Beyond Checklists: The Impact of ChatGPT on Tense–Aspect Accuracy in Iraqi EFL Writing

Assistant Lecturer Sajad Essam Aluaibi
Department of Political Science, College of Law and Political Sciences
Al-Iraqia University
Baghdad, Iraq
Email: sajadessamaleibi@aliraqia.edu.iq

Assistant Lecturer Thulfiqar Abdulameer Hameed
Department of English, College of Arts
University of Al-Qadisiyah
Al-Diwaniyah, Iraq
Email: zulfiqar.abd@qu.edu.iq

Abstract

This quasi-experimental study examined whether tightly supervised, tense-focused use of ChatGPT during revision improves Iraqi EFL students' verb-tense accuracy beyond checklist-guided self/peer editing. Two intact classes (N = 70; CEFR B1–B2) completed parallel pre- and post-test writing tasks over eight weeks of in-class drafting and revision. The experimental class used ChatGPT solely for tense/aspect diagnostics and brief explanations, whereas the control class relied on a teacher checklist. Tense accuracy was defined as 100 − (errors per 100 words)-in other words, higher scores indicate fewer tense errors. An ANCOVA on post-test accuracy, controlling for pre-test performance, showed a large advantage for the ChatGPT group, $F(1, 67) = 101.36$, $p < .001$, partial $\eta^2 = .60$; estimated marginal means indicated an adjusted difference of ≈ 11.5 points. Convergent gains analyses likewise favored ChatGPT: improvement ≈ 23.46 vs. 13.20 points (difference ≈ 10.26; Welch's $t \approx 7.94$, $p < .001$; Hedges' $g \approx 1.88$). Assumption checks were satisfactory. Because the post-test was completed without devices, improvements reflect independent control rather than tool dependence. Findings align with SLA accounts emphasizing noticing and pushed output and indicate that constrained, explanation-oriented LLM feedback can scale effective written corrective feedback in large, resource-constrained EFL classes. Implications for practice include brief, supervised AI-assisted revision windows paired with student-owned rewriting; limitations concern intact-class assignment and a single grammatical target.

Keywords: ChatGPT; EFL writing; tense/aspect; written corrective feedback

ما بعد قوائم التحقّق: أثر ChatGPT في دقّة الزمن والهيئة في كتابة متعلّمي الإنجليزية كلغةً أجنبية في العراق

م.م. سجاد عصام العيبي
قسم العلوم السياسية، كلية القانون والعلوم السياسية
الجامعة العراقية

العـــدد 20A شباط 2026
No.20A Feb. 2026
المجلة العراقية للبحوث الإنسانية والإجتماعية والعلمية
Iraqi Journal of Humanitarian, Social and Scientific Research
Print ISSN 2710-0952 Electronic ISSN 2790-1254

بغداد، العراق
البريد الإلكتروني: sajadessamaleibi@aliraqia.edu.iq
م. م. ذوالفقار عبد الأمير حميد
قسم اللغة الإنجليزية، كلية الآداب
جامعة القادسية
الديوانية، العراق
البريد الإلكتروني: zulfiqar.abd@qu.edu.iq

الملخص

تهدف هذهِ الدراسةُ شبهُ التجريبية ما إذا كان الاستخدامُ الخاضعُ للإشراف والمركّز على الزمن/الهيئة لـنموذج الذكاء الاصطناعي ChatGPT أثناء مرحلة المراجعة يحسّن دقّةَ أستخدام الأزمنة لدى متعلمي اللغة الإنجليزية في العراق، مقارنةً بالمراجعة الإرشادية القائمة على قوائم التحقق والمراجعة الذاتية او مع الزملاء. شارك 70 متعلمًا مقسمين الى شعبتين في مهام كتابةٍ متوازية قبلية وبعدية على امتداد ثمانية أسابيع من الدراسة والكتابة الصفّية والمراجعة. استخدمت المجموعة التجريبية ChatGPT حصراً لتشخيص قضايا الزمن/السمة وتقديم شروحٍ موجزة، في حين اعتمدت المجموعة الضابطة على قائمة تحقق من المعلم. جرى تعريف دقّة الأزمنة بوصفها 100 (عدد الأخطاء لكل 100 كلمة)، أي إن الدرجة الأعلى تعني أخطاء أقل في استخدام الزمن. أظهر تحليل التباين ANCOVA في الدرجات البعدية (مع التحكم في الأداء القبلي) أفضليةً كبيرةً لمجموعة ChatGPT: $F(1, 67) = 101.36$, $p < .001$, partial $\eta^2 = .60$، وأشارت المتوسطات الحدّية المقدّرة إلى فرقٍ معدَّل يقارب 11.5 نقطة. كما تقاربت تحليلات المكاسب في الاستنتاج ذاته: تحسّنت مجموعة ChatGPT بنحو 23.46 نقطة مقابل 13.20 نقطة للمجموعة الضابطة (الفرق ≈ 10.26؛ اختبار Welch's $t \approx 7.94$, $p < .001$؛ حجم الأثر Hedges' $g \approx 1.88$). وكانت افتراضات النمذجة مرضية. ولأن الاختبار البَعدي أُنجز من دون استخدام أدواتٍ رقمية، فإن التحسّن يعكس تحكّماً مستقلاً لا اعتماداً على الأداة. تتسق النتائج مع منظورات اكتساب اللغة الثانية التي تؤكد دور "الملاحظة" و"الدفع إلى الإنتاج"، وتُشير إلى أن التغذية الراجعة من النماذج اللغوية الكبرى عندما تكون مقيدةً ومعززة بالشرح يمكن أن توسّع نطاق فعالية التصويب الكتابي المركّز في الصفوف الكبيرة محدودة الموارد. وتشمل الدلالات التطبيقية اعتماد نافذةٍ قصيرةٍ للمراجعة بمساعدة الذكاء الاصطناعي تحت إشراف المعلّم مع إبقاء تنفيذ التعديلات بيد الطالب؛ بينما تتمثل محددات هذه الدراسة في استخدام شعبٍ قائمةٍ (دون تعيينٍ عشوائي فردي) والتركيز على هدفٍ نحوي واحد.

الكلمات المفتاحية: ChatGPT؛ الكتابة في اللغة الانجليزية؛ الزمن/الهيئة؛ التغذية الراجعة التصحيحية الكتابية

1. Introduction

Classical methodologies are still continuously used in teaching English to Iraqi students, where the Grammar-Translation Method (GTM) continues to be employed in many Iraqi universities and schools. As is well known in this method, teachers focus on grammar explanations and translation activities instead of using English as it is in real-life situations. As one recent analysis describes, "EFL teaching in Iraq was based on the Grammar Translation Method (GTM) of language teaching." (Hassan et al., 2015, p. 1). This traditional, form-focused education-while systematic-often fails to help students apply grammatical rules when writing extended texts. In simpler terms, classroom practice does not always align with the authentic demands of writing (Richards and Rodgers, 2001).

It shows that one of the most obvious problem areas for Iraqi EFL learners is verb-tense use. Previous studies have claimed that many Iraqi EFL college students are

having trouble managing their tense levels in the writing of English. As Al-Shujairi and Tan (2017) note, “students have serious problems with the usage of verb tenses, articles, and prepositions” (p.122). In his research (Faeq, 2023), he realized that students constantly mix up certain tenses; for instance, he observed that many of them confuse the simple past and present perfect. In addition, Al-Musawi and Kareem (2024) found that errors in the tense persisted even after several years of formal instruction. In sum, tense and aspect errors are not occasional slips and it is also an ongoing defect of Iraqi EFL students´ English writing. Verb tense is more than a matter of grammatical correctness when it comes to getting right, it is an essential element for coherence in writing. The proper use of the use of tenses further helps the readers understand the order of events, and when they happen, and, the connections between them, making sense of the relationships between, the events as well as the timing and the order in the sequence. The misuse of tense makes reading unclear (Ellis, 2003). Tense is more than a grammatical aspect; it is a communication. As Bardovi-Harlig (2000) pointed out “discourse is a central influence on the distribution of tense–aspect morphology” (p. 335). That is, mastering the tense and aspect requires understanding the situation of usage of form to portray time and subtlety in communication, not having the rule knowledge separately. Since the 1970s, in a series of second language acquisition (SLA) papers, many theories have recommended taking a form-oriented approach to improve accuracy (e.g., verb tense). In Schmidt’s words: “intake is that part of the input that the learner notices” (1990, p. 139). There is less chance of future output correcting errors that went unnoticed. In the same vein, Swain’s (1995) Output Hypothesis asserts producing language (spoken or written) forces learners to engage with language at a higher level, developing interlanguage. According to Swain, being compelled to generate output in the L2 “may force the learner to move from semantic processing to syntactic processing” (1995, p. 128), indicating that students paying attention to how to express their meaning must contend with grammatical form, not just general meaning. By generating output and obtaining feedback, learners are able to explore their language hypotheses and refine their usage accordingly. In the absence of the corrective mechanisms, errors that learners of any language make may become entrenched over time; Selinker (1972) called it "fossilisation", which is a stage where learners stop progressing in specific aspects of language accuracy. In the field of tense usage, fossilisation refers to the repeated and habitual errors in tenses that prove resistant to further educational efforts. As one description of fossilisation explains, learners may “stop learning while their internalised rule system contains rules different from those of the target [language]” (as cited in Ellis, 2008, p. 963). Targeted intervention and feedback are therefore crucial for preventing fossilised errors in verb tenses from taking hold.

The British Council report on understanding English language teaching and learning in the Iraqi context showed that classroom conditions often limit the ability to provide the sort of intensive, individual feedback that is vital for

resolving persistent grammar difficulties. English as a foreign language classes in Iraq are often large and inadequately equipped, with teachers bearing heavy responsibilities. "Class sizes were often much higher than 30" (Borg & Capstick, 2024, p. 42). Under such circumstances, providing detailed, personalised feedback on each student's writing is extremely difficult. As a result, students might learn rules in theory but receive insufficient corrective practice in applying them in their writing. In many cases, learners leave the classroom with only a superficial understanding of grammar and not enough feedback cycles to internalise accurate usage. There is a clear need for tools or techniques that can supplement the teacher's feedback and provide students more opportunities to notice and correct their errors in writing.

In such cases, it is evident that technology must play a role. Recent advances in generative artificial intelligence-such as ChatGPT-enable natural language dialogue and can substantially support writing. Unlike earlier Automated Writing Evaluation (AWE) systems that primarily flagged errors, ChatGPT can interact with learners, diagnose problems, explain underlying rules, and suggest context-sensitive revisions; in other words, it functions as a conversational tutor rather than a mere error detector. Prior research on AI-based feedback is, on the whole, encouraging-that is, the evidence to date points to meaningful improvements in accuracy and metalinguistic awareness.

in his study about the usefulness of ChatGPT in supporting language learning Mahapatra (2024) found a positive impact of ChatGPT on students' academic writing skills; also, he found that students' perceptions of the feedback were "overwhelmingly positive" (p. 1). In large classrooms (as in the case of Iraq) and with good training, "ChatGPT can be a good feedback tool" that helps to fill the gaps left by limited teacher feedback (Mahapatra, 2024, p.1). Researchers found that AI tools can help students write better and feel more confident. Early classroom results suggest that AI feedback gets students more involved in revising and eases their worries about writing (Shi et al., 2025; Li, 2025).

There remains a clear gap in the literature. While prior work has examined automated feedback and AWE tools in language learning (e.g., Ranalli, Link, & Chukharev-Hudilainen, 2018), no studies have, to our knowledge, investigated tense-focused use of ChatGPT to improve verb-tense accuracy-nor have they done so in the Iraqi EFL context. In other words, both the target construct (tense accuracy via a generative AI tutor) and the setting (Iraqi learners) are underexplored. This study seeks to address that gap.

The central research question guiding this inquiry is:

To what extent does supervised, tense-focused use of ChatGPT during revision improve Iraqi EFL students' verb-tense accuracy, compared with checklist-guided self/peer editing?

To answer this question, we conducted a quasi-experimental study with two classes at Al-Mustafa Private Institute for Language Education (Diwaniyah, Iraq). The experimental group revised using ChatGPT with carefully designed prompts

targeting tense consistency and accuracy under instructor supervision; the control group revised via self-editing and teacher checklists. Both groups completed a pretest and a parallel posttest writing task. By comparing the gains in tense accuracy and the adjustment of basic performance, we evaluate whether generative AI feedback helps bridge the ongoing gap between knowledge of rules and their precise application in extended writing.

1.1. The Role of Noticing and Output: SLA Theory Revisited

The effect of the ChatGPT intervention can be interpreted through established SLA perspectives on noticing and output. According to Schmidt's (1990) Noticing Hypothesis, durable learning requires attention to a specific form in context; in other words, learners must consciously register the target feature at the moment of use. In our classes, AI feedback made tense–aspect problems salient precisely when students were revising. For example, one student wrote, "When I reach the market, I realized I forget my wallet." During revision, ChatGPT flagged the tense mismatch and explained that a past-time narrative requires reached and had forgotten-that is, simple past for the sequence of events and past perfect to mark prior occurrence. This focused prompt highlighted a form that might otherwise have gone unchecked. Additionally, a number of participants reported that ChatGPT tended to catch problems that they had not noticed as errors; in doing so, the tool encouraged that type of conscious awareness that the notice account contends is necessary for change. Swain's (1995) Output Hypothesis provides support for this understanding. Both groups generated drafts and went on to revise them, although the revision work they produced varied in quality. For the ChatGPT condition, learners were given immediate, individualized explanations, and this created an instantaneous conversation about learners' language use. Learners then developed a hypothesis (original sentence) and received feedback (confirmation or correction) before reformulating the sentence. This cycle-produce → notice → explain → revise-cohorts the way that output is expected to support development by driving learners beyond meaning-only cognition and toward form–function mapping in context. Conversely, the control group's checklist method had learners or peers recognize problems they simply missed, which would likely limit the opportunities for metalinguistic reflection. There is a concomitant issue of fossilization (Selinker, 1972): recurrent, discourse-inappropriate patterns that develop resistance to instruction when feedback is sporadic or not specific. The same tense error was evident in a number of control scripts, indicating that some of the patterns remained unchallenged in the posttest context. In the experimental group, repeated, focused reminders curtailed the reappearance of the same error across drafts, which is consistent with the idea that regular, specific correction can counteract incipient fossilization by interrupting entrenched habits.

Many works indicated that WCF is effective when focused on a single target and when explanations are understandable (Ellis, 2009; Bitchener & Knoch, 2010). Our intervention is a modern instantiation of focused WCF: ChatGPT was constrained to tense/aspect and asked to provide brief, comprehensible rationales rather than full rewrites. The large adjusted effect and the transfer to an independent posttest without AI suggest that learners internalized the target patterns rather than merely editing with assistance. In short, the findings are theoretically coherent: noticing made the problem visible, pushed output made the correction stick, and focused WCF delivered these processes at scale within normal classroom constraints.

2. Literature Review

2.1. Theoretical Foundations of Tense and Aspect in SLA

Verb tense and aspect have long been central concerns in second language acquisition (SLA). Mastery entails more than memorizing conjugation rules; in other words, learners must grasp how tense–aspect choices encode meaning in context. Research suggests that learners typically acquire tense–aspect contrasts gradually. For example, they often secure basic tense contrasts (e.g., present vs. past) before they reliably deploy more complex systems such as the progressive and perfect (Ellis, 2003). Put differently, tense learning is incremental: a learner may supply correct forms in drills yet fail to use them appropriately in extended discourse until later developmental stages.

A key implication is that learners must acquire not only the forms of tense–aspect but also their discourse functions. Bardovi-Harlig (2000) argues that the distribution of tense–aspect morphology is shaped by discourse; that is, knowing the structure of the past perfect does not guarantee knowing when a narrative context requires it. Tense–aspect acquisition therefore reflects an interplay of form, meaning, and use (Bardovi-Harlig, 2000). Early on, learners may default to the simple past in all past-time contexts because the conditions of use for the present perfect or past perfect are not yet internalized (Ellis, 2003). Over time, with sufficient input and feedback, these mappings are refined step by step-and learning changes step from rule knowledge to context-sensitive control. In this refinement, attention is an essential element. Schmidt (1990) notes in his Noticing Hypothesis that input only becomes intake when learners notice certain characteristics, meaning awareness precedes durable change. If a student habitually writes “She go to school yesterday,” declarative knowledge of the “-ed” rule is not enough; progress typically occurs when the learner perceives the gap between “go” and the target “went”. Instruction and feedback that make tense–aspect choices salient, that is, that highlight the relevant contrasts at the moment of use, facilitate this noticing.

Production (output) also matters for consolidating grammatical development. Swain (1995) argues that attempting to express meaning in the L2 can push learners from primarily semantic processing to syntactic processing; put

differently, trying to say something often reveals what one cannot yet say and prompts revision. For example, when narrating a story, a learner might begin in the past, shift inadvertently to the present for background information, and then-through feedback or self-monitoring-recognise the tense-sequence inconsistency. Opportunities to produce language with feedback provide chances to test hypotheses about form–function relations and, in turn, to refine output toward target-like use.

Without such feedback, some errors risk becoming fossilised-relatively stable features of interlanguage that persist despite exposure (Selinker, 1972; Han, 2013). In tense–aspect, fossilisation may appear as the routine use of the simple present for past events (e.g., "Last year I visit London") and the failure to correct that pattern over time. Preventing fossilisation requires deliberate practice and targeted corrective feedback (Lightbown & Spada, 2013). Empirical work shows that focused feedback on specific grammatical targets can yield durable gains; for instance, Bitchener and Knoch (2010) reported improvements that were retained over ten months. The challenge, however, is practical: in many classrooms-particularly in the Iraqi context-large class sizes and limited time make it difficult to provide the sustained, individualised feedback that tense–aspect development requires.

2.2. Technology-Enhanced Feedback in L2 Writing

Over the past two decades, technology has increasingly supported language learning, especially by providing feedback on writing. Automated writing evaluation (AWE) tools (e.g., Grammarly, Criterion, MyAccess) apply natural language processing to flag errors and generate scores. In both L1 and L2 contexts, AWE delivers immediate feedback and performs well on many surface features (e.g., subject–verb agreement, simple tense forms). However, significant limitations are also identified in the literature. For these reasons Chen and Cheng (2008) show that AWE feedback focuses on text length, surface correctness, and offers little explanation or contextualization. However, as other analyses later indicate: AWE tends to raise red flags and suggest improvements but fails to elucidate how a form is faulty or how it fits in with a discourse context (Grimes & Warschauer, 2010). As Ranalli et al. (2018) point out, automated feedback often classifies errors in short-form and without explicit, metalinguistic explanation (p. 670), in which case it is seen giving the error less explicitness and interaction from a human feedback perspective. It's simple AWE can inform learners what is wrong but rarely triggers conversation about the underlying rule or the discourse conditions of appropriate use, in other words, it identifies symptoms without engaging with causes. The introduction of large language model (LLM) tools like ChatGPT represents a new generation of writing support. Instead of merely flagging problems, LLMs can be asked to diagnose a problem, briefly explain the problem, suggest a fix and answer follow-up questions; in other words, they can maintain an instructional conversation rather than an alert, one-

off. Early results are encouraging: Bonner, Lege, and Frazier (2023) suggest that ChatGPT tailors feedback according to learning capabilities and this would be more attainable-that is, learners' support is more matched to their problem. Practically, an LLM can provide multiple interactive feedback within one sitting - valuable where teacher grading burdens are high. Meanwhile, generative AI raises well-founded concerns, too. At such instances, students may be relying heavily on AI even going so far as to produce whole essays, which runs counter to the goals of writing instruction - for instance, to impart independent autonomy and mastery (i.e., making an instant shortcut, not a scaffold). Mahapatra (2024) describes the significance of explicit training and classroom policies so that AI is a learning aid. A second concern is about accuracy: LLMs can misdiagnose mistakes or provide misleading rationales, especially given vague prompts. Hence, teacher supervision is critical: for example, instructors can vet the quality of prompts, sample AI feedback to check for accuracy, and provide reinforcement when appropriate, thereby calibrating the inputs and the outputs for ethical use (Zhang & Zou, 2023). However, the promise is potentially huge in low-resource settings: AI can offer personalised, real-time feedback while teachers focus on higher order concerns - content, organisation and coherence - and targeted coaching. In conclusion, AWE is useful for quick, shallow feedback but does not feature interpretive and interactive capabilities, or in a word, facilitate deeper learning. LLMs like ChatGPT can provide that interaction, given constraints and oversight. This study implements these safeguards: ChatGPT performs tense/aspect diagnostics and brief explanations, students revise their own text (not get rewritten by the AI), and learning is evaluated on a device-free posttest. That is to say, the design also uses AI as a scaffold, not a replacement, to protect the academic integrity and help build the improvements of accuracy in tense measurement that extend over time beyond the immediate task.

3. Methodology

3.1 Research Design

This study, as we mentioned before, has employed a quasi-experimental pretest and posttest design with two classes serving as comparison groups. Because individual random assignment was not feasible in this instructional setting, entire classes were assigned to conditions reflecting common practice in educational research (Creswell & Creswell, 2018). The experimental group received ChatGPT-assisted revision constrained to tense and aspect feedback; the control group used teacher-provided checklists with self/peer editing. Both groups completed a pretest at the start of the study and a posttest after eight weeks. We used ANCOVA with posttest tense accuracy as the dependent variable, group as the independent variable, and pretest accuracy as a covariate; in other words, group effects are evaluated after controlling for initial levels, which is more rigorous than comparing unadjusted posttests (Tabachnick & Fidell, 2019). As a complementary check, we also compared gain scores (post − pre) between groups and reported effect sizes-partial η^2 for the ANCOVA and Hedges' g for gains.

Assumptions (homogeneity of regression slopes, normality of residuals, and homogeneity of variance) were examined prior to interpretation; that is, model prerequisites were verified before drawing conclusions.
In sum, the design tests whether ChatGPT-assisted revision yields measurable improvements in tense accuracy beyond what could be attributed to prior proficiency or chance.

3.2 Participants

Participants were 70 Iraqi EFL learners enrolled in two classes at the Al-Mustafa Private Institute for Language Education (Al-Diwaniyah, Iraq). For each class there were 35 students and each class was assigned to the experimental or control condition. Ages ranged from 17 years to 26 years. A placement assessment in relation to the CEFR identified those who were classified as B1–B2, which indicated that they were able to manage routine communication in English but still experienced difficulty with grammatical accuracy in extended writing. Most participants studied English for six or more years (i.e. with a large amount of exposure to grammar rules but relatively fewer opportunities for communicative use), reflecting a typical situation for Iraqi EFL pupils (Al-Shujairi & Tan, 2017). The classes consisted of a mix of genders (39 female, 31 male). A brief survey of background data indicated that approximately 20% had utilised AI tools like ChatGPT previously, usually for casual vocabulary look-ups or translation-in other words, for most students, this study was their first structured use of ChatGPT for revision feedback. For comparability, the same instructor taught both classes, using the same syllabus, task order and schedule. That is, there was only a systematic difference among groups that differed in the way of revision-ChatGPT-assisted vs checklist/peer editing-thus minimising potential confounds around teaching style, curricular content etc.

3.3 Materials

The study utilized writing tasks and resources intended to promote tense–aspect use instead of generic writing practice. The pretest and posttest prompts were parallel in structure and length (each 180–200 words). The pretest asked students to narrate “A memorable day in your past,”, which primarily elicited past-time forms. The posttest prompted students to write about "Plans and Expectations for the Next Five Years," which encouraged the use of future-oriented and present verb forms, such as "will" or "going to," for plans, and, when appropriate, the present perfect tense for summarising prior achievements. This pairing sampled multiple tense categories across tasks rather than a single structure.
Between the pretest and posttest, both groups completed weekly practice tasks (short narratives, opinion paragraphs, and descriptive pieces) that encouraged contrastive use of simple vs. progressive and simple vs. perfect forms. The control group received a teacher-designed tense checklist for self/peer editing. The experimental group accessed ChatGPT on institution-approved devices (laptops or phones), with use restricted to tense–aspect feedback only. Students were not permitted to use ChatGPT for vocabulary enrichment, content generation, or

translation-that is, the tool was constrained to form-focused support. A standard prompt was provided to maintain focus; for example:

"Please examine my paragraph for verb tense and aspect only. List each tense/aspect issue, provide a brief explanation, and suggest a correction. Do not rewrite the paragraph or add ideas."

In other words, the prompt channelled the AI toward diagnostic, metalinguistic feedback rather than text production. This constraint ensured that AI feedback facilitated noticing of tense–aspect issues while learners retained ownership of their revisions.

3.4 Instruments

Tense–aspect accuracy coding. Following established guidance for written corrective feedback (Bitchener & Ferris, 2012), a coding scheme categorized tense-related errors into four types:

1. Wrong tense (e.g., I go yesterday → I went yesterday).
2. Wrong aspect (e.g., I was finish my work → I was finishing my work).
3. Auxiliary misuse (e.g., He don't went → He didn't go).
4. Tense-sequence shift (unmotivated shifts within a narrative).

Each text received an accuracy score defined as:

$$\textbf{\textit{Accuracy}}\ (\%) = \mathbf{100} - \left(\frac{\textbf{\textit{Errors}}}{\textbf{\textit{Total words}}} \times \mathbf{100}\right)$$

Example: 200 words with 6 tense errors → 6 ÷ 2 = 3 errors per 100 words → Accuracy = 97%.

Background questionnaire. Prior to the intervention, students completed a brief survey (age, gender, years of English study, prior AI exposure) to contextualize results and identify potential covariates (e.g., previous AI use).

Post-study attitude survey (experimental group). After the intervention, experimental-group students rated perceived usefulness of ChatGPT, clarity of explanations, and comfort with AI in language learning. These responses provided learner-perspective evidence to complement quantitative outcomes**.**

Interrater reliability for the coding procedure is reported in (3.6.)

3.5. Procedure

The study ran for eight weeks as follows:

- Week 1 (Orientation & Pretest). Students were briefed on the study and provided informed consent. The pretest writing task was administered under exam-like conditions: handwritten, no dictionaries or devices.
- Weeks 2–7 (Weekly Drafting & Revision). Learners completed six in-class writing tasks (150–180 words). Drafts were then revised using condition-specific procedures:
  - Experimental group. Students used ChatGPT on supervised devices. Use was restricted to tense/aspect feedback with a standardised prompt, e.g., "Please examine my writing for verb tense and aspect only. Identify any errors, provide a

brief explanation, and suggest a correction; do not rewrite my text or add ideas". Instructors monitored screens to ensure all queries remained within scope. Control group. Students revised using a teacher-provided tense checklist (irregular verbs, tense sequence, auxiliaries). Teachers circulated to support task management and clarify checklist items but did not provide direct corrections.

- Week 8 (Posttest). Students completed the posttest writing task under the same exam-like conditions as the pretest (handwritten, no tools).

After collection, all scripts were anonymised and coded by the rating team. This procedure held tasks, teacher, schedule, and time-on-task constant across groups; the only planned difference was the type of feedback used during revision.

3.6 Raters and Reliability

Two MA-qualified instructors in applied linguistics-each with several years' experience teaching grammar and writing in Iraqi EFL contexts-served as independent raters. Their familiarity with local error patterns supported consistent identification of tense–aspect issues; in other words, they were well positioned to recognise recurrent forms of misuse.

Before scoring the main dataset, the raters completed a training and calibration session using 10 practice scripts not included in the analysis. They applied the tense-error coding scheme, discussed discrepancies, and refined operational definitions until the agreement stabilised-that is, the coding rules were clarified and consistently applied.

Interrater reliability was estimated with Cohen's κ, which assesses agreement beyond chance. A κ of .80 or higher is typically interpreted as strong agreement (McHugh, 2012). On a set of 20 randomly selected study scripts, κ = .84, indicating high reliability. When disagreements arose in the main scoring, the raters reconciled by discussion to consensus; put differently, final scores reflect a shared judgement rather than individual bias.

Chapter 4: Results

4.1. Descriptive Statistics

Both classes improved from pretest to posttest, but the ChatGPT class improved more overall.

- Control (no ChatGPT): from 54.71% to 67.91% - average gain 13.20 points.
- Experimental (with ChatGPT): from 59.03% to 82.49% - average gain 23.46 points.

Thus, the ChatGPT class improved by **about 10 points more** than the control class on average. At baseline, the experimental group's pretest mean was slightly higher, but this difference was not statistically significant; Table 1 summarizes group means and gains.

Table 1

Descriptive statistics for verb-tense accuracy (%) by group

Group	Pretest M (SD)	Posttest M (SD)	Gain M (SD)
Control (n=35)	54.71 (SD 9.81)	67.91 (SD 8.75)	13.20 (SD 4.33)
Experimental (n=35)	59.03 (SD 9.84)	82.49 (SD 8.06)	23.46 (SD 6.30)

Note. Accuracy = 100 − (tense errors per 100 words). Gains = Post − Pre.

In short, both groups improved, but the ChatGPT group's improvement was much larger. This pattern is also clear in Figure 1.

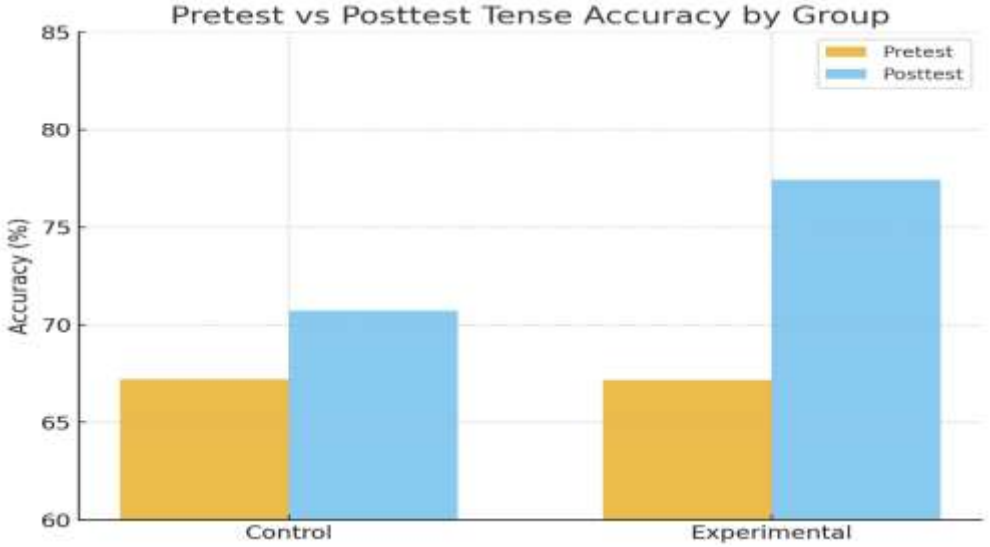

Figure 1. pre-test vs post-test tense accuracy by group

Both bars rise from pretest to posttest, but the experimental group's posttest bar is substantially higher than the control group, mirroring the gains reported in Table 1.

4.3 ANCOVA results

To make a fair comparison, we compared groups' posttest scores while controlling for pretest. The analysis showed a clear and large advantage for the ChatGPT group, $F(1, 67) = 101.36$, $p < .001$, partial $\eta^2 = .60$. In practical terms, the revision method (ChatGPT vs. checklist) accounts for about 60% of the meaningful differences in posttest tense accuracy after adjusting for baseline-an educationally large effect. As expected, pretest score was also strongly related to posttest, $F(1, 67) = 154.57$, $p < .001$, partial $\eta^2 = .70$; however, the ChatGPT advantage remains even when two students begin at the same pretest level.

Table 2

ANCOVA summary: Posttest ~ Group + Pretest

Source	SS	df	F	p	Partial η²
Group	2200.18	1	101.36	< .001	.60
Pretest Accuracy	3355.17	1	154.57	< .001	.70
Residual	1454.32	67	-	-	-

To illustrate the adjusted difference, estimated marginal means (EMMs) showed that, at the average pretest, the control group would be expected to score about 69.5% at posttest (95% CI ≈ 67.9–71.1), whereas the ChatGPT group would be expected to score about 80.9% (95% CI ≈ 79.4–82.5). In other words, holding starting level constant, the ChatGPT condition is associated with an ≈11.5-point higher posttest score.

Assumptions were acceptable: homogeneity of regression slopes (Pre × Group) F = 2.22, p = .141; residual normality (Shapiro–Wilk W = .966, p = .058); and homogeneity of variance (Levene's: Post p = .648; Gain p = .190). These checks support the validity of the ANCOVA results.

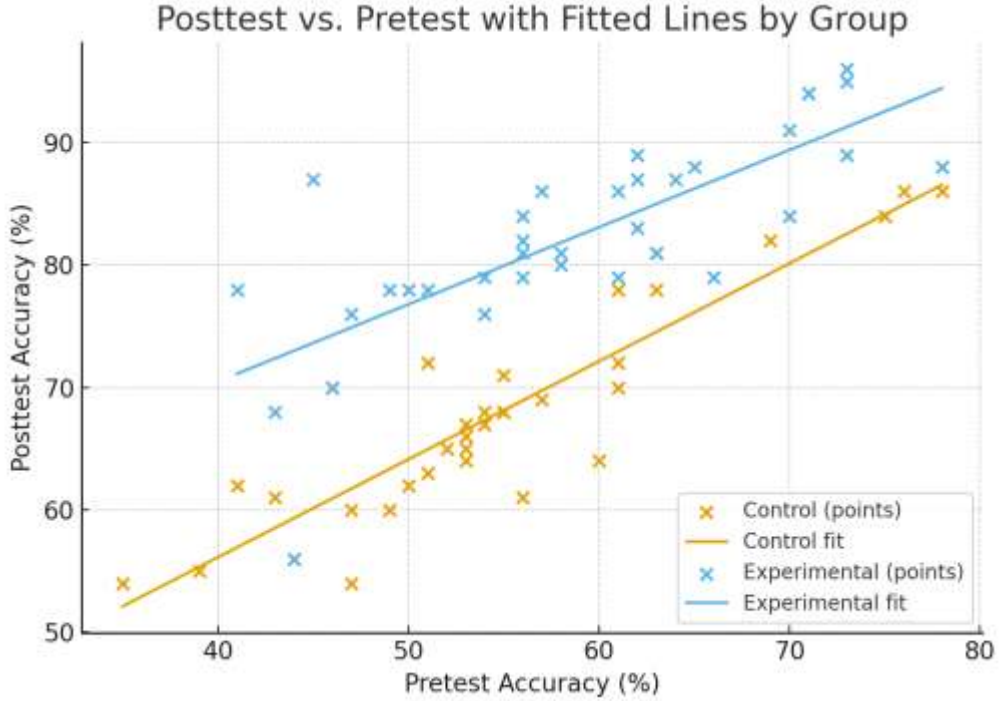

Figure 2. *Posttest vs. pretest with fitted lines by group (ANCOVA visualization).*

As shown in Figure 2, the fitted line for the ChatGPT group lies consistently above the control line across the pretest range, reinforcing the adjusted advantage reported in Table 2.

4.4 Supplementary Analysis

To show improvement directly, we compared gain scores (post − pre) between groups. The ChatGPT group gained about 10 points more on average than the control group (mean difference = 10.26; 95% CI [7.67, 12.85]). This difference was statistically significant and very large in magnitude, Welch $t(\approx 60.25) = 7.94$, $p < .001$, Hedges' g = 1.88. In practical terms, larger improvements were typical-not exceptional-in the ChatGPT condition, which converges with the ANCOVA results.

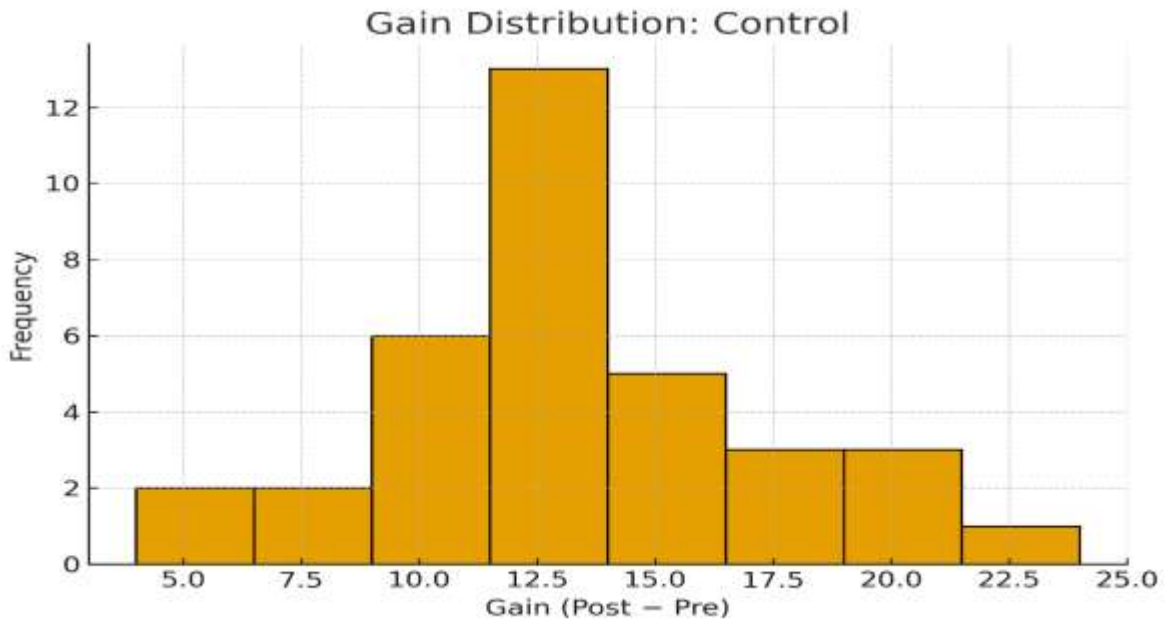

Figure 3a. *Gain distribution (post − pre), control group.*
Most control students improved by around 10–15 points. The bars bunch near the middle, showing steady but modest progress, with few very large jumps

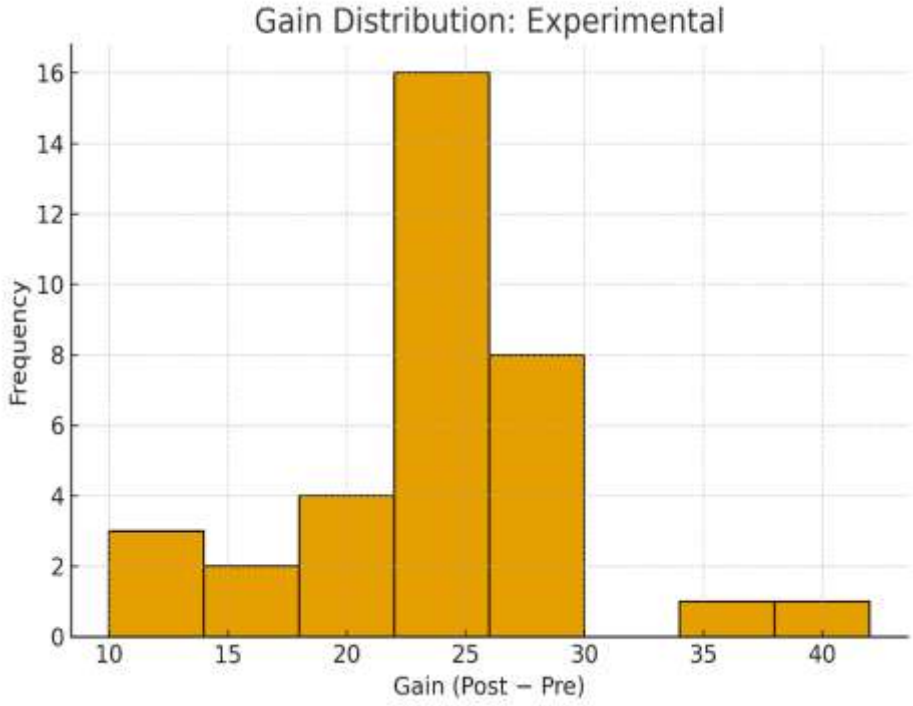

Figure 3b. *Gain distribution (post − pre), experimental group.*

In the ChatGPT class, most students gained in the low-to-mid 20s. The whole plot shifts to the right compared with the control, showing bigger typical improvements

4.5 Summary of Results

Both classes improved from pretest to posttest, but the ChatGPT class improved substantially more. The control group rose from 54.71% to 67.91% (gain 13.20),

whereas the experimental group rose from 59.03% to 82.49% (gain 23.46). The gain difference was therefore ≈10.26 points in favor of the ChatGPT group.

An ANCOVA adjusting for pretest confirmed a large group effect, $F(1, 67) = 101.36$, $p < .001$, partial $\eta^2 = .60$. Estimated marginal means at the sample's average pretest further illustrate the adjusted difference: 69.5% for the control group (95% CI ≈ 67.9–71.1) versus 80.9% for the ChatGPT group (95% CI ≈ 79.4–82.5), an advantage of ≈11.5 points. Assumption checks (homogeneity of regression slopes, approximately normal residuals, comparable variances) were satisfactory.

A supplementary gains analysis converged with the ANCOVA: the mean gain difference was 10.26 points (95% CI [7.67, 12.85]), Welch $t(\approx 60.25) = 7.94$, $p < .001$, Hedges' $g = 1.88$ (very large). In practical terms, larger improvements were typical-not exceptional-in the ChatGPT condition. Since the posttest was done without devices, these gains indicate greater independent tense control rather than dependency on the tool.

4.6 Discussion

In this study I inquired whether tense-focused supervised ChatGPT use during revision was superior to checklist-guided self/peer editing. The evidence makes for a clear conclusion. Once pretest control (adjusted for all variance), ChatGPT group had significantly better posttest accuracy (partial $\eta^2 = .60$), and unadjusted gains were also significantly greater (≈ 23.46 vs. 13.20; difference ≈ 10.26; $g = 1.88$). You confirmed assumptions, and you performed nonparametric checks that, according to your interpretation of the data, are in agreement suggesting the results are robust, and not a distributional artifact. The ANCOVA plot indicates that students who began at the same point consistently had higher levels of achievement after revising with ChatGPT. The general experimental gain distribution shifted rightward -indicating that larger gains in all classes was typical. Pedagogically, a short, tightly scoped AI-assisted revision step produced instructionally meaningful change within eight weeks and typical classroom time. Significantly, the protocol kept AI's use mostly focused on tense/aspect diagnosis and brief explanations, and had students write their own revisions and take part in the posttest themselves without devices. This design is learning support, not tool dependency.

4.7 Linking the Findings to Previous Research

The findings correspond with the previous research on focused written corrective feedback (WCF), directing feedback towards a singular target generally produces more significant and permanent enhancement than dispersing focus across multiple forms. (Ellis, 2009; Sheen, 2007; Bitchener & Knoch, 2010). By restricting feedback to tense/aspect and requiring brief explanations, the present study operationalized that focus and produced a large adjusted effect (partial $\eta^2 = .60$), consistent with claims that focused attention reduces cognitive load and strengthens form–meaning–use links (Bitchener & Ferris, 2012).

The phenomenon also corresponds with core SLA theory. This pattern is consistent with Schmidt's (1990) Noticing Hypothesis where explanatory feedback makes the relevant form salient, which assists learners in identifying discrepancies between their output and target usage. Corresponding with Swain's (1995) Output Hypothesis, students then revised their own sentences, probably facilitating internalization. The observed right-shift in the experimental group's gain distribution is exactly the kind of cumulative improvement expected when noticing and pushed output interact across repeated revision cycles.

Compared with earlier AWE systems that mostly flag errors and offer limited rationale (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Ranalli et al., 2018), a guardrailed LLM can provide brief, context sensitive explanations and respond to follow-ups when used under teacher supervision. The crucial difference here is conceptual, students received a concise why and then made the fix themselves. In the Iraqi EFL context, where tense errors are well-documented (Al-Shujairi & Tan, 2017) and classes are large, this approach offers a scalable way to add individualized, explanation-oriented feedback without displacing teacher focus on content and organization.

Overall, the findings converge with prior research on focused WCF and extend it by showing that a guardrailed LLM can deliver that focus at scale while preserving independent performance on a device-free posttest. When AI is used to explain and focus it can act as a reliable amplifier of effective feedback practice.

Conclusion

This study provides evidence that brief, supervised, and tightly focused use of ChatGPT during revision yields instructionally large gains in tense–aspect accuracy for Iraqi EFL learners. In other words, when AI feedback is constrained to diagnosis and concise explanation-and students must implement changes themselves-learners notice form–meaning mismatches more reliably and revise more productively. Crucially, the advantage persisted after controlling for initial proficiency, and posttest writing (completed without AI) showed that improvements carried over to independent performance rather than tool-dependent drafting.

These findings align with research on focused written corrective feedback and with SLA perspectives that emphasise noticing and pushed output; that is, targeted prompts at the moment of need help convert input to intake and nudge learners from semantic to syntactic processing. Practically, the approach is feasible in crowded classrooms: a standard prompt, a short supervised feedback window, and student-owned rewriting allow teachers to reallocate attention to higher-order concerns (for example, coherence and organisation) while maintaining academic integrity.

At the same time, boundaries matter. The design's strength lies in its constraints (scope-limited prompts, teacher oversight), not in unrestricted text generation. Limitations include intact-class assignment and a single grammatical target.

Future work should examine durability over longer intervals, transfer to additional forms (that is, articles and prepositions), and boundary conditions for effective scaffolding (for example, proficiency level, prompt design, and oversight intensity). Put differently, when used to explain and focus-not to generate text-ChatGPT functions as a reliable amplifier of beneficial feedback practice in EFL writing.

References

Al-Musawi, D. M., & Kareem, D. K. (2024). Difficulties faced by Iraqi EFL learners in using past tense. *Morfologi: Jurnal Ilmu Pendidikan, Bahasa, Sastra dan Budaya, 2*(5), 297–305. https://doi.org/10.61132/morfologi.v2i5.993

Al-Shujairi, Y. B. J., & Tan, H. (2017). Grammar errors in the writing of Iraqi English language learners. *International Journal of Education & Literacy Studies, 5*(4), 122 130. https://doi.org/10.7575/aiac.ijels.v.5n.4p.122

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Blackwell.

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge. https://doi.org/10.4324/9780203832400

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten-month investigation. *Applied Linguistics, 31*(2), 193 214. https://doi.org/10.1093/applin/amp016

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology, 23*(1), 23–41. https://doi.org/10.56297/BKAM1691/WIEO1749

Borg, S., & Capstick, T. (2024). *Iraq – Understanding English language teaching and learning*. British Council. https://doi.org/10.57884/P0XB-4S40

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112. https://doi.org/10.64152/10125/44145

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal, 1*(1), 3–18.

Faeq, F. M. (2023). Iraq preparatory school students' errors in using the present perfect simple tense and the past simple tense. *Journal of Current Researches on Educational Studies, 13*(1), 61–72.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6).

Han, Z. (2013). Forty years later: Updating the fossilization hypothesis. *Language Teaching, 46*(2), 133–171. https://doi.org/10.1017/S0261444812000511

Hassan, R. F., et al. (2015). Examining the potential integration of Second Life platform into Iraqi language education programme. *International Journal of Trend in Research and Development* [Preprint]. http://repo.uum.edu.my/19534/

Lightbown, P. M., & Spada, N. (2013). *How languages are learned* (4th ed.). Oxford University Press.

Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature, 2*(1), 122–152. https://doi.org/10.1515/dsll-2025-0007

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed-methods intervention study. *Smart Learning Environments, 11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge University Press.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Selinker, L. (1972). Interlanguage. *IRAL: International Review of Applied Linguistics in Language Teaching*, *10*(1–4), 209–232. https://doi.org/10.1515/iral.1972.10.1-4.209

Shi, H., Chai, C. S., & Zhou, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2454541

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144). Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Zhang, R., Zou, D., & Cheng, G. (2023). Chatbot-based learning of logical fallacies in EFL writing: Perceived effectiveness in improving target knowledge and learner motivation. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2220374