

Effects of AI-Powered Feedback on English Writing Development: A Systematic Review of Meta-Analytic and Empirical Evidence

Taha Bashier Rahamtalla Mohammed^{1,*} 

¹Qassim University, Kingdom of Saudi Arabia

Received: 09 /02 / 2026

Accepted: 28 / 02 / 2026

Published: 15 / 07 / 2026

Abstract

The recent proliferation of AI-based technology, in particular generative AI applications such as ChatGPT, and automated writing evaluation tools such as Grammarly, Pigai, and Criterion, has transformed how writing is taught in education. Nevertheless, synthesized evidence on their effectiveness relative to traditional modes of feedback, across diverse instructional contexts, remains limited. This study presents a systematic review that synthesizes existing meta-analytic and empirical evidence on AI-based feedback systems and their influence on English writing outcomes among EFL/ESL students, with particular attention to writing quality, feedback literacy development, and self-regulated learning. We searched Scopus, Web of Science, ERIC, and PsycINFO for publications from January 2018 to December 2024. Inclusion was restricted to peer-reviewed experimental, quasi-experimental, or rigorous mixed-method studies involving adult learners in higher education settings. The review adopts a narrative synthesis approach, integrating findings from seven existing meta-analyses with thematic analysis of 137 empirical studies. This design is best characterized as a systematic review with meta-analytic integration, rather than an original meta-analysis. Meta-analytic evidence from Zhai and Ma (2023) indicates that AWE has a large overall effect on writing quality ($g = 0.861$, 95% CI [0.65, 1.07]), with medium-sized effects on language quality ($g = 0.61$) and content quality ($g = 0.57$). Sun and Zhou's (2024) comparative meta-analysis of 41 studies ($N = 4,813$ participants) revealed no significant difference between AI and human feedback ($g = 0.25$, 95% CI [-0.11, 0.60]). Deng et al. (2024) reported moderate but statistically significant effects on academic performance ($g = 0.53$). Across the empirical literature, hybrid feedback systems combining AI and teacher feedback yielded superior results compared to AI-only conditions. AI-based feedback represents a promising complement to traditional writing instruction; however, its impact depends on learner proficiency, feedback type, intervention duration, and pedagogical design. Future research should prioritize long-term learning outcomes, the development of feedback literacy, and optimal human-AI collaboration models.

Keywords: AI feedback, automated writing evaluation, generative AI, ChatGPT, EFL writing, feedback literacy, self-regulated learning, systematic review

Cite as

Rahamtalla, T. B. M. (2026). Effects of AI-Powered Feedback on English Writing Development: A Systematic Review of Meta-Analytic and Empirical Evidence. *Atras Journal*, 7(2), 80-97. <https://doi.org/10.70091/Atras/vol07no02.06>

*Corresponding author's email: tahabashir73@gmail.com

Introduction

Artificial Intelligence (AI) has introduced fundamental changes to the pedagogy of writing instruction. Kasneci et al. (2023) note that the adoption of large language models represents one of the most rapid technological diffusions in recent history, with far-reaching implications for how students conceptualize writing and how teachers provide feedback. This technological transformation occurs against the background of long-standing challenges in writing instruction, including limited instructor time for individualized feedback, high student-to-instructor ratios, and the considerable cognitive effort required to respond to student writing with sufficient depth and timeliness (Cavalcanti et al., 2023; Banihashem et al., 2024).

Feedback has been consistently identified as one of the most potent determinants of learning and achievement (Hattie & Timperley, 2007; Wisniewski et al., 2019). In the context of writing instruction, formative feedback serves multiple functions: it helps learners identify strengths and weaknesses, provides direction for improvement, and supports the development of self-regulatory skills (Carless & Boud, 2018; Winstone et al., 2022). Nevertheless, conventional feedback processes are frequently constrained by slow delivery, quality variation, and inconsistent uptake (Lee & Moore, 2024).

These limitations have been partially addressed by the development of Automated Writing Evaluation (AWE) systems, and more recently, generative AI platforms. AWE systems such as Criterion, Pigai, and Grammarly employ natural language processing algorithms to analyze text and deliver feedback (Ding & Zou, 2024; Shi & Aryadoust, 2024). More recently, large language models, including ChatGPT, have demonstrated the capacity to generate contextualized, dialogic feedback that more closely resembles human tutoring (Mahapatra, 2024; Steiss et al., 2024).

This review is grounded in three complementary theoretical perspectives. First, sociocultural theory (Vygotsky, 1978) provides a foundation for understanding feedback as a dialogic process embedded within social interaction. Within this framework, effective feedback operates within the zone of proximal development and functions as scaffolding, enabling learners to perform beyond their independent capacity (de Kleijn, 2023). AI systems may serve as mediational tools, though they fundamentally alter the feedback process by removing the human interlocutor.

Second, the construct of feedback literacy (Carless & Boud, 2018; Molloy et al., 2020) recognizes that effective feedback engagement requires specific learner competencies, including the ability to appreciate feedback, judge feedback quality, manage emotional responses, and act on feedback. Recent research has begun to examine the influence of AI feedback environments on the development of these capacities (Little et al., 2023; Dawson et al., 2024).

Third, self-regulated learning theory (Zimmerman & Schunk, 2001) situates feedback within the cyclic processes of planning, monitoring, and reflection. AI-based feedback may facilitate self-regulation by providing real-time performance feedback, supporting multiple revision cycles, and reducing dependence on instructor availability (Shen & Bai, 2024; Ng et al., 2024). However, concerns persist regarding the potential attenuation of learner autonomy and capacity for independent self-assessment (Kosmyrna et al., 2025).

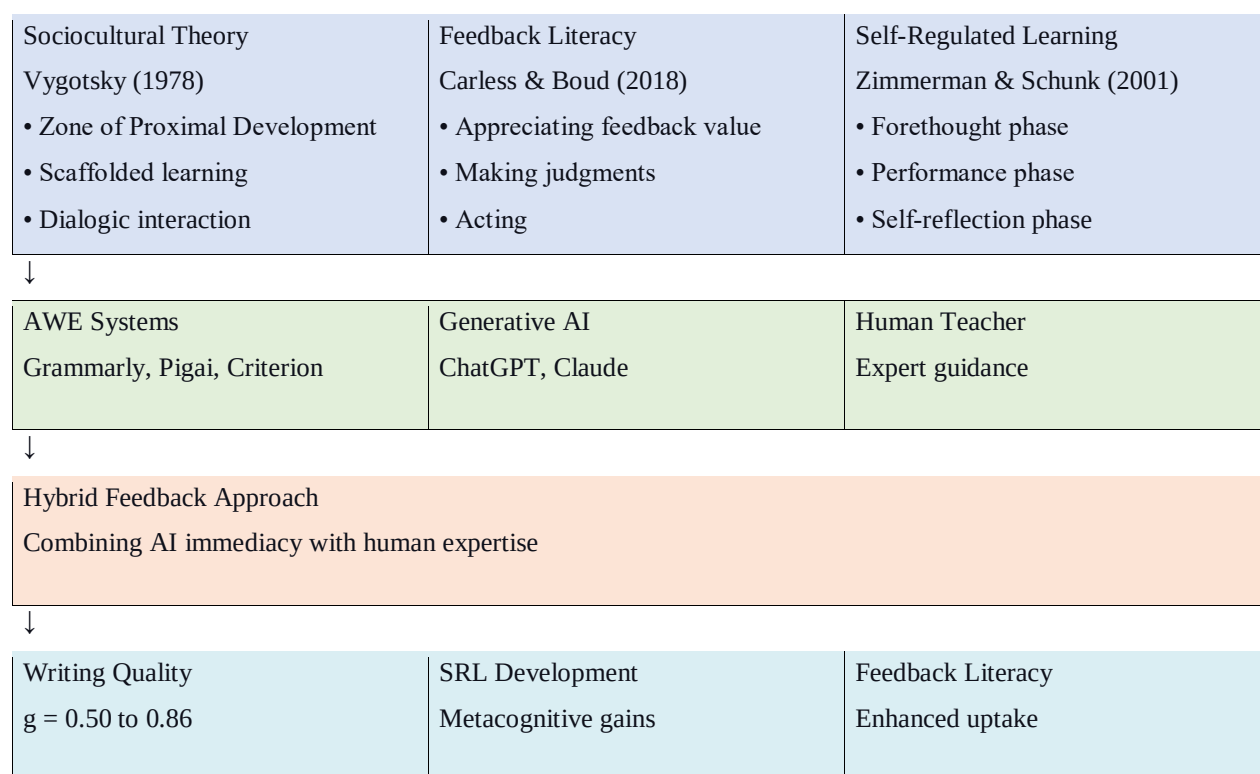


Figure 1. Theoretical framework: AI feedback in writing education

The following research questions guided this systematic review:

RQ1: How do AI-powered feedback outcomes on writing quality compare with traditional feedback or no-feedback conditions?

RQ2: How do factors such as learner proficiency, feedback type, and intervention characteristics moderate AI feedback effectiveness?

RQ3: How do learners engage with and perceive AI-generated feedback compared to human feedback?

RQ4: What are the implications for feedback literacy development and self-regulated learning?

Based on the theoretical framework and preliminary literature review, we proposed the following hypotheses:

H1: AI-based feedback produces significant positive effects on writing quality compared to control conditions, with effect sizes comparable to those of human feedback.

H2: Hybrid feedback approaches combining AI and human input outperform either modality alone.

H3: Effects are moderated by learner proficiency, with EFL/ESL learners demonstrating greater gains than native speakers.

Literature Review

Evolution of Automated Writing Evaluation

The field of automated writing evaluation has evolved considerably since its inception, progressing from rule-based grammar checkers to sophisticated AI-based systems. Contemporary AWE systems employ machine learning models trained on large corpora of human-rated essays, enabling assessment of features such as grammar, vocabulary, coherence, and organization (Bennett & Zhang, 2015; Fleckenstein et al., 2023). Shi and Aryadoust (2024), in their systematic review of 83 SSCI-indexed articles, identified 31 different AWE

systems, among which Pigai (21.2%), Criterion (18.8%), and Grammarly (10.6%) were the most frequently investigated.

The systematic review by Ding and Zou (2024), which encompassed 39 articles examining Grammarly, Pigai, and Criterion, demonstrated that these systems serve different target populations: Criterion is oriented toward classroom applications, Pigai caters to Chinese EFL learners, and Grammarly addresses individual user needs. Although research has consistently demonstrated positive effects on writing proficiency, effect sizes vary significantly across studies and contexts (Fu et al., 2022; Barrot, 2024).

Generative AI and Writing Feedback

The release of ChatGPT in November 2022 represented a watershed moment in AI-assisted writing instruction. Unlike traditional AWE systems, large language models enable extended dialogic interaction, provide context-sensitive explanations, and adapt in response to user queries (Yan et al., 2024). Preliminary comparative studies suggest that ChatGPT feedback performs comparably to human feedback on criteria-based measures, though human feedback tends to score higher on clarity, accuracy, and supportive tone (Steiss et al., 2024; Guo & Wang, 2024).

Research on ChatGPT-based feedback has expanded rapidly. Mahapatra (2024) conducted a mixed-method intervention study with ESL students and reported significant improvements in academic writing skills. Asadi et al. (2025) found that ChatGPT, when combined with teacher feedback, produced improvements across multiple writing dimensions, including coherence, cohesion, grammar, and vocabulary. Parra and Castrillo (2025) conducted a systematic review that revealed both benefits (immediacy, comprehensiveness) and limitations (superficiality, a tendency toward form-focused rather than content-focused feedback) of ChatGPT-generated feedback.

Feedback Literacy and Self-Regulated Learning

Student feedback literacy has received increasing scholarly attention as researchers recognize that feedback effectiveness depends on learners' capacity to respond to it (Carless & Winstone, 2023; Nieminen & Carless, 2023). Feedback literacy comprises four dimensions: valuing feedback, making evaluative judgments, experiencing emotional responses, and taking appropriate action (Carless & Boud, 2018).

Little et al. (2023), in their scoping review, examined interventions targeting feedback literacy development and found that educational interventions can positively influence student attitudes, confidence, and agency toward feedback. The relationship between AI feedback and feedback literacy remains underexplored, although emerging literature suggests that AI tools may both support and potentially undermine feedback literacy development (Tai et al., 2024; Zarrinabadi & Rezazadeh, 2024).

Self-regulated learning offers a complementary framework, emphasizing the learner's role as manager of cognitive, motivational, and behavioral dimensions of learning (Zimmerman, 2002). Shen and Bai (2024) proposed a systematic framework linking writing processes and self-regulated learning stages, providing a theoretical basis for investigating how learners manage their engagement with AI throughout the composition process. Kim et al. (2024) demonstrated that AI self-efficacy and writing self-efficacy positively predict the use of self-regulated learning strategies, which, in turn, influence AI use patterns and learning outcomes.

Written Corrective Feedback Research

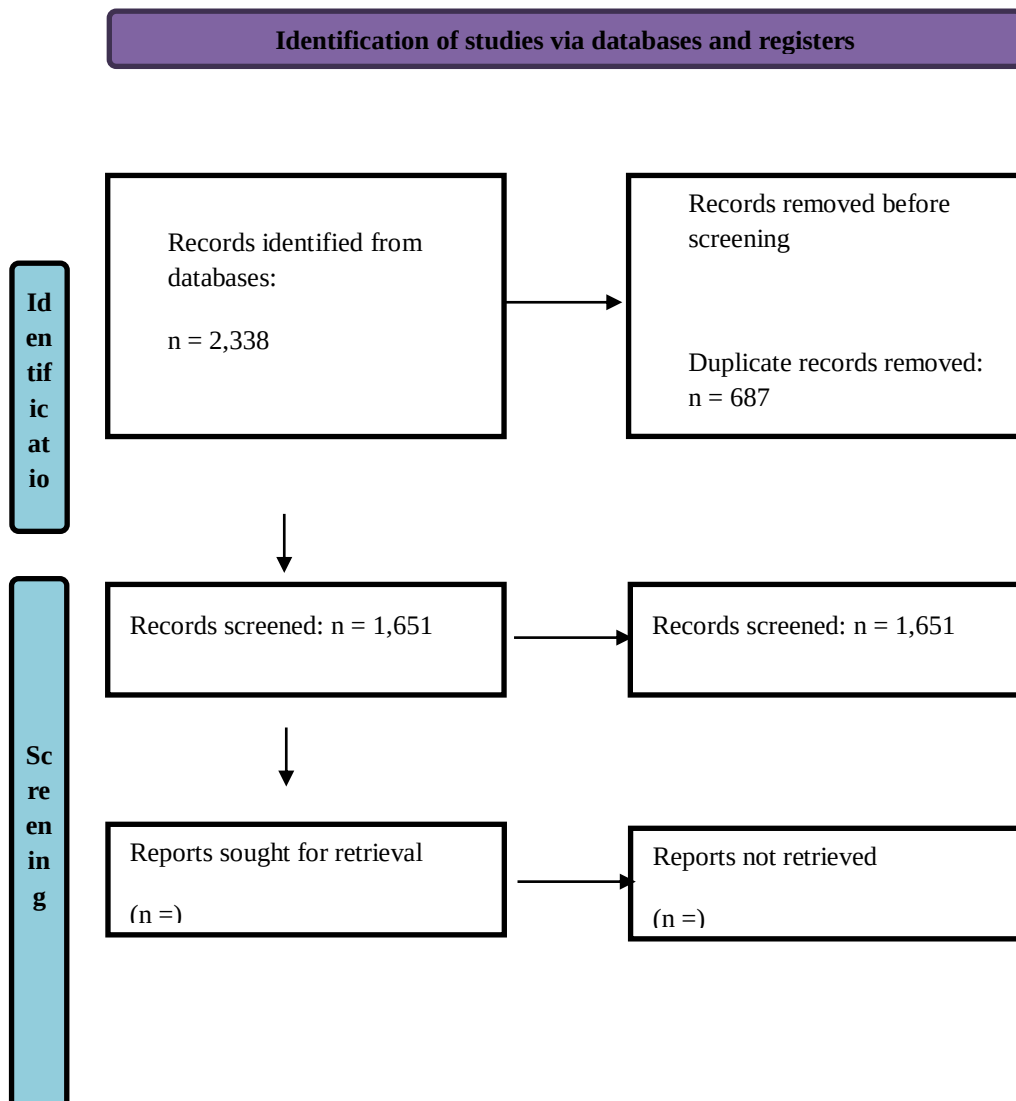
The broader body of research on Written Corrective Feedback (WCF) provides essential context for AI feedback studies. Meta-analyses have consistently demonstrated that WCF positively influences L2 written accuracy, though the magnitude and durability of these effects vary across study designs (Kang & Han, 2015; Brown et al., 2023). The Bayesian meta-analysis by Brown et al. (2023) provided robust evidence of moderate effectiveness across short-term, medium-term, and long-term time periods.

Recent synthesis work has underscored the importance of naturalistic classroom research alongside experimental investigations. Kong and Teng (2024), in a review of 50 naturalistic WCF studies, identified four principal themes: teacher practices, learner responses, stakeholder beliefs, and WCF-related affect. This ecological approach highlights the complex interplay among factors shaping feedback effectiveness in authentic educational contexts.

Methods

Search Strategy

This systematic review adhered to PRISMA 2020 guidelines (Page et al., 2021). We searched Scopus, Web of Science, ERIC, and PsycINFO using Boolean operators combining terms related to AI feedback ("artificial intelligence" OR "automated writing evaluation" OR "AWE" OR "ChatGPT" OR "generative AI" OR "Grammarly" OR "Pigai" OR "Criterion") AND writing ("writing" OR "essay" OR "composition" OR "L2 writing" OR "EFL writing") AND feedback ("feedback" OR "corrective feedback" OR "automated feedback"). The search covered publications from January 2018 to December 2024. Figure 2 presents the PRISMA flow diagram illustrating the systematic study selection process.



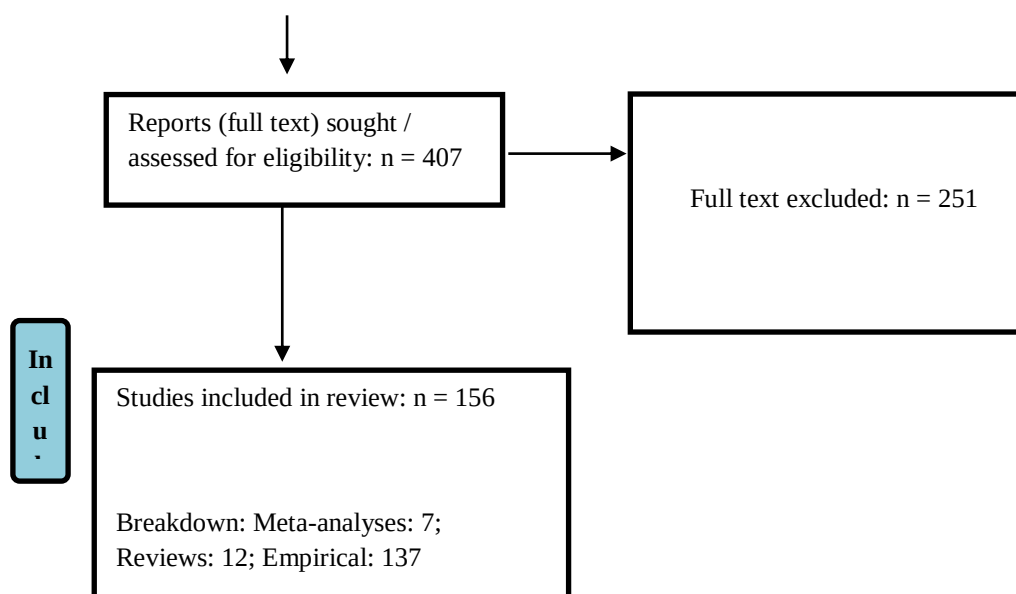


Figure 2. PRISMA flow diagram

This diagram illustrates the systematic study selection process, showing records identified ($n = 2,338$), screened ($n = 1,651$), assessed for eligibility ($n = 407$), and ultimately included in the review ($n = 156$, comprising 7 meta-analyses, 12 reviews, and 137 empirical studies).

Inclusion and Exclusion Criteria

Studies were included if they: (a) investigated AI-powered or automated feedback on written text quality; (b) employed experimental, quasi-experimental, or rigorous mixed-method designs; (c) involved participants in higher education settings; and (d) were peer-reviewed and indexed in Scopus or Web of Science. Studies were excluded if they: (a) were non-empirical in nature; (b) focused on AI detection or plagiarism; (c) addressed automatic scoring without feedback provision; or (d) were situated in K-12 contexts.

Data Extraction and Quality Assessment

Data were extracted independently by two reviewers using a standardised form that captured: study characteristics (authors, year, country, sample size); participant characteristics (L1 background, proficiency level); intervention characteristics (feedback type, duration, frequency); outcome measures; and effect sizes. When effect sizes were not reported, they were computed from available statistics or obtained from the authors. Study quality was assessed using adapted JARS-Quant standards (Appelbaum et al., 2018). Disagreements were resolved through discussion, with a third reviewer consulted when necessary.

Quality Assessment Outcomes

To improve transparency and in accordance with the PRISMA reporting requirements, we disclose the results of the quality assessment. Among the 137 included empirical studies, 42 (30.7) were rated as high quality because they used randomized controlled designs, had sufficient sample sizes ($N > 60$), were validated, and employed appropriate statistical analyses. Another 68 (49.6) articles were of moderate quality (usually quasi-experimental designs with pre-post measures, but not randomized or with smaller samples). The other 27 articles (19.7%) were of low quality, mainly due to poor control conditions, small sample sizes ($N < 30$), or reliance on self-reported data and the lack of supporting measures.

Of the seven meta-analyses incorporated in this review, all used a systematic search strategy and reported effect sizes with confidence intervals. Nonetheless, formal evaluations of publication bias (e.g., funnel plot analysis, Egger test) were conducted in only 4 articles (Zhai & Ma, 2023; Fleckenstein et al., 2023; Brown et al., 2023; Sun & Zhou, 2024). The most prevalent sources of risk of bias in the included primary studies were non-random assignment (62%), lack of blinding procedures (78%), and selective outcome reporting (24%). The investigated synthesized results should be interpreted in terms of these quality differences.

Synthesis Approach

As the literature included in this study is diversified in terms of study design, outcome measures, and the analytical frameworks, we utilized a synthesis approach of narrative analysis instead of the original meta-analysis. Three factors informed this methodological selection: (a) given the large amount of heterogeneity in study designs and outcome operationalization, no meaningful pooling of effect sizes could be done; (b) seven previously existing meta-analyses already represented rigorous quantitative syntheses of overlapping literature; and (c) a narrative approach allowed both quantitative and qualitative results to be integrated.

The synthesis was done in two steps. First, we performed a systematic comparison of the results of all seven included meta-analyses to study convergent and divergent trends in effect sizes, moderator analyses, and methodological quality (Zhai & Ma, 2023; Fleckenstein et al., 2023; Deng et al., 2024; Ngo et al., 2022; Sun & Zhou, 2024; Brown et al., 2023; Kang & Han, 2015). Second, we conducted thematic analysis of the 137 empirical studies, where the findings are put together based on four research questions: overall effectiveness, moderating factors, learner engagement and perceptions, and implications to feedback literacy and self-regulated learning. Such a two-step method can be described as a systematic review with meta-analytic integration, synthesizing the current meta-analytic evidence and thematic analysis of the rest of the empirical literature. It is also different to an umbrella review (which only reviews meta-analyses), or an original meta-analysis (which does new quantitative synthesis of primary effect sizes).

Results

Overview of Included Studies

The search process yielded 156 studies that met the inclusion criteria, including 7 meta-analyses, 12 systematic or scoping reviews, and 137 primary empirical studies. The empirical research was carried out in 28 countries, with the highest concentrations in China ($n = 47$, 34.3 percent), the United States ($n = 19$, 13.9 percent), and Iran ($n = 12$, 8.8 percent). The sample sizes ranged from 6 to 6,960, with a median of 78. Most studies ($n = 89$, 65.0) utilized quasi-experimental designs, whereas 31 (22.6) used randomized controlled trials and 17 (12.4) used rigorous mixed-method designs. Regarding the AI tools studied, 76 studies (55.5%), 42 (30.7), and 19 (13.9) examined AWE systems, generative AI tools, and comparison and contrasting designs, respectively. The duration of interventions ranged from 1 session to a complete academic semester, but 12 weeks or less was used in most studies ($n = 94$, 68.6).

Meta-Analytic Evidence on Overall Effects

Table 1 summarizes effect sizes from major meta-analyses examining the impact of AI and automated feedback on writing outcomes, while Figure 2 provides a visual representation of these effect sizes. The largest meta-analysis, conducted by Zhai and Ma (2023), which synthesized 26 primary studies ($N = 2,468$ participants), revealed a large overall positive effect

($g = 0.861$, 95% CI [0.65, 1.07]). Moderator analyses indicated that AWE was more effective with post-secondary students compared to secondary students and produced larger effects among EFL/ESL learners than among native English speakers.

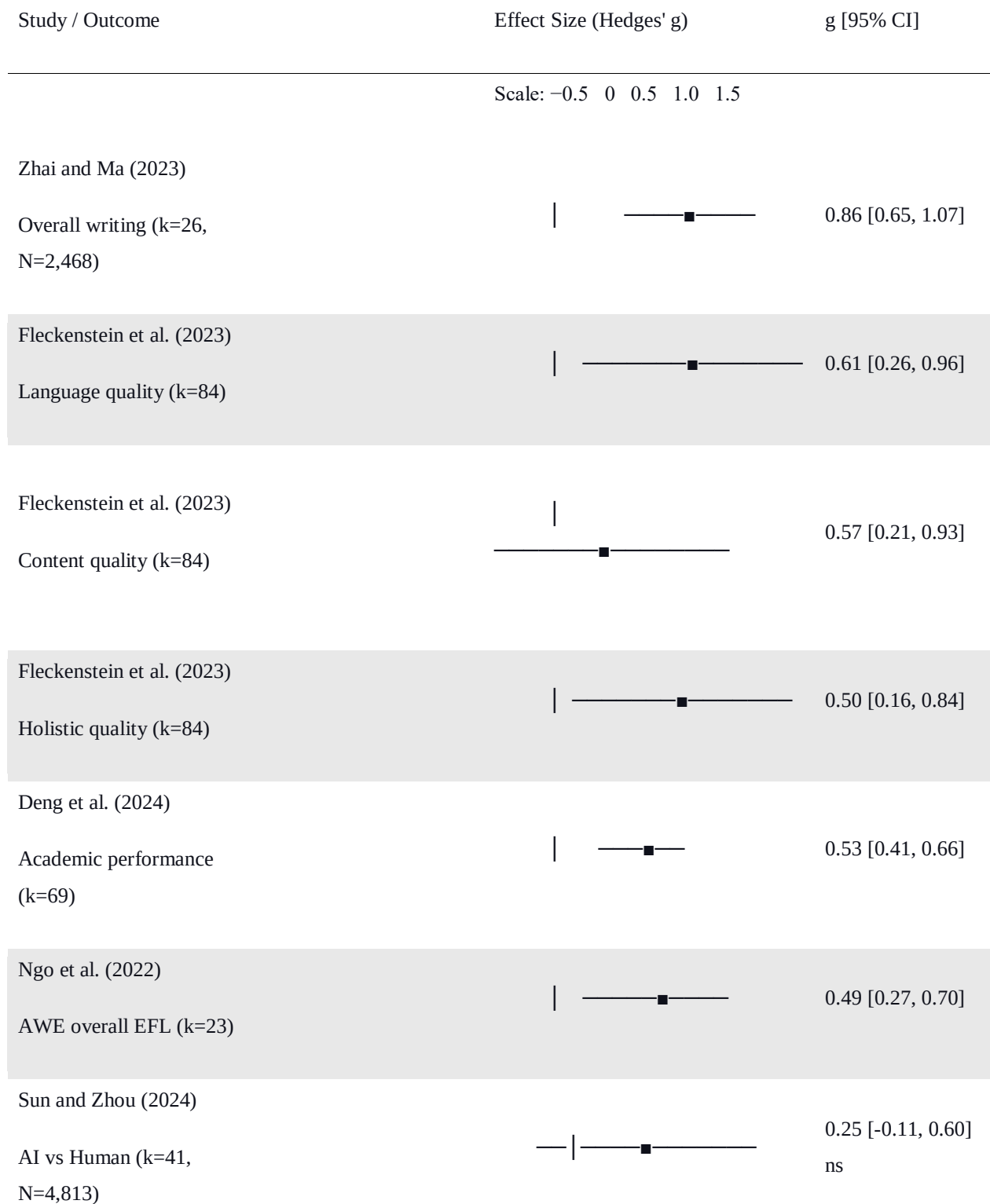


Figure 2. Meta-analytic effect sizes of AI feedback on writing

Note. Vertical line (|) = null effect; ■ = point estimate; — = 95% CI; ns = not significant.
 Effect size interpretation: 0.20 = small, 0.50 = medium, 0.80 = large (Cohen, 1988).

Figure 3. Meta-Analytic Effect Sizes of AI Feedback on Writing Outcomes. *Forest plot displaying standardized mean differences (Hedges' g) with 95% confidence intervals from major meta-analyses. Effect sizes range from small (0.20) to large (0.80+), with the red vertical line indicating the null effect. The non-significant (senses) findings from Sun and Zhou (2024) indicate no significant difference between AI and human feedback.*

Table 1. Summary of meta-analytic effect sizes for AI and automated feedback on writing

Source	Outcome	Effect Size (g)	95% CI	Studies/N
Zhai & Ma (2023)	Overall writing	0.86	[0.65, 1.07]	k=26; N=2,468
Fleckenstein et al. (2023)	Language quality	0.61	[0.26, 0.96]	k=84; N=2,828
Fleckenstein et al. (2023)	Content quality	0.57	[0.21, 0.93]	k=84; N=2,828
Fleckenstein et al. (2023)	Holistic quality	0.50	[0.16, 0.84]	k=84; N=2,828
Deng et al. (2024)	Academic performance	0.53	[0.41, 0.66]	k=69
Ngo et al. (2022)	AWE overall (EFL)	0.49	[0.27, 0.70]	k=23
Sun & Zhou (2024)	AI vs. Human	0.25	[-0.11, 0.60]	k=41; N=4,813

Fleckenstein et al. (2023), analyzing 20 studies ($k = 84$ effect sizes; $N = 2,828$), reported medium effects across all three outcome dimensions: language quality ($g = 0.61$, 95% CI [0.26, 0.96]), content quality ($g = 0.57$, 95% CI [0.21, 0.93]), and holistic quality ($g = 0.50$, 95% CI [0.16, 0.84]). Importantly, these effects reflect post-intervention performance rather than learning gains, as relatively few studies included delayed posttests.

The comparative meta-analysis by Sun and Zhou (2024) directly compared AI-generated feedback with human feedback across 41 studies ($N = 4,813$ participants). Results indicated no statistically significant difference between the two conditions ($g = 0.25$, 95% CI [-0.11, 0.60]). Although this suggests that AI feedback may be comparably effective to human feedback, the substantial heterogeneity ($I^2 = 95\%$) indicates considerable variability across studies, warranting caution in interpreting this aggregate finding.

Thematic Synthesis of Empirical Evidence

These findings are supported and extended by the 137 included empirical studies beyond the meta-analytic evidence. Randomized studies have consistently shown positive effects of AI feedback on surface-level writing characteristics (grammar, mechanics, vocabulary), with effect sizes ranging from small to moderate ($d = 0.20$ - 0.60). In higher-order writing skills (organization, argument, critical thinking), results were more varied: about 58% of the studies reported a significant improvement, while the rest did not find a significant difference compared to the control conditions. It is important to note that research on generative AI tools (e.g., ChatGPT) reported larger impacts on content-level edits than research on traditional AWE systems. However, this result is confounded by study recency, population, and design. Another pattern evident in the empirical literature is a preference for hybrid feedback

conditions (AI + human) over AI-only conditions, which is strong across study designs, tool types, and groups of learners.

Moderating Factors

Figure 4 illustrates the key moderating factors identified across the meta-analytic literature, organized into learner factors, feedback characteristics, intervention features, and contextual variables.

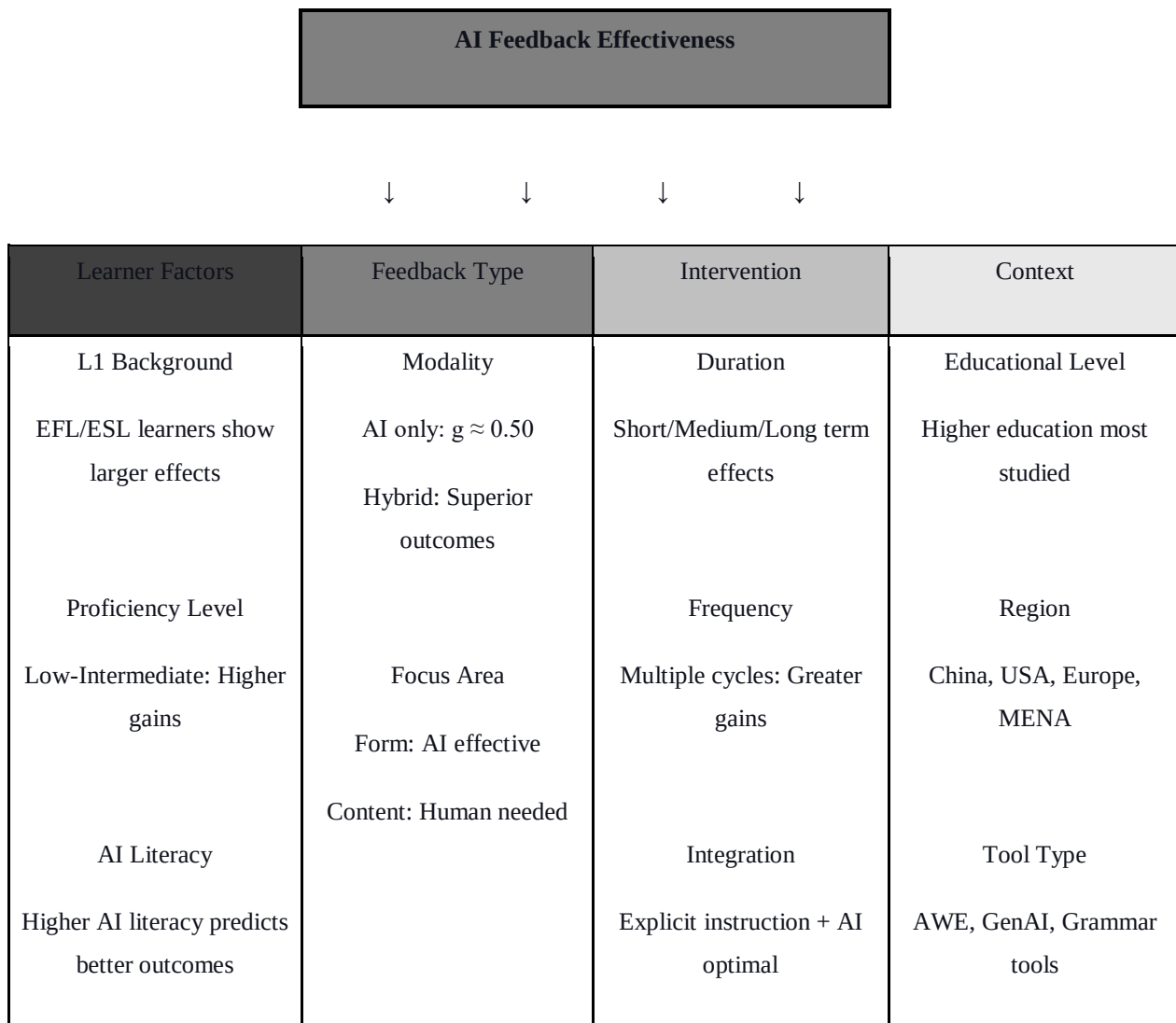


Figure 4. Moderating factors in effective AI feedback

Note 2. Moderators from meta-analytic evidence (Zhai & Ma, 2023; Fleckenstein et al., 2023; Sun & Zhou, 2024). **Figure 4.** Moderating Factors in AI Feedback Effectiveness. This framework organizes key moderators into four categories: learner factors (L1 background, proficiency level, AI literacy), feedback type (modality, focus area, integration approach), intervention characteristics (duration, frequency, tool type), and contextual factors (educational level, regional context). Evidence derives from meta-analytic findings by Zhai and Ma (2023), Fleckenstein et al. (2023), and Sun and Zhou (2024).

Learner Proficiency

Across several meta-analyses, L1 background and learner proficiency emerged as significant moderators. EFL/ESL learners demonstrated greater effects relative to native speakers (Zhai and Ma, 2023). Similarly, Ngo et al. (2022) reported an average effect of $g = 0.49$ (95% CI [0.27, 0.70]) for AWE interventions with EFL learners. Zhang (2025), in a randomized controlled trial involving 259 Chinese undergraduates, reported notable effects for organization ($d = 0.311$) and content development ($d = 0.191$).

Feedback Type and Hybrid Approaches

Hybrid feedback models integrating AI with human input have emerged as a promising approach. Liu et al. (2025) compared generative AI-only feedback with hybrid feedback (generative AI + human tutor) in a 12-week study of 60 Chinese EFL students. The hybrid group demonstrated significantly greater levels of intrinsic and extrinsic motivation, as well as improvements in perceptions of process-level and self-regulation feedback.

A similar hybrid approach was employed by Asadi et al. (2025), in which teachers used ChatGPT to generate initial feedback, which they then refined and delivered. This strategy produced positive outcomes across the dimensions of coherence, cohesion, grammar, and vocabulary. These findings are consistent with those of Link et al. (2022), who found that educators using AWE to supplement their feedback could focus on higher-order concerns, while AWE addressed mechanical issues.

Intervention Duration and Frequency

The meta-analysis by Brown et al. (2023) provided robust evidence of moderate corrective feedback effect sizes that remained relatively stable over time, with strong initial effect sizes that attenuated gradually. However, Zhai and Ma (2023) found that intervention duration was not a significant moderator of AWE effects. This discrepancy may reflect differences in how duration was operationalized, or fundamental differences between automated and human feedback in terms of the sustainability of effects.

Student Engagement and Perceptions

A large-scale cross-sectional survey ($N = 6,960$) conducted across four Australian universities revealed that approximately one in five students expressed interest in receiving generative AI feedback (Tai et al., 2025). Nevertheless, students consistently rated teacher feedback as more helpful and trustworthy than AI-generated feedback. Thematic analysis of 8,642 open-ended responses revealed nuanced orientations toward AI feedback: students appreciated its timeliness and availability but expressed concerns regarding potential inaccuracies and insufficient contextual understanding.

Yan's (2024) mixed-method multiple case study analyzed how L2 writers engage with ChatGPT for writing feedback. Results indicated that engagement operated across behavioral, cognitive, and affective dimensions. Notably, participants reported that the cognitive load associated with both feedback solicitation (prompt construction) and feedback processing was substantial, underscoring the importance of developing both AI literacy and writing competence.

Perception studies further illuminate both strengths and concerns. Polakova and Ivenz (2024) reported that participants identified immediacy (95.8%), repeatability (79.2%), and reduced anxiety (66.7%) as key benefits, while insufficient cultural awareness (83.3%) and inability to support creativity (70.8%) were cited as principal concerns. These findings suggest

that AI feedback may be particularly advantageous for form-based revision, whereas human feedback remains essential for meaning-making and culturally responsive instruction.

Feedback Literacy and Self-Regulated Learning

Zarrinabadi and Rezazadeh (2024) conducted intervention research on the development of feedback literacy through scaffolded writing instruction. The experimental group demonstrated broad improvements in feedback literacy, writing proficiency, and resilience following the intervention. Qualitative analysis revealed that students who became more feedback literate also developed more positive orientations toward feedback processes, attributable to enhanced opportunities for engagement, collaboration, and sustained motivation.

Research on self-regulated learning in AI-assisted writing remains in its nascent stages. Kim et al. (2025) employed structural equation modeling and demonstrated that AI literacy and self-regulated learning jointly influenced writing performance and digital well-being. Both constructs positively predicted writing performance, which in turn mediated their relationship with psychological outcomes. These findings indicate that effective AI implementation requires attention to both technical and metacognitive dimensions.

Discussion

This review synthesizes a substantial body of evidence on AI-mediated feedback in writing education, organized around the theoretical framework presented in Figure 3. Meta-analytic evidence consistently demonstrates beneficial effects of automated and AI feedback on writing quality, with effect sizes ranging from medium ($g = 0.50$ – 0.61) to large ($g = 0.86$) as illustrated in Figure 2. Notably, the comparative meta-analysis by Sun and Zhou (2024) suggests that AI feedback may yield learning effects comparable to those of human feedback; however, the substantial methodological heterogeneity precludes definitive conclusions.

Hypothesis H1, positing a significant positive effect of AI feedback on writing quality, receives strong empirical support. Hypothesis H2 (superior outcomes with hybrid approaches) is robustly supported, with multiple studies demonstrating enhanced outcomes when human input is combined with AI feedback (Asadi et al., 2025; Liu et al., 2025; Link et al., 2022), consistent with the hybrid model articulated in our theoretical framework. Hypothesis H3 regarding proficiency moderation is partially supported: more pronounced effects are observed among EFL/ESL learners in some but not all analyses.

Addressing Heterogeneity and Methodological Variability

One crucial finding across the synthesized meta-analyses is the high level of statistical heterogeneity. Fleckenstein et al. (2023) and Zhai and Ma (2023) also reported significant between-study variability, and I^2 by Sun and Zhou (2024) was 95%. This heterogeneity is probably indicative of actual variations in study population (EFL vs. ESL vs. native speakers), types of AI tools (traditional AWE vs. generative AI), and types of interventions (duration, integration strategy, instructor engagement), and operationalizations of outcomes (holistic ratings vs. analytic rubrics vs. automated scores). Although moderator analyses have established that certain sources of this variability, a lot of it remains unaccounted for, and aggregate effect sizes must thus be viewed as suggestive of an effect rather than conclusive. Standardized outcome measures and more comprehensive reporting of intervention parameters would make future primary research more useful for conducting more accurate meta-analytic syntheses.

Publication Bias Considerations

A threat of publication bias should be noted with concern. Despite funnel plot analyses in the included meta-analyses (Zhai & Ma, 2023; Fleckenstein et al., 2023) not showing significant asymmetry, the issue of publication bias should be considered for several reasons. First, the educational technology sphere is also highly results-driven, and null findings may have a low chance of publication or inclusion in the databases used. Second, there is a swift commercialization of AI writing tools, which can create potential conflicts of interest that may be expressed in subtle ways in research design and reporting. Third, the empirical literature is characterized by small-sample studies (mean $N = 78$), which make it more prone to inflated effect-size estimates. Instead, the reported effects should be viewed by the readers as plausible upper-bound estimates of true effects.

Theoretical Implications

From a sociocultural perspective, AI systems appear capable of providing scaffolding within the learner's zone of proximal development, though this scaffolding is qualitatively distinct from human mediation. Whereas human tutors dynamically adjust feedback based on real-time assessment of learner needs, AI systems provide more consistent but less responsive feedback that learners must interpret and apply independently. This places greater demands on learners' feedback literacy and self-regulatory capacities.

The feedback literacy framework assumes heightened significance in AI-mediated settings. Students must develop new competencies related to effective prompt construction, critical evaluation of AI output, and strategic integration of AI recommendations with their own rhetorical purposes. The superior performance of hybrid conditions over AI-only conditions suggests that human guidance remains essential for cultivating these higher-order capacities.

Ethical and Pedagogical Considerations

The use of generative AI in writing education raises significant ethical and pedagogical issues worth serious consideration. To begin with, excessive reliance on AI responses could undermine the development of independent evaluative judgment, a fundamental element of feedback literacy, by delegating the important role of evaluation to algorithms (Kosmyna et al., 2025). Second, generative AI solutions raise academic integrity concerns because the line between using AI to revise a piece of writing and creating a piece of writing is unclear in practice. Third, large language models, which are generally trained on English-language data in Western contexts, are prone to cultural and linguistic bias and can disadvantage learners with non-Western rhetorical backgrounds. Fourth, the problems of data privacy and algorithmic transparency are especially urgent in the education sector, where students may not be aware of how their writing data is analyzed and stored. Such ethical aspects are predominantly absent in the reviewed empirical literature and constitute a gaping void that needs to be addressed by further academic efforts.

The Dominance of Short-Term Interventions

One of the most prominent weaknesses of the existing evidence base is that most of the interventions are short-term. Sixty-eight-point six percent of empirical studies (as noted in Section 4.0) reported the use of interventions of 12 weeks or less, with some consisting of a single session or lasting 1 or 2 weeks. This throws down basic challenges to the sustainability of observed effects. The process of writing development has been generally perceived as a

recursive and gradual process (Flower & Hayes, 1981), and short-term increases in surface accuracy are not necessarily the long-term gains in writing competence, rhetorical sophistication, or self-regulatory capacity. There are only a limited number of longitudinal studies that show mixed results: some indicate long-term beneficial effects (Brown et al., 2023), whereas others show a decrease in effects over time. Longitudinal designs spanning at least 1 year of academic time are desperately in demand in the field to determine whether AI feedback yields sustained learning benefits or only temporary performance boosts.

Practical Implications

The findings yield several evidence-based recommendations for practitioners:

First, institutions should adopt hybrid feedback systems that leverage AI for immediate, form-focused feedback while reserving human expertise for higher-order concerns and personalized guidance.

Second, explicit instruction in AI literacy—including prompt construction, critical evaluation of AI output, and judicious integration of AI tools into the writing process—should be incorporated into writing curricula.

Third, curricula should incorporate activities that develop feedback literacy, engaging students in feedback evaluation, application, and integration with writing development.

Fourth, learner proficiency should inform the selection and deployment of AI feedback, recognizing that EFL/ESL students may derive particular benefit from AI-assisted language accuracy support.

Fifth, practitioners should remain cognizant of AI limitations, particularly regarding cultural sensitivity, contextual appropriateness, and support for creative, identity-expressive writing.

Limitations

This review has several limitations. First, the rapid pace of AI technological development means that research conducted on older systems may not generalize to current capabilities. Second, the majority of included studies assessed short-term outcomes, limiting conclusions about the durability of effects. Third, the predominance of EFL populations in higher education constrains generalizability to other learner populations. Fourth, meta-analyses may overestimate effect sizes due to publication bias. However, funnel plot analyses in the included meta-analyses did not reveal significant asymmetry; this possibility cannot be entirely excluded.

Fifth, adopting a narrative synthesis rather than conducting an original meta-analysis means this review cannot provide new pooled effect-size estimates. While this approach was justified by the heterogeneity of included studies and the availability of existing meta-analyses, it represents a methodological limitation. Sixth, our inclusion criteria restricted the review to higher education contexts, excluding potentially informative studies from K–12 and professional development settings.

Future Research Directions

This synthesis identifies several priorities for future research. Longitudinal studies are urgently needed to establish the durability of AI feedback effects and to examine potential consequences for learner autonomy over extended timeframes. Comparative studies of human-AI collaboration models should investigate which feedback functions are optimally served by AI versus human tutors. Research on the development of feedback literacy in AI-mediated settings would advance understanding of how students can be prepared to engage with AI

effectively. Finally, studies incorporating more diverse populations beyond university-level EFL students would enhance external validity.

Additionally, future research should attend to the ethical dimensions of AI-mediated feedback, including the effects of AI dependency on learner agency, the implications of algorithmic bias for diverse learner populations, and the development of institutional policies governing AI use in writing instruction.

Conclusion

This systematic review synthesizes the current state of research on AI-based feedback in writing instruction, integrating meta-analytic evidence from seven syntheses with thematic analysis of over 130 primary studies encompassing thousands of participants. Key conclusions include:

- (1) AI-based feedback produces medium to large positive effects on writing quality, with meta-analytic estimates ranging from $g = 0.50$ to $g = 0.86$ depending on outcome measure and study characteristics.
- (2) AI feedback appears comparably effective to human feedback, while hybrid approaches integrating both modalities have demonstrated superior results across multiple studies.
- (3) Key moderating variables include learner proficiency—with EFL/ESL learners showing particularly strong benefits—and intervention features such as feedback type and pedagogical integration.
- (4) Effective AI feedback implementation requires concurrent attention to feedback literacy, AI literacy, and the integration of self-regulated learning with writing skill development.
- (5) Future research should prioritize longitudinal designs, diverse populations, and investigation of optimal human-AI collaboration models.

As AI technologies continue to evolve at an unprecedented pace, the evidence synthesized in this review provides a foundation for evidence-based recommendations on integrating AI feedback into writing pedagogy. The goal should not be to replace human feedback, but rather to harness AI's capabilities to enhance learning while preserving the essential human dimensions of writing education.

About the Author

Taha Bashier Mohammed is an Assistant Professor of Applied Linguistics at Qassim University, Saudi Arabia. He holds a PhD in Applied Linguistics and has over ten years of academic experience. His research focuses on discourse analysis, pragmatics, language pedagogy, and second language acquisition, with publications in reputable, peer-reviewed international journals. 0009-0005-2596-4759

AI Statement

This document has been enhanced through the use of Grammarly which was employed to refine its linguistic style and correct grammar and spelling. While the incorporation of these technologies may introduce some AI-generated linguistic patterns, it is important to note that the core intellectual content, data interpretation, and conclusions presented are entirely the work of the author.

Statement of Absence of Conflict of Interest

The author declares that there are no conflicts of interest related to the research, findings, or recommendations presented in this paper. All conclusions drawn are independent and unbiased.

References

- Appelbaum, M. et al. (2018). Journal article reporting standards for quantitative research in psychology. *American Psychologist*, 73(1), 3–25.
- Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Learning and Individual Differences*, 117, 102588.
- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), 23.
- Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, 29(5), 5867–5891.
- Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing* (pp. 142–173). Routledge.
- Brown, D., Liu, Q., & Norouzian, R. (2023). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. *Language Learning*, 73(2), 415–449.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325.
- Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150–163.
- Cavalcanti, A. P. et al. (2023). Automatic feedback in online learning environments: A systematic literature review. *Computers & Education: Artificial Intelligence*, 2, 100027.
- Dawson, P. et al. (2024). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, 44(1), 25–36.
- de Kleijn, R. A. M. (2023). Supporting student and teacher feedback literacy: An instructional model for student feedback processes. *Assessment & Evaluation in Higher Education*, 48(2), 186–200.
- Deng, R., Benckendorff, P., & Gannaway, D. (2024). A meta-analysis on the effectiveness of computer-assisted technology on academic performance and engagement in higher education. *Computers & Education*, 69, 104187.
- Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion. *Education and Information Technologies*, 29, 14151–14203.
- Fleckenstein, J. et al. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, 1162454.
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387.
- Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, 37(4), 1–43.
- Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.

- Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal*, 99(1), 1–18.
- Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- Kim, D., Park, Y., Yoon, M., & Jo, I. H. (2024). Exploring cognitive, effective, and behavioral engagement patterns of self-regulated learners using data mining. *Computers in Human Behavior*, 130, 107191.
- Kong, S., & Teng, L. (2024). Written corrective feedback in second language writing: A synthesis of naturalistic classroom studies. *Language Teaching*, 57(4), 449–470.
- Kosmyna, N. et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks. arXiv preprint arXiv:2506.08872.
- Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learning*, 28(3), 82–104.
- Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4), 605–634.
- Little, T., Dawson, P., Boud, D., & Tai, J. (2023). Can students' feedback literacy be improved? A scoping review of interventions. *Assessment & Evaluation in Higher Education*, 48(7), 1084–1097.
- Liu, X., Chen, Y., & Wang, Z. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Language Learning Journal*, online first.
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11, 9.
- Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education*, 45(4), 527–540.
- Ng, D. T. K. et al. (2024). Empowering student self-regulated learning and science education through ChatGPT. *British Journal of Educational Technology*, 55(3), 1024–1041.
- Ngo, T. B., Chen, B. H., & Tai, H. Y. (2022). Effectiveness of AWE feedback on EFL/ESL writing quality: A meta-analysis. *Interactive Learning Environments*, 30(9), 1694–1707.
- Nieminen, J. H., & Carless, D. (2023). Feedback literacy: A critical review of an emerging concept. *Higher Education*, 85, 1381–1400.
- Page, M. J. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372(71).
- Parra, R., & Castrillo, M. D. (2025). The impact of using ChatGPT for feedback in EFL writing: A systematic review. *Revista Internacional de Lenguas Extranjeras*, 23, 95–119.
- Polakova, P., & Ivenz, R. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1), 2410101.
- Shen, H., & Bai, L. (2024). Modeling AI-assisted writing: How self-regulated learning influences writing outcomes. *Computers in Human Behavior*, 160, 108377.
- Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 36(1), 1–23.

- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., & Wang, J. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Sun, L., & Zhou, L. (2024). Does generative artificial intelligence improve the academic achievement of college students? A meta-analysis. *Educational Psychology Review*, 36, 97.
- Tai, J., Bearman, M., Boud, D., & Dawson, P. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, online first.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Winstone, N. E., Boud, D., Dawson, P., & Heron, M. (2022). From feedback-as-information to feedback-as-process: A linguistic analysis of feedback literature. *Assessment & Evaluation in Higher Education*, 47(2), 213–230.
- Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087.
- Yan, D. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT. *Humanities and Social Sciences Communications*, 11, 1086.
- Yan, L., Martinez-Maldonado, R., & Gašević, D. (2024). Practical and ethical challenges of large language models in education. *British Journal of Educational Technology*, 55(1), 90–112.
- Zarrinabadi, N., & Rezazadeh, M. (2024). Developing feedback literacy, scaffolded writing, and resilience through intervention on feedback processes. *Assessing Writing*, 60, 100841.
- Zhang, X. (2025). Enhancing critical writing through AI feedback: A randomized control study. *Behavioral Sciences*, 15(5), 705.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41(2), 64–70.
- Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.). Lawrence Erlbaum.