，将整个“练习训练”的过程安排在“不学习第一章”的阶段里，这样安排的主要原因如下：

（1）符合研究目的。教师在第一章授课完成后便进入下一章，而学生对于第一章内容的掌握水平是各不相同的，学生可以通过个性化练习自主提升，符合本研究的初衷。

（2）控制干扰变量。第二章“一元二次方程”和第三章“正比例函数和反比例函数”的主要内容与第一章的主要内容联系不大，在学习完第一章之后到前三章总复习之前的这段时间进行实验，尽可能地控制了教师授课内容对后测成绩的影响以保证前后测成绩的变化是由练习训练引起。

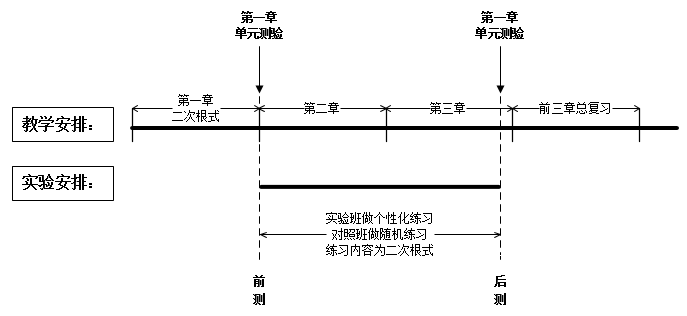


图2 实验安排

### 6.2.5 练习形式

实验时所有的练习题采用纸质作答形式，每张纸上印有学生的姓名，即学生“定制”的练习。与教师商讨后决定每次练习出五道题，即个性化练习和随机练习每次的题量均为五道题。采用纸质形式主要因为预实验时采用在线做题形式效果不佳。实验学校拥有自己的错题库网站，教师可以在网站上针对某一或某些学生布置练习并设置截止时间，是非常好的实验平台。今年六月底至七月份初，在该平台上使用其他章节的内容进行了预实验，但发现学生作业的完成率很低，一方面可能因为是暑假期间，学生不愿意做题，另一方面也可能是家长对学生上网的限制。故在十月至十一月的正式实验时，没有采用该平台，所有的练习题是通过计算机出题后再打印出来给学生做，即采用纸质练习。这也突显了题组形式的一大优点——可以线下做题。

### 6.2.6 辅助工具

（1）实验辅助工具：本实验前期准备及实验进行过程中的主要辅助工具为R和BILOG软件。实验过程中个性化练习和普通练习的出题、计算学生theta值部分，都使用自主编制的程序来完成，所有程序基于R 3.3.0软件编制。建设题库时使用BILOG 3.0软件对题目参数进行估计。

（2）数据分析工具：R 3.3.0、SPSS 21.0、EXCEL 2013软件。

## 6.3 题库的建设

本实验中的题库建设以项目反应理论为指导，具体步骤包括题目的选择、获得作答矩阵、数据与模型的拟合、参数估计、筛选题目入库等。

### 6.3.1 题目的来源

从教师推荐的质量较高的练习册和试卷中选出160道题，题目的内容均关于二次根式，题型包括单选题、填空题、计算题。由于本文采用的是0-1评分模型，故将计算题改为填空题，即题库中的题目类型只包含单选题和填空题。

### 6.3.2 获取学生的作答结果

收取全年级六个班共137名学生的练习册和试卷，从以上160道题中选出所有学生均作答的题目132道，再剔除作答不认真的同学6名，最后共记录131名学生对132道题目的作答结果。记录方式为答对记1、答错记0，对于计算题是当成填空题按照最终结果来判定对错，最后得到了131行132列的0-1作答矩阵。

### 6.3.3 模型拟合与参数估计

根据获取的0-1作答矩阵，使用BILOG软件进行3PLM参数估计。估计时假设数据符合三参数Logistic模型，BILOG会自行对每个项目进行拟合优度的检查，根据拟合统计量及其显著性结果，删除不拟合的题目10道，记录剩下122道题的区分度a、难度b和猜测系数c。

### 6.3.4 筛选题目入库

剔除难度参数b值不在[-3,+3]范围内的题目8道，最终入库的题目有114道。题库中参数的分布如表1所示。可以发现，题目的难度偏低，主要是因为很难的题目教师没有布置给所有学生做，故在获取学生的作答结果时这部分题因作答不完全已被删除。使用R软件的shapiro.test()命令进行正态检验，显示参数b的分布近似正态分布，其概率分布曲线如下图3所示，中等难度（b的范围在[-1.5,1.5]）的题目占总题数的76%。

表1 题库中的参数分布

|  |  |  |  |
| --- | --- | --- | --- |
|  | 区分度a | 难度b | 猜测系数c |
| 均值 | 1.238 | -0.901 | 0.176 |
| 标准差 | 0.328 | 0.999 | 0.020 |
| 最小值 | 0.648 | -2.983 | 0.119 |
| 最大值 | 2.461 | 2.813 | 0.220 |

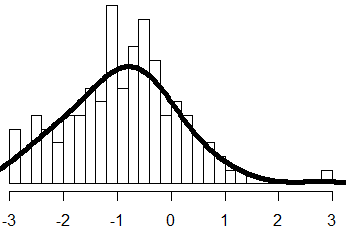


图3 题目难度参数的概率分布曲线

## 6.4 实验过程

实验主要分为两部分工作，一部分是用R语言程序计算theta值和出题，另一部分是去学校发练习、收练习、发批改好的练习。每次练习的题量为五道题，共进行了四次练习，具体实验过程（如图4）描述如下。

（1）实验开始：在实验开始前，已有前测成绩，正式实验从第一次出题开始；

（2）第一次出题：根据学生之前的单元练习作答结果估算出初始theta值，实验班第一次出题时根据初始theta值采用最大Fisher信息量法从题库中选出五道题，对照班中每个学生均是从题库中随机选出五道题；

（3）第一次发练习：打印题目，印有每个学生的姓名，去学校发给学生做；

（4）收练习并批改：一般隔一天去学校收练习并进行批改，批改时将学生做错的题目的正确答案写在旁边，即告诉学生正确答案；

（5）反馈：将批改后的练习再发给学生，并要求学生自己订正，写出解题过程；

（6）计算theta值：将学生的作答结果输入到R软件中，使用自己编写的R语言程序计算每个学生的theta估计值；

（7）出题：实验班的每个学生根据步骤（6）得到的theta估计值，采用最大Fisher信息量法从题库中选出五道题，而对照班的每个学生仍是随机选五道题；

（8）发练习：打印题目，印有每个学生的姓名，去学校发给学生做，注意此时学生已经订正好上一次的练习题。即步骤（6）~（7）可以与步骤（5）同时进行，但这三个步骤一定发生在步骤（4）和步骤（8）之间；

（9）循环步骤（4）~（8），直至四次练习完成，循环在步骤（6）处停止。

（10）实验结束：实施后测，记录后测成绩，实验结束。

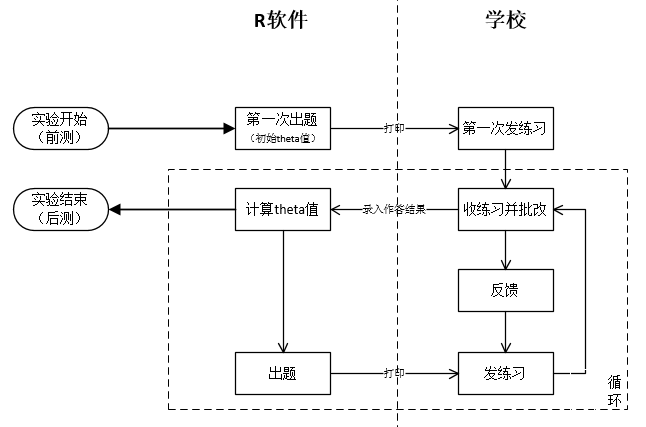


图4 实验过程

## 6.5 实验结果与分析

### 6.5.1 学生能力水平值的整体变化情况

实验过程中已提到，四班和一班分别进行了四次练习，每次练习完成后，均根据学生的作答结果来估算学生的能力水平值即theta值，分别记为theta1、theta2、theta3、theta4；而第一次练习出题时依据的初始theta值，记为theta0。本节将对这五个theta值的变化情况进行分析，必要时以four\_和one\_对两个班的theta值变量进行区分。另由于本节还涉及模拟数据，故称实验过程中获得的theta值数据为实验theta值，而称模拟得到的theta值数据为模拟theta值。学生的真实能力水平称为真实theta值，一般情况下无从得知，因此我们才需要对真实theta值进行估算，实验theta值和模拟theta值都是真实theta值的估计值。

#### 6.5.1.1 实验theta值的变化情况

（1）实验theta值的概率密度分布情况

首先，使用R软件的ggplot2包画出四班和一班各theta值的概率密度函数图（如图5），可以看出各theta值分布近似正态分布。

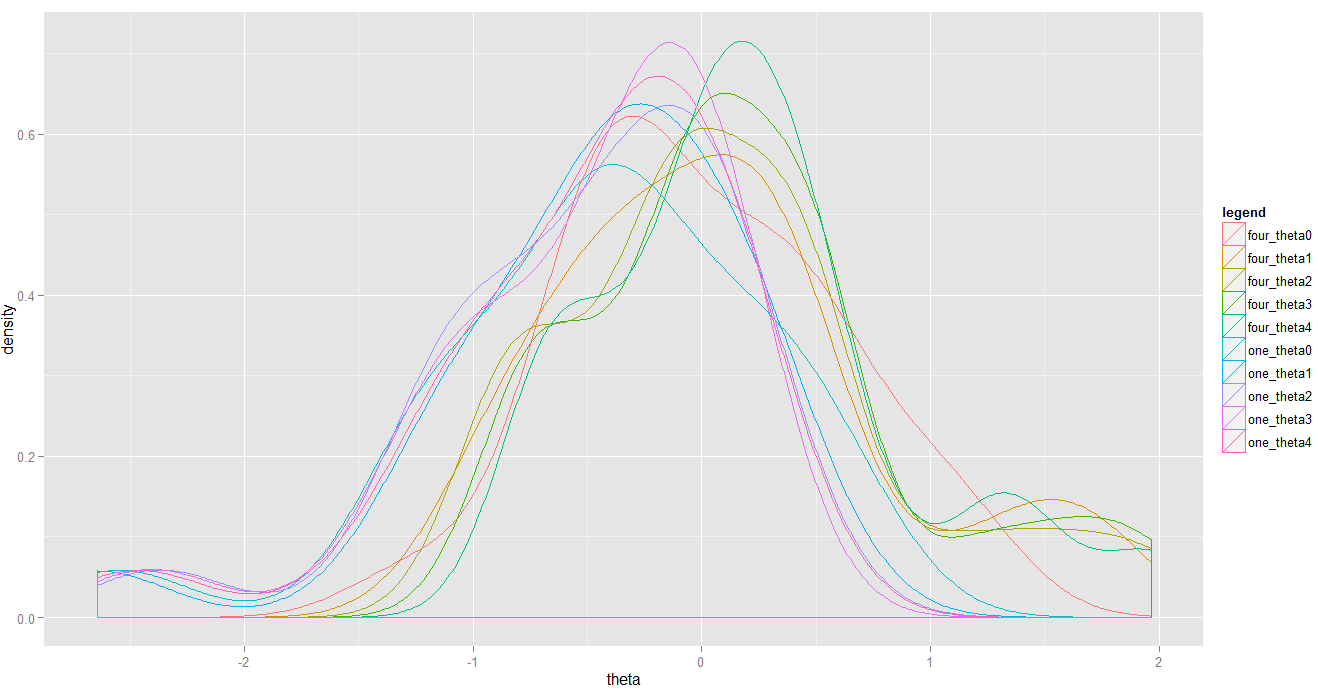


图5 两个班各theta值的概率密度函数图

（2）描述性统计

两个班级各theta值的描述性统计指标如表2所示。初步来看，四班的均值有变化，一班的均值变化不明显；四班和一班的标准差均变化不大；四班的最小值有变化，一班的最小值变化不明显；四班和一班的最大值均有变化。

表2 两个班各theta值的描述性统计指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 班级 | theta值 | 均值 | 标准差 | 最小值 | 最大值 |
| 四班 | theta0 | 0.006 | 0.597 | -1.238 | 1.175 |
| theta1 | 0.08 | 0.727 | -0.998 | 1.718 |
| theta2 | 0.11 | 0.732 | -0.888 | 1.918 |
| theta3 | 0.18 | 0.717 | -0.828 | 1.923 |
| theta4 | 0.21 | 0.685 | -0.745 | 1.965 |
| 一班 | theta0 | -0.441 | 0.737 | -2.549 | 0.637 |
| theta1 | -0.466 | 0.68 | -2.644 | 0.452 |
| theta2 | -0.493 | 0.635 | -2.369 | 0.346 |
| theta3 | -0.502 | 0.628 | -2.41 | 0.148 |
| theta4 | -0.49 | 0.642 | -2.466 | 0.253 |

结合两个班级各theta值的箱形图（如图6所示），可以更加直观地看出四班的theta值的四分位点和中位数均有变化，整体呈上移的趋势；而一班的各项指标有一定的波动，整体显示出微弱的下降趋势。此外，箱形图还可看出四班有两个学生的水平比一般学生高，而一班有一个学生的水平比一般学生低。

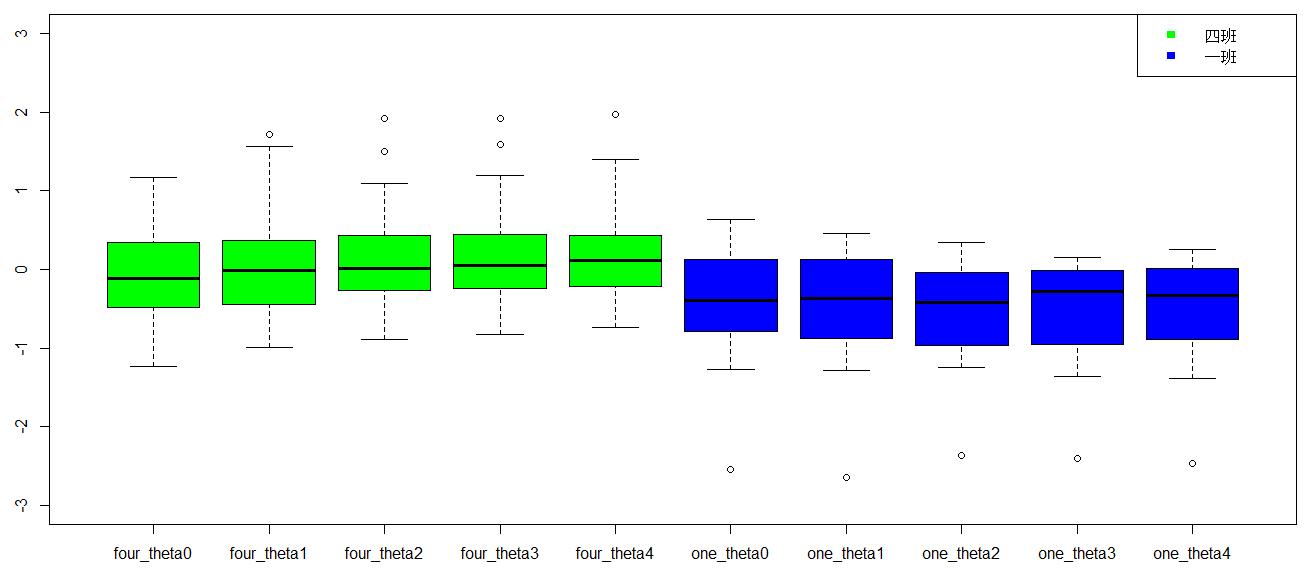


图6 两个班各theta值的箱形图

（3）两个班实验theta均值的变化对比

如果说最大值和最小值只是一个值，具有偶然性，那么均值是最具效力的工具。两个班级的theta平均值变化情况如图7所示，从theta0到theta4，相比一班而言，四班有较为明显的上升趋势；一班却有微弱的下降趋势；四班和一班之间的差距在逐渐加大。

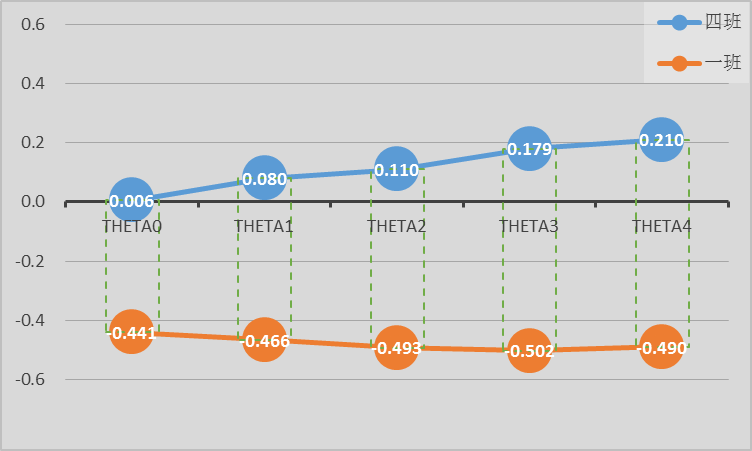


图7 两个班theta均值的变化对比情况

#### 6.5.1.2 实验与模拟theta值的对比变化情况

通过上一小节的分析，我们是否就可以得出结论，认为个性化练习可以提高学生的真实能力水平值，而随机练习对学生的真实能力水平值有微弱的降低作用呢？按照常理来说，随机练习对学生的真实能力水平值应该有微弱的提高作用，即使没有提高也不应该有降低作用，为什么一班学生的实验theta值会降低呢？不禁让人产生疑问。笔者提出了三个可能的原因：第一，学生的真实能力水平确实下降了；第二，学生作答不认真，没有表现出真实水平；第三，恰好随机出的题对于学生偏难，使得作答结果不理想，以致于估计出的实验theta值偏低。同理，四班学生的实验theta值会提高也有三个原因：第一，学生的真实能力水平确实上升了；第二，学生通过找答案等方式得到正确的作答结果；第三，个性化练习选出来的题对于学生偏简单，使得作答结果较理想，以致于估计出的实验theta值偏高。对于一班，通过对学生作业的观察，可以排除第二个原因；对于四班，个性化练习也只是以普通作业的形式发放，并不是考试，学生不会为了获得好的成绩而去找答案，而且每个学生的题目也不一样，也不会因为相互抄作业而提高作答表现，故第二个原因也可以排除。下面将分别对一班学生能力水平降低、四班学生能力水平提高的第三个原因进行检验。

本小节将通过模拟学生作答的方式来排查第三个原因。具体做法是，假设学生的真实能力水平没变，即还是theta0的水平（这里以初始theta值作为真实能力水平的估计值），以theta0的水平来模拟学生作答这四次练习的题目，模拟作答的算法描述如下：

根据题目的三参及学生的当前能力真值算出答对该题的概率p

产生服从均匀分布（）的随机数r

如果r小于等于p：

判定为答对，记为1

否则：

判定为答错，记为0

通过模拟作答的算法即可得到学生的模拟作答矩阵，根据模拟作答矩阵得到学生的模拟theta值。若模拟theta值与实验theta值相差不大，说明学生的真实能力水平没有发生变化。为了减小偶然误差，笔者进行了30次模拟作答，也就是说，对于每一次练习，每个学生都有30个模拟theta值，取这30个模拟theta值的平均值作为该生本次练习的最终模拟theta值，接下来讨论的模拟theta值就是指最终的模拟theta值。

（1）一班的实验theta值与模拟theta值对比

一班实验theta值和模拟theta值的描述性统计指标对比如图8所示，为了使图形更为直观，均值和标准差使用柱形图，且坐标轴在右侧；最大值和最小值使用折线图，且坐标轴在左侧。容易发现，一班的实验theta值和模拟theta值数据的各项指标差别均不大，尤其是均值，几乎没有发生变化。这一结果说明了图3中微弱的下降趋势可能是由题目对学生来说稍偏难或其他偶然因素引起的，实质上一班学生的真实能力水平几乎没变。于是，可得出实验结论：在练习题量不多的情况下，随机练习对于提高或降低学生的真实能力水平均没有显著作用。（2）四班的实验与模拟theta值对比

既然一班学生实验theta值的下降趋势是“假”的，那么四班学生实验theta值的上升趋势会不会也是“假”的呢？依然采用上述模拟作答的方式来检验。

四班实验theta值和模拟theta值的描述性统计指标对比如图9所示，添加次坐标轴的方式同图8，且为了保证涨降幅度视觉效果上的一致性，次坐标轴的刻度也与图8一致。从图9可知，四班学生的模拟theta值几乎没变，四班学生实验theta值却逐渐提高，且实验theta值和模拟theta值的差距在逐渐增大，说明四班学生的真实能力水平确实提高了。于是，可得出实验结论：即使练习题量不多，个性化练习可以提高学生的真实能力水平值，充分体现了提高的“效率”，这才是有意义的“有效作业”。

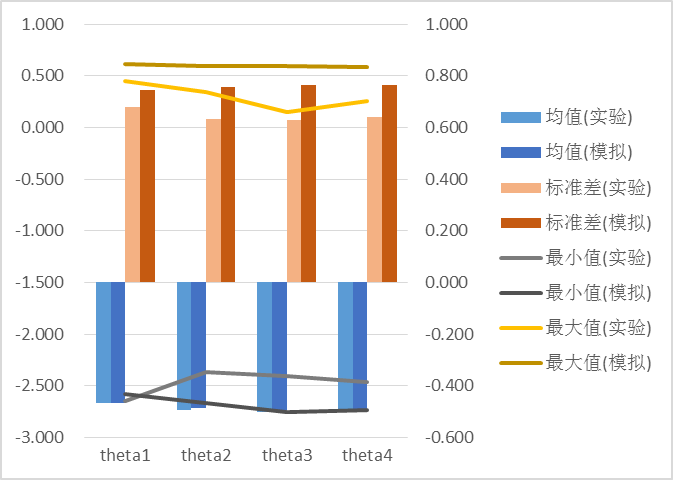


图8 一班实验和模拟theta值的描述性统计指标对比

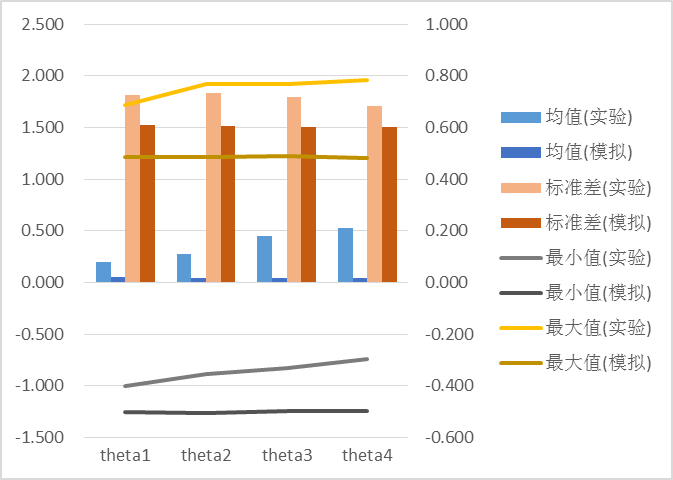


图9 四班实验和模拟theta值的描述性统计指标对比

### 6.5.2 学生成绩的整体变化情况

为了验证练习干预对学生成绩的影响，教师出了一份与之前的单元测验（前测）难度相当的考卷，测试的内容仍是二次根式，记录两个班级的成绩作为后测。四班和一班的前后测成绩描述性统计指标如表3所示，明显看出，无论是前测还是后测，四班的平均分都高于一班的平均分，且四班的标准差都低于一班的标准差，说明四班的成绩不仅高且分布更集中；练习干预后，一班的平均分上升了1.4分，四班的平均分上升了5.2分。本小节将使用SPSS 21.0软件对实验班和对照班的前后测成绩差异进行显著性检验，即T检验。

表3 两个班前后测成绩的描述性统计指标

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 班级 | 前后测 | 均值 | 标准差 | 最小值 | 最大值 |
| 一班 | 前测 | 67.5 | 13.291 | 46 | 85 |
| 后测 | 68.9 | 11.992 | 41 | 90 |
| 四班 | 前测 | 80.2 | 8.434 | 61 | 92 |
| 后测 | 85.4 | 8.215 | 70 | 100 |

#### 6.5.2.1 实验班和对照班前测成绩差异比较

对实验班和对照班的前测成绩进行独立样本t检验，结果如表4所示，F=7.957，P=0.007<0.05，说明两个班级前测成绩的方差是不齐的，即否认方差相等的假设，选用方差不相等时的t检验结果，t=-3.806，P=0.001<0.05，说明四班和一班的前测成绩有显著性差异，即两个班的水平有明显差异。

表4 实验班和对照班前测成绩独立样本t检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **独立样本检验** | | | | | | |
|  | | 方差方程的Levene检验 | | 均值方程的t检验 | | |
| F | Sig. | t | df | Sig.(双侧) |
| 前测 | 假设方差相等 | 7.957 | .007 | -3.806 | 42 | .000 |
| 假设方差不相等 |  |  | -3.806 | 35.554 | .001 |

#### 6.5.2.2 实验班、对照班各自前后测成绩差异比较

从表4已知，实验班和对照班不是平行班级，意味着不能直接对两个班的后测成绩进行比较，可以检验两个班的前后测成绩差异显著性。分别对实验班和对照班的前测、后测成绩进行配对样本t检验，结果如表5所示，四班t=3.208，P=0.004<0.05，说明实验班的后测成绩与前测成绩有显著性差异；一班t=1.605，P=0.123>0.05,说明对照班的后测成绩与前测成绩没有显著差异。基本可以表明，个性化练习对学生的成绩有显著作用，而随机练习对学生的成绩没有显著作用。

表5 实验班和对照班前测、后测成绩配对样本t检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **成对样本检验** | | | | | | |
|  | | 成对差分 | | t | df | Sig.(双侧) |
| 均值 | 标准差 |
| 对1 | 四班后测-四班前测 | 5.136 | 7.511 | 3.208 | 21 | .004 |
| 对2 | 一班后测-一班前测 | 1.455 | 4.251 | 1.605 | 21 | .123 |

#### 6.5.2.3 实验班和对照班后测与前测成绩差值差异比较

为了进一步验证实验假设，对实验班和对照班后测与前测成绩差值进行方向性假设检验，假设实验班提高的分数显著大于一班提高的分数。用实验班和对照班的后测成绩减去前测成绩得到两组差值，对这两组差值进行独立样本t检验（单尾检验），结果如表6所示。F=3.058，P=0.088>0.05，说明两个班级前后测成绩差值的方差是齐的，选用方差相等时的t检验结果，t=2.001，P=.026<0.05，说明假设成立，四班提高的分数显著大于一班提高的分数，即个性化练习可以有效提高学生的成绩。

表6 实验班和对照班的后测与前测成绩差值独立样本t检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **独立样本检验（单尾检验）** | | | | | | |
|  | | 方差方程的Levene检验 | | 均值方程的t检验 | | |
| F | Sig. | t | df | Sig.(单侧) |
| 后测-前测 | 假设方差相等 | 3.058 | .088 | 2.001 | 42 | .026 |
| 假设方差不相等 |  |  | 2.001 | 33.202 | .027 |

### 6.5.3 个性化练习对不同成绩段学生的影响的尝试性分析

6.5.1和6.5.2的分析结果显示，从全班整体水平来看，个性化练习可以有效提高学生的theta值和成绩。本节将尝试分析，个性化练习是否对每个成绩段的学生都有作用。

一班和四班前后测分数的分布情况如图10所示，后测与前测相比可以发现：从高分来看，80分及以上的人数，四班明显增多而一班几乎没变；从低分来看，四班的前测分数最低也处于60-70分数段，该分数段学生的后测分数均提高到70分以上，但一班60分以下的人数也减少了两个人。仅从分布图的人数变化并不能作出结论，为进一步分析，笔者在综合考虑两个班的分数和班级人数的基础上，将成绩分为60分以下、60-80分（含60分不含80分）、80分及以上三段，其中由于四班的前测分数中60分以下的人数为零，故接下来只对60-80分、80分及以上两个分数段的学生进行分析。

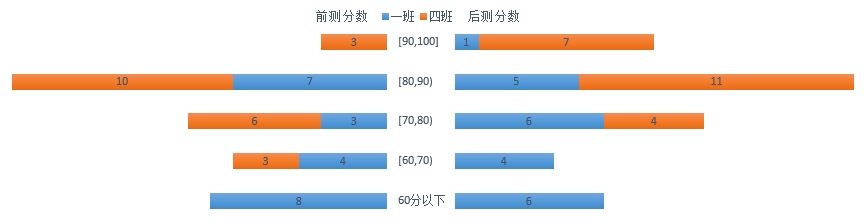


图10 两个班各分数段学生的前后测分数分布情况

#### 6.5.3.1 个性化练习对各分数段学生theta值的影响

对于60-80分、80分及以上两个分数段，一班和四班的theta平均值变化情况如图11所示，实线表示实验theta值，虚线表示模拟theta值，可以发现：除一班60-80分数段的学生以外，两个班级各分数段学生的实验theta值变化趋势，均类似图7中班级整体实验theta值变化趋势；无论是60-80分数段还是80分及以上的分数段，四班和一班之间的差距都在逐渐加大；无论是60-80分数段还是80分及以上的分数段，四班的实验theta值与模拟theta值之间的差距也在逐渐加大，且60-80分数段的学生增长的幅度比80分及以上分数段的学生更高。可见，个性化练习对于不同分数段学生的theta值都有提高，但提高的程度不同。

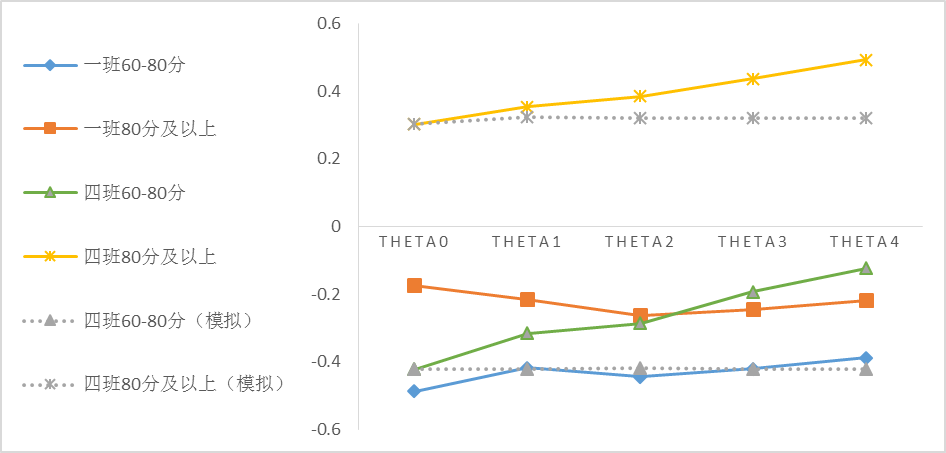


图11 两个班各分数段学生的theta均值变化情况

#### 6.5.3.2 个性化练习对各分数段学生成绩的影响

（1）个性化练习对60-80分数段学生成绩的影响

类似于6.5.2.3，对两个班60-80分数段的学生，后测与前测成绩差值进行方向性假设检验，假设四班60-80分的学生提高的分数显著大于一班60-80分的学生提高的分数。独立样本t检验（单尾检验）的结果如表7所示。F=2.732，P=0.121>0.05，说明两个班级前后测成绩差值的方差是齐的，选用方差相等时的t检验结果，t=2.902，P=.006<0.05，说明假设成立，对于60-80分数段的学生，四班提高的分数显著大于一班提高的分数，即个性化练习可以有效提高60-80分数段学生的成绩。

表7 两个班60-80分的学生后测与前测成绩差值独立样本t检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **独立样本检验（单尾检验）** | | | | | | |
|  | | 方差方程的Levene检验 | | 均值方程的t检验 | | |
| F | Sig. | t | df | Sig |
| 后测-前测 | 假设方差相等 | 2.732 | .121 | 2.902 | 14 | .006 |
| 假设方差不相等 |  |  | 3.225 | 10.275 | .005 |

（2）个性化练习对80分及以上分数段学生成绩的影响

对两个班80分及以上这一分数段的学生，后测与前测成绩差值进行方向性假设检验，假设四班80分及以上的学生提高的分数显著大于一班80分及以上的学生提高的分数。独立样本t检验（单尾检验）的结果如表8所示。F=1.273，P=0.274>0.05，说明两个班级前后测成绩差值的方差是齐的，选用方差相等时的t检验结果，t=1.069，P=.150>0.05，说明假设不成立，对于80分以上这一分数段的学生，实验班提高的分数没有显著大于对照班。

表8 两个班80分及以上的学生后测与前测成绩差值独立样本t检验结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **独立样本检验（单尾检验）** | | | | | | |
|  | | 方差方程的Levene检验 | | 均值方程的t检验 | | |
| F | Sig. | t | df | Sig.(单侧) |
| 后测-前测 | 假设方差相等 | 1.273 | .274 | 1.069 | 18 | .150 |
| 假设方差不相等 |  |  | 1.196 | 16.612 | .125 |

对于这一结果，可能的解释是，80分及以上本身就是一个较高的分数，提升的空间有限，所以显著提高有些困难。笔者分别对两个班级80分及以上分数段的学生，计算后测减去前测分数的差值的均值，得到四班均值为1.92，一班均值为-0.14，即四班80分及以上的学生，后测成绩相比前测成绩平均分提高了近2分，而一班80分及以上的学生，后测成绩相比前测成绩平均分数基本没变。说明实验班的分数还是有提高的，只是提高不够显著，也与图11中显示的信息相符，即80分及以上分数段的学生提高的程度不如60-80分数段的学生。

另一种可能的解释是，试卷分数对学生水平的估计不如能力值精确。相同分数的学生能力水平值不一定相同，也就是说，也许学生的水平提高了但提高的幅度小，考卷没有考出学生水平的变化。

**第七章 总结与展望**

## 7.1 研究总结

本文尝试将自适应测试思想应用于个性化练习中，并进行了实证研究。笔者将围绕以下四个方面展开总结：

（1）个性化练习的目标和需求

本文对个性化练习的概念进行了界定，认为个性化练习的目标是用最少的题来最有效地提高学生的知识水平，而使用何种个性化练习推荐机制来达到该目标是需要解决的关键问题。

（2）自适应测试的思想应用于个性化练习的理由

自适应测试的核心思想是，选出对学生来说既不是太简单又不是太难的题让学生作答，可为估计学生能力水平提供最大的信息，是甄别学生能力水平最好的题。虽然这样的题是与学生能力水平相当，但不代表学生一定会做，实质上，这样的题目对于学生来说，答对和答错的概率几乎是对半。笔者认为，这样的题借助反馈的影响力是最有助于学生提高的题。

（3）自适应测试思想应用于个性化练习的探索

笔者提出将自适应测试的思想应用于个性化练习。首先，对自适应测试思想应用于个性化练习进行可行性分析；然后，根据个性化练习的需求，从学生水平估计的准确性、自适应出题、检查题目与修改答案、允许线下做题这四个方面对CAT和MST进行比较，并确定将CAT的选题算法与MST的题组形式相结合作为个性化练习的选题机制；同时，增加对学习效果有强大影响力的反馈因素，以期帮助学生更好地消化个性化练习选择的题目。于是，形成了本文提出的应用自适应测试思想的个性化练习。

（4）开展准实验研究进行效果检验

为验证应用自适应测试思想的个性化练习对提高学生的能力水平和成绩的有效性开展准实验研究。而题库的建设是准实验研究的前提。笔者从教师推荐的质量较高的练习册和试卷中选出160道题；获取学生作答结果时，删除作答不完全的题目28道；模型拟合检验时，删除不拟合的题目10道；根据参数估计结果，剔除难度参数b值不在[-3,+3]范围内的题目8道；经过一系列筛选，最后剩余114道题入库。使用R软件的shapiro.test()命令进行正态检验，显示参数b的分布近似正态分布，中等难度（b的范围在[-1.5,1.5]）的题目占总题数的76%，题库的整体质量达标。

准实验研究采用不相等实验组对照组前测后测设计，选取上海市某中学同一教师教的两个原始班级开展实验，练习内容为数学学科初二年级“二次根式”章节，实验班采用本文提出的应用自适应测试思想的个性化练习，对照班采用随机出题。研究结果表明：（1）整体而言，在练习题量不多的情况下，随机练习对于提高学生的能力水平和成绩均没有显著的作用；而应用自适应测试思想的个性化练习可以有效提高学生的能力水平和成绩；（2）分成绩段而言，个性化练习对于各分数段学生的theta值和成绩都有提高，但提高的程度不同；80分及以上分数段的学生提高的幅度不如60-80分数段的学生。

## 7.2 研究局限与不足

由于本人的研究水平以及实验客观条件的限制，本研究还存在许多不足之处。

首先，关于题库建设方面。由于人力物力的限制，题库中的题是从教师推荐的质量较好的练习册和试卷中选取，并不是由专家命制，题目的质量有待考察。估计题目参数时，只有137名学生的作答结果，学生人数还不够多，但这已然是该校初二年级所有的学生。经过层层筛选之后，最终只有114道题进入题库，题库不够大。

第二，实验人数少。只有两个班级参与实验，每个班级只有22人，实验样本偏小。而且，在探索分析个性化练习对不同成绩段学生的影响时，受到了小样本的限制。四班学生均在中上水平，前测成绩在60分以下的人数为零，故无法对该成绩段进行实验组和对照组的对比分析。

第三，练习次数不多。为了尽可能地控制干扰变量，笔者将实验安排在学习完“二次根式”章节之后、在期中考试复习之前的这一段时间内。由于是纸质形式发放练习，故发练习、收练习、给予反馈都需要时间，期间学生还有外出实践课等，故最后在这段时间内，一共只进行了四次练习。如果增加练习次数，可以从实验数据中挖掘更多的信息，实验结果也会更有说服力。

第四，本文提出的基于自适应测试思想的个性化练习，是将CAT的选题策略与MST的题组形式相结合，但CAT的选题策略有很多，笔者在实验过程中采用的是便于实现的最大Fisher信息量法，并没有对CAT的多种选题算法进行比较，择优按需选取。

## 7.3 研究展望

个性化练习是个性化学习的重要组成部分，但由于教师的时间与精力有限，使得个性化练习在教学过程中的实践受到了限制。故笔者认为，技术支持下的基于不同目标的个性化练习系统的研发会成为主要的研究方向，而其中练习的推荐机制是最核心的研究点。本文就是该方向的一次尝试，以提高学生的知识水平为目标，借鉴自适应测试思想形成个性化练习的推荐机制，并用实验研究证实了其有效性。笔者希望今后的研究可以从以下三个方面继续展开：

（1）扩大实证研究的范围。不仅扩大实验样本，也要增加实验时长。一方面是进一步证实个性化练习的有效性；另一方面，想通过个性化练习与普通练习的长期对比，观察学生提高的幅度变化。换句话说，随着练习题量的增加，个性化练习是否对各个层次的学生一直有效，普通练习是否达到一定题量开始对学生提高有帮助，提高到一定程度是否会饱和等。

（2）加入认知诊断。本文的实证研究是以某一章的内容为例，根据数学能力水平来出题，然而一章的内容包含多个知识点，诊断学生当前的知识掌握情况可以更好地指导教学。故后期笔者将尝试采用计算机化自适应认知诊断测验（简称CD-CAT）思想，或者根据学生能力水平出题的同时考虑到学生的知识点掌握情况，来制定个性化练习的推荐机制，并与之前的推荐机制作比较。

（3）进一步证实反馈在个性化练习中的重要作用。已有许多研究表明，反馈对学习效果的重要影响力。笔者想要证实的是，自适应测试思想选题与作答反馈的结合不是两者的简单相加。笔者认为，如果说自适应测试思想选的题相比于普通题目，对学生的能力水平的提高作用高出一个值a的话，那么自适应测试思想选题与作答反馈的结合相比于普通题目与作答反馈的结合，对学生的能力水平的提高作用高出的值应该大于a。用公式形式可以表示为：若[自适应测试思想选题 - 普通练习 = a]，那么[自适应测试思想选题&反馈 - 普通练习&反馈 > a]。当然这只是笔者的猜测，希望可以在后续研究中证实。

**参考文献**

[1] Bangert-Drowns R L, Kulik C L C, Kulik J A, et al. The instructional effect of feedback in test-like events[J]. Review of educational research, 1991, 61(2): 213-238.

[2] Barton M A, Lord F M. An upper asymptote for the three-parameter Logistic item response model[J]. ETS Research Report Series, 1981:1-20.

[3] Binet A, Simon T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux[J]. L'année Psychologique, 1905, 11(1): 191-244.

[4] Birnbaum A. Some latent train models and their use in inferring an examinee's ability[M]// Statistical Theories of Mental Test Scores. 1968:395-479.

[5] Bock, R., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431-444.

[6] Brusilovsky P, Millán E. User models for adaptive hypermedia and adaptive educational systems[C]//The adaptive web. Springer-Verlag, 2007: 3-53.

[7] Carlson S. ETS finds flaws in the way online GRE rates some students[J]. Chronicle of Higher Education, 2000, 47(8): A47.

[8] Chang H H, Ying Z. To weight or not to weight? Balancing influence of initial items in adaptive testing[J]. Psychometrika, 2008, 73(3): 441-450.

[9] Chen S Y, Ankenmann R D, Chang H H. A comparison of item selection rules at the early stages of computerized adaptive testing[J]. Applied Psychological Measurement, 2000, 24(3): 241-255.

[10] Čisar S M, Čisar P, Pinter R. Evaluation of knowledge in Object Oriented Programming course with computer adaptive tests[J]. Computers & Education, 2016, 92: 142-160.

[11] Clariana R B, Wagner D, Murphy L C R. Applying a connectionist description of feedback timing[J]. Educational Technology Research and Development, 2000, 48(3): 5-22.

[12] Corbett A T, Anderson J R. Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2001: 245-252.

[13] Embretson S E, Reise S P. Item response theory[M]. New Jersey:Psychology Press, 2013:266.

[14] Hattie J, Timperley H. The power of feedback[J]. Review of educational research, 2007, 77(1): 81-112.

[15] Hendrickson A. An NCME instructional module on multistage testing[J]. Educational Measurement: Issues and Practice, 2007, 26(2): 44-52.

[16] Hornke L F. Item response times in computerized adaptive testing[J]. Psicológica, 2000, 21(1-2): 175-189.

[17] Hua-Hua Chang，Jiahe Qian，ZhiLiang Ying. a-Stratified Multistage Computerized Adaptive Testing with b Blocking[J].Applied Psychological Measurement，2001，25（4）：333-341.

[18] Hua-Hua Chang，ZhiLiang Ying．a-Stratified Multistage Computerized Adaptive Testing[J]．Applied Psychological Measurement，1999，23（3）：211-222.

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

[20] Jodoin M G, Zenisky A, Hambleton R K. Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes[J]. Applied Measurement in Education, 2006, 19(3): 203-220.

[21] Kingsbury G G, Houser R L. Assessing the utility of item response models: Computerized adaptive testing[J]. Educational Measurement: Issues and Practice, 1993, 12(1): 21-27.

[22] Kulhavy R W, Stock W A. Feedback in written instruction: The place of response certitude[J]. Educational Psychology Review, 1989, 1(4): 279-308.

[23] Lord F M. A broad-range tailored test of verbal ability[J]. Applied Psychological Measurement, 1977, 1(1): 95-100.

[24] Lord F M. Maximum likelihood and Bayesian parameter estimation in item response theory[J]. Journal of Educational Measurement, 1986, 23(2): 157-162.

[25] Lord F M. Robbins-Monro procedures for tailored testing[J]. Educational and Psychological Measurement, 1971.

[26] Lord F. A theory of test scores[J]. Psychometric monographs, 1952,7.

[27] Lord F. Some test theory for tailored testing[C].//Computer assisted instruction, testing, and guidance. New York:Harper & Row,1970: 139–183

[28] Lu H, Hu Y, Gao J. The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing[J]. Computers & Education, 2016, 100: 45-55.

[29] Lu J. Personalized e-learning material recommender system[C]//International conference on information technology for application. 2004: 374-379.

[30] Luecht R, Brumfield T, Breithaupt K. A testlet assembly design for adaptive multistage tests[J]. Applied Measurement in Education, 2006, 19(3): 189-202.

[31] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease[J]. 1959,22:719-748.

[32] Marsh E J, Lozito J P, Umanath S, et al. Using verification feedback to correct errors made on a multiple-choice test[J]. Memory, 2012, 20(6): 645-653.

[33] Metcalfe J, Kornell N, Finn B. Delayed versus immediate feedback in children’s and adults’ vocabulary learning[J]. Memory & Cognition, 2009, 37(8): 1077-1087.

[34] Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models[J]. Applied Psychological Measurement, 2000, 24(1): 50-64.

[35] Owen R J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing[J]. Journal of the American Statistical Association，1975，70：351-356.

[36] Pashler H, Cepeda N J, Wixted J T, et al. When does feedback facilitate learning of words?[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2005, 31(1): 3-8.

[37] Quellmalz E S, Pellegrino J W. Technology and testing[J]. Science, 2009, 323(5910): 75-79.

[38] Rasch G. Probabilistic models for some intelligence and achievement tests[M]. Copenhagen: Danish Institute for Educational Research, 1960.

[39] Samejima F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika monograph supplement, 1969.

[40] Samejima F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika monograph No.17, 1969, 34.

[41] Shute V J. Focus on formative feedback[J]. Review of educational research, 2008, 78(1): 153-189.

[42] Stocking M L, Lewis C. Controlling item exposure conditional on ability in computerized adaptive testing[J]. Journal of Educational and Behavioral Statistics, 1998, 23(1): 57-75.

[43] Sympson J B, Hetter R D. Controlling item-exposure rates in computerized adaptive testing[C]//Proceedings of the 27th annual meeting of the Military Testing Association.San Diego CA：Navy Personnel Research and Development Center，1985: 973-977.

[44] Triantafillou E, Georgiadou E, Economides A A. The design and evaluation of a computerized adaptive test on mobile devices[J]. Computers & Education, 2008, 50(4): 1319-1330.

[45] Van der Linden, W. J., & Glas, C. A. W. Computerized adaptive testing: Theory and practice[M]. Dordrecht: Kluwer Academic, 2000:viii.

[46] Vispoel W P, Hendrickson A B, Bleiler T. Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results[J]. Journal of Educational Measurement, 2000, 37(1): 21-38.

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

[48] Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450.

[49] Wingersky M S, Lord F M. An investigation of methods for reducing sampling error in certain IRT procedures[J]. Applied Psychological Measurement, 1984, 8(3): 347-364.

[50] Wise S L. A Critical Analysis of the Arguments for and against Item Review in Computerized Adaptive Testing[R]. New York: the Annual Meeting of the National Conference on Measurement in Education,1996.

[51] Xu X, Chang H, Douglas J. A simulation study to compare CAT strategies for cognitive diagnosis[C]//annual meeting of the American Educational Research Association, Chicago. 2003.

[52] Yan, D., von Davier, A. A., & Lewis, C. Computerized multistage testing: Theory and applications[M]. Boca Raton:CRC Press, 2014: xix.

[53] Yao L, Boughton K. Multidimensional linking for tests with mixed item types[J]. Journal of Educational Measurement, 2009, 46(2): 177-197.

[54] Yi Q, Zhang J, Chang H. Identifying practical indices for enhancing item pool security[C]//annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada. 2005.

[55] Yi Q，Chang H H．a-Stratified design with content blocking[J]．British Journal of Mathematical and Statistical Psychology，2003，56（2）：389-378.

[56] Ying Cheng. Computerized Adaptive Testing - New Development and Application[D]：[Doctoral Dissertation]. Urbana- Champaign：University of Illinois at Urbana- Champaign，2008:7-8.

[57] Zheng Y, Chang H H. Multistage testing, on-the-fly multistage testing, and beyond[M]. // In Y. Cheng, & H-H. Chang (Eds.), Advancing methodologies to support both summative and formative assessments. Charlotte, NC: Information Age Publishing, 2014: 21-39.

[58] Zheng Y, Chang H H. On-the-fly assembled multistage adaptive testing[J]. Applied Psychological Measurement, 2015, 39(2): 104-118.

[59] 郭辰.高中英语诊断性练习系统的需求分析、模块设计和知识体系构建[D].中国科学技术大学,2014:22.

[60] 国家中长期教育改革和发展规划纲要（2010-2020）[DB/OL]. http://www.hprc.org.cn/gsgl/zggk/guijiajg/xiangguanwenjian/201007/t20100730\_56482.html,2016-11-12.

[61] 简小珠,戴海琦,张敏强,彭春妹. CAT选题策略分类概述[J]. 心理学探新,2014,05:446-451.

[62] 蒋一君,邱飞岳,刘迎春,吴皖赣.基于错题库的个性化练习生成模型研究[J].中国教育信息化,2011,08:74-76.

[63] 罗照盛.项目反应理论基础[M].北京：北京师范大学出版社，2012:91+95+19+5.

[64] 王睿. 基于计算机化多阶段自适应测验的路由规则比较及其应用研究[D].江西师范大学,2015.

[65] 王文泉.错题管理系统中个性化推荐练习算法的设计与实现[J].中国教育信息化,2016,11:67-70+92.

[66] 维纳.控制论[M]. 郝季仁,译.北京:科学出版社,1983:98.

[67] 吴皖赣.基于知识地图的试题组织及个性化练习评价研究与应用[D].浙江工业大学,2012.

[68] 张华华,程莹. 计算机化自适应测验(CAT)的发展和前景展望(续)[J]. 考试研究,2005,02:22-41.

[69] 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,2005,01:12-24.

[70] 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.

[71] 张文波.中小学教育信息化发展新阶段问题的现状及对策研究[J].中国电化教育,2014,05:39-43.

[72] 张心.CAT常用能力估计方法比较及其优化：能力综合估计方法开发[D].江西师范大学,2014.

**后记**

“真舍不得这里啊”，当我发出如此感慨时，小伙伴说“等盲审过了再说舍不得”……三年一眨眼就过去了，但是回想第一天来学校的场景，又觉得很久远，只能说时间过得又快又慢。很幸运当初能够保研进入华东师范大学学习，有机会参与国际会议，有机会聆听中外教授的精彩学术报告，使我在专业知识水平增长的同时，眼界也更加开阔。感谢华东师范大学，给予我如此美好的三年。

时间悄然而逝，却在不知不觉中收获成长。记得第一次汇报英文文献熬了一个通宵，我永远记得那天晚上的感受：逃避没有用，怕也没有用，焦虑只会让你更加无措，唯有冷静地去解决问题。从那天起，我开始相信，只要你愿意花时间去学，你就可以做到。这是我研究生三年学习期间一直信奉的准则。出乎意料的是，当初觉得会一直深刻印在脑海中的记忆，现在却很浅很浅了。陡然发现，曾经完成起来如此吃力的事情，现在却已很平常，这就是进步吧。人生因为需要超越，才会成长。很享受这三年来，从早到晚在图书馆的日子。

研究生生涯的学习离不开导师的指引。感谢张华华教授带领我进入计算机自适应测试的研究领域，每次见面时您的点拨都让我受益匪浅，欣赏您年轻幽默的心态，喜欢您让我们提问引发我们思考，更震惊于您一针见血的解答。感谢顾小清教授带领我进入学习分析的研究领域，第一次见您时就被您的微笑融化，后来加入e2lab团队学习，更是折服于您干练、高效、拼搏的工作风格和端庄大方的人格魅力。感谢陈伟运老师领我参与项目，不会忘记您耐心的指导和认真负责的态度。感谢冷静老师和吴忭老师在学习和生活中的解疑答惑。感谢已成为老师的胡艺龄师姐在项目和论文完成过程中给予的帮助。真的非常感谢你们，得到你们的指导，是我的幸运。

同时，我也要向我的小伙伴们致谢。感谢刘妍师姐帮助我寻找毕设的实验对象，感谢郑隆威师兄在代码学习方面的指导，感谢石小恋、胡思畅、崔家琪、姚媛媛、冯仰存以及其他e2lab成员的帮助。感谢我的室友蔡海云、郝珍艺、李佳赟，一起在宿舍疯傻玩闹，一起在图书馆拼搏奋斗。感谢我的同学们，陪伴我走过这三年，因为有你们，回忆才充满色彩。还要特别感谢仙霞高中的朱老师对我完成毕设的支持和帮助。

感谢我的家人给予我的关心和支持。朋友们说“从你的性格能看出你拥有一个幸福的家”，是的，我拥有一个温暖又幸福的家，爸妈对我爱而不溺。感谢你们的付出，没有你们就没有现在的我。

想感谢的人还有很多，我都会记在心里。这三年，是我至今成长最多的三年。初来时的消极情绪已无影无踪，我学会了主动去探索自己感兴趣的领域，提升了自主学习的能力，明白了馅饼不是天上掉下来的，现在的我是过去的我努力换来的，未来的我需要现在的我继续努力。研究生生涯结束了，但是学习生涯并没有结束，不管以后做什么样工作，走什么样的路，都会保持一直在学习的心态。

2017年3月13日于华东师大图书馆

1. 国家中长期教育改革和发展规划纲要（2010-2020）[DB/OL]. http://www.hprc.org.cn/gsgl/zggk/guijiajg/xiangguanwenjian/201007/t20100730\_56482.html,2016-11-12. [↑](#footnote-ref-1)
2. 张文波.中小学教育信息化发展新阶段问题的现状及对策研究[J].中国电化教育,2014,05:39-43. [↑](#footnote-ref-2)
3. 蒋一君,邱飞岳,刘迎春,吴皖赣.基于错题库的个性化练习生成模型研究[J].中国教育信息化,2011,08:74-76. [↑](#footnote-ref-3)
4. 郭辰.高中英语诊断性练习系统的需求分析、模块设计和知识体系构建[D].中国科学技术大学,2014:22. [↑](#footnote-ref-4)
5. 王文泉.错题管理系统中个性化推荐练习算法的设计与实现[J].中国教育信息化,2016,11:67-70+92. [↑](#footnote-ref-5)
6. 吴皖赣.基于知识地图的试题组织及个性化练习评价研究与应用[D].浙江工业大学,2012. [↑](#footnote-ref-6)
7. Binet A, Simon T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux[J]. L'année Psychologique, 1905, 11(1): 191-244. [↑](#footnote-ref-7)
8. Lord F. A theory of test scores[J]. Psychometric monographs, 1952,7. [↑](#footnote-ref-8)
9. Lord F. Some test theory for tailored testing[C].//Computer assisted instruction, testing, and guidance. New York:Harper & Row,1970: 139–183 [↑](#footnote-ref-9)
10. Triantafillou E, Georgiadou E, Economides A A. The design and evaluation of a computerized adaptive test on mobile devices[J]. Computers & Education, 2008, 50(4): 1319-1330. [↑](#footnote-ref-10)
11. Xu X, Chang H, Douglas J. A simulation study to compare CAT strategies for cognitive diagnosis[C]//annual meeting of the American Educational Research Association, Chicago. 2003. [↑](#footnote-ref-11)
12. Čisar S M, Čisar P, Pinter R. Evaluation of knowledge in Object Oriented Programming course with computer adaptive tests[J]. Computers & Education, 2016, 92: 142-160. [↑](#footnote-ref-12)
13. Chen S Y, Ankenmann R D, Chang H H. A comparison of item selection rules at the early stages of computerized adaptive testing[J]. Applied Psychological Measurement, 2000, 24(3): 241-255. [↑](#footnote-ref-13)
14. Stocking M L, Lewis C. Controlling item exposure conditional on ability in computerized adaptive testing[J]. Journal of Educational and Behavioral Statistics, 1998, 23(1): 57-75. [↑](#footnote-ref-14)
15. Samejima F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika monograph supplement, 1969. [↑](#footnote-ref-15)
16. Hornke L F. Item response times in computerized adaptive testing[J]. Psicológica, 2000, 21(1-2): 175-189. [↑](#footnote-ref-16)
17. Lu H, Hu Y, Gao J. The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing[J]. Computers & Education, 2016, 100: 45-55. [↑](#footnote-ref-17)
18. Yi Q, Zhang J, Chang H. Identifying practical indices for enhancing item pool security[C]//annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada. 2005. [↑](#footnote-ref-18)
19. Yao L, Boughton K. Multidimensional linking for tests with mixed item types[J]. Journal of Educational Measurement, 2009, 46(2): 177-197. [↑](#footnote-ref-19)
20. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease[J]. 1959,22:719-748. [↑](#footnote-ref-20)
21. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models[J]. Applied Psychological Measurement, 2000, 24(1): 50-64. [↑](#footnote-ref-21)
22. Van der Linden, W. J., & Glas, C. A. W. Computerized adaptive testing: Theory and practice[M]. Dordrecht: Kluwer Academic, 2000:viii. [↑](#footnote-ref-22)
23. Yan, D., von Davier, A. A., & Lewis, C. Computerized multistage testing: Theory and applications[M]. Boca Raton:CRC Press, 2014: xix. [↑](#footnote-ref-23)
24. Quellmalz E S, Pellegrino J W. Technology and testing[J]. Science, 2009, 323(5910): 75-79. [↑](#footnote-ref-24)
25. Rasch G. Probabilistic models for some intelligence and achievement tests[M]. Copenhagen: Danish Institute for Educational Research, 1960. [↑](#footnote-ref-25)
26. Birnbaum A. Some latent train models and their use in inferring an examinee's ability[M]// Statistical Theories of Mental Test Scores. 1968:395-479. [↑](#footnote-ref-26)
27. Barton M A, Lord F M. An upper asymptote for the three-parameter Logistic item response model[J]. ETS Research Report Series, 1981:1-20. [↑](#footnote-ref-27)
28. 罗照盛.项目反应理论基础[M].北京：北京师范大学出版社，2012:91+95+19+5. [↑](#footnote-ref-28)
29. 维纳.控制论[M]. 郝季仁,译.北京:科学出版社,1983:98. [↑](#footnote-ref-29)
30. Shute V J. Focus on formative feedback[J]. Review of educational research, 2008, 78(1): 153-189. [↑](#footnote-ref-30)
31. Hattie J, Timperley H. The power of feedback[J]. Review of educational research, 2007, 77(1): 81-112. [↑](#footnote-ref-31)
32. Kulhavy R W, Stock W A. Feedback in written instruction: The place of response certitude[J]. Educational Psychology Review, 1989, 1(4): 279-308. [↑](#footnote-ref-32)
33. Bangert-Drowns R L, Kulik C L C, Kulik J A, et al. The instructional effect of feedback in test-like events[J]. Review of educational research, 1991, 61(2): 213-238. [↑](#footnote-ref-33)
34. Pashler H, Cepeda N J, Wixted J T, et al. When does feedback facilitate learning of words?[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2005, 31(1): 3-8. [↑](#footnote-ref-34)
35. Marsh E J, Lozito J P, Umanath S, et al. Using verification feedback to correct errors made on a multiple-choice test[J]. Memory, 2012, 20(6): 645-653. [↑](#footnote-ref-35)
36. Corbett A T, Anderson J R. Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2001: 245-252. [↑](#footnote-ref-36)
37. Metcalfe J, Kornell N, Finn B. Delayed versus immediate feedback in children’s and adults’ vocabulary learning[J]. Memory & Cognition, 2009, 37(8): 1077-1087. [↑](#footnote-ref-37)
38. Clariana R B, Wagner D, Murphy L C R. Applying a connectionist description of feedback timing[J]. Educational Technology Research and Development, 2000, 48(3): 5-22. [↑](#footnote-ref-38)
39. Lu J. Personalized e-learning material recommender system[C]//International conference on information technology for application. 2004: 374-379. [↑](#footnote-ref-39)
40. Brusilovsky P, Millán E. User models for adaptive hypermedia and adaptive educational systems[C]//The adaptive web. Springer-Verlag, 2007: 3-53. [↑](#footnote-ref-40)
41. Ying Cheng. Computerized Adaptive Testing - New Development and Application[D]：[Doctoral Dissertation]. Urbana- Champaign：University of Illinois at Urbana- Champaign，2008:7-8. [↑](#footnote-ref-41)
42. Embretson S E, Reise S P. Item response theory[M]. New Jersey:Psychology Press, 2013:266. [↑](#footnote-ref-42)
43. Lord F M. Robbins-Monro procedures for tailored testing[J]. Educational and Psychological Measurement, 1971. [↑](#footnote-ref-43)
44. Owen R J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing[J]. Journal of the American Statistical Association，1975，70：351-356. [↑](#footnote-ref-44)
45. Lord F M. A broad-range tailored test of verbal ability[J]. Applied Psychological Measurement, 1977, 1(1): 95-100. [↑](#footnote-ref-45)
46. Sympson J B, Hetter R D. Controlling item-exposure rates in computerized adaptive testing[C]//Proceedings of the 27th annual meeting of the Military Testing Association.San Diego CA：Navy Personnel Research and Development Center，1985: 973-977. [↑](#footnote-ref-46)
47. Hua-Hua Chang，ZhiLiang Ying．Global Information Approach to Computerized Adaptive Testing[J]．Applied Psychological Measurement，1996，20（3）：213-229. [↑](#footnote-ref-47)
48. Hua-Hua Chang，ZhiLiang Ying．a-Stratified Multistage Computerized Adaptive Testing[J]．Applied Psychological Measurement，1999，23（3）：211-222. [↑](#footnote-ref-48)
49. Hua-Hua Chang，Jiahe Qian，ZhiLiang Ying. a-Stratified Multistage Computerized Adaptive Testing with b Blocking[J].Applied Psychological Measurement，2001，25（4）：333-341. [↑](#footnote-ref-49)
50. Yi Q，Chang H H．a-Stratified design with content blocking[J]．British Journal of Mathematical and Statistical Psychology，2003，56（2）：389-378. [↑](#footnote-ref-50)
51. 简小珠,戴海琦,张敏强,彭春妹. CAT选题策略分类概述[J]. 心理学探新,2014,05:446-451. [↑](#footnote-ref-51)
52. 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,2005,01:12-24. [↑](#footnote-ref-52)
53. 张华华,程莹. 计算机化自适应测验(CAT)的发展和前景展望(续)[J]. 考试研究,2005,02:22-41. [↑](#footnote-ref-53)
54. Wingersky M S, Lord F M. An investigation of methods for reducing sampling error in certain IRT procedures[J]. Applied Psychological Measurement, 1984, 8(3): 347-364. [↑](#footnote-ref-54)
55. Samejima F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika monograph No.17, 1969, 34. [↑](#footnote-ref-55)
56. Bock, R., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431-444. [↑](#footnote-ref-56)
57. Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450. [↑](#footnote-ref-57)
58. Lord F M. Maximum likelihood and Bayesian parameter estimation in item response theory[J]. Journal of Educational Measurement, 1986, 23(2): 157-162. [↑](#footnote-ref-58)
59. 张心.CAT常用能力估计方法比较及其优化：能力综合估计方法开发[D].江西师范大学,2014. [↑](#footnote-ref-59)
60. Kingsbury G G, Houser R L. Assessing the utility of item response models: Computerized adaptive testing[J]. Educational Measurement: Issues and Practice, 1993, 12(1): 21-27. [↑](#footnote-ref-60)
61. Zheng Y, Chang H H. Multistage testing, on-the-fly multistage testing, and beyond[M]. // In Y. Cheng, & H-H. Chang (Eds.), Advancing methodologies to support both summative and formative assessments. Charlotte, NC: Information Age Publishing, 2014: 21-39. [↑](#footnote-ref-61)
62. Luecht R, Brumfield T, Breithaupt K. A testlet assembly design for adaptive multistage tests[J]. Applied Measurement in Education, 2006, 19(3): 189-202. [↑](#footnote-ref-62)
63. Jodoin M G, Zenisky A, Hambleton R K. Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes[J]. Applied Measurement in Education, 2006, 19(3): 203-220. [↑](#footnote-ref-63)
64. 王睿. 基于计算机化多阶段自适应测验的路由规则比较及其应用研究[D].江西师范大学,2015. [↑](#footnote-ref-64)
65. Carlson S. ETS finds flaws in the way online GRE rates some students[J]. Chronicle of Higher Education, 2000, 47(8): A47. [↑](#footnote-ref-65)
66. Chang H H, Ying Z. To weight or not to weight? Balancing influence of initial items in adaptive testing[J]. Psychometrika, 2008, 73(3): 441-450. [↑](#footnote-ref-66)
67. Wise S L. A Critical Analysis of the Arguments for and against Item Review in Computerized Adaptive Testing[R]. New York: the Annual Meeting of the National Conference on Measurement in Education,1996. [↑](#footnote-ref-67)
68. Wainer H. Some practical considerations when converting a linearly administered test to an adaptive format[C]. Educational Measurement: Issues and Practice, 1993, 12(1): 15-20. [↑](#footnote-ref-68)
69. Hendrickson A. An NCME instructional module on multistage testing[J]. Educational Measurement: Issues and Practice, 2007, 26(2): 44-52. [↑](#footnote-ref-69)
70. Vispoel W P, Hendrickson A B, Bleiler T. Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results[J]. Journal of Educational Measurement, 2000, 37(1): 21-38. [↑](#footnote-ref-70)
71. Zheng Y, Chang H H. On-the-fly assembled multistage adaptive testing[J]. Applied Psychological Measurement, 2015, 39(2): 104-118. [↑](#footnote-ref-71)
72. 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. [↑](#footnote-ref-72)