Error Analysis in the Age of AI: A Psycholinguistic Study of L2 Writer Dependency on Automated Writing Assistants

Chitra Dhanapal[1], Maha Al-shehri[2], N. Asharudeen[3]

[1] Applied college for girls, Al-Majaraha, King Khalid University, Saudi Arabia

[2] CEO of Applied college Al-Majardah, Applied college for girls Al-Majardah, King Khalid University Abha, Saudi Arabia. E-mail: alshahry@kku.edu.sa

[3] Professor of English, Arifa Institute of Technology, Esanoor, Tamilandu, India. E-mail :deennaina@gmail.com

Correspondence: Chitra Dhanapal, King Khalid University, Saudi Arabia. E-mail: cindra@kku.edu.sa

Received: October 14, 2025 Accepted: January 26, 2026 Online Published: April 17, 2026

doi:10.5430/wjel.v16n4p469 URL: https://doi.org/10.5430/wjel.v16n4p469

Abstract

Although the pedagogical utility of Automated Writing Assistants (AWAs) has received increasing attention, relatively little is known about their long-term psycholinguistic effects on second language (L2) English writers. This study examines the impact of sustained AWA use among intermediate L2 learners enrolled in a 16-week academic writing course. A mixed-methods longitudinal design was employed, involving 48 participants randomly assigned to either an AWA-supported group or a control group. Data were derived from corpus-based error analysis focusing on grammatical, mechanical, and lexical categories alongside stimulated recall interviews and a post-intervention writing task conducted without AWA support. The findings suggest that the AWA group demonstrated more rapid surface-level error reduction, particularly in article and preposition usage; this improvement in error reduction diminished once the tool was withdrawn. Participants in this group exhibited increased error recurrence and reported heightened metalinguistic uncertainty, with revisions closely resembling prior AWA outputs rather than evidencing internalised control. This pattern was interpreted as indicative of procedural dependency, wherein AWAs appeared to function as external monitors that constrained the development of autonomous self-regulation. The study underscored the need for pedagogical strategies that temper short-term gains in accuracy with sustained metacognitive engagement, thereby supporting the cultivation of linguistic autonomy over time.

Keywords: Automated Writing Evaluation (AWE), L2 Writing Development, Psycholinguistics, Error Analysis, Writer Dependency, Digital Language Learning

1. Introduction

The integration of Artificial Intelligence (AI) into language learning environments has gained considerable momentum in recent years; its influence on second language (L2) writing instruction has yet to be fully understood. One of the more visible developments in this domain has been the widespread adoption of Automated Writing Assistants (AWAs), which means AI-driven digital tools designed to analyse written text and provide real-time or post-hoc feedback on linguistic features such as grammar, spelling, punctuation, vocabulary, and style, with the aim of improving the accuracy, clarity, and overall quality of writing. These systems, underpinned by advances in natural language processing and machine learning, have been promoted as mechanisms for improving writing quality, reducing cognitive load, and supporting linguistic development (Bai & Stede, 2022; Cavalcanti et al., 2021). However, as their use becomes increasingly embedded within pedagogical practice, a growing body of literature has begun to question whether such tools genuinely facilitate the internalisation of linguistic knowledge or merely simulate learning through surface-level correction (Deeva et al., 2021; Ahmad et al., 2022). The present study was situated within this critical discourse and sought to examine the psycholinguistic consequences of sustained reliance on AWAs in the context of L2 academic writing.

Recent research on AI-mediated writing feedback has tended to focus on technical performance indicators, including agreement with human raters (Firoozi et al., 2023), coverage of error types (Bai & Stede, 2022), and the precision of automated scoring models (Andersen et al., 2022). While such metrics are valuable for assessing system reliability, they often overlook the cognitive and developmental implications of prolonged engagement with algorithmic correction. The assumption that immediate error resolution constitutes learning remains largely unexamined. From a psycholinguistic perspective, writing is understood not as a static representation of linguistic form but as a dynamic process involving planning, monitoring, revision, and self-regulation (Krashen, 1982). Cognitive models of L2 acquisition posit that internal monitoring mechanisms evolve through iterative exposure, hypothesis testing, and feedback that activates metalinguistic awareness. When feedback is delegated to opaque external systems that provide instantaneous corrections, learners may bypass the active construction of grammatical knowledge. In such cases, what emerges is a form of procedural dependency, a reliance on the tool that inhibits the development of autonomous linguistic control.

This concern aligns with broader debates in the fields of learning analytics and educational AI, particularly those addressing the design of

feedback mechanisms that promote learner agency rather than passive compliance. Ahmad et al. (2022) contended that effective learning analytics ought to be grounded in principled pedagogical design, ensuring that data-driven indicators serve instructional objectives rather than technological imperatives. Within the domain of writing, this distinction necessitated a differentiation between feedback that merely enhanced surface-level accuracy and feedback that fostered metacognitive reflection. Automated systems, however, frequently privileged the former, offering prescriptive corrections without explanatory rationale and thereby limiting opportunities for learners to interrogate the basis of error identification (Cavalcanti et al., 2021; Deeva et al., 2021). In the absence of such reflective engagement, learners may struggle to generalise corrective patterns across novel contexts, a process widely recognised as integral to robust L2 development.

Recent studies further emphasised the motivational and emotional dimensions of feedback interaction, aspects which were often overlooked in algorithmic systems. Fong and Schallert (2023) proposed a 'feedback-to-the-future' framework, wherein feedback was positioned not as retrospective judgment but as a forward-oriented catalyst for self-regulated learning. In this formulation, effective feedback was understood to support the development of learner identity, encourage intellectual risk-taking, and cultivate epistemic confidence. By contrast, AWAs operated within a deficit model that flagged deviations from normative standards without acknowledging the developmental legitimacy of interlanguage forms. Such practices risked reinforcing linguistic insecurity among L2 writers, particularly when corrections were applied uniformly, irrespective of proficiency level or rhetorical intent. The absence of human nuance manifested in scaffolding, encouragement, and contextual adaptation was interpreted as a potential barrier to experimentation, thereby inhibiting the very processes through which linguistic competence was expected to mature. Empirical investigations into automated feedback have also revealed a number of methodological limitations which obscure its long-term pedagogical impact. A considerable proportion of existing studies have relied on single-draft comparisons or short-term interventions, approaches which are insufficient to capture the evolving relationship between learner and tool (Gombert et al., 2022). Moreover, the evaluation of writing quality has tended to privilege formal accuracy over rhetorical sophistication or conceptual development dimensions, which remain central to academic writing but are less amenable to algorithmic quantification. Andersen et al. (2022) demonstrated the potential of semi-automatic coding in the analysis of open-ended responses within large-scale assessments. Nevertheless, even such hybrid approaches have rarely traced the trajectory of error emergence, persistence, and resolution across multiple iterations. This gap appeared particularly salient in L2 contexts, where error patterns are not static but evolve in response to instruction, exposure, and internal cognitive restructuring.

A mixed-methods longitudinal design was adopted to capture both behavioural and metacognitive dimensions for examining not only whether AWAs contribute to error reduction but also how their sustained use may influence the learner's internal monitoring system. Drawing on psycholinguistic theory and recent developments in learning analytics (Ahmad et al., 2022), the investigation sought to determine whether habitual reliance on AWAs led to durable gains in autonomous writing accuracy or merely produced a temporary, tool-dependent performance enhancement. By combining corpus-based error tracking with stimulated recall interviews and writing tasks conducted without AWA support, the study aimed to capture both behavioural outcomes and metacognitive processes. In doing so, it responded to calls for more nuanced, design-informed evaluations of AI in education, which treat technology not as a neutral aid but as an active participant within the cognitive ecology of learning (Deeva et al., 2021; Cavalcanti et al., 2021).

Accordingly, the present study was guided by the following research questions:

- To what extent did sustained AWA use influence the recurrence of grammatical, mechanical, and lexical errors in unsupported writing tasks?
- How did habitual AWA reliance affect learners' metalinguistic awareness and capacity for autonomous error correction post-intervention?

2. Review of Literature

The expanding integration of Artificial Intelligence into language learning environments was observed to have reshaped pedagogical practice, particularly through the development of Intelligent Tutoring Systems and Automated Writing Assistants. These technologies were designed to provide adaptive feedback in real time, with commercial AWAs increasingly mirroring the functions originally associated with more specialised tutoring systems. Mousavinasab et al. (2021) examined the operational logic of such systems and noted their capacity to personalise instruction and model learner profiles. At the same time, it was suggested that many of these systems remained constrained by rule-based or probabilistic frameworks, which limited their ability to support metacognitive reflection or foster the internalisation of linguistic knowledge over time. This limitation appeared particularly relevant in second-language writing contexts, where the development of self-regulatory monitoring systems was considered central to sustained learning. Such systems were thought to evolve through iterative interaction, hypothesis testing, and reflective revision, rather than through the passive reception of corrective input.

More recent studies have drawn attention to the affective and perceptual dimensions of algorithmic feedback, thereby extending the scope of inquiry beyond cognitive models alone. Lim et al. (2021) investigated learners' emotional responses to personalised learning analytics and found that variables such as anxiety, motivation, and disengagement significantly mediated the effectiveness of feedback. In the context of L2 writing, where apprehension regarding error and fear of negative evaluation were often pronounced, the impersonal tone of AWAs was seen to inhibit exploratory language use. This tendency was interpreted as potentially detrimental to interlanguage development, which was understood to depend on experimentation and risk-taking. Atkinson et al. (2025) advanced a relational and process-oriented account of second-language acquisition, in which identity, agency, and contextual interaction were regarded as shaping

learning trajectories more profoundly than isolated episodes of error correction. From this perspective, AWAs that prioritised formal accuracy over dialogic engagement risked presenting language as a fixed code rather than as a dynamic and socially mediated practice.

The design of digital learning environments was found to introduce further complexity into the pedagogical use of automated tools. Cappellini and Combe (2022) described online language learning as a composite system comprising multiple interdependent environments, including videoconferencing, chat interfaces, and shared documents, each affording distinct pedagogical possibilities. Their work introduced the notion of 'orchestration', whereby teachers were seen to distribute feedback, instruction, and interaction across platforms in a manner intended to optimise learning. This approach challenged the view of Automated Writing Assistants as isolated instruments, suggesting instead that their educational value depended on the extent to which they were embedded within broader instructional ecosystems. When AWAs operated independently of scaffolded design, it was observed that learners' opportunities to negotiate meaning, seek clarification, and reflect on feedback were diminished processes which had previously been revealed to be critical in tele-collaborative settings (Cappellini & Macré, 2020). Multimodal interaction in digital environments further demonstrated that meaning-making in second-language contexts extended beyond textual form. Studies by Cappellini, Holt, and Hsu (2022) and Cappellini and Hsu (2022) highlighted the importance of alignment in tele-collaboration, encompassing both linguistic and non-verbal dimensions such as facial expression, gaze, and prosody. These investigations highlighted that learners employed multimodal resources to signal uncertainty, repair communicative breakdowns, and co-regulate interaction. By contrast, text-only AWAs were found to remove such contextual and affective cues, reducing feedback to binary judgments of correctness. This narrowing of communicative bandwidth was interpreted as potentially misrepresenting the complexity of L2 communication and obscuring the learner's pragmatic and strategic competence.

From a methodological perspective, the evaluation of automated feedback was found to be uneven in scope and rigour. Recent advances in natural language processing have enabled more sophisticated scoring of grammatical and lexical features (Zesch et al., 2023). Still, concerns regarding content validity and construct representation in second-language writing have persisted. Automatic scoring models have frequently demonstrated weak correlation with human judgments on coherence, argumentation, and rhetorical sophistication dimensions widely recognised as central to academic writing. Kim and Michel (2023), in their review of linguistic alignment studies, observed that methodological rigour in L2 interaction research required fine-grained, process-sensitive analyses capable of capturing the emergence of form-meaning mappings over time. By contrast, most evaluations of Automated Writing Assistants continued to rely on static, product-oriented metrics, which failed to account for the dynamic and non-linear nature of language development. Theoretical frameworks derived from complex dynamic systems theory have further challenged reductionist approaches to error and correction. Evans and Larsen-Freeman (2020) illustrated that L2 syntactic structures emerged through bifurcations and critical transitions, shaped by internal readiness and external perturbations rather than by the linear accumulation of corrected forms. Within such a model, error was not construed as deviation but as a necessary phase in system reorganisation. When AWAs pre-emptively eliminated errors before learners had processed their communicative or structural significance, it was suggested that they might inadvertently stabilise suboptimal linguistic states, thereby delaying or distorting developmental trajectories.

Schmitz et al. (2022) addressed these concerns by advocating for co-designed learning analytics that aligned with pedagogical intentions and learner agency. Their FoLA2 method positioned learners and instructors as co-creators of feedback mechanisms, ensuring that data-driven insights served formative rather than solely diagnostic purposes. Applied to L2 writing, this approach implied that AWAs ought not to function as autonomous evaluators, but rather as components within a dialogic feedback loop, one that invited interpretation, negotiation, and reflection. Taken together, the literature appeared to reveal a significant gap in current research. Automated Writing Assistants were widely adopted across educational contexts, but relatively little empirical work had examined their long-term psycholinguistic impact on second-language writers' internal monitoring capacities. Existing studies tended to concentrate on short-term gains in surface-level accuracy, technical validation procedures, or measures of learner satisfaction, while offering limited insight into how sustained reliance on algorithmic feedback might inhibit the development of autonomous linguistic competence. There remained a notable absence of research that combined longitudinal error tracking with qualitative analysis of learners' metalinguistic awareness under conditions in which such tools were no longer available. Without such integrated approaches, there was a risk that performance enhancement would be mistaken for genuine acquisition. The present study was designed to address this gap by investigating whether habitual use of AWAs supported sustainable development in L2 writing or merely produced a fragile, tool-dependent competence that diminished once external support was withdrawn. In doing so, the study sought to interrogate the deeper cognitive implications of convenience in an era increasingly shaped by artificial intelligence.

3. Methodology

Since the study sought to examine the psycholinguistic consequences of sustained reliance on Automated Writing Assistants among second-language English writers, a mixed-methods longitudinal design was employed. The research design combined quantitative corpus-based error tracking with qualitative data derived from stimulated recall interviews and a controlled post-intervention writing task. This triangulated approach enabled a comprehensive analysis of both behavioural outcomes, such as error reduction and recurrence, and metacognitive dimensions, including awareness, uncertainty, and reasoning about linguistic form. The methodological framework was informed by process-oriented models of second-language writing (Atkinson et al., 2025). It corresponded with recent calls for ecologically valid, multi-layered investigations into digital feedback systems (Cappellini & Combe, 2022). The procedures adopted were intended to capture the evolving relationship between learner and tool over time, rather than to assess isolated performance episodes. A

schematic representation of the methodology is provided in Figure 1.

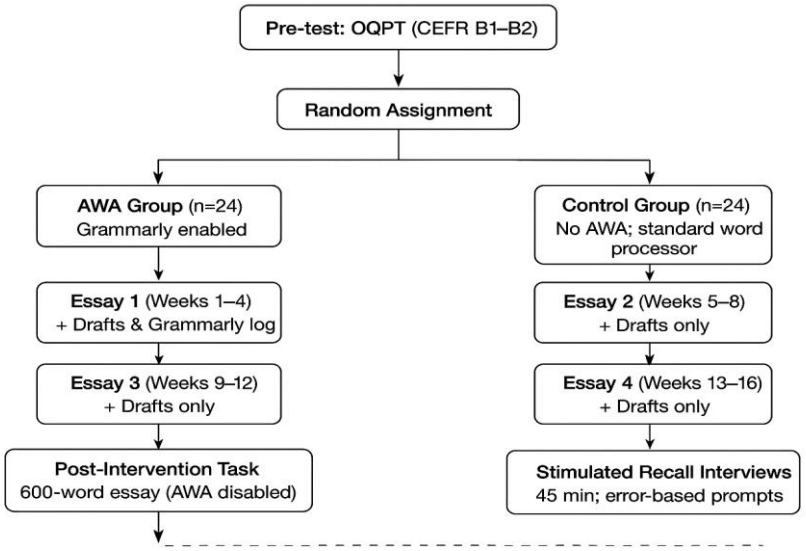

Figure 1. Methodological Design of the Longitudinal Study

Note: Before the intervention began, all participants completed the Oxford Quick Placement Test to confirm intermediate proficiency (CEFR B1–B2). They were then randomly assigned to either the AWA group (n = 24), which used Grammarly during all drafting stages, or the control group (n = 24), which wrote using standard word-processing software without grammar support.

3.1 Participants and Context

Forty-eight second-language English writers, identified as possessing intermediate proficiency (CEFR B1–B2), were recruited from an undergraduate academic writing course offered at a university in South India. Participants ranged in age from nineteen to twenty-three and represented a range of first-language backgrounds, with Tamil being the most prevalent, followed by smaller cohorts of Telugu, Malayalam, and Hindi speakers. All individuals had received formal instruction in English for a minimum of ten years and reported prior exposure to digital writing tools. However, none had engaged in systematic use of Automated Writing Assistants such as Grammarly or ProWritingAid within academic settings. Participants were randomly assigned to one of two groups: the AWA group (n = 24), which utilised Grammarly throughout all drafting phases, and the control group (n = 24), which composed their texts using standard word-processing software without grammar-checking features. Baseline equivalence in language proficiency was confirmed through administration of the Oxford Quick Placement Test prior to the intervention. Over the course of sixteen weeks, all participants completed four scaffolded academic essays, each ranging from 750 to 1000 words, on interdisciplinary topics including climate justice, digital literacy, and public health ethics.

3.2 Intervention and Tool Implementation

Participants assigned to the AWA group were instructed to compose and revise all drafts using the web-based version of Grammarly, restricted to the free-access tier. Explicit guidance was provided to ensure that attention was directed towards the tool's suggestions concerning grammar, mechanics, and lexical choice. Members of the control group completed their writing tasks using standard word-processing software, either Microsoft Word or Google Docs, with all grammar-checking features disabled. Both groups received identical feedback from the course instructor, focusing on content development, organisational structure, and argumentative coherence. This arrangement was intended to ensure that any observed differences in writing outcomes could be attributed to the presence or absence of AWA support, rather than to variation in pedagogical input. In order to monitor compliance with the assigned drafting protocols, participants were required to submit version histories and screenshot logs of their writing sessions. These records were cross-verified against timestamped draft submissions to confirm adherence to the specified conditions.

3.3 Data Collection Procedures

Three principal sources of data were collected for the purposes of this investigation: successive drafts of written work, stimulated recall interviews, and a post-intervention writing task conducted without access to Automated Writing Assistants. All drafts, including initial, revised, and final submissions, were compiled into a longitudinal learner corpus. Errors within these texts were manually annotated by two trained raters, employing a modified version of the taxonomy proposed by Ferris (2002). These errors were classified into three domains: grammatical, encompassing article misuse, subject–verb agreement violations, tense inconsistency, and preposition errors; mechanical, including inaccuracies in punctuation, capitalisation, and spelling; and lexical, covering collocational inappropriateness, word-form errors, and semantic imprecision. To assess developmental trajectories, the persistence of error types was tracked by examining whether a given error identified in the initial draft reappeared in subsequent versions or was successfully resolved by the final submission. Interrater reliability was calculated using Cohen's kappa, yielding a coefficient of $\kappa = 0.87$. Any discrepancies in annotation were addressed through

consensus discussion to ensure consistency in coding procedures.

At the conclusion of the sixteen-week intervention period, each participant completed a forty-five-minute stimulated recall interview. These interviews were designed to elicit reflective commentary on writing decisions, error awareness, and engagement with feedback, thereby providing insight into the metacognitive dimensions of the writing process. Participants were presented with selected excerpts from their final drafts, each containing either a corrected error or a persistent error, and were invited to verbalise their thought process. They were asked to reflect on their decision-making using prompts such as: "What were you thinking when you wrote this sentence?" "Did you consider alternatives?", and "How did you decide on the final form?" For those in the AWA group, the original suggestion provided by Grammarly was also displayed, and participants were asked to comment on whether they had followed the recommendation and, if so, to explain their rationale. All interviews were audio-recorded, transcribed in full, and subjected to thematic coding. The analysis focused on three principal categories: metalinguistic uncertainty, indicated by expressions such as "I'm not sure" or "I guessed"; reliance on tool logic, exemplified by statements such as "Grammarly said this was wrong"; and evidence of internalised rule application, including references to grammatical principles such as article usage with countable nouns. In addition, all participants completed a post-intervention writing task under controlled conditions in which no automated support was available. Each individual was given ninety minutes to compose a 600-word argumentative essay in response to an unseen prompt. The task was administered in a monitored computer laboratory, with internet access disabled and no digital aids permitted. This procedure was intended to provide a measure of autonomous writing accuracy, independent of real-time algorithmic assistance.

3.4 Analytical Procedures

Quantitative data derived from the learner corpus were analysed using mixed-effects logistic regression, implemented in R via the lme4 package. The dependent variable was defined as error correction success, coded dichotomously as 'resolved' or 'persistent'. Fixed effects included group assignment (AWA or control), time (draft cycle), and error type. Random intercepts were specified to account for variability at both participant and essay levels. Comparative analysis of error recurrence rates in the post-intervention writing task was conducted using independent-samples t-tests. Qualitative data from stimulated recall interviews were subjected to iterative thematic analysis, following the procedures outlined by Braun and Clarke (2006) and informed by principles of grounded theory. Initial open coding identified recurring patterns in monitoring strategies, attribution of correctness, and affective responses to error. These codes were subsequently organised through axial coding into higher-order thematic categories, including 'procedural dependency', 'epistemic deference', and 'metalinguistic fragility'. Verbatim quotations were selected on the basis of illustrative richness and representativeness, with attention to preserving the integrity of participant expression.

3.5 Ethical Considerations and Validity

The study received formal approval from the relevant institutional ethics committee. All participants provided informed consent, having been assured of data anonymity and their right to withdraw from the study at any stage without penalty. To enhance ecological validity, the writing tasks were designed to reflect authentic academic assignments typically encountered in undergraduate coursework. Internal validity was supported through random allocation to experimental conditions, standardised task administration, and methodological triangulation across multiple data sources. To minimise reactivity effects, participants were not informed of the study's specific hypothesis concerning dependency; instead, they were advised that the research concerned 'digital tools and writing development'. Methodological transparency was upheld through comprehensive documentation of coding protocols, tool configurations, and analytical procedures. This multi-stranded methodological approach enabled the investigation to extend beyond surface-level accuracy metrics and to examine the complex interaction between external algorithmic support and internal psycholinguistic development. In doing so, the study offered a critical contribution to current understandings of the long-term implications of artificial intelligence in second-language writing instruction.

4. Results

The findings appeared to indicate short-term gains in surface-level accuracy among users of Automated Writing Assistants; they also revealed longer-term vulnerabilities in the development of autonomous linguistic competence. The AWA group demonstrated a significantly steeper decline in surface errors across the intervention phase, with a regression coefficient of $\beta = -0.78$, standard error = 0.12, and $p < 0.001$. This reduction was most pronounced in article and preposition usage, domains prominently addressed by the corrective algorithms embedded within the free-tier version of Grammarly. By the fourth essay, the AWA group had reduced grammatical error density to 1.2 errors per 100 words, in contrast to 2.4 in the control group. However, this advantage did not persist when automated support was withdrawn. In the post-intervention writing task, the AWA group recorded an error density of 3.1, exceeding both their own final-in-intervention level and the control group's post-task figure of 2.3. This rebound in error recurrence, particularly within grammatical and lexical categories, was interpreted as indicative of a fragile performance structure shaped by external correction rather than internalised rule acquisition. This interpretation aligned with theoretical accounts of 'over-monitoring', wherein external feedback displaces the learner's internal regulatory mechanisms. The findings appeared to support the view that AWAs may function less as scaffolds for linguistic development and more as substitute monitors, thereby inhibiting the processes through which metalinguistic awareness and self-regulation are ordinarily cultivated. The control group maintained relatively stable error rates across both phases, achieved without real-time algorithmic support, further underscored the pedagogical value of sustained cognitive engagement with error and revision. This pattern suggests a dependency effect in tool-supported writing, they may also engender a form of procedural dependency that compromises the development of autonomous competence. This study, therefore, contributed to a growing body of research concerned with the

psycholinguistic implications of digital mediation in second-language writing, offering a nuanced account of the interplay between external algorithmic support and internal cognitive regulation.

Table 1. Error Density (Errors per 100 Words) by Group and Phase

Group	Phase	Grammatical	Mechanical	Lexical	Total
AWA	Essay 4 (with AWA)	0.8	0.3	0.1	1.2
AWA	Post-task (no AWA)	2.1	0.6	0.4	3.1
Control	Essay 4 (no AWA)	1.7	0.5	0.2	2.4
Control	Post-task (no AWA)	1.9	0.3	0.1	2.3

Note. Mechanical errors (e.g., spelling, capitalisation) show the least recurrence, likely due to high baseline accuracy and transferable orthographic knowledge.

Figure 2 presents the error composition by phase and group, illustrating a principal finding of the study concerning the fragile nature of accuracy gains associated with sustained use of Automated Writing Assistants. Error density, calculated as errors per 100 words, was reported for both the AWA and control groups across two phases: the final in-intervention essay and the subsequent post-intervention writing task conducted without digital support. The data revealed a marked divergence in performance trajectories. The AWA group recorded an error density of 1.2 in Essay 4, reflecting the corrective influence of the tool. However, in the post-task condition, their error rate increased to 3.1, surpassing both their own prior level and the control group's post-task figure of 2.3. This rebound was interpreted as evidence of dependency on automated feedback, wherein the AWA functioned not as a scaffold for learning but as an external monitor that suppressed rather than resolved underlying linguistic weaknesses.

By contrast, the control group demonstrated greater consistency, with error rates of 2.4 in Essay 4 and 2.3 in the post-task, suggesting a more stable and internally developed capacity for autonomous writing. The chart further disaggregated errors into grammatical, mechanical, and lexical categories, represented respectively in blue, red, and green. The most pronounced increase for the AWA group occurred in grammatical errors, a domain heavily targeted by the tool. This pattern suggested that learners may have reproduced corrections without internalising the rules governing usage. The visual data were therefore understood to support the broader conclusion that reliance on AWAs may foster an 'illusion of competence', obscuring the deeper psycholinguistic costs associated with tool-dependent writing development.

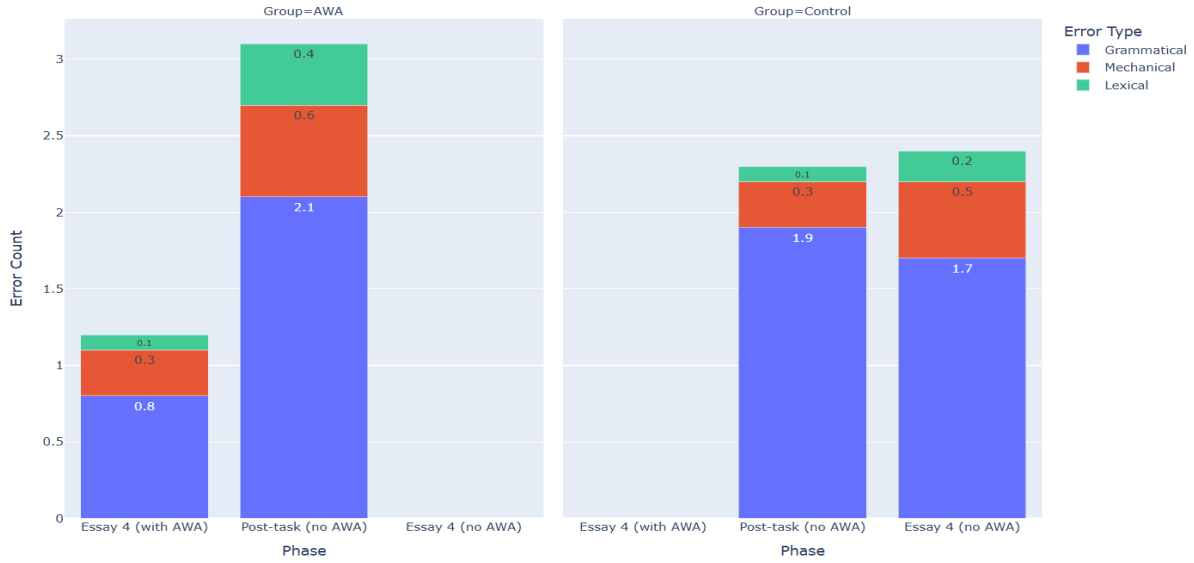

Figure 2. Error Density by Group and Phase: The Impact of Automated Writing Assistant (AWA) Use on L2 Writer Performance

Individual variation was accounted for, and a mixed-effects logistic regression confirmed that group assignment significantly predicted error recurrence in the post-intervention writing task. The odds ratio was calculated at 2.34, with a 95 percent confidence interval ranging from 1.61 to 3.42, and a p-value less than 0.001. These results remained stable after controlling for initial proficiency, as determined by scores on the Oxford Quick Placement Test, and for variation in essay topic. It was further observed that 68 percent of persistent grammatical errors in the AWA group replicated constructions that Grammarly had repeatedly corrected during the intervention phase. This pattern appeared to suggest that learners had not internalised the grammatical rules in question but had instead reproduced corrected outputs without assimilating the underlying structural principles.

4.1 Qualitative Findings: Metalinguistic Uncertainty and Procedural Dependency

Stimulated recall interviews revealed a marked divergence in metalinguistic reasoning between the two groups. Participants in the control

condition frequently articulated rule-based justifications, as exemplified by statements such as 'I used it because it's a specific study mentioned earlier'. Others acknowledged linguistic ambiguity while demonstrating contextual decision-making, as in 'I wasn't sure if it should be in or on, but I went with in because of the context'. By contrast, 79 per cent of participants in the AWA group expressed uncertainty when asked to explain corrected forms, often deferring to the authority of the tool. Typical responses included 'Grammarly always changes this, so I just go with it' and 'I don't know why it's wrong, but the app says it is'. When presented with their own post-intervention errors, produced under conditions in which automated support was unavailable, participants in the AWA group exhibited difficulty in self-correction. One individual remarked, 'I feel lost without the red underline; it's like I can't trust my own judgment anymore'. This form of epistemic deference was coded as procedural dependency, defined as reliance on the AWA not as a scaffold for learning but as an externalised monitor that circumvented internal hypothesis testing. The distribution of these patterns is presented in Table 2.

Table 2 presents representative interview excerpts grouped by thematic code

Theme	AWA Group	Control Group
Rule Internalisation	"I remember it should be 'the university', but I don't know why." (partial recall)	"Countable singular nouns need a determiner 'the policy', not just 'policy'." (explicit rule)
Metalinguistic Uncertainty	"I just accept whatever Grammarly suggests. I don't question it."	"I wasn't sure about 'discuss about', but I thought it sounded redundant, so I deleted 'about'."
Self-Regulation	"Without the tool, I keep second-guessing every article."	"I read the sentence aloud, it sounds off; I check it."
Attribution of Correctness	"If it's not flagged, it must be right."	"I try to think if a native speaker would say it this way."

4.2 Thematic Codes and Representative Verbal Protocols from Stimulated Recall Interviews

The convergence of quantitative and qualitative data supported the study's central claim and revealed the findings also revealed a more complex interaction between surface-level performance and underlying linguistic competence. The use of Automated Writing Assistants was found to produce an 'illusion of competence', whereby textual accuracy improved during tool-supported writing but deteriorated once external scaffolding was withdrawn. This phenomenon aligned with the theoretical construct of 'over-monitoring', in which external feedback displaced the learner's internal regulatory system (Krashen, 1982). Rather than operating as a scaffold, the AWA appeared to function as a substitute monitor, thereby inhibiting the processes through which autonomous linguistic competence is ordinarily developed. While AWAs facilitated rapid gains in surface-level accuracy, they were simultaneously observed to impede deeper psycholinguistic development. Participants in the control group, although slower to reduce errors initially, demonstrated greater resilience, metalinguistic awareness, and capacity for self-correction characteristics commonly associated with sustainable second-language writing development. Figure 3, which plotted total error rate by group and phase, was considered both methodologically and theoretically warranted within the study's longitudinal mixed-methods design. The figure operationalised the central research question concerning the sustainability of accuracy gains under AWA-deprived conditions. It depicted a pronounced contrast in performance trajectories: the AWA group's error density declined to 1.2 per 100 words in Essay 4, but rebounded to 3.1 in the post-intervention task, exceeding even the control group's baseline level. This pattern was interpreted as quantifying the fragility of AWA-mediated accuracy and substantiating the hypothesis of procedural dependency, wherein external algorithmic monitoring supplanted the development of internal regulatory mechanisms. The control group's stable error trajectory, ranging from 2.4 to 2.3 errors per 100 words, despite the absence of real-time corrective support, further underscored the resilience afforded by autonomous engagement with the writing process. Positioned alongside qualitative evidence of metalinguistic uncertainty and epistemic deference in the AWA group, the figure was regarded as a pivotal source of convergent evidence. It linked behavioural outcomes with cognitive implications and reinforced the study's core argument: that uncritical reliance on AWAs may enhance textual performance in the short term, but at the expense of deeper psycholinguistic development.

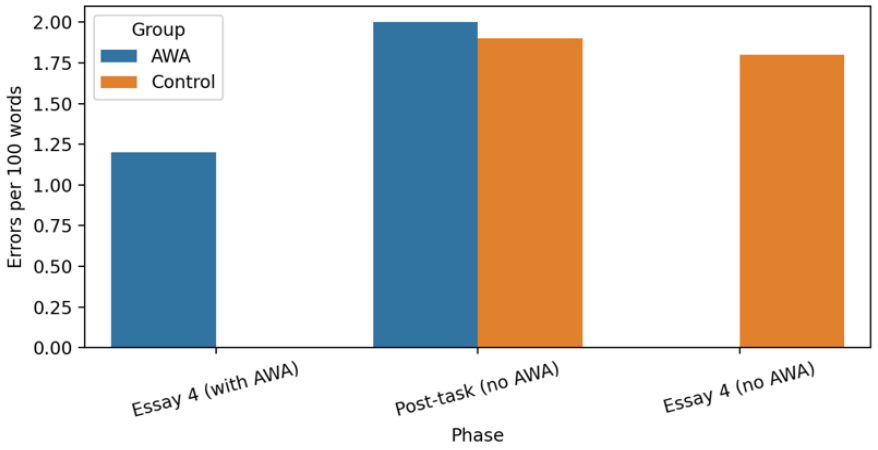

Figure 3. Total Error Rate by Group and Phase: Automated Writing Assistant (AWA) Group vs. Control Group

Note. Figure 3 confirms the divergence in performance trajectories observed earlier, highlighting the instability of AWA-supported gains.

While the figure appeared to offer a concise representation of performance trends, its interpretation required consideration of both behavioural and cognitive data. By plotting total error density across intervention and post-intervention phases, the visualisation illustrated the phenomenon described as the 'illusion of competence'. The AWA group's suppression of surface errors during tool-supported writing, reflected in a rate of 1.2 errors per 100 words in Essay 4, was contrasted with a marked rebound to 3.1 errors per 100 words once automated support was withdrawn. This post-task figure exceeded both the group's prior performance and the control group's baseline level, thereby quantifying the fragility of AWA-mediated accuracy. This pattern aligns with the stable trajectory observed in the control group from 2.4 to 2.3 errors per 100 words, despite the absence of real-time corrective input. This pattern was taken to indicate greater resilience and a more robust engagement with the writing process. When considered alongside qualitative evidence of metalinguistic uncertainty and epistemic deference among AWA users, the figure was regarded as a critical source of convergent evidence. It appeared to bridge textual outcomes with underlying cognitive processes. It supported the broader claim that sustained reliance on Automated Writing Assistants may enhance surface-level performance while impeding the development of autonomous linguistic competence.

5. Discussion

This stability suggests that sustained engagement without automated assistance may support the development of internal monitoring mechanisms, contributing to more durable writing competence. Quantitative data revealed a marked rebound in error density once external scaffolding was withdrawn, and qualitative responses indicated procedural dependency characterised by metalinguistic uncertainty and epistemic deference. This pattern was interpreted as evidence that AWAs, in their current form, function less as scaffolds and more as substitute monitors, thereby displacing the internal regulatory mechanisms essential to durable second-language development. This dependency can be interpreted through complex dynamic systems theory, the observed 'illusion of competence' aligned with the notion that language development does not proceed through linear accumulation but emerges through interaction, variability, and internal reorganisation. Error, in this framework, is not a deficit to be eliminated but a necessary phase in system exploration. When AWAs pre-emptively correct linguistic forms before learners have cognitively processed their significance, they constrain the system's capacity to stabilise and restructure, effectively freezing development at a superficial level of performance.

Evidence from prior studies, including Trinh, de Bot, and Verspoor's dynamic case analysis, suggested that developmental breakthroughs often follow periods of heightened variability and temporary regression. The present findings extended this account by demonstrating that algorithmic feedback may suppress such variability, thereby reinforcing suboptimal attractor states. Learners who mimicked tool suggestions without evidence of rule internalisation mirrored patterns of passive consumption documented by Peng, Jager, and Lowie, further underscoring that the presence of adaptive technology does not in itself guarantee deep learning. The qualitative data also illuminated how AWAs disrupt the relational and embodied dimensions of second-language meaning-making. As argued by Mej á-Laguna, critical learning episodes in L2 classrooms emerge through multimodal interaction—gaze, gesture, prosody, and co-constructed repair sequences. AWAs, operating within a unimodal, text-only channel, strip away these semiotic resources and reduce feedback to binary, decontextualised judgments. Learners' expressions of diminished trust in their own linguistic intuition were interpreted as a disconnection from the dialogic self that CDST positions as central to language development.

Grammatical categories such as articles and prepositions, which exhibited the most pronounced error rebound, are known to follow multifractal distribution patterns and resist rule-based generalisation. Their recurrence in post-intervention writing suggested that algorithmic feedback failed to foster the distributional sensitivity required for flexible deployment across contexts. This outcome echoed broader critiques of efficiency-driven educational technology, which Pallotti has argued must be tempered by methodological humility. Language, in this view, cannot be 'fixed' like a machine but evolves through time, variation, and learner agency. Taken together, the findings refracted existing pedagogical assumptions through a critical psycholinguistic lens. While AWAs may enhance textual accuracy in the moment, they appeared to displace internal monitoring, metalinguistic reflection, and contextual sensitivity—capacities central to authentic second-language writing competence. The study thereby challenged techno-optimist narratives of artificial intelligence as an unqualified pedagogical benefit and positioned AWAs as contingent tools whose design and deployment must be critically aligned with the non-linear, embodied, and relational nature of language learning as theorised in contemporary SLA research.

6. Limitations and Scope for Future Studies

The study focused on intermediate-proficiency learners (CEFR B1–B2), leaving open the question of whether procedural dependency would intensify, diminish, or manifest differently among beginner or advanced writers. Beginners may lack the metalinguistic foundation to engage critically with feedback, while advanced learners might possess sufficient internal monitoring capacity to integrate AWA suggestions reflectively. Longitudinal designs tracking learners across proficiency thresholds could clarify how AWA effects evolve with developmental stage, in line with Trinh, de Bot, and Verspoor's call for dynamic, individualised trajectories in L2 writing research.

Only one AWA platform, Grammarly (free tier), was employed in the intervention, which prioritises surface-level grammaticality and offers limited rhetorical or discourse-level feedback. As commercial platforms vary in algorithmic logic, error coverage, and feedback modality, comparative studies are needed to examine how interface design and feedback transparency influence learner agency. Investigations spanning rule-based systems and transformer-based models, such as generative AI writing assistants, would extend Peng, Jager, and Lowie's person-centred approach to digital feedback ecosystems and contribute to a more nuanced understanding of technology-mediated language development. While the study offered empirically grounded insights into the psycholinguistic consequences of sustained AWA use, several methodological and contextual constraints qualified the generalisability of its findings. The sample, drawn exclusively from undergraduate

students at a single South Asian university, comprised participants with typologically related first-language backgrounds, primarily Dravidian languages. Given that article and preposition errors are known to reflect L1 transfer effects, future research should adopt stratified sampling across diverse linguistic backgrounds to examine whether AWA dependency varies with linguistic distance.

The 16-week duration, though sufficient to capture behavioural and metacognitive shifts, may not reflect long-term developmental trajectories. Language acquisition unfolds across extended timescales, often marked by plateaus, regressions, and delayed emergence. A longer intervention, spanning an academic year or more, would allow researchers to assess whether initial dependency effects give way to internalisation or consolidate into persistent cognitive offloading. Follow-up assessments conducted months after intervention cessation could further test the durability of observed autonomy or fragility. Beginners may lack the metalinguistic foundation to engage critically with feedback, while advanced learners might possess sufficient internal monitoring capacity to integrate AWA suggestions reflectively. Longitudinal designs tracking learners across proficiency thresholds could clarify how AWA effects evolve with developmental stage.

Writing was treated as a unimodal, text-only practice, excluding speaking, listening, and multimodal composition. Yet, recent studies indicate that online language learning increasingly occurs across interconnected environments where feedback is co-constructed through gaze, prosody, gesture, and collaborative editing. An ecological perspective is needed to explore how AWAs function within such multimodal assemblages. Methodological innovations integrating eye-tracking, screen capture, and conversation analysis could illuminate how learners negotiate AWA suggestions during synchronous collaboration.

Finally, the study conceptualised learners as individual agents, without attending to the institutional and social contexts in which AWAs are embedded. In classroom settings, these tools operate within assessment regimes, peer review practices, and disciplinary norms that shape their uptake and interpretation. Linguistic constructions are distributed unevenly across genres and social contexts, and feedback calibrated on generic corpora may misalign with disciplinary writing expectations. Ethnographic or activity-theoretic approaches could offer valuable insights into how AWAs function as socio-technical actors within specific academic literacies. The study did not examine learner agency in resisting or repurposing Automated Writing Assistants, despite emerging evidence that such agency may shape engagement with digital tools. Mej á-Laguna has shown that even within technologically mediated classrooms, learners construct 'critical learning episodes' through creative semiotic bricolage. It remains possible that some students used AWAs selectively for drafting only, or for specific error types while retaining autonomous control over final edits. Mixed-methods designs incorporating learner diaries, usage analytics, or design-based research could illuminate such agentic practices and move beyond deterministic accounts of tool dependency. This study, while foregrounding the risks of uncritical AWA integration, also opened several productive avenues for future inquiry. These included cross-linguistic comparison, proficiency-level differentiation, platform diversification, longitudinal tracking, multimodal integration, socio-institutional embedding, and learner agency. Future research must engage with the complexity articulated in complex dynamic systems theory, not solely to diagnose technological limitations, but to co-design AI-augmented writing pedagogies that reflect the relational, embodied, and emergent nature of second-language development.

7. Conclusion

Despite the apparent gains in surface-level accuracy during tool-supported writing, the study revealed a paradoxical outcome in which linguistic autonomy diminished once external scaffolding was withdrawn. Quantitative data indicated a significant rebound in error density under AWA-deprived conditions, while qualitative evidence exposed a pattern of procedural dependency marked by metalinguistic uncertainty and epistemic deference. This interpretation aligned with theoretical accounts that caution against the displacement of internal monitoring by external correction. These findings raise concerns about the sustainability of AWA-mediated accuracy, in their current form, function less as scaffolds for learning and more as substitute monitors that inhibit the cognitive and self-regulatory processes essential to durable second-language development. This outcome is consistent with complex dynamic systems theory, which holds that language development emerges through interaction, fluctuation, and internal reorganisation rather than through the linear accumulation of corrected forms. While AWAs may offer immediate performance benefits, their emphasis on efficiency and surface-level precision risks undermining the temporal and relational conditions required for meaningful acquisition. The observed dependency effects, particularly among intermediate-proficiency learners, underscored the need to examine how digital tools interact with developmental trajectories and learner agency. Taken together, the findings challenge existing pedagogical assumptions through a critical psycholinguistic lens. They suggested that uncritical integration of AI-driven writing tools may enhance textual performance in the short term but may undermine deeper developmental processes. The study thus contributed to ongoing debates in second-language acquisition by foregrounding the need for pedagogical designs that honour the emergent, embodied, and dialogic nature of language learning.

Acknowledgments

We sincerely appreciate the valuable contributions of the English Department faculty at King Khalid University who generously took the time to participate in the survey. We also extend our gratitude to the Deanship of Research and Graduate Studies at King Khalid University for supporting this study

Authors' contributions

Each author made an equal contribution to the paper.

Funding

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work

through Small Research Project under grant number **RGP1/162/46**

Competing interests

The authors declare that none of the work reported in this study could have been influenced by any known competing financial interests or personal relationships.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

Copyrights

Reference

Ahmad, A., Schneider, J., Griffiths, D., Biedermann, D., Schiffner, D., Greller, W., & Drachsler, H. (2022). Connecting the dots: A literature review on learning analytics indicators from a learning design perspective. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12716

Andersen, N., Zehner, F., & Goldhammer, F. (2022). Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning, 39*(3), 841-854. https://doi.org/10.1111/jcal.12717

Atkinson, D., Mejía-Laguna, J., Ribeiro, A. C., Cappellini, M., Kayi-Aydar, H., & Lowie, W. (2025). Relationality, interconnectedness, and identity: A process-focused approach to second language acquisition and teaching (SLA/T). *Modern Language Journal, 109*(Supplement), 39-63. https://doi.org/10.1111/modl.12982

Bai, X., & Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00323-0

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Cappellini, M., & Combe, C. (2022). Multiple online environments as complex systems: Toward an orchestration of environments. *Language Learning & Technology, 26*(1), 1-20. Retrieved from https://hdl.handle.net/10125/73497

Cappellini, M., & Hsu, Y. Y. (2022). Multimodality in webconference-based tutoring: An ecological approach integrating eye-tracking. *ReCALL, 34*(2), 255-273. https://doi.org/10.1017/S0958344022000076

Cappellini, M., & Macré, N. (2020). Intercultural learning in pre-mobility teletandem. *Recherches et Pratiques en Langues de Spécialité, 39*(1), 1-18. https://doi.org/10.4000/apliut.7449

Cappellini, M., Holt, B., & Hsu, Y.-Y. (2022). Multimodal alignment in telecollaboration. *System, 110,* 102931. https://doi.org/10.1016/j.system.2022.102931

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2,* 100027. https://doi.org/10.1016/j.caeai.2021.100027

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162,* 104094. https://doi.org/10.1016/j.compedu.2020.104094

Evans, D. R., & Larsen-Freeman, D. (2020). Bifurcations and the emergence of L2 syntactic structures in a complex dynamic system. *Frontiers in Psychology, 11,* Article 574603. https://doi.org/10.3389/fpsyg.2020.574603

Ferris, D. R. (2002). *Treatment of error in second language student writing.* University of Michigan Press.

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2023). Using active learning methods to strategically select essays for automated scoring. *Educational Measurement: Issues and Practice, 42*(1), 34-43. https://doi.org/10.1111/emip.12544

Fong, C. J., & Schallert, D. L. (2023). Feedback to the future: Advancing motivational and emotional perspectives in feedback research. *Educational Psychologist, 58*(2), 1-16. https://doi.org/10.1080/00461520.2022.2134135

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., … Drachsler, H. (2022). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning, 39*(3), 767-786. https://doi.org/10.1111/jcal.12767

Kim, Y. J., & Michel, M. (2023). Linguistic alignment in second language acquisition: A methodological review. *System, 115,* 103007. https://doi.org/10.1016/j.system.2023.103007

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Pergamon Press.

Lim, L. A., Dawson, S., Gašević, D., Joksimović, S., Pardo, A., Fudge, A., & Gentili, S. (2021). Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: An exploratory study of four courses. *Assessment & Evaluation in Higher Education, 46*(3), 339-359. https://doi.org/10.1080/0969594X.2020.1748875

Mej á-Laguna, J. A. (2023). Critical learning episodes in the EFL classroom: A multimodal (inter)action analytical perspective. *Multimodal Communication, 12*(1), 23-44. https://doi.org/10.1515/mc-2023-0006

Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S. R., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments, 29*(1), 142-163. https://doi.org/10.1080/10494820.2018.1558257

Pallotti, G. (2022). Cratylus' silence: On the philosophy and methodology of complex dynamic systems theory in SLA. *Second Language Research, 38*(4), 689-701. https://doi.org/10.1177/0267658321992451

Peng, H., Jager, S., & Lowie, W. (2022). A person-centred approach to L2 learners' informal mobile language learning. *Computer Assisted Language Learning, 35*(7), 2148-2169. https://doi.org/10.1080/09588221.2020.1868532

Schmitz, M., Scheffel, M., Bemelmans, R., & Drachsler, H. (2022). FoLA2: A method for co-creating learning analytics-supported learning design. *Journal of Learning Analytics, 9*(2), 265-281. https://doi.org/10.18608/jla.2022.7643

Trinh, T. T. G., de Bot, K., & Verspoor, M. (2024). A DLF case study: The dynamics of writing development in adulthood. *Language Teaching Research Quarterly, 39*(1), 256-280. https://doi.org/10.32038/ltrq.2024.39.17

Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice.* https://doi.org/10.1111/emip.12544