العدد3 No 3
2026 ايار May2026
مجلة ساوة للدراسات الانسانية
Sawah Journal of Humanities
E-ISSN 3106 - 8324 P-ISSN 3106 - 8316
مجلة محكمة تصدر عن
جامعة ساوة الاهلية

The Impact of Artificial Intelligence-Based Tools on EFL Students' Academic Writing :A Linguistic and Pedagogical Analysis

Assistant teacher Wasan shakir Hamoodi
Wasan. shakir. hamoodi @ec.edu.iq

Abstract

This study examines how three AI-based writing tools, ChatGPT (GPT-4o), Google Gemini, and QuillBot, positively affect the academic development of undergraduate students who are learning to write English as a foreign language (EFL) within the context of Iraqi higher education. The central claim of this study is to show how the use of AI tools can positively affect the development of specific writing conventions used in EFL academic writing (lexical diversity, syntactic complexity and textual cohesion). To ensure that the effects of AI tools on EFL academic writing are preserved as independent, maintain the authenticity of the writers' voices and develop skills independent of reliance on AI tools, the integration of systematic instructional practices is necessary. Using a pre-test/post-test quasi-experimental/mixed-methods design to examine the effects of AI tools on EFL academic writing over twelve weeks with 100 undergraduate assessments (experimental group n = 50 with structured AI tool training monitored through an AI Usage Log [AIUL], control group n = 50 with traditional instruction), a priori power analysis determined sample size using G*Power 3.1.9.7 (power = .85, α = .05, d = .65; Faul et al., 2009). The quantitative data were analysed using paired-samples t-tests, ANCOVA, and Pearson correlations between the five dependent variables (lexical diversity [Multi-Token Linkage Density, MTLD, via TAALES 2.0; Kyle et al., 2021], syntactic complexity [Mean Length of T-units, MLT, via Syntactic Complexity Analysis; Lu, 2010], global cohesion (Text Cohesion Analysis, TCA 2.0; Crossley et al. 2020), argument quality (Writing Quality History and Reference Summary [WQHRS]; McNamara et al., 2015]), and writer voice (Writer Assessment and Rating Scale [WARS]; Ivanìc, 2004; Hyland, 2019]), and the Cohen's d effect sizes used to report data (Cohen, 1988). To measure reliance on AI tools, a new measure labelled AI Dependency Index (ADI) was developed. Qualitative data were collected from semi-structured interviews conducted with each of the twelve (N = 12) EFL writing instructors to identify themes within the interview data through reflexive thematic analysis (Braun and Clarke, 2022). Results of the study's statistical analysis demonstrated that the experimental group made a large and significant improvement compared to the control group in lexical diversity (d = 1.33), syntactic complexity (d = 1.51), and global cohesion (d = 1.60); and moderate but non-significant improvement in argument quality (d = 0.58) compared to the control group; however, the experimental group had similar levels of writer voice (d = 0.07) as the control

group (d = 0.29, p = .024). The mean score of the ADI was 0.68 (indicating moderate-to-high reliance on AI tools) and negatively correlated with writer voice (r = -0.51, p < 0.001) and argument quality (r = -0.34, p = 0.016). Instructors' interview themes were aligned with the statistical analysis of data. The outcomes of the research led to the development of the Pedagogically Mediated AI Writing Framework (PMAWF), a pedagogical framework that comprises four phases of instruction and uses the principle of Assessment Realignment in the instructional design, as a result of the surface-deep dissociation indicated in the quantitative data for each of the five measures of EFL academic writing.

Keywords: artificial intelligence, EFL academic writing, AI dependency index, ChatGPT, Google Gemini, QuillBot, lexical diversity, syntactic complexity, writer voice, argument quality, PMAWF, Iraqi writing, sociocultural theory

أثر أدوات الذكاء الاصطناعي على الكتابة الأكاديمية لطلاب اللغة الإنجليزية كلغة أجنبية: تحليل لغوي وتربوي

م.م. وسان شاكر حمودي
Wasan.shakir.hamoodi@ec.edu.iq

ملخص

تدرس هذه الدراسة كيف تؤثر ثلاث أدوات كتابة قائمة على الذكاء الاصطناعي، وهي-ChatGPT (GPT-4o و Google Gemini وQuillBot، إيجابًا على التطور الأكاديمي لطلاب المرحلة الجامعية الذين يتعلمون الكتابة باللغة الإنجليزية كلغة أجنبية في سياق التعليم العالي العراقي. وتتمثل الفكرة الرئيسية لهذه الدراسة في إظهار كيف يمكن لاستخدام أدوات الذكاء الاصطناعي أن يؤثر إيجابًا على تطوير قواعد الكتابة الأكاديمية المستخدمة في اللغة الإنجليزية كلغة أجنبية (التنوع المعجمي، والتعقيد النحوي، والترابط النصي). ولضمان الحفاظ على تأثير أدوات الذكاء الاصطناعي على الكتابة الأكاديمية، والحفاظ على أصالة أسلوب الكاتب، وتطوير المهارات بمعزل عن الاعتماد على أدوات الذكاء الاصطناعي، من الضروري دمج ممارسات تعليمية منهجية. باستخدام تصميم شبه تجريبي/مختلط الأساليب للاختبار القبلي/البعدي لفحص آثار أدوات الذكاء الاصطناعي على الكتابة الأكاديمية للغة الإنجليزية كلغة أجنبية على مدار اثني عشر أسبوعًا مع 100 تقييم جامعي) المجموعة التجريبية ن = 50 مع تدريب منظم على أدوات الذكاء الاصطناعي تمت مراقبته من خلال سجل استخدام الذكاء الاصطناعي[AIUL] ، المجموعة الضابطة ن = 50 مع التعليم التقليدي(، حدد تحليل القوة المسبق حجم العينة باستخدام) G*Power 3.1.9.7 القوة = 0.85، α = 0.05، d = 0.65؛ Faul et al. ، 2009 .(تم تحليل البيانات الكمية باستخدام اختبارات t للعينات المزدوجة، وتحليل التغاير(ANCOVA) ، ومعاملات ارتباط بيرسون بين المتغيرات التابعة الخمسة (التنوع المعجمي [كثافة الارتباط متعدد الرموز، MTLD، عبرTAALES 2.0 ؛ كايل وآخرون، 2021]، والتعقيد النحوي [متوسط طول وحداتT ، MLT، عبر تحليل التعقيد النحوي؛ لو، 2010]، والتماسك الكلي (تحليل تماسك النص، TCA 2.0؛ كروسلي وآخرون، 2020)، وجودة الحجة (ملخص تاريخ جودة الكتابة والمراجع [WQHRS]؛ ماكنمارا وآخرون، 2015])، وصوت الكاتب (مقياس تقييم وتصنيف الكاتب[WARS] ؛ إيفانيك، 2004؛ هايلاند، 2019])، بالإضافة إلى أحجام تأثير كوهين d المستخدمة في عرض البيانات

العدد3 No 3
2026 ايار May2026
مجلة ساوة للدراسات الانسانية
Sawah Journal of Humanities
E-ISSN 3106 - 8324 P-ISSN 3106 - 8316
مجلة محكمة تصدر عن
جامعة ساوة الاهلية

(كوهين، 1988). لقياس الاعتماد على تم تطوير مقياس جديد يُسمى مؤشر الاعتماد على الذكاء الاصطناعي (ADI)لأدوات الذكاء الاصطناعي. جُمعت بيانات نوعية من مقابلات شبه منظمة مع كلٍّ من مدرّسي الكتابة الاثني عشر (ن = ١٢) للغة الإنجليزية كلغة أجنبية، وذلك لتحديد المواضيع الرئيسية في بيانات المقابلات من خلال التحليل الموضوعي التأملي (براون وكلارك، ٢٠٢٢). أظهرت نتائج التحليل الإحصائي للدراسة أن المجموعة التجريبية حققت تحسنًا كبيرًا وهامًا مقارنةً بالمجموعة الضابطة في التنوع المعجمي (د = ١.٣٣)، والتعقيد النحوي (د = ١.٥١)، والتماسك العام (د = ١.٦٠)؛ وتحسنًا متوسطًا ولكنه غير هام في جودة الحجة (د = ٠.٥٨) مقارنةً بالمجموعة الضابطة؛ ومع ذلك، كان لدى المجموعة التجريبية مستويات مماثلة من أسلوب الكتابة (د = ٠.٠٧) مثل المجموعة الضابطة (د = ٠.٢٩، ق = ٠.٠٢٤). بلغ متوسط درجة مؤشر الاعتماد على الذكاء الاصطناعي ٠.٦٨ (مما يشير إلى اعتماد متوسط إلى عالٍ على أدوات الذكاء الاصطناعي) وكان له تأثير سلبي ارتبطت هذه النتائج ارتباطًا وثيقًا بأسلوب الكاتب(r = -0.51 ، p < 0.001)وجودة الحجة(r = -0.34 ، p = 0.016). وتوافقت محاور مقابلات المُدرّسين مع التحليل الإحصائي للبيانات. أدت نتائج البحث إلى تطوير إطار الكتابة المدعوم بالذكاء الاصطناعي(PMAWF) ، وهو إطار تربوي يتألف من أربع مراحل تعليمية، ويستخدم مبدأ إعادة تنظيم التقييم في التصميم التعليمي، وذلك نتيجةً للانفصال بين السطح والعمق الذي أشارت إليه البيانات الكمية لكل مقياس من المقاييس الخمسة للكتابة الأكاديمية باللغة الإنجليزية كلغة أجنبية.

الكلمات المفتاحية: الذكاء الاصطناعي، الكتابة الأكاديمية باللغة الإنجليزية كلغة أجنبية، مؤشر الاعتماد على الذكاء الاصطناعي، ChatGPT، Google Gemini، QuillBot، التنوع المعجمي، التعقيد النحوي، أسلوب الكاتب، جودة الحجة، PMAWF، الكتابة العراقية، النظرية الاجتماعية الثقافية

1. Introduction

Writing in English as a Foreign Language (EFL), involves careful coordination of words, grammar and writing a strong, academic argument (Hyland, 2019; Wingate, 2023). For Arab students whose first language has different grammar rules, alphabets and writing conventions than English, there are difficulties with writing in an academic genre that cannot be solved by traditional methods (Bitchener & Storch, 2016; Ferris, 2021). Tools that use artificial intelligence to help students write, such as ChatGPT, Google Gemini and Quillbot, provide a higher level of help than previous types of automated writing systems (Warschauer et al., 2023). However, using these tools as a substitute for real thinking can be considered as displacing your knowledge base from your own thinking to computer-generated text (Poole, 2023). Hockly (2023) highlights that introducing these tools into EFL classrooms should happen after careful planning and a set of structured guidelines; not just allowed or prohibited in class.

The existing literature is generally characterised as containing all-encompassing claims regarding whether AI tools positively or negatively affect writing outcomes without clear statements specifying what aspects are positively/negatively affected (Warschauer, Heller, et al, 2023). This paper tests a narrower scope of the above general claim; that is, that AI tools positively affect specific features of writing identified in and incorporated into the writing itself: lexical diversity, syntactic

complexity, and cohesion; and have either a neutral or negative impact on writer voice and argument quality when mediated in an unstructured manner – resulting in reliance and risk of losing long-term independence as writers..The inquiry has three research questions for structuring it:RQ1. Are AI tool users achieving greater pre to post achievement across the five measures of writing quality compared to their counterparts who were instructed using traditional methods?RQ2. What is an individual's level of dependency on AI tools, and how does that dependency correlate to their learning outcomes?RQ3. What are the perceptions of instructors who teach English as a Foreign Language (EFL) with regards to the effects of AI tools used in the writing process?The purpose of this study is to provide quasi-experimental evidence, an AI Dependency Index (ADI), a triangulation of qualitative and quantitative data, and the Pedagogically Mediated AI Writing Framework (PMAWF).

2. Literature Review

2.1 AI Tools in EFL Academic Writing: A Functional Typology

AI writing tools today can be distinguished based on their major functions and at what stage of the writing process they are most useful for users. Using a process of predicting subsequent word(s), ChatGPT (GPT-4o) is capable of generating high-quality academic work by using minimal information to create an appropriate type of writing in an academic genre. Within the EFL writing process, the primary uses of ChatGPT are found during the pre-writing and first-drafting stages of the writing process; idea-generating, argument-building, and genre-modeled content are the three ways that ChatGPT assists writers in the EFL writing process. Google Gemini, another writing tool, differs from ChatGPT because it uses a system of multiple modalities while providing higher-order synthesis of knowledge (Google DeepMind, 2023) than ChatGPT, making it very suitable for writers who are working on revising texts by adding more detailed information (expanding on the content); and last, QuillBot, which is a neural machine that uses a corpus of clean paraphrase pairs for text paraphrasing, produces altered forms of sentences or clauses that have enough potential variation both in their word choice(s) and syntax to enable writers to do the type of editing known as fluency editing. Each of the three different types of tools discussed above serves a different pedagogical purpose in providing users with the right type of AI writing help, and therefore cannot be used interchangeably in the planning of learning experiences for students and teachers. The functional differences between the three tools are shown in Table 1, and each of the tools has significant implications for learning and achieving outcomes in the writing process. Based on the body of the literature in automated writing evaluation (AWE), the interaction quality between users/students and AI will be more important than the specific functionality of any individual AI writing tool when determining if an individual will make developmental gains while using

the tool during the writing process (Li & Link, 2016; Ranalli, 2018; Stevenson & Phakiti, 2019). The theoretical framework for developing a means of measuring lexical richness in work produced by the use of an AI writing tool is evaluated by using Kyle's (2020) framework of measuring lexical richness and is operationalized by the MTLD index used in the present study.

Table 1

AI Tools, Primary Functions, and Pedagogical Roles in the PMAWF

AI Tool	Primary Function	Writing Stage	Pedagogical Role in PMAWF
ChatGPT (GPT-4o)	Text generation, ideation, argument scaffolding	Pre-writing & Drafting	Genre modelling; idea generation; draft comparison
Google Gemini	Research synthesis, factual elaboration, multimodal content	Pre-writing & Revision	Content scaffolding; academic source integration
QuillBot	Neural paraphrasing, sentence restructuring, fluency editing	Revision & Editing	Syntactic variation; lexical upgrading; register adjustment

2.2 Empirical Evidence on AI Tool Effects in L2 Writing

Even before 2022, the published research literature examining the impact of AI on L2 learners' academic writing has grown rapidly in quantity and scope. For example, Crossley et al. (2020) demonstrated that students who receive automated feedback on their writing tend to develop greater lexical sophistication while writing in a second language (L2). This provides a valuable and necessary methodological foundation for using TAALES 2.0 and TAACO 2.0 in the current study, which aims to explore how the use of AI tools may impact L2 learners' academic writing. In addition, Zhang and Hyland (2023) provided further support for the use of automated feedback in L2 academic writing, as their study found that using automated feedback interactively increases clause-level syntactic elaboration in intermediate EFL writers (e.g., increases in subordination density and nominal phrase elaboration). These findings suggest that the use of Lu's (2010) Syntactic

Complexity Analyzer to measure syntactic complexity in the current study is warranted.

Lastly, McNamara et al. (2015) provide the methodological foundation for the automated essay scoring methodology used in the WQHRS argumentation sub-score in this current study.

I also reviewed the research literature related to use of Generative LLMs to produce AI-assisted essays. AI-assisted essays have been shown to demonstrate greater lexical diversity (Mizumoto & Eguchi, 2023) and syntactic elaboration (Lu & Ai, 2015) than students who do not use AI to help them write their papers. Researchers have shown that as students use AI to help them write their essays, they produce fewer writer-specific stance markers than when they do not use AI (Hyland, 2019).
To further investigate the question of AI dependency, researchers have established that second language (L2) writers with lower levels of proficiency are more likely to adopt text that they have generated with the assistance of AI, with little or no modification (Warschauer et al., 2023).
Regarding quality of argument and writers' voices, Wingate (2023) argues that, in general, AI-generated writing is depersonalized and institutionally generic with respect to the disciplinary stance they take toward their underlying content, whereas Poole (2023) provides qualitative evidence to suggest that students who have adopted the use of AI report a lack of a unique writing voice (i.e., they feel as though they have "borrowed voice"). These findings are reflected in the proposed Writer's Voice Rating Scale (WVRS) in the current study, which is based on two theoretical works regarding academic voice (Ivanič, 2004; Hyland, 2019).

2.3 AI Dependency and Academic Integrity

Dependence upon AI use in academia exists at an intersection that is not entirely reducible solely to existing, traditional forms or discourses of academic integrity. Integrity-related scholarship/literature has largely focused on issues relating to AI-text-based submissions in terms of whether those submissions are detectable or punishable (Cotton et al., 2023; Liang et al., 2023). This paper situates AI dependency within a lens of developmental learning. The foundational arguments for technology dependency in language learning are provided by Kessler (2018), who argues that the initial consideration is not whether technology is present but what role technology plays within the context of supporting or supplanting learner cognitive effort; this distinction has direct implications for the four-dimensional operationalization as described within the ADI. Additionally, Deane (2021) has identified cognitive processes that involve effortful composition, revision, and argument development (i.e., cognitive processes most critical to developing writing ability)—the very processes that high-reliance forms of AI use bypass. Lastly,

Cotton et al. (2023) documented hundreds of students (internationally) reporting confusion concerning the limitations surrounding acceptable uses of AI; as such, there is a compelling need for institutions to create frameworks that address the development of dependency rather than simply prohibit the use of AI.

3. Theoretical Framework

The theoretical architecture of this study integrates three complementary frameworks: Vygotskian sociocultural theory, Swain's Output Hypothesis, and genre-based pedagogy.

3.1 Sociocultural Theory and AI Mediation

The Vygotskian Sociocultural Theory (SCT) views learning as mediated, or mediated through cultural tools that develop cognitive abilities for the learner in the Zone of Proximal Development (ZPD; Vygotsky, 1978). Lantolf and Thorne (2006) have shown that writing tools are also a type of psychological artifact and provide cognitive processes for the writer. Generative AI tools are a new type of mediated artifact because they can produce generative, evaluative, and transformational written texts with the writer in real time. Therefore, when the mediated artifact produces a large portion of the written text the Vygotskian premise that the learner continues to be the primary agent of epistemic knowledge may be called into question. The concept that the way an individual engages with an AI-generated output can create a disconnect between a person's understanding and their learning ability is known as "epistemic displacement" (Poole, 2023) and serves as the basis for the ADI as a measurement tool and for the Staged Independence guideline of the PMAWF..

3.2 The Output Hypothesis

As per Swain (2006), using a productive output (an output produced by the learner that exceeds the learner's current level of knowledge) is the driving force behind second language syntactic development through practicing hypothesis testing and metalinguistic reflection. Using AI-generated outputs as a substitute for the effort to produce an output will bypass these mechanisms used to develop language. This follows the hypotheses of the study regarding the differences in productivity related to AI, where the positive impacts of surface features of language will be attributable to the features included in the AI output, while the effects on the quality of the writer's voice and the quality of their arguments (which require for the learner to be engaged with their own knowledge) will be neutral to negative.

3.3 Genre-Based Pedagogy and Assessment Realignment

According to Martin & Rose and Rose & Martin (2012), explicit genre instruction is the place where students learn about the rhetorical conventions, Move structures, and evaluative stance of academic writing through genre-based pedagogy. This use of AI tools for critical comparison of genre exemplars allows for the development of metalinguistic awareness through the Critical AI Interactivity principle. The

Assessment Realignment principle (see Wingate, 2023) was created as a direct response to the dissociation of surface and deep achieved in the results of this study. The use of multimodal evidence (such as process portfolios and compositions written during class) will require a redesign of writing assessments to focus on AI-resistant competencies: analytical argumentation, supporting evidence, and disciplinary stance.

4. Methodology

4.1 Research Design

To conduct this research, a pre-test/post-test quasi-experimental design was used with instructor semi-structured interviews as an embedded qualitative strand (Creswell & Plano Clark, 2018). The quantitative strand will be the main evidence used to examine the effect of AI tools on the writing quality, while the qualitative strand provides the instructor's account of the writing quality that is being investigated and extends the quantitative findings through methodological triangulation. During pre-tests and post-tests, both treatment and control groups received the same argumentative essay tasks, allowing the elicitation conditions to be compared.

4.2 Sample Size Determination: A Priori Power Analysis

The initial sample size was generated through an a priori statistical power analysis that makes use of G*Power3.1.9.7 (Faul et al., 2009). The medium-to-large effect size of d = 0.65 was selected due to similar L2 writing intervention projects (Crossley et al., 2020; Mizumoto & Eguchi, 2023). The analysis produced 44 participants per group at a power of $1 – \beta = .85$ and a two-tailed $\alpha = .05$. The eventual sample size ended up being 50 participants in the experimental condition and 50 participants in the control condition (N = 100); this provided slightly over 13.6% of buffer space from the estimated attrition rate. The effect sizes reported throughout by Cohen (1988) use the following thresholds: .20 = small; .50 = medium; and .80 = large.

4.3 Participants and Setting

The research sample consisted of 100 (n=58 females; n=42 males) undergraduate study participants (ages 19-23; M=21.1, SD=1.1) enrolled in the Academic Writing in English course for the 2023-2024 academic year at the College of Education for Humanities, University of Baghdad, Department of English. All of the participants were native Arabic speakers, had received formal instruction in English for between 8 and 12 years, and had placement test scores that fell within the B1 and B2 range of the Common European Framework of Reference (CEFR). Proficiency data (B1: n = 28/group; B2: n = 22/group) were controlled for stratified random assignment using a baseline independent samples t-test; therefore, the equivalent proficiency level of the two sample groups was confirmed ($t(98) = .34$, $p = .736$). Twelve (7 male; 5 female) EFL writing instructors with a mean teaching

experience of 11.3 years were also recruited to participate in qualitative interviews. Ethical approval for the study was granted by the University of Baghdad Research Ethics Committee (Ref: UoB-REC-2023-114); written informed consent was obtained from the participants.

4.4 Intervention Procedures

This 12-week intervention consisted of three phases. During Phase 1 (Weeks 1-4), experimental participants took four training workshops related to the functional differentiation of ChatGPT, Google Gemini, and QuillBot, how to evaluate critical AI outputs, AIUL completion processes, and how to design prompts for specific genres (Hockly, 2023). The control group will be provided an equivalent number of classes but will not have access to AI tools. During Phase 2 (Weeks 5-9), experimental group participants produced three graded argumentative essays through AIUL protocols and included reflective annotations of how they compared the AI-generated section to the section they authored, as well as a metalinguistic explanation of the rationale behind each revision they made. Both groups will be writing on the same topics during the study. During Phase 3 (Weeks 10-12), participants had progressively less access to AI tools: Week 10 QuillBot was restricted to being used only for post-draft revisions; Week 11 ChatGPT was restricted to pre-writing only; and Week 12 all AI tools will be prohibited. All participants will conduct the posttest under the same conditions, not using any AI tool.

4.5 Writing Tasks and Topic Equivalence

All pre/post-task essays of 550-650 words on the same contested social/educational issue required evidence-based argumentation, expression of disciplinary stance, and structured counter-arguments. The prompts were drawn randomly from a pool of 12 researchers' case-limited prompts rated equally by 3 expert raters (Krippendorff's $\alpha = .83$). The same topics were used for each group within the intervention as a means of controlling for effects of familiarity with the topic.

4.6 The AI Usage Log and AI Dependency Index

Experimental participants utilizing the AI Usage Log (AIUL) were required to record four aspects of their use in the AIUL for each essay session. They were required to record the frequency of their AI usage (sessions per task), the prompt type they used (Generative / Corrective / Paraphrase), the total amount of time they used the AI (minutes, based upon the time stamps) and their post-AI text revision rate (percentage of text modified before submitting the essay, based upon WDiff analysis). The structure of the AIUL can be seen in Table 2.

Table 2

AI Usage Log (AIUL): Dimensions, Operationalisations, and High-Reliance Thresholds

Log Dimension	Operational Definition	Measurement Unit	Instrument	High-Reliance Threshold
Frequency of Use	No. of AI interaction sessions per essay task	Count per task	Browser log / Self-report diary	≥5 sessions = high reliance
Prompt Type	Nature of command: generative / corrective / paraphrase	Categorical coding (3 types)	Prompt journal	Generative >50% = reliance risk
Duration of Use	Total minutes per AI interaction session per task	Minutes per essay	Time-stamp log	>30 min = overuse threshold
Post-AI Revision Rate	Proportion of AI text modified by student before submission	% word-level changes (WDiff)	Track Changes / WDiff analysis	<10% modifications = minimal agency

.

The ADI is a weighted composite that was developed using the following formula: ADI = 0.25(Frequency_norm) + 0.30(Generative_Prompt_ratio) + 0.20(Duration_norm) + 0.25(1 - PostAI_Revision_rate). Each of the elements in the formula was min-max normalised from 0 to 1 to enable the relative weighting of each of the elements using theoretical models of the relative impact of the use of a prompt as opposed to being a scaffold (i.e., either an indicator of AI replacement or scaffold) on how the use of that element would affect the ongoing development of that individual as an online learner. An ADI of 0.70 or higher was operationally defined as having high dependence on AI.

4.7 Dependent Variable Measures

A variety of five attributes were measured using essays completed at both the start and end of a course: (a) Lexical Variety - MTLD, as measured with TAALES Online 2.0 (Kyle et al., 2021) by the method described in Kyle et al. (2020) for evaluating lexical richness of a text sample with acceptable robustness when measured across different lengths of text. (b) Syntactic Complexity - MLT, as measured using Lu's (2010) SCA, as this measure was chosen for its ability to detect change in L2 development with a population found at the intermediate proficiency level. (c) Global Cohesion, as measured by semantic overlap indexes through TAACO 2.0 (Crossley et al., 2020), as this measure had been shown to have the best predictive value with L2 corpora. (d) Argument Development - Argumentation Criteria Sub-Score from the WQHRS, as rated using items proposed by McNamara et al., (2015) using independent raters who were blind to both the treatment, as well as the order of the essays being rated, produced an overall inter-rater reliability of Cohen's $\kappa = .87$. (e) Writer's Voice will also be evaluated using a five-point rubric called the WVRS that was developed for this study to measure writer's personal stance; evaluative lexis density; authorial position; and hedging vs. boosting of the writer's voice; based on Ivanič (2004) and Hyland (2019), the WVRS will also produce an overall inter-rater reliability of Cohen's $\kappa = .83$..

4.8 Qualitative Data Collection and Analysis

Twelve instructor participants each took part in a separate semi-structured interview (45 to 65 minutes) to gather quantitative and qualitative data. The main areas of interest included: How do instructors perceive the impact of AI tools on the quality of student written work? How do instructors perceive students' metalinguistic behaviours relative to their use of AI tools? Are instructors concerned about academic integrity/dependence? How do instructors believe they can best respond pedagogically to student's use of AI? All interviews were recorded and transcribed verbatim. A reflexive thematic analysis (Braun & Clarke, 2022) was conducted on the interview data, and a second researcher independently coded 25% of transcripts (Cohen's kappa = .80).

4.9 Statistical Analysis

Paired t-tests (or Wilcoxon signed-rank tests as applicable) were utilized for testing the differences between pre and post within each group. ANCOVA was performed upon groups to test for difference between groups of the posttest while controlling for pre-test scores as covariates. The effect size for all comparisons was also calculated (using Cohen's d, 1988). Pearson’s correlation coefficient was utilized to examine the relationship between the ADI and the five outcome measures. Statistical significance was set at a=.05. Analyses were conducted in R Version 4.3 (R Core Team, 2023) and SPSS Version 28..

5. Results

5.1 Baseline Equivalence and Descriptive Overview

For each of the dependent variables, the analysis found no significant differences between the pre-test mean scores of each group (all p values > .05), confirming that both groups were relatively equivalent prior to treatment. Descriptive statistics, along with results of pairwise t-tests conducted within each group, can be found in Table 3 for all five measures and for the ADI.

Table 3

Pre-test and Post-test Descriptive Statistics, Paired t-test Results, and Effect Sizes by Group

Measure	Group	Pre M (SD)	Post M (SD)	t (df=49) / p	d
MTLD (Lex. Diversity)	Experimental (n=50)	64.3 (11.2)	88.7 (10.4)	9.41, p<.001***	1.33
MTLD (Lex. Diversity)	Control (n=50)	63.8 (10.9)	69.2 (11.7)	2.18, p=.034*	0.31
Syntactic Complexity (MLT)	Experimental	8.4 (1.3)	11.2 (1.1)	10.67, p<.001***	1.51
Syntactic Complexity (MLT)	Control	8.3 (1.4)	9.1 (1.3)	2.72, p=.009**	0.38
Global Cohesion (TAACO)	Experimental	0.34 (0.07)	0.51 (0.06)	11.34, p<.001***	1.60
Global Cohesion (TAACO)	Control	0.33 (0.06)	0.37 (0.07)	2.30, p=.026*	0.33
Argument Quality (WQHRS)	Experimental	2.9 (0.6)	3.4 (0.5)	4.12, p<.001***	0.58
Argument Quality (WQHRS)	Control	2.8 (0.6)	3.2 (0.5)	3.21, p=.002**	0.45
Writer Voice	Experimental	3.1 (0.7)	3.0 (0.8)	−0.61,	0.07

(WVRS)				p=.544 (ns)	
Writer Voice (WVRS)	Control	3.0 (0.7)	3.3 (0.6)	2.04, p=.047*	0.29
AI Dependency Index (ADI)	Experimental only	—	0.68 (0.14)	—	—

Table 4 presents ANCOVA between-group post-test comparisons controlling for pre-test scores, reporting partial η^2 with thresholds from Cohen (1988).

Table 4

ANCOVA Between-Group Post-test Comparisons Controlling for Pre-test Scores

Dependent Variable	F(1, 97)	p	Partial η^2	Sig.	Favours
MTLD — Lexical Diversity	41.23	<.001	.30	***	Exp.
MLT — Syntactic Complexity	67.89	<.001	.41	***	Exp.
Global Cohesion (TAACO)	73.14	<.001	.43	***	Exp.
Argument Quality (WQHRS)	1.87	.175	.02	ns	—
Writer Voice (WVRS)	5.23	.024	.05	*	Control

5.2 Lexical Diversity (MTLD)

The experimental group experienced significant improvements in MTLD from pre- to post-test based on a paired-sample t-test analysis ($t(49) = 9.41$, $p < .001$. $d = 1.33$), which is a large effect size according to Cohen's (1988) benchmarks. Even though there was a statistically significant change for the control group, the change was less than that of the experimental group ($t(49) = 2.18$, $p = .034$, $d = 0.31$). ANCOVA results demonstrated there was a large effect size difference between groups at post-test ($F(1, 97) = 41.23$, $p < .001$, partial $\eta^2 = .30$) highlighting that using the AI tool to complete an essay provided a greater increase in lexical

diversity than conventional instruction methods. This data supports previous findings using corpus data, which demonstrate that AI-assisted essays draw on a more expansive lexical domain compared to unassisted EFL writing (Mizumoto & Eguchi, 2023), naming the deployment of Academic Word List vocabulary by the outputs of ChatGPT and Google Gemini as part of this increase using the TAALES framework (Kyle, 2020; Kyle et al., 2021).

5.3 Syntactic Complexity (MLT)

The participants in this study gained a lot on their MLT (Mean Length of T-unit) from test 1 to test 2 demonstrated by an extremely high pre-to-post MLT growth (i.e., $t(49) = 10.67$; $p < .001$; $d = 1.51$) increasing from 8.4 to 11.2 words per T-unit. The control group showed statistically significant but small MLT growth (i.e., $t(49) = 2.72$; $p = .009$; $d = 0.38$). The ANCOVA indicates there was a strong and statistically significant difference between the experimental participants and the control group (i.e., $F(1,97) = 67.89$; $p < .001$; partial $\eta^2 = .41$). In particular, the increases in MLT were concentrated in the areas of subordinate clause embedding and nominal phrase elaboration (Lu & Ai, 2015; Zhang & Hyland, 2023). This pattern can be explained theoretically as the exposure to the AI-generated writing acting as more comprehensible input (Ellis, 2015), thus giving learners of English as a foreign language (EFL) exposure to syntactic structures beyond what they could produce independently.

5.4 Text Cohesion (Global Overlap)

The outcome of the experimental group showed that there was a very large increase from pretest to posttest in terms of cohesiveness or connectedness ($t (49) = 11.34$, $p < 0.001$, $d = 1.60$), with the semantics of the essays being associated with each other increased from an average of 0.34 to an average of .51. While there were also significant, yet slight increases in this area for the control group, ($t (49) = 2.30$, $p = 0.026$, $d = 0.33$), the ANCOVA analysis demonstrated a significant difference between the two groups ($F (1, 97) = 73.14$, $p < 0.001$, part $\eta^2 = 0.43$). In particular, high-ADI essays (with an ADI of at least 0.80) show that authors of these high-ADI essays tended to over-rely on a very limited selection of high-frequency connectors to create surface cohesion, rather than developing a cohesive (or "coherent") structure to their essays through the creation of a semantic-like structure (Crossley et al., 2020).

5.5 Argument Quality (WQHRS)

Both groups had statistically significant improvements with regards to their pre- to Post-test scores within their respective (Experimental: $t(49)=4.12$, $p<0.001$, $d = 0.58$; Control: $t(49)=3.21$, $p=0.002$, $d = 0.45$) groups. Of paramount importance is the fact that ANCOVA did not identify any statistically significant differences in the two groups at the Post-test, as evidenced by the following F statistic: ($F(1, 97)=1.87$, $p=0.175$, partial $\eta^2=0.02$) or d = approximately 0.12, respectively. Thus,

the fact that the experimental group significantly excelled on raw measures relative to the control group is one the largest envelope of conclusive quantitative data produced by this study, since AI tool-assisted argumentation did not produce any significant change in the level of argument quality beyond what would have otherwise been produced through standard (traditional) instruction, in agreement with the conclusion of McNamara et al. (2015) that argument quality requires cognitive processes that AI cannot support directly.

5.6 Writer Voice (WVRS)

The experimental group did not demonstrate any statistically significant improvement between the pre- and post-intervention scores ($t(49) = -0.61$, $p = 0.544$, $d = 0.07$), with an average score drop from 3.1 to 3.0. The control group did, however, demonstrate a statistically significant positive change ($t(49) = 2.04$, $p = 0.047$, $d = 0.29$) in their average scores from 3.0 to 3.3. Furthermore, the ANCOVA analysis produced a statistically significant between-group effect in favour of the control group ($F(1, 97) = 5.23$, $p = 0.024$, partial $\eta^2=0.05$). This result supports Wingate's (2023) view that AI-generated prose produces a depersonalised type of register that has the potential to replace a developing authorial voice, and is also in accordance with Ivanič's (2004) theoretical framework that supports voice as being produced through a writing process that includes an identity investment and an effortful engagement with the writing task.

5.7 AI Dependency Index (ADI)

The mean ADI for all intervention essays was $M = 0.68$ ($SD = 0.14$), which is approaching a high-reliance threshold of 0.70; analysis of component prompts showed that the majority of participants (62%) relied extensively on generative prompts when drafting. Furthermore, post-AI revision rates were very low ($M = 18.3\%$, $SD = 9.4\%$). Moreover, ADI was negatively and significantly correlated with WVRS post-test scores ($r = -0.51$, $p < 0.001$) and with post-test argument quality scores ($r = -0.34$, $p = 0.016$). In contrast, ADI had a non-significant positive correlation with MTLD ($r = 0.28$, $p = 0.051$). Overall, higher levels of AI reliance among participants were found to be specifically associated with poorer rhetorical-developmental dimensions of writing quality (Kessler, 2018).

5.8 Qualitative Findings: Instructor Perceptions

Reflexive thematic analysis (Braun & Clarke, 2022) of the twelve instructor interview transcripts generated four superordinate themes presented in Table 5, with frequency data and linkage to corresponding quantitative findings.

Table 5

Instructor Qualitative Themes, Representative Excerpts, Frequencies, and Links to Quantitative Findings

Theme	Representative Excerpt	Frequency	Linked Quantitative Finding
AI as surface-level corrector	'Students produce better sentences but cannot explain why they are correct.' (Instructor 3)	9/12 instructors	MTLD↑ (d=1.33); WVRS ns
Diminished metalinguistic awareness	'When I ask why they chose a structure, they say the AI told them so.' (Instructor 7)	11/12 instructors	ADI↑ → Voice↓ (r=−.51)
Borrowed voice / inauthenticity	'The essay sounds like no student I know. It is fluent but it is not theirs.' (Instructor 11)	10/12 instructors	WVRS: Exp. unchanged; Control↑
Increased drafting confidence	'Weaker students are willing to write more. The blank page no longer terrifies them.' (Instructor 5)	7/12 instructors	MTLD d=1.33; Cohesion d=1.60

The major theme of lessened awareness of metalanguage (11 out of 12 faculty members) aligns with the ADI–voice association ($r = -0.51$, $p < 0.001$): those students who depend heavily on AI also have less of a writer's voice and, as instructors note, think less about their linguistic choices and have less ability to explain or justify those choices. The only clearly positive theme of writing, more confidence in drafting (7 out of 12) is backed up by the significant average group gains at the surface level, which confirms that the stabilizer effect of AI scaffolding on the affective barrier to initially writing does not develop similar cognitive-rhetorical abilities that lead to creating and supporting authentic academic arguments.

6. Discussion

6.1 The Surface-Deep Dissociation

The pattern of results across five dependent variables reveals a theoretically coherent and practically significant dissociation between AI tool effects on surface linguistic features and on deeper rhetorical-developmental dimensions. For lexical

diversity, syntactic complexity, and cohesion, the experimental group demonstrated large-to-very-large ANCOVA effects (partial η^2 = .30, .41, .43 respectively), consistent with the computational corpus literature (Crossley et al., 2020; Zhang & Hyland, 2023), and attributable in part to the lexical richness mechanisms identified by Kyle (2020) and validated by Kyle et al. (2021). For argument quality, no significant between-group effect emerged (p = .175, partial η^2 = .02). For writer voice, the control group significantly outperformed (p = .024, partial η^2 = .05). The qualitative instructor data—particularly the 'surface-level corrector' theme endorsed by 9/12 instructors—provide a convergent explanatory account: students produce better sentences without understanding why they are better, undermining the metalinguistic development that underlies independent writing competence (Ranalli, 2018).

This evidence of surface-level dissociation can be explained through Social Cultural Theory to illustrate that the AI tool mediates effective construction of target-like linguistic form (Lantolf & Thorne, 2006; Vygotsky, 1978) but does not mediate the development of argumentation and authorial voice, which requires strenuous cognitive engagement with both effort and identity investment, and are therefore not facilitated through the use of AI-generated output (Swain, 2006). Additionally, the existence of a significant negative correlation between Authorial Voice and AI use ($r = -.51$) demonstrates it is not a pure between group difference: within the experimental group, AI use directly predicts stagnation across dimensions that require authentic epistemic investment.

6.2 The PMAWF: Operationalised Framework

The PMAWF is operationalised as a four-stage classroom-deployable model grounded in the empirical findings and theoretical frameworks of this study. Figure 1 presents the complete framework.

Figure 1. Pedagogically Mediated AI Writing Framework (PMAWF)

STAGE 1 AI Tool Input	STAGE 2 Guided Interaction	STAGE 3 Language Development	STAGE 4 Independent Writing
▸ ChatGPT (GPT-4o)	▸ Critical AI Interactivity	▸ Lexical Diversity	▸ Writer Voice
▸ Google Gemini	▸ AI Usage Log	▸ Syntactic Complexity	▸ Reduced AI Reliance
▸ QuillBot	▸ Metalinguistic Reflection	▸ Cohesion Quality	▸ Autonomous Competence

Pedagogical Mediation Layer: ZPD Scaffolding | ADI Monitoring |

Note. Stage progression is contingent on ADI monitoring. The Pedagogical Mediation Layer is a continuous supervisory function across all stages, encompassing ZPD scaffolding (Vygotsky, 1978), ADI monitoring, and Assessment Realignment.

In Stage 1, AI Tools N1,N2, and N3 are activated according to each tool's specified operating parameters (Table 1) and are designed to facilitate Critical AI Interactivity for students receiving Junior genre instruction. This occurs prior to the use of AI generated content in student essays and takes place as part of the Deconstruct Phase of Martin and Rose (2012)'s Deconstruction model within an AI mediated context. In Stage 2, Guided Interaction utilizes Metalinguistic Scaffolding wherein students record their AIUL and make reflective annotations post-session related to the linguistic reasoning for accepting or rejecting their AI UL(e.g., Swain's, 2006) definition of languaging. Stage 3 records the development of target-like Surface Features as a result of practice with AI Mediated input. The critical pedagogical challenge is maintaining the quality of Argument and Voice in relation to the emergence of Surface Features through argument mapping and peer review tasks that require independent engagement in Rhetorical Argumentation. Stage 4 implements the Staged Independence model (and Assessment Realignment Principle), whereby Summative Assessments must assess AI Resistant Skills/Abilities, and demonstrate development through Multi Modal Evidence including Process Portfolios and In-Class Compositions in addition to essay submissions (Wingate, 2023). The Author Voice Data indicate that under the current 12-week protocol, Stage 4 transfer has not been fully completed; therefore, the Staged Independence Timeline should be extended for any future implementations.

6.3 AI Dependency: Pedagogical and Policy Implications

The average AI-Driven Instruction (ADI) of 0.68 indicates high risk of harm in teaching. The main factor driving this risk is the high proportion of students requesting long form texts via generative prompts as opposed to using AI to correct their work or rewrite it in their own way. This shows that there is a need for students to develop skills in how to create prompts that will help them develop their own ideas rather than taking away from what they already had; this is a pedagogical goal (Kessler, 2018; Hockly, 2023). The Argument Against Prohibiting AI (PMAWF) is supported by empirical data showing that banning AI would take away from the confidence a teacher gained when working with AI in their classroom and therefore placing many students from language-minority

backgrounds at a disadvantage. The Assessment Realignment standard allows for integrity concerns to be addressed without creating additional prohibitions on the use of AI.

6.4 Limitations and Future Research

Generalizability has limitations for a number of reasons. First, the single site limits external validity (both institutional and geographical); this would be greatly enhanced by replicating at multiple sites across universities in Iraq .Second, the 12-week intervention does not allow for any conclusions about the longitudinal consolidation or decay of the AI-mediated improvements; to answer these questions, researchers will need to conduct longitudinal studies that attribute students over the course of an entire academic year. Third, while the WVRS achieved an acceptable level of interrater reliability (κ =.82), researchers need to validate this tool on a wider variety of EFL populations. Finally, the predictive validity structure of the ADI requires additional validation independent of the other two variables (WVRS and ADI). Finally, future research should employ keystroke logging and verbal protocols to better understand the microgenetic processes that occur in writing cognition as a result of interacting with artificial intelligence tools.

7. Conclusion

The controlled, power-analyzed quasi experimental evidence (including both pre-test/post-test results) presented in this study, when triangulated with instructor qualitative information, supports the conclusions that large effects are obtained in surface linguistic features of EFL student performance (in terms of vocabulary diversity, syntactic complexity, and cohesion) as the outcome of using AI writing tools (ChatGPT, Google Gemini, and QuillBot) and that no statistically significant independent effect was noted on the argument quality of writers using AI tools. On average, between .02 and .05 were the coefficients of association between the AI Dependency Index (M= 0.68) and the writing dimensions specified by the rhetorical-developmental principles for writers. Bilateral converging qualitative data from the instructors showed that students using AI writing tools produced writing that visually looked better but had not internalized the requisite metalinguistic and rhetorical principles to achieve the writing competence characteristic of authentic academic writing..

The Pedagogically Mediated AI Writing Framework (PMAWF), which has been designed to support EFL (English as a Foreign Language) instructors in implementing various teaching strategies, has been developed through four stages – AIUL and ADI Instruments used in the classroom and ‘Assessment Realignment’ also known as ‘Assessment Realigning’ are the key components of the PMAWF. AI tools are very useful for educational purposes (with the exception of certain cases); however, they are not sufficient to teach EFL writing skills in an effective way because the success of these tools depends exclusively on how well they have

been integrated into the writing process by using appropriate pedagogical methods or practices. Academic writing instruction aims to produce writers who will be able to think critically and creatively using logical reasoning, persuasively based on their unique experiences, and express themselves clearly in all their respective areas of research and expertise. The writer cannot rely solely on generative AI (an example of a type of AI) to facilitate the writing process. Therefore, educators must use the integration of generative AI into their curricula as a means of facilitating and supporting the writer's ability to think critically and creatively and to engage in analytical reasoning without losing their creative abilities by using appropriate pedagogical practices or methods..

References

- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52–62. https://doi.org/10.61969/jai.1337500
- Bitchener, J., & Storch, N. (2016). Written corrective feedback for L2 development. Multilingual Matters. https://doi.org/10.21832/9781783095056
- Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications. https://doi.org/10.4135/9781529682045
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.5555/3495724.3495883
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
- Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications. https://doi.org/10.15353/cjlt.v13i3.309
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2020). Predicting human judgments of essay quality in high and low stakes contexts: A comparison of

physical and virtual raters. International Journal of Artificial Intelligence in Education, 30(3), 424–452. https://doi.org/10.1007/s40593-020-00201-5

- Deane, P. (2021). The key practices underlying effective writing. ETS Research Report Series, 2021(1), 1–37. https://doi.org/10.1002/ets2.12311
- Ellis, N. C. (2015). Implicit and explicit learning of languages: Looking back and projecting forward. In P. Rebuschat (Ed.), Implicit and explicit learning of languages (pp. 419–431). John Benjamins. https://doi.org/10.1075/sibil.48.21ell
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149
- Ferris, D. R. (2021). Written corrective feedback in second language acquisition and writing studies. Language Teaching, 54(1), 94–109. https://doi.org/10.1017/S0261444819000247
- Google DeepMind. (2023). Gemini: A family of highly capable multimodal models [Technical report]. https://doi.org/10.48550/arXiv.2312.11805
- Hockly, N. (2023). Artificial intelligence in English language teaching: The good, the bad and the ugly. RELC Journal, 54(2), 445–452. https://doi.org/10.1177/00336882231168504
- Hyland, K. (2019). Second language writing (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108635547
- Ivanič, R. (2004). Discourses of writing and learning to write. Language and Education, 18(3), 220–245. https://doi.org/10.1080/09500780408666877
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274
- Kessler, G. (2018). Technology and the future of language teaching. Foreign Language Annals, 51(1), 205–218. https://doi.org/10.1111/flan.12318
- Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), The Routledge handbook of vocabulary studies (pp. 454–476). Routledge. https://doi.org/10.4324/9780429291586-29
- Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgments. Language Assessment Quarterly, 18(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

- Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and the genesis of second language development. Oxford University Press. https://doi.org/10.1017/S0272263108080510
- Li, J., & Link, S. (2016). Examining the effects of automated writing evaluation on EFL learners' revision behaviours and their writing outcomes. CALICO Journal, 33(1), 1–23. https://doi.org/10.1558/cj.v33i1.26895
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu
- Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with different L1 backgrounds. Journal of Second Language Writing, 29, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003
- Martin, J. R., & Rose, D. (2012). Learning to write, reading to learn: Genre, knowledge and pedagogy in the Sydney School. Equinox. https://doi.org/10.1558/isbn.9781908049988
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35–59. https://doi.org/10.1016/j.asw.2014.09.002
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050
- Poole, R. (2023). Student-facing generative artificial intelligence in English for academic purposes. Journal of English for Academic Purposes, 66, Article 101315. https://doi.org/10.1016/j.jeap.2023.101315
- R Core Team. (2023). R: A language and environment for statistical computing (Version 4.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
- Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? ELT Journal, 72(1), 42–51. https://doi.org/10.1093/elt/ccx018
- Rose, D., & Martin, J. R. (2012). Learning to write, reading to learn. Equinox. https://doi.org/10.1558/isbn.9781908049988
- Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland & F. Hyland (Eds.), Feedback in second language writing:

Contexts and issues (2nd ed., pp. 125–141). Cambridge University Press. https://doi.org/10.1017/9781108569125

- Swain, M. (2006). Languaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.), Advanced language learning: The contribution of Halliday and Vygotsky (pp. 95–108). Continuum. https://doi.org/10.5040/9781474212427.ch-005
- Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4
- Wahle, J. P., Ruas, T., Foltýnek, T., Meuschke, N., & Gipp, B. (2022). Identifying machine-paraphrased plagiarism. In A. Düsterhöft, M. Klettke, & K.-D. Schewe (Eds.), Modelling, computation and optimization in information systems and management sciences (pp. 393–404). Springer. https://doi.org/10.1007/978-3-030-97008-6_35
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. P. (2023). The affordances and contradictions of AI-generated text for writers of academic English. Journal of Second Language Writing, 62, Article 101071. https://doi.org/10.1016/j.jslw.2023.101071
- Wingate, U. (2023). What does 'academic writing' mean in the era of AI? Journal of English for Academic Purposes, 65, Article 101287. https://doi.org/10.1016/j.jeap.2023.101287
- Zhang, Z., & Hyland, K. (2023). Student engagement with automated feedback in L2 writing: Authenticity, authority, and expertise. Computers and Education, 193, Article 104683. https://doi.org/10.1016/j.compedu.2022.104683