文史哲论丛
Literature, History and Philosophy
ISSN Print: 3079-5400
ISSN Online: 3079-5419
https://doi.org/10.63313/LHP.8058

Rasch模型检测AI生成“跨文化能力考试模拟题”效度的实验分析 —— 以常州工程职业技术学院为例

娄立国

常州工程职业技术学院通识教育部，江苏 常州

发布日期：2026年9月07日

摘 要

跨文化能力考试自2023年问世以来，关注度持续提升，但配套题库资源仍显不足。本研究提出AI赋能题库资源建设的假设，以豆包生成的跨文化能力考试模拟题客观题部分为对象开展测试实验。通过运用Rasch 模型对实验数据进行整体质量、怀特图、单维性及拟合度分析，结果表明豆包生成的模拟题质量达到可接受标准。研究认为，豆包可作为题库资源建设的有效赋能工具。

关键词

Rasch模型；跨文化能力考试； CAK框架； AI生成模拟题；效度

Validity Test of “AI-Generated Intercultural Competence Test Simulants” under the CAK Framework: A Rasch Model-Based Case Study of Changzhou Vocational Institute of Engineering

Liguo Lou

Department of General Education, Changzhou Institute of Engineering Technology, Changzhou 213164, Jiangsu Province, China

Abstract

文章引用: 娄立国. (2026). Rasch 模型检测 AI 生成 “跨文化能力考试模拟题” 效度的实验分析 —— 以常州工程职业技术学院为例. 文史哲论丛, 2(3), 28 - 40. ISSN Print: 3079-5400, ISSN Online: 3079-5419.
DOI: 10.63313/LHP.8058

Since its launch around 2023, the Intercultural Competence Test (ICT) has gained increasing attention; however, supporting test item bank resources remain limited. This study hypothesizes that AI can empower the construction of ICT item bank resources. An experiment was conducted to test the objective items of an ICT simulation test generated by Doubao AI. By applying the Rasch model to analyze the overall quality of the experimental data, Wright Map, test unidimensionality, and item fit, we found that the simulation test items generated by Doubao met acceptable quality standards. Therefore, Doubao can be used to empower the construction of ICT item bank resources.

Keywords

Rasch Model; Intercultural Competence Test; CAK framework; AI-generated simulation test; Validity

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

1. 引言

跨文化交际能力（Intercultural Communicative Competence, ICC）已成为 21 世纪外语教育的核心目标之一。为系统测评与培养该能力，彭仁忠、吴卫平等[1]提出了“CAK 模型”，将 ICC 划分为跨文化交流技能（Communicative Skills）、跨文化认知技能（Cognitive Skills）、跨文化态度（Attitude）、跨文化意识（Awareness）、本国文化知识（Knowledge of Self）及外国文化知识（Knowledge of Others）六个维度。该模型与 Byram[2] 51 [3] 46-50、Deardorff[4] 247-250 等国际学者提出的 ICC 构成要素（语言技能、文化知识、开放态度与文化意识）高度契合。此后，张红玲与吴诗沁[5] 6-7 构建了跨文化能力教学参考框架，对从小学至大学各学段学生应达成的跨文化能力水平进行阶梯式细化，为跨文化能力培养提供了系统化参照。2023 年，全国跨文化能力考试（Intercultural Competence Test, ICT）正式推出并持续升温，然而 CAK 模型在一线教学实践中的直接应用仍处于起步阶段，指导学生备考的题库资源有限，学界对跨文化能力考试题库建设的探讨亦不足。

在当前人工智能技术广泛普及的时代背景下，AI 工具已深度嵌入学习与工作场景，借助 AI 提升效率已成为不可逆转的趋势。基于此，本研究尝试将 AI 应用于跨文化能力考试题库建设，采用豆包生成试题并开展测试实验。选择豆包的主要原因在于，该工具在学生群体中使用频率最高，体现出较强的教学辅助功能。该结论源于一项前期调研：研究团队通过 QQ 群向我校 2025 级 10 个行政班级（总人数逾 400 人）发放腾讯问卷（调研时间为 2026 年 1 月 3 日），共回收有效问卷 165 份，问卷浏览量为 187 次。调研结果见图 1。



Figure 1. Diagram of frequently used AI tools

图 1. 问卷题目-你最经常用的 AI 工具是什么？

问卷数据支持了以豆包为出题工具的选择。为使豆包准确理解任务需求，研究团队首先通过多轮对话进行预热，提出若干答案已知的问题，包括：（1）什么是跨文化能力考试？（2）什么是 CAK 模型？（3）您是否了解《跨文化能力考试大纲（2023 年试行版）》？该大纲是否包含跨文化能力考试样题？预热对话旨在达到两个目的：其一，引导豆包按用户需求输出内容；其二，检验豆包生成内容是否存在“幻觉”。经核对，豆包的答复与预设答案一致，表明其理解准确且输出可靠。随后向豆包上传“跨文化能力考试初级样题”（载于《跨文化能力考试大纲（2023 年试行版）》）及“跨文化能力考试模块介绍”（载于彭仁忠编《跨文化能力考试备考指南》）两份 PDF 文件，并下达指令：

“我已将跨文化能力考试初级样卷发给你，请模仿它生成一套与之类似的模拟考卷，难度、效度严格向它靠拢。注意：要提供参考答案。文件‘跨文化能力考试模块介绍’供你参考，有助于提高你生成题目的质量。”

豆包在短时间内生成了模拟试题及参考答案。经人工审阅发现题目整体质量较高，未发现明显的事实性错误。然而，即便通过了人工审核，我们对豆包所生成试题的效度仍存疑虑——毕竟，这些题目仅为权威样题的仿制品，能否真正用于跨文化能力考试备考辅导，尚缺乏实证依据。为回应这一疑问，本研究开展了测试实验，拟借助 Rasch 模型对所得数据进行分析，检验主观假设。

2. 文献综述

2.1. Rasch 模型在跨文化能力测评中的应用

Rasch 模型是项目反应理论（IRT）的核心模型之一，由 Georg Rasch 于 1960 年提出，主要用于考试、量表等心理与教育测量领域。该模型的核心优势在于能够将“被试能力”与“试题难度”置于同一等距量表（logit scale）上进行量化比较，从而实现客观测量[6]15 [7]44。

在跨文化能力测评领域，Rasch 模型已得到应用。彭仁忠等[8]运用该模型对 2021 年全国跨文化能力考试（ICT）初级测试数据进行分析，从测试维度、信度、拟合效果和项目适配度等方面系统检验了试卷质量。Garzón-Umerenkova 等人[9]在拉丁美洲五个国家开展的情绪能力问卷（ECQ）跨文化验证研究中，整合使用验证性因素分析、Rasch 建模和探索性图分析三种方法，结果表明 Rasch 分析能够有效评估项目功能、检验 Likert 量表的适用性，并识别潜在的跨文化差异。Kianinezhad 等[10]在 C-test 语言能力测试的信度比较研究中，将经典信度指标与 Rasch 模型、Mokken 量表等现代心理测量方法进行对比，发现 Rasch 分析能够提供关于项目难度和被试信度的精细化信息，从而更全面地评估测试的心理测量学属性。

上述研究为 Rasch 模型在跨文化能力测评中的应用提供了方法论支撑。尽管 Rasch 模型在国际学术界应用广泛，但在国内仍处于推广阶段，在跨文化能力测评领域的实证研究尚不多见。本研究将 Rasch 模型引入 AI 生成跨文化能力试题的效度检验，是较新颖的尝试。

2.2. AI 生成试题的效度研究现状

随着大语言模型的快速发展，AI 辅助试题生成已成为教育测评领域的研究热点。Polat 等[11]在医学教育评估研究中，系统考察了 ChatGPT 4o 生成的多项选择题难度参数与模拟多阶段测试数据的匹配程度。结果显示，AI 生成的难度参数与模拟数据之间存在中等程度的相关性，但 AI 系统性地高估了试题难度（平均偏差 0.240 logits），且对容易题目难度估计尤为不准。研究者指出，尽管 ChatGPT 作为初步试题生成工具展现出应用潜力，但在实际应用前仍需经验校准和专家审核，因为系统性偏差表明 AI 缺乏真实学生表现数据的反馈。与此同时，李峰、郭嘉悦等[12]基于 PISA 2022 创造性思维测评框架，探索利用大语言模型辅助生成情境化任务的技术路径。其研究发现 AI 生成的题目具有良好的区分度、拟合度与内部效度，并认为大语言模型可有效提升情境化命题效率，但其生成的题目仍需依赖人工审校与优化。上述研究呈现出对 AI 生成试题质量的不同判断：一方发现系统性偏差，强调校准与审核的必要性[11] 11；另一方则报告了良好的测量学指标，但仍强调人工优化不可或缺[12] 84。“褒贬不一”的研究格局凸显了将 AI 整合到教育测评中时进行严格测量验证的必要性。迄今为止，我们尚未发现将 Rasch 模型应用于 AI 生成跨文化能力试题效度检验的研究。本题立足于这一研究空白，尝试引入 Rasch 模型，对 AI 生成跨文化能力考试模拟题的效度进行检验。

3. 研究方法

3.1. 研究对象

研究对象包括试题与被试两部分。试题部分基于跨文化能力考试（ICT）初级权威样题，由 AI 工具“豆包”生成 75 道客观题，从中抽取 30 道构成微型试卷。被试部分为来自我校艺术设计、智能制造、化工制药、建筑工程等学院的 514 名大一、大二学生，均表示自愿参与测试实验。

3.2. 研究目标

本研究旨在回答以下问题：豆包生成的跨文化能力考试模拟题能否有效用于跨文化能力考试辅导？具体研究目标：（1）检验 AI 生成模拟试卷的整体质量，即信度与效度；（2）通过怀特图分析被试能力与试题难度的匹配程度；（3）检验试卷是否满足单维性；（4）分析试题与被试拟合情况。

3.3. 实验工具

Rasch 分析软件 Winsteps；豆包；外研社 iTest 智能测评云平台（用于试题发布与数据采集）。

3.4. 数据分析方法

采用 Rasch 模型对数据进行分析。具体操作如下：使用 Winsteps 软件进行整体质量分析、怀特图绘制、单维性检验及项目拟合分析，生成相关统计指标。运用 Rasch 模型核心理论框架（Bond & Fox, 2015；Bond et al., 2021；Linacre, 1998）及彭仁忠团队相关研究（彭仁忠等，2022）结论与标准，对数据进行解读，评估 AI 生成试题的测量学性能。

4. 实验步骤及有关说明

第一步：生成模拟试题

用“豆包”生成一套完整的跨文化能力考试初级模拟题。关键操作：首先通过多轮对话与豆包进行交互，使其充分理解研究目的与任务需求；随后上传《跨文化能力考试大纲（2023 年试行版）》中的初级样题，令其参照样题的结构与难度生成模拟试题，并提供参考答案。

第二步：构建微型试卷

从豆包生成的 75 道客观题中随机抽取 30 道，组成微型测试卷，每题 1 分，满分 30 分。试卷规模的确定基于以下考量：若采用整套模拟题进行测试，预计耗时约 120 分钟，高职学生难以全程保持专注；若题目数量少于 30，则样本代表性不足，难以有效检验试卷质量。30 道题的微型试卷可在 15 分钟左右完成，既能保证测量的有效性，又能避免被试疲劳。试卷的维度分布（见表 1）如下：C 维度（技能类）共 11 题，侧重考查在具体情境中得体沟通、理解并适应文化差异的能力；A 维度（态度/意识类）共 4 题，题目考查学生对文化差异的反思与开放态度；K 维度（知识类）共 15 题，考查本国及他国文化知识。

Table 1. Distribution of CAK Model Dimensions in Simulation Test Questions (N=30)

表 1：模拟试题 CAK 模型维度分布（N=30）

主维度	子维度	数量	题号及内容概括
C	Cognitive skills	5	Q18，日本客户间接沟通；Q20，阿拉伯客户用餐忌猪肉；Q24，沙特礼拜朝向麦加；Q27，日-中翻译风格调整；Q29，韩国等级文化的应对
	Communicative skills	6	Q16，法国餐厅拒绝奶酪的得体方式；Q17，加拿大室友聚会打扰学习的处理；Q19，西班牙家庭晚餐品尝特色菜；Q21，俄罗斯商务伙伴见面礼仪；Q22，沙特与女性同事交流；Q23，中英拥抱礼仪回应
A	Attitude	1	Q28，巴西人身体接触的理解
	Awareness	3	Q25，中美休闲差异的理解；Q26，澳中家庭观差异的理解；Q30，爱尔兰圣帕特里克节绿色含义
K	Knowledge of self	6	Q1，京杭大运河始建于隋朝；Q2，五四运动爆发于 1919 年；Q4，陶行知与全面发展教育思想；Q5，《诗经》是中国最早诗歌总集；Q6，九年义务教育；Q7，端午节纪念屈原
	Knowledge of others	9	Q3，二战同盟国成员；Q8，美国独立战争；Q9，工业革命发源于英国；Q10，古埃及文明发源于尼罗河；Q11，常春藤盟校成员；Q12，英国 G5 大学；Q13，美国本科入学考试；Q14，印度忌用左手递食物；Q15，泰国禁忌黄色

第三步：组织测试

将 30 道试题导入外研社 iTest 智能测评云平台，面向所选定的被试班级发布测试。测试前向学生说明诚实作答对保证研究有效性的重要意义，并明确禁止使用 AI 辅助答题。

第四步：清洗数据

对回收的答题数据进行质量筛查，剔除无效样本。无效样本的判定标准：（1）敷衍作答，即两分钟内交卷且得分低于 6 分；（2）疑似 AI 辅助作答，即十分钟内交卷且得分高于 24 分。测试卷满分为 30 分，其中 0–6 分界定为低分段，24–30 分界定为高分段。经统计，共回收答卷 514 份，剔除无效样本 91 份，得有效样本 423 份。

第五步：分析数据

借助于专业的 Rasch 分析软件 Winsteps 对样本数据进行分析，获取试卷整体质量、怀特图、单维性及项目拟合等关键数据，并使用 Bond, T. G.的 Rasch 模型分析指标对数据进行解读。

5. 研究结果与讨论

5.1. 试卷整体质量检验

将 423 份有效样本的答题数据导入 Winsteps 软件进行 Rasch 模型分析，得整体质量检测数据，如图

2 所示。

Standardized Residuals N(0,1) Mean: .01 S.D.: 1.01
Time for estimation: 0:0:0.197
Processing Table 0
ICT test data.xlsx

PERSON	423 INPUT	423 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	14.7	30.0	-.03	.43	1.00	-.1	1.01	.0
S.D.	5.7	.0	.95	.07	.14	1.1	.22	1.1
REAL RMSE	.43 TRUE SD	.84 SEPARATION	1.94 PERSON RELIABILITY	.79				

ITEM	30 INPUT	30 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	207.3	423.0	.00	.11	1.00	-.2	1.01	.1
S.D.	50.6	.0	.59	.01	.12	2.6	.19	2.6
REAL RMSE	.11 TRUE SD	.58 SEPARATION	5.21 ITEM	RELIABILITY	.96			

Figure 2. Overall Quality Assessment of the Test Paper
图 2. 试卷整体质量检测

首先，从模型整体拟合来看，残差标准化后均值为 0.01，标准差为 1.01，非常接近标准正态分布，表明 Rasch 模型与数据整体拟合良好[13] 7。其次，被试分析结果显示：被试能力估计值（Measure）均值为-0.03 logits（SD=0.95），与项目平均难度（0.00 logits）基本匹配，说明样本整体能力水平适中。被试的内部拟合（Infit MNSQ=1.00，ZSTD=-0.1）和外部拟合（Outfit MNSQ=1.01，ZSTD=0.0）均在可接受范围内，ZSTD 均值接近 0，表明被试的作答模式与模型预期基本一致（MNSQ 可接受范围 0.5 - 1.5，ZSTD 可接受范围-2 - 2[7] 60）。被试信度为 0.79，分离度为 1.94。根据 Bond 等[7] 60（2021: 60）的标准，信度大于 0.7 尚可接受，但分离度低于 2.0，说明被试能力未能被有效区分为显著不同的层次。这一结果可能与样本同质性较强有关（高职生英语水平较为接近），而非题目难度设置不当所致。彭仁忠等[8] 15 在跨文化能力考试的 Rasch 分析中也报告了类似的信度与分离度（考生分离度 2.49，远低于试题分离度 11.93），并建议结合怀特图进一步判断被试与题目的匹配情况。第三，项目信度高达 0.96，分离度为 5.21，表明题目难度层次丰富（至少可区分为 5 个难度等级），能够稳定区分不同难度水平的题目[7] 60。项目的 Infit 和 Outfit MNSQ 均值分别为 1.00 和 1.01，均在理想范围内，但 ZSTD 的标准差较大（2.6），提示少数题目可能存在拟合异常（通常|Z|>2 需关注[14] 124）。对于拟合异常题目的进一步诊断，需结合项目特征曲线进行分析[8] 17。综上，在整体上模拟题具有较高的项目信度和合理的难度分布，模型拟合度达到可接受标准。

5.2. 被试-项目怀特图分析

Winsteps 软件生成的被试-项目怀特图反映了以下情况：首先，被试能力分布。被试能力估计值分布在-2 至+3 logits 之间，呈中间多、两端少的近似正态分布。能力均值（M）位于 0 logit 附近，表明样本整体能力水平适中，与项目平均难度基本匹配[7] 59。大部分被试集中在-1 至+1.5 logits 的区间内，表明中等能力水平的被试占多数。其次，项目难度分布。项目难度估计值覆盖-1.5 至+2.5 logits 的范围，与被试能力分布范围基本重合。难度均值（M）同样位于 0 logit 附近，说明整份试卷的平均难度与样本的平均能力匹配良好，有利于获得较高的测量信度（Boone et al., 2014：139）。题目按难度从高到低排序为：Q18 > Q20, Q27 > Q19 > Q30 > Q23, Q22, Q28, Q4, Q7 > Q17, Q29 > Q10, Q21 > Q11, Q25, Q26 > Q1, Q8 >

Q15, Q16, Q24, Q9 > Q6 > Q13 > Q12, Q3, Q5 > Q2 > Q14。难度层次排序反映了跨文化能力不同维度（知识、技能、意识）的相对难度差异，可为后续教学重点和试题优化提供参考[8] 18。第三，项目与能力的匹配程度。从整体来看，项目难度较好地覆盖了被试能力的主要区间（-1 至+1.5 logits），这意味着大多数被试都能找到与自己能力水平相当的题目，测验具有良好的区分能力[15] 126。然而，在能力高端（>2.5 logits）和能力低端（< -1.5 logits）存在一定的“空隙”——几乎没有题目能精确测量这些极端被试的能力。这可能导致对顶尖学生和基础薄弱学生的测量精度略有不足，即轻微天花板效应（ceiling effect）和地板效应（floor effect）[16] 37。彭仁忠等[8] 16 在跨文化能力测试的怀特图分析中也发现了类似现象，并建议在后续版本中优化题目难度分布，以覆盖更广泛的能力区间。第四，标准差位置分析。左侧的 S 线和 T 线分别表示距离能力均值一个标准差（约 ±1 logit）和两个标准差（约 ±2 logit）的位置。右侧项目难度的 S 线和 T 线表明，大多数题目落在均值附近一个标准差内，少数极端题目（如 Q18）超出了两个标准差，这进一步验证了前述项目与能力匹配的结论。

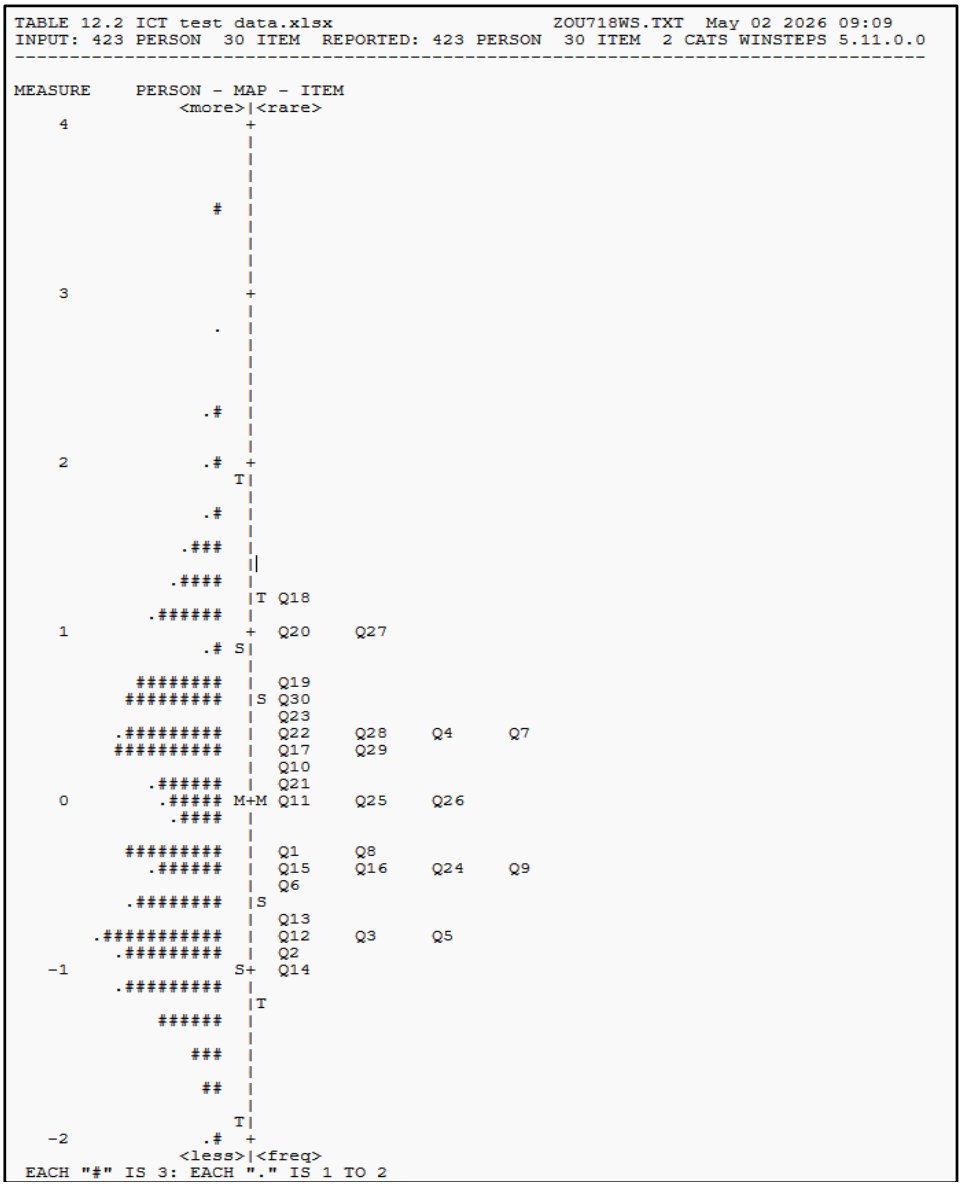


图 3. ICT 模拟题被试-项目关系分布

Figure 3. Distribution of Person-Item Relationships for ICT Simulation Test Questions

5.3. 单维性检测分析

Winsteps 软件生成了单维性检测数据，提取核心数据如表 2 所示。

表 2. 测验单维性检验摘要

Table 2. Summary of Unidimensionality Test for the Assessment

指标	数值	说明
总观测方差	37.8	原始数据总变异（特征值单位）
测量主成分解释方差	7.8 (20.6%)	Rasch 维度(能力+难度)解释的部分
未解释残差总方差	30.0 (79.4%)	模型未能解释的部分
第一对比(最大残差维度)	2.5 (6.7%)	特征值 2.5，解释残差方差的 6.7%
第一对比负载极端项目	正极(负载>0.3)：Q12(0.50)，Q9(0.50)，Q5(0.46)，Q13(0.38)，Q8(0.36)，Q14(0.35)，Q2(0.31)，Q3(0.29)，Q6(0.25) ，多为知识类题目（K-self/K-other） 负极（负载<-0.3)：Q23(-0.40)，Q20(-0.40)，Q30(-0.35)，Q19(-0.35)，Q29(-0.34)，Q18(-0.33)，Q26(-0.29)，Q17(-0.28)，Q28(-0.26)，Q22(-0.23)，Q7(-0.22) ，多为技能/意识类题目（C/A）	
后续对比特征值	第 2：1.8 (4.8%) 第 3：1.6 (4.3%) 第 4：1.4 (3.8%) 第 5：1.4 (3.6%)	均小于 2，解释方差低于 5%，属随机噪声
残差相关较高项目对	正相关：Q5 与 Q9 (0.21) 负相关：Q4 与 Q8 (-0.22)，Q9 与 Q29 (-0.21)，Q12 与 Q23 (-0.21)	绝对值均小于 0.3，局部依赖微弱

单维性（unidimensionality）要求测验中的所有题目主要测量同一种潜在特质[7] 35。研究采用残差主成分分析（PCAR）检验测验的单维性，研究发现：

首先，测验的单维性基本成立。第一对比（最大残差维度）的特征值为 2.5，解释残差方差的 6.7%。关于单维性的判断标准，要求第一对比特征值小于 2.0[7] 255。某些实证应用中也有人使用更宽松的经验性参考值（例如 <□3），但这并非正式统计阈值或官方推荐。Bokhary 等[17] 10 在开发儿童视觉生活质量量表（CVLS）的 Rasch 分析中，报告了完全相同的第一对比特征值——2.5，并据此判断量表具有良好的稳定性（原文为“... the first contrast of 2.5, which demonstrated good stability”，笔者将此陈述理解为单维性良好）。这为本研究结果提供了一个实证参照。Bakhary 研究的样本数（参与实验儿童）101 人，远低于本研究的样本数 423 人。但特征值大于 2 的事实让我们警惕可能存在第二维度。本研究中第一对比特征值为 2.5，虽略高于 2.0 的严格标准，但低于 3.0 的宽松标准。后续对比特征值均小于 2（第 2 至第 5 对比分别为 1.8、1.6、1.4、1.4），解释方差均低于 5%，表明不存在有意义的第二维度[7] 284-286。

其次，第一对比的负载模式反映内容差异。第一对比中，正负载项目（Q12、Q9、Q5、Q13、Q8、Q14、Q2、Q3、Q6）几乎全部为知识类题目（K-self/K-other），而负负载项目（Q23、Q20、Q30、Q19、

Q29、Q18、Q26、Q17、Q28、Q22、Q7）主要为技能/意识类题目（C/A）。这一负载模式反映了题目内容类型（知识 vs. 技能/意识）的差异。特征值（2.5）虽不足以威胁测验的单维性，但提示未来可进一步优化题目的难度匹配，以优化考生结构[8] 16。

第三，局部独立性良好。局部独立性（local independence）是 Rasch 模型的另一重要假设，要求控制主特质后，题目残差之间不应存在显著相关[7] 256。残差相关系数显示，所有项目对之间的相关绝对值最大为 0.22（Q4 与 Q8 的负相关）。根据 Christensen 等[18] 4 的研究，残差相关系数 0.3 在学界中也常被视为局部独立的标准，即低于 0.3 即可视为局部独立。因此，本测验满足局部独立性假设[7] 286。

综上，可以判断本测验基本满足 Rasch 模型的单维性假设，可用于跨文化能力的有效评估。第一对比负载的知识/技能分化虽提示内容差异，但未形成独立的第二维度，不影响测验的整体效度。

5.4. 项目拟合分析

项目拟合分析旨在检验每个题目是否与 Rasch 模型的预期一致，是评估测验质量的重要环节。彭仁忠等[8]指出，项目拟合检验能够有效识别需要优化的题目，为试题修订提供实证依据。拟合判断标准采用学界公认的参考值：Infit/Outfit MNSQ 可接受范围为 0.5 - 1.5，ZSTD 可接受范围为-2 至+2[7] 244-245。若某题拟合不佳（Infit/Outfit MNSQ 超出 0.5 - 1.5 或 ZSTD 超出±2），可能表明题目存在歧义、测量了不同潜在特质，或存在猜测效应[7] 64-65 [13] 3。

Winsteps 软件生成的 30 道题目拟合情况按题目难度排列如下图：

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
18	112	423	1.16	.12	1.02	.31	1.10	1.16	.35	.38	76.6	76.1	Q18
27	121	423	1.03	.12	1.41	6.57	1.66	6.79	-.04	.39	64.5	74.7	Q27
20	123	423	1.01	.12	1.06	1.02	1.11	1.38	.32	.39	73.5	74.4	Q20
19	144	423	.73	.11	1.03	.55	1.12	1.74	.35	.39	71.2	71.2	Q19
30	154	423	.61	.11	1.03	.62	1.09	1.42	.36	.40	72.6	70.0	Q30
23	165	423	.48	.11	1.06	1.30	1.07	1.20	.34	.40	67.4	69.0	Q23
4	168	423	.44	.11	.94	-1.35	.91	-1.56	.46	.40	71.9	68.7	Q4
22	171	423	.41	.11	.96	-1.03	.97	-.47	.43	.40	71.9	68.5	Q22
28	172	423	.40	.11	1.00	-.09	1.00	.05	.40	.40	69.3	68.4	Q28
7	174	423	.37	.11	1.04	1.02	1.04	.80	.36	.40	66.4	68.3	Q7
29	178	423	.33	.11	1.01	.26	1.00	.06	.39	.40	68.3	68.1	Q29
17	181	423	.29	.11	1.07	1.81	1.09	1.54	.32	.40	66.7	67.9	Q17
10	192	423	.17	.11	.98	-.55	.97	-.55	.42	.40	69.7	67.5	Q10
21	198	423	.10	.11	1.10	2.53	1.11	1.99	.30	.40	61.9	67.4	Q21
26	207	423	.00	.11	1.06	1.54	1.04	.79	.34	.39	61.7	67.2	Q26
25	208	423	-.01	.11	.91	-2.49	.89	-2.18	.48	.39	71.9	67.2	Q25
11	209	423	-.03	.11	.86	-3.92	.82	-3.50	.53	.39	75.9	67.2	Q11
8	229	423	-.25	.11	1.19	4.72	1.42	6.53	.17	.39	59.6	67.2	Q8
1	237	423	-.34	.11	1.15	3.80	1.29	4.53	.22	.38	60.8	67.3	Q1
15	238	423	-.35	.11	.99	-.13	.98	-.38	.39	.38	69.5	67.4	Q15
16	238	423	-.35	.11	.88	-3.26	.84	-2.89	.50	.38	71.9	67.4	Q16
9	240	423	-.37	.11	.89	-3.00	.86	-2.43	.49	.38	74.7	67.4	Q9
24	246	423	-.44	.11	.90	-2.55	.90	-1.57	.47	.38	74.5	67.6	Q24
6	249	423	-.48	.11	.87	-3.37	.85	-2.48	.50	.38	75.2	67.7	Q6
13	264	423	-.65	.11	.88	-2.94	.82	-2.69	.49	.37	73.8	68.5	Q13
3	274	423	-.77	.11	1.07	1.64	1.08	1.08	.29	.36	68.3	69.2	Q3
12	274	423	-.77	.11	.88	-3.04	.83	-2.44	.49	.36	73.5	69.2	Q12
5	276	423	-.79	.11	.91	-2.12	.89	-1.45	.45	.36	74.9	69.4	Q5
2	288	423	-.94	.11	.88	-2.63	.81	-2.41	.48	.35	73.5	70.8	Q2
14	289	423	-.95	.11	.89	-2.54	.80	-2.50	.47	.35	75.2	71.0	Q14
MEAN	207.3	423.0	.00	.11	1.00	-.24	1.01	.05			70.2	69.1	
P.SD	50.6	.0	.59	.00	.12	2.55	.19	2.57			4.7	2.3	

图 4. 30 道题目拟合情况

Figure 4. Fit Statistics for the 30 Test Items

据上述学界公认的参考值将拟合异常题目按照内部拟合数值（Infit MNSQ）大小排序整理，见下表 3：

表 3. 拟合异常题目排序

Table 3. Ranking of Items with Misfit

题号	难度 (Measure)	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	所属 CAK 维度	拟合状况
拟合异常题目							
Q27	1.03	**1.41**	6.6	1.66	6.8	C-Cognitive	**最异常**
Q8	-0.25	**1.19**	4.7	**1.42**	6.5	K-other	**异常**
Q1	-0.34	**1.15**	3.8	**1.29**	4.5	K-self	**异常**
Q21	0.10	**1.10**	2.5	**1.11**	2.0	C-Communicative	**异常**
Q25	-0.01	**0.91**	-2.5	**0.89**	-2.2	Awareness	**异常**
Q11	-0.03	**0.86**	-3.9	**0.82**	-3.5	K-other	**异常**
Q16	-0.35	**0.88**	-3.3	**0.84**	-2.9	C-Communicative	**异常**
Q9	-0.37	**0.89**	-3.0	**0.86**	-2.4	K-other	**异常**
Q6	-0.48	**0.87**	-3.4	**0.85**	-2.5	K-self	**异常**
Q12	-0.77	**0.88**	-3.0	**0.83**	-2.4	K-other	**异常**
Q14	-0.95	0.89	-2.5	0.80	-2.5	K-other	
Q2	-0.94	0.88	-2.6	0.81	-2.4	K-self	
正常参照题目							
Q18	1.16	1.02	0.3	1.10	1.2	C-Cognitive	正常

从维度关联性分析，K 维度题目拟合异常占比最高。从表 4 可见，30 道题目中有 12 道（40%）拟合指标异常，其中 K 维度题目 8 道，占比 67%；C 维度 3 道，占比 25%；A 维度 1 道，占比 8%。该情况表明 K 维度（知识类）题目答题数据最“不稳定”，这首先与 K 维度题目（总数 15 道）占试卷比重最高（50%）有关，此外也表明 K 维度容易出现极端答题倾向，即掌握某知识点的被试可轻易正确作答，而未掌握某知识点的被试多采用随机猜测模式快速答题。这提示我们知识类题目对高职生而言区分度有限。此外，C 维度虽只有 3 道拟合异常，但 Q27 拟合最差的事实反映技能类题目的反应变异性较大。

从 MNSQ 值特征分析，主要呈现“欠拟合”与“过拟合”两种模式[7] 243。项目拟合分析主要依据 Infit 和 Outfit MNSQ 统计量。根据 Rasch 模型的理论，当 MNSQ > 1.0 时，表明题目存在欠拟合（underfit），即数据反应模式比模型预期更具随机性、噪声过多；当 MNSQ < 1.0 时，表明题目存在过拟合（overfit），即数据反应模式比模型预期更具可预测性、过于理想化[19] 5。学界通常采用 MNSQ 介于 0.5 - 1.5 作为可接受范围，超出此范围则需结合 ZSTD 值进一步诊断[7] 244-245。

首先，MNSQ > 1.5 为“欠拟合”题目，即反应不可预测。从表 5 知 Q27（日-中翻译风格调整）的 Outfit MNSQ = 1.66，ZSTD = 6.8，是拟合最差的题目,也是试卷中仅次于 Q18 的第二难题，仅 121 名（占样本总体 28%）被试答对。该题如下：

Q27. You are a Chinese translator working for a Japanese company. When translating for a Japanese leader and a Chinese client, what is the most appropriate way to communicate? (初级/C 维度) （答案：D）

A. Use direct and concise language for both sides

B. Use indirect and euphemistic language for both sides

C. Use euphemistic language for the Japanese leader and direct language for the Chinese client

D. Adjust the language style according to the characteristics of both sides

该题考察认知技能维度，要求被试理解跨文化翻译中需根据双方文化特点调整沟通风格。其拟合不佳的可能原因包括：1）题目情境对高职生过于抽象；2）选项设计可能存在歧义，如生词过多导致读不懂题，如“euphemistic”和“concise”很难为高职生所熟悉，导致被试猜测[7] 64-65；3）某些被试存在文化刻板印象，导致反应模式异常。彭仁忠等[8] 17 发现，认知技能类题目容易因选项干扰性不足而产生拟合偏差。我们认为 Q27 的情况恰恰相反，其选项的干扰性过强，也让被试无所适从。

MNSQ < 0.5 为“过拟合”题目，反应过于可预测。多道题目（Q11、Q16、Q9、Q6、Q12 等）的 Infit/Outfit MNSQ 在 0.86–0.91 之间，ZSTD 在-3.9 至-2.2 之间。这类题目虽未低于 0.5 的严格下限，但显著偏低的 MNSQ 和负向 ZSTD 提示题目可能过于简单或存在冗余[7] 244。这些题目主要分布在 K-other（Q11、Q9、Q12）和 K-self（Q6）维度，与怀特图中知识类题目普遍偏易的观察一致。

项目拟合分析表明，模拟题整体上可接受，但 12 道题目（占总题量 40%）需关注：Q27、Q8、Q1、Q21 需优先优化——可检查题干表述清晰度、添加生词注释、视情况减少或增加选项干扰性。Q11、Q16、Q9、Q6、Q12 等题目拟合偏低但未严重偏离——可考虑增加难度或保留作为基础题。对于拟合正常的极端题目（Q14 最易、Q18 最难），可根据测验目的决定是否保留——若需精确测量中等能力群体，可考虑替换；若需扩展测量范围，则可保留。Wu 和 Adams[20] 85 指出，拟合统计量反映的是“相对拟合”而非“绝对拟合”，不应机械地据此剔除题目，因为这可能导致优质题目被移出，而应作为诊断指标，结合内容分析做出审慎判断。彭仁忠等[8] 18 也指出少数题目因难度与目标群体不匹配而产生拟合偏差，可结合内容分析进行修订。

6. 结论

本研究基于 Rasch 模型，从试卷整体质量、项目-被试匹配（怀特图）、单维性以及项目拟合度四个维度，对 AI（豆包）生成的跨文化能力考试模拟题客观题部分进行了系统检验。发现如下：第一，模型整体拟合良好。残差标准化均值为 0.01，标准差为 1.01，与标准正态分布的理想值（0，1）几乎完全一致，表明 Rasch 模型与实验数据之间不存在系统性偏差，模型建构合理。第二，试卷满足单维性假设。残差主成分分析显示，第一对比特征值为 2.5，低于 3.0 的经验临界值，后续对比特征值均小于 2，表明试题主要测量同一潜在特质——跨文化能力，测验结构效度可接受。第三，被试与试题匹配度总体可接受，但存在优化空间。怀特图分析显示，被试能力与试题难度分布范围基本重合，但能力高端与低端区间存在测量“空隙”，建议在后续题库建设中适当增加高难度与低难度题目，以覆盖更广泛的能力区间，提升测量精度。第四，项目拟合分析识别出 12 道拟合异常题目，占总题量 40%。基于上述四点，可以认为豆包所生成的这套跨文化能力考试初级样题的效度达到合格水平。

回顾本测试实验，得出结论如下：第一，数据清洗对实验分析至关重要。本研究剔除 91 份无效样本（占比 17.7%）后，模型拟合指标达到合格水平，表明严格的数据筛查是 Rasch 分析有效性的重要前提。第二，试卷的结构对教学实验结果有影响。本研究中试卷的 30 道题目中 K 维度题目占比 50%，比例过高。8 道 K 维度题目拟合异常，占拟合异常题目总量的 67%。由此可知如果将 C、A、K 三个维度题目的

占比控制在更加合理的范围内，可以预期实验结果会更加理想。基于上述结论，得出一个推测：AI 生成试题的质量高度依赖人机交互水平。AI 辅助题库建设的有效性，最终取决于使用者是否具备科学驾驭 AI 的能力——只有通过精准的提示词设计、有效的多轮对话以及必要的人工审校，才能充分发挥 AI 在试题生成中的赋能作用。

本研究的局限性亦需明确：第一，测试实验仅检验了客观题部分（客观题占考试分值 70%），未涉及主观题（占考试分值 30%）。第二，测验数据的真实性难以达到 100%。尽管在测试过程中反复向学生强调诚实答题对于教学研究的重要性，但不排除少数学生在测试过程中借助于 AI 工具答题的现象。

基金项目

2025 年度江苏高校哲学社会科学研究项目“基于 CAK 模型的高职英语‘跨文化交际案例库’建设研究（项目编号：30230100125023）”阶段性成果；2025 年常州工程职业技术学院教改课题“AI 融入大学英语教学：优势、风险与优化策略研究（编号：25JY025）”阶段性成果。

参考文献

[1] 彭仁忠,付容容,吴卫平.新时代背景下跨文化外语教学理论模型和实践模型研究[J].外语界,2020,(04):45-53.

[2] Byram M. ‘Cultural awareness’ as vocabulary learning[J]. Language learning journal, 1997, 16(1): 51-57.

[3] Byram, M. Teaching and assessing intercultural communicative competence (Revised edition) [M]. Bristol: Multilingual Matters, 2021.

[4] Deardorff, D. K. The identification and assessment of intercultural competence as a student outcome of internationalization[J]. Journal of Studies in International Education, 2006, 10(3), 241-266.

[5] 张红玲,吴诗沁.外语教育中的跨文化能力教学参考框架研制[J].外语界,2022,(05):2-11.

[6] Bond, T. G., & Fox, C. M. Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.) [M]. Routledge, 2015.

[7] Bond, T. G., Yan, Z., & Heene, M. Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.) [M]. Routledge, 2021.

[8] 彭仁忠,朱崇光,吴卫平.基于 Rasch 模型的跨文化能力测试质量研究[J].外语界,2022,(05):12-19+79.

[9] Umerenkova A G, Alba M, Cendales B, et al. Integrated validation of a cross-cultural emotional competence scale using Confirmatory Factor Analysis, Rasch Modeling and Exploratory Graph Analysis[J]. MethodsX, 2025: 103627.

[10] Kianinezhad N, Kianinezhad M. Comparative evaluation of C-test reliability using classical and modern psychometric methods[J]. Language Education & Assessment, 2025, 8: 2279-2279.

[11] Polat M, Karadag E. AI-generated biochemistry test item parameters in MST test conditions[J]. BMC Medical Education, 2025, 25(1): 1705.

[12] 李峰,郭嘉悦,胡新雨,等.大语言模型辅助情境化命题模式探索——以创造性思维测评为例[J].中国考试,2025,(09):76-86.

[13] Linacre J M. Detecting multidimensionality: which residual data-type works best?[J]. Journal of outcome measurement, 1998, 2: 266-283.

[14] 蔡旻君&马光仲. Rasch 模型教育测量应用基础[M]. 北京：中国科学技术出版社，2021.

[15] Boone W J, Staver J R, Yale M S. Rasch analysis in the human sciences (Vol. 10, pp. 978-94) [M]. Dordrecht: Springer, 2014.

[16] Terwee C B, Bot S D M, De Boer M R, et al. Quality criteria were proposed for measurement properties of health status questionnaires[J]. Journal of clinical epidemiology, 2007, 60(1): 34-42.

[17] Bokhary K A, Suttle C, Alotaibi A G, et al. Development and validation of the 21-item Children's Vision for Living Scale (CVLS) by Rasch analysis[J]. Clinical and Experimental Optometry, 2013, 96(6): 566-576.

[18] Christensen K B, Makransky G, Horton M. Critical values for Yen’s Q 3: Identification of local dependence in the Rasch model using residual correlations[J]. Applied psychological measurement, 2017, 41(3): 178-194.

[19] Smith A B, Rush R, Fallowfield L J, et al. Rasch fit statistics and sample size considerations for polytomous data[J].

BMC medical research methodology, 2008, 8(1): 33.

[20] Wu M, Adams R. Applying the Rasch model to psycho-social measurement: A practical approach[M]. Melbourne: Educational Measurement Solutions, 2007.