中高级汉语二语写作评价量表构建及验证

Development and Validation of a Rubric for Intermediate-Advanced Chinese as a Second Language Writing

徐顺锦[1]
Shunjin XU
中国上海师范大学对外汉语学院
International College of Chinese Studies, Shanghai Normal University, China
沙特阿拉伯米克达德・本・阿姆尔中学
Al-Miqdad Bin Amr Intermediate School, Saudi Arabia
1927004541@qq.com

骆慧歆[2]
Huixin LUO
中国广西民族大学文学院
College of Literature, Guangxi Minzu University, China
沙特阿拉伯塔布克第三十二女子中学
The 32nd Middle School for Girls, Saudi Arabia
leslie980324@163.com

DOI: http://doi.org/10.5281/zenodo.16630688

摘　要 本研究基于汉语二语写作教学需求，结合多部大纲与相关评价标准，设计了一份分项式写作评价量表，并通过专家评审与评分实证验证其信效度。研究结果表明，该量表涵盖词汇、语法、篇章结构与内容等核心维度，能有效区分中高级学习者写作水平，提高评分者间的一致性，并在教师和学生评分员之间均保持稳定。尽管分项式评分较整体式更为复杂，但在诊断写作问题、提供多维度反馈等方面具有明显优势，可为中高级汉语二语写作教学评价与改进提供重要参考。

关键词 汉语二语写作；评价量表；信度验证

Abstract This study develops an analytic rubric for intermediate and advanced learners of Chinese as a second language (CSL), drawing on multiple guidelines and evaluation standards. Expert reviews and empirical scoring confirmed the rubric's reliability and validity. Covering key dimensions such as vocabulary, grammar, organization, and content, the rubric effectively differentiates learner proficiency levels and improves inter-rater consistency across both teacher and student raters. Though more complex than holistic scoring methods, the analytic rubric provides enhanced diagnostic insights and multi-dimensional feedback, offering valuable guidance for the assessment and improvement of intermediate- to advanced-level CSL writing instruction.

收稿日期：2024-11-26
作者简介：[1] 徐顺锦，上海师范大学对外汉语学院博士研究生，沙特塔布克米克达德・本・阿姆尔中学中文教师。
[2] 骆慧歆，广西民族大学文学院硕士研究生，沙特塔布克第三十二女子中学中文教师。

Keywords Chinese as a Second Language writing; Rubric; Reliability validation

一、引言

评价量表（rubrics）是一种由教师或专业人员开发的评价工具，旨在通过具体、可操作的描述性语言，按照不同等级标准对学习者的表现进行衡量（罗晓杰等，2010；Reddy & Andrade, 2010; Panadero & Jonsson, 2020）。该工具能够有效评估复杂能力（如写作能力），并提供多维度的表现分析（East, 2009; Andrade et al., 2009; Panadero & Jonsson, 2013）。传统上，评价量表主要应用于总结性测评和高风险的大规模测试中，能够提高评分员评分的一致性和效率，同时增强评估的透明度（Jonsson & Svingby, 2007）。近年来的研究表明，评价量表同样适用于日常教学与学习场景，不仅有助于教师了解学习者的水平、跟踪其学习进展并提供有效反馈（Stevens & Levi, 2005; Panadero & Jonsson, 2013），同时也可帮助学习者明确教师期望、识别自身不足，从而实现更具针对性的学习改进（Andrade et al., 2009）。

在课堂评估中，评价量表具有多方面优势，例如降低学习焦虑、增强反馈效果、支持自我调节（self-regulation），并提高学习者成绩（Panadero & Jonsson, 2013）。在二语写作教学中，评价量表对学习者写作能力的评估应用也愈发普遍（East, 2009）。然而，现有二语写作评价量表多用于总结性测试或大规模考试，其内容设计可能难以满足课堂写作中以诊断为目的的评价需求（邹绍艳、范劲松，2019）。相较之下，课堂写作评价量表应具备更强的针对性与细致性，以满足特定课堂或作业需求（Weigle, 2002）。因此，为特定课程、教学目标或不同语言水平研发适用的课堂写作评价量表，不仅能提高评分和反馈的效率，还能有效促进学习者写作能力的提升。

目前常见的写作评价量表类型包括整体式（holistic）、分项式（analytic）、主特质式（primary trait）和多特质式（multiple-trait）。其中，主特质式和多特质式通常为特定写作任务量身定制，虽能提升评分效度，但因设计复杂、应用成本较高，难以在常规教学中广泛推广。整体式评价基于对文本的总体印象进行评分，操作简单且效率高，但难以解释得分依据或提供详细反馈（Weigle, 2002）。分项式评价则通过设定多个维度和评价标准，对写作的不同方面进行综合评估，评分信度高，且能提供详尽的诊断信息，有助于评价培训及为学习者提供具体反馈。综合而言，分项式评价量表更适用于二语课堂写作评估。

在国内二语写作测评领域，分项式评价量表的构建研究主要集中于英语学界，涉及大规模写作测试（李清华，2014；邹绍艳，2017）、专门用途写作（王丽，2016；吴雪峰，2018）以及课堂写作（白丽茹，2012；马晓梅等，2022）。相比之下，汉语二语学界在评价量表构建方面的研究较为有限。

本研究拟构建一份适用于中高级汉语学习者课堂写作测评的分项式评价量表，并对其信效度进行验证。我们期望该量表的开发为一线教师提供参考，帮助其制定符合自身课堂需求的评价工具，从而促进汉语二语学习者写作能力的发展。

二、研究设计

（一）研究问题

1. 中高级汉语二语写作评价量表应从哪些维度对文本进行评估？如何进行评估？

2. 构建的汉语二语写作评价量表的信效度如何？能否缩小评分员之间的评分差距？

（二）参与人员

专家评估组：本研究邀请了 7 位在中高级汉语写作教学与研究方面具有长期经验的汉语教师填写评价量表有效性问卷并提出改进建议。受邀专家大多拥有五年以上的汉语教学经验，尤其侧重于中高级汉语二语写作课程的课堂教学与相关研究，能够对不同水平的汉语学习者写作表现进行较为精准的评估与分析。在年龄与性别构成方面，本研究力求覆盖不同群体（男性教师 2 名，女性教师 5 名；年龄分布在 27 岁至 55 岁之间），以尽可能多元的视角为评价量表的修订提供依据。

评分员一组：基于研究初步评阅的需求及研究规模限制，本研究选取了两位在中高级汉语写作课程具有较长时间教学与评估经验的评分员，参与写作样本的初步分析。两位评分员均为汉语国际教育专业的硕士或博士研究生，并已在中高级写作课堂担任兼职教师或研究助理超过两年，具备扎实的写作教学基础与评分实务经验。在初步评分过程中，他们通过多次讨论交流，归纳不同分数组作文的典型特征，并为后续评估维度的设定及各维度的具体描述提供了详实依据。

评分员二组：在量表信度验证阶段，共有 10 位评分员参与评分任务。其中 5 位为拥有五年以上中高级汉语教学经验的专职教师，熟悉中高级写作教学大纲与评分要点；另外 5 位则是汉语国际教育专业的硕士研究生，当前均以助教或兼职教师身份参与中高级写作课程教学，具备一定的实操与评改经验。该团队在年龄、性别构成上同样较为多元（其中评分员男性 4 名，女性 6 名；年龄分布在 22 岁至 45 岁之间）。在统一接受培训、熟悉评分细则后，评分员对写作样本进行独立评阅，以确保评分过程的科学性与结果的可靠性。

（三）研究工具

量表有效性问卷设置了五个评价题目，分别为“评价维度能够很好地反映学习者的写作能力”“各维度的分级描述语足够清晰”“能够更有效地对课堂写作进行评价和反馈”“能够提高评分的客观性和一致性”以及“能够简化评分过程”。每个题目采用四级评分制，从“1=非常不同意”到“4=非常同意”。

本研究使用 SPSS 24.0 作为主要分析工具，用于评价量表一致性的分析，并对评分差异进行独立样本 *t* 检验。

（四）研究步骤

本研究分为两个主要阶段。第一阶段，根据汉语二语教学与测试相关大纲文件中关于“写作能力”的描述，结合相关文献分析以及评分员一组对写作文本的评阅结果，提取评价维度并收集相应的描述语。在此基础上，设计初始评价量表并确定各维度的等级划分。

第二阶段旨在对评价量表进行验证。首先，邀请 7 位具有多年教学经验的汉语教师对初始评价量表的效度进行判断，并提出修改建议。研究人员根据这些建议对量表进行调整与完善。然后，由 10 位评分员使用修改后的评价量表对六篇不同水平的写作文本进行评分。研究人员利用 SPSS 对评分结果进行一致性分析，以验证量表的信效度。

三、评价量表的构建

（一）评价维度的确定

理想情况下，评价量表的各维度应与语言能力模型相一致。然而，现有语言能力模型在二语写作评价中的实际应用仍面临诸多挑战（Knoch, 2009）。因此，本研究借鉴 Jin & Mak（2013）的研究，对《汉语水平等级标准与语法等级大纲(1996)》《国际汉语能力标准(2007)》《国际汉语教学通用课程大纲(2008, 2014 r.v.)》及《国际中文教育中文水平等级大纲(2021)》等汉语二语教学与测试大纲进行了梳理，从中提取出普遍认可的写作评价特征（见表 1）。在实际文本分析过程中，我们发现以下问题尚需进一步明确：一是如何界定词汇得体性及其与准确性的界限；二是写作规范的特征及其是否应纳入评价维度；三是修辞手法在写作评价中的作用；四是内容层面的评价标准。本研究将对此深入探讨。

表 1 大纲中写作能力的评价特征

评价维度	评价特征	1996 大纲	2007 标准	2008/2014 大纲	2021 等级标准
汉字	准确性	√			√
词汇	准确性	√	√	√	√
	复杂性	√			√
	得体性	√	√	√	√
句子	准确性	√	√	√	√
	复杂性	√		√	√
流利性	写作速度	√			√
篇章结构	语篇连贯	√	√	√	√
写作规范	格式规范	√	√	√	√
内容	叙事清晰	√	√		√
	观点明确	√	√	√	√
	论据充分	√	√	√	√
	创造性			√	
	新颖性			√	
	完整性				√
修辞	复杂性			√	√

关于词汇得体性，Hymes（1972）将其视为交际语言能力的重要组成部分，强调语言运用需符合特定语境。《中国英语能力等级量表》亦将得体性纳入语用范畴，涵盖语体、语域与文化参照等方面（韩宝成、黄永亮，2018）。然而，在二语写作中，词汇得体性缺乏公认的操作性定义，难以建立明确、可量化的评价标准，且对评分者的语用判断力与专业素养要求较高。尤其在中高级阶段，学习者虽具备一定的词汇积累，但得体性作为高阶语用能力，

往往超出常规课堂写作训练的目标与能力范围。因此，处于可评估性与稳定性的考虑，本研究暂不将词汇得体性纳入评价体系。

在写作规范方面，英语写作评价中通常通过拼写、大小写、标点、缩进和段落数量等量化指标加以体现（Kennedy & Thorp，2007）。相关研究指出，规范性在学术和商务写作中尤为关键，甚至影响语言的交际功能（徐昉，2013）。不过，吴雪峰等（2018）认为，“写作规范”可作为补充性维度，根据具体写作任务灵活调整。考虑到本研究对象为中高级汉语二语学习者，其已基本掌握写作规范，且部分规范（如段落与标点）已在其他维度中间接覆盖，故本研究亦不单独设定“写作规范”作为评价维度。

关于修辞手法，《国际中文教育中文水平等级标准》等大纲建议高阶写作适度使用修辞。然而，实际研究发现，大多数二语学习者在写作中较少运用修辞，这可能与教学中对修辞训练的系统性不足有关（侯颖，2012）。此外，中高级写作教学普遍更侧重词汇准确性、句法复杂性与篇章结构，修辞能力则被视为进阶但非核心的技能。基于对学习者语料的初步分析，本研究未将修辞手法纳入主要评分维度，以避免评分标准过于主观。

关于是否应将内容纳入写作能力评价，学界长期存在分歧。祝秉耀（1984）和张宝林（2009）认为，汉语二语写作教学应侧重书面语表达能力，因此写作的重点应落在语言表达形式上，内容仅需与题目相关即可。然而，近年来多项研究日益强调语言形式与内容质量的综合评价（辛平，2007；Kuiken & Vedder，2017），认为内容是衡量写作完成度与交际效果的重要维度。文秋方（2007）进一步指出，内容评价可涵盖话题展开、逻辑性和任务完成度，并应根据文体设定标准。本研究在此基础上，结合学习者语料，对内容维度进行了简化，聚焦记叙文和议论文两种常见体裁，并从立意明确性（记叙文）、叙事清晰性（记叙文），任务相关性和完成度等方面进行界定，以增强评价的操作性与适切性。

综上所述，基于相关大纲要求、现有文献和理论框架，并结合本研究对象的实际写作表现，最终确定了本研究写作评价量表的评分维度与核心特征（见表 2）。

表 2　评价量表的评价维度与特征

一级维度	二级维度	三级维度
语言要素	词汇	·复杂性 ·准确性
	语法	·复杂性 ·准确性
	汉字	·准确性
	标点符号	·准确性
篇章结构	组织	·清晰性
	衔接	·连贯性
作文内容	内容	·立意明确（记叙文） ·叙事清楚（记叙文） ·论点明确（议论文） ·论据充分（议论文）

	任务完成情况	·相关性 ·任务完成度

（二）等级描述语的生成

生成等级描述语需要解决两个关键问题。首先是等级划分。研究表明，评分标准将等级划分设定在 5 至 9 个之间时信度较高（Myford, 2002）。North（2003）认为，等级划分应兼顾区分学习者能力水平和操作可行性：既要足以体现学习者的进步，又不宜过多，以免增加评分难度。针对特定群体的写作评分，Knoch（2011）指出，当评价对象集中于单一能力水平时，三至四个等级已足够。基于本研究聚焦中高级汉语学习者，初步将各评价维度划分为四个等级，并通过相应的描述语加以细化。

其次是等级描述语的语言风格与表达策略。描述语的具体性与客观性不仅是提高评分信度的关键因素，也为教学反馈提供有力支持（Knoch，2011）。描述语的来源包括文献、采样和独立撰写三种方式。文献法通过查阅语言能力标准、教学大纲及评分标准等材料，建立描述语库；采样法则通过向教师、学习者和评分员征集描述语；独立撰写则由专家或有经验教师根据实际需求编写描述语。

本研究结合文献法和评分员撰写生成描述语。在文献法的运用过程中，为确保评价量表构建的科学性和系统性，本研究除参考了四部具有代表性的汉语二语教学大纲外，还采集了两种语言能力量表、十种二语写作评价量表及一部阅卷员培训手册，重点关注其中针对中高级写作能力的描述。[1] 这些文献均在二语写作领域具有代表性与权威性，并且对中高级水平写作技能的培养目标、内容要求和评价指标有相对完整的阐述。通过比对这些文献中不同等级写作能力的描述与分析，本研究得以从宏观层面梳理适用于中高级写作水平的核心指标，为后续评价量表的构建提供了系统而可比的依据。研究人员对质量较高的描述语予以保留，对不符合本研究需要的内容进行改写，并将英文描述语翻译为中文。最终，共收集 135 条描述语。与此同时，评分员一组通过作文评价实践又补充撰写了 20 条描述语，形成总数为 155 条的描述语库。

无论描述语来源如何，其内容与形式均需符合规范性要求（方绪军、杨惠中，2017）。North（2000）在《欧洲语言共同参考框架》（CEFR）研究中提出，描述语应具备以下特征：从正面描述学习者的能力，避免负面表达；内容应具体明确，避免模糊或歧义；语言应清晰易懂，不使用晦涩术语；表达应简洁明了；内容具有相对独立性，避免互相依赖。这些原则对于评价量表的描述语同样适用。

基于上述要求，本研究对所采集的描述语进行了系统处理，包括参数标定、重复内容合并、表述风格同一等操作。最终依据四个等级划分和学习者的实际写作表现，完成第一版评价量表的等级描述语设计。

[1] 本研究除参考《欧洲语言共同框架》《中国英语能力等级量表》《ACTFL 写作能力等级描述》《IELTS 写作评分标准》《TOEFL iBT 写作评分标准》《剑桥英语写作评估标准》《大学英语四、六级写作评分标准》《高校英语专业四、八级写作评分标准》外，还采集了《ESL composition profile》《TEEP attribute writing scales》（Weigle, 2002）、《汉语水平考试（HSK）写作评分标准》（聂丹，2009）、《诊断性写作评分量表》（史剑雄、王佶旻，2022）以及《汉语考试阅卷员培训手册》等。

四、信效度检验

（一）效度验证

通过专家及有经验的教师对评价量表进行构念效度验证，是一种常用且有效的方法。本研究邀请七位具有丰富教学经验的汉语教师组成专家小组，参与量表验证与修订工作。专家小组收到的材料包括：第一版评价量表、一份评价量表有效性问卷，以及 8 篇依据量表评分的作文样本。专家根据材料填写问卷，从五个方面评估评价量表的效度，并提出改进建议。

本阶段的修订并非一次性完成，而是在多轮反馈与调整的基础上持续推进的。专家小组在多次讨论后，对评分维度及其描述语达成共识，并最终完成有效性问卷填写。专家反馈的重点涉及评分复杂度与准确度是否一致、描述语的独立性与必要性等问题，这些意见均已在量表设计中予以回应。

根据专家小组填写的问卷结果，各题项的平均数与标准差如表 3 所示。除“T5”外，其余四个题目的均值均高于 2.5，表明专家对评价量表的效度具有较高认可。然而，在“简化评分过程”方面，部分教师提出分项式评分需逐一审阅各维度，相较整体式评分更加耗时费力；Knoch（2007）也曾指出，分项式评价量表在实际使用中往往需要投入更多时间与精力。不过，也有教师反映，尽管初期操作较为繁琐，随着熟悉程度的提升，评分效率会逐步提高。

表 3 评价量表效度统计

题项	均值	标准差
T1：很好地反映学习者写作能力	3.714	0.452
T2：分级描述语足够清晰	3.571	0.495
T3：可以更好地对课堂写作进行评价和反馈	3.857	0.350
T4：可以提高评分的客观性和一致性	3.286	0.452
T5：可以简化评分过程	2.428	0.495

此外，本研究还发现，不少教师在以往的作文评价中更多依赖直觉或经验进行整体性评改，或仅作简单纠错反馈。相较而言，分项式评分对他们而言操作更为复杂。因此，如何提升分项式评价量表的评分效率，将是后续研究的重要方向。

综上所述，本研究构建的评价量表具有较高的构念效度。经吸收专家小组意见并多次修改与完善后，形成了最终的评价量表（见表 4）。

表 4　中高级汉语二语写作评价量表

一级维度	二级维度	三级维度	分级描述语
语言要素	词汇	复杂性	4 使用丰富的高级词汇（如成语、习语） 3 使用中高级词汇及部分固定词组，少量重复 2 主要使用基础与部分中级词汇，有一定重复 1 使用的多为基础词汇，并频繁重复
		准确性	4 词语使用准确，仅有个别错误 3 较为准确，有少量词汇错误 2 基本准确，有较多词汇错误 1 有大量词汇错误
	语法	复杂性	4 使用多种复杂句式（如多重复句），形式多样

			3 可使用较长或复杂的关系复句和特殊句式，形式较丰富 2 仅能使用简单句式和简单复句（如因果复句），形式单一 1 多为简短句式，缺乏多样性
		准确性	4 句式结构准确，仅个别语法错误 3 较为准确，有少量语法错误 2 基本准确，有较多语法错误 1 有大量语法错误
	汉字	准确性	4 书写准确，仅有个别错别字 3 较为准确，有少量错别字 2 基本准确，有较多错别字 1 有大量错别字
	标点符号	准确性	4 标点使用准确，仅有个别错误 3 较为准确，有少量标点错误 2 基本准确，有较多标点错误 1 有大量标点错误
篇章结构	组织	清晰性	4 段落安排合理，层次清晰，结构完整（开头点题、结尾有效总结主题等） 3 结构较为清晰，开头和结尾基本完善，中间层次略显不足 2 基本合理，但可能缺少开头或结尾，层次不够明晰 1 缺乏有效分段，结构不完整
	衔接	连贯性	4 善用多种衔接手段（并列、递进、连词、代词等），行文流畅 3 使用衔接手段增强连贯性，但种类有限，整体流畅 2 仅能使用简单衔接（连词、代词），有重复或漏用 1 仅能使用少量连词连接句子，文段连接松散。
作文内容	内容	（记叙）	4 叙事完整、主题明确，能用细节或心理描写突出主题 3 叙事较清晰，主题明确但细节不足 2 能大体叙述事件，主题不够充分 1 叙述简单，内容单薄，主题不明
		（议论）	4 观点明确、论据充足，逻辑清晰 3 观点较明确、论据较充分，个别论点展开不足 2 论据有限，论据支撑不足 1 仅提出少量论点，基本无论证
	任务完成情况	任务完成度	4 完全完成任务，内容充分回应题目要求 3 基本完成任务，内容回应较完整 2 部分完成任务，存在偏离 1 未完成任务，明显偏离题目要求

（二）信度验证

评价量表的信度通常可分为评分者间信度与评分者内信度。在教育测量学与统计学领域，信度历来被视为评价工具科学性的核心指标（Bachman & Palmer, 1996）。其中，评分者间信度指当两名或多名评分者依据相同标准对同一表现进行评分时，其评分结果的一致程度，反映了评价量表的基本保障和基础特征；评分者内信度则指同一评分者在不同时间对同一表现进行评分时的一致性。

为检验本研究所设计评价量表的信度，研究团队采用了如下三步流程：首先，研究者根

据评价量表对所有作文样本进行初评，并从中选取两组具有代表性的样本，分别标记为评分样本 A（8 篇作文，包括 4 篇记叙文、4 篇议论文，附评分）与评分样本 B（6 篇作文，包括 3 篇记叙文、3 篇议论文，未附评分）。这两组样本在各维度的初评得分上均存在显著差异。其次，评分员以评分样本 A 为练习素材，熟悉量表各评分维度，并在遇到疑问时与研究者讨论。研究者将讨论结果进行汇总后统一反馈给所有评分员。随后，评分员依据量表对评分样本 B 进行独立打分并提交结果。最后，研究者收集评分数据，采用肯德尔和谐系数（Kendall's W）分析多人评分的一致性，并以独立样本 *T* 检验比较教师评分员与学生评分员之间的评分差异。

在此过程中，尽管本研究涉及记叙文与议论文两种文体，研究在评分时并未对二者进行区分，而是统一纳入一致性分析。分析结果显示，文体差异并未显著影响评分员的一致性，说明量表适用于不同文体的写作评价。整体评分的肯德尔和谐系数达到 0.886，显著高于信度判断阈值 0.8（见 Bachman & Palmer, 1996），表明量表具有良好信度。进一步分组分析发现，教师评分员（W=0.891）与学生评分员（W=0.908）的肯德尔系数均高于 0.8，进一步验证了评分一致性。

与此同时，教师评分员的平均评分（*M*=30.133，*SD*=0.850）略高于学生评分员（*M*=28.667，*SD*=2.695），这一差异可能源于教师评分员更倾向于从整体脉络出发，对内容表达做出略高评价。然而，独立样本 *t* 检验（*t*=1.161，*p*=0.198>0.05）显示，两组评分之间评分差异并不显著，表明在统一量表的指导下，即使评分者的背景与经验存在差异，评分结果依然表现出较高的一致性。

信度分析结果显示，本研究所构建的评价量表能够有效提高评分一致性并减少评分偏差。即使评分员缺乏教学经验，在量表指导下也能实现较高的评分准确性与一致性。此外，记叙文与议论文两类文体差异在评分一致性上并无显著差异，进一步验证了该量表在不同文体课堂写作评分中的普适性与信度稳定性。综上所述，本研究设计的量表在实际评分中的信度表现良好，不仅具备良好的可行性，也体现出其设计的科学性与实用性。

表 5 评分员间一致性统计

评分员	肯德尔系数	平均值	标准差
评分员二组	0.886	29.400	2.035
教师评分员	0.891	30.133	0.850
学生评分员	0.908	28.667	2.695

五、结语

近年来，评价量表在二语写作教学中的应用日益广泛，已被广泛用于纠错反馈（Ene & Kosobucki, 2016）、学生自评与同伴互评（Wang, 2014）等教学环节。本研究旨在为中高级汉语二语学习者设计一套具有实用性与诊断性的写作评价量表，既作为有效的作文评分工具，又可帮助学习者更清晰地认识自身写作能力，促进自评与互评的开展，推动教学、学习与评

价的一体化融合。尽管如此，本研究仍存在一些局限性。首先，对作文文本特征的分析与统计不足，内容维度的描述语尚不够全面，未能全面覆盖写作任务中的语义展开与逻辑组织。其次，当前量表尚未涵盖初级学习者的写作特点与需求，因为难以适用于更广泛的教学对象。此外，评分流程相对复杂，对评分者的操作能力提出较高要求，从而影响了其在教学实践中的推广效率。

为此，未来研究应在文本特征分析与统计方面进一步深入，充分利用语料库及其他自然语言处理工具，对词汇、句法结构与篇章组织等指标进行更系统的定量与定性分析。一方面，这有助于揭示不同语言水平学习者的真实写作特征；另一方面，也能为量表的修订与完善提供更坚实的数据基础。其次，内容维度的描述语应更加丰富和多元，尤其应根据不同文体（如记叙文、说明文、议论文等）的特点制定相应的评价标准，以提升量表在多种写作情境下的适用性。第三，为扩大量表的受众范围，应充分考虑初级学习者的写作需求，关注其常见的语言错误类型与写作思维模式，通过适当的调整或补充，使初级学习者亦能从量表中获得有效支持。第四，在评分操作层面，应简化量表的评分流程，例如开发更直观的评分指南或在线评分工具，让教师和学习者在使用过程中更便捷高效，从而提升教学实践的可操作性与适用性。与此同时，研究还应尽可能扩大样本量，纳入来自不同背景、不同语言水平的学习者作文，以增强量表的普适性与信度，并为量表的持续修订与信度验证提供更全面的实证支持。最后，随着自动化技术的不断发展，将评价量表与自动化评改技术相结合也是下一步的重要方向，通过将量表嵌入智能评改系统，可实现对写作过程与结果的即时反馈，从而为课堂教学与个性化指导提供即时支持，并在此基础上迭代优化量表设计，为汉语二语写作教学的效率与质量提升奠定更加坚实的理论与技术基础。

参考文献

白丽茹. (2012). 大学英语写作中同伴互评反馈模式测量评价表的编制. *现代外语, 35*(2), 184–192+220.

方绪军, 杨惠中. (2017). 语言能力等级量表的效度及效度验证. *外国语（上海外国语大学学报）, 40*(4), 2–14.

国家对外汉语教学领导小组办公室汉语水平考试部. (1996). *汉语水平等级标准与语法等级大纲*. 高等教育出版社.

国家汉办/孔子学院总部. (2014). *国际汉语教学通用课程大纲*. 北京语言大学出版社.

国家汉语国际推广领导小组办公室. (2007). *国际汉语能力标准*. 外语教学与研究出版社.

韩宝成, 黄永亮. (2018). 中国英语能力等级量表的研制——语用能力的界定与描述. *现代外语, 41*(1), 91–100+146–147.

侯颖. (2012). 汉语国际教育中的修辞教学. *海外华文教育*, (1), 17-23. https://doi.org/10.14095/j.cnki.oce.2012.01.005

教育部国家语委. (2021). *国际中文教育中文水平等级标准*.

李清华. (2014). *高校英语专业四级测试写作评分标准的设计与效度研究*. 科学出版社.

罗晓杰, 曾家延, 董方圆. (2010). 英语学科作文分析式评分量规及其研制. *课程・教材・教法, 30*(12), 61–65. https://doi.org/10.19877/j.cnki.kcjcjf.2010.12.015

马晓梅, 史晓婷, 芦畅, 李荣. (2022). 基于课堂测评的英语写作同伴互评量表研发及验证. *西安外国语大学学报, 30*(1), 56-62. https://doi.org/10.16362/j.cnki.cn61-1457/h.2022.01.010.

聂丹. (2009). 汉语水平考试(HSK)写作评分标准发展概述. *云南师范大学学报（对外汉语教学与研究版）, 7*(06),15–20. https://doi.org/10.16802/j.cnki.ynsddw.2009.06.003

史剑雄, 王佶旻. (2022).诊断性写作测试评分量表的构建研究. *语言文字应用*, (03), 114–123. https://doi.org/10.16499/j.cnki.1003-5397.2022.03.001.

王丽. (2016). *商务英语写作能力量表的开发与效度研究*(博士学位论文). 上海交通大学. https://doi.org/10.27307/d.cnki.gsjtu.2016.004839.

文秋芳. (2007). “作文内容”的构念效度研究——运用结构方程模型软件 AMOS 5 的尝试. *外语研究*, (3), 66–71+112.

吴雪峰. (2018). *概要写作评分量表研究：开发研制与效度验证*(博士学位论文). 上海外国语大学.

辛平. (2007). 基于语言能力构想的作文评分标准及其可操作性研究. *暨南大学华文学院学报*, (3), 19–24+42. https://doi.org/10.16131/j.cnki.cn44-1669/g4.2007.03.012.

徐昉. (2013). 学习者英语学术写作格式规范的认知调查报告. *外语教学, 34*(2), 56–60. https://doi.org/10.16362/j.cnki.cn61-1023/h.2013.02.021.

张宝林. (2009). “汉语写作入门”教学模式刍议. *语言教学与研究,* (3), 54–59.

祝秉耀. (1984). 浅谈汉语写作课. *语言教学与研究,* (1), 96–105.

邹绍艳, 范劲松. (2022).大学英语四级考试写作测评量表效度研究. *外国语文, 35*(3), 148–156.

邹绍艳. (2017). *大学英语四级写作测试分项评分量表的制定及其效度研究*(博士学位论文). 上海交通大学. https://doi.org/10.27307/d.cnki.gsjtu.2017.000035.

Andrade, H. L., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research, 102*(4), 287–302. https://doi.org/10.3200/JOER.102.4.287–302

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88–115. https://doi.org/10.1016/j.asw.2009.04.001

Ene, E., & Kosobucki, V. (2016). Rubrics and corrective feedback in ESL writing: A longitudinal case study of an L2 writer. *Assessing Writing, 30*, 3–20. https://doi.org/10.1016/j.asw.2016.06.003

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin Books.

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work?. *Language Testing, 30*(1), 23–47. https://doi.org/10.1177/0265532212442637

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kennedy, C., & Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS academic writing task. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in Speaking and Writing Assessment* (pp. 316–378). Cambridge University Press.

Knoch, U. (2007). The development and validation of an empirically-developed rating scale for academic writing [Doctoral dissertation]. *University of Auckland*.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in*

Education, 15(2), 187–215. https://doi.org/10.1207/S15324818AME1502_04

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph, 24.* Educational Testing Service.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 100329. https://doi.org/10.1016/j.edurev.2020.100329

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Stevens, D. D., & Levi, A. J. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus Publishing.

Wang, W. (2014). Students' perceptions of rubric-referenced peer feedback on EFL writing: A longitudinal inquiry. *Assessing Writing, 19*, 80–96. https://doi.org/10.1016/j.asw.2013.11.008

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

（责任编辑：韩瑞宝）