成果引用与转载证明材料

1、成果引用情况：

通过查询中国期刊网发现本成果被引用次数 2 次，分别是何壮等 2012 年发表于《中国考试》第 10 期第 18-24 页，引用本成果；罗洪刚等 2012 年发表于《黔南民族师范学院学报》第 4 期第 47-52 页，引用本成果。

2、成果转载情况：

（1） “中国社会科学引文索引”（CSSCI）全文收录；

（1） 国务院发展研究中心信息网（国研网）全文转载：
http://www.drcnet.com.cn/eDRCnet.common.web/DocSummary.aspx?docid=2933762&leafid=109

（2） 维普资讯网：
http://www.cqvip.com/QK/96925X/201206/42353900.html

（3） 中国知网（CNKI）：
http://www.cnki.com.cn/Article/CJFDTotal-JYYJ201206013.htm

（4） 豆丁网： http://www.docin.com/p-445897661.html

（5） 道客巴巴网：http://www.doc88.com/p-987350873052.html

（6） 吾喜杂志网：http://wuxizazhi.cnki.net/Search/JYYJ201206013.html

文献检索证明

作者姓名：

作者单位：

该作者 2012 年发表的文献“Rasch 模型在研究生入学考试质量分析中的应用”被中文社会科学引文索引（CSSCI）收录。

检索结果见附件。

特此证明

证明单位：贵州师范大学图书馆

证明人：

二〇一三年五月十三日

【文件序号】11G0522012060009
【来源篇名】Rasch模型在研究生入学考试质量分析中的应用
【英文篇名】The Application of Rasch Model in the Quality Analysis of Graduate Entrance Examination
【来源作者】赵守盈/何妃霞/陈维/罗杰/关丹丹
【作者姓名拼音】
【文章类型】论文
【基金】贵州省教育厅自然科学重点项目(黔教科(2007)19号)/贵州省高等学校教学质量与教学改革工程重点项目(高教发[2011]28-1)/贵州师范大学精品课程“心理测量”建设项目
【期刊】教育研究
【第一机构】
【机构名称】50001////教育部考试中心,100084
【学科分类】教育
【第一作者】
【中图类号】G40
【年代卷期】2012,33(6):61-65
【标引词】研究生/入学考试/Rasch模型/质量分析
【基金类别】9/8/9

参考文献：
1.教育部考试中心.2010年全国硕士研究生入学统一考试心理学专业基础综合考试大纲.北京：高等教育出版社，2009
2.Rasch,Georg.Probabilistic models for some intelligence and attainment tests.Copenhagen: Institute of Educational Research, 1960
3.Wright,B.D..Rasch models overview.Journal of Applied Measurement.2000.(1)
4.Bond,T.G..Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.).Mahwah,N.J.: Erlbaum, 2007
5.Rasch,Georg.Objective comparisons.UNESCO Seminar.Voksenasen, Oslo, 1964
6.晏子.心理科学领域内的客观测量——Rasch模型之特点及发展趋势.心理科学进展.2010.(18)
7.Coe,Rebert.Comparability of GCSE examinations in different subjects: an application of the Rasch model.Oxford Review of Education.2008.(5)
8.Clements,Douglas H..Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment.Educational Psychology.2008.(28)
9.Randall,Jennifer.Examining Teacher Grades Using Rasch Measurement Theory.Journal of Educational Measurement.2009.(1)
10.彭聃龄.普通心理学.北京：北京师范大学出版社，2004
11.朱滢.实验心理学.北京：北京师范大学出版社，2000

CSSCI

中文社会科学引文索引
Chinese Social Sciences Citation Index

重新选择数据库 被引文献检索

命中结果1篇
检索表达式：LY12,:Z1=/赵守盈/

检索字段： 作 者 检索词： 赵守盈 检索 清除检索

序号	来源作者	来源篇名	期刊	年代卷期	全文
1		Rasch模型在研究生入学考试质量分析中的应用	教育研究	2012,33(6):61-65	显示

选择所有 显示所选 下载所选

转跳至第 1 屏

来源篇名	Rasch模型在研究生入学考试质量分析中的应用
英文篇名	The Application of Rasch Model in the Quality Analysis of Graduate Entrance Examination
来源作者	
作者机构	贵州师范大学教育科学学院,550001////教育部考试中心,100084
文献类型	论文
学科类别	教育学
中图类号	G643
基金项目	贵州省教育厅自然科学重点项目(黔教科(2007)19号)/贵州省高等学校教学质量与教学改革工程重点项目(高教发[2011]28-1)/贵州师范大学精品课程“心理测量”建设项目
来源期刊	教育研究
年代卷期	2012,33(6):61-65
标引词	研究生/入学考试/Rasch模型/质量分析

参考文献

1 教育部考试中心.2010年全国硕士研究生入学统一考试心理学专业基础综合考试大纲..北京:高等教育出版社,2009

2 Rasch,Georg.Probabilistic models for some intelligence and attainment tests..Copenhagen:Institute of Educational Research,1960

3 Wright,B.D..Rasch models overview.Journal of Applied Measurement.2000.(1)

4 Bond,T.G..Applying the Rasch model:Fundamental measurement in the human sciences (2nd ed.)..Mahwah,N.J.:Erlbaum,2007

5 Rasch,Georg.Objective comparisons.UNESCO Seminar.Voksenasen,Oslo,1964

6 晏子.心理科学领域内的客观测量——Rasch模型之特点及发展趋势.心理科学进展.2010.(18)

7 Coe,Rebert.Comparability of GCSE examinations in different subjects:an application of the Rasch model.Oxford Review of Education.2008.(5)

8 Clements,Douglas H..Development of a measure of early mathematics achievement using the Rasch model:the Research-Based Early Maths Assessment.Educational Psychology.2008.(28)

9 Randall,Jennifer.Examining Teacher Grades Using Rasch Measurement Theory.Journal of Educational Measurement.2009.(1)

10 彭聃龄.普通心理学..北京:北京师范大学出版社,2004

11 朱滢.实验心理学..北京:北京师范大学出版社,2000

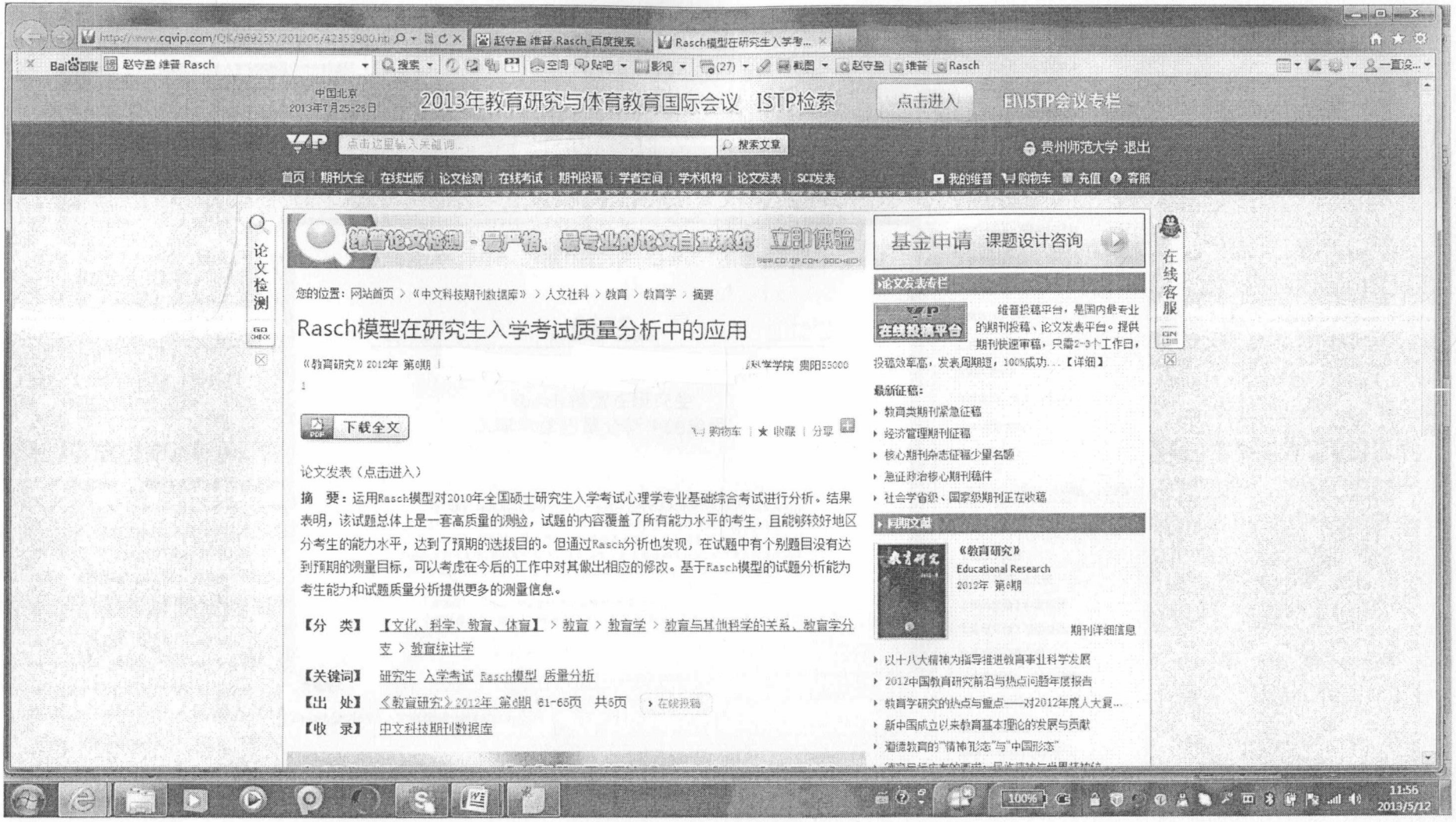
被中国知网转载


CNKI知网空间
www.cnki.com.cn
主办：中国知网
全球最大的数字图书馆
充值中心
购买充值卡
帮助中心
送卡上门
下载阅读器
English
首页
期刊全文库
学位论文库
会议论文库
学术百科
吾喜杂志
工具书
CNKI学问
基础科学 | 工程科技I辑 | 工程科技II辑 | 医药卫生科技 | 信息科技 | 农业科技 | 哲学与人文科学 | 社会科学I辑 | 社会科学II辑 | 经济管理
输入关键词
搜文献
搜期刊
在您毕业前参加考试。
GRE
立即注册
《教育研究》2012年06期
加入收藏 获取最新
Rasch模型在研究生入学考试质量分析中的应用
【摘要】：运用Rasch模型对2010年全国硕士研究生入学考试心理学专业基础综合考试进行分析。结果表明,该试题总体上是一套高质量的测验,试题的内容覆盖了所有能力水平的考生,且能够较好地区分考生的能力水平,达到了预期的选拔目的。但通过Rasch分析也发现,在试题中有个别题目没有达到预期的测量目标,可以考虑在今后的工作中对其做出相应的修改。基于Rasch模型的试题分析能为考生能力和试题质量分析提供更多的测量信息。
【作者单位】：
【关键词】：研究生入学考试
【基金】：贵州省教育厅自然科学重点项目(项目编号:黔教科(2007)019号) 贵州省高等学校教学质量与教学改革工程重点项目"基于PBL理论改进心理教育测量教学改革研究"(项目批准号:黔高教发[2011]28-1) 贵州师范大学精品课程"心理测量"建设项目的阶段性成果
【分类号】：G643
【正文快照】：
一、引言我国的硕士研究生入学心理学专业基础综合考试是为高等院校和科研院所招收心理学学科的硕士研究生而设置的具有选拔性质的统一入学考试科目。自2007年起,全国心理学专业硕士研究生考试实行全国统一考试(简称"统考"),目前绝大多数拥有心理学硕士点的高等院校和科研
下载全文
更多同类文献
机构专属电子阅览室
免费创建啦!
立即申请>>
相关期刊
>外国教育研究
>兰州大学学报(社会科学版...
>江西师范大学学报(哲学社...
>哈尔滨工业大学学报
>昭乌达蒙族师专学报
>比较教育研究
>中华人民共和国国务院公报
>山东政报
>合肥工业大学学报(自然科...
>北京师范大学学报(社会科...
相关机构
>《每日新闻》
相关作者
没有考不上的研究生
暑期50天魔鬼集训营
跨考考研
关闭
11:47
2013/5/12

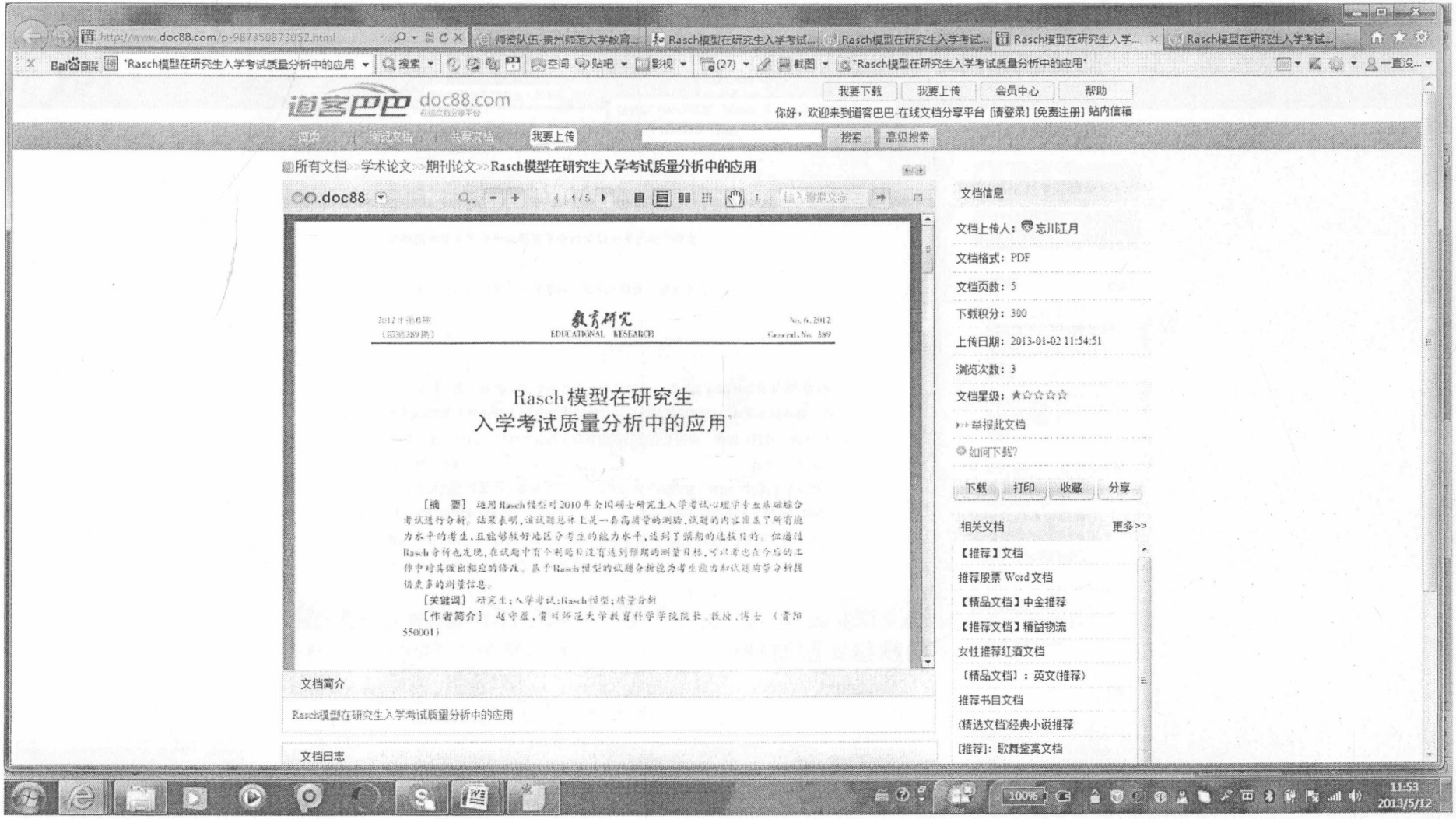
被吾喜杂志网转载


免费创建《机构专属电子阅览室》
立即申请
吾喜杂志
中国知网镜下电子杂志网站
登 录
免费注册
用QQ帐号登录
吾喜首页
个人阅览室
机构阅览室
吾喜美文
吾喜教育
免费杂志
搜索
搜杂志
搜文章
来源：《教育研究》2012年第6期 作者：赵守盈;何妃霞;陈维;罗杰;关丹丹;
选择字号 大 中 小
Rasch模型在研究生入学考试质量分析中的应用
分享到:
一、引言我国的硕士研究生入学心理学专业基础综合考试是为高等院校和科研院所招收心理学学科的硕士研究生而设置的具有选拔性质的统一入学考试科目。自2007年起,全国心理学专业硕士研究生考试实行全国统一考试(简称"统考"),目前绝大多数拥有心理学硕士点的高等院校和科研院所都参加了统考。实行统一入学考试的目的是为了能够更科学、公平、有效地测量、评价、鉴别考生掌握心理学学科大学本科阶段专业基础知识、基本理论、基本方法的水平和分析问题、解决问题的能力,它对"优化初试,强化复试",确保心理学学科——硕士研究生的入学质量等具有重要的意义和作用。[1]心理学统一考试的试题由教育部考试中心统一命制,在每次考试前,教育部考试中心都会发布每年的考试大纲,即《全国硕士研究生入学统一考试心理学专业基础综合考试大纲》。考试大纲既是命题的依据和方向,也是考生进行专业复习的指导性资料。作为心理学硕士研究生入学考试改革的一项新措施,截止到2010年,我国心理学硕士研究生入学统考已经实行了4年,统考试题的质量如何,测量的目标和试题的结构是否符合大纲要求,各个心理学学科比例分配是否.....
[继续阅读本文]
下载本文
阅读本刊
开通阅览室
如何获取本文>>
如何获取本刊>>
教育研究
2012 · 6
教育研究
主办：中央教育科学研究所
出版：教育研究杂志编辑部
出版周期：月刊
出版地：北京市
阅读
订阅
目录
本期目录
零风险读研
不过退款!
跨考考研
WWW.KUAKAO.COM
关闭
电话：0851-4354777

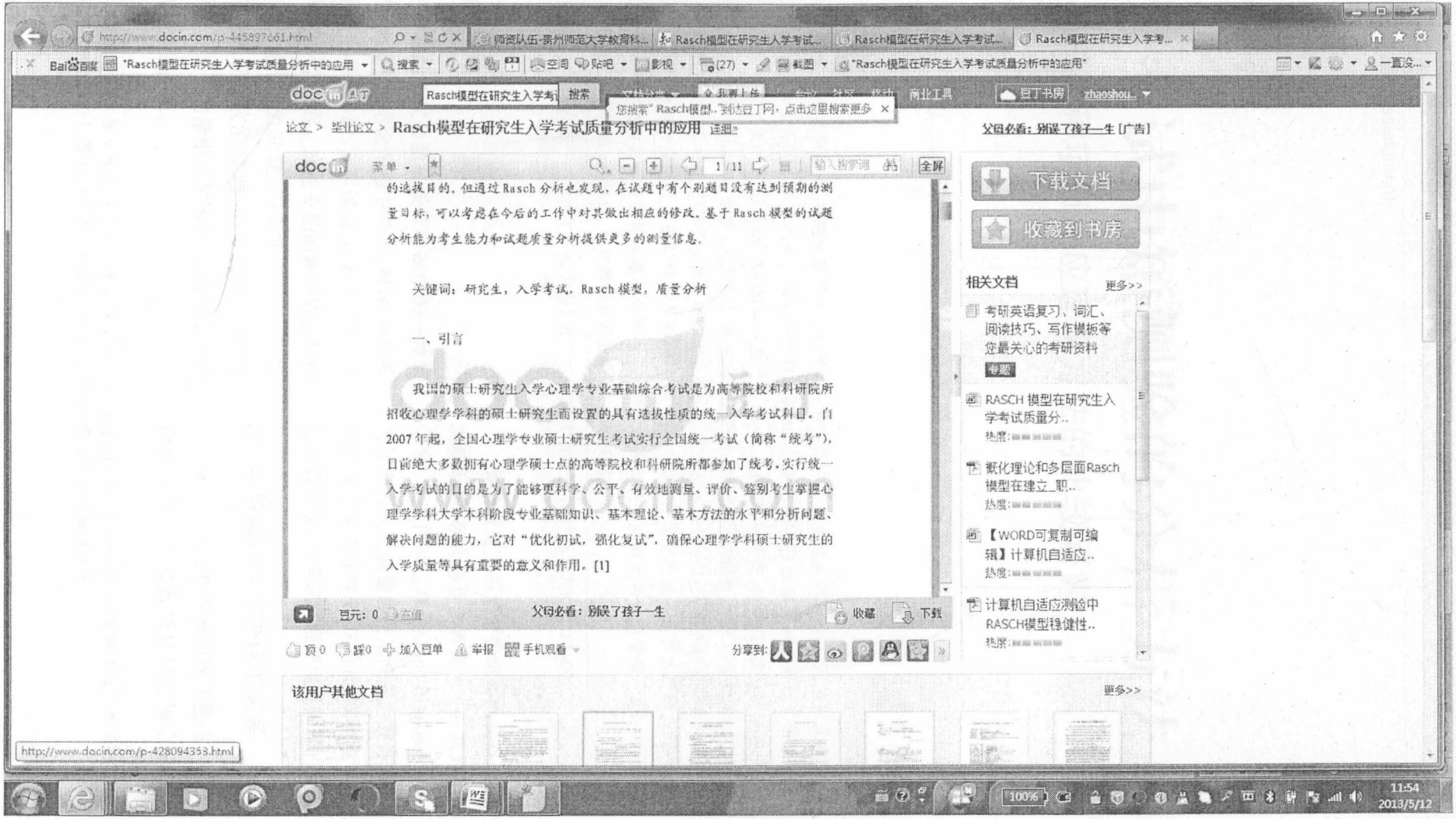
被维普期刊网转载


Rasch模型在研究生入学考试质量分析中的应用
《教育研究》2012年 第6期
下载全文
论文发表（点击进入）
摘 要：运用Rasch模型对2010年全国硕士研究生入学考试心理学专业基础综合考试进行分析。结果表明，该试题总体上是一套高质量的测验，试题的内容覆盖了所有能力水平的考生，且能够较好地区分考生的能力水平，达到了预期的选拔目的。但通过Rasch分析也发现，在试题中有个别题目没有达到预期的测量目标，可以考虑在今后的工作中对其做出相应的修改。基于Rasch模型的试题分析能为考生能力和试题质量分析提供更多的测量信息。
【分 类】 【文化、科学、教育、体育】 > 教育 > 教育学 > 教育与其他科学的关系、教育学分支 > 教育统计学
【关键词】 研究生 入学考试 Rasch模型 质量分析
【出 处】 《教育研究》2012年 第6期 61-65页 共5页
【收 录】 中文科技期刊数据库
基金申请 课题设计咨询
论文发表专栏
维普投稿平台，是国内最专业的期刊投稿、论文发表平台。提供期刊快速审稿，只需2-3个工作日，投稿效率高，发表周期短，100%成功...【详细】
最新征稿：
教育类期刊紧急征稿
经济管理期刊征稿
核心期刊杂志征稿少量名额
急征政治核心期刊稿件
社会学省级、国家级期刊正在收稿
同期文献
《教育研究》
Educational Research
2012年 第6期
期刊详细信息
以十八大精神为指导推进教育事业科学发展
2012中国教育研究前沿与热点问题年度报告
教育学研究的热点与重点——对2012年度人大复...
新中国成立以来教育基本理论的发展与贡献
道德教育的"精神形态"与"中国形态"

被道客巴巴网转载


道客巴巴 doc88.com
我要下载 我要上传 会员中心 帮助
你好，欢迎来到道客巴巴-在线文档分享平台 [请登录] [免费注册] 站内信箱
首页 浏览文档 共享文档 我要上传 搜索 高级搜索
所有文档>>学术论文>>期刊论文>>Rasch模型在研究生入学考试质量分析中的应用
教育研究
EDUCATIONAL RESEARCH
Rasch模型在研究生
入学考试质量分析中的应用
[摘 要] 运用Rasch模型对2010年全国硕士研究生入学考试心理学专业基础综合考试进行分析。结果表明，该试题总体上是一套高质量的测验，试题的内容涵盖了所有能力水平的考生，且能够较好地区分考生的能力水平，达到了预期的选拔目的。但通过Rasch分析也表明，在试题中有个别题目没有达到预期的测量目标，可以考虑在今后的工作中对其做出相应的修改。基于Rasch模型的试题分析能为考生能力和试题质量分析提供更多的测量信息。
[关键词] 研究生；入学考试；Rasch模型；质量分析
[作者简介] 赵守盈，贵州师范大学教育科学学院院长、教授、博士（贵阳 550001）
文档信息
文档上传人：忘川江月
文档格式：PDF
文档页数：5
下载积分：300
上传日期：2013-01-02 11:54:51
浏览次数：3
文档星级：★☆☆☆☆
举报此文档
如何下载?
下载 打印 收藏 分享
相关文档 更多>>
【推荐】文档
推荐股票 Word文档
【精品文档】中金推荐
【推荐文档】精益物流
女性推荐红酒文档
[精品文档]：英文(推荐)
推荐书目文档
(精选文档)经典小说推荐
[推荐]：歌舞鉴赏文档
文档简介
Rasch模型在研究生入学考试质量分析中的应用
文档日志
11:53
2013/5/12

被豆丁网转载


Rasch模型在研究生入学考试质量分析中的应用
论文 > 毕业论文 > Rasch模型在研究生入学考试质量分析中的应用
的选拔目的，但通过Rasch分析也发现，在试题中有个别题目没有达到预期的测量目标，可以考虑在今后的工作中对其做出相应的修改。基于Rasch模型的试题分析能为考生能力和试题质量分析提供更多的测量信息。
关键词：研究生，入学考试，Rasch模型，质量分析
一、引言
我国的硕士研究生入学心理学专业基础综合考试是为高等院校和科研院所招收心理学学科的硕士研究生而设置的具有选拔性质的统一入学考试科目。自2007年起，全国心理学专业硕士研究生考试实行全国统一考试（简称“统考”），目前绝大多数拥有心理学硕士点的高等院校和科研院所都参加了统考。实行统一入学考试的目的是为了能够更科学、公平、有效地测量、评价、鉴别考生掌握心理学学科大学本科阶段专业基础知识、基本理论、基本方法的水平和分析问题、解决问题的能力，它对“优化初试，强化复试”，确保心理学学科硕士研究生的入学质量等具有重要的意义和作用。[1]
下载文档
收藏到书房
相关文档
该用户其他文档

教育考试中短测验的分析方法
——基于两种项目反应理论方法的比较研究

摘要：教育考试中专题、短测验等形式是命题的一种主要方式。对这类测验的分析，可以从参数项目反应理论和非参数项目反应理论入手。本研究分别选取Rasch模型和Mokken模型对某高三文科综合地理试卷进行分析比较。使用Winsteps和Xcalibre软件进行Rasch分析，得到难度、信息量、项目功能差异等参数；使用MSP软件进行Mokken分析，得到正答率和同质性系数。比较两种结果，得出以下结论：(1)非参数项目反应理论以正答率对题目排序与参数项目反应理论以难度排序一致；(2)而有个别不符合参数项目反应理论标准的题目对提高测验质量同样有意义，不应被删除；(3)进行维度检验和题目筛选时，非参数项目反应理论标准比参数项目反应理论标准更加严格；(4)两种理论的项目功能差异检验结果一致。

关键词：教育测量；短测验；非参数项目反应理论；摩根模型

【中图分类号】G405　　【文献标识码】A　　【文章编号】1005-8427(2012)10-0018-7

1　前言

采用专题的方式命题是现代考试中的一种常见的方式，如高考英语全国卷中听力、阅读、写作等内容均可视为一个专题；文科综合试卷中，政治、历史、地理各成一个专题。这些专题的题目数量都很少，如英语试卷中听力20题、阅读20题、写作2题；文科综合选择题中政治12题、历史12题、地理11题。短测验在教育测量中非常流行，测验质量一直是命题者关注的问题。

命题质量关系到评价结果的客观公正，为确保测验质量，教育测量学者提出了一系列的方法。近年来，以项目反应理论(Item Response Theory, IRT)为代表的现代测量理论逐渐成为教育测量的主流，帮助教育者通过难度、区分度、猜测度、一致性等了解试卷质量，对考试实践产生了深远的影响。

2　两种项目反应理论的关系

按照项目分析时所用统计量的不同，项目反应理论可以分为参数项目反应理论(Parameter Item

本文为贵州省高等学校教学质量与教学改革工程重点项目"基于PBL理论改进心理教育测量教学改革研究"(项目批准号：黔高教发[2011]28-1)；贵州师范大学精品课程"心理测量"建设项目阶段性成果。

Response Theory, P-IRT)和非参数项目反应理论(Non-parameter Item Response Theory, NIRT)。P-IRT模型以区分度a、难度b、猜测度c、能力参数θ、信息量I等统计量为参数进行项目分析。常见的P-IRT模型有Rasch模型、Logistic模型、等级反应模型等。N-IRT使用正答概率的次序、哥特曼错误数、同质性系数H(coefficients of homogeneity)等指标进行项目分析。目前在教育测量中运用最为广泛的N-IRT模型是摩根模型(Mokken Model)[1]。

P-IRT多应用于大试题量、大样本的测验中。在处理短测验、小样本数据时P-IRT存在很大的误差,N-IRT理论可以弥补这一缺陷[2],帮助研究者全面了解测验的质量。两种理论都遵循IRT的基本假设:潜在特质单维、被试作答局部独立、项目特征曲线(item characteristic curve,ICC)单调递增。在摩根模型中,如果测验数据满足三个假设,就构成了单调同质模型(monotonely homogeneous model, MH)[3]。其ICC曲线类似于P-IRT中的Logistic模型:每个题目的ICC曲线都满足单调递增,但由于区分度不同,ICC曲线可能相交(图1-1)。如果数据拟合MH模型,说明被试能力与试题得分之间单调相关。P-IRT中用拟合指数等进行维度检验,拟合差则说明测量结果中可能受到了目标特质之外的其它因素的影响。如Rasch模型中的Outfit MNSQ和Infit MNSQ,理想值为1,越接近理想值拟合越好,测验过程没有受到潜在特质之外的因素影响[4]。

如果测验数据满足这三个假设且不同题目的ICC曲线不相交(N-IRT中称之为题目间单调)这就构成了摩根模型中的双重单调模型(doubly monotone model, DM)[3]。DM模型可以用来对试卷进行项目功能差异(differential item functioning ,DIF)检验。一份优秀的试卷要求试题难度排序具有不变性的特点。即对于同一群体的不同子群体(如考试中的男、女两个子群体),按照正答率对试题排序,排序结果应当一致。出现不一致的情况则表明不同子群体在同一题目上的正答率不同,这些题目可能存在DIF。P-IRT中也有许多方法进行DIF检验。体现在ICC曲线上,不同被试子群体的ICC曲线不重合(图1-3),曲线越不重合,DIF越严重。

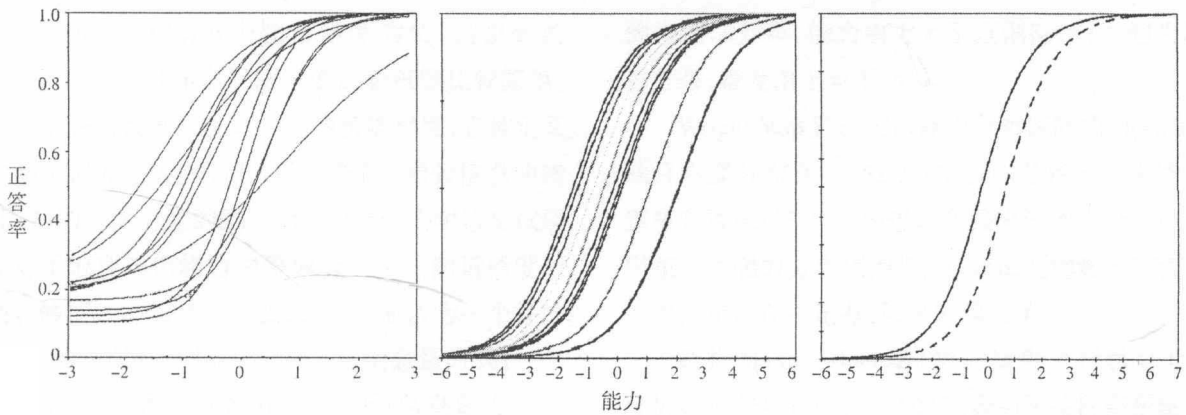
3 分析实例

3.1 研究样本及数据来源

本研究的样本是贵州省贵阳市某高三文科班学生。研究数据为贵阳市2011年一模文综考试的地理部分,共11个题目。样本量为194人,其中,男生71人,女生123人。

3.2.1 Rasch分析结果

Rasch模型是一种单参数模型,本研究选择该



1-1 MH模型　　1-2 DM模型　　1-3 项目功能差异

图1 ICC曲线

模型对数据进行P-IRT分析，并与N-IRT的分析结果进行比较。Rasch分析采用Winsteps软件。利用Rasch模型对试卷进行分析可以得到难度b、信息量I、拟合指数等参数(表1)。

表1 Rasch分析结果

题目	难度b	Infit MNSQ	Outfit MNSQ	M-H	p
1	-0.14	0.88	0.89	0.98	0.96
2	2.08	1.05	0.95	0.92	0.88
3	0.16	1.12	1.15	0.68	0.36
4	-1.13	0.93	0.83	0.84	0.73
5	-0.46	1.01	0.94	0.74	0.51
6	-0.82	0.93	0.85	0.79	0.64
7	-1.52	0.94	0.67	1.00	0.99
8	0.05	1.12	1.12	1.16	0.73
9	-1.36	1.02	0.97	0.46	0.10
10	2.04	1.18	1.47	1.68	0.36
11	1.11	0.9	0.83	1.60	0.37
平均数		1.00	0.97		

Rasch分析结果显示Infit MNSQ 均值为1.00、Outfit MNSQ均值为0.97。拟合指数等于或接近理想值1[5]，说明数据与模型拟合良好，测量过程没有受到目标特质之外的因素影响。整套试题测量的特质为地理知识能力。

一般认为试题的难度应在[-2,2]之间，难度太大(>2)或太小(<-2)的题目对潜在特质的测量效用不大[6]。这套试题中有两个题目(题目2、10)的难度大于2，超出上述标准。对剩余9个题目难度作进一步分析，发现有6个题目难度为负，占总数的66%。这说明对样本群体来说这套试题比较简单。测验信息函数表示能力估计的精确程度，它被定义为测量误差平方[SE $(\theta)^2$]的倒数[7]。测验信息曲线(图2)的峰值对应的能力值为-0.35，表明这套试题在对于地理知识能力中等偏低的学生测量精度最高。峰值处的信息量为2.125。一般认为一个好的试卷，测验误差应当在0.25以下，信息量为16；一个更好的试卷，测验误差在0.2以下，信息量为25[8]。地理试卷的测验信息未达到上述标准。测验信息

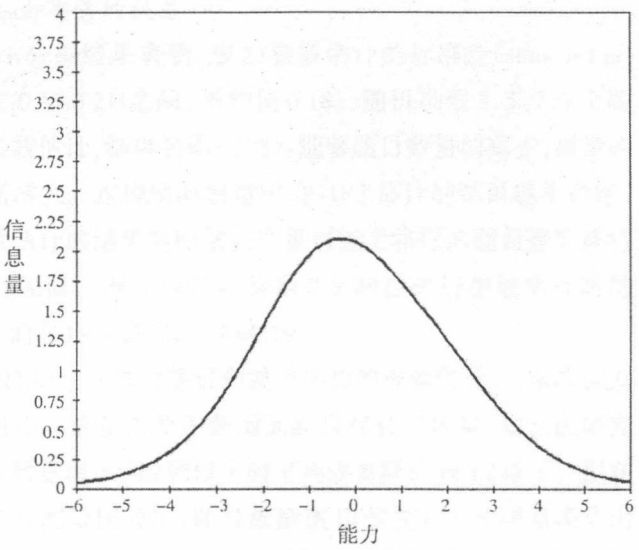


图2 Rasch分析信息曲线

量是全部题目信息量加总后得到的，整套试卷信息量太低可能与每个题目信息量太低、题目数量太少有直接关系。此次分析只涉及了客观题部分，但一套完整的试卷除此之外还有简答、论述、综合等主观题，应当结合整套试卷判断试题质量，而不是简单按照参数标准删除或修改题目。

Rasch模型还可以比较题目难度与被试能力的分布，常见的Rasch分析软件都以Wright Map的方式输出结果。Wright Map中通过对数转换，将被试能力和题目难度转换成同一单位——Logit，这样就可以在同一坐标系中比较被试和题目[9]。图3中左侧为被试分布，能力由上至下逐渐降低；右侧为题目分布，难度由上至下递减。

Wright Map显示被试能力分布区间约为[-2,3]，题目难度分布在[-1.52,2.08]；表明题目难度未能涵盖所有被试(图3)。理想的测验应该是测验项目集中在学生能力分布周围[10]。这套试题的难度与被试能力分布存在一定差异，试题偏简单。

选择Mantel-Haenszel法对试卷进行性别DIF分析。若题目p值小于0.05则表明该题目存在显著的DIF。结果表明(表1)整套试卷不存在性别DIF。

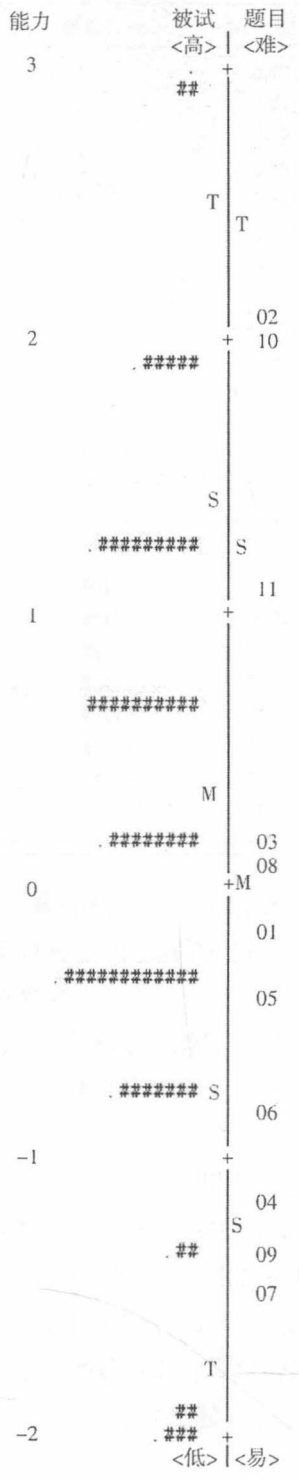


图3 题目难度与被试能力

3.2.2 Rasch测量的误差

Rasch分析结果表明(表2)参数估计的标准差(standard error,SE)在[0.17,0.21]之间,平均值0.18。随机抽取3、5、7、9个题目进行参数估计,结果表明(表2):随着题目数量的减少,测量的误差逐渐增大。这说明测验越短,P-IRT估计的结果越不稳定,按照参数估计的结果判断题目质量可能会将好的题目删除或将差的题目保留。例如题目4,抽取7个题目进行参数估计时的SE=0.21,抽取11个题目时SE=0.19。

P-IRT追求的是对题目和能力参数的准确估计,力求将误差降低到最小。这就需要不断增加题目和被试数量,如此次研究中将题目数量增加到9题以上时平均误差降低到0.2以下。但在实际的教育测量情境中,题目数量在10题左右的短测验经常出现。尤其是当题目与模型拟合较差时,参数估计的误差更大,结果更不稳定。

针对P-IRT的这一局限,有研究者提出了N-IRT模型作为补充[11]。Mokken模型是最具代表性的非参模型之一,它以同质性系数、正答率、哥特曼错误数等统计量进行项目分析。这些统计量(如正答率次序)不受题目数量的影响[12]。

3.3 Mokken分析结果

利用Mokken模型对试卷进行分析,常用的统计量称为同质性系数H。共有三种类型的同质性系数:题目i与题目j之间的同质性系数H_{ij}、题目i与剩余题目的同质性系数H_i、全部题目的同质性系数H。H值越高,测验总分对被试潜在特质的排序越准确,Mokken提出:H_{ij}应大于0,H_i和H至少为0.3。$0.3 \leq H < 0.4$表明试卷的测量准确程度较弱;$0.4 \leq H < 0.5$表明试卷测量准确程度中等;$0.5 \leq H \leq 1$时,试卷测量准确程度强;$H < 0.3$表明试卷不合格[13]。

Mokken分析主要从两个方面进行:对试卷的维度进行分析、筛选题目;DIF检验。本研究使用MSP5软件对数据进行Mokken分析。

Mokken模型利用同质性系数进行维度检验,当数据拟合MH模型,同时满足以下两个条件时,题目所测量的是同一特质。任意两个项目之间同质性系数$H_{ij} > 0$;特定题目与剩余题目间同质性系数$H_i > 0.3$[1]。分析结果显示(表3):11个题目中有7个题目达到上述标准,这些题目测量的是同一潜在特质,另外4

表2 随机筛选题目参数估计时的标准差

题目	11题	筛选9题	筛选7题	筛选5题	筛选3题
1	0.17		0.18		0.21
2	0.21	0.21	0.21	0.23	0.23
3	0.17	0.17		0.17	0.20
4	0.19		0.21		
5	0.17	0.18		0.18	
6	0.18		0.20		
7	0.21	0.21	0.22		
8	0.17	0.17		0.19	
9	0.20		0.21	0.23	
10	0.20	0.20	0.20		
11	0.17	0.17			
平均值	0.18	0.15	0.20	0.20	0.21

表3 Mokken分析结果

	题目	题目2	题目11	题目1	题目6	题目4	题目9	题目7
H_{ij}	题目2							
	题目11	0.09						
	题目1	0.11	0.43					
	题目6	0.57	0.36	0.40				
	题目4	0.69	0.60	0.36	0.36			
	题目9	0.31	0.48	0.39	0.25	0.24		
	题目7	0.50	0.51	0.54	0.42	0.33	0.32	
	H_i	0.31	0.39	0.39	0.37	0.38	0.32	0.41

个题目(题目3、5、8、10)测量的可能不是地理能力，或测量过程受到了其它因素影响。这些题目应当删除或改进。

筛选后的整套试卷同质性系数H= 0.37，表明利用试卷对学生的地理能力进行测量，准确程度接近中等。

如果数据拟合DM模型，就可以通过比较不同子群体题目正答率次序进行DIF检验。性别DIF检验结果显示(表4)，男生组数据中有7个题目与DM模型拟合，女生组有5个题目拟合。以正答率为指标分别对这些题目进行排序，男女生两组的排序结果相同，且正答率非常接近；题目不存在性别上的差异。

3.4 两种分析方法的比较

将P-IRT中的题目按照难度值由高到低排列，并与N-IRT中按照正答率由低到高排列的结果比较。结果证明二者是等效的，即难度越高的题目，正答率越低。这一结果与其他研究者的结论一致[14]。

表4 项目功能差异结果

题目	男生组		女生组	
	正答率P	H_i	正答率P	H_i
题目2	0.17	0.54		
题目11	0.37	0.45	0.31	0.49
题目1	0.54	0.40	0.58	0.43
题目6	0.63	0.44	0.71	0.36
题目9	0.66	0.43		
题目4	0.69	0.47	0.75	0.32
题目7	0.76	0.56	0.79	0.38

在Rasch模型中依据难度筛选题目，结果有9个题目达到统计学要求，2个题目需要改进或删

除。在Mokken模型中，依据同质性系数筛选题目，结果有7个题目达到统计学要求，4个题目需要改进或删除。对比两个结果，在题目筛选上Mokken模型比Rasch模型更加严格。

需要特别指出的是对题目2和题目10的分析结果。在Rasch模型下，两个题目的难度相当，分别为2.08、2.04，均稍高于Rasch标准。在Mokken模型下，项目2的同质性系数H_i=0.54、P=0.54，是符合Mokken标准中难度最大的题目。这表明在P-IRT中参数不合格或处在合格与不合格分界处的题目对潜在特质的测量仍然有用。这有可能是Rasch在分析短测验试卷时的不稳定性造成的。另有研究者指出这类题目反映出被试的得分概率与潜在特质之间的关系可能不是Logistic类型，而是简单的非递减函数关系。这些题目对提高测验的质量意义也很大，以往研究中，简单的按照参数标准删除或修改的做法有待商榷。对于超出Rasch标准很高的题目是否也会出现类似现象，由于此次研究中未出现这类题目，这一问题需要在之后的研究中做进一步的讨论。

从维度检验的结果来看，Rasch模型下单维的数据，在Mokken模型下不一定单维，这表明后者对数据的要求更为严格。在重要考试中，可以采用Mokken模型进行维度检验，保证试题质量。

两种理论的DIF检验结果一致。Mokken模型下进行DIF检验的过程要将各分组的数据分别进行处理，结果发现有些题目在整体处理中能与模型拟合，分组后却并不一定能与模型拟合。这类信息是Rasch分析无法得到的，但却对命题非常重要。

P-IRT分析可以估计出准确的题目参数，依照相应的标准评价题目和试卷；N-IRT分析只能得出正答率及其次序、同质性系数、哥特曼错误数。项目分析时使用最多的方法是排序。排序方式没有参数标准精确，但对P-IRT是个重要的补充。

P-IRT更适用于大规模的测验，题量越大、被试越多，参数估计结果越准确、稳定。而在教育测量中常会遇到由少数题目组成的短测验或被试数量很少的情况。尤其是高考等重要考试中，以专题、短测验形式命题已经成为命题的主流。N-IRT为这类测验的分析提供了思路，可以为测验的准确性和公平性提供重要参考。它在题量小、被试少的测验分析上所表现出的优势备受研究者青睐[15]。N-IRT对计算机自适应测验的题库建设也具有深远的意义。在题库建设阶段施测的被试越少，越能有效地降低题目的曝光率。

4 结语

两种理论比较体现出来的差异，其原因是多方面的。既有理论本身的原因，也有所运用软件的原因。Rasch分析的结果比较丰富与其软件更为成熟也有一定的关系。目前基于P-IRT计算机软件不论是数量还是商业化程度都远超基于N-IRT的软件。相关软件的开发也将会是N-IRT理论发展的一个突破口。

近年来N-IRT的研究取得了长足的进步，但受其项目分析结果不够精确等特点的限制，研究者在实际应用中将其多作为P-IRT的补充。随着模型和算法的不断完善，它将逐渐减少人们对P-IRT的依赖。两种理论互补，共同提高教育测量的质量。

参考文献

[1] 张军. 非参数项目反应理论在维度分析中的运用及评价[J]. 心理学探新，2010(3): 80-83.

[2] 辛涛. 项目反应理论研究的新进展[J]. 中国考试，2005(7): 18-21.

[3] Van Schuur W H. Mokken scale analysis: between the Guttman scale and parametric item response theory[J]. Political Analysis. 2003, 11(2): 139-163.

[4] 晏子. 心理科学领域内的客观测量——Rasch模型之特点及发展趋势[J]. 心理科学进展，2010(8): 1298-1305.

[5] Smith Jr E V, Others. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.[J]. Journal of applied measurement. 2002, 3 (2): 205.

[6] 余嘉元. 项目反应理论及其应用[M]. 南京: 江苏教育出版社. 1992.
[7] 杨建原,柏桧,赵守盈. 计算机自适应测验开发的程序研究[J]. 中国考试, 2012(3): 3-7.
[8] 涂冬波,蔡艳. 信息函数在标准参照测验中的应用研究[J]. 江西师范大学学报(自然科学版),2005, 29(2): 167-172.
[9] 赵守盈,何妃霞,陈维,等. Rasch 模型在研究生入学考试质量分析中的应用[J]. 教育研究, 2012(6): 61-65.
[10] 张金勇,何妃霞. 教育测试中学生能力水平与测验项目难度的 Rasch模型分析——个体能力与题目难度之间的对应关系[J]. 当代教育科学,2012(12): 11-14.
[11] 刘欣,徐海波. 国外非参数项目反应理论的回顾与展望[J]. 统计教育,2002(1): 43-44.
[12] Engelhard Jr G. Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken[J]. Measurement. 2008, 6(3): 155-189.
[13] Mokken R J. A theory and procedure of scale analysis[M]. Mouton The Hague, 1971.
[14] 雷新勇. 非参数项目反应理论模型及其在教育考试中的应用[J]. 考试研究,2006(3): 53-71.
[15] Junker B W, Sijtsma K. Nonparametric item response theory in action: An overview of the special issue[J]. Applied Psychological Measurement. 2001, 25(3): 211-220.

作者单位

Item Analysis of Short Test in Educational Testing: Comparative Study on Parameter and Non-parameter Item Response Theory

HE Zhuang, YUAN Shuli and ZHAO Shouying

Abstract: As one of the significant types of tests, the test project and short test are popular in educational testing. Parameter and non-parameter item response theory being the starts, these tests were under analysis. Compared was the geography paper in inaugurated arts taken by some senior three students. During this comparison the Rasch and Mokken model were respectively selected. For analyzing software Winsteps and Xcalibre were utilized to analyze item parameters in Rasch model. Analyzed in detail were the parameters of difficulty, differential item functioning and information curve. Software MSP was for the purpose of analyzing items in Mokken model. Besides, the statistics of accurate rate and coefficients of homogeneity were also analyzed in detail. Finally, four conclusions were arrived at as the following: (1)The estimate results of difficulty between non-parameter and parameter item response theory were equivalent. (2)Those items, which failed to fit parameter item response theory, succeeded in non-parameter item response theory. (3)Non-parameter item response theory is more rigorous than parameter item response theory in dimension testing and item screening. (4) The result was equivalent in the detection for differential item functioning between non-parameter and parameter item response theory.

Keywords: Educational Measure; Short Yest; Non-parameter Item Response Theory; Mokken Model