

Checkpointed Evaluation in AI-Assisted Integrated Reading-to-Write Revision: Conscious-Access Routines and No-AI Transfer

Hesham Aldamen

Department of English Language and Literature, School of Foreign Languages, The University of Jordan, Amman, Jordan

Mohamad Almashour

Department of English Language and Literature, School of Foreign Languages, The University of Jordan, Amman, Jordan

Marwan Jarrah*

Department of English Language and Literature, School of Foreign Languages, The University of Jordan, Amman, Jordan

Correspondence

Email: m.jarrah@ju.edu.jo

Abstract

Integrated reading-to-write (RTW) tasks in English for Academic Purposes (EAP) ask students to work carefully with sources. Writers must keep the original meaning, connect claims to evidence, and manage stance and voice when using several texts. Generative AI can make revision easier, but it may also produce wording that sounds convincing without enough evidence. This qualitative longitudinal multiple-case study examined how EFL university writers used conscious-access routines during AI-assisted revision and whether these routines later appeared in no-AI writing. Twenty-four third- and fourth-year English majors in a 16-week EAP course completed four RTW cycles. Each cycle included no-AI drafting and a 60-minute Gemini-supported revision session. Two revision conditions kept tool access and time-on-task the same. The checkpointed condition required uptake decisions, short reasons, and confidence ratings. The non-checkpointed condition allowed students to evaluate suggestions as they wished. Six data sources produced 384 coded revision episodes. The findings also showed clear differences between the two conditions. Evidence-first routines increased across cycles, while fluency-first routines decreased. No-AI milestone writing showed more evidence-first instances in the checkpointed condition at Weeks 8 and 12. Structured checkpointing may therefore help students use AI more responsibly by checking sources and evidence before accepting fluent suggestions.

ARTICLE HISTORY

Received: 20 February 2026

Revised: 12 September 2026

Accepted: 22 September 2026

KEYWORDS

Integrated Reading-to-Write, Generative AI, Metacognition, Conscious Access, English for Academic Purposes

How to cite this article (APA 7th Edition):

Aldamen, H., Almashour, M., & Jarrah, M., (2026). Checkpointed evaluation in AI-assisted integrated reading-to-write revision: Conscious-access routines and no-AI transfer. *Language Teaching Research Quarterly*, 56, 84–107. <https://doi.org/10.32038/ltrq.2026.56.05>

Introduction

Integrated reading-to-write tasks are central to English for Academic Purposes (EAP) because they treat academic literacy as evidence-based meaning making across texts. Writers select, connect, and reshape source ideas into an argument while keeping claims within what the evidence permits; outside sources are often restricted (Cumming et al., 2005; Plakans, 2009). EAP performance is judged not only by linguistic accuracy and coherence, but also by fidelity to source meaning and clear links between evidence and claims (Baxter & Neumann, 2023; Chuang & Yan, 2023). Revision is therefore a point at which writers must improve language and organization without losing source meaning, overstating the evidence, or weakening ownership of the text.

Student revision now routinely involves generative AI, and this adds one more layer to the pressure of integrated writing. These tools can help smooth out phrasing or improve rhetorical flow, but fluent wording can be convincing even if the sources cannot support it. In an integrated task, that can produce drift in the meaning of the source material, claims pushing beyond the source, or even simply polished wording that lacks the appropriate evidential weight. AI wording can also cause shifts in hedging, stance, and self-positioning that the student writer would not necessarily choose (Hyland, 2005; Sandstead & Kibler, 2025). Simultaneously, research on L2 writing shows that generative AI can support reflection and a feeling of agency when its use is carefully planned (Michel et al., 2025; Moorhouse et al., 2025; Wang et al., 2025). The common practical problem here is how writers decide what to accept when integrated reading-to-write revision must still be true to the source and evidence.

Our concern is with the self-awareness routines of EFL writers who revise integrated reading-to-write drafts with AI, in an advanced university EAP course. Self-awareness routines here refers to the conscious monitoring students can do or talk about, such as attending to a problem or potentiality, crosschecking it against what the task asks of their text and what the source evidence allows, and then using an explicitly stated cause and a confidence judgment to self-regulate acceptance of the opportunity (Fleming, 2024; Chu et al., 2023; Zanesco et al., 2024). This kind of conscious checking becomes visible in integrated writing when students have to go back to a source, become more narrow in their claims, modify their hedging to reflect the evidence available, or turn down fluent language that is not supported.

Two AI revision ecologies were compared. Both granted writers the same level of access to AI and the same amount of time, yet they differed in what writers had to do prior to accepting an AI suggestion. Each ecology operated throughout a 16-week EAP semester

at a public university in Jordan, during which third- and fourth-year English majors completed four integrated reading-to-write cycles alongside separate no-AI tasks that were designed to test transfer. In Condition A, the transparent and checkpointed ecology, each AI suggestion block prompted an uptake decision plus a brief justification tied to meaning fidelity, evidence alignment, stance or voice, and/or clarity and academic style, along with a confidence rating regarding preservation of the source meaning. In Condition B, writers were not subject to any checkpoints and faced no transparency; they had the same time limits and AI access, but their evaluations were solely their responsibility. The call for more caution regarding the impact of task design and accountability on evaluative practice in AI-mediated revision is supported by this comparison (Michel et al., 2025; Moorhouse et al., 2025; Sandstead & Kibler, 2025; Wang et al., 2025; Yan, 2024). The study addresses four research questions:

RQ1: What conscious-access routines do EFL writers enact during AI-assisted revision of integrated reading-to-write drafts?

RQ2: How do these routines differ between transparent checkpointed and opaque non-checkpointed revision ecologies?

RQ3: How do routines develop across four writing cycles within the semester?

RQ4: Which routines show evidence of transfer to no-AI writing (initial drafts and delayed transfer tasks)?

Literature Review

Integrated Reading-to-Write, Source Use, and Stance

Integrated reading-to-write tasks frame academic literacy as evidence-based meaning making under clear constraints. Writers must choose and connect ideas from the provided sources, turn them into a coherent argument, and keep claims within what those sources support (Cumming et al., 2005; Plakans, 2009). This makes such tasks especially relevant to EAP, where university writing usually requires knowledge claims to be built through engagement with texts rather than personal experience alone. Definitions of “integration” vary, but pedagogy and assessment share a concern with accountable discourse synthesis and with writers’ ability to select, organize, and connect information across texts (Knoch & Sitajalabhorn, 2013; Plakans, 2009; Yang & Plakans, 2012).

How writers use sources appears closely tied to the quality of integrated writing. Stronger argumentative performances tend to keep claims within a tighter scope and to ground them more clearly in source-based reasoning (Chuang & Yan, 2023). Research on source-based writing reaches a similar conclusion: proficiency differences are visible not only in linguistic accuracy but also in the way source ideas are positioned for particular rhetorical purposes (Baxter & Neumann, 2023). Studies of reliability and comparability with writing-only tasks likewise support the relevance of reading-to-write constructs for EAP pedagogy and academic-purpose assessment (Cumming et al., 2005; Gebril, 2010).

Writing from sources also creates a familiar tension. Students need to draw on source language while producing discourse that is recognizably their own. For novice writers, textual borrowing may be part of development, but problems arise when copying or patchwriting takes the place of comprehension and reformulation, because meaning fidelity and attribution can then be compromised (Keck, 2006). Borrowing also varies substantially across individuals in integrated writing assessment. If extensive reuse is not detected, it may make proficiency appear stronger than it is (Weigle & Parker, 2012).

What matters, then, is not simply how much source language is reused but how writers put source material to work. Source use is responsible for the variance in scores in large-scale assessment studies, where stronger texts are those that were selected and merged intentionally, rather than solely relying on narrow copying. Revision concerns that require attention are fidelity and evidential accountability, as taught through EAP. The original meaning should not be compromised by paraphrasing, as a reformulation cannot make lacked credibility claims without upholding the original intent. The EAP's evidence alignment strategy is characterized by both informative and rhetorical elements. Position and engagement resources enable writers to convey degrees of certainty, align themselves with readers as they see them, and handle personal responsibility (Hyland, 2005). The quantity of evidence that can be presented as a whole task restricts the extent to which he or she can express his position. To preserve the credibility of an argument, it is necessary to reevaluate hedging, attribution, and claim scope during revision (Chuang & Yan, 2023; Hyland, 2005). This is a critical situation. In this proof-reading task, citation methods are utilized. Citations go beyond formatting, as they acknowledge concepts and sources and help writers establish relationships with each other. Hence, these are called "citations". Petrić (2007) discovered that the attribution functions of higher- and lower-rated graduate writing varied across different levels, emphasizing the role of citation strategy in wider academic proficiency. Integrative writing involves identifying source material and emphasizing how different sources together support a claim, rather than solely relying on citation form for corrections.

AI-Mediated Revision and Feedback Engagement

Despite the potential benefits of advanced AI in wording, organization, and fluency, it cannot enhance the proof of concept aspect of a revision. Recent research has emphasized the question of how learners respond to AI feedback, rather than whether it is present. The use of AI-led writing can foster deliberation and agency (Michel et al., 2025; Moorhouse et al., 2025) while fluent output can also be adopted without verification to alter meaning or perspective (Sandstead & Kibler, 2025). According to Dong (2024), engagement with AI feedback is a combination of cognitive, affective, and ethical duties. The emphasis on mediation, engagement, and transfer beyond assisted performance is emphasized by Pitychoutis and Spathopoulou (2026), while Kostka and McMartin-Miller (2026) highlighted the importance of basic AI literacy. The central idea of this work is the

evaluative stage, where writers must evaluate a suggestion against its source intent, evidential scope, and intended stance.

This evaluation step can be characterized as such by mindful access. Authors detect a potential modification, appraise it, and regulate the subsequent steps.' Faith can determine the acceptance, check or revision of a suggestion (Fleming, 2024), especially when writers are required to explain their decision and provide an explanation (Maniscalco & Lau, 2012). Additionally, Attention is also important because in AI-assisted RTW movement requires moving between a draft, source texts and generated output. The following are examples of conscious access: source rechecking, refusal of fluent but inaccurate wording; adjustment of hedging; and explicit justification of uptake in relation to evidence and ownership.

Revision Ecologies, Routinization, and Transfer

Whether evaluation has to be made visible is an important difference among AI-permitted revision settings. Research on generative AI in L2 writing shows that classroom framing matters: outcomes differ depending on whether students are expected to deliberate, engage critically, and account for the decisions they make (Michel et al., 2025; Moorhouse et al., 2025; Wang et al., 2025). From an ecological perspective, revision environments can therefore differ in a simple but consequential respect: some require uptake decisions and reasons to be externalized, while others allow that evaluation to remain implicit.

Certain forms of assessment can become a routine through the use of repeated integrated writing. The use of self-regulatory and synthesis strategies by writers is utilized, with performance being linked to their utilization (Yang & Plakans, 2012). When AI-assisted revision repeatedly prompts writers to verify and justify their actions, the stability of source checking and stance calibration can be improved over time (Chuang & Yan, 2023; Plakans & Gebril, 2013). The importance of this potential is heightened in EAP, where writing without AI is frequently evaluated as an independent measure of competence. The reinforcement of conscious-access routines in structured revision should persist in no-AI work, including early drafts and postponed tasks (Fleming, 2024; Yan, 2024). Research on RTW and research on AI-mediated writing both examine learner regulation from different perspectives, but they all deal with the same issue. In RTW studies, more effective writing is often associated with disciplined source selection, synthesis, identification, and alignment of evidence and claims (Chuang & Yan, 2023; Plakans & Gebril, 2013). According to GenAI research, swift linguistic assistance can enhance revision speed while promoting over-reliance or weak evaluation (Barrot, 2024; Dong, 2024). Concerns about learner engagement, mediation, durability, and independent transfer have been identified in AI feedback work but remain unresolved (Kostka & McMartin-Miller, 2026). The process that connects these two areas of research is less well-known, as researchers use limited evidence to regulate AI suggestions, while

repetitive practice can transform routineness into a norm and reveal uncertainty even after AI support is removed. This study monitors evaluative choices during AI-assisted RTW cycles and then scrutinizes their presence in subsequent no-AI writing.

Method

Design and Context

EFL university writers utilized a qualitative, longitudinal, and multiple-case process design to investigate the implementation of conscious-access routines during their revision of integrated reading-to-write tasks with AI. Writing about revision was viewed as an evaluation process, with writers expected to match evidence to meaning, control viewpoint and tone, direct attention or respond to AI suggestions (Baxter & Neumann, 2023; Chuang & Yan, 2023; Sandstead & Kibler, 2025) and decide what to do with AI recommendations. Ancillation of conscious access was operationalized through observable and reportable monitoring that linked uptake decisions with evidence and confidence (Fleming, 2024; Chu et al., 2023). [Note 1].

The contrastive design ensured that AI access and time-ontask were identical, but there was some variation in the need for evaluative monitoring (Table 2). The writers of Condition A were obligated to document an uptake decision, provide a brief explanation and report confidence after every AI suggestion block. With the same AI access, condition B was operational in the identical revision window without any additional checkpoints. Consent was stratified based on their pretest performance and year level after the Week 1 pretEST using block randomization with a random number generator. These resulting assignments would be used for all four cycles. An examination was conducted to determine if the inclusion of evaluation as mandatory and visible had any influence on writing habits, such as accountability talks and revisions (Moorhouse et al., 2025; Sandstead & Kibler, 2025). This setup also provided insight into this.

The study was conducted over a 16-week semester in an advanced academic writing course in the Department of English at a public university in Jordan. Participants were third- and fourth-year English majors. Of the 60 students enrolled, 24 consented to research data collection and analysis, and only their data were included. Ethical approval was obtained before data collection. Participation was voluntary, had no effect on grades, and could be withdrawn without penalty. Non-research alternatives were available for additional activities such as interviews. Participants were assigned anonymized identifiers (P1, P2, P3, and so forth), and identifying details were removed from transcripts and analytic excerpts.

Task and Revision Conditions

All tasks used an integrated reading-to-write format. Students read two thematically related sources and produced a 450–650 word argumentative synthesis without using outside sources. The tasks were designed to elicit discourse synthesis and claim–

evidence alignment rather than recall, reflecting the need in integrated writing to monitor source meaning and evidential scope (Baxter & Neumann, 2023; Chuang & Yan, 2023). Task prompts, source packets, and task rules were retained separately as study materials. Data collection ran from Week 1 through Week 12. No-AI writing was sampled at three points: the Week 1 pretest, the Week 8 posttest, and the Week 12 delayed transfer task. These points maintained continuity in the construct and provided the basis for examining transfer. Four intervention cycles were completed between Weeks 2 and 7, each using the two-meeting sequence shown in Table 1. Students drafted without AI in Meeting 1 and then revised with AI for a fixed 60 minutes in Meeting 2. The no-AI texts from Weeks 1, 8, and 12 were later coded for textual evidence of evidence-first behavior. Interpretation of Week 12 transfer also used contemporaneous self-report from the weekly micro-reflection.

Table 1*Semester Timeline, Task Workflow, and Data-Collection Points*

Week(s)	Activity	Tool Access	Data Collected
1	Pretest writing task	No AI	Writing product
2–7 (4 cycles)	Meeting 1: Draft from two sources (450–650 words)	No AI	Draft (versioned when available)
2–7 (4 cycles)	Meeting 2: Revise Meeting 1 draft (60 min)	Gemini, per condition	Screen recording; AI prompt/output log; revised text
Weekly during Weeks 1–12	Micro-reflection prompts	No AI required	Weekly micro-reflection
8	Posttest writing task	No AI	Writing product
12	Delayed transfer writing task	No AI	Writing product

Note. Week 12 planning actions mentioned in the micro-reflections, including scratch-paper comparison tables, are treated as self-reported strategy descriptions rather than as planning artifacts collected by the researchers.

Gemini was available to both conditions during Meeting 2. To reduce variation in request type, all students worked from the same prompt template, which restricted the tool to revision rather than whole-text generation. They entered the task prompt, Source 1 and Source 2, and the relevant draft excerpt. Requests had to remain source-grounded, avoid material beyond the two provided texts, point to supporting source locations, and acknowledge uncertainty where support was unclear. The interaction logs preserved the prompts and outputs that students actually used. Across sessions, generation settings were kept constant using the most deterministic option available in the interface.

Writers in Condition A, the transparent checkpointed condition, made an explicit decision on each AI suggestion block: adopt, adapt, or reject. They then wrote one sentence explaining that decision in terms of meaning fidelity, evidence alignment, stance or voice, or clarity and academic style, and rated from 0 to 100 how confident they were that source meaning and evidential fit had been preserved. Condition B used the same AI

access and the same revision window, but checkpointing was absent and evaluation remained discretionary (Michel et al., 2025; Moorhouse et al., 2025). Table 2 sets out what was held constant and what differed; the precise checkpoint wording and log fields are retained in the logging forms.

Table 2

Condition Features for AI-Assisted Revision Ecologies

Feature	Condition A: Transparent Checkpointed	Condition B: Opaque Non-Checkpointed
AI platform	Gemini	Gemini
Time-on-task	60-minute revision window	60-minute revision window
Prompt template	Required (prompt template)	Required (prompt template)
Sources available	Two provided sources only	Two provided sources only
Logged per AI block	Prompt + full output (logging forms)	Prompt + full output (logging forms)
Required evaluation fields	Uptake decision + justification + confidence (logging forms)	None required
Evaluation visibility	Externalized at time of uptake	Discretionary, often implicit

Data Sources and Procedures

Across the 24 cases, data collection combined synchronized traces of the writing process with writing products and participant explanations (Chu et al., 2023; Fleming, 2024; Zanesco et al., 2024). Screen activity, AI interaction records, drafts, revisions, milestone texts, weekly micro-reflections, and stimulated recall were brought together so that a routine identified in one source could be checked against behavioral, textual, and reported evidence from the others. Table 3 summarizes the role of each source in the analysis.

Each Meeting 2 session was screen-recorded in full. The recordings captured cursor movement, AI interaction, source consultation, editing, moments of attention drift, and subsequent reorientation; system audio was not collected. These traces formed the primary behavioral record for deciding where one revision episode ended and another began. They also made it possible to link on-screen behavior to particular textual changes by showing when a writer moved among the draft, the source packet, and Gemini before accepting, adapting, rejecting, or independently rewriting material.

For each Gemini request, participants logged the complete prompt and complete output together with a timestamp and the location in the draft. Condition A added the uptake decision, the one-sentence justification, and the confidence rating. Because both conditions worked from the same template, which restricted assistance to revision and required grounding in Source 1 and Source 2, variation in request type was reduced. The logs could therefore be used to examine uptake, verification, and later textual change at a relatively fine level of detail (Baxter & Neumann, 2023; Chuang & Yan, 2023).

The writing archive included every no-AI Meeting 1 draft, its corresponding Meeting 2 revision, and any intermediate versions that were available. We also retained the no-AI

milestone texts from Weeks 1, 8, and 12. These materials allowed comparisons between draft and revision within the AI-assisted cycles, supported trace-to-text linking at specific locations in the draft, and provided separate evidence-first coding for writing produced without AI.

The weekly micro-reflection was deliberately brief. Participants recorded one decision they felt confident about and why, one point of difficulty or confusion, and one verification routine. These entries complemented the behavioral traces with contemporaneous accounts of monitoring, confidence, attention regulation, and occasional planning. When participants described Week 12 planning actions, those descriptions were analyzed as self-report rather than as observed composing behavior.

Table 3

Data Sources, Timing, and Analytic Contribution

Data Source	When Collected	What it Captures	Primary Analytic Use
Screen recordings (Meeting 2)	Each cycle, Meeting 2	On-screen evaluation behavior, source consultation, AI interaction, edits	Episode segmentation; trace evidence for verification, uptake, and attention regulation (AI-assisted revision)
AI interaction logs (logging forms)	Each cycle, Meeting 2	Full prompts, full AI outputs, timestamps, targeted draft location; plus checkpoints in Condition A	Link AI blocks to episodes and text changes; analyze uptake decisions and rationales
Writing artifacts (cycle texts)	Each cycle	Meeting 1 draft; Meeting 2 revised product; intermediate versions when available	Trace-to-text linking; examine outcomes of verification, stance calibration, and evidence alignment
No-AI milestone writing products	Weeks 1, 8, 12	No-AI integrated writing products at baseline, posttest, and delayed transfer	Code textual indicators of evidence-first behavior across milestones; support transfer comparisons
Weekly micro-reflections	Weeks 1–12	Brief self-reports on monitoring, confidence, difficulty, verification (and occasional planning descriptions)	Longitudinal complement; triangulate rationale and attention patterns; interpret self-reported strategy carryover
Stimulated recall interviews (interview protocol)	Selected cycles, post Meeting 2	Trace-anchored explanations of decisions, verification, stance, ownership, attention	Rationale interpretation; strengthen trace-talk-text triangulation

Stimulated recall was used to obtain explanations tied to specific moments in the trace, particularly moments involving evidence alignment, meaning fidelity, stance and voice ownership, or regulation of attention. Time-stamped screen recordings and AI logs served as anchors for the interviews, strengthening both interpretability and trace-to-talk triangulation (Chu et al., 2023; Fleming, 2024; Zanesco et al., 2024). Because participants were responding to concrete clips, they could explain not only what they

changed but why a suggestion had seemed safe, risky, authoritative, or inconsistent with the sources.

The stimulus-recognition of all 24 participants was possible, but the level of interviewing was influenced by the case type. To obtain longitudinal depth, we conducted interviews with eight core cases - four in each condition – after all four cycles of the meetings. Eight individuals, each with a specific condition, made up the rotating cases and were interviewed in two cycles to capture early through middle experience. Within-case development and cross- case variation in AI uptake and revision routines were both emphasized while also being balanced with feasibility in this design (Table 4).

Interviews followed Meeting 2 immediately, lasted no longer than 25 minutes, and were audio-recorded under anonymized IDs. Participants could respond in Arabic, with brief English keywords elicited only when needed to maintain analytic consistency.

Each interview drew on three to five short clips taken from the participant's own trace. The clips sampled rapid uptake, explicit source checking, stance or hedging decisions, a clear change of mind, and, where visible, attention drift followed by recovery. Participants were asked what they were trying to revise, what they checked in Source 1 or Source 2, why they chose a particular degree of certainty, whether the wording still felt like their own, and how they redirected attention when it had drifted.

Table 4

Stimulated Recall Sampling Plan by Case Type and Condition

Case Type	Condition A	Condition B	Interview Coverage
Core Cases	4	4	4 cycles (after Meeting 2 each cycle)
Rotating Cases	8	8	2 cycles (selected to capture early, middle, later experiences)

Arabic interviews were transcribed verbatim. When excerpts were reported in English, they were rendered as analytic glosses and checked by bilingual researchers. All data were kept on an encrypted university drive with restricted access, and the screen recordings remained within the research team.

Data Analysis and Trustworthiness

The analysis did not involve a single endpoint, but rather relied on the evaluation process to determine the output of an AI-aided revision. The process orientation is in line with the research on integrated writing and metacognitive control. Revision decision episodes were split into Meeting 2 data, and no-AI texts from Weeks 1, 8, and 12 were analyzed separately to identify evidence-first behavior textual clues (Baxter & Neumann, 2023; Chuang & Yan, 2023; Fleming, 2024).

Table 5*Analytic Unit, Segmentation Outputs, and Coding Layers*

Component	Definition/Output	Evidence Used
Unit of analysis (Meeting 2)	Revision decision episode: a bounded segment in which a writer encounters a revision opportunity, evaluates it through observable moves, commits to an action, and leaves a trace in the evolving text	Screen trace; AI logs; text versions
Unit of analysis (no-AI milestones)	Coded instance in milestone text: a sentence-level or clause-level segment in Weeks 1/8/12 writing products where evidence use, evidential bounds, or claim-evidence alignment is explicit enough to code	Milestone writing products (Weeks 1/8/12)
Episode triggers	AI output, self-initiated source check, or purposeful rewrite targeting claim, citation mapping, stance calibration, or coherence	Screen trace; AI logs
Segmentation output	Time-stamped episode map per Meeting 2 session for sequencing and density comparisons across cycles and conditions	Screen trace
Coding layers	Action codes (what happened) and rationale codes (why the decision was made)	Screen trace; Condition A checkpoints; stimulated recall; micro-reflections
Trace-to-text linking	Episode-level link from on-screen actions and AI blocks to specific before-after text changes	Screen trace; AI logs; drafts/revisions

No-AI milestone texts were coded at sentence or clause level when evidential bounds, justification-driven claim framing, or calibrated epistemic stance were explicit enough to identify. Percentages reported for these milestones therefore represent the share of coded instances classified as evidence-first.

In Meeting 2 the basic unit was the episode involving a revision decision. Such an episode would start when a writer began to evaluate a particular revision target and could be caused by AI output, a self-initiated source check, or by deliberately rewriting a claim, a citation, a stance choice, or a coherence problem. It would come to an end when the wording for that target had settled down and when the writer's attention had been directed towards something else. The same boundary rules were used in both conditions. Two layers of coding were employed. The action codes explained what could be observed in the trace, such as uptake, source rereading, draft-to-source comparison, claim narrowing, hedging adjustment, citation repair, rewriting, rollback, and the presence of attention markers. The rationale codes dealt with the reasons for the decisions, covering meaning fidelity, evidential alignment and scope, stance and voice ownership, clarity and academic style, the perception of AI authority, and time pressure. The rationale evidence was obtained from the Condition A checkpoints, stimulated recall, and micro-reflections, and the accounts of attention were interpreted in light of meta-aware noticing and recovery. A thematic analysis focusing on recurring patterns of reasoning was carried out

in accordance with a phase-based method (Braun & Clarke, 2006). The initial categories in the codebook were based on the study's constructs, but the early sessions from both conditions were examined repeatedly and this was used inductively to improve the categories. Thus, it was possible to include new patterns without losing sight of the constructs which had framed the study.

Each episode was connected via its draft location to the corresponding screen trace, AI log, and the text changes before and after the episode. In Condition A, the decision to take up the option, the justification for that decision, and the confidence rating were all attached to the same episode. Since Condition B did not call for an explicit evaluation, uptake was only inferred in those cases where the trace supported it and was associated with the nearest relevant textual change. Thus, episode-level comparability was maintained without making any claims about reasoning that had not been observed.

The analysis was carried out on two levels. Firstly, the core cases were tracked over the cycles in order to see how they developed individually. Secondly, cross-case analysis looked at how frequently routines occurred, the order in which they occurred, and the way in which the participants explained them in the different conditions. It was on the basis of these comparisons that the conscious-access routine typology was developed.

Trustworthiness was achieved by connecting interpretations to a number of different types of episode-level evidence. Screen recordings, AI logs, writing artifacts, micro-reflections, and stimulated recall were combined in such a way that any interpretive claim could be associated with a time-stamped segment of the trace, the corresponding entry in the AI system, and a related textual change. When it was available, stimulated recall was used as an extra verification. This decreased the reliance on retrospective explanations alone and ensured that cases were kept in which what the participants said did not completely agree with what was shown on screen.

Before actually writing the code, two coders who were bilingual in Arabic and English worked together on a subset of the sessions from both conditions in order to calibrate the segmentation and to clarify the differences between the codes. Any disagreements were discussed and recorded with a senior adjudicator, and the changes made to the codebook were kept in versioned form. The reliability checks focused on those judgments which could be directly seen in the traces, such as episode boundaries, trigger type, verification moves, and, where the evidence permitted in Condition B, the uptake category.

Episodes which went against the developing pattern were deliberately selected rather than being excluded. All the decisions made regarding sampling, the rules used for segmentation, any revisions to the codebook, the way disagreements were settled, the thematic memos and the comparisons between episodes were recorded in the audit trail (Braun & Clarke, 2006). The analysis of negative cases helped to distinguish between

patterns that occurred repeatedly and those that were isolated, and it prevented overinterpretation when a completed checkpoint failed to be confirmed by substantive verification. The interviewing and coding were carried out by bilingual researchers who were familiar with the instructional setting, and reflexive memos were prepared in order to examine whether the expectations of the institution might be influencing the participants' replies and to tell the difference between interpretations which were supported by evidence and those which still required evidence.

Findings

The analysis incorporates both process and product evidence drawn from 24 participants observed over four AI-assisted cycles and three no-AI assessment points. The materials available comprised Condition A checkpoint logs; screen recordings of the 60-minute Meeting 2 revision sessions; AI logs that included prompts and Gemini outputs; drafts from Meeting 1 and revisions from Meeting 2; no-AI products from Weeks 1, 8, and 12; weekly micro-reflections; and stimulated recall interviews grounded in the trace (see Table 3). Procedures were developed through thematic analysis of checkpoint justifications, reflections, and recall, and were then validated against trace indicators and textual outcomes (Braun & Clarke, 2006).

Altogether, 384 revision episodes were coded, yielding an average of 16 episodes per participant. Episode counts were balanced across condition, with 192 episodes in each group. This structure enabled comparison between the two conditions, tracking of changes across cycles, and linkage of process evidence to performance in the no-AI writing tasks.

Conscious-Access Routines

Six recurring routines were identified in the 384 revision episodes. They differed in the criterion that writers appeared to prioritize when judging AI output. Their definitions, principal trace indicators, and condition distributions are summarized in Table 6.

The sharpest thematic difference concerned what controlled uptake: linguistic plausibility or source evidence. Evidence-first writers treated AI output as provisional. P15, for instance, rejected a claim of a 50% productivity increase when the source supported only “up to 15%,” and P2 changed “must” to “may” because the stronger modality went beyond Source 2. Fluency-first episodes followed the opposite sequence, with polished wording often accepted before any source check. Compliance and time-pressure routines showed that completing the visible task could also take priority over evidential checking. Mixed episodes were especially useful because some writers initially relied on AI wording and then returned to source-based evaluation. The six themes therefore differ mainly in what writers treated as the decisive basis for uptake, not in whether they used AI.

Table 6*Thematic Classification and Distribution of Conscious-Access Routines*

Routine	Primary Evaluative Focus	Principal Indicators	Checkpointed	Non-Checkpointed
Evidence-first evaluator	Meaning and scope	fidelity evidential Source before claim rejection of unsupported content	checking uptake; narrowing; 82	34
Fluency-first acceptor	Flow and academic register	Rapid uptake; minimal immediate source consultation	24	78
Stance-first editor	Voice and epistemic positioning	Hedging, modality adjustment, removal of excessive certainty	41	35
Compliance-first editor	Visible task requirements	Citation placement or rule satisfaction without consistent scope verification	22	26
Time-pressure shortcut	Completion under constrained time	Accelerated uptake and reduced checking late in the session	12	18
Mixed routine	Shifting evaluative priorities	Movement between fluency, stance, and evidence checking within an episode	11	1
Total			192	192

Condition Differences

Differences between the distribution of evidence-first and fluency-prep episodes were most evident in both ecologies. The non-checkpointed condition resulted in 34 out of 192 episodes, with only 82 being evidence-first. The pattern was reversed, with 24 episodes in Condition A and 78 in Conditioning B. Fluency-first uptake was observed. This meant that in Condition A, source verification was more evident and so the condition B relied more on linguistically plausible wording than any other factor. Differences in prevalence do not correspond to specific behavioral categories, but rather to different ones. A few authors in Condition B underwent advanced verification, while some checkpoints completed in condition A lacked extensive analysis.

When comparing the two groups, there is a similar contrast: evidence-first episodes were found to be more prevalent in Condition A (82/192 vs 34/192) while fluency-First episodes are more common in Concise B (78/198 versus 24/192). The screen lines were oriented in the same way. In Condition B, there were more cases of "so-called" insertion without an immediate check of the sources, and in Condition A, writers returned to their sources with greater visibility.

In Condition A, checkpoint logs made the evaluative moment clear. Acceptances that were accepted had the lowest confidence, while adaptations had a higher confidence and adoptions with the highest confidence. Writers commonly displayed high confidence when dealing with evidential problems such as overclaiming, false causality, hallucinated details, misattribution, or overgeneralization of scope. This was true even for the most confident authors. The justification for lower confidence-based adoptions was often based on their academic vocabulary, flow, or citation coverage. One interpretation of this pattern is that checkpointing made evidential concerns into explicit points of control rather than implicit (Fleming, 2024).

In Condition B, evaluation was frequently arbitrary and did not always come to a conclusion during the revision process. A few writers claimed they held onto an AI proposal despite uncertainty due to its academic sound, while others believed a suggestion was valid because the source packet was included in the prompt. Rapid-insertion behavior in the traces is consistent with these accounts. Despite the absence of evidence-based work, some writers employed AI output to identify sources and source disagreements. Despite this, the use of evidence-first methods was less common in the more opaque ecology.

Development Across Cycles

Change across the four cycles was gradual rather than abrupt. Evidence-first episodes became more frequent over time, while fluency-first uptake became less common.

The writing artifacts show a parallel shift. Categorical claims were increasingly replaced by more bounded inferences and more carefully calibrated scope.

- P1 (Cycle 2): Meeting 1: “Social media makes everyone sad.” Meeting 2: “Studies suggest social media use is associated with decreased well-being.” Week 12: “The correlation between screen time and mood is statistically significant but small.”
- P10 (Cycle 3): Meeting 1: “AI will take all the jobs.” Meeting 2: “Automation is projected to impact specific labor categories disproportionately.” Week 12: “Technological displacement varies significantly across manual and cognitive tasks.”

Transfer to No-AI Writing

The Week 8 posttest, completed without AI under controlled conditions, offers evidence that some of the routines used during AI-assisted revision were also present when the tool was removed.

- Evidence-first transfer: P1 (A): “While Source 2 highlights economic growth, it is important to note this was limited to urban centers.”
- Evidence-first transfer: P23 (B): “The findings in 2020 show a 15% increase, but this should be viewed as a temporary spike.”

- Stance-first transfer: P2 (A): “The data could be interpreted as a shift in behavior, though further longitudinal study is required.”
- Fluency-first transfer: P5 (B): “The technological revolution has fundamentally altered the way that modern societies function.”

Comparable patterns appeared in other Week 8 excerpts. In other words, the routines identified in the AI-assisted sessions were not confined to those sessions; they were also visible in writing produced immediately afterward without AI.

By Week 12, evidence of carryover appeared in both the no-AI texts and participants' own accounts of regulation. P12, for example, described “checkpointing” sentences without Gemini by asking whether each claim could be justified from the texts, echoing the verification and justification routines observed earlier. In the milestone writing, evidence-first and stance-first transfer took the form of bounded inference and calibrated epistemic language. Fluency-first transfer looked different: an academic register could remain even when evidential limits weakened, as illustrated by P5's description of urbanization as an “inevitable result.”

Table 7 shows how the proportion of coded evidence-first instances in the no-AI milestone texts changed over time in the two conditions.

Table 7

No-AI Evidence-First Rates Across Milestones by Condition (Weeks 1, 8, and 12)

Milestone	Condition A	Condition B
Week 1 Baseline	12%	14%
Week 8 Posttest	68%	31%
Week 12 Delayed Transfer	62%	28%

Baseline rates were low and similar in the two groups. After the AI-assisted cycles, however, the proportion of evidence-first instances was higher in Condition A at Week 8 and remained higher at Week 12. This pattern accords with the greater number of evidence-first episodes observed during AI-assisted revision and with participants' reports that checkpointing prompted source rechecking and regulation through justification.

Discussion

Interpreting Checkpointed Evaluation

During integrated reading-to-write (RTW) revision, the writers were evaluated on whether to adopt, adapt, or reject their AI suggestions. The resulting routines corresponded closely to known and familiar RTW requirements: synapses across texts; attribution, calibration of evidential strength; and control of stance in academically acceptable prose (Cumming et al., 2005; Knoch & Sitajalabhorn, 2013; Plakans, 2009;

Yang & Plakans, 2012). Fluent candidate language from AI makes the consequences of these RTW behaviors more apparent, rather than solely using them.

The RTW construct utilizes evidence-first evaluation because AI output is treated as a proposal that must still be comparable to the source's meaning, scope, and attribution. For instance, in order to construct a discourse from an integrated task, the same sort of claim-evidence work is required to reject an overstated modality or faussibility or causality (Chuang & Yan, 2023; Plakans & Gebril, 2013). It tackles one of the most common weaknesses of integrated writing, which is that it can conceal misrepresentation, overgeneralization, or problematic borrowing when language is seemingly plausible (Weigle & Parker, 2012). According to the present data, evidence-first routines were generally linked to narrower claims, better-calibrated modality, and clearer evidential limits, which also resulted in stronger source-based writing (Baxter & Neumann, 2023; Plakans, 2009).

Fluency-first acceptance reversed those priorities. Writers attended first to register and flow, leaving evidential control more exposed. Academic tone is a legitimate concern in EAP revision (Hyland, 2005), yet polished AI wording can invite acceptance on the basis of plausibility when checking is optional, particularly late in a timed session. The persistence of fluency as a stylistic imprint without equally stable verification is consistent with descriptions of AI-generated academic prose that sounds appropriate while remaining only weakly tied to evidential constraints (Goulart et al., 2024). From an RTW perspective, fluency support is not itself the problem. The difficulty arises when ready access to fluent text changes what the writer monitors first.

Stance-first editing brings a different pressure into view. RTW writers must position themselves interpersonally while keeping that position compatible with the evidence and with a sense of ownership over the text. Stance and engagement resources are central to managing claim strength and caution (Hyland, 2005), but AI can complicate that work by supplying a ready-made voice. Some writers responded by reducing certainty, reclaiming wording, or reshaping evaluative language to fit the position they intended to take, which is consistent with concerns about voice in AI-mediated L2 writing (Sandstead & Kibler, 2025). Stance, on this account, is part of responsible evidence representation rather than a purely stylistic choice.

Compliance-first editing shows how easily source work can be reduced to visible mechanics, especially the presence of citations. Integrated tasks require evidence to be functionally incorporated and accurately attributed, not simply accompanied by a citation marker (Petrić, 2007; Plakans & Gebril, 2013). AI can make surface compliance quick, but citation placement may then become detached from evaluation of meaning if it is treated primarily as something required by the rubric. That problem predates AI,

although rapid AI assistance can make weakly warranted but apparently “citation-correct” sentences easier to produce.

Conscious access to resources is dependent on their availability in the present, as suggested by time-pressure shortcuts and mixed routines.[p]. Some writers favored completion or textual quality over verification when time became limited. This was an advantage. Mixed routines commonly required an initial acceptance of fluent language and a subsequent reference to sources. However, they were not always successful. This is similar to adjustment of strategy in integrated writing, wherein understanding and planning are balanced with revision under constraint (Yang & Plakans, 2012). The point of verification in AI-assisted revision can determine whether the meaning remains stable or shifts.

The contrast between the checkpointed and non-checkpointing ecologies is best understood as a result of the way evaluation was integrated into the task. The condition A requirement included a decision, reasoning for meaning fidelity (i.e, alignment of evidence), stance or voice, and confidence report. With the same AI access and time, Condition B was left with the option to optionally evaluate. Routine sets are in line with the possibility that compulsory assessment can modify writers' predetermined tasks, even when the tool and writing task remain unchanged. Despite the absence of AI, it is important to carefully interpret evidence that comes from milestone-text indicators and contemporaneous micro-reflection, not process traces of no-AI composing sessions. Therefore, careful interpretation is necessary.

This view is consistent with metacognitive accounts, in which the use of evaluation and monitoring are partly dependent on cues within the context (Fleming, 2024). The confidence can influence subsequent control, such as the examination of evidence or the modification of a belief (Maniscalco & Lau, 2012). Rejection and evidence-based correction were more likely to be accepted in these data, while adoptions justified by flow or compliance were less likely. Although not necessarily incorrect, the adoptions had a weaker basis for evidence, which was still adequate for stylistic refinement but susceptible to criticism in an RTW project where accuracy relied on sources. Additionally, Recent feedback research provides a useful comparison.' The advantages of using feedback in AI are influenced by the instructional design, as stated by Soori et al. (2026), while Sheng and Ganapathy (2026) show that feedback effects differ across various designs and outputs. It's not as much about determining whether feedback is effective as it is about managing it. The amount of feedback given to learners does not affect checkpointing, but rather the evaluative work that precedes it.

Negative cases hold equal significance. The completion of a checkpoint didn't always mean meticulously verified, and some writers in the vague condition did extensive evidence work without being required to document it.

Routinization and No-AI Transfer

Evidence-first episodes were more common over the course of the four cycles, while fluency-prep posts were less common.⁷ Repeated task demands are a suitable method of routinizing. RTW research indicates that writers can construct more efficient strategy repertoires by incorporating texts under similar conditions repeatedly (Gebriel, 2010; Yang & Plakans, 2012). Repetition and a standardized prompt template that kept AI assistance sources may have helped writers to identify the moments that require verification and regulate uptake.

The change in development wasn't solely due to fluency. Writers were more inclined to consider the boundaries of scope, modality, and attribution, which is in line with research on RTW that suggests source use and argument quality are more important than just surface correctness. A similar concept is found in L2 feedback research (Mao et al., 2024), where learning depends on the process of reviewing and altering suggestions rather than mechanically applying them. The capacity of learning is dependent on literacy, not its acceptance, although AI may offer feedback (Muñoz Muñoz et al., 2025; Yan, 2024).

One possible mechanism is suggested by the attention evidence. Scholars described a pattern of "catching pause, pausing and returning to the sources" as follows: "This is like self-caught monitoring in task-performance research (Chu et al., 2023; Zanesco et al., 2024). The teaching of this method requires the use of verification triggers. Stopping and checking can be taught to students at moments where numbers, causal language, universal quantifiers, and claims that go beyond the source are present.

The transfer was not uniform and occurred in the absence of AI writing. The milestone products showed evidence-first indicators that increased from baseline to Week 8 and maintained high levels in Week 12, particularly in the checkpointed condition. Some external prompts may have become part of their own regulation, as reported by several participants during Week 12. Metacognitive control may have shifted towards questions that ask writers if they can justify a sentence from retrieved text packets and based their claims on evidence (Fleming, 2024; Maniscalco & Lau, 2012). The integrated writing construct is well-suited to this because routines prioritize meaning fidelity and evidential alignment over surface correction. Self-reported micro-Reflection data persist in Week 12 planning actions, rather than being viewed as evidence from observed processes. As no-AI tasks were not process-traced, the pattern should be considered consistent with transfer and not as a direct indication that regulation has been internalized.

A new type of carryover also surfaced. Evidence control may not have maintained the fluency-first patterns in the writing. However, they could still maintain this style. Students may make confident or broad statements without AI if an academic style similar to AI becomes a genre model (Goulart et al., 2024). In integrated assessment, accuracy in

representing and synthesising the source material is crucial, as opposed to being solely dependent on complex prose.

Stance-first transfer makes the picture more intricate. Once they have incorporated AI as a rival source of authorial wording, writers may begin to scrutinize voice ownership as part of revision. This can be beneficial when it leads to more disciplined hedging and better stance calibration (Hyland, 2005), but it can also bring about increased mental workload when academic phrasing is incompatible with ownership. When working with sources, the pedagogical value of a stance should be focused on its ability to accurately reflect evidence, not personal preference, as it must remain within the text's accuracy.

Practical and Pedagogical Implications

The pedagogical implication is to organize AI use around routines rather than around the tool itself. A simple sequence is to require a decision on each suggestion—adopt, adapt, or reject—followed by a brief rationale and a confidence judgment. Because that sequence targets meaning fidelity, evidence alignment, and stance calibration, it can later be used as a self-check even when AI is absent (Plakans, 2009; Yang & Plakans, 2012). Citation teaching can follow the same logic by asking students to identify the exact source segment that warrants both the content and the strength of a claim (Petrić, 2007; Plakans & Gebril, 2013). Teachers can also make certainty inflation, universal quantifiers, and AI-induced shifts in stance explicit objects of attention (Hyland, 2005; Sandstead & Kibler, 2025). Used this way, AI is less a route to polished prose than a dense source of feedback through which accountable uptake can be practiced (Mao et al., 2024; Yan, 2024).

Limitations and Future Research

These findings need to be read within several limits. The sample was drawn by consent from a single EAP course and may therefore include a disproportionate number of students already comfortable with AI. The study used only Gemini and constrained its use through a shared prompt template, so the findings cannot automatically be extended to other platforms or more open-ended AI use. No-AI transfer was inferred from writing products and micro-reflections rather than from process traces, and stimulated recall is inevitably open to some reconstruction. Comparability across cases was strengthened by limiting the design to one instructional context, one AI platform, two-source RTW tasks, and a fixed revision procedure, but those same choices narrow the scope of inference. Condition differences should therefore be read as ecological affordances rather than causal proof. Future studies could examine checkpointing across proficiency levels, disciplines, AI systems, and source configurations, add finer measures of source-text alignment to process tracing, and ask whether comparable routines develop through peer or teacher feedback.

Conclusion

This research investigated the impact of AI-assisted revision on integrated reading-to-write tasks for EFL writers, by virtue of the conscious-access routines that evaluated, accepted, modified, or rejected suggestions. In various sources and 384 coded revision episodes, authors revisited the fundamental issues of source-based EAP writing, such as preserving the meaning of their source material while maintaining evidence validity, managing ownership of voice and claim status. The emergence of concerns was evident in six routines, including evidence-first verification, fluency-preached acceptance, stance-before-editing, compliance-focused citation management, time-pressure shortcuts and mixed routine. These problems demonstrate how evaluative behavior can be shaped by repeated work. Transparent environments were the most noticeable contrast compared to non-transparent ones. It was found that mandating the recording and explanation of an uptake decision involved more intensive scrutiny based on evidence-based monitoring and closer attention to scope, attribution, and unwarranted certainty. Without these checkpoints, fluency-first uptake was more common and AI wording was treated more as a linguistic template, especially when time was limited. This resulted in this phenomenon. The difference need not be treated as deterministic: both surface-level compliance and careful verification were present in each context. The frequency of each pattern under the two instructional arrangements was not consistent. The inclusion of carryover was selective in no-AI writing. The evidence-first instances increased from baseline and remained elevated at the time of delayed transfer, while the checkpointed condition showed higher rates. Based on reports of self-checkpointing and planning, it is possible that some writers had begun to use structured evaluation as part of their self regulation. Fluency-oriented patterns were sometimes observed in academic settings without equally unproven proof. There is a risk of mispedagogy in the possibility that assertive language can outperform the evidence presented to support it. The transfer of stance-focused routines highlights a shared problem, as ownership of voice is essential for accurately representing and calibrating evidence. EAP teaching places less emphasis on constructing sophisticated language than on providing structured evaluations. A more effective approach to this is through the use of AI. By using checkpointing, the same competencies as integrated assessments are maintained and used to keep evidence alignment, fidelity, calibration (standing up), and view, all in one place.

ORCID

 <https://orcid.org/0009-0004-5127-5546>

 <https://orcid.org/0009-0007-8971-5879>

 <https://orcid.org/0000-0003-1528-7326>

Publisher's Note

The claims, arguments, and counter-arguments made in this article are exclusively those of the contributing authors. Hence, they do not necessarily represent the viewpoints of

the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

Acknowledgements

The authors thank the students who participated in the study and contributed their time and experiences.

Funding

This research received no external funding.

CRedit Authorship Contribution Statement

Hesham Aldamen: Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing

Mohamad Almashour: Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing

Marwan Jarrah: Conceptualization, Methodology, Supervision, Project Administration, Writing – Review & Editing

Generative AI Use Disclosure Statement

ChatGPT was used during revision for language editing, condensation, and formatting. It was not used for data collection, coding, analysis, or analytic decision-making. The authors reviewed all AI-assisted changes and take full responsibility for the manuscript.

Ethics Declarations

The study was conducted in accordance with the Declaration of Helsinki. The Institutional Review Board of the University of Jordan approved the study (Decision No. 437/2025, 13 May 2025). All participants provided informed consent, participation was voluntary, and the data were anonymized.

Consent for Publication

All participants provided informed consent for their anonymized data to be used in this research and for the findings to be published. No identifying personal information is included in the manuscript.

Competing Interests

The authors declare no competing interests.

Data Availability

Participant traces, screen recordings, and interviews are not publicly available because of ethics and consent restrictions. Non-identifying materials may be requested from the corresponding author, subject to institutional approval. Study instruments will be deposited in IRIS where permitted.

References

- Barrot, J. S. (2024). Leveraging ChatGPT in the writing classrooms: Theoretical and practical insights. *Language Teaching Research Quarterly*, 43, 43–53. <https://doi.org/10.32038/ltrq.2024.43.03>
- Baxter, G., & Neumann, H. (2023). Source-based writing of the high- and low-proficiency adolescent writer in the high-school L2 classroom. *Journal of Second Language Writing*, 62, Article 101064. <https://doi.org/10.1016/j.jslw.2023.101064>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Chu, M. T., Marks, E., Smith, C. L., & Chadwick, P. (2023). Self-caught methodologies for measuring mind wandering with meta-awareness: A systematic review. *Consciousness and Cognition*, 108, Article 103463. <https://doi.org/10.1016/j.concog.2022.103463>
- Chuang, P.-L., & Yan, X. (2023). Connecting source use and argumentation in L2 integrated argumentative writing performance. *Journal of Second Language Writing*, 60, Article 101003. <https://doi.org/10.1016/j.jslw.2023.101003>
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10(1), 5–43. <https://doi.org/10.1016/j.asw.2005.02.001>
- Dong, L. (2024). ChatGPT in language writing education: Reflections and a research agenda for a ChatGPT feedback engagement framework. *Language Teaching Research Quarterly*, 43, 121–131. <https://doi.org/10.32038/ltrq.2024.43.07>
- Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75, 241–268. <https://doi.org/10.1146/annurev-psych-022423-032425>
- Gebriel, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15(2), 100–117. <https://doi.org/10.1016/j.asw.2010.05.002>
- Goulart, L., Matte, M. L., Mendoza, A., Alvarado, L., & Veloso, I. (2024). AI or student writing? Analyzing the situational and linguistic characteristics of undergraduate student writing and AI-generated assignments. *Journal of Second Language Writing*, 66, Article 101160. <https://doi.org/10.1016/j.jslw.2024.101160>
- Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2), 173–192. <https://doi.org/10.1177/1461445605050365>
- Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, 15(4), 261–278. <https://doi.org/10.1016/j.jslw.2006.09.006>
- Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing*, 18(4), 300–308. <https://doi.org/10.1016/j.asw.2013.09.003>
- Kostka, I., & McMartin-Miller, C. (2026). Reflections on using GenAI for multilingual writing feedback: Implications for practice. *Feedback Research in Second Language*, 4, 16–27. <https://doi.org/10.32038/frsl.2026.04.03>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Mao, Z., Lee, I., & Li, S. (2024). Written corrective feedback in second language writing: A synthesis of naturalistic classroom studies. *Language Teaching*, 57(4), 449–477. <https://doi.org/10.1017/S0261444823000393>
- Michel, M., Bazhutkina, I., Abel, N., & Strobl, C. (2025). Collaborative writing based on generative AI models: Revision and deliberation processes in German as a foreign language. *Journal of Second Language Writing*, 67, Article 101185. <https://doi.org/10.1016/j.jslw.2025.101185>
- Moorhouse, B. L., Wan, Y., Wu, C., Wu, M., & Ho, T. Y. (2025). Generative AI tools and empowerment in L2 academic writing. *System*, 133, Article 103779. <https://doi.org/10.1016/j.system.2025.103779>
- Muñoz Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System*, 134, Article 103805. <https://doi.org/10.1016/j.system.2025.103805>
- Petrić, B. (2007). Rhetorical functions of citations in high- and low-rated master's theses. *Journal of English for Academic Purposes*, 6(3), 238–253. <https://doi.org/10.1016/j.jeap.2007.09.002>
- Pitychoutis, K. M., & Spathopoulou, F. (2026). AI-mediated written corrective feedback in L2 writing: What changes, what doesn't, and what we still don't know. *Feedback Research in Second Language*, 4, 6–15. <https://doi.org/10.32038/frsl.2026.04.02>

- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561–587. <https://doi.org/10.1177/0265532209340192>
- Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, 22(3), 217–230. <https://doi.org/10.1016/j.jslw.2013.02.003>
- Sandstead, M. G., & Kibler, A. K. (2025). Voice in L2 writing in the age of AI. *Journal of Second Language Writing*, 69, Article 101212. <https://doi.org/10.1016/j.jslw.2025.101212>
- Sheng, Y., & Ganapathy, M. (2026). A systematic review of written corrective feedback and dynamic written corrective feedback research in ESL/EFL contexts. *Language Teaching Research Quarterly*, 54, 151–192. <https://doi.org/10.32038/ltrq.2026.54.07>
- Soori, A., Khojasteh, L., Kotamjani, S. S., & Mojarad, S. (2026). Optimizing EFL article accuracy with hybrid AI-teacher feedback. *Language Teaching Research Quarterly*, 54, 102–124. <https://doi.org/10.32038/ltrq.2026.54.05>
- Wang, C., Tian, Z., & Christiansen, M. S. (2025). Toward critically and humanely grounded futures for L2 writing in the GenAI era: Sparking dialogue and critical experimentation. *Journal of Second Language Writing*, 69, Article 101229. <https://doi.org/10.1016/j.jslw.2025.101229>
- Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, 21(2), 118–133. <https://doi.org/10.1016/j.jslw.2012.03.004>
- Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology*, 28(1), 1–19. <https://doi.org/10.64152/10125/73597>
- Yang, H.-C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, 46(1), 80–103. <https://doi.org/10.1002/tesq.6>
- ZanESCO, A. P., Van Dam, N. T., Denkova, E., & Jha, A. P. (2024). Measuring mind wandering with experience sampling during task performance: An item response theory investigation. *Behavior Research Methods*, 56(7), 7707–7727. <https://doi.org/10.3758/s13428-024-02446-9>