人工智能反馈对大学生学术英语写作表现的影响：元分析

Effects of AI Feedback on Students' Academic Writing Performance in Higher Education: A Meta-analysis

罗文筠[1*]，詹颖[1]
[1]香港教育大学课程与教学学系
[*]wyluo@s.eduhk.hk

【摘要】 人工智能 （AI）反馈在英语写作教学中的应用越来越广泛，但是学界对人工智能反馈能否有效促进英语写作教学却一直存在争议。本研究对 38 篇 AI 反馈实证研究进行了三水平元分析，探究其在大学生学术英语写作表现中的效果及影响因素。结果显示，AI 反馈具有中等效应，AI 反馈工具、反馈内容、反馈模式、人机混合反馈中的序列均对 AI 反馈的学习效应有显著影响。本研究结果对大语言模型时代下英语写作教学具有一定的启示，特别是在如何优化 AI 反馈工具的设计和使用方面。
【关键词】 人工智能反馈；大学生学术英语写作表现；三水平元分析；人机混合反馈

Abstract: *Feedback provided by artificial intelligence is increasingly common in English academic writing, yet controversies exist regarding its effectiveness. This study conducted a three-level meta-analysis of 38 valid studies to explore AI feedback's impact on students' academic English writing performance and its influencing factors in the higher education setting. Results indicated a moderate effect size, with significant influences from AI feedback tools, content, mode, and the sequence of human-AI hybrid feedback mode. These findings offer valuable insights for English writing instruction in the AI era, particularly in optimizing the design, implementation, and integration of AI feedback tools into pedagogical practices.*
Keywords: AI feedback, student academic writing performance, three-level meta-analysis, human-AI hybrid feedback mode

1. 引言

生成式人工智能（GenAI，如 *ChatGPT*、*Deepseek* 和 *Copilot*）的兴起进一步推动了 AI 写作反馈的发展。GenAI 基于大型语言模型（LLM），能够生成接近人类水平的文本，为学习者提供实时反馈，帮助其在写作方面取得进步。然而，AI 反馈的效果仍存在争议。自动写作评估(AWE)工具在话语层面（如内容和组织）表现较弱，且缺乏与学习者的互动(Chen & Cui, 2022)；聊天机器人和机器翻译则因对上下文理解不足(Guo et al., 2024)。GenAI 虽然有一定的效果，但也面临着知识转移不足、道德问题以及抑制思辨能力发展的挑战(Fan et al., 2024)。

鉴于现有研究结果的多样性，有必要通过元分析系统量化 AI 反馈对学生英语学术写作表现的影响，以更全面地评估其效果和局限性，为未来的反馈实践提供科学依据。因此，本文提出以下研究问题：1）人工智能反馈对大学生学术英语写作表现的总体效应如何？2）哪些因素会影响人工智能反馈对大学生学术英语写作表现的效应？

本文中 AI 反馈的操作定义基于 Carless 和 Boud (2018)提出的定义进行了改编。AI 反馈的操作定义为：学习者主动寻求、接收并理解由 AI 生成的关于其学习表现的信息，并利用这些信息优化其学习策略的过程。而学术英语写作的操作定义指的是在大学阶段用于评估学生英语能力的写作形式(Zhai & Ma, 2023)。

2. 研究方法

2.1. 文献检索

文献检索于 2024 年 12 月进行。选取的数据库有：Web of Science，Scopus 和 EBSCO。为确保全面地检索相关文献，并未对时间范围和文献类型进行限制。论文筛选的关键词与“AI 反馈”“写作表现”“高等教育”有关。我们还采用了人工检索，范围包括综述类文章及教育技术类期刊。通过以上检索策略，去重后最终检索到英文论文 601 篇。

2.2. 文献筛选

文献筛选标准如下：1）必须是在高等教育语境下使用了 AI 反馈的实证研究 2）必须是组间或组内比较的实验研究设计 3）必须包含计算效应量所需的统计信息 4）必须通过文献质量评估。文献质量评估采用了改编版的 Effective Public Health Practice Project (EPHPP) 工具(Thomas 等，2004；Yan 等，2022)，涵盖了五个维度：参与者选择偏倚、研究设计、混杂因素、数据收集方法以及参与者退出情况。每个维度均按照“强”“中”“弱”三个等级进行评分。若有一个“弱”评级则被评为质量好的研究，两个“弱”为质量中等，超过两个“弱”则为质量较差的研究 (Yan 等，2022)。质量评估由两位在读教育学博士生进行，对于存在分歧的部分，两位评估者通过讨论达成一致。评定为质量较差的研究则被排除。经筛选，纳入文献总计 39 篇。

2.3. 文献编码

本研究对纳入的文献按照以下方面进行编码：1）所使用的 AI 反馈工具 2）是否有对受试进行反馈工具使用的培训 3）反馈类型 4）反馈内容 5）反馈来源 6）人机混合反馈中的序列。编码过程由两位教育学博士生进行，一致性检验为 Cohen’s alpha = .88，对于存在歧义的地方，两位编码人员进行讨论以达到一致意见。

2.4. 数据分析

我们采用 Cohen’s *d* (Cohen, 1988)来计算效应量指标,并将 Cohen’s *d* 转换为 Hedges’ *g* (Hedges, 1981)，因为其能够在小样本研究中提供无偏差的效应量估计(Lipsey & Wilson, 2001)。写作表现的测量通常包括多个变量。因此，一个研究通常会报告多个效应量，而这些效应量之间可能存在依赖性(Cheung, 2014)。本研究采用了三水平元分析因其能够避免分析中的潜在偏差(Ngo et al., 2024)。

除了探究 AI 反馈对学生英语学术写作成绩影响的总体效应外，还对调节变量进行分析。为保证研究结果的稳健性，我们进行了异常值检测。在纳入的文章中，有一篇文章的效应量为 $g = 52.32$ (Rad et al., 2023)，超出了($\bar{x} - 3SD$, $\bar{x} + 3SD$)的范围(Acuna & Rodriguez, 2004)，因此将其移除。最终纳入本元分析中的研究有 38 篇。

3. 研究结果

3.1. 总体效应量

表 1 中呈现了 AI 反馈对大学生学术英语写作的总体效应量。根据三水平元分析的结果，总体效应量为 $g = 0.78$，达到中等效应(Plonsky & Oswald, 2014)。

表 1 效应量及异质性检验结果

	加权效应量			95% 置信区间		异质性						
	k	*g*	*SE*	*Lower*	*Upper*	*Q*	*df*	*p*	τ^2_{level3}	I^2_{level3}	τ^2_{level2}	I^2_{level2}
三水平	221	0.78	0.21	0.36	1.19	3,915.71	220	< .0001	0.67	66.51%	1.46	30.4%

3.2. 调节变量检验

本研究进行了调节变量检验，调节变量的分析结果如表 2 所示。值得注意的是，在 AI 反馈工具、反馈内容、反馈模式及混合反馈中的序列四个调节变量中发现了显著的调节作用。

在 AI 反馈工具中，GenAI 的效应量最大($g = 1.85, p < 0.001$)，其次是 AWE ($g = 0.44, p < 0.01$)，聊天机器人 ($g = 0.04, p > 0.05$), 及机器翻译 ($g = -0.04, p > 0.05$)。GenAI 的表现优于其他三种工具。在不同的反馈内容下，整体反馈($g = 1.95, p < 0.001$)的效应量大于局部反馈($g = 0.56, p < 0.01$)及两者并存的反馈 ($g = 0.82, p < 0.05$)。

在反馈模式中，人机混合反馈的效应量($g = 1.04, p < 0.05$)高于仅使用 AI 工具进行反馈($g = 0.24, p > 0.05$)。由于人机混合反馈的效应量比仅 AI 反馈的效应量更大，我们对人机反馈中的序列进行了进一步的比较。如表 2 所示，序列 AI-人工-AI ($g = 2.72$)的效应量最大，其次是人工-AI-人工($g = 0.91$)，紧接着是 AI-人工($g = 0.76$)，效应量最小为人工-AI ($g = 0.26$)。

该分析还考察了其他调节变量，但未发现有显著的调节作用。提供使用 AI 工具的培训($g = 0.83$，$p < 0.01$)比未提供培训($g = 0.65$，$p < 0.05$)产生了更大的效应量。此外，仅提供评语($g = 1.18$，$p < 0.01$)比混合反馈模式（即评分加评语，$g = 0.51$，$p < 0.05$）的效应量更大。

表 2 调节效应分析

调节变量		**效应数量**	**效应量 [95%置信区间]**	***t* 值**	***p* 值**
AI 反馈工具	GenAI	55	1.85 [1.11, 2.59]	$t_{(3, 217)} = 3.76$	0.01*
	AWE	162	0.44 [-0.002, 0.88]		
	聊天机器人	1	0.04 [-2.6, 2.68]		
	机器翻译 MT	3	-0.04 [-2.34, 2.26]		
给学生提供 AI 工具培训	有提供	164	0.83 [0.33, 1.32]	$t_{(1, 219)} = 0.14$	0.71
	无提供	57	0.65 [-0.15, 1.45]		
反馈类型	评语	91	1.18 [0.57, 1.79]	$t_{(2, 219)} = 2.97$	0.09
	评语+评分	130	0.51 [0.002, 1.01]		
反馈内容	整体反馈	16	1.95 [0.99, 2.91]	$t_{(2, 218)} = 3.83$	0.02*
	局部反馈	143	0.56 [0.09, 1.04]		
	混合	62	0.82 [0.24, 1.41]		
反馈模式	仅 AI 反馈	86	0.24 [-0.13, 0.88]	$t_{(1, 219)} = 6.72$	0.01*
	人机混合反馈	135	1.04 [0.57, 1.45]		
人机混合反馈中的序列	AI-人工	82	0.76 [0.37, 1.14]	$t_{(3, 131)} = 4.14$	0.008**
	人工-AI	17	0.26 [-0.22, 0.74]		
	AI-人工-AI	21	2.72 [0.45. 4.31]		
	人工-AI-人工	15	0.91 [-0.09, 1.74]		

4. 讨论与总结

本研究对 38 项 AI 反馈实证研究进行了元分析。结果显示，AI 反馈的效应值为 0.78，说明 AI 反馈能提高大学生英语写作表现。通过调节变量检验，我们发现四个对总效应量有显著贡献的调节变量，分别是 AI 反馈工具、反馈内容、反馈模式及人机混合反馈中的序列。

该元分析为融合智能(hybrid intelligence)这一概念提供了支撑。融合智能由 Akata 等(2020)提出，通过人类与机器智能的互补优势，以实现那些仅靠人类或机器单独无法完成的任务。本研究中，人机混合反馈对总体效应影响更大，可以归因于人机混合反馈可以达到人工与 AI 之间的意义协商，进而达到融合智能。通过观察人机混合反馈模式下序列之间的差异，我们也可以发现 AI-人工-AI 及人工-AI-人工的这两组迭代序列会比线性序列更加有效。根据 Nguyen 等(2024)的讨论，线性序列倾向于将 AI 仅仅用作信息的补充，缺乏 AI 与学生之间深度互动，从而影响写作效果。

本研究进一步表明 AI-人工-AI 这一序列比人工-AI-人工所产生的作用更大，自我决定理论(Ryan & Deci, 2020)可用于解释这一现象。AI 作为提供即时反馈，帮助学生了解自己写作中表现较好和需要改进的方面，学生可自行决定是否采纳 AI 所提供的反馈，满足其胜任需求和自主需求。随后，人工反馈能够提供情感支持和进一步的指导，满足其关系需求。最后，学生通过 AI 的评估明确看到自己的成长，进一步强化胜任感。若先有人工反馈，学生可能因早期收到主观评价而产生防御心理，从而抑制自主需求及胜任需求。

本研究还发现，整体反馈和混合反馈的有效性较高，而局部反馈效果中等，然而 Ngo 等(2024)发现整体和局部反馈效果显著，而混合反馈效果较弱。差异可能源于 AI 反馈工具的不同。Ngo 等主要使用 AWE 工具，而本研究还采用了生成式 AI、聊天机器人和机器翻译，这些工具在提供局部和整体反馈方面表现更全面。混合反馈可能因同时要求语法纠正和结构建议而导致学生产生认知负荷，而整体反馈更受学生重视，因为其能直接影响文章整体质量。

AI 反馈工具的不同也对总体效应量产生了显著影响，其中 GenAI 的效应量较其他工具更大。GenAI 基于大量语言数据训练，能够通过对话生成反馈，支持迭代改进。相比之下，AWE、聊天机器人和机器翻译依赖预定义规则，缺乏对自然语言的深度理解。

对 AI 工具的使用进行培训和反馈类型这两个变量的调节作用并不显著。然而，有培训的效应量大于无培训。在 GenAI 时代，培养学生如何使用 AI 工具显得尤为重要。此外，反馈类型的效果也显示出差异：仅提供评语的反馈比“评分+评语”反馈更有效。评语能够帮助学生深入理解作品并促进有意义的反思。

综上所述，本研究为 AI 反馈在写作教学中的应用提供了实证支持，并强调了人机协作、迭代反馈序列以及生成式 AI 的重要性。未来研究可进一步探索如何通过系统化培训帮助学生更有效地使用 AI 工具，以及如何优化反馈设计（如结合情感支持和个性化指导）来最大化 AI 反馈的教育价值。此外，在 GenAI 时代，教师的角色将更多体现在与 AI 的协作中。如何适应这一角色转变，通过运用 AI 技术提升自身及学生的反馈素养，也值得进一步探索。

参考文献

Acuna, E., & Rodriguez, C. (2004). *A meta-analysis study of outlier detection methods in classification.* Department of Mathematics, University of Puerto Rico at Mayaguez. https://www.researchgate.net/publication/228728761_A_meta_analysis_study_of_outlier_detection_methods_in_classification

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/mc.2020.2996587

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257–278. https://doi.org/10.1111/j.2044-8317.1988.tb00901.x

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing*, *57*, 100915. https://doi.org/10.1016/j.jslw.2022.100915

Cheung, M. W. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13544

Guo, K., Li, Y., Li, Y., & Chu, S. K. W. (2024). Understanding EFL students' chatbot-assisted argumentative writing: An activity theory perspective. *Education and Information Technologies*, *29*(1), 1–20. https://doi.org/10.1007/s10639-023-12230-5

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.

Ngo, T. T., Chen, H. H., & Lai, K. K. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: a three-level meta-analysis. *Interactive Learning Environments*, *32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

Plonsky, L., & Oswald, F. L. (2014). How big is "Big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. https://doi.org/10.1111/lang.12079

Rad, H. S., Alipour, R., & Jafarpour, A. (2023). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: a case of Wordtune application. *Interactive Learning Environments*, *32*(9), 5020–5040. https://doi.org/10.1080/10494820.2023.2208170

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Thomas, B., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing*, *1*(3), 176–184. https://doi.org/10.1111/j.1524-475x.2004.04006.x

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review*, *37*, 100484. https://doi.org/10.1016/j.edurev.2022.100484

Zhai, N., & Ma, X. (2023). The Effectiveness of Automated Writing Evaluation on Writing Quality: A Meta-Analysis. *Journal of Educational Computing Research*, *61*(4), 875–900. https://doi.org/10.1177/07356331221127300