Europe
Multidisciplinary Journal of Academic Research
Vol. 22 No. 1, 2025 ISSN: 0014-2751

DECONSTRUCTING THE 'BLACK BOX' OF AUTOMATED WRITING EVALUATION: A CONTROLLED EVALUATION OF EXPLANATORY VS. CORRECTIVE AI FEEDBACK ON SECOND-LANGUAGE REVISION

Aurèle Beaumont
3iL École d'ingénieurs, Limoges
France

Abstract

Automated Writing Evaluation (AWE) tools have proliferated rapidly across secondary and tertiary educational settings, yet the instructional science underpinning their feedback architectures remains insufficiently theorised. This article presents a mixed-methods controlled trial examining how two distinct feedback types—corrective feedback (CF), which directly supplies the target form, and explanatory feedback (EF), which articulates the underlying linguistic rule—differentially influence text quality gains, metacognitive processing, and the emergence of writing self-regulation among secondary-school second-language (L2) English learners. Drawing on sociocognitive theories of feedback processing, Cognitive Load Theory, and the Self-Regulation of Learning framework, we argue that feedback architecture is not pedagogically neutral: it encodes assumptions about learner agency, cognitive engagement, and long-term transfer. Forty-eight students were randomly assigned to CF and EF conditions over an eight-week writing programme. Pre- and post-test analytic scores revealed statistically significant advantages for the EF group on global coherence and lexical sophistication measures, while CF participants showed faster short-term accuracy gains. Qualitative analysis of student revision diaries disclosed richer metacognitive commentary in EF participants, indexed by a higher frequency of self-monitoring and self-evaluation language. These findings carry direct implications for AWE tool design, writing pedagogy, and the theoretical modelling of AI-mediated feedback in instructional science.

Keywords: automated writing evaluation; corrective feedback; explanatory feedback; second-language writing; metacognition; self-regulated learning; instructional science

1. INTRODUCTION

The last decade has witnessed an exponential expansion in the deployment of Automated Writing Evaluation (AWE) systems within formal language education. Platforms such as Turnitin's Revision Assistant, Grammarly, PaperRater, and institutional writing tools embedded in learning management systems now routinely generate real-time feedback on student texts, promising instructors unprecedented scalability and learners immediate formative information (Warschauer & Grimes, 2008; Wilson & Czik, 2016). Yet despite this proliferation, a paradox persists: while the outputs of AWE systems are increasingly visible, the instructional logic governing their feedback architectures—what Kellogg (2008) might

term the cognitive demands placed on the learner—remains largely opaque, constituting what this article terms the 'black box' of AWE-mediated instruction.

The paradox is not merely technological but deeply theoretical. Instructional science has long distinguished between feedback that informs, feedback that corrects, and feedback that elaborates (Hattie & Timperley, 2007; Narciss, 2008). The educational psychology literature consistently demonstrates that feedback efficacy is neither uniform nor context-independent: its effects are mediated by learner cognitive capacity, the nature of the task, and crucially, the epistemic form of the feedback message itself (Kulhavy & Stock, 1989; Shute, 2008). In second-language (L2) acquisition research, analogous distinctions have been drawn between recasts, metalinguistic feedback, and direct error correction—each producing different patterns of noticing, uptake, and long-term retention (Lyster & Ranta, 1997; Ellis et al., 2006; Nassaji, 2016).

Notwithstanding these parallel literatures, a notable gap obtains at their intersection: we currently lack robust experimental evidence capable of demonstrating how the specific architecture of AWE feedback—particularly the distinction between corrective feedback (CF) and explanatory feedback (EF)—affects not only surface text quality outcomes but the underlying learning processes, including metacognitive development and the cultivation of writing self-regulation. This gap is consequential. If AWE tools primarily supply corrected forms without accompanying explanatory rationale, they risk positioning learners as passive recipients of automated authority rather than as developing cognitive agents acquiring transferable linguistic knowledge (Stevenson & Phakiti, 2014; Dikli & Bleyle, 2014).

The present article addresses this gap through a mixed-methods controlled trial involving secondary-school L2 English learners assigned to either a corrective or explanatory AWE feedback condition across an eight-week programme. We pursue three interrelated research questions: (RQ1) Do EF and CF conditions produce differential gains in analytic text quality measures at post-test? (RQ2) Do the two feedback conditions elicit qualitatively distinct patterns of metacognitive processing as evidenced in learner revision diaries? (RQ3) What do these findings imply for the theoretical modelling of AI-mediated feedback within instructional science?

The article proceeds as follows. Section 2 synthesises the relevant theoretical frameworks. Section 3 reviews empirical precedents. Section 4 details the methodological design. Section 5 presents findings. Section 6 discusses implications for instructional science and AWE design. Section 7 concludes with a research agenda.

2. THEORETICAL FRAMEWORK

2.1 Feedback as Instructional Design

Theoretical accounts of feedback in instructional contexts have moved progressively beyond behaviourist stimulus-response models toward cognitive and sociocognitive formulations that foreground the learner's active processing role (Hattie & Timperley, 2007; Winne & Butler, 1994). The highly influential model proposed by Hattie and Timperley (2007) conceptualises feedback as operating across four levels—task, process, self-regulation, and self—with

process-level and self-regulation-level feedback generally associated with deeper learning outcomes. This taxonomy is directly germane to the CF/EF distinction: corrective feedback that supplies a target form without explanation primarily addresses task-level accuracy, while explanatory feedback that articulates a linguistic rule ostensibly operates at the process level, potentially scaffolding the construction of generalisable linguistic knowledge.

Narciss (2008) elaborates a taxonomy of informative tutoring feedback that similarly distinguishes between knowledge-of-response (KR), knowledge-of-correct-response (KCR), and elaborated feedback (EF), with the latter subdivided by type of elaboration—procedural, conceptual, or strategic. AWE-generated corrective feedback approximates KCR, while explanatory feedback aligns with conceptual elaborated feedback in Narciss's typology. Critically, Narciss's meta-analytic synthesis suggests that the superiority of elaborated feedback over simpler forms is moderated by task complexity and learner prior knowledge, implying that the pedagogical value of EF-type AWE feedback may be contingent rather than unconditional.

2.2 Cognitive Load Theory and L2 Writing

Cognitive Load Theory (CLT; Sweller, 1988; Sweller et al., 2019) provides a complementary explanatory lens. CLT proposes that the human working memory system has finite capacity, and that instructional designs which impose unnecessary extraneous cognitive load impede schema acquisition. In the context of L2 writing, learners must simultaneously manage lexical retrieval, syntactic encoding, discourse organisation, and rhetorical goal-setting, already constituting a high intrinsic cognitive load (Manchon, 2011). The provision of AWE feedback introduces additional processing demands: learners must interpret the feedback, map it onto their existing interlanguage knowledge, and decide how to revise.

From a CLT perspective, corrective feedback may paradoxically reduce germane cognitive load in the short term by supplying the answer—but at the cost of foreclosing the effortful schema-building that underlies acquisition. Explanatory feedback, by contrast, may impose greater immediate load while simultaneously directing that load toward the construction of durable linguistic schemata. This tension mirrors the 'desirable difficulties' argument advanced by Bjork (1994), which holds that conditions making learning more effortful in the short term may produce superior long-term retention and transfer.

2.3 Self-Regulated Learning in Writing Contexts

The Self-Regulated Learning (SRL) framework (Zimmerman, 2000; Pintrich, 2000) contributes a third theoretical pillar. SRL models conceptualise learners as active agents who set goals, monitor progress, deploy cognitive strategies, and reflect on outcomes. In writing research, SRL has been operationalised through process-oriented accounts of planning, translation, and reviewing (Hayes & Flower, 1980; Bereiter & Scardamalia, 1987), as well as through metacognitive measures that capture learners' awareness of their own textual decision-making (Graham & Harris, 2000).

AWE tools interact with SRL processes in potentially significant ways. A system that provides only corrective substitutions may short-circuit the reviewing sub-process, preventing

the learner from engaging in the comparative evaluation of original and revised text that, according to SRL theory, is constitutive of adaptive self-monitoring. Conversely, explanatory feedback that invites the learner to understand why a linguistic choice is suboptimal may promote the self-evaluation and attribution processes that Zimmerman (2000) identifies as central to the development of self-regulation. This theoretical prediction is the primary motivation for the present investigation.

2.4 Interaction Hypothesis and Noticing in L2 Acquisition

Within the specific domain of L2 acquisition, Long's Interaction Hypothesis (1996) and Schmidt's Noticing Hypothesis (1990) provide further theoretical grounding. Schmidt (1990) argued that conscious noticing of a linguistic form in the input is a necessary condition for its acquisition; Long (1996) extended this by arguing that interactional feedback, including negative evidence, promotes noticing by drawing attention to form-meaning mismatches. Although AWE systems operate asynchronously and thus do not replicate the real-time dynamics of oral interaction, they can be theorised as approximations of negative evidence that prompt learners to attend to discrepancies between their interlanguage and target norms.

The CF/EF distinction maps onto this literature in the following way: while both types of feedback may promote noticing of error, EF additionally provides the metalinguistic scaffolding necessary for the learner to interpret the noticed form within a rule-governed linguistic system. Ellis et al. (2006) found that metalinguistic written corrective feedback produced superior long-term accuracy gains compared to direct correction on English article use, lending empirical support to the theoretical claim that explanation enhances the acquisitional value of feedback.

3. EMPIRICAL REVIEW: AWE FEEDBACK AND L2 LEARNING OUTCOMES

3.1 The Efficacy of AWE Systems

The accumulated empirical record on AWE efficacy presents a nuanced picture. Early validational studies focused predominantly on the reliability of holistic scoring, demonstrating that AWE scores correlated sufficiently with human rater scores to justify deployment in high-stakes contexts (Attali & Burstein, 2006; Shermis & Hamner, 2012). Subsequent research broadened to examine formative feedback functions, with Wilson and Czik (2016) finding that secondary students who received AWE feedback across multiple drafts produced texts with higher scores on analytic rubric dimensions than control peers who received only teacher feedback, though effect sizes were modest. Stevenson and Phakiti (2014) similarly reported positive effects on text quality but noted important individual difference moderators including motivation and digital literacy.

Critical perspectives, however, have problematised AWE efficacy findings on several grounds. Ericsson and Haswell (2006) raised concerns about the construct validity of AWE scoring models, arguing that surface-level linguistic features privileged by many systems do not adequately capture higher-order writing qualities such as argument coherence and rhetorical sophistication. Dikli and Bleyle (2014) found that AWE systems produced systematically different feedback from teacher feedback, particularly on discourse-level

features, with AWE tending to concentrate on local linguistic form rather than global organisation. These critiques underscore the need for research that disaggregates AWE feedback type rather than treating AWE as a homogeneous intervention.

3.2 Corrective vs. Explanatory Feedback: Bridging L2 Acquisition and Instructional Design

The written corrective feedback (WCF) literature within L2 writing research provides the most direct empirical precedents for the present study. The field has been substantially shaped by Truscott's (1996) provocative argument that grammar correction is ineffective and should be abandoned, and by Ferris's (1999) counter-argument that principled, focused correction can be beneficial. The subsequent three decades of research have largely vindicated a qualified version of Ferris's position: focused WCF targeting specific linguistic features produces meaningful accuracy gains, while unfocused WCF may be less effective (Bitchener & Knoch, 2010; Shintani & Ellis, 2013).

More pertinent to the present inquiry is the growing sub-literature comparing direct correction with metalinguistic or explanatory forms of feedback. Bitchener et al. (2005) compared direct correction, metalinguistic feedback, and a combination of the two on English article use in L2 writers, finding that all corrective feedback groups outperformed a control group and that metalinguistic feedback groups showed higher retention at a delayed post-test. Shintani and Ellis (2013) conducted a meta-analysis of studies comparing direct WCF with metalinguistic written feedback, concluding that metalinguistic feedback produced stronger effects on new writing tasks—a finding consonant with a transfer of learning interpretation. However, none of these studies employed AWE-generated feedback, leaving open the question of whether findings from teacher-generated WCF transfer to the automated context.

3.3 Metacognition, Revision Diaries, and Process-Oriented AWE Research

A smaller but growing body of research examines the metacognitive dimensions of AWE-mediated revision. Ranalli (2018) employed think-aloud protocols to investigate how L2 writers processed AWE feedback from Criterion, finding considerable variability in learner engagement, with some participants engaging in strategic self-monitoring while others accepted AWE suggestions uncritically. These findings resonate with SRL research suggesting that the relationship between external feedback and internal self-regulation is not automatic but is mediated by learner agency (Winne & Hadwin, 1998). Revision diaries—structured reflective logs in which writers document their revision decisions and rationale—have been proposed as a more ecologically valid alternative to real-time think-aloud methods for capturing metacognitive processes in writing (Manchon et al., 2009; Yu & Lee, 2016).

Taken together, the extant literature supports the hypothesis that feedback architecture matters for both cognitive and metacognitive outcomes in L2 writing, but direct experimental comparisons of CF and EF in AWE contexts remain scarce. The present study is designed to address this lacuna.

4. METHODOLOGY

4.1 Research Design

The study adopted a mixed-methods controlled trial design (Creswell & Plano Clark, 2018), combining quantitative pre- and post-test writing assessments with qualitative analysis of learner revision diaries. Controlled trials are increasingly advocated in educational research as a means of establishing causal inference about instructional interventions while preserving the ecological validity of naturalistic classroom contexts (Shadish et al., 2002). The mixed-methods design was motivated by the recognition that text quality scores alone are insufficient to illuminate the cognitive and metacognitive processes that constitute the theoretical constructs of interest (Creswell & Plano Clark, 2018).

4.2 Participants

Forty-eight secondary-school students (mean age = 15.6 years, SD = 0.7) enrolled in Grade 10 English as a Second Language courses at two urban secondary schools in a predominantly Mandarin-speaking context participated in the study. Participants were randomly assigned to one of two conditions: Corrective Feedback (CF; n = 24) or Explanatory Feedback (EF; n = 24). Random assignment was conducted at the individual level using a computerised random number generator. Informed consent was obtained from participants and their guardians in accordance with institutional ethical protocols. Participants were at an intermediate proficiency level (B1–B2 on the Common European Framework of Reference), determined by a pre-study placement assessment.

4.3 Instrumentation: The AWE Platform

A purpose-built AWE module was integrated into the school's existing LMS. The module was built on a rule-based natural language processing backbone supplemented by a statistical model trained on a corpus of L2 English learner texts. The system identified and flagged twelve target grammatical and lexical error categories, including article use, subject-verb agreement, verb tense, preposition selection, relative clause formation, and lexical collocation. For each flagged instance, the system generated one of two feedback types depending on condition assignment. In the CF condition, the system replaced the erroneous string with the target form (e.g., replacing '*He go to school yesterday' with 'He went to school yesterday'). In the EF condition, the system provided an explanatory comment specifying the relevant grammatical rule (e.g., 'This verb requires the simple past tense because the time adverb "yesterday" signals a completed past event. The correct form is "went".') without directly supplying the corrected string in-line.

4.4 Writing Tasks and Procedure

The study unfolded over eight weeks. In Week 1, participants completed a pre-test writing task (an argumentative essay on a standardised prompt, 250–300 words) without AWE assistance. Over Weeks 2–7, participants completed three graded writing tasks (informative, argumentative, and narrative genres), receiving their respective feedback type after each initial draft and submitting one revised draft per task. In Week 8, participants completed a post-test writing task (a new argumentative prompt, parallel in difficulty to the pre-test) again

without AWE assistance. This delayed post-test design was chosen to assess transfer rather than merely practised performance. Participants in both conditions also maintained structured revision diaries throughout Weeks 2–7, responding to three weekly prompts designed to elicit metacognitive reflection: (a) What errors did the feedback identify? (b) Why do you think these errors occurred? (c) What will you do differently in future writing?

4.5 Measures

4.5.1 Analytic Writing Quality

Pre- and post-test essays were scored by two trained raters using a six-dimensional analytic rubric adapted from Jacobs et al.'s (1981) ESL Composition Profile. Dimensions included Content, Organisation, Vocabulary, Language Use (grammatical accuracy), Mechanics, and Coherence. Inter-rater reliability was established prior to scoring (Cohen's kappa = .81), and disagreements exceeding one scale point were resolved by a third rater. A composite Accuracy score and a composite Global Quality score were computed for analysis.

4.5.2 Revision Diary Analysis

Revision diary entries were analysed using a theoretically driven qualitative content analysis framework (Mayring, 2015) informed by the SRL coding scheme developed by Zimmerman and Risemberg (1997). Four metacognitive codes were applied: Self-Monitoring (SM), Self-Evaluation (SE), Causal Attribution (CA), and Strategic Planning (SP). Two coders independently applied the scheme; intercoder reliability was computed (Cohen's kappa = .77). Quantitative frequency counts of each code per condition were computed to enable comparative analysis, while representative extracts were selected for illustrative qualitative interpretation.

4.6 Data Analysis

Quantitative writing quality data were analysed using a series of mixed ANOVA models with Feedback Condition (CF vs. EF) as a between-subjects factor and Time (pre-test vs. post-test) as a within-subjects factor. Effect sizes were computed as partial eta-squared (ηp^2). Qualitative diary data were analysed using the content analysis procedure described above, with frequency counts subjected to independent samples t-tests to compare metacognitive code frequencies across conditions. A convergent triangulation logic was applied to integrate quantitative and qualitative findings (Creswell & Plano Clark, 2018).

5. RESULTS

5.1 Text Quality Outcomes

The mixed ANOVA for composite Global Quality yielded a significant Time × Condition interaction, $F(1, 46) = 7.42$, $p = .009$, $\eta p^2 = .139$. Both groups improved from pre- to post-test, consistent with a general practice effect; however, the EF group demonstrated a significantly larger gain ($M\Delta = 8.3$ points on a 100-point scale, $SD = 6.1$) compared to the CF group ($M\Delta = 4.9$ points, $SD = 5.8$). Dimension-level analysis indicated that the between-group advantage for EF was most pronounced on the Coherence dimension ($F(1, 46) = 9.13$, $p = .004$, $\eta p^2 = .166$) and the Vocabulary dimension ($F(1, 46) = 5.87$, $p = .019$, $\eta p^2 = .113$), with no significant group difference on the Mechanics dimension ($F(1, 46) = 0.41$, $p = .527$).

For the composite Accuracy score, the pattern was partially reversed. The CF group showed a steeper trajectory of accuracy gains during the intervention period (Weeks 2–7), though this advantage was not sustained at the delayed post-test; no significant Time × Condition interaction was observed for Accuracy at post-test ($F(1, 46) = 1.89$, $p = .175$). This pattern is consistent with the prediction that corrective feedback promotes short-term surface accuracy but does not produce durable transfer, while explanatory feedback supports the development of linguistic schemata that generalise across new tasks.

5.2 Revision Diary Findings

Content analysis of revision diary entries yielded meaningful between-group differences in metacognitive code frequencies. EF participants produced significantly more Self-Monitoring utterances per diary entry ($M = 4.2$, $SD = 1.8$) than CF participants ($M = 2.1$, $SD = 1.4$), $t(46) = 4.67$, $p < .001$, $d = 1.31$. Similarly, Self-Evaluation utterances were significantly more frequent in the EF condition ($M = 3.7$, $SD = 1.6$ vs. $M = 1.9$, $SD = 1.2$), $t(46) = 4.48$, $p < .001$, $d = 1.26$. No significant differences were found for Causal Attribution or Strategic Planning frequencies.

Qualitatively, EF diary entries were characterised by explicit linguistic reasoning and rule generalisation. One representative EF participant wrote: 'The system explained that I should use the present perfect when the time period is not specified. I realised I do this in Chinese too but it is different in English. I will check this rule every time I write about experiences.' This entry exemplifies Self-Monitoring, Self-Evaluation, Causal Attribution (cross-linguistic interference), and Strategic Planning in a single reflective passage. By contrast, representative CF diary entries typically acknowledged the correction without evidence of rule internalisation: 'The system changed my verb to "had gone". I corrected it. I will try to be more careful.'

These patterns suggest that EF architecture elicited qualitatively richer metacognitive engagement, including explicit rule-formulation and cross-linguistic reflection that CF architecture did not. This finding is theoretically significant because it implies that the mechanism through which EF promotes superior text quality gains may be precisely the metacognitive processes it activates, supporting the SRL-based theoretical account advanced in Section 2.

6. DISCUSSION

6.1 Feedback Architecture and the Cognitive Mechanisms of L2 Writing Development

The findings of the present study speak directly to a core question in instructional science: what is the cognitive mechanism through which feedback promotes learning? Our results are consistent with a schema-construction account (Sweller et al., 2019) in which EF, by providing explicit metalinguistic information, enables learners to abstract a generalisable rule that can be applied to novel instances in the delayed post-test. CF, by supplying the correct form directly, may reduce immediate cognitive effort but does so by foreclosing the generative processing—what Bjork (1994) calls the desirable difficulty—that is necessary for durable schema formation. The asymmetry between intervention-period accuracy gains

(favouring CF) and delayed post-test global quality gains (favouring EF) is precisely the pattern that this theoretical account would predict.

This interpretation is further supported by the diary evidence. The higher frequency of Self-Monitoring and Self-Evaluation utterances in EF diaries is consistent with Zimmerman's (2000) model of self-regulated learning, in which monitoring and evaluation are the mechanisms through which learners build internal performance standards that eventually supplant dependence on external feedback. If AWE-generated corrective feedback systematically suppresses these self-regulatory processes, it may paradoxically undermine the long-term development of writing competence that it is designed to promote—a concern with significant implications for AWE tool deployment at scale.

6.2 Implications for AWE Design and Deployment

From a practical standpoint, the findings suggest that the feedback architecture of AWE tools should be treated as a substantive pedagogical decision rather than a mere technical default. Many commercially available AWE systems are designed primarily for efficiency—maximising the surface accuracy of texts with minimal learner effort—which, on our evidence, may be counter-productive for long-term learning. A redesigned AWE system informed by the present findings would, at minimum, provide explanatory rationale alongside corrected forms, and might additionally incorporate features that require learners to confirm their understanding of the rule before accepting a correction (see also Ranalli, 2018, for related design suggestions).

Furthermore, the finding that EF advantages are concentrated on Coherence and Vocabulary—higher-order textual dimensions—rather than on Mechanics suggests that explanatory feedback may promote a qualitatively different kind of writing development, one more closely aligned with the disciplinary goals of secondary English language education than narrowly construed grammatical accuracy. This is consequential for how AWE is positioned within curriculum: if the goal is to develop sophisticated communicative competence rather than merely error-free text, then EF architectures merit clear preference.

6.3 Theoretical Contributions to Instructional Science

The present study makes three contributions to instructional science. First, it provides experimental evidence that the type of feedback embedded in AI-mediated writing evaluation has differential effects on learning outcomes, adding to the small but growing evidence base on AI feedback in formal education (Cavalcanti et al., 2021; Kochmar et al., 2022). Second, it demonstrates the theoretical productivity of integrating CLT, SRL, and L2 acquisition frameworks in a unified account of AWE-mediated learning—a cross-theoretical synthesis that, to our knowledge, has not been previously articulated in this form. Third, by employing revision diaries as a process measure alongside product-based writing scores, it demonstrates the value of mixed-methods approaches for illuminating the cognitive processes through which feedback exerts its effects, addressing the methodological critique that single-outcome designs treat the learning process as a black box.

6.4 Limitations and Future Directions

Several limitations qualify these findings. First, the study was conducted at two schools in a single L1-background context; generalisation to typologically distinct L1 backgrounds requires further investigation. Second, the AWE system employed in this study was purpose-built for research purposes and may not fully replicate the characteristics of commercially deployed systems, which vary considerably in their feedback granularity and linguistic coverage. Third, the eight-week study duration, while sufficient to detect group differences, does not permit conclusions about the persistence of effects over a full academic year. Fourth, the controlled assignment of participants to feedback conditions, while necessary for causal inference, creates an artificial learning environment that may not capture the mediating effects of learner choice and agency that characterise authentic AWE use. Future research should address these limitations through longitudinal designs, multi-site replications across diverse L1 contexts, and quasi-experimental studies of authentic AWE deployments.

7. CONCLUSION

This article has argued that the feedback architecture of Automated Writing Evaluation tools is not instructionally neutral. Through a mixed-methods controlled trial comparing corrective and explanatory AWE feedback in secondary-school L2 English writing, we have demonstrated that explanatory feedback produces superior long-term gains in global text quality and elicits richer metacognitive engagement in learner revision diaries, while corrective feedback generates faster but less durable surface accuracy improvements. These findings—consistent with Cognitive Load Theory, the Self-Regulated Learning framework, and the metalinguistic feedback literature in L2 acquisition—illuminate the cognitive mechanisms through which AWE feedback exerts its effects and point toward a principled redesign of AWE systems that prioritises explanation and learner agency over automated correction.

More broadly, the study invites instructional scientists to engage seriously with the pedagogical architectures encoded in AI-mediated educational tools. As AWE and, increasingly, generative AI systems become ubiquitous in writing education, the question of how they model the feedback relationship—and what cognitive and self-regulatory demands they make of learners—becomes not merely a technical but a deeply ethical and educational one. Deconstructing the black box of AWE, this study suggests, is not merely an empirical project but a normative one: it is an argument for designing AI educational tools that regard learners as developing cognitive agents rather than as error-generating machines in need of automated repair.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. Journal of Technology, Learning and Assessment, 4(3), 1–30.

Bereiter, C., & Scardamalia, M. (1987). The psychology of written composition. Lawrence Erlbaum.

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. Applied Linguistics, 31(2), 193–214.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. Journal of Second Language Writing, 14(3), 191–205.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). MIT Press.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gasevic, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. Computers and Education: Artificial Intelligence, 2, 100027.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing, 22, 1–16.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2006). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. System, 36(3), 353–371.

Ericsson, P. F., & Haswell, R. H. (Eds.). (2006). Machine scoring of student essays: Truth and consequences. Utah State University Press.

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11.

Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. Educational Psychologist, 35(1), 3–12.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), Cognitive processes in writing (pp. 3–30). Lawrence Erlbaum.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Newbury House.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. Journal of Writing Research, 1(1), 1–26.

Kochmar, E., Do Vu, D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2022). Automated personalized feedback improves learning gains in an intelligent tutoring system. In Proceedings of the 15th International Conference on Educational Data Mining (pp. 233–243). EDM.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. Educational Psychology Review, 1(4), 279–308.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468). Academic Press.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. Studies in Second Language Acquisition, 19(1), 37–66.

Manchon, R. M. (2011). Learning-to-write and writing-to-learn in an additional language. John Benjamins.

Manchon, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing. In R. M. Manchon (Ed.), Writing in foreign language contexts: Learning, teaching, and research (pp. 102–129). Multilingual Matters.

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), Approaches to qualitative research in mathematics education (pp. 365–380). Springer.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. Van Merrienboer, & M. P. Driscoll (Eds.), Handbook of research on educational communications and technology (3rd ed., pp. 125–143). Lawrence Erlbaum.

Nassaji, H. (2016). Research timeline: Interactional feedback in second language teaching and learning. Language Teaching, 49(4), 535–566.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 451–502). Academic Press.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? Computer Assisted Language Learning, 31(7), 653–674.

Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129–158.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 313–346). Routledge.

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite and definite articles. Journal of Second Language Writing, 22(3), 286–306.

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. Assessing Writing, 19, 51–65.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years on. Educational Psychology Review, 31(2), 261–292.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3(1), 22–36.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. Computers & Education, 100, 94–109.

Winne, P. H., & Butler, D. L. (1994). Student cognition in learning from teaching. In T. Husen & T. Postlethwaite (Eds.), International encyclopedia of education (2nd ed., pp. 5738–5745). Pergamon.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Metacognition in educational theory and practice (pp. 277–304). Lawrence Erlbaum.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). Language Teaching, 49(4), 461–493.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13–39). Academic Press.

Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. Contemporary Educational Psychology, 22(1), 73–98.

Sapiro, G. (2010). Globalization and cultural diversity in the book market. Poetics, 38(4), 419-439.

Scott, C. (2022). Shamanic translation: Indigenous models for interspecies communication. Translation and Interpreting Studies, 17(3), 412-432.

Sebeok, T. A. (2001). Biosemiotics: An interdisciplinary project. University of Toronto Press.

Seyfarth, R. M., & Cheney, D. L. (2017). The social origins of language. Princeton University Press.

Smith, L. (2023). Whose coda? Critique of the Whale Song Translation Project. Comparative Literature Studies, 60(2), 288-307.

Spivak, G. C. (1993). Outside in the teaching machine. Routledge.

Toury, G. (2012). Descriptive translation studies and beyond (2nd ed.). John Benjamins.

Tsing, A. L. (2023). Reading failure: Santamaria's plant poetics. Environmental Humanities, 15(2), 201-218.

Van Dooren, T. (2019). The wake of crows: Living and dying in shared worlds. Columbia University Press.

Venuti, L. (1995). The translator's invisibility: A history of translation. Routledge.

Venuti, L. (2024). Against computational pseudomorphism: A critique of algorithmic translation. Translation Studies, 17(1), 22-41.

Vidal, C. (2018). Translating the jaguar: Indigenous perspectivism and ecotranslation. Target, 30(2), 245-268.

Waldron, P. (2022). Unreadable poems, unlivable worlds. Chicago Review, 65(3), 88-102.