

智能测评环境下公众英语演讲能力评价量表的开发与检验

Development and Validation of a Rubric for Assessing English Public Speaking

Competence in an AI-assisted Automated Assessment Environment

王艾琳¹, 郑春萍^{1*}, 沈彤彤¹, 鹿萍¹

¹ 北京邮电大学

*zhengchunping@bupt.edu.cn

【摘要】 公众英语演讲能力是学生综合素质发展的核心能力之一，但在智能技术赋能教育评价的趋势下，现有的评价量表难以适配人机协同的评估需求。本研究旨在开发并验证一份公众英语演讲能力的评价量表，通过自主研发的多模态教学评价平台，采集了 110 名学生一学期三次演讲的教师评价与机器多模态数据。结果表明：(1) 构建的量表包含 5 个维度，共 21 个子维度，且具有良好的信效度；(2) 在 10 个可量化子维度上，机器自动评分与教师人工评价呈显著正相关。本研究开发的量表不仅为高校英语演讲教学提供了科学的测评工具，也为人机协同的评估模式提供了实证依据。

【关键词】 量表开发；公众演讲；语言教学；自动测评

Abstract: Public English speaking skills are one of the core competencies for the development of students' comprehensive quality. However, amid the trend of intelligent technology-empowered educational evaluation, existing evaluation scales struggle to meet the assessment needs of human-machine collaboration. This study aims to develop and validate an evaluation scale for public English speaking skills. Through a self-developed multimodal teaching evaluation platform, teacher evaluations and machine multimodal data of three speeches delivered by 110 students over one semester were collected. The results show that: (1) The constructed scale includes 5 dimensions and a total of 21 sub-dimensions, with good reliability and validity; (2) For 10 quantifiable sub-dimensions, there is a significantly positive correlation between automatic machine scoring and manual teacher evaluation. The scale developed in this study not only provides a scientific assessment tool for college English speaking teaching but also offers empirical evidence for the human-machine collaborative assessment model.

Keywords: Rubric development, English public speaking, Language teaching, Intelligent assessment

1. 研究背景

公众英语演讲是高校学生综合素质发展的核心能力之一，涵盖语言技能、跨文化交际、批判性思维与情绪调控等综合素养（De Paola et al., 2021; Morreale et al., 2006）。因此，科学、全面的评估工具是诊断学习者能力短板、提升教学干预有效性的前提。

从公众英语演讲能力的评价维度来看，早期研究聚焦单一语言维度，后续虽拓展至“语言+非语言”的框架发展，例如，Mercer 等（2017）提出的口语素养四维框架，将口语能力划分为物理维度、语言维度、认知维度与社交情感维度，但仍普遍忽视“演讲焦虑与自我调控”，且对“视觉辅助材料设计能力”覆盖不足，难以实现精准测评（Brunfaut et al., 2025）。

近年来，人工智能与多模态学习分析技术的发展推动着公众英语演讲评价向人机协同模式演进。现有自动化评价研究主要聚焦语音特征（如发音准确性、流畅度）与简单非语言特征（如面部表情）的提取与评价，借助 ASR 语音识别、OpenFace 面部特征提取等技术实现了部分维度的机器评分（Burton, 2026）。但当前人机协同评价模式的构建仍面临挑战：一是缺乏与机器评价适配的公众英语演讲评价量表，并验证人机评价在各维度的一致性；二是现有研究多聚焦单一模态特征的提取，未实现语音、视觉、文本等多模态特征的融合分析，难以全面支撑综合能力的机器评价。

因此，开发一份适配智能测评环境的公众英语演讲能力评价量表，能够为人机协同评估模式的实施提供重要工具，对提升公众英语演讲教学质量具有重要的理论与实践意义。

2. 研究设计

本研究选取中国北方某高校 110 名本科非英语专业学生（男生 19 人，女生 91 人，平均 19.22 岁），邀请 3 名资深公众英语演讲讲师，共同参与量表构建与内容效度评审。研究流程遵循专家访谈与德尔菲法，初拟 5 个核心维度、21 条子指标；随后基于自研多模态公众英语演讲评价平台（该平台依托多模态智能评估技术与大语言模型，自主研发面向演讲语言表达、演讲展现与情绪调控三大维度的多模态智能评估模型），采集 16 周内 4 次演讲的语音、语言及多模态数据。最终通过 110 名学生的教师评分数据（660 份）、机器评分数据（130 份），使用 SPSS 26.0 与 AMOS 24.0 进行数据处理。

3. 研究结果

3.1. 量表效度检验结果

本研究首先进行探索性因子分析（EFA），分析前适用性检验结果表明数据符合因子分析前提（ $\chi^2=2242.88$ ， $df=210$ ， $p<0.001$ ）。采用主成分分析法并经最大方差旋转后，共提取 5 个特征值大于 1 的公因子，其累计方差解释率达到可接受水平，初始量表中 3 个未能形成有效因子结构的子维度予以剔除后，各子维度因子载荷均满足标准，题项与因子归属明确（见表 1）。验证性因子分析（CFA）结果进一步表明，模型拟合指标均符合统计学推荐标准，聚合效度与区分效度检验均满足判断标准，进一步验证了量表的结构效度（见表 2）。

表 1 探索性因子分析结果

维度与子维度	因子 1	因子 2	因子 3	因子 4	因子 5	α 系数	均值	标准差
因子 1：演讲焦虑自我调控	-	-	-	-	-	0.92	4.38	0.34
面部表情调控	0.76	-	-	-	-	-	-	-
眼神行为调控	0.85	-	-	-	-	-	-	-
肢体动作调控	0.81	-	-	-	-	-	-	-
因子 2：口语能力	-	-	-	-	-	0.90	3.19	0.35
准确性	-	0.82	-	-	-	-	-	-
流畅度	-	0.64	-	-	-	-	-	-
清晰度	-	0.85	-	-	-	-	-	-
表现力	-	0.56	-	-	-	-	-	-
因子 3：演讲稿写作能力	-	-	-	-	-	0.88	4.19	0.29
词汇运用	-	-	0.72	-	-	-	-	-
句式运用	-	-	0.72	-	-	-	-	-
演讲结构	-	-	0.66	-	-	-	-	-
相关性	-	-	0.67	-	-	-	-	-
适切性	-	-	0.60	-	-	-	-	-
逻辑性	-	-	0.71	-	-	-	-	-
论证力	-	-	0.67	-	-	-	-	-
因子 4：演讲展现能力	-	-	-	-	-	0.95	3.86	0.41
展现性面部表情	0.87	-	-	-	-	-	-	-
展现性肢体动作	-	-	-	0.81	-	-	-	-
展现性眼神接触	-	-	-	0.84	-	-	-	-
因子 5：视觉辅助材料设计能力	-	-	-	-	-	0.94	4.17	0.33
布局	-	-	-	-	0.83	-	-	-
字体大小	-	-	-	-	0.77	-	-	-
配色方案	-	-	-	-	0.79	-	-	-
内容	-	-	-	-	0.86	-	-	-

注：整体 α 系数=0.95，累计方差解释率=79.24%，N=110

表 2 验证性因子分析结果

潜在变量	组合信度 (CR)	平均方差提取值 (AVE)	χ^2/df	CFI	NFI	IFI	RMR	参考标准	结果判断
演讲焦虑自我调控	0.94	0.84	1.81	0.94	0.88	0.94	0.07	CR $\geq$ 0.60;	优秀
口语能力	0.92	0.74						AVE $\geq$ 0.40;	优秀
演讲稿写作能力	0.91	0.60						χ^2/df 1.0-2.0	优秀
演讲展现能力	0.95	0.79						CFI/IFI $\geq$ 0.90	优秀
视觉辅助材料设计能力	0.94	0.76						NFI $\geq$ 0.80	优秀
								RMR $\leq$ 0.08	

注：1. 各潜在变量 AVE 的平方根均大于其与其他潜在变量的相关系数，判别效度良好；2. 整体模型收敛效度与拟合度均达标，五因子结构稳定。

3.2. 量表信度检验结果

信度检验结果显示，量表整体及各维度信度均达标（见表 3）。内部一致性信度层面，整体 Cronbach's α 系数为 0.95，各维度介于 0.88-0.95 之间，均达优秀标准；评分者间信度层面，相邻同意率均值 94.9%、加权 Kappa 系数 0.76，表明评分一致性良好，量表评价标准具有较强操作性。

表 3 量表信度检验结果

评价指标	Cronbach's α 系数	相邻同意率 (%)	加权 Kappa 系数	参考标准	结果判断
演讲焦虑自我调控	0.92	94.9	0.76**	Cronbach's $\alpha \geq 0.8$ （优秀）； 相邻同意率 $\geq 85\%$ （良好）； 加权 Kappa ≥ 0.75 （良好）； ** $p < 0.001$	优秀
口语能力	0.90				优秀
演讲稿写作能力	0.88				优秀
演讲展现能力	0.95				优秀
视觉辅助材料设计能力	0.94				优秀
量表整体	0.95	-	-		优秀

注：1. 相邻同意率、加权 Kappa 系数为 2 名评分者的评分一致性结果；2. **表示 $p < 0.001$ ，差异具有极显著统计学意义。

3.3. 人机评价相关性结果

人机评价相关性分析结果表明，10 个可机器评估子维度与教师人工评价均呈现极显著正相关（ $p < 0.01$ ），相关系数介于 0.73~0.87 之间，验证了机器评价在这些维度上的可靠性（见表 4）。其中，自我调控肢体运动、口语准确性、写作词汇能力 3 个子维度的相关系数最高（ $\rho = 0.87$ ），说明机器在非语言肢体动作捕捉、语言准确性判断及词汇复杂度分析上与人工评价的一致性最优；演讲结构合理性维度相关性相对较低（ $\rho = 0.73$ ），但仍达显著相关标准，可能与文本逻辑衔接的质性判断难度较高有关。不同评估技术支撑下的机器评价均能较好匹配人工评价结果，为后续构建人机协同的评价模式提供了核心实证依据。

表 4 人机评价相关性结果

可机器评估子维度	相关系数 (ρ)	显著性水平
自我调控面部运动	0.83**	$p < 0.01$
自我调控目光行为	0.77**	$p < 0.01$
自我调控肢体运动	0.87**	$p < 0.01$
口语准确性	0.87**	$p < 0.01$
口语流畅性	0.86**	$p < 0.01$
口语清晰度	0.84**	$p < 0.01$
写作词汇能力	0.87**	$p < 0.01$
写作句子能力	0.78**	$p < 0.01$
演讲结构合理性	0.73**	$p < 0.01$
写作内容相关性	0.83**	$p < 0.01$

注：**表示 $p < 0.01$ ，差异具有极显著统计学意义；10 个可机器评估子维度与人机相关系数均 ≥ 0.73 ，表明机器评分与教师评分均呈显著正相关。

为清晰呈现本研究核心成果，现将经信效度检验的智能测评环境下公众英语演讲能力评价量表展示如下（表 5），其中标注“√”的子维度为可机器评估维度。

表 5 公众英语演讲能力评价量表

核心维度	子维度	子维度描述（核心评价要点）
演讲焦虑自我调控	自我调控面部运动√	演讲中通过面部表情调整缓解焦虑，保持自然状态
	自我调控目光行为√	通过稳定目光接触调控焦虑，避免回避或过度紧张注视
	自我调控肢体运动√	控制肢体动作（如手势、站姿），减少焦虑引发的多余动作
口语能力	准确性√	词汇、语法使用准确，发音标准，无明显错误
	流畅度√	表达连贯，停顿合理，无过多重复或卡顿
	清晰度√	语音洪亮清晰，语调自然，易于听众理解
	表现力	语调富有变化，能通过语音传递情感与表达重点
演讲稿写作能力	词汇运用√	词汇选择恰当、丰富，符合演讲主题与语境
	句式运用√	句式多样、结构正确，表达简洁有力
	演讲结构√	开篇引入、主体论证、结尾总结逻辑清晰，层次分明
	相关性√	内容紧扣主题，无偏离核心议题的表述
	适切性	内容难度、语言风格适配听众与演讲场景
	逻辑性	论点明确，论据与论点关联紧密，推理严谨

演讲展现能力	论证力	论据充分可信，能有效支撑观点，说服力强
	展现性面部表情	通过表情传递情感，增强演讲感染力，适配内容基调
视觉辅助材料设计能力	展现性肢体动作	手势、站姿等动作自然协调，辅助强化表达内容
	展现性眼神接触	与听众保持合理眼神互动，增强沟通感与参与感
	布局	PPT等材料排版整洁、层次清晰，重点突出
	字体大小	字体、字号适配演示场景，易于远距离阅读
	配色方案	色彩搭配协调舒适，无视觉干扰，适配主题
	内容	辅助材料内容简洁精炼，与演讲内容互补，不冗余

注：标注“√”的子维度为可机器评估维度，相关系数均 ≥ 0.73 ($p < 0.01$)，与教师人工评价呈极显著正相关，可纳入“机器初评”环节；未标注子维度需人工精评。

4. 结论

本研究构建并验证了智能测评环境下公众英语演讲能力评价量表，补充了现有评价工具在演讲焦虑调控、视觉辅助设计等维度的覆盖缺口 (Palmour, 2024)，明确区分了机器可自动评估与需人工判定的条目，10 个可机评子维度与教师人工评价呈极显著正相关，可由机器完成初评；论证力、适切性等质性子维度需人工精评。此外，量表构建的五维度框架，尤其是演讲焦虑调控与视觉辅助设计评价，适配数字化演讲教学场景，为多模态技术驱动的人机协同测评模式提供了标准化工具。

此外，本研究仍然存在以下局限：样本仅覆盖单一高校外国语言文学专业学生，未涉及不同学科背景、不同英语水平的群体，量表普适性需进一步验证；研究未纳入跨文化演讲场景，对跨文化交际要素的评价覆盖不足。未来研究可从三方面推进：扩大样本范围，纳入多学科、多院校学生，完善量表普适性检验；新增跨文化演讲评价维度，补充文化适配性等指标；优化机器评价算法，探索论证力、逻辑性等质性维度的量化路径，进一步提升人机协同评价的全面性。

参考文献

Brunfaut, T., & Kormos, J. (2025). Assessing Multimodal Viewing-to-Write Constructs: Task Design, Performance, Processing, and Rating. *Language Assessment Quarterly*, 22(4–5), 429–459. <https://doi.org/10.1080/15434303.2025.2596374>

Burton, J. D. (2026). The salience and function of nonverbal behavior in L2 speech evaluations. *TESOL Quarterly*, tesq.70073. <https://doi.org/10.1002/tesq.70073>

De Paola, M., Lombardo, R., Pupo, V., & Scoppa, V. (2021). Do women shy away from public speaking? A field experiment. *Labour Economics*, 70, 102001.

Mercer, N., Warwick, P., & Ahmed, A. (2017). An oracy assessment toolkit: Linking research and development in the assessment of students’ spoken language skills at age 11-12. *Learning and Instruction*, 48, 51–60. <https://doi.org/10.1016/j.learninstruc.2016.10.005>

Morreale, S. P., Hugenberg, L. W., & Worley, D. W. (2006). The basic communication course at U.S. colleges and universities in the 21st century: Study VII. *Communication Education*, 55(4), 415-437. 55(4), 415-437.

Palmour, L. (2024). Assessing speaking through multimodal oral presentations: The case of construct underrepresentation in EAP contexts. *Language Testing*, 41(1), 9–34. <https://doi.org/10.1177/02655322231183077>