Pure Mathematics 理论数学, 2024, 14(11), 280-288
Published Online November 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.1411395

Hans汉斯

基于Rasch模型的高中数学测验命题特点分析

姜　容[1]，杨雪芹[2]，刘元勋[1]，何　壮[3]

[1]贵阳市第三实验中学，贵州　贵阳
[2]贵阳学院材料科学与工程学院，贵州　贵阳
[3]贵阳学院教育科学学院，贵州　贵阳

收稿日期：2024年9月20日；录用日期：2024年10月22日；发布日期：2024年11月29日

摘　要

Rasch分析具有参数客观性、等距性和被试独立性等特征在试卷特点和品质分析中独具特色，与经典测量理论互补，为教育考试数据分析提供了多元的视角。本研究以高中数学测验为例，从测量目标的单维性、试卷的拟合、题目的拟合、学生能力与题目难度的怀特图、题目参数的气泡图等内容开展了分析，并进一步探讨了结果应用中需要关注的问题。

关键词

Rasch模型，命题特点，高中数学

Characteristics Analysis of High School Mathematics Test Based on Rasch Model

Rong Jiang[1], Xueqin Yang[2], Yuanxun Liu[1], Zhuang He[3]

[1]Guiyang No.3 Experimental Middle School, Guiyang Guizhou
[2]College of Materials Science and Engineering, Guiyang University, Guiyang Guizhou
[3]School of Education and Science, Guiyang University, Guiyang Guizhou

Received: Sep. 20th, 2024; accepted: Oct. 22nd, 2024; published: Nov. 29th, 2024

Abstract

Rasch analysis has the characteristics of parameter objectivity, isometry and subject independence in the characteristics and quality analysis of the examination paper. It is complementary to the

文章引用: 姜容, 杨雪芹, 刘元勋, 何壮. 基于 Rasch 模型的高中数学测验命题特点分析[J]. 理论数学, 2024, 14(11): 280-288. DOI: 10.12677/pm.2024.1411395

classical measurement theory, and provides a diversified perspective for the educational examination data analysis. Taking the high school mathematics test as an example, this paper analyzed the unidimensionality of the measurement target, the fitting of the test paper, the fitting of the questions, the white diagram of the students ability and the topic difficulty, and the bubble map of the topic parameters, and further discussed the problems to be paid attention to in the application of the results.

Keywords

Rasch Model, Proposition Feature, High School Mathematics

1. 引言

命题特点和品质是科学评价教学效果、学生能力的首要前提。随着教育测量与评价领域最新技术和理念的不断普及，命题特点和品质分析逐渐成为了教育考试数据分析中的重要内容。当前，在命题特点和品质分析领域主流的理论基础有经典测量理论和项目反应理论，两种理论互为补充，各有优势。经典测量理论发展较早，在教育评价领域发挥了重要作用。随着教育测量学技术的发展，经典测量理论的不足也逐步显现，例如在难度计算过程中，以正确率、得分率作为难度固然容易操作，但却受到抽样群体的影响，即参数估计过程中存在“被试依赖性”；以总分作为学生能力参数，它是将学生在整套试卷上所有题目的得分相加而来，以得分率作为题目难度参数，则是以所有学生在该题目上的作答情况统计而来，两个参数的估计过程相互独立，题目难度和学生能力不能直接比较，影响了对数据的进一步挖掘分析。上世纪中叶，为解决上述难题，测量学家提出了项目反应理论，以 Rasch 模型为代表的项目反应理论以其样本独立性、难度客观性、等距性等特点弥补了经典测量理论的不足[1]。Georg Rasch 于 1960 年提出 Rasch 模型，Andrich 于 1978 年将其拓展为 Rasch 等级模型(Rasch-Andrich Rating Scale Model, RSM)。RSM 可以用于处理多值计分数据，对数据的要求比较严格，如题目间的等级间距(Thresholds)要相同、梯难度和选项难度需要满足递增趋势。RSM 分析适用于心理测量领域常用的李克特量表。1982 年 Masters 将 RSM 拓展为 Rasch 分部计分模型(Rasch-Masters Partial Credit Model, PCM) (Masters, 1982)。PCM 题目的选项可以随机排列，各题目可以有一套单独的 Thresholds。PCM 更适用于教育测量中的多项选择题和主观题。

本研究以高中数学测验为例，使用 Rasch 模型对命题的特点和品质做详细分析，探讨从不同视角评价命题的技术及结论应用。

2. 研究方法

2.1. 高中数学试卷

本研究使用的高中数学试卷共计 28 题，其中客观题(选择题、填空题) 16 题，主观题 12 题。客观题以 O01~O16 表示，主观题以 S1701~S2202 表示，其中，S1701 表示第 17 题第 1 问，以此类推。

2.2. 样本选择

以某高中三年级全体学生为样本，本次考试，共获得有效数据 785 人。

2.3. 数据分析方法

数据管理使用 Excel，数据分析软件为 Winsteps 5.4.3。该软件为 Rasch 模型专用的数据分析软件，本次数据分析选择了分部计分模型(Partial Credit Model, PCM)，在软件的参数估计设定中，将客观题得分均转换为 0-1 计分、主观题得分转换为等级计分，题目的平均难度设定为 0，主要数据分析内容包括：单维性、试卷拟合、题目拟合、怀特图、气泡图等内容。

3. 数据分析结果

3.1. 单维性检验

单维性检验分析针对命题的总体品质，主要用于评价命题是否紧扣学科主题，单维性特征较好，说明学生在考试过程中有且仅有数学能力这一种潜在特质影响作答表现。

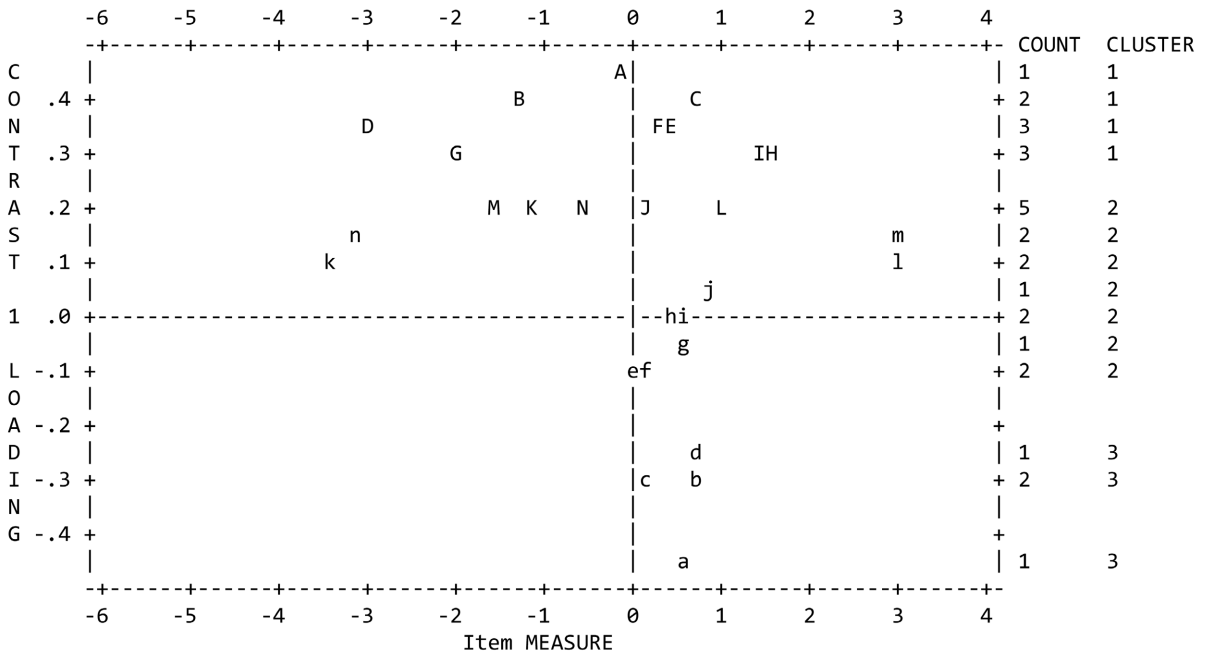

Figure 1. Unidimensionality test
图 1. 单维性检验

本次测验的单维性检验结果如图 1 所示，横坐标代表题目的难度级别，纵坐标代表题目得分和可能影响因素(潜在特质)之间的相关关系值，字母 A、B、C、D 和 a、b、c、d 等分别代表一道题目。可以发现，大部分题目的相关关系值都集中在[−0.4, +0.4]之间，符合 Rasch 模型相关理论中对单维性的要求。但 A、a 代表的 2 个题目的相关关系值超出[−0.4, +0.4]的建议范围，不符合单维性检验的要求，表明 O07、S1902 这 2 道题目不只受单一因素的影响，可能存在其它潜在特质影响了学生作答。

Rasch 模型中还有一种基于方差的单维性检验方法。结果如图 2 所示。I 为由题目解释的方差，1 为未解释方差主成分分析中第 1 主成分，如果 I > 1，则说明题目解释的方差高于残差中第一主成分解释的方差，也就是说数据中即使存在第二个能力维度对测量产生影响，也不影响测量结果。图 2 数据表明，I 为 23.5%，1 为 2.6%，I 远高于 1，说明 A、a 虽然受到多个潜在特质的影响，但并未影响整体测试结果。简而言之，试题符合 Rasch 模型的单维性要求，能够进行后续分析。

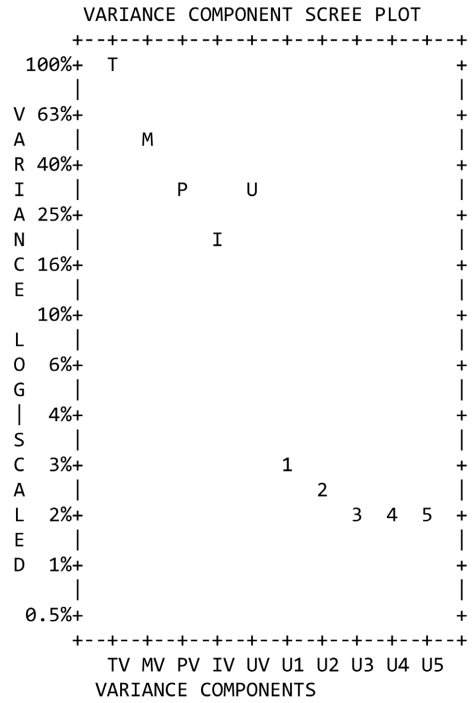

Figure 2. Principal component analysis
图 2. 主成分分析

3.2. 试卷拟合

对 785 名学生的数据进行整体拟合检验，结果如表 1 所示。

Table 1. Quality inspection
表 1. 整体质量检测

	数量	INFIT		OUTFIT		Speration	Reliability
		MNSQ	ZSTD	MNSQ	ZSTD		
被试(Student)	785	1.07	0.1	1.04	0.1	2.43	0.85
项目(Item)	28	1.02	0.1	1.07	−0.1	14.31	1.0

Rasch 模型还提供了区分度指标，其含义与经典测量理论的区分度相似。区分度是指测验题目能够在多大程度上区分所要测量的心理品质，区分度值越高，表明测验题目能够越好地将不同能力水平的被试区分开来。Rasch 模型中反应区分度的指标称为 Speration，相关理论建议该值应至少大于 2 [2]。Speration 区分度代表该组试题能够将学生分成多少个能力不同的群组。本评测试卷项目区分度为 14.31，取值较高。与之对应，学生区分度为 2.43，取值也达到了相关理论建议的水平。从试卷和学生的两个角度，数据均表明试卷能够区分不同能力水平的学生。

与 Speration 配套的是 Rasch 信度，用于评价测验结果的内部一致性，即试卷中的题目是否考查同一种能力，信度系数越高表明测验的结果越一致、稳定和可靠，相关理论建议的理想值为 1，大于 0.7 表明测信度较高[3]。本评测试卷题目整体信度为 1.0，表明题目的整体信度较高；学生整体信度为 0.85，取值较高。试题和学生两个视角的统计结果均表明试题整体信度较好。

Outfit MNSQ 为未加权均方拟合统计量，Infit MNSQ 为加权均方拟合统计量，ZSTD 是 MNSQ 的标

准化形式。这些都是 Rasch 模型中常用的拟合指标，该项用于评价原始数据与模型预测数据之间的一致性，一致性越高，代表原始数据越符合模型假设，数据与模型的“拟合”越好。Outfit MNSQ 对异常数据敏感，Infit MNSQ 对题目难度与学生能力数据敏感。Rasch 分析结果显示 Infit MNSQ 均值为 1.02、Outfit MNSQ 均值为 1.07，拟合指数等于或接近理想值 1，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。ZSTD 的取值越接近 0 越好[4]，分析结果显示被试和题目的该项拟合指数均值在[−0.1, 0.1]之间，接近理想值。可以看出，本次试题项目整体拟合度较好，符合 Rasch 模型理论要求。

3.3. 题目拟合

本次测验，试卷的 28 个题目拟合统计信息如表 2 所示。

Table 2. Item fit
表 2. 题目拟合

题目	难度	标准误	Infit MNSQ	OutfitMNSQ	相关系数
O01	−1.522	0.121	0.99	0.96	0.18
O02	−3.122	0.246	0.98	0.79	0.29
O03	−5.10	0.086	0.97	0.95	0.23
O04	−3.063	0.239	0.99	0.77	0.28
O05	−2.016	0.148	0.97	0.87	0.27
O06	−1.333	0.112	0.96	0.86	0.30
O07	−0.135	0.079	0.93	0.91	0.33
O08	−0.411	0.073	0.98	0.98	0.18
O09	−0.379	0.073	0.99	0.98	0.18
O10	0.037	0.077	1.02	1.03	0.07
O11	−0.137	0.075	0.98	0.97	0.20
O12	−3.369	0.280	0.99	0.86	0.24
O13	−1.103	0.103	0.96	0.91	0.27
O14	−0.982	0.073	0.96	0.95	0.25
O15	3.026	0.123	0.99	0.93	0.13
O16	2.997	0.121	0.99	0.97	0.12
S1701	0.301	0.014	0.84	0.67	0.56
S1702	0.694	0.011	0.95	1.02	0.53
S1801	−0.005	0.023	1.32	3.82	0.39
S1802	0.587	0.013	1.12	1.13	0.48
S1901	0.190	0.017	1.15	1.04	0.48
S1902	0.618	0.010	1.09	1.17	0.54
S2001	0.782	0.011	1.22	1.38	0.45
S2002	0.760	0.011	1.04	1.08	0.51
S2101	0.521	0.014	1.03	1.02	0.53
S2102	1.393	0.018	0.87	0.79	0.47
S2201	0.859	0.012	1.18	1.18	0.46
S2202	1.608	0.025	0.97	0.95	0.36

有研究者建议，Infit MNSQ 和 Outfit MNSQ 的取值应当在[0.8, 1.2]之间，较为宽松的标准认为取值应当在[0.5, 1.5]之间[5]。作为高风险考试，应当选择较为严格的拟合标准[6]。统计数据表明，各题目 Infit

MNSQ 的取值范围在[0.84, 1.32]之间，大多数题目都在可接受的范围内，表明数据与模型拟合较好。但S2001、S1801 参数值为 1.22、1.32 (>1.20)，说明学生在回答这个题目时，可能有低能力水平的学生正确回答了题目，而部分高能力水平的学生错误地回答了这个题目。Outfit MNSQ 的取值范围[0.67, 3.82]之间，其中 O02、O04、S1701、S1801、S2001 和 S2102 参数值分别为 0.79、0.77、0.67、3.82、1.38 和 0.79，均不同程度地偏离了正常范围。除 S1801、S2001 两题外其它四道题的 Infit MNSQ 的取值范围在[0.80, 1.20]范围内，综合两个拟合参数，可以认为 O02、O04、S1701 和 S2102 四个题目在可接受范围内，S1801、S2001 题确实受到其它因素干扰。

相关系数(CORR.)代表试题与试题测量目标的拟合程度，有研究者建议该相关系数的最低可接受水平为 0.03，相关系数越高，说明题目与试题的测量目标越接近。可以看出，所有测验项目的相关系数都是正向的，说明测验项目与测验目标一致，测量相同的潜在特质。且主观题因测量内容丰富，其相关系数总体较高，客观题第 10 题的相关系数在整套试卷中最低，仅为 0.07。

3.4. 怀特图

怀特图(Wright Map)也称为学生——题目图，能够利用 Rasch 量尺的特性，直观地展现项目难度与被试能力、被试与被试、项目与项目之间的关系。本次测验的怀特图如图 3 所示。

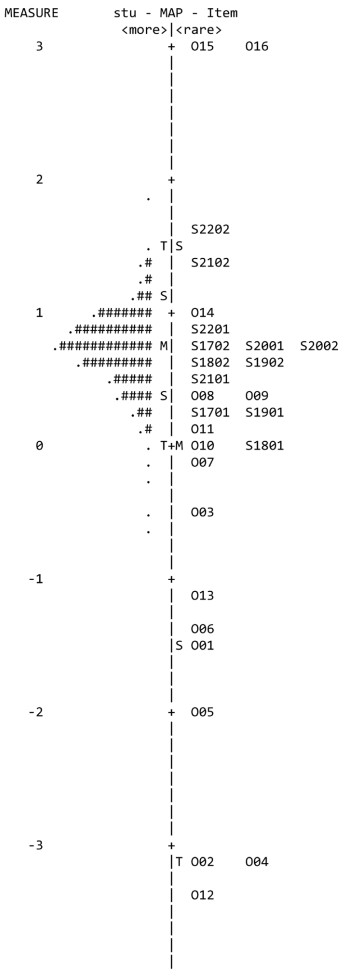

Figure 3. Wright map
图 3. 怀特图

图 3 中线是 Rasch 量尺，量尺上的刻度为 Logit，是试卷难度与学生能力水平对比的重要媒介，中线左侧代表学生能力水平的分布情况，中线右侧代表试题难度的分布情况。M 为 Mean 的缩写，分别代表学生能力和题目难度的平均水平；S 是 One Standard Error 的缩写，含义为距离均值的一个标准差；T 是 Two Standard Error 的缩写，含义为距离均值的两个标准差，刻度尺由上往下，对应学生的能力水平逐步降低，对应项目的难度水平也逐步减小；被试与被试间的间隔代表了彼此间能力的差异，项目与项目的间隔代表彼此间难度的差异，距离越近，差异越小。

图左侧每一个“#”代表 13 个被试，每一个“·”代表 1 至 12 个被试。学生的能力水平处于[−0.6, 1.8]之间，能力分布范围为 2.4 Logit，平均值为 0.772 Logit。与之相对应，题目难度平均值被设定为 0，二者之差为 0.772 Logit，说明该评测试卷对于被试来说整体难度偏低。

图右侧为题目，从分布位置上开看，题目 O16 难度最大，为 3.016 Logit，O12 难度最小，为−3.396 Logit，题目难度的分布范围为 6.3 Logit。

学生能力均值较试题难度均值高 0.772 Logit，由此可以看出学生的能力水平相对高于测验项目的难度水平。也就是说，其项目难度设计与学生的实际水平之间不太吻合，难度偏低。尤其是第 1、2、3、4、5、6、12、13 题，均处在较为简单的区域，且这一区域并没有学生分布。这表明本次考试，存在一定数量的简单题。与之相对应的，测验中较难的题目也偏少，对于能力大于 1 的学生，仅有第 21 题第 2 问和第 22 题第 2 问两个题目难度相对应。

总体来看容易的试题偏多，试题之间的难度水平差距较大，难度中等的试题分布较为集中，多数题目分布在[0, 1]之间。与之相对应，学生的能力分布却极为集中，主要分布在[0, 1.5]之间，题目难度和学生能力分布匹配度不佳，且有 11 个题目分布在上述区间之外，太多的题目难度与学生能力分布的主体不对应，未能更好地发挥评价作用，不利于对不同能力水平的学生做出很好的区分。

3.5. 气泡图

气泡图是用来综合评价拟合、测量误差的图形，绘图简单且结果直观，因此受到相关研究者的青睐。本次测验各题目的气泡图如图 4 所示。

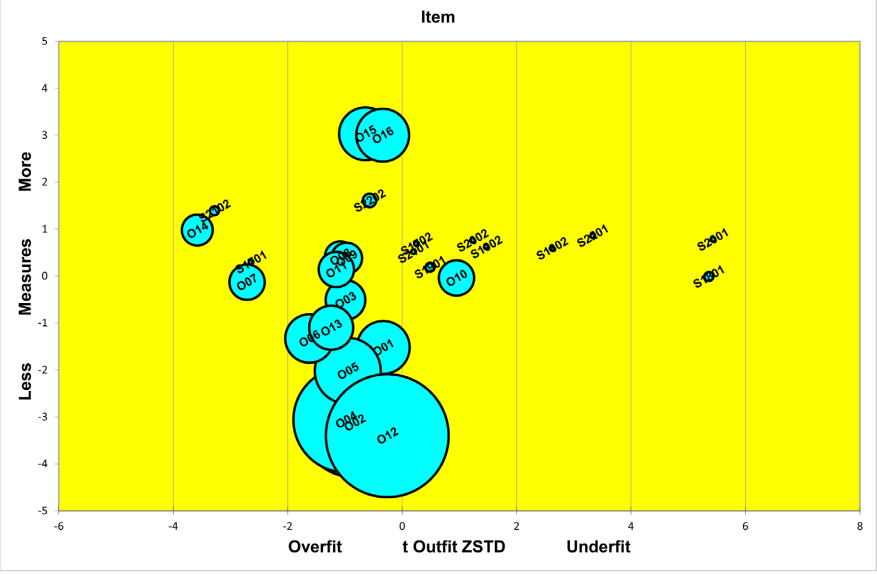

Figure 4. Bubble chart
图 4. 气泡图

图 4 中用气泡代表每个题目，气泡的大小代表 Rasch 标准误的大小。气泡越小，说明该测验对学生能力水平的估计越精确。纵轴为 Measure 气泡纵轴分布的位置代表试题难度参数，试题越靠近顶端，难度水平越大。在理想的情况下，项目应靠近气泡图的中轴线。从图中可以看出，多数气泡分布在[0, 1]区间附近，表明这些题目的难度非常接近。纵轴为 Outfit MNSQ ZSTD，代表题目的拟合，接近 0 则代表拟合较好，靠左表示过渡拟合、靠右代表不拟合，结合气泡图进行分析，试卷中大部分试题落在了可接受区间[−2, 2]内，但 O14、O07、S1701、S2102 为过拟合，代表这些项目与模型过度拟合；S1801、S1802、S2001、S2201 为不拟合，存在“能力高的学生做错低难度题目”“能力低的学生答对高难度题目”的情况；O02、O04 和 O12 气泡较大，表明其难度估计的误差较大，其对应测量结果的精准性较小。

4. 数据分析结论

4.1. 试卷命题特点

整套试卷的难度相对学生群体偏低，有 9 个客观题累计 45 分的题目对于参加本次考试的学生群体来说难度较低。同时，在高水平学生对应的难度区间上，仅有 2 个主观题累计 11 分的题目与学生能力相匹配。整套试卷的命题特点以简单题目为主，高难度题目极少。

4.2. 试卷整体命题质量

受难度分步的影响，因低难度题目较多，在评价中发挥的作用较小，同时，高难度题目较少，对高水平学生的区分度不高。因此，整套试卷的区分度不高，虽然达到了相关理论建议的水平，但若作为选拔性测试(常模参照测验)，其区分度较难达到评价目的的要求。若作为标准参照测验，题目的难度以课程标准要求为依据，测验结果则起到了评价教学效果的作用，题目难度和区分布可不做要求。

4.3. 试题命题质量

单维性检验结果表明，部分题目单维性检验结果较差，说明该题目的作答过程中，受到了除数学能力之外的潜在特质的影响，如猜测、阅读理解等。涉及的题目包括了主观题和客观题各 1 个，累计 11 分。另有两个主观题目的拟合指数较差，累计 12 分，1 个客观题的相关系数较低，累计 5 分。这些题目存在较大的测量误差，或其测量内容与测量的主要目标——数学能力存在较大的差异，需要根据题目内容和学生作答表现做进一步分析。

5. 讨论和结果应用建议

Rasch 模型的分析仅针对数据特点，参数的优劣并不代表命题质量的绝对好坏。对命题特点的分析还应该结合题目内容、测验目的等做综合判断。如标准参照测验以考察学生知识掌握情况为目的，以课程标准的难度要求为依据，对测验和题目的难度及其分布并无固定要求，仅需要关注单维性、拟合等指标。但对于常模参照测验，其主要目的是区分不同能力水平学生，因此还应当特别关注难度和能力的分布关系、区分度等指标。试卷和题目评价的结果不是独立于应用存在的，任何结论都应当结合评价的目的、命题设计等要素出发，在实际的应用工作中，需要提防只关注拟合指标的唯参数论。

Rasch 模型分析为命题特点和质量分析提供了新的视角，在实际教学工作中需要和经典测量理论协同使用，才有可能产生更高的价值。

基金项目

本研究为 2022 年贵阳市教育科学规划重点课题(编号 GYZD22019)；贵州省教育科学规划课题《五育并举视域下高中生增值评价实践研究》(编号 2022B212)阶段性成果；贵州省高等学校教学内容和课程体

系改革项目：“会测善用”导向的师范生教育测量与评价课程设置与建设实践(2023250)。

参考文献

[1] 李静璇, 王秋红, 何壮, 袁淑莉. Rasch 模型在初等教育阶段试卷质量分析领域的应用[J]. 贵阳学院学报(社会科学版), 2022, 17(3): 87-92.

[2] 柏毅, 朱文琴, 陈慧珍. Rasch 模型在试卷质量分析中的应用——以小学科学六年级技术与工程素养评测试卷为例[J]. 教育测量与评价, 2019(1): 25-31.

[3] 肖月, 桑芝芳. Rasch 模型在物理学业质量评价中的应用研究[J]. 物理通报, 2021(6): 119-123.

[4] 何壮, 赵守盈. 技能评分项目裁判员评分结果的多面 Rasch 模型分析——项目反应理论在体育运动领域的应用[J]. 成都体育学院学报, 2014, 40(3): 43-48.

[5] 何壮, 袁淑莉, 赵守盈. 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[J]. 中国考试, 2012(10): 18-24.

[6] 何壮, 袁淑莉, 余水, 任敏. 心理测量在高风险考试分析中的应用[J]. 贵阳学院学报(社会科学版), 2020, 15(2): 114-118.