Advances in Education 教育进展, 2026, 16(7), 378-385
Published Online July 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.1671379

Hans汉斯

英语写作智能评估工具的比较研究

——以批改网与DeepSeek为例

郭艳宾

新疆大学外国语学院，新疆 乌鲁木齐

收稿日期：2026年6月5日；录用日期：2026年7月6日；发布日期：2026年7月13日

摘　要

本研究旨在比较传统自动写作评估工具“批改网”与新兴大语言模型“DeepSeek”在评估英语专业学生作文档次时的评分效度。本研究直面一线外语教学中写作评估“高需求–低反馈”的核心痛点，以113篇英语专业二年级学生的作文为样本，以TEM4分项评分标准为效标，采用相关分析与典型案例分析法，对比了批改网与DeepSeek在作文评分中的效度。研究发现：(1) DeepSeek总分与教师总分呈强正相关($r = 0.890, p < 0.01$)，而批改网仅为中等相关($r = 0.429, p < 0.01$)；(2) DeepSeek总分与教师内容分高度相关($r = 0.827, p < 0.01$)，批改网与教师内容分相关性较弱($r = 0.292, p < 0.01$)，与教师语言分为中等相关($r = 0.552, p < 0.01$)；(3) 典型案例分析进一步揭示，批改网无法识别跑题作文等根本性内容问题，其反馈集中于词汇、句法等表层语言特征，而DeepSeek在各分数段上均与教师判断更为一致，其反馈能够触及论证逻辑、结构层次等内容层面。上述发现表明，DeepSeek在写作评估的整体效度及对内容维度的把握上均显著优于批改网，为教师在技术选择和教学设计上提供了清晰的决策依据。

关键词

二语写作评估，批改网，DeepSeek，评分效度

A Comparative Study of Automatic English Writing Assessment Tools

—A Case Study of “Pigai” and “DeepSeek”

Yanbin Guo

College of Foreign Languages, Xinjiang University, Urumqi Xinjiang

Received: June 5, 2026; accepted: July 6, 2026; published: July 13, 2026

文章引用: 郭艳宾. 英语写作智能评估工具的比较研究[J]. 教育进展, 2026, 16(7): 378-385.
DOI: 10.12677/ae.2026.1671379

Abstract

This study aims to compare the scoring validity of the traditional automated writing assessment tool "Pigai" and the emerging large language model "DeepSeek" in evaluating the quality of English major students' compositions. Addressing the core challenge of "high demand, low feedback" in front-line foreign language writing instruction, this study uses 113 essays written by second-year English majors as the sample, adopts the TEM4 analytic rating scale as the criterion, and employs correlation analysis and typical case analysis to compare the validity of Pigai and DeepSeek in essay scoring. The findings reveal that: (1) DeepSeek scores show a strong positive correlation with teacher scores ($r = 0.890$, $p < 0.01$), while Pigai shows only a moderate correlation ($r = 0.429$, $p < 0.01$); (2) DeepSeek scores are highly correlated with content scores given by teacher ($r = 0.827$, $p < 0.01$), whereas Pigai shows a weak correlation with content scores given by teacher ($r = 0.292$, $p < 0.01$) and a moderate correlation with language scores given by teacher ($r = 0.552$, $p < 0.01$); (3) typical case analysis further reveals that Pigai fails to identify fundamental content issues such as off-topic essays, with its feedback primarily focusing on surface-level language features like vocabulary and syntax, while DeepSeek demonstrates greater consistency with teacher judgments across all score levels, with its feedback addressing content dimensions such as argumentation logic and structural organization. These findings indicate that DeepSeek significantly outperforms Pigai in both overall scoring validity and the assessment of content quality, providing a clear empirical basis for teachers in technology selection and instructional design.

Keywords

L2 Writing Assessment, Pigai, DeepSeek, Scoring Validity

1. 研究背景

写作评估在英语写作教学体系中是一个核心环节。然而，在传统的人工批改模式下，多数教师因为繁重的教学任务与模式耗时较长等因素限制，对学生的作文批改往往流于形式，进而形成了“高需求–低反馈”的矛盾。这已经严重阻碍了写作教学效率的提升。由于人工智能的迅猛发展，教师开展作文评分的方式也随之变化。从传统纯人工评分逐步借助机器评分再到人工智能参与，这一发展反映了语言能力测评需求的不断提高。作为国内早期上线的英语作文自动批改在线系统，批改网自 2011 年投入使用以来，已累计批改作文接近 10 亿篇。该系统基于云计算技术，通过计算学生作文和标准语料库之间的距离，即时生成作文得分和语言分析结果。新一代通用大语言模型 DeepSeek 秉持“开放、可控的开源治理模式”，依托“国家超算中心与民营 AI 实验室协同运演进化的机制”，催生“技术与社会共享协同演进的涌现效应”，在写作评估方面展现出强大能力，使其成为教师作文批改的得力助手。因此，选取智能化的作文批改工具来破解学生作文批改高需求和教师因繁重教学任务导致反馈不足的矛盾十分紧迫。具体而言，批改网深耕英语写作评估领域，经长期数据积累和算法优化，在语法纠错、评分一致性等方面具有专业优势；而 DeepSeek 作为通用模型，在语义理解、创意写作、多模态处理等方面展现出更强的综合能力。简而言之，批改网更侧重于语言表达规范性的评价，而 DeepSeek 在内容深度、逻辑结构、创意表达等方面能够提供更全面的评估。两种工具的功能差异为探索“人机协同”的写作评估模式提供了重要

契机。本研究拟选取 50 篇学生作文作为研究样本，通过对比批改网的分数与 DeepSeek 的批改反馈，为教师提供智能化的批改支持，从而提升作文批改效率以及教学效果。

2. 文献综述

2.1. 作文评估

写作能力评估的理论与实践经历了从整体性评价向分析性评价的重要转向。早期的整体式评分主要依赖评分者对作文的整体印象给出单一分数，这种方式尽管操作便捷，但其评分标准较为模糊、主观性强、评分者间一致性较低，且难以向教学提供具体、可操作的反馈[1]。

在此背景下，分项式评分标准逐渐成为主流，尤其在高风险、大规模的标准化测试中得到广泛应用。分项评分将写作能力分解为多个可观察、可测量的维度，从而提升了评估的透明度、客观性与诊断价值。中国高校英语专业四级考试(TEM4)评分标准的演变，正是这一趋势的典型体现。李清华[2]通过对比 TEM4 整体与分项评分标准的实证研究发现，分项标准在多个方面显著优于整体标准：在评分者间一致性上，分项标准的肯德尔和谐系数(0.746)高于原整体标准(0.673)；在评分者内部一致性方面，分项标准下无非拟合项目，而整体标准中存在 1 项；更重要的是，93%的评分者认为分项标准能为写作教学提供更详细的反馈信息。目前 TEM4 采用的分项评分标准将写作能力划分为“内容”(10 分)、“结构”(3 分)与“语言”(7 分)三个维度。其中，“内容”维度关注思想切题性、观点清晰度与论证充分性，被视为写作能力的核心与灵魂[2]。

然而，内容维度的评估恰恰最具挑战。相较于“语言”(语法、词汇)与“结构”(衔接、组织)等相对客观的维度，“内容”评估涉及对思想深度、逻辑严密性及论证有效性的判断，要求评估者具备深层的语义理解与高阶认知处理能力[1]。这种内在的主观性与认知复杂性，使得内容评估成为检验任何评估工具效度的关键试金石。

2.2. 机器辅助写作评分

以批改网为代表的系统，通过自然语言处理与机器学习技术，实现了对作文的量化分析与自动评分[3]。这类工具显著提升了评估效率，能够以稳定的标准处理大规模文本，并提供即时反馈，为传统写作教学注入了新的技术动力。

然而，随着应用场景的不断拓展，学者们逐渐认识到这类系统在评估效度上的局限性，尤其是它们在内容理解与逻辑判断方面的不足。谢耀晶[4]基于 Coh-Metrix 工具对中国学生 550 篇议论文的分析发现，批改网的评分与句法复杂性指标(如主动词前平均词数)呈现显著关联，但这些指标与人工评分中的内容质量维度却缺乏对应关系。值得注意的是，系统评分与“相邻句子句法相似度”呈负相关($r = -0.300$)，意味着批改网倾向于为句法变化丰富的作文打出高分，而这一形式特征并不能直接反映思想内容的深度与逻辑性。高健民[5]对 104 篇大学英语作文的深入分析进一步揭示了系统的评分倾向。研究发现，批改网评分与人工评分相关性较弱($r = 0.410$)，且系统分数普遍偏高。更重要的是，其评分机制明显倚重词汇多样性($r = 0.764$)与文本长度($r = 0.636$)等可量化的表面特征，而在句法复杂度、语言准确性等维度上则缺乏稳定关联。这表明，当前系统仍停留于对语言形式的表层捕捉，尚未建立起对写作内容实质的有效评估能力。

2.3. 人工智能辅助作文评分

随着生成式人工智能的突破，以 ChatGPT、DeepSeek 为代表的大语言模型开始被探索用于写作评估与反馈。与基于规则和特征匹配的传统批改系统不同，以 ChatGPT、DeepSeek 为代表的大型语言模

型，为写作教学带来了更具“对话感”和“思想性”的智能支持。这类模型能基于对上下文的理解，不仅指出语言错误，更能评估内容的逻辑性、观点的深度，并提供具体修改建议，其反馈方式更接近一位经验丰富的写作辅导者。在教学实践中，LLM 的实用性正得到验证。魏爽和李璐遥[6]发现，教师在教学中引入 ChatGPT 后，其反馈能覆盖从基础的语法修正、到中级的句法润色、乃至高阶的观点拓展等多个层面。尤其值得注意的是，学生可以与之进行多轮追问互动，从而获得动态、个性化的指导，这种交互性是传统自动化工具难以实现的。在国内教学语境下，以 DeepSeek 为例的最新探索揭示了新的可能。马睿朵[7]对比了 DeepSeek 与“批改网”在实际高职英语写作教学中的应用。研究发现，批改网的反馈侧重于语言形式的准确性(如拼写、语法和固定搭配)，而 DeepSeek 的反馈则更偏向于写作内容的建构性(如逻辑漏洞的指出、段落衔接的建议、中式表达的优化)。具体而言，DeepSeek 能够提供完整的修改范例，并通过提问引导学生深入思考，在教学中扮演了从“纠错者”到“协作者”的角色转变。

对于一线教师而言，这种差异意味着新的工具选择与角色定位。LLM 能够辅助处理文章结构和思想内容等更耗时的深度反馈，让教师得以从繁重的形式批改中部分解放，更专注于启发思维和组织高阶教学活动。然而，在实际应用中，教师也需注意引导学生批判性使用 AI 反馈，并通过精心设计提示词来获得更贴合教学目标的建议[7]，同时警惕其可能存在的逻辑谬误或理解偏差[8]。

综上所述，批改网评分偏重词汇和句法等表层语言特征[4] [5]，而 DeepSeek 则展现出在文章内容和逻辑深度上进行评价的潜力[6] [7]。然而，这两种工具在同一评分任务中，其评分结果与专业分项评分标准之间的关联尚缺乏实证检验。为深入探讨此问题，并为实际教学中如何选择和使用这两种工具提供依据，本研究围绕以下两个问题展开：

在评价英语专业学生作文时，批改网的自动评分与 DeepSeek 的智能评分，哪一个与教师根据 TEM4 评分标准给出的总分更接近？

这两种工具的评分是否分别与教师对内容和语言的评价有特定的关联？

3. 研究方法

3.1. 研究样本

本研究选取了国内某高校英语专业二年级学生三个平行班在两个常规写作练习中提交的 113 篇有效作文作为分析样本。其中，初始阶段 50 篇，作文围绕指定话题《The Central Tension in Love: Security versus Autonomy》展开；为进一步增强样本代表性与外部效度，后续补充收集 63 篇，为同一课程的另一写作任务《The Role of Responsibility》。所有作文均通过批改网平台提交，保持了真实的教学场景。

3.2. 数据收集

首先，教师评分由研究者本人依据《高校英语专业四级考试(TEM4)写作评分标准》进行。该标准采用分项评分方式，将写作能力划分为内容(满分 10 分)、语言(满分 7 分)与结构(满分 3 分)三个维度。在评分过程中，对每一篇作文都进行了独立审阅，分别给出三个维度的得分，最终汇总为 20 分制的教师总分，以此建立一个稳定的评判基准。

之后，为自动化评分系统评分。在批改网布置作文题目时，评分设置直接采用其系统设定的 20 分制。对于大语言模型 DeepSeek，研究者通过指令进行引导：在输入作文文本前，先提供 TEM4 评分标准的具体描述，并发送作文题目，发出指令“请依据上述标准，对该作文的整体质量进行评分，满分为 20 分”，由此获取 DeepSeek 生成的 20 分制总分，此举旨在使 DeepSeek 的评分任务与人类评分员的认知框架尽可能对齐。

3.3. 评分者信度控制

为保证教师评分的客观性与可靠性，研究者本人依据《高校英语专业四级考试(TEM4)写作评分标准》对全部 113 篇作文进行独立评分，分别给出“内容”(满分 10 分)、“语言”(满分 7 分)、“结构”(满分 3 分)三个维度的分数，并汇总为 20 分制的总分。检测评分者间的一致性，随机抽取 30%的样本(n = 34)，邀请第二位具有英语写作教学经验的评分者(某高校英语专业讲师，教龄 6 年)依据相同的 TEM4 标准进行独立评分。第二位评分者在评分前接受了标准培训，熟悉评分细则和流程，且在整个评分过程中不参考研究者的原始分数。采用组内相关系数(ICC)对两位评分者的分数从内容、结构、语言三个方面进行一致性检验，选用双向混合效应模型、绝对一致类型，结果显示：内容维度的单一度量 ICC 为 0.791 (95% CI: 0.629~0.888)，语言维度为 0.837 (95% CI: 0.698~0.915)，结构维度为 0.756 (95% CI: 0.565~0.870)。其中，ICC 值在 0.75~0.90 之间为“良好”，0.90 以上为“优秀”。本研究中三个维度均达到良好水平，表明两位评分者的评分具有较高的一致性，所以后续分析中以研究者本人的评分作为教师评分基准是可靠的。

3.4. 数据分析

数据分析为定量相关分析为主，质性案例分析为辅。首先针对问题一：使用皮尔逊积差相关分析，分别计算批改网总分、DeepSeek 总分与教师总分之间的相关系数，以评估二者与人类整体判断的一致性。针对问题二：系统比较四组相关系数的强弱模式，分别计算批改网总分与教师内容分、教师语言分的相关系数，以及 DeepSeek 总分与教师内容分、教师语言分的相关系数。此外，将选取在 AI 评分与教师评分上呈现显著差异的 1~2 篇作文作为典型案例，对比分析其获得的自动化反馈文本，从而为量化统计结果提供具体的质性分析。

4. 结果与讨论

4.1. 描述性统计

首先，本研究对三项总分(教师人工总分、DeepSeek 总分、批改网总分)进行了描述性统计分析，以了解其整体分布情况，结果见表 1。

Table 1. Descriptive statistics for total scores assigned by teachers, DeepSeek and Pigai
表 1. 教师人工总分、DeepSeek 与批改网总分的描述性统计

名称	样本量	最小值	最大值	平均值	标准差	中位数
总分(人工)	113	0.000	17.500	14.049	2.539	14.500
总分(批改网)	113	13.000	18.000	16.257	1.009	16.500
总分(DeepSeek)	113	0.000	18.500	13.522	2.266	14.000

如表 1 所示，三项总分在分布上呈现出清晰差异。批改网的平均分(M = 16.257)显著高于教师评分(M = 14.049)，其中位数也更高，这表明批改网在整体评分上存在明显的宽松倾向。相比之下，DeepSeek 的平均分(M = 13.522)略低于教师评分，其评分分布与教师评分更为接近，但整体略为严格。从离散程度(标准差)来看，教师评分的离散度最大(SD = 2.539)，说明教师在区分作文优劣时给出的分数跨度更大；而批改网系统的评分则更为集中(SD = 1.009)，这暗示其评分策略相对保守，区分度稍弱。

4.2. 相关性统计

为回答第一个研究问题，比较了批改网与 DeepSeek 总分与教师总分的整体一致性，相关分析结果见

表 2。

Table 2. Correlations between Pigai, DeepSeek and teacher-assigned total scores
表 2. 批改网、DeepSeek 与教师总分的一致性

	总分
批改网	0.429**
DeepSeek	0.890**

注：*p < 0.05, **p < 0.01。

如表 2 所示，二者与教师总分的一致性存在巨大差异。DeepSeek 总分与教师总分达到了强正相关(r = 0.890)，且结果非常显著(p < 0.01)。这意味着，DeepSeek 在判断作文整体质量上，与教师的观点高度一致。 相反，批改网总分与教师总分呈中等程度正相关(r = 0.429)，但相关性显著低于 DeepSeek。这表明，仅凭批改网的总分，几乎无法可靠地预测教师对作文的整体评价。

为回答问题二，比较了 DeepSeek、批改网作文总分与人工批改的内容、语言分数的相关性，相关性分析见表 3：

Table 3. Correlations between Pigai, DeepSeek and teacher-assigned content and language scores
表 3. 批改网、DeepSeek 与教师批改内容、语言分数的相关性

		内容	语言
批改网	相关系数	0.292**	0.552**
	p 值	0.002	0.000
	样本量	113	113
DeepSeek	相关系数	0.827**	0.764**
	p 值	0.000	0.000
	样本量	113	113

注：*p < 0.05, **p < 0.01。

如表 3 所示，DeepSeek 总分与教师评定的内容维度分达到了强正相关(r = 0.827, p < 0.01)。这表明，当教师认为一篇作文内容出色时，DeepSeek 也极有可能给出高分，两者在“内容好坏”的标准上高度一致。值得注意的是，DeepSeek 总分与教师语言维度分也呈中等至强正相关(r = 0.764, p < 0.01)。这显示它并非只关注内容，其在语言质量上的判断也与教师有相当高的共识。相比之下，批改网总分与教师语言维度分仅呈中等正相关(r = 0.552, p < 0.01)。

4.3. 典型个案对比分析

在高分作文中以一篇教师分数为 17.5 的作文为例，教师给出评语“总结完整准确概括材料矛盾，并明确提中立的观点，责任既塑造也揭示了内在性格；结构完整，词汇句式多变，非常好”；DeepSeek 给出 18 分的高分并指出“完美区分两种观点，语言准确、完整、精炼，抓住了核心比喻‘mold’‘pressure cooker’；立场成熟且有深度‘both shape and reveal, with inherent nature as the foundation’；论证逻辑严密，层层递进；清晰的四段结构，段落功能分明；语言质量很高，词汇丰富且精准，句式丰富…”。相比之下，批改网尽管给出 17 的高分，但其反馈停留在语言表层“有效地采用了多种衔接方法，层次清晰；

作者词汇基础扎实，拼写也很棒；作者句法知识扎实，可适当增加从句的使用。”由此可见，DeepSeek在高分作文的评估上和教师有更高的一致性，两者都对高质量的论证做出肯定，而批改网以语言特征为主，难以发现内容层面的优势。

在中等分数作文中以一篇教师分数为12.5的作文为例，DeepSeek给出与教师完全一致的评分(内容：6；结构：1.5；语言：5)指出学生作文结构存在问题，主体论述部分不充分且不完整，语言用词基本准确，存在个别拼写错误，句子结构基本正确，表达清晰，而批改网给出15.5的评分并评价“应增加文中从句的使用；作者词汇基础扎实，拼写也很棒；文章层次不够分明，缺少组织，需注意文章整体的组织结构。”这一案例进一步印证了批改网的给分宽松的倾向，同时也表明DeepSeek在中分段作文上依然能够保持与教师判断的高度一致。

此外，在低分组案例中，某篇作文严重跑题，与题目要求无关，教师均给予0分，DeepSeek也给出0分，然而，批改网却给出了16.5分的高分，并评价“作者英语功底扎实，可适当调整文中从句数量；作者词汇基础扎实，拼写也很棒；文章层次较为清晰。”这一案例进一步揭示了批改网的缺陷：其评分模型无法识别作文内容与题目的相关性，仅依据语言形式特征(如词汇、句长、篇幅)进行评分，导致对跑题作文产生了严重虚高的评分。

4.4. 讨论

本研究发现，DeepSeek在预测作文整体质量上与教师评分强正相关($r = 0.890$)，而批改网相关性较低($r = 0.429$)。这一结果与马睿朵[7]研究中观察到两类工具各有侧重的结论一脉相承，但将对比推向了更本质的层面——整体效度。它挑战了“专业自动化评分工具在给出总分上必然更可靠”的潜在假设。事实上，高健民[5]的研究早已指出批改网评分与人工评分仅为弱相关($r = 0.410$)，且其评分主要反映词汇和流利度。本研究以英语专业学生为样本，获得了更强有力的证据($r = 0.429$)，表明在要求更高的专业写作语境下，批改网评分与教师判断的脱节可能更为严重。正如何旭良[9]指出批改网分数显著高于人工批改作文分数。相反，DeepSeek表现出的高效度，印证了魏爽、李璐遥[6]对类似大语言模型(ChatGPT)在理解与生成反馈方面潜力的判断，并证明这种潜力可以有效转化为对作文整体质量的精准评估。

相关性数据进一步揭示了两类系统评估逻辑的根本差异，这为它们截然不同的效度表现提供了直接解释。本研究证实，批改网总分与教师内容分弱相关($r = 0.292$)，这与其设计原理密切相关。正如谢耀晶[4]所揭示，批改网评分与一系列句法复杂性指标(如名词短语修饰语数量)显著相关，而这些指标与人工评分的内容质量关联微弱。这说明，批改网的模型本质上是对有限文本特征的加权计算，它擅长捕捉“名词前有多少修饰语”这类形式特征，却无法理解这些修饰语所服务的“论点是否深刻”。即便在语言维度，其相关性($r = 0.552$)与DeepSeek相比较低，这可能意味着，它所依赖的词汇、句法特征库[5]，与教师综合评价语言准确性、得体性及与内容的契合度时所运用的复杂标准，存在一定差距。

而DeepSeek则展现出一种“意义驱动”的评估模式。其总分与教师内容分的强相关($r = 0.827$)是其核心的特征，表明其评分深度关联于对作文思想与逻辑的理解。这与马睿朵[7]发现其擅长“段落语义连贯分析”和“逻辑漏洞补救”的结论完全吻合。更有启发性的是，其与教师语言分的较强相关($r = 0.764$)并非独立存在。DeepSeek是在理解全文意义的过程中，将地道的语言视为有效表达的一部分进行整体评估，而非剥离语境地检查孤立的语言点。这使其在语言评估上也比仅作“句法错误纠错”[7]的批改网更贴近教师的综合性判断。

此外，我们应该重新审视两个工具的角色。鉴于批改网评分效度有限且严重偏向形式特征，不应再将其总分视为权威的学业评价依据，应回归其工具属性，将其用作辅助学生自查基础语言错误(如拼写、基础语法)的“电子校对器”，其反馈需经教师或学生批判性审视。而DeepSeek具有作为高阶思维协作

者的潜力。教师可借鉴[6] [7]的思路，设计提示词，引导学生利用它进行内容深化(如：“请批判我第三个论点的说服力”)、结构优化(如：“为我这篇作文提供一个更佳的逻辑框架”)和语言润色(在具体语境中寻求更地道的表达)，从而将写作反馈的重点从纠错转向思维提升。在此背景下，教师的核心任务不再是提供全部反馈，而是设计融合智能工具的学习流程，教会学生如何与 AI 有效互动以获取高质量反馈，并最终帮助学生整合、判断与内化这些反馈。这正如李清华[2]所强调的，好的评分标准(及评价工具)应能为教学提供积极的反拨效应。

5. 结论

本研究通过对 113 篇英语专业学生作文的评分进行系统对比与分析，发现在英语专业写作评估中，以 DeepSeek 为代表的大语言模型评分效度明显优于批改网这类传统自动化评分系统。具体而言，DeepSeek 总分与教师总分的相关系数达到 0.890 (强相关)，而批改网仅为 0.429 (中等相关)。这一差异表明这两类系统的评估逻辑存在根本区别：批改网的评分主要依赖于对词汇复杂度、句长、篇幅等表层语言特征的量化计算，而 DeepSeek 则能够在大规模语料训练的基础上，对作文的内容相关性、论点清晰度、论证充分性等深层语义特征进行整体性评判。同时，这也在一定程度上反映了未来写作评估技术的发展趋势。

此外本研究也存在几点局限：首先，样本虽涵盖两个写作题目，但仍来自同一所学校、同一年级，样本的代表性有限，可能限制结论的推广范围；其次，本研究主要聚焦于评分效度的量化对比，对于 DeepSeek 所生成反馈内容的质量(如准确性、针对性、可操作性)，以及这些反馈能否真正促进学生修改行为、提升写作能力等问题，尚未进行深入探究。针对上述局限，未来研究可从以下方向展开：一是扩大样本来源，涵盖不同年级、不同院校、不同写作任务的学生作文，以验证本研究结论的稳定性与推广性；二是深入比较不同大语言模型(如 ChatGPT、豆包、文心一言等)在写作评估中的表现差异；三是开展教学干预研究，探究学生如何与 DeepSeek 等工具进行互动，以及这种互动对其写作修改行为与写作能力的实际影响。

综上所述，本研究通过数据揭示了批改网与 DeepSeek 在写作评估中的效度差异，为英语写作教学中智能工具的选择与使用提供了参考。未来研究可在更丰富的教学情境中进一步验证和完善这些发现。

参考文献

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

[2] 李清华. TEM4 写作分项式评分标准与整体式评分标准对比研究[J]. 外语测试与教学, 2014(3): 11-20.

[3] Attali, Y. and Burstein, J. (2006) Automated Essay Scoring with E-Rater® V.2. *Journal of Technology, Learning, and Assessment*, **4**, 1-30.

[4] 谢耀晶. 英语议论文作文质量与句法复杂性的相关关系研究——基于人工评分、句酷批改网和 Coh-Metrix 的分析[J]. 海外英语, 2020(2): 100-102+121.

[5] 高健民. 批改网英语作文自动评分系统评分质量研究[J]. 哈尔滨学院学报, 2021, 42(7): 102-105.

[6] 魏爽, 李璐遥. 人工智能辅助二语写作反馈研究——以 ChatGPT 为例[J]. 中国外语, 2023, 20(3): 33-40.

[7] 马睿朵. DeepSeek 与批改网写作批改效能对比研究[J]. 计算机时代, 2025(7): 62-65.

[8] Bang, Y., Cahyawijaya, S., Lee, N., *et al*. (2023) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *Proceedings of the* 2023 *International Joint Conference on Natural Language Processing* (*IJCNLP*), Nusa Dua, 1-4 November 2023, 675-718.

[9] 何旭良. 句酷批改网英语作文评分的信度和效度研究[J]. 现代教育技术, 2013, 23(5): 64-67.