

基于Rasch模型的教育考试主观题数据分析

何 壮, 刘天容, 袁淑莉

贵阳学院教育科学学院, 贵州 贵阳

收稿日期: 2024年9月3日; 录用日期: 2024年10月5日; 发布日期: 2024年10月12日

摘 要

本研究运用Rasch模型对某次心理学课程考试的8个主观题进行了分析。结果显示数据与模型拟合达到Rasch理论要求。分析结果包括题目阈值参数、考生能力、信息量等。结果表明主观题的8个题目是一组高质量的题目。分析结果为今后进一步修订和完善评价方案、改进命题、提高学业诊断质量提供相关的测量学参考依据。

关键词

Rasch模型, 主观题, 项目分析

Data Analysis of Subjective Questions in Educational Examinations Based on the Rasch Model

Zhuang He, Tianrong Liu, Shuli Yuan

College of Education Science, Guiyang University, Guiyang Guizhou

Received: Sep. 3rd, 2024; accepted: Oct. 5th, 2024; published: Oct. 12th, 2024

Abstract

This study used the Rasch model to analyze 8 subjective items from a psychology course exam. The results show that the data fit the model according to the requirements of Rasch theory. The analysis results include threshold parameters, students' abilities and test information. The results show that 8 subjective questions form a high-quality set of questions. The analysis results provide relevant measurement references for further revision and improvement of the evaluation scheme, improvement of test items, and enhancement of the quality of academic assessment.

Keywords

Rasch Model, Subjective Item, Item Analysis

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来以项目反应理论(Item Response Theory, IRT)为代表的教育测量理论逐渐成为教育测量研究的主流,与经典测量理论(Classical Test Theory, CTT)互相补充,帮助教师更加全面地了解试题特征和学生知识掌握情况。IRT 理论研究取得的进展对考试的实践产生了深远的影响。与 CTT 相比,IRT 的优点在于项目参数、考生能力等估计不受考生样本影响。

Rasch 模型是最常见的 IRT 模型之一,在心理学、社会学、管理学领域,基于该模型对李克特量表题目、0~1 计分题目的分析研究已经非常成熟。但在教育考试领域,尤其是针对主观题的分析仍不多见。与客观题相比主观题考察范围更广,考察结果更具外部效度[1]。因此对主观题的项目分析非常有必要。因此,本研究拟通过对 Rasch 模型的回顾和实例分析,系统介绍该模型在教育考试主观题分析中的应用。

2. Rasch 模型的基本理论及其发展

2.1. Rasch 模型

丹麦数学家 Georg Rasch 于 1960 年提出 Rasch 模型,该模型是一种用于测量被试潜在特质的概率模型。Rasch 模型的基本原理认为被试正确作答题目的概率可以用个体能力 θ 与该题目难度 δ 的一个简单函数来表示[2]:

$$f(P_{ni1}) = \theta_n - \delta_i$$

其中, P_{ni1} 表示被试 n 回答第 i 个题目得 1 分的概率。在 0~1 计分题目中,得 1 分表示被试答对该题目。函数表示被试答对该题的概率 P_{ni1} 取决于被试能力 θ_n 及题目难度 δ_i 。

被试能力 θ_n 和题目难度 δ_i 的取值范围为 $[-\infty, +\infty]$, 所以 $\theta_n - \delta_i$ 的取值范围也是 $[-\infty, +\infty]$ 。但正答率的取值范围仅为 $[0, 1]$ 。Rasch 利用自然对数进行转换,就得到了基本 Rasch 模型:

$$P_{ni1} = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}}$$

2.2. 多值计分 Rasch 模型

基本 Rasch 模型只能用于处理 0~1 计分的题目,常见的选择题、判断题就属于这种类型。但考试的简答题、计算题,心理量表的李克特量表都属于多值计分题目。以简答题为例,被试回答满分为 5 分的题目,得分可能有 0~6,共 6 种。基本 Rasch 模型显然无法处理这类数据。

Masters 将基本 Rasch 模型拓展为 Rasch 分部计分模型(Rasch-Masters Partial Credit Model, PCM) [3]:

$$\text{Log} \frac{P_{nij}}{P_{ni(j-1)}} = \theta_n - (\delta_i + \tau_{ij})$$

P_{nij} 表示被试 n 在第 i 个题目上得 j 分的概率; $P_{ni(j-1)}$ 表示他在该题目上得 $j-1$ 分的概率; θ_n 表示这名被试的能力; δ_i 表示第 i 个题目的整体难度; τ_{ij} 表示该题目第 j 等级的梯难度(step calibration)或阈值参数(threshold difficulty)。有 X 种得分的可能, 就有 $X-1$ 个阈值参数。基本 Rasch 模型其实是 PCM 的一个特例, 当一个题目只有两种得分的可能(0 或 1)时, 就只有一个阈值参数。一套试卷中, 由于满分不同, 可能的得分种类也不尽相同, 所以阈值参数也各不相同。

PCM 的另一个特例是等级积分模型(Rating Scale Model, RSM) [4]:

$$\text{Log} \frac{P_{nij}}{P_{ni(j-1)}} = \theta_n - (\delta_i + \tau_j)$$

心理测验的评分最常用的是李克特等级量表, 这种量表中所有题目的等级相同, 就意味着这些题目共用一套阈值参数。如 SCL90 是典型的李克特 5 级量表, 所有题目“没有、很轻、中等、偏重、严重”5 个选项的阈值参数应当相等。所以 RSM 与 PCM 公式唯一的区别就是 τ_j 的下标没有表示题目的 i , 这表示所有题目共用一套阈值参数。

关于阈值参数是否具有顺序性的问题, 说法不一。有研究者认为阈值参数可以不具顺序性, 出现非顺序性主要是因为分布在某些等级上的被试在全体被试中所占较少[5]。也有研究者认为阈值参数不具备顺序性就代表题目的质量不好, 应当修订或剔除[6]。事实上, 优质的题目, 其阈值参数具备顺序性特征, 并且相邻两个阈值参数之间的距离不能太小, 整个阈值参数的跨度应当足够大, 这样才能有效区分不同能力水平的被试[7]。

理论的发展为将 Rasch 引入教育测量领域提供了可能。Rasch 模型在欧美的教育测量界已经非常流行, 国内近些年来开始有研究者将 Rasch 模型运用到写作之外的主观题(简答题、综合题)项目分析中[8]。国外研究表明 PCM 更适用于教育测量中的多项选择题和主观题[9]。

3. 研究方法

3.1. 数据来源

本文用到的数据是某高校心理学课程考试数据。研究主要针对试卷的主观题: 76~83 题, 共 8 个题目。其中, 简答题 76~80 题, 每题满分 10 分, 评分等级为 0 到 10 分; 综合题 81~83 题每题满分 30 分, 评分等级为 0 到 30 分。有效数据 2075 名考生。

3.2. 数据处理

采用 SPSS 对收集到的原始数据进行管理, 使用 Facets3.70 软件, 选择 PCM 对数据进行分析, 研究得到的主要结果包括考生能力、拟合指数、试题阈值参数、测验信息量等。

3.3. 数据——模型拟合

Rasch 模型认为只有当数据与模型预测值拟合时测量的结果才足够客观、准确。分析得到两个主要的拟合统计量加权均方拟合统计量(Infit MNSQ)、未加权均方拟合统计量(Outfit MNSQ)。由于 Outfit MNSQ 更容易受到极值数据的影响, 研究者主要以 Infit MNSQ 作为判断数据与模型拟合的指标[10]。拟合指数为 1 说明观察值与模型预测值完全拟合; 大于 1 说明观察值与模型预测值之间存在随机偏差, 称为非拟合; 小于 1 说明观察值间差异小于模型预测, 称为过度拟合。有研究者建议在主观题项目评价中, Infit MNSQ 和 Outfit MNSQ 的取值应在[0.6, 1.4]范围内比较好, 越靠近 1 越好[11]。

考生的拟合结果显示, 有 18 名考生数据出现过度拟合、52 名考生数据出现非拟合, 占考生总体的 3.37%, 低于 Rasch 理论中 5% 的要求[12]。因此可以认为考生数据与模型拟合。题目的拟合情况见表 1,

Infit MNSQ 的取值在[0.82, 1.25]之间；Outfit MNSQ 的取值范围在[0.82, 1.29]之间。

从这些结果看，题目与模型拟合，可以使用 Rasch 模型对这组主观题进行项目分析。

Table 1. Item fitness

表 1. 题目拟合

题目	Infit MNSQ	Outfit MNSQ	平均分	标准差	相关系数	考查知识点
item76	0.82	0.84	4.77	3.16	0.67	普通心理学
item77	1.24	1.29	3.16	2.38	0.38	普通心理学
item78	0.91	0.86	3.74	3.54	0.62	发展与教育心理学
item79	0.91	0.92	4.58	2.68	0.61	实验心理学
item80	0.99	0.99	4.16	2.90	0.58	实验心理学
item81	1.25	1.26	13.66	5.74	0.59	普通心理学
item82	0.83	0.82	12.76	7.85	0.75	发展与教育心理学
item83	0.97	1.20	5.31	6.21	0.59	心理统计与测量

4. 结果与讨论

4.1. 阈值参数

PCM 中将相邻两条选项反应曲线(Category response curves, CRC)之交点命名为阈值参数(Threshold difficulty)。对于有 X 种得分可能的题目，可以得到 $X - 1$ 个参数。0 分和 1 分的 CRC 曲线交点称为 δ_1 ，依次类推。以第 76 题为例： $\delta_1 = -1.15$ ，代表 0、1 两条 CRC 曲线的交点对应的能力水平为-1.15。它的意义在于：如果能力为-1.15 的考生回答该题目，得 0 分或得 0 分以上的可能各占 50%；能力小于-1.15 的考生回答该题，得 0 分的概率大于 50%，反之亦然。如果没有考生获得某一分数，就不存在此选项的 CRC 曲线。如第 82 题，在这一题目上没有考生得 29 分，因此该题目就没有 δ_{29} 。同样第 83 题上没有考生得满分，该题目就没有 δ_{30} 。各题目的阈值参数估计结果如表 2 所示。

Table 2. Parameter estimation

表 2. 参数估计结果

Threshold	分布范围	跨度
item76	[-1.15, 0.8]	1.95
item77	[-1.18, 1.36]	2.54
item78	[-0.5, 0.61]	1.11
item79	[-1.4, 1.87]	3.27
item80	[-1.08, 1.76]	2.84
item81	[-1.69, 1.8]	3.49
item82	[-0.98, 2.82]	3.8
item83	[-0.68, 1]	1.68

第一个阈值参数对应的能力水平越低说明该题目得分越容易。 δ_1 是考生是否得分的分界点。能力为 δ_1 的考生，不得分和得分的概率各 50%。第 81 题的 $\delta_1 = -1.69$ ，在所有题目中最低。这表示考生的能力只要高于-1.69 就有可能得分。第 81 题：“阐述马斯洛的需要层次理论，并就下图(图略)马斯洛需要层次演进图说明几种需要之间的关系。”该题目要求先简述马斯洛的需要层次理论，然后根据给出的图说明各种需要之间的关系。但图片已经给出了马斯洛的需要层次演进图，对回答第一问有提示作用。这就降低了该题的难度，另外马斯洛需要层次理论是大纲中明确指出考生应当掌握的内容。受上述因素影响，

81 题为最容易得分的题目。

最难得分的题目为第 78 题：“简述德韦克的成就目标定向理论。”题目的 $\delta_1 = -0.5$ ，即考生能力应高于 -0.5 才有可能得分。这一概念虽是发展与教育心理学中的知识点，但并不像需要层次理论那样在大纲中明确指出。考生在备考时，尤其跨专业考生，在备考时如果只以大纲为准而不对课本知识充分了解，容易漏掉某些重要概念，最终出现了这样的结果。

将各题目的第一个阈值参数从小到大排列，最易得分的题目依次为：81、79、77、76、80、82、83、78。

最后一个阈值参数对应的能力水平越小说明该题得满分越容易。第 78 题的 $\delta_{10} = 0.61$ ，在所有题目中最低。这说明考生能力只要高于 0.61，就有可能得到满分。需要指出的是在前边的分析中发现，这一题目同样也是得分最难的题目。对考生的作答情况分析发现，该题目的平均分为 3.74，标准差为 3.54；得 0 分的考生有 30.1%，得高分(8~10 分)有 21.2%。表明在这个题目上，学生的得分差异很大。熟悉这一概念的考生有可能得到很高的分数。没有复习到的考生很难得分。心理学专业的教学中，成就目标定向理论属于老师经常强调的非常重要的内容，但这一知识点在大纲中却未明确体现。这就可能造成本专业考生回答优于跨专业考生的情况。另外，这一题目的阈值参数对应的能力范围在各题目中最小，阈值跨度仅为 1.11 个 logit 单位，对能力在 $[-0.5, 0.61]$ 区间外的考生测量误差较大。

最难得满分的题目为第 82 题：“试述弗洛伊德和艾里克森的心理发展观，并比较其异同”。该题目的 $\delta_{30} = 2.82$ ，即考生能力必须高于 2.82 才有可能得到满分。这一题目的参考答案大致可以分为三个部分：弗洛伊德的心理发展观、艾森克的心理发展观、两者的异同。要获得满分就需要考生答出三个部分的共十多条内容，比较困难。但另一方面，由于答案包含的内容较多，考生能够回忆起来的内容也多。所以虽然这一题目得满分很困难，但得到一个中等的分数还是比较容易的。统计结果显示该题目的平均得分为 12.76，在全部题目中属于较高的。

以最后一个阈值参数为标准将全部题目排列，最容易得满分的题目依次为：78、76、83、77、80、81、79、82。

此外，在 Rasch 模型中被试能力与题目阈值参数越接近对考生的能力估计越准确。这就要求题目的阈值参数分布得越分散越好。整套题目的阈值参数分布在 $[-1.69, 2.82]$ 之间，最高和最低阈值参数之间相差 4.51 个 logit 单位。

从单个题目来看第 79、81、82 题的阈值跨度最大，分别为 3.27、3.49、3.8。与其他题目相比，这些题目对能力分布在阈值范围内的被试能力估计更加准确，对整组试题测量结果的贡献最大。阈值跨度较小的两个题目为第 78、83 题。

4.2. 相关系数

计算题目得分和剩余全部题目得分之间的相关系数(表 1)。该相关系数代表该题目测量的能力与剩余题目测量的能力是否一致。相关系数应大于 0.3，相关性越高，说明两者的测量目标越接近。分析结果表明相关系数均在可接受范围内，除第 77 题之外均为中高度相关，即这些题目的测量目标高度一致。第 77 题：“根据感觉相互作用的现象，说明为什么重感冒患者往往会感到食而无味”。该题考察的是感知觉关系，需要考生在对知识点深入理解基础上分析题目，才能给出比较合理的答案。答题过程中需要运用课本知识之外的内容，因此该题的相关系数最低，但也达到了 Rasch 理论的要求。

4.3. 考生能力与题目阈值分布

图 1 将考生能力和题目阈值呈现在同一 Rasch 尺度下，方便研究者了解两者之间的关系。图的最左

侧为 Rasch 量尺，第二列为考生能力分布情况，右侧分别列出了 8 个题目的阈值分布情况。图 1 显示考生能力的分布在 $[-3.57, 0.83]$ 之间。题目的难度分布在 $[-1.69, 2.82]$ 区间。题目的阈值并没有完全涵盖考生的能力分布范围，即阈值与考生能力分布不相匹配。但绝大多数考生能力集中在 $[-1, 0.8]$ 之间，题目的阈值涵盖了这一部分的考生。表明在主观题较少的前提下，试卷仍然可以精确测量绝大多数考生能力。

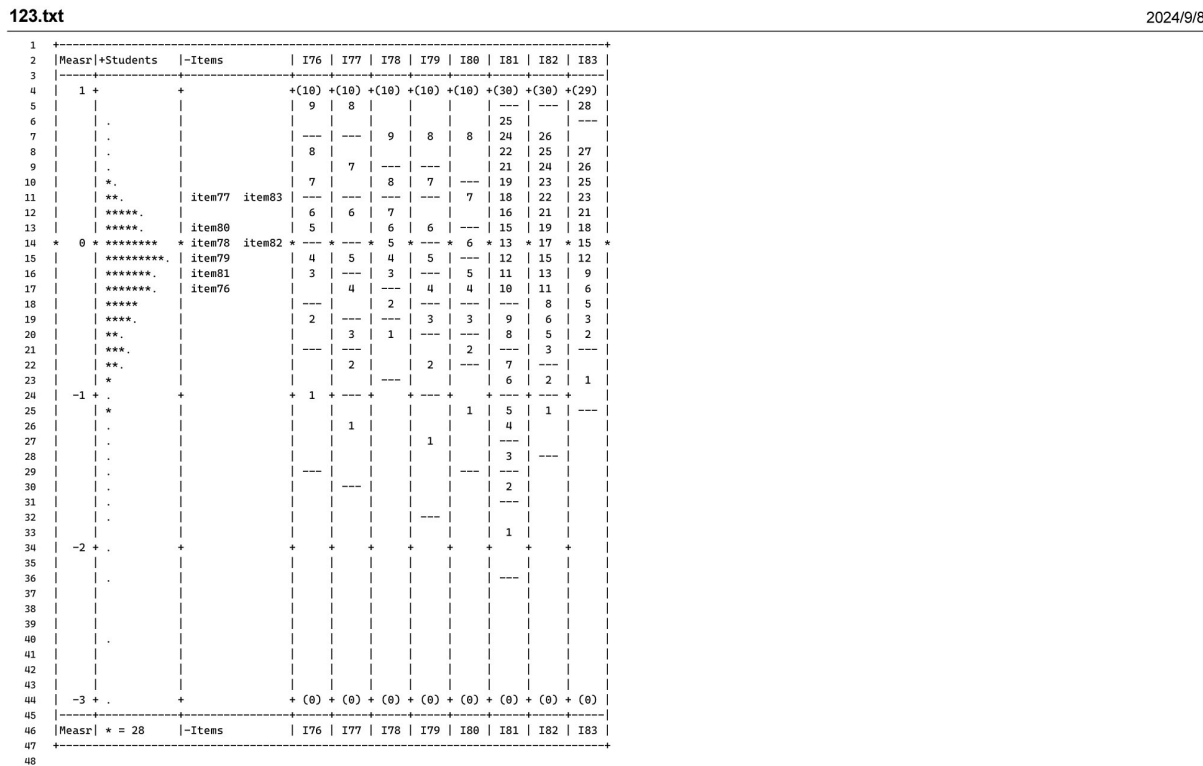


Figure 1. Wright map
图 1. 怀特图

4.4. 信息量

信息量是用于衡量测验精度的指标。现已被广泛应用于测验编制、测量精度评价、确定计分权重等方面[13]。信息量越大测量精准度越高。题目对不同能力水平考生的测量误差不同，IRT 将信息量定义为测量标准误平方的倒数，因此测量标准误差越小信息量越大。

若将本次考试作为选拔性考试，能力很高的考生肯定可以达到选拔要求，能力很低的考生则肯定被淘汰。中等能力考生分布在录取获晋级标准附近，能否精区分这部分考生是保证测验质量的关键。8 个题目的信息曲线峰值为 58.36，对应的考生能力为 0.12。说明利用这组题目对中等能力水平的被试进行测量时精度最高。

第 79、81、82 题的阈值跨度比较大，这类题目能够精确测量考生能力的范围较大，这就为测验提供了更多有价值的信息。图 2 下方的一条曲线为这三个题目的信息量，可以看出这三个题目的信息量占到了全部信息量的一半以上，对整个测验贡献很大。这三个题目的内部一致性系数为 0.689，整套题目的内部一致性系数为 0.826。两者仅相差 0.137，也能够说明这三个题目对精确测量考生能力有很大帮助。

有研究者认为测验信息量大于 16 时质量较好[14]。信息函数在 $[-0.67, 0.85]$ 间信息量大于 16，能够对中等能力考生的能力作出准确估计。

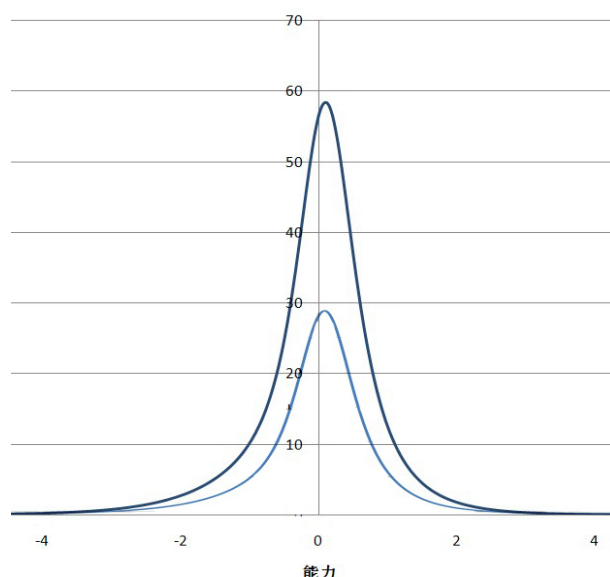


Figure 2. Item information
图 2. 测验信息量

5. 结论

8 个题目涉及 3 个普通心理学知识点、2 个发展与教育心理学知识点、2 个实验心理学知识点、1 个心理统计与测量知识点(表 1)。除第 78 题外, 简答题的知识点在考试大纲中都作了明确规定, 答案要点在参考书中都能找到。考生只需要在理解的基础上进行整合加工, 提炼出简洁的答案就可以正确作答, 这说明简答题要求的主要是考生对考察内容的理解与记忆。综合题需要考生结合所学的相关知识作答, 考察知识的综合运用能力。除第 83 题外, 其他题目的阈值跨度都达到 3 个 logit 单位以上, 能够准确区分较宽范围内的考生。整组题目涵盖了大纲规定的全部科目, 既有心理学理论也有研究方法。综合分析结果表明整组试题的质量很高, 能够很好地区分考生的能力。为客观、公正、准确地选拔提供了很多有效信息。

Rasch 理论为主观题项目分析提供了新的方法, 是对已有方法的重要补充, 为提高命题和测量质量提供了很多有用信息。Rasch 理论可以运用到大型考试的项目分析中, 为提高测评质量提供参考依据。

基金项目

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建设与建设实践(2023250); 贵州省教育科学规划课题: 基于新课标的小学生数学问题解决认知机制研究(2022B009)。

参考文献

- [1] 李中权, 孙晓敏, 张厚粲, 等. 多面 Rasch 模型在主观题评分培训中的应用[J]. 中国考试(研究版), 2008(1): 26-31.
- [2] Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Danmarks paedagogiske Institut.
- [3] Masters, G.N. (1982) A Rasch Model for Partial Credit Scoring. *Psychometrika*, **47**, 149-174.
<https://doi.org/10.1007/bf02296272>
- [4] Andrich, D. (1978) A Rating Formulation for Ordered Response Categories. *Psychometrika*, **43**, 561-573.
<https://doi.org/10.1007/bf02293814>
- [5] Wright, B.D. and Stone, M.H. (1979) Best Test Design. Rasch Measurement.

-
- [6] Andrich, D. (1988) A General Form of Rasch's Extended Logistic Model for Partial Credit Scoring. *Applied Measurement in Education*, **1**, 363-378. https://doi.org/10.1207/s15324818ame0104_7
 - [7] Van Dam, N.T., Earleywine, M. and Borders, A. (2010) Measuring Mindfulness? An Item Response Theory Analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences*, **49**, 805-810. <https://doi.org/10.1016/j.paid.2010.07.020>
 - [8] 李静璇, 王秋红, 何壮, 等. 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.
 - [9] Yan, Z. and Bond, T.G. (2011) Developing a Rasch Measurement Physical Fitness Scale for Hong Kong Primary School-Aged Students. *Measurement in Physical Education and Exercise Science*, **15**, 182-203. <https://doi.org/10.1080/1091367x.2011.590772>
 - [10] 何壮, 袁淑莉, 余水, 等. 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.
 - [11] Linacre, J.M. (1994) Many-Facet Rasch Measurement. MESA Press.
 - [12] 张洁. PETS 三级口语考试评分误差研究——结合定量统计和定性描述的方法[J]. 外语测试与教学, 2012(2): 33-42.
 - [13] 余嘉元. 项目反应理论及其应用[M]. 南京: 江苏教育出版社, 1992.
 - [14] 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.