International Education Forum, 2026, Volume 4, Issue 7
http://ojs.bbwpublisher.com/index.php/IEF
Online ISSN: 2981-8605
Print ISSN: 3083-4902

Generative AI and Evidentiary Displacement in Second Language Writing: A Critical Integrative Review of Feedback, Composing Processes, and Assessment Validity

Bin Wang*

Shanghai Aurora College, Shanghai 201908, China

***Corresponding author:** Bin Wang, b.wang@aurora-college.cn

Abstract: Generative artificial intelligence (GenAI) has become embedded in second language (L2) writing not only as a source of correction, but also as a participant in planning, drafting, translating, revising, evaluating, and genre modeling. This critical integrative review argues that the central pedagogical problem raised by GenAI is evidentiary rather than merely technological. In much classroom and assessment practice, learner texts continue to function as privileged evidence for engagement, authorship, composing decisions, and writing ability. GenAI unsettles this relation because it can produce ideas, structure, language, feedback, and evaluation from minimal input while leaving no reliable trace in the submitted text. This review conceptualizes this problem as evidentiary displacement: evidence of learning is not destroyed, but redistributed across less visible processes of prompting, selecting, evaluating, revising, rejecting, and integrating machine output. Drawing on post-2022 work on GenAI-mediated L2 writing and longer-standing scholarship on feedback engagement, composing processes, and assessment validity, the review examines displacement across three sites: feedback, composing, and assessment. It argues for selective process evidence—revision rationales, decision-point annotations, feedback engagement records, and learners' accounts of their choices—while also cautioning that documentation can increase workload, reproduce inequities, invite surveillance, and become strategically performed rather than pedagogically meaningful.

Keywords: Second language writing; Generative AI; Evidentiary displacement; Feedback engagement; Composing processes; Assessment validity; Process evidence

Online publication: August 13, 2026

1. Introduction

L2 writing scholarship has long treated writing as more than the production of accurate written products. Process accounts made planning, translating, reviewing, and revising visible as objects of inquiry; later

social, post-process, and classroom-based perspectives connected writing to mediation, audience, genre, identity, institutional context, and language resources [1–5]. Feedback research likewise cautions against reading textual revision as a direct measure of learning, since the pedagogical force of feedback depends on how writers notice, interpret, resist, appropriate, and later reuse it [6–9]. In assessment, validity work has made a related point in a different register: judgments about writing ability require warranted interpretations of performance, not simply the existence of a text to be scored [10–13].

Yet institutional practice has not always absorbed these insights. Teachers still infer engagement from revised drafts, researchers often reconstruct composing from products and process traces, and assessors regularly warrant claims about writing ability from final texts produced under nominally specified task conditions. The text has never been transparent evidence. It has, however, been a practical evidentiary anchor: a stable artifact around which instruction, feedback, research, and certification could be organized. GenAI makes the fragility of that arrangement harder to ignore.

The issue is not that GenAI is wholly unprecedented. Machine translation, automated writing evaluation, dictionaries, corpora, peer review, writing centers, and teacher feedback have all mediated L2 writing in consequential ways. Research on machine translation has already questioned authorship, language learning, translanguaging, and the politics of tool use [14–15]. Automated writing evaluation research has similarly shown that machine feedback requires validation of the inferences users draw from it [16]. GenAI nevertheless intensifies and consolidates forms of assistance that were previously more distributed, partial, or easier to contextualize. A single conversational interface can generate ideas, organize arguments, translate, draft, revise, evaluate, and polish language from minimal input, often without leaving a reliable signature in the submitted text.

The early JSLW dialogue on AI-generated text therefore framed the issue through tensions rather than simple affordances. Warschauer et al. identified affordances and contradictions for ESL/EFL writers [17]; Pecorari, Sasaki, Kubota, Kang and Yi, and Praphan and Praphan extended the discussion to plagiarism, L1 resources, multimodality, equity, and classroom policy [18–22]. Recent syntheses have broadened the field further, mapping learner perceptions, feedback practices, prompting strategies, assessment applications, and emerging AI literacies [23–28]. These accounts are useful. They also show, however, why another catalog of benefits and risks would add little.

This review advances a more specific claim. One central consequence of GenAI-mediated L2 writing is evidentiary displacement: the learner's final text can no longer, by itself, sustain the same inferences about engagement, authorship, composing decisions, and writing development that L2 writing pedagogy has often drawn from it. Evidence of learning is not eliminated. It is redistributed across less visible and harder-to-interpret activities: prompting, selecting, evaluating, revising, rejecting, adapting, and integrating machine output. The review asks how this displacement operates in feedback, composing processes, and assessment validity, and what kinds of pedagogical evidence might respond to it without turning writing instruction into documentation for its own sake.

This framing deliberately avoids two tempting reductions. One reduction treats GenAI as an integrity problem to be solved mainly through detection and prohibition. Another treats GenAI as a productivity tool whose value can be measured primarily by better scores, cleaner syntax, or more positive learner perceptions. Both positions can be locally reasonable, but neither is adequate for L2 writing pedagogy. Writing instruction is concerned with how learners develop judgment over time: how they formulate meanings, respond to

feedback, revise for readers, and assume responsibility for language choices. GenAI intervenes at precisely those points. The evidentiary question, then, is not a peripheral administrative matter. It reaches into how the field defines learning, authorship, and assessment under conditions of mediated text production.

The contribution is thus intentionally restrained. The review does not propose that all L2 writing pedagogy be reorganized around AI, nor does it treat process documentation as a universal remedy. It offers a conceptual vocabulary for a problem that teachers already encounter when a polished text, a revised draft, or an impressive portfolio no longer gives adequate grounds for knowing what the learner did or learned. By naming this problem as evidentiary displacement, the article brings together debates that often remain separated: feedback uptake, process-oriented composing, AI literacy, authorship, and assessment validity. The value of the concept lies in whether it helps the field ask more precise questions of empirical studies and classroom designs.

2. Review design and positioning

The article is a critical integrative review. This design is appropriate when a developing area contains empirical studies, conceptual commentary, methodological discussion, and policy-oriented work that cannot be reduced to a single effect-size question. Integrative reviews synthesize heterogeneous literatures in order to clarify concepts, identify tensions, and develop a research agenda, rather than to estimate prevalence or aggregate outcomes [29–31]. GenAI-mediated L2 writing is such a field: the literature is recent, uneven across educational contexts, and distributed across feedback, AI literacy, writing process, translation, academic integrity, and assessment.

The review uses post-2022 peer-reviewed research on GenAI in L2 writing as its primary corpus, while drawing on foundational scholarship in L2 writing feedback, composing process research, technology-mediated writing, and assessment validity to interpret the implications of that corpus. This layered approach is necessary because GenAI is new as a public-facing technology but not new as a problem of mediation. The purpose is not to claim exhaustive coverage of every study published since the release of ChatGPT. Rather, the review reads a targeted body of recent work through an evidentiary lens and asks what kind of learning, authorship, and ability can be inferred when AI participates in text production.

The organizing logic differs from tool-centered syntheses. Instead of separating ChatGPT feedback, automated evaluation, machine translation, prompt engineering, and AI literacy as adjacent topics, the article treats them as sites where a common evidentiary problem appears. Feedback, composing, and assessment were selected because each depends on a relationship between visible textual evidence and less visible learner activity. In each site, GenAI weakens familiar inferences without making evidence irrelevant. The analytic task is therefore to identify where evidence has moved, what it can and cannot warrant, and how pedagogy might elicit it responsibly.

The review is also positioned against a recurring weakness in fast-moving AI scholarship: the tendency to let the tool define the object of inquiry. In L2 writing, a study of ChatGPT, automated feedback, or machine translation is rarely only a study of a tool. It is also a study of writing tasks, language ideologies, classroom authority, and the evidence through which learning is recognized. This article therefore treats GenAI as a pressure point on existing pedagogical assumptions rather than as an autonomous cause of change.

This methodological stance carries a limitation that should be stated rather than hidden. Because the review is conceptually driven, it does not claim to settle how often particular forms of GenAI use occur, nor does it rank interventions by effect size. The value of the synthesis depends instead on whether the proposed concept clarifies relations that remain dispersed across adjacent literatures. To reduce the risk of conceptual overreach, the analysis repeatedly returns to three questions: what inference is being made from the learner text, what GenAI mediation makes less secure, and what alternative evidence could support a more cautious interpretation. These questions provide the warrant for the selection of studies and the structure of the review.

3. Conceptualizing evidentiary displacement

Texts in L2 writing classrooms are rarely only texts. A revised paragraph may be read as evidence of feedback uptake. A series of drafts may be read as evidence of developing control over genre, stance, organization, or grammar. A polished final essay may be read as evidence of writing ability. These readings are institutionally useful, but they are interpretive acts rather than neutral observations. They depend on assumptions about task conditions, assistance, learner agency, and the relation between product and process.

GenAI changes the evidentiary distribution. A fluent text may reflect independent development, effective use of AI suggestions, full machine generation, translation from the writer's L1, peer support, or a mixture of these. A revision may reflect careful engagement with feedback, mechanical acceptance of an AI rewrite, or a student's attempt to reconcile teacher comments with machine-generated alternatives. A genre-appropriate introduction may show genre learning, a well-designed prompt, or imitation of an AI-generated model. The point is not that learning has disappeared. It is that the final text no longer specifies what sort of learning or decision-making occurred.

Evidentiary displacement has four dimensions. First, it concerns the object of inference: engagement, authorship, rhetorical judgment, linguistic development, or writing ability. Second, it concerns the source of evidence: final text, draft history, interaction log, reflection, conference, classroom observation, or controlled performance. Third, it concerns the site of displacement: prompting, selecting, evaluating, rejecting, revising, and integrating AI output. Fourth, it concerns interpretive authority: learners, teachers, assessors, researchers, and institutions may treat the same trace differently. **Figure 1** summarizes this movement from product-centered evidence to distributed process evidence.

The term displacement is chosen carefully. It does not mean that evidence becomes unavailable, nor that teachers can no longer make any judgments about students' writing. It means that the location and interpretability of evidence change. A product may still reveal audience awareness or textual coherence, but it may not reveal whether those features were planned, adopted, prompted, or understood. A prompt log may reveal interaction, but not necessarily reflection. A student explanation may reveal reasoning, but it may also be retrospective rationalization. Evidentiary displacement therefore names a problem of warrant rather than a simple movement from bad evidence to good evidence. The practical response has to be triangulation, not replacement.

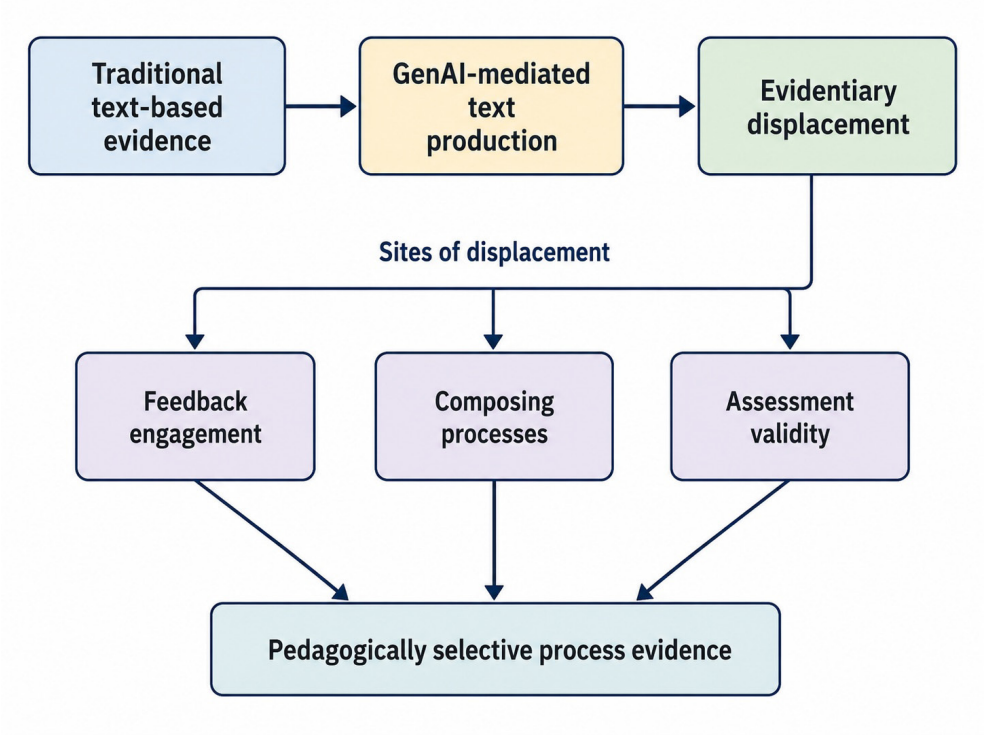

Figure 1. Evidentiary displacement in generative-AI-mediated L2 writing. Final products and revisions remain useful evidence, but GenAI mediation relocates some evidence of learning into process traces and learner explanations.

The concept does not render product evidence obsolete. Final texts still matter: they show rhetorical effects, linguistic choices, genre positioning, and audience orientation. Under AI-mediated composing, however, final products have to be interpreted together with the conditions of their production and with evidence of the learner's role in decision-making. **Table 1** applies this logic to feedback, composing, and assessment validity.

This qualification matters for classroom feasibility. A theory that makes final texts irrelevant would be unusable and empirically implausible. Teachers do not have unlimited access to students' mental activity, and most writing assessment will continue to involve products. The more modest claim is that product evidence has become more conditional. Its evidentiary value depends on task design, declared and permitted assistance, opportunities for explanation, and the relationship between the text and other artifacts. In this sense, GenAI does not create the need for process-oriented pedagogy; it makes the cost of ignoring process evidence more visible.

Table 1. Evidentiary displacement across feedback, composing processes, and assessment validity

Site	Conventional evidence	GenAI displacement	More defensible evidence
Feedback	Revision is often read as uptake.	Textual change may result from machine rewriting rather than learner engagement.	Learner interpretations, accepted/ rejected suggestions, revision rationales.
Composing	Drafts and final texts are treated as traces of writer decisions.	Planning, wording, structure, and revision may move into human-AI interaction.	Prompt comparisons, decision-point annotations, draft commentary, oral explanations.
Assessment	Final text quality supports claims about writing ability.	Production conditions become opaque, weakening text-to-ability inference.	Triangulated product-process evidence aligned with the assessment purpose.

Note: The table identifies the inference that becomes less secure when GenAI participates in text production.

4. Feedback: From textual uptake to evidenced engagement

Feedback has always depended on inference. Teachers comment on texts, students revise, and the revised draft is often treated as evidence that feedback was understood. L2 writing research has complicated that assumption. Hyland and Hyland describe feedback as interpersonal and contextual [8]; Ferris links corrective feedback to development rather than immediate correction alone [6]; Han and Hyland show that engagement has cognitive, behavioral, and affective dimensions [7]. Feedback literacy scholarship reaches a similar conclusion from another direction: learners must appreciate feedback, judge its relevance, manage affect, and take action [32–34]. A changed text, then, is only partial evidence of feedback engagement.

GenAI makes this partiality more visible. AI feedback can be fast, extensive, personalized, and available outside class. Students may prefer it for proofreading and editing because it is immediate and less socially exposing than peer response [35]. Comparative work also shows that GPT-assisted feedback can be useful but uneven, especially when it offers fluent reformulations without pedagogically precise explanation [36]. Recent reviews of GenAI feedback studies identify recurring concerns about prompt transparency, relevance, reliability, learner dependence, and the thinness of product-only measures [24, 26].

The evidentiary issue is sharper than the question of whether AI feedback improves writing quality. When a student submits a revised draft after consulting GenAI, textual improvement may not indicate feedback engagement. It may show that the student accepted a machine rewrite. Conversely, limited textual change may conceal careful evaluation, selective rejection, and emerging rhetorical judgment. Yeung's study is instructive because it combined drafts, screen recordings, stimulated recall, and interviews to examine behavioral, cognitive, and affective engagement with GenAI-supported automated writing evaluation feedback [37]. Such designs show that evidence of engagement often lies in interpretation and justification, not only in the changed surface of the text.

The same difficulty affects the teacher's work. If AI feedback supplies extensive correction before a teacher sees the draft, the teacher may be responding to an already mediated text. Some errors may have disappeared, but the underlying developmental need may remain. Other features may have been improved in ways the learner cannot reproduce or explain. This changes the diagnostic value of drafts. A teacher may need to ask not only what is wrong with the text, but what the student did with available feedback and what kind of support produced the observed quality. That question is uncomfortable because it complicates efficient grading routines. It is nevertheless central if feedback is to remain developmental rather than merely editorial.

Pedagogy should therefore move from textual uptake to evidenced engagement. This does not require collecting every prompt or policing every revision. A more defensible approach is to identify a small number of consequential feedback decisions. Students might annotate AI suggestions they accepted and rejected, compare teacher and AI feedback on the same passage, or write a short rationale explaining how a revision addressed audience, genre, evidence, or stance. These artifacts do not prove learning in a simple sense, and they can be performed strategically. Still, they provide better grounds for pedagogical inference than the revised text alone. They also teach students that feedback is not information to be consumed, but advice to be interpreted.

Such practices also help distinguish different kinds of improvement. A grammar correction may indicate local editing; a restructuring of evidence may indicate rhetorical reconsideration; a rejected AI suggestion may indicate that the student has detected a mismatch between generic advice and task-specific purpose.

These distinctions are often flattened when feedback studies focus only on pre- and post-revision quality. They matter because L2 writing development is not exhausted by textual improvement in a single draft. It also involves the growth of criteria, judgment, and metalinguistic awareness. GenAI can support these capacities, but only when students are required to make their reasoning available for response.

5. Composing processes: From product traces to accountable authorship

Process-oriented L2 writing research already provides a language for the second site of displacement. Writing involves planning, formulation, monitoring, reviewing, and revision; it is also shaped by discourse communities, collaboration, material tools, and institutional purpose [1–5, 38–39]. In conventional process research, draft histories, keystrokes, screen recordings, stimulated recall, and interviews are used because the final product cannot display the route by which it was made. GenAI widens the gap between route and product.

Recent process-rich studies illustrate the point. Hwang et al. found that learners used ChatGPT differently across the composing process, with early uses for idea generation and later uses for checking, polishing, examples, and continued development [40]. Yang and Lin show how EFL students used GenAI in translanguaging practices, drawing on L1 resources for planning and problem solving before transforming ideas into English [41]. Michel et al. document revision and deliberation in collaborative writing with GenAI model texts, while de Oliveira and dos Santos show how AI-generated mentor texts can support genre-based pedagogy when they become objects of analysis rather than templates for imitation [42–43]. Ou et al. and Wang and Wang extend the discussion to critical GAI literacy, emphasizing evaluation, ownership, and ethical positioning [44–45]. These studies suggest that AI-mediated composing is not a single activity. It is a shifting set of negotiations among linguistic resources, machine output, task demands, and the writer's emerging judgment.

Authorship consequently becomes an empirical and pedagogical question. It is not enough to ask whether a human or a machine wrote a sentence. Textual authorship concerns the origin of words and clauses; rhetorical authorship concerns who shaped purpose, audience, organization, stance, and emphasis; accountable authorship concerns whether the learner can explain, revise, and take responsibility for those choices. These dimensions may diverge. A learner may rely on AI wording while retaining rhetorical control. Another may type every sentence but outsource the conceptual structure. A third may use AI heavily and still learn from comparing alternatives. The submitted text cannot adjudicate these cases alone.

This account complicates familiar classroom binaries. A handwritten or AI-free draft is not automatically evidence of mature authorship, just as an AI-assisted draft is not automatically evidence of outsourcing. The relevant pedagogical issue is the distribution of decision-making. Which problems did the learner identify? Which resources were consulted? Which alternatives were considered? What was rejected, and why? These questions are more demanding than asking whether AI was used, but they are closer to the intellectual work of writing. They also avoid reducing multilingual writers' tool use to suspicion, a risk that becomes especially acute when polished English is treated as improbable for some students but normal for others.

This is also where equity enters the argument. Sasaki warns against treating EFL writers' L1 resources as deficits, and Jiang et al. similarly caution that machine translation should not be treated as politically neutral [14, 19]. Darvin pushes the analysis toward critical digital literacies, where platform access, interface design, language hierarchies, and institutional policy shape what writers can do [46]. Sandstead and Kibler add

that voice itself must be reconsidered when AI-generated language enters multilingual writing [47]. Broader AI literacy frameworks can support this discussion, but L2 writing requires domain-specific attention to rhetorical choice, language learning, and textual responsibility [48–49].

The pedagogical implication is not that students should produce an audit trail of every interaction. More useful is decision-point evidence: moments where writers identify a consequential choice, describe available alternatives, explain the option selected, and reflect on its rhetorical or linguistic effects. Such evidence might include a short prompt comparison, an AI-output critique, an annotated revision, a draft history with commentary, or a brief oral explanation. The aim is not to reconstruct a total composing history. It is to make visible enough of the writer's judgment to support learning-oriented inference.

For research, this means that methods must be matched to the kind of authorship being investigated. Screen recordings and interaction logs can document textual movement, but they do not by themselves reveal ownership or understanding. Stimulated recall and interviews can elicit reasoning, but they are retrospective and shaped by what learners think researchers or teachers want to hear. Text analysis can identify differences between student and AI output, but it cannot settle the pedagogical meaning of those differences. A stronger design would combine traces, accounts, and products, while treating all three as partial. This methodological modesty is especially needed in a field where technologies and user practices change faster than research cycles.

6. Assessment validity: From integrity policing to validity argument reconstruction

The assessment implications of GenAI are often framed as academic integrity. That concern is real, but it is too narrow. In validity theory, scores and judgments are warranted only when evidence supports the intended interpretation and use [10–12]. A writing sample can support a claim about ability only if the construct, task conditions, forms of assistance, scoring criteria, and consequences are sufficiently specified. This is why automated writing evaluation, even before GenAI, required argument-based validation rather than simple claims of efficiency [13, 16].

GenAI weakens the inference from text quality to writing ability when the conditions of text production are opaque. Studies comparing AI-generated and student writing show that textual features alone may not reliably settle questions of provenance [50]. Work on AI statements and institutional policies also shows that generic declarations do not necessarily address the realities of linguistically diverse writers [51]. Research on ChatGPT as an assessment companion suggests potential for accuracy assessment and rating support, but also raises questions about reliability, construct representation, and appropriate use [52–53]. The problem, therefore, is not only whether students used AI. It is what the submitted performance can validly be taken to mean.

Rubrics are affected as well. Criteria such as grammar, lexical range, organization, and coherence were never pure measures of independent competence, but under AI-mediated conditions they become still more ambiguous. A student may receive high marks for language accuracy because the text has been polished by AI, while another may demonstrate stronger developing control through a less polished but more independently reasoned draft. This does not imply that accuracy and organization should be abandoned as criteria. It does suggest that criteria need to be linked to task conditions. If AI editing is allowed, assessment must decide whether it values the resulting text, the learner's ability to evaluate the edit, or both.

Detection-based responses are especially weak from a validity perspective. Even where detection tools appear administratively attractive, they do not explain what the writer understood, how the text was

produced, or whether assistance was permitted by the task. They also risk shifting attention from construct representation to suspicion management. A detector may claim that a passage is likely AI-generated, but the pedagogical question is more specific: did the learner make the rhetorical and linguistic decisions the assessment is intended to capture? If that question cannot be answered from the text, the assessment design needs better evidence, not simply a stronger detection instrument.

A validity-oriented response begins with assessment purpose. If the goal is to assess independent timed writing, AI is construct-irrelevant assistance and should be controlled. If the goal is to assess writing development in a course where AI use is taught, AI may be part of the instructional context and should be documented. If the goal is to assess academic or professional writing under contemporary conditions, the construct may include the ability to prompt, evaluate, revise, verify, and take responsibility for AI-mediated text. **Table 2** summarizes these distinctions. The categories are not exhaustive, but they make explicit a point often lost in policy discussions: the same AI use can be construct-irrelevant in one assessment and construct-relevant in another.

Table 2. Assessment purposes, GenAI status, and evidence needed for validity arguments

Assessment purpose	Construct claim	GenAI status	Evidence needed
Independent L2 writing	Learner can plan, draft, and revise without AI.	Controlled or prohibited.	Supervised writing; comparison with coursework.
AI-supported development	Learner can use AI to revise and learn from feedback.	Permitted under defined conditions.	Drafts, feedback records, revision rationales.
AI-mediated academic writing	Learner can compose responsibly with AI in authentic tasks.	Integrated into the construct.	Final text, AI-use disclosure, decision-point notes, explanation of choices.
Diagnostic instruction	Teacher identifies needs and next steps.	Varies by diagnostic focus.	Short in-class writing, reflective comparison, targeted conference.

Note: The table treats GenAI use as construct-relevant or construct-irrelevant depending on the intended score interpretation.

This perspective does not make process evidence a simple solution. AI-use statements, prompt logs, reflection notes, and oral explanations can all be incomplete, fabricated, or performed for the teacher. They may also increase workload and anxiety. Still, when aligned with a clear assessment purpose, they can provide better support for interpretation than a final text alone. The task is to triangulate rather than to detect. A final essay, a short in-class diagnostic sample, a revision rationale, and an oral explanation may together warrant a more cautious claim about development than any single artifact. Wang, Tian, and Christiansen call for critically and humanely grounded experimentation in GenAI-era L2 writing; validity reconstruction is one way to make that call operational [54].

A validity argument also needs to consider consequences. If institutions respond by moving all assessment to closed-book, AI-free timed writing, they may protect one inference while narrowing the construct of writing to a decontextualized performance. If they respond by allowing unrestricted AI use without process evidence, they may produce attractive portfolios with little basis for judging development. Neither response is adequate across purposes. The more defensible path is plural: some tasks should diagnose independent control; others should assess supported revision, genre learning, and responsible AI-mediated composing. The validity question is not whether AI is present, but whether its presence is specified, taught, and aligned with the claim being made.

7. Pedagogical reconstruction around selective process evidence

Reading feedback, composing, and assessment through evidentiary displacement reframes the problem for pedagogy. The question is not whether GenAI should be banned, permitted, or celebrated in the abstract. Institutions that teach, respond to, and certify writing still need warranted evidence of learner development. The more difficult question is where such evidence now resides, how it can be elicited without excessive monitoring, and how it can be interpreted with attention to context.

Selective process evidence offers a practical but limited response. It is selective because not every interaction should be recorded. It is process evidence because it focuses on decisions, not merely products. It is pedagogical because its purpose is to develop judgment, not only to protect assessment. In feedback, students can be asked to explain how they interpreted a small number of comments from teacher, peer, or AI sources. In composing, they can annotate a decision point where AI changed their plan, language, or stance. In assessment, they can provide an AI-use disclosure that specifies the functions used—idea generation, translation, feedback, genre modeling, editing—and then explain responsibility for the final text.

This approach can be built into ordinary assignments without making the course revolve around AI. A process note might be limited to 150 words. A revision rationale might ask for two accepted suggestions and one rejected suggestion. A conference might last five minutes and focus on one paragraph rather than the whole paper. A portfolio might include one AI-assisted artifact and one short supervised diagnostic sample. Small designs of this kind are less spectacular than comprehensive digital tracking, but they are more likely to survive the conditions of real teaching. They also keep the focus on writing decisions rather than on technological display.

These practices need to be designed with restraint. A classroom that requires exhaustive logs may reward students who are adept at curating process records rather than those who are developing as writers. It may also place heavier burdens on teachers and disadvantage learners with limited access, weaker metalinguistic vocabulary, or less familiarity with institutional expectations. Documentation is not neutral. It can become surveillance, and it can reproduce inequity. For this reason, process evidence should be short, task-specific, and directly tied to the construct being taught or assessed.

Teacher judgment remains central. Process evidence does not interpret itself, and AI tools should not be positioned as neutral arbiters of student development. Teachers need criteria for reading process notes, for distinguishing explanation from compliance, and for responding to students who use AI in different ways. They also need institutional support. Without time, policy clarity, and professional discussion, process evidence can become another layer of uncompensated labor. The conceptual argument of this review therefore has a practical edge: evidentiary displacement cannot be addressed by asking individual teachers to monitor more. It requires task designs and policies that make interpretation feasible.

Policy language should therefore be tied to pedagogy rather than written only as compliance rules. A statement that AI is allowed, banned, or must be disclosed says little unless students know what kinds of help are legitimate for a particular task and why. For one assignment, translation may be prohibited because the purpose is to diagnose independent sentence-level control; for another, translation may be allowed because the purpose is to develop argument and genre awareness. Policies that do not name the learning purpose behind restrictions are unlikely to guide students' decisions. They may also encourage strategic concealment, which further displaces evidence from teacher view.

The conceptual shift is from product policing to accountable authorship. Students should not learn merely to hide or disclose AI use. They should learn to question AI output, compare alternatives, revise

for audience and purpose, notice language patterns, and justify decisions. Prompting is not inherently educational; it becomes educational when it supports inquiry into meaning, genre, evidence, and style. AI literacy is similarly insufficient if detached from writing development. What matters for L2 writing is whether learners can use AI without outsourcing the rhetorical and linguistic judgment that writing instruction is meant to cultivate.

8. Research agenda and conclusion

Evidentiary displacement suggests several lines of inquiry. The field needs more process-rich studies that combine draft histories, screen recordings, interaction logs, stimulated recall, interviews, and classroom observation, while recognizing that these traces remain partial and socially shaped. Product-score studies and perception surveys are still useful, especially in early mapping work, but they cannot explain how learners negotiate AI output during composing. Research on feedback should examine engagement, not only feedback quality or revision outcomes. Research on authorship should distinguish textual, rhetorical, and accountable authorship rather than treating AI use as a binary problem. Assessment research should build and test validity arguments for different purposes: independent writing, AI-supported development, AI-mediated academic writing, and diagnostic classroom assessment.

Longitudinal work is particularly needed. Many current studies capture initial encounters with tools whose novelty may affect both enthusiasm and anxiety. The field knows less about whether repeated GenAI use changes writers' tolerance for difficulty, their willingness to revise without machine assistance, their sense of voice, or their ability to transfer feedback into later writing. Research should also examine teachers' interpretive practices. If process evidence is to be used, how do teachers read it, what do they trust, and how do their judgments differ across contexts? These questions are not merely methodological. They concern the social life of evidence in classrooms.

Contextual breadth is also needed. Much current work is concentrated in higher education and in contexts where students and teachers have relatively stable access to commercial tools. Evidentiary displacement will look different in large compulsory writing classes, secondary schools, vocational programs, exam-oriented systems, multilingual classrooms, and lower-resource institutions. In such settings, the demand for process evidence may conflict with class size, grading load, institutional policy, or uneven access to paid AI services. A theory of AI-mediated L2 writing that does not account for these conditions risks mistaking the practices of well-resourced classrooms for general pedagogy. Future studies should therefore examine not only what learners do with AI, but also what forms of evidence teachers can realistically request and interpret.

The agenda also requires attention to limits. Process evidence can be fabricated. It may become administrative work that displaces writing itself. It may be easier to implement in well-resourced universities than in large classes, vocational settings, or under-supported programs. It may also privilege students who already know how to narrate their learning in institutionally valued terms. These concerns do not invalidate process evidence, but they prevent it from becoming a new orthodoxy. Pedagogical design should ask how much evidence is enough, whose evidence counts, and what forms of documentation are proportionate to the instructional purpose.

GenAI does not make final texts meaningless. It makes their evidentiary status more conditional. In feedback, textual revision cannot be read straightforwardly as uptake. In composing, the product is an

incomplete trace of decision-making. In assessment, text quality cannot warrant claims about writing ability unless the conditions of production and the intended construct are specified. Evidence of learning has not disappeared; it has moved. The task for L2 writing pedagogy is to design conditions under which that evidence can be elicited, interpreted, and used with care. The goal is not to police final texts into purity, nor to perfect them through machine assistance. It is to cultivate writers who can make, explain, evaluate, and take responsibility for rhetorical and linguistic decisions when composing with GenAI.

Disclosure statement

The author declares no conflict of interest.

References

[1] Flower L, Hayes JR, 1981, A Cognitive Process Theory of Writing. College Composition and Communication, 32(4): 365–387. https://doi.org/10.2307/356600

[2] Hayes JR, 2012, Modeling and Remodeling Writing. Written Communication, 29(3): 369–388. https://doi.org/10.1177/0741088312451260

[3] Manchón RM, 2011, Learning-to-Write and Writing-to-Learn in an Additional Language. John Benjamins, Amsterdam.

[4] Matsuda PK, 2003, Process and Post-Process: A Discursive History. Journal of Second Language Writing, 12(1): 65–83. https://doi.org/10.1016/S1060-3743(02)00127-3

[5] Storch N, 2013, Collaborative Writing in L2 Classrooms. Multilingual Matters, Bristol.

[6] Ferris DR, 2010, Second Language Writing Research and Written Corrective Feedback in SLA. Studies in Second Language Acquisition, 32(2): 181–201. https://doi.org/10.1017/S0272263109990490

[7] Han Y, Hyland F, 2015, Exploring Learner Engagement with Written Corrective Feedback in a Chinese Tertiary EFL Classroom. Journal of Second Language Writing, 2015(30): 31–44. https://doi.org/10.1016/j.jslw.2015.08.002

[8] Hyland K, Hyland F, 2006, Feedback on Second Language Students' Writing. Language Teaching, 39(2): 83–101. https://doi.org/10.1017/S0261444806003399

[9] Zhang ZV, Hyland K, 2018, Student Engagement with Teacher and Automated Feedback on L2 Writing. Assessing Writing, 2018(36): 90–102. https://doi.org/10.1016/j.asw.2018.02.004

[10] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, Standards for Educational and Psychological Testing. American Educational Research Association.

[11] Kane MT, 2013, Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1): 1–73. https://doi.org/10.1111/jedm.12000

[12] Messick S, 1989, Validity. in Linn RL (ed.), Educational Measurement, 3rd ed., American Council on Education/Macmillan, 13–103.

[13] Weigle SC, 2002, Assessing Writing. Cambridge University Press, Cambridge.

[14] Jiang L, Yu R, Zhao Y, 2024, Theoretical Perspectives and Factors Influencing Machine Translation Use in L2 Writing: A Scoping Review. Journal of Second Language Writing, 64: Article 101099. https://doi.org/10.1016/j.jslw.2024.101099

[15] Lee SM, Kang N, 2024, Effects of Machine Translation on L2 Writing Proficiency: The Complexity,

Accuracy, Lexical Diversity, and Fluency. Language Learning & Technology, 28(1): 1–19. https://doi.org/10.64152/10125/73585

[16] 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

[17] Warschauer M, Tseng W, Yim S, et al., 2023, The Affordances and Contradictions of AI-Generated Text for Writers of English as a Second or Foreign Language. Journal of Second Language Writing, 2023(62): Article 101071. https://doi.org/10.1016/j.jslw.2023.101071

[18] Pecorari D, 2023, Generative AI: Same Same but Different? Journal of Second Language Writing, 2023(62): Article 101067. https://doi.org/10.1016/j.jslw.2023.101067

[19] Sasaki M, 2023, AI Tools as Affordances and Contradictions for EFL Writers: Emic Perspectives and L1 Use as a Resource. Journal of Second Language Writing, 2023(62): Article 101068. https://doi.org/10.1016/j.jslw.2023.101068

[20] Kubota R, 2023, Another Contradiction in AI-Assisted Second Language Writing. Journal of Second Language Writing, 2023(62): Article 101069. https://doi.org/10.1016/j.jslw.2023.101069

[21] Kang J, Yi Y, 2023, Beyond ChatGPT: Multimodal Generative AI for L2 Writers. Journal of Second Language Writing, 2023(62): Article 101070. https://doi.org/10.1016/j.jslw.2023.101070

[22] Praphan PW, Praphan K, 2023, AI Technologies in the ESL/EFL Writing Classroom: The Villain or the Champion? Journal of Second Language Writing, 2023(62): Article 101072. https://doi.org/10.1016/j.jslw.2023.101072

[23] Barrot JS, 2023, Using ChatGPT for Second Language Writing: Pitfalls and Potentials. Assessing Writing, 2023(57): Article 100745. https://doi.org/10.1016/j.asw.2023.100745

[24] Crosthwaite P, Sun S, 2026, Generative AI and L2 Written Feedback Studies: A Scoping Review. RELC Journal, 57(1): 207–219. https://doi.org/10.1177/00336882251386530

[25] Feng H, Li K, Zhang LJ, 2025, What Does AI Bring to Second Language Writing? A Systematic Review (2014-2024). Language Learning & Technology, 29(1): 1–27. https://doi.org/10.64152/10125/73629

[26] Li M, 2024, Leveraging ChatGPT for Second Language Writing Feedback and Assessment. International Journal of Computer-Assisted Language Learning and Teaching, 14(1): 1–11. https://doi.org/10.4018/IJCALLT.360382

[27] 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

[28] Yang C, Li R, Yang L, 2025, Revisiting Trends in GenAI-Assisted Second Language Writing: Retrospect and Prospect. Journal of Educational Computing Research, 63(7–8): 1819–1863. https://doi.org/10.1177/07356331251367309

[29] Snyder H, 2019, Literature Review as a Research Methodology: An Overview and Guidelines. Journal of Business Research, 2019(104): 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[30] Torraco RJ, 2005, Writing Integrative Literature Reviews: Guidelines and Examples. Human Resource Development Review, 4(3): 356–367. https://doi.org/10.1177/1534484305278283

[31] Whittemore R, Knafl K, 2005, The Integrative Review: Updated Methodology. Journal of Advanced Nursing, 52(5): 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

[32] Boud D, Molloy E, 2013, Rethinking Models of Feedback for Learning: The Challenge of Design. Assessment & Evaluation in Higher Education, 38(6): 698–712. https://doi.org/10.1080/02602938.2012.691462

[33] Carless D, Boud D, 2018, The Development of Student Feedback Literacy: Enabling Uptake of Feedback.

Assessment & Evaluation in Higher Education, 43(8): 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

[34] Yu S, Zhang ED, Liu C, 2022, Assessing L2 Student Writing Feedback Literacy: A Scale Development and Validation Study. Assessing Writing, 2022(53): Article 100643. https://doi.org/10.1016/j.asw.2022.100643

[35] Allen TJ, Mizumoto A, 2024, ChatGPT over My Friends: Japanese English-as-a-Foreign-Language Learners' Preferences for Editing and Proofreading Strategies. RELC Journal. https://doi.org/10.1177/00336882241262533

[36] Lin S, Crosthwaite P, 2024, The Grass Is Not Always Greener: Teacher vs. GPT-Assisted Written Corrective Feedback. System, 2024(127): Article 103529. https://doi.org/10.1016/j.system.2024.103529

[37] Yeung S, 2025, University Students' Engagement with Generative AI-Supported Automated Writing Evaluation (AWE) Feedback. Journal of Second Language Writing, 2025(68): Article 101203. https://doi.org/10.1016/j.jslw.2025.101203

[38] Storch N, 2005, Collaborative Writing: Product, Process, and Students' Reflections. Journal of Second Language Writing, 14(3): 153–173. https://doi.org/10.1016/j.jslw.2005.05.002

[39] Leijten M, Van Waes L, 2013, Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. Written Communication, 30(3): 358–392. https://doi.org/10.1177/0741088313491692

[40] Hwang H, Chang X, Sun J, 2025, Generative AI Is Useful for Second Language Writing, but When, Why, and for How Long Do Learners Use It? Journal of Second Language Writing, 2025(69): Article 101230. https://doi.org/10.1016/j.jslw.2025.101230

[41] Yang W, Lin C, 2025, Translanguaging with Generative AI in EFL Writing: Students' Practices and Perceptions. Journal of Second Language Writing, 2025(67): Article 101181. https://doi.org/10.1016/j.jslw.2025.101181

[42] Michel M, Bazhutkina I, Abel N, et al., 2025, Collaborative Writing Based on Generative AI Models: Revision and Deliberation Processes in German as a Foreign Language. Journal of Second Language Writing, 2025(67): Article 101185. https://doi.org/10.1016/j.jslw.2025.101185

[43] de Oliveira LC, dos Santos AE, 2025, Using AI-Text Generated Mentor Texts for Genre-Based Pedagogy in Second Language Writing. Journal of Second Language Writing, 2025(67): Article 101184. https://doi.org/10.1016/j.jslw.2025.101184

[44] Ou AW, Khuder B, Franzetti S, et al., 2024, Conceptualising and Cultivating Critical GAI Literacy in Doctoral Academic Writing. Journal of Second Language Writing, 2024(66): Article 101156. https://doi.org/10.1016/j.jslw.2024.101156

[45] Wang C, Wang Z, 2025, Investigating L2 Writers' Critical AI Literacy in AI-Assisted Writing: An APSE Model. Journal of Second Language Writing, 2025(67): Article 101187. https://doi.org/10.1016/j.jslw.2025.101187

[46] Darvin R, 2025, The Need for Critical Digital Literacies in Generative AI-Mediated L2 Writing. Journal of Second Language Writing, 2025(67): Article 101186. https://doi.org/10.1016/j.jslw.2025.101186

[47] Sandstead M, Kibler A, 2025, Voice in L2 Writing in the Age of AI. Journal of Second Language Writing, 2025(69): Article 101212. https://doi.org/10.1016/j.jslw.2025.101212

[48] Ng DTK, Leung JKL, Chu SKW, et al., 2021, Conceptualizing AI Literacy: An Exploratory Review. Computers and Education: Artificial Intelligence, 2021(2): Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

[49] Xu L, Zhang L, Ou L, et al., 2025, The Development and Validation of a Scale on Student AI Literacy in L2 Writing: A Domain-Specific Perspective. Journal of Second Language Writing, 2025(69): Article 101227. https://doi.org/10.1016/j.jslw.2025.101227

[50] Goulart L, Matte ML, Mendoza A, et al., 2024, AI or Student Writing? Analyzing the Situational and Linguistic

Characteristics of Undergraduate Student Writing and AI-Generated Assignments. Journal of Second Language Writing, 2024(66): Article 101160. https://doi.org/10.1016/j.jslw.2024.101160

[51] Dang A, Wang H, 2024, Ethical Use of Generative AI for Writing Practices: Addressing Linguistically Diverse Students in U.S. Universities' AI Statements. Journal of Second Language Writing, 2024(66): Article 101157. https://doi.org/10.1016/j.jslw.2024.101157

[52] Mizumoto A, Shintani N, Sasaki M, et al., 2024, Testing the Viability of ChatGPT as a Companion in L2 Writing Accuracy Assessment. Research Methods in Applied Linguistics, 3(2): Article 100116. https://doi.org/10.1016/j.rmal.2024.100116

[53] Pfau A, Polio C, Xu Y, 2023, Exploring the Potential of ChatGPT in Assessing L2 Writing Accuracy for Research Purposes. Research Methods in Applied Linguistics, 2(3): Article 100083. https://doi.org/10.1016/j.rmal.2023.100083

[54] Wang C, Tian Z, Christiansen MS, 2025, Toward Critically and Humanely Grounded Futures for L2 Writing in the GenAI Era: Sparking Dialogue and Critical Experimentation. Journal of Second Language Writing, 2025(69): Article 101229. https://doi.org/10.1016/j.jslw.2025.101229