iWrite 与 DeepSeek 的评分信度与反馈内容对比分析

胡婕妤

武汉纺织大学外国语学院　湖北武汉

【摘要】随着大语言模型性能的不断优化，大语言模型逐渐被教育者用来批改作文和提供反馈，与专门的作文评阅系统 iWrite 相比，其评分信度与反馈性能如何？是否能成为一款可以信赖的评分与反馈工具。为探究此问题，本研究以国内某大学国际合作办学院系中艺术专业大二两个班的 46 篇雅思作文为样本，对比分析 iWrite 与 DeepSeek 的评分信度与反馈内容，以期为教育工作者在选择评分与反馈工具时提供借鉴。

【关键词】iWrite；DeepSeek；英语写作；评分信度；反馈

【基金项目】2021 年度湖北省高等学校哲学社会科学研究项目青年项目（项目编号：21Q110）：动态系统理论视域下的人机反馈对二语写作发展影响的研究；2022 年武汉纺织大学教学研究项目（项目编号：2022JY082）：语料库驱动的体裁分析法在通用学术英语写作教学中的应用研究

【收稿日期】2025 年 10 月 15 日　【出刊日期】2025 年 11 月 15 日　【DOI】10.12208/j.sdr.20250263

Comparison of scoring reliability and feedback content between iWrite and DeepSeek

Jieyu Hu

School of Foreign languages, Wuhan Textile University, Wuhan, Hubei

【**Abstract**】With the constant improvement of the performance of Large Language Models(LLMs), LLMs are gradually employed by teachers to score students' writing and provide feedback for them. Compared with the professional Automated Essay Scoring system such as iWrite, the scoring reliability and the performance of generating feedback of LLMs are unclear. It remains doubtful whether these LLMs can be used as reliable tools for scoring and providing feedback. In order to answer this question, this study conducts a comparative analysis of iWrite and DeepSeek, evaluating their scoring reliability and feedback performance on IELTS writing tasks completed by 46 sophomore Art majors in a Chinese-foreign cooperative university program. It aims to provide some insights into choosing automated scoring and feedback tools for teachers and researchers.

【**Keywords**】iWrite; DeepSeek; English writing; Scoring reliability; Feedback

1 引言

英语写作教学的一大难题就是繁重的作文批改工作。学生稍显薄弱的英语写作水平需要教师对学生的习作做出有针对性的反馈，而国内大班英语教学的课堂模式使得及时和高效的反馈对老师而言压力颇大。再者，作文的评阅工作也是一个主观性较强的工作，作文的评改及分数的客观性受到批改教师的情绪、身体状态的影响。为解决这一问题，作文自动评阅系统（Automated Essay Scoring，以下简称 AES）应运而生，使得大规模考试作文的自动批改成为现实，也为英语写作教学提供了强有力的自动批改反馈工具。国内较早且使用较多的 ASE 有批改网和 iWrite 英语写作教学与评阅系统（以下简称 iWrite）。具有业内领先机评效果的 iWrite 是由北京航空航天大学梁茂成教授团队研发，建立在“基于语法规则的简约模型”和“基于深度学习的统计模型”双核联动之上，依托 iWrite 中国英语学习者语料库，能够对提交作文从语言、内容、篇章结构、技术规范四个维度进行机器智能评阅。最新版本的 iWrite 已提升至三核联动并融入 AI 分析，极大提升了该引擎的纠错性能。

除了 AES 之外，随着 2022 年 11 月 ChatGPT 的横空出世，大语言模型也开始被用来做一些作文批改与反馈工作。2025 年 1 月 20 日国内自主研发的

大语言模型深度求索 DeepSeek 问世，其以优越的性能，较低的训练成本和开放的公众使用权限，在国际上引起了广泛关注（高乔，2025）[1]。有研究表明与国际主流模型相比，DeepSeek 在中文语义理解、学术文本评估和教育场景的适应性方面，具有独特的优势（Guo 等）[2]。国内已有学者对 DeepSeek 在学术文本及中英文作文评分与反馈中进行了探索。据此，本研究想继续沿着前人的步伐，探讨专有英语写作教学评阅系统 iWrite 与通用大语言模型 DeepSeek 对作文的自动评分及反馈的表现，旨在评估两种系统的评分信度、描写它们的反馈特性，以期为教师在教学及学生在自主学习中，依据自己的需求选择何种平台或模型进行评分与反馈提供参考与佐证。

2 文献综述

2.1 iWrite 英语写作教学与评阅系统

国内对于 iWrite 的研究主要及集中在两个方面：利用 iWrite 进行写作教学与反馈的效果研究和 iWrite 自动评分信效度研究。张福慧等（2019）对比分析了 68 名大学英语学习者在 Peerceptive 在线同伴互评平台、QQ 写作互评平台和 iWrite 平台上的自我调节性写作学习效果，发现 iWrite 组在语言层面修改幅度最大[3]。王昕和李钦萌（2023）以 34 名英语专业大学生为研究对象，探究了“iWrite+线上教师反馈+线上同伴反馈”的多元反馈模式的反馈效果，发现学生对 iWrite 系统的反馈吸收率最高[4]。此外，也有学者从 iWrite 系统的评分信效度进行研究。李艳玲和田夏春（2018）以 645 篇“国际人才英语考试”的实考作文为语料，通过一致性方法和一致率算法对比分析 iWrite 2.0 的机器评分和人工评分的信度，并认为“iWrite2.0 机器评分几乎可与人工评分相媲美”（p.75）[5]。马小森（2024）通过分析 iWrite 对 54 名大二学生提交的四六级作文的反馈，及对比人工发评分和机评分数，探析了 iWrite 系统的反馈能力及评分信度，指出 iWrite 反馈的正确率很高，但复杂错误的识别率有待提高；对于评分信度，此研究表明 iWrite 的评分明显高于人工评分，但其评分信度基本可以接受[6]。

2.2 大语言模型用于作文自动评阅

随着大语言模型的出现及其在教育领域的广泛应用，相关研究不断涌现。在 DeepSeek 问世之前，关于大语言模型在语言教学及写作领域的研究主要关注 ChatGPT。国外 Mizumoto & Eguchi （2023）探讨了 ChatGPT 对托福作文进行评分的信度问题，认为 ChatGPT 可以有效地作为一款一致且高效的评分工具[7]。国内殷小娟和林庆英（2024）以 45 篇非英语专业大学生的英语作文为样本，对比分析了 ChatGPT、国内的批改网、iWrite 和冰果这三款 AES 系统，发现 ChatGPT 的评分效度虽不如这三个系统，但可以应用到写作教学的阶段性评价中；此外，此研究还指出 ChatGPT 的反馈优势在于反馈内容[8]。

2025 年初，国内自主研发的大语言模型 DeepSeek 自推出以来收到各界一致好评，国内外学者也开始注意到 DeepSeek 在教育领域的应用研究。英语教育方面，冯庆华（2025）从学习评估、语料分析、译文研究和风格预设四个方面探讨了 DeepSeek 在翻译教学与研究中的创新应用[9]。写作教学方面，张天成（2025）对比分析了经微调的 DeepSeek R1 与 GPT-4o 模型在自动评估奥克兰大学诊断性学术英语测评的写作文本中的表现，经与人工评分员的评分结果对照后，发现两类模型的评分均达到与人类专家高度一致的可靠性水平[10]。

综上所述，以往对于 iWrite 的研究主要聚焦其自身的评分信效度或者反馈促学作用，而有关 DeepSeek 在英语教育领域的研究主要关注翻译，已有写作领域的相关研究是对国外学生的作文文本评估，用的模型为微调后的 DeepSeek（张天成，2025）[10]。在研究 DeepSeek 对国内学生写作文本的评估方面，仅有马睿朵（2025）对比分析了 DeepSeek 和批改网在批改大学英语作文时的效能差异[11]，但此研究并没有分析 DeepSeek 的评分信度。此外，具有业内领先水平且被众多高校使用的 iWrite 系统，还鲜有人将其与 DeepSeek 进行横向对比研究，因此，本研究拟对比分析 iWrite 与 DeepSeek 在国内学生雅思作文评分中的信度及生成反馈的表现。

3 研究设计

3.1 研究问题

本研究拟回答以下两个问题：

（1）iWrite 和 DeepSeek 的自动评分的信度如何？

（2）iWrite 和 DeepSeek 在反馈内容与班级成绩分析方面有何异同?

3.2 研究样本

本研究共收集了 46 篇笔者所教两个班级（国内

某高校国际合作办学艺术专业大二）学生在 iWrite 平台上所提交的雅思真题作文的练笔。题型为雅思作文任务 2 的议论文，话题为“工作的唯一目的是否为了挣钱”，字数为 250 字。在任务截止前，学生可以无限次提交。

3.3 研究过程

笔者首先登录 iWrite 平台，下载两个班级的作文成绩 excel 表和学生提交的最后一版作文，并在 iWrite 平台上对每一位同学的作文进行 AI 分析，将 AI 生成的反馈复制粘贴到 word 文档，以 iWrite（学生姓名）命名保存。

接着笔者将学生的作文文档对应姓名以序号编号，打印出来，分别由笔者和另外一位写作教师进行人工评分。笔者和这位评分教师均具有英语语言文学专业的硕士学位和研究生学历，分别在高校从事英语教学 15 年和 19 年，从事写作教学 8 年和 6 年，参加过由前雅思考官进行的雅思评分培训，对雅思作文评分规则非常熟悉。参照殷小娟和林庆英（2024）[8]，为减少人为因素的影响，本研究取 2 位教师对同一篇作文评分的平均成绩作为人工评分的最终成绩。

最后，笔者让网页版 DeepSeek 对学生作文进行评分。本研究给 DeepSeek 下达的操作指令如下：“附件上传的作文是对雅思作文题目 The only reason for people working hard is to earn money. To what extent do you agree or disagree? Give reasons for your answer and include any relevant examples from your own knowledge or experience. Write at least 250 words.的回应，请根据雅思作文评分标准对其评分并点评。”然后笔者通过附件上传学生作文的 PDF 文档。待 DeepSeek 生成评分及反馈评语之后，笔者分别将评分输入到 excel 表格，并将评语复制粘贴到 word 文档，以 DeepSeek（学生姓名）命名。之后，笔者先关闭之前对话框，再开启新的对话框，输入指令并上传附件得到第二个学生的作文分数和反馈。以此方式，笔者共得到 2 个班 46 个同学的作文分数和 46 份反馈报道。

3.4 数据处理

虽然笔者在 iWrite 平台上选择打分公式为雅思作文，但 iWrite 系统生成的机评分数还是与雅思作文的评级分数形式不一样。因此，笔者首先对 iWrite 评分进行处理：（1）有 AI 评分的，取机评和 AI 评分的均值，或者结合 AI 评分给出的分数区间取值（例如，机评分数是 6.3，AI 评分为 6.0-6.5，那么最终分数为 6.5）；（2）AI 没有提供评分的，对机评分数按照雅思作文分数要求处理。具体处理原则为：对于机评分数中小数位的数值，0.25 以下约等于 0，0.25 以上约等于 0.5，0.75 以下约等于 0.5，0.75 以上约等于 1。教师师评分数和 DeepSeep 评分分数为标准化的雅思作文分数，无需再次处理。

3.5 数据分析

（1）评分一致性分析

根据先前的研究（董艳云，祁昕阳，马晓梅，2024；殷小娟，林庆英，2024）[8,12]，由于本研究样本量只有 46（46<50），笔者首先采用 SPSS 27.0 对三组数据进行夏皮洛-威尔克 Shapiro-Wilk 检验来检验数据是否服从正态分布，以决定是选择皮尔逊 Pearson 相关系数还是斯皮尔曼 Spearman 相关系数来分析三组评分间的一致性，检验结果见表 1。

表 1 三组评分的正态性检验

组别	个数	平均值	标准差	Shapiro-Wilk 检验	
				统计量	显著性
iWrite	46	5.815	.6088	.860	<0.01
DeepSeek	46	5.663	.5874	.834	<0.01
教师	46	5.630	.4647	.873	<0.01

由表 1 可见，iWirte 评分、DeepSeek 评分及教师评分的统计量均呈现出显著性（p<0.01），拒绝零假设（数据正态分布），说明三组评分数据都不服从正态分布。因此，本研究采用斯皮尔曼相关系数分析三组评分的一致性，斯皮尔曼相关系数分析结果见表 2。

根据 Akoglu（2018，p.92）的标准，斯皮尔曼系数为 0.00-0.19 是极弱相关，0.20-0.39 是弱相关，0.40-0.59 是中等相关[13]。表 2 中，iWrite 和教师评分的相关系数是中等相关且相关性显著（p<0.05），

说明 iWrite 的评分和教师的评分趋势比较一致，iWrite 的评分比较接近教师评分。DeepSeek 和教师评分的相关系数是弱相关且相关性不显著（p>0.05），表明 DeepSeek 和教师评分的趋势一致性不明显。也就说，教师评分差的作文，DeepSeek 可能评分好，这可能与 DeepSeek 在对作文进行评分时与教师评分时对评分标准的具体把握不一致有关。笔者比对学生作文样本发现，对于句法和词汇表现优秀，但内容深度和文章内在连贯性弱的作文，DeepSeek 评分要高于教师。

（2）评分差异性分析

由于斯皮尔曼相关系数分析得出 iWrite 评分与教师评分相关性显著但却中等相关，而 DeepSeek 和教师评分相关性弱，本研究继续分析三组评分的差异性。参考（林莉兰，2021；陈曦，胡中峰，2025）[14,15]，比较 iWrite 评分、DeepSeek 评分与教师评分之间的差异性，本研究首先采用 SPSS 27.0 对三组数据进行了配对样本 t 检验（n>30），结果如表 3 所示。

表 2　斯皮尔曼相关系数分析结果

序号	组别	斯皮尔曼相关系数	显著性（双尾）
1	iWrite-教师	.452*	.002
2	DeepSeek-教师	.152*	.313
3	iWrite-DeepSeek	.437*	.002

表 3　iWrite、DeepSeek 和教师评分配对样本 t 检验结果

序号	组别	平均值	标准差	T 值	显著性（双尾）
1	iWrite-教师	.1848	.5996	2.090	.042
2	DeepSeek-教师	.0326	.6617	.334	.740
3	iWrite-DeepSeek	.1522	.6043	1.708	.095

从表 3 可以看出，iWrite 评分与教师评分的差异显著（p<0.05），但差距均值 0.1848 非常小，小于雅思作文评分级别最小单位 0.5，说明虽然 iWrite 评分和教师评分之间有差距，但这种差距比较小，因而，iWrite 评分与教师评分差异不大，评分比较一致。据 DeepSeek 与 iWrite 配对 t 检验的结果可知 t 值是 0.33，p 值是 0.740。根据若 t 值接近零且 p 值较大，则表明评分系统在不同时间或不同评估者间的一致性较高（陈曦，胡中峰，2025，p.65）[15]，说明 DeepSeek 评分组与教师评分组在对学生作文评分时，差异均值比较一致，即 DeepSeek 给分严厉度与教师比较一致。

4 结果与讨论

4.1 评分信度

通过斯皮尔曼相关系数分析和配对样本 t 检验，本研究发现 iWrite 评分和教师评分在组内评分趋势上比较一致，虽然配对样本 t 检验发现 iWrite 评分与教师评分差异较显著，但差距均值 0.1848 不会对雅思作文分数的级别 band 产生影响或影响甚微（升降 0.5 个 band）。因比，本研究认为 iWrite 的评分与教师评分比较一致，iWrite 的评分信度高，这与李艳玲和田夏春（2018）[5]、马小森（2024）[6]的发现一致。

通过配对样本 t 检验，本研究发现 DeepSeek 与教师评分一致性较高，但之前的斯皮尔曼相关系数分析表明 DeepSeek 评分与教师评分相关性弱，说明两者评分不一致。笔者结合学生作文样本推断，这一差异可能是由于 DeepSeek 与教师在学生分数区间的中段评分比较一致，而在两端的极值评分差异性较大造成的。因此，本研究认为 DeepSeek 对作文的评分信度较 iWrite 差，DeepSeek 用于评分时需要人工把关。

4.2 反馈内容

iWrite 的反馈内容主要包括总分、星级评分（从语言、内容、篇章结构和技术规范四个维度）、条目式罗列的概述性评语（从词汇、句式、连贯与流畅、语法、拼写错误等方面），以及对作文逐句的语法纠正性批改。反馈页面左边的工具栏有语言、内容和 AI（新增）三项，分别提供语言错误的分类统计，切题度与连贯度柱状图的计量信息。AI 分析会从作文

的结构、连贯性、对题目的回应和语言表达四个维度对文章的优缺点进行点评，并给予总体评价，部分 AI 分析还给出了类似与雅思作文评分模式的四个维度的单项分和作文总分。所以，如果学生觉得 iWrite 反馈页面初始的评语点评不够细致具体， 可以尝试 AI 分析提供更翔实更全面的反馈。

DeepSeek 的反馈大致可以分为四部分：评分、详细点评、修改建议、总结/示范修改（有些反馈无）。DeepSeek 的作文打分是基于雅思作文评分标准的四项小分（任务回应，连贯与衔接、词汇资源、语法范围与准确性）和总分，其评分从数据呈现模式上更加公正透明，并且每项小分有相应的点评信息。详细点评和修改建议分别从文章思想、论证深度、词汇与句式多样性、连贯与衔接这些方面点评优点与指出不足。此外，对有些段落或全文 DeepSeek 甚至直接提供修改后的版本以供参考，如图 1。

总之，iWrite 在语法修改层面上有更多的优势，提供整句的语法纠错服务，部分伴有元语言解释，更适用于语法基础比较薄弱的同学修改作文中的语法错误，夯实英语基础。而且 iWrite 还有作文不同版本之间的对比功能，可以清晰记录学生英文写作水平的提升轨迹。而对于词汇和句式修改而言，笔者认为 DeepSeek 更具优势。DeepSeek 能够提供具体句子的润色版本，如图 1，和可替代词汇，如图 2。

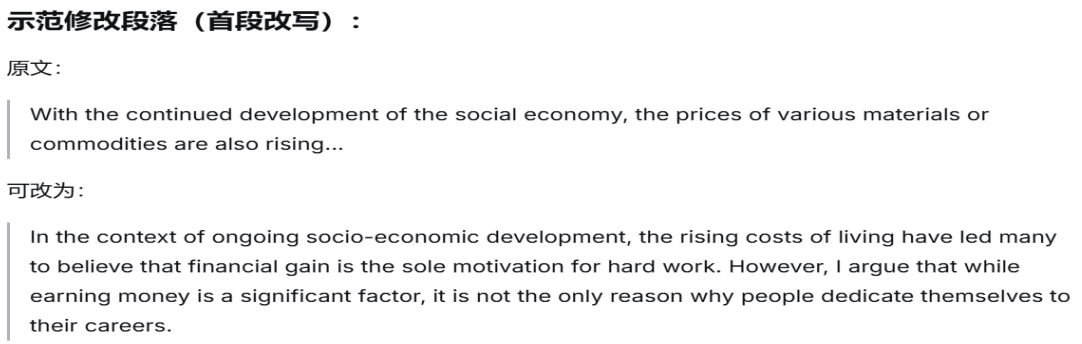

图 1　DeepSeek 句子修改

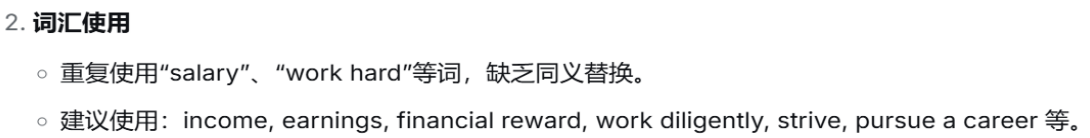

图 2　DeepSeek 词汇反馈

4.3 班级成绩分析

DeepSeek 可以批量上传文件（目前版本 50 个以内）进行批改，通过设定指令，支持对整体成绩以 excel 表格形式输出。通过指令 DeepSeek 可以继续对班级成绩进行整体分析，并从整体表现、高分学生名单（总分高分和四个单项都高分两种统计）、各维度分析，最后提出有操作性意义的总结，例如在本研究中，DeepSeek 指出该班学生写作的核心问题在于语法准确性和词汇使用的精准度，并分别就 5.5 分以下学生和 6.5 分以上学生给出了教学建议。

iWrite 教师页面，有专门的成绩分析和批量批改设置，方便老师进行班级的成绩分析和记录。此外，iWrite 还能对整个班级进行错误类型统计、作文文本客观特征分析和学习情况统计，并支持 excel 表格导出。

整体而言，笔者认为从批量批改、学习成绩记录和管理方面，iWrite 更加方便快捷，但是从班级成绩深度解读、现有写作问题梳理和提供具有可操性的教学建议方面，DeepSeek 的输出更有参考价值，当然 DeepSeek 生成成绩分析报告的质量很大程度上取决于操作者输入的指令。

5 总结与建议

从信度分析可以看出，无论是斯皮尔曼相关系数分析还是配对 t 检验，iWrite 与教师评分的一致性较高，可以作为一款让人信赖的作文自动评分工具供教师使用；配对 t 检验分析显示 DeepSeek 评分与

教师评分的松紧尺度比较一致，但是斯皮尔曼相关系数却表明 DeepSeek 评分在组内高低分顺序趋势与教师评分相关性弱，这说明 DeepSeek 的评分信度值得商榷，教师在选用 DeepSeek 作为作文自动评分工具时，还需把一下关。在反馈内容方面，iWrite 的优势在于提供细致的语言层面的语法纠错功能，而 DeepSeek 在评分的维度和反馈评语的细致性方面要优于 iWrite；其次，DeepSeek 的句子润色和词汇多样性反馈表现优于 iWrite。在班级成绩分析方面，iWrite 胜在方便老师管理成绩和记录学生写作水平变化轨迹，而 DeepSeek 更能提出有针对性和实操性的教学建议。

本研究中，笔者对 DeepSeek 的评分指令并没有经过优化处理，这对 DeepSeek 的评分信度有一定影响。参照董艳云等（2024）测试 GPT-4 对雅思作文任务 2 评估的探索实验[14]，后期研究可以结合指令工程和小样本训练，以提高 DeepSeek 在作文评分任务上的表现，从而可以探究 DeepSeek 和教师评分之间的信度是否可以更加优化。

参考文献

[1] 高乔.中国人工智能创新你何以令海外惊叹 [N].人民日报海外版,2025-02-15(006).

[2] Guo. D.,Yang. D.,Zhang, H., et al. Deepseek-R1: Incentivizing Reasoning Capability in LLms Via Reinforcement Learning[R/OL]. (2025-01-29) [2025-09-15]. https://arxiv.org/abs/2501.12948

[3] 张福慧,李文滔,龙宓吟,高瑛. 基于三个技术平台的自我调节性写作学习效果对比研究 [J].外语电化教学, 2019, (10):22-26.

[4] 王昕,李钦萌.英语专业大学生学术英语写作线上多元反馈模式探索[J]. 外语研究,2023,(4):44-50.

[5] 李艳玲,田夏春.iWrite 2.0 在线英语作文评分信度研究[J].现代教育技术, 2018, 28(2):75-80.

[6] 马小森.AES 系统 iWrite 反馈能力及评分信度研究[J]. 海外英语, 2024,(3):99-101.

[7] Mizumoto, A. & Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring [J]. *Research Methods in Applied Linguistics*, 2023, 2(2): Article 100050.

[8] 殷小娟,林庆英.ChatGPT 与 AES 系统对大学英语写作的反馈效度比较[J]. 闽江学院学报,2024,(3):78-92.

[9] 冯庆华. DeepSeek 在翻译教学与研究中的创新应用[J]. 中国翻译,2025,(2):58-67.

[10] 张天成.大语言模型与人类评分员的对比研究[J].外语测试与教学,2025,(3):31-38,58.

[11] 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.

[12] 董艳云,祁昕阳,马晓梅.基于 GPT-4 的英语写作自动化评估探索----以雅思写作任务 2 为例子[J].语言测试与评价,2024,(2):13-30.

[13] Akoglu, H.User's guide to correlation coefficients[J]. *Turkish Journal of Emergency Medicine*,2018,18(3) : 91-93.

[14] 林莉兰.基于电子档案袋测评的评分者间信度分析报告[J].西安外国语大学学报,2021,29(4) : 67－72.

[15] 陈曦,胡中峰.基于 DeepSeek 的智能评分:效度、信度与可行性研究[J].高教探索,2025,(3):62-67.