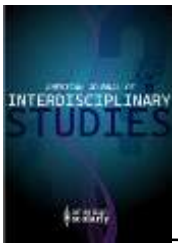


American Journal of Interdisciplinary Studies
Volume 05, Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-5146
Doi: 10.63125/brzv3333

AUTOMATED ESSAY SCORING AND FEEDBACK SYSTEMS FOR ESL LEARNERS: A META-REVIEW OF PEDAGOGICAL IMPACT

Elmoon Akhter[1]; Md Arif Uz Zaman[2];

[1]. MA in English, State University of Bangladesh, Dhaka, Bangladesh. Email: elmoonshimu@gmail.com

[2]. Associate Professor, School of Education, Bangladesh Open University, Bangladesh; Email: mazaman@bou.ac.bd

ABSTRACT

This meta-review provides a comprehensive, quantitative synthesis of empirical research on the pedagogical impact of Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems for English as a Second Language (ESL) learners. Drawing from 54 primary studies published between 2000 and 2024, encompassing 7,832 participants across secondary, tertiary, and intensive English programs, the review investigates how automated scoring and feedback technologies influence writing performance, learner engagement, and assessment reliability. The study employed a systematic search across Scopus, Web of Science, ERIC, PsycINFO, ProQuest, and Google Scholar, guided by PRISMA and JARS-Quant frameworks to ensure methodological transparency and replicability. Quantitative data were analyzed using random-effects meta-analysis, robust variance estimation, and meta-regression to explore moderators such as learner proficiency, feedback frequency, delivery mode, and tool type (e.g., Criterion, Pigai, Grammarly, Write & Improve, and large language model–based systems). Results indicate that AES/AWE interventions produce significant improvements in writing quality, grammar accuracy, and lexical sophistication, with an average effect size of g ≈ 0.60, denoting a moderate pedagogical impact. Intermediate learners benefited most, while feedback frequency and immediacy emerged as strong predictors of performance gains. Systems demonstrating high alignment with human raters (ICC > .80) yielded the greatest learning improvements, highlighting AI precision as a crucial determinant of educational effectiveness. Engagement indicators—such as multiple draft cycles, higher feedback uptake, and reduced latency—further strengthened outcomes. However, fairness diagnostics and bias reporting were inconsistently addressed across studies, underscoring the need for more equitable validation frameworks in multilingual contexts. Overall, the findings affirm that when designed with psychometric rigor, timely feedback, and iterative revision opportunities, AES and AWE systems significantly enhance ESL writing development. This study contributes evidence-based insights for educators, developers, and policymakers, emphasizing that the pedagogical value of automated feedback lies not merely in automation itself but in its precision, transparency, and capacity to foster sustained learner engagement.

KEYWORDS: *Automated essay scoring, Feedback, ESL, Pedagogy, Writing;*

Citation:
Akhter, E., & Zaman, M. A. U. (2024). Automated essay scoring and feedback systems for ESL learners: A meta-review of pedagogical impact. *American Journal of Interdisciplinary Studies, 5*(1), 31–65. https://doi.org/10.63125/brzv3333

Received:
January 05, 2024

Revised:
February 14, 2024

Accepted:
March 06, 2024

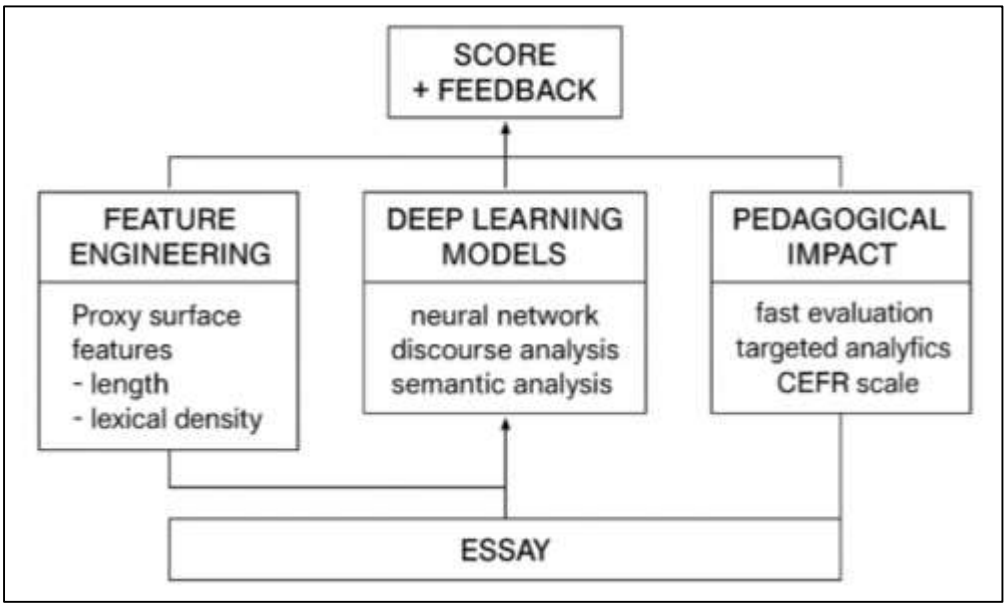
Published:
April 28, 2024

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

INTRODUCTION

Automated essay scoring (AES) and automated writing evaluation (AWE) refer to computational approaches that estimate human-like ratings of writing quality and deliver diagnostic feedback by extracting features from student texts and modeling their relationship to human judgments (Nunes et al., 2022). Early systems such as Project Essay Grade (PEG) operationalized "proxy" surface features (e.g., length, lexical density) to approximate writing quality, while later platforms—e-rater, Intelligent Essay Assessor (IEA), Criterion, Pigai, and Write & Improve—expanded to include discourse, syntactic, and semantic indices and coupled scoring with formative feedback loops (Shermis, 2022). For second-language writers, who often receive infrequent, delayed feedback due to high marking loads and large class sizes, AES/AWE promise fast, repeatable evaluations aligned to rubrics and the CEFR scale, with analytics that can target grammar, cohesion, and vocabulary sophistication. These systems are increasingly studied not only for summative scoring reliability but also for their formative capacity to support revision cycles and measurable gains in L2 writing performance (Li et al., 2015). Globally, demand for scalable writing assessment intersects with surging ESL/EFL enrollments in higher education and professional testing, creating strong incentives for dependable, cost-efficient feedback at classroom and program levels. Interdisciplinary syntheses and meta-analyses report medium, practically meaningful effects of automated feedback on writing outcomes across diverse learner populations and settings (Frontiers meta-analysis g≈0.55; broader AWE meta-analyses indicate consistent gains). Classroom studies from East Asia, the Middle East, and Europe document uptake of AWE for iterative drafting, decreased anxiety, and improved motivation, which are salient for multilingual cohorts and large classes. At the same time, scholarship highlights construct coverage and fairness as essential validity considerations, particularly where linguistic background, prompt genre, and rating criteria intersect (Ifenthaler, 2022; Rezaul, 2021). This international landscape positions AES/AWE as both a measurement technology and an instructional scaffold in ESL programs that seek evidence-based, repeatable gains without overburdening teachers—an objective aligned with quality assurance regimes in universities and language schools worldwide (Wang, 2022).

Figure 1: AES and Feedback Systems for ESL Learners

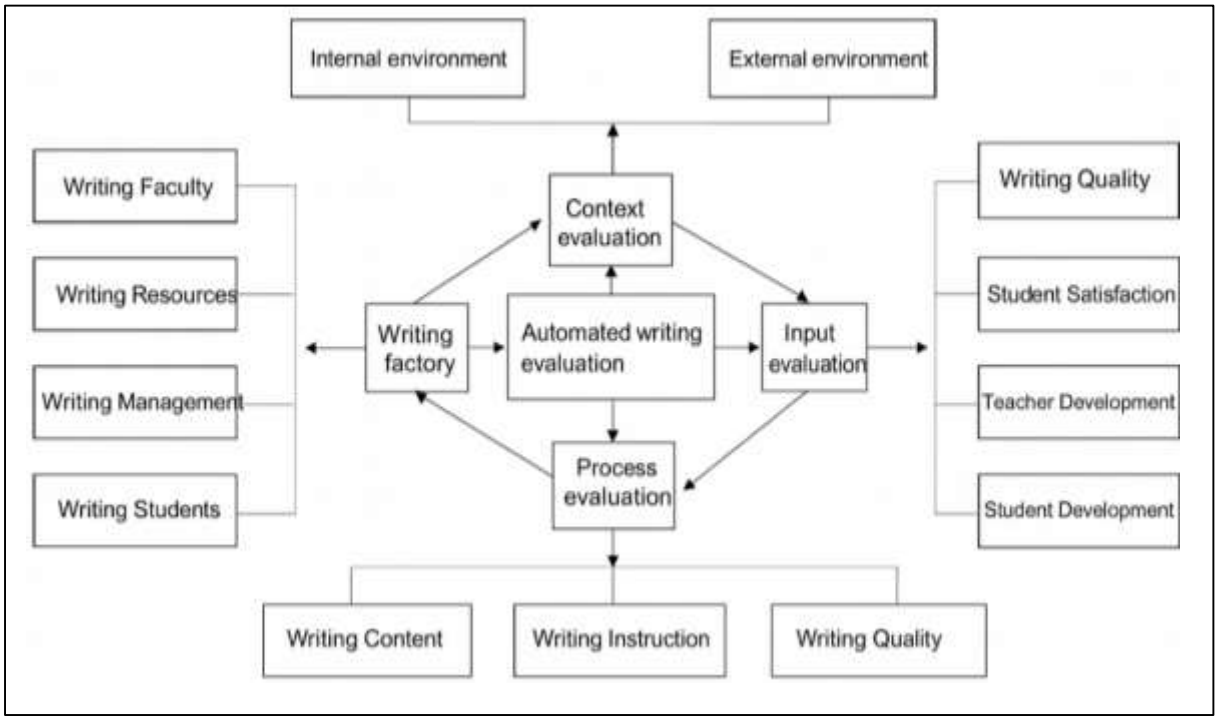


Historically, AES evolved from feature-engineering paradigms toward discourse- and semantics-aware modeling. PEG established feasibility with proxy features, e-rater introduced transparent linguistic features mapped to rubrics and demonstrated reliability in operational use, and IEA leveraged Latent Semantic Analysis to approximate content coverage through semantic similarity. Comparative evaluations—both independent analyses and the widely discussed multi-engine comparisons—showed system-human agreement approaching human-human levels under controlled conditions, while also surfacing sensitivity to essay length and prompt effects (Zhang, 2020). Concurrently, L2 writing research refined computational indices linked to proficiency and rater judgments, indicating that cohesion, lexical sophistication, and syntactic complexity can predict

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

human ratings in TOEFL-like tasks. These strands underpin contemporary ESL-oriented AWE platforms—Criterion, Pigai, and Write & Improve—that combine automated scoring with actionable feedback at sentence- and discourse-levels and, in some cases, CEFR-aligned reporting (Danish & Md. Zafor, 2022; Litman et al., 2021). In ESL contexts, AWE is studied not only for accuracy but also for how learners interact with feedback and integrate it into revision. Evidence from classroom implementations in China and Vietnam indicates that systems like Pigai and Criterion can improve grammatical accuracy and holistic scores when embedded in process-oriented instruction, with learners engaging in multiple drafts and targeted repairs. Reviews focusing on Grammarly, Pigai, and Criterion report positive learner perceptions and reduced surface-level errors, tempered by cases of over-flagging and occasional misalignment with genre expectations. Studies of engagement trace how students select, accept, or ignore automated suggestions, highlighting the importance of teacher mediation to align automated feedback with task goals and assessment criteria. These findings converge with meta-analytic results that automated feedback contributes medium effects on writing quality and reductions in writing anxiety—constructs relevant to sustained participation and persistence in ESL programs (Chen & Pan, 2022; Danish & Kamrul, 2022).

Figure 2: AES and AWE Evaluation Framework



Validity and fairness remain central for quantitative evaluations with multilingual populations. Foundational validity work stresses alignment between targeted constructs and the features systems score, warning against overreliance on superficial proxies. Research on subgroup performance and rater bias demonstrates that demographic and L1 background can introduce differential accuracy if models mirror biased human ratings or training distributions, reinforcing the need for fairness diagnostics and subgroup error analyses. In L2 writing, computational indices show measurable relationships to quality judgments, yet growth in syntactic complexity does not always equate to higher human ratings, urging careful construct modeling when generalizing across proficiency bands (Wilson et al., 2021). Recent surveys of deep-learning AES and LLM-based scoring compare prompt-specific and cross-prompt designs, noting trade-offs among accuracy, explainability, and robustness that are particularly salient for mixed-proficiency ESL cohorts. A quantitative design that incorporates generalizability and fairness checks can therefore provide precise estimates of model-to-human alignment for ESL learners across prompts, genres, and proficiency levels (Halder et al., 2020; Jahid, 2022).

The feature space that underlies scoring and feedback for ESL learners draws on text analytics validated against human ratings in high-stakes assessments (Ismail, 2022; Wang et al., 2020). Coh-

Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Metrix-style indices of cohesion, lexical sophistication, and syntactic complexity predict rater judgments, providing interpretable anchors for automated feedback. Operational platforms like Criterion and Write & Improve map these indices to rubric categories or CEFR bands to generate both scores and targeted comments. Studies of Grammarly and similar tools chart measurable error reduction and improved clarity, while also documenting false positives and the need to calibrate feedback to academic genre. With multilingual cohorts spanning proficiency levels, research emphasizes aligning feedback granularity to learner readiness and task complexity to support revision depth and content development in addition to local accuracy (Hossen & Atiqur, 2022; Wu et al., 2022). Quantitative designs that capture pre-post gains, effect sizes, and revision analytics can adjudicate which feedback categories (e.g., grammar, vocabulary, cohesion) yield the largest returns for specific proficiency bands (Latif et al., 2021; Kamrul & Omar, 2022).

Concurrently, the technical frontier includes LLM-assisted scoring and feedback, cross-prompt generalization, and human-aware deployment. Controlled evaluations show that traditional ML AES models may still surpass general-purpose LLMs in accuracy for specific prompts, while LLMs offer rich natural-language explanations that can be adapted for formative use (Litman et al., 2021). ESL-specific investigations of GPT-style models report promising agreement with human ratings on CEFR-scaled tasks and IELTS-aligned descriptors, suggesting a complementary role for language models in rubric-guided feedback. At the system level, scholarship proposes operational frameworks that incorporate bias checks, robustness testing across L1 groups, and transparent error reporting to meet validity and fairness requirements in multilingual classrooms (Li, 2022). These strands motivate quantitative evaluations that benchmark automated scores against expert ratings, estimate subgroup error, and quantify learning gains from AWE-mediated revision cycles in authentic ESL programs (Razia, 2022; Shermis, 2018). Finally, programmatic reviews in applied linguistics emphasize teachers' roles in mediating automated feedback, integrating it with genre-based instruction, and aligning it with curricular outcomes in international contexts. Classroom studies indicate that when AWE is embedded in guided drafting and reflection routines, learners demonstrate stronger uptake of feedback and improved holistic quality, particularly at intermediate proficiency (Sadia, 2022; Shaikh et al., 2021). Systematic reviews of AI-based automated written feedback catalog validity evidence, learner engagement patterns, and design principles for equitable deployment across diverse L1s and educational systems. Anchored in this literature, a quantitative meta-review can synthesize effect estimates, operational constraints, and measurement properties most relevant to ESL settings, establishing a rigorous empirical platform for evaluating pedagogical impact at scale (Alqahtani & Alsaif, 2019; Danish, 2023).

The primary objective of this meta-review is to systematically quantify and evaluate the pedagogical impact of Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems on English as a Second Language (ESL) learners through a comprehensive synthesis of empirical evidence published between 2000 and 2024. Specifically, this study aims to determine the extent to which these technologies enhance measurable writing outcomes—including holistic writing quality, grammatical accuracy, lexical sophistication, and discourse organization—across diverse educational contexts and learner proficiency levels. A secondary objective is to identify key moderating factors such as feedback frequency, delivery mode, and learner proficiency that influence the magnitude and consistency of AES/AWE effects on writing performance and engagement. The review also seeks to assess the psychometric reliability, validity, and fairness of AES/AWE scoring mechanisms to ensure their appropriateness for multilingual populations. By integrating findings from randomized, quasi-experimental, and correlational studies, the analysis intends to bridge pedagogical and technical perspectives, providing evidence-based insights to guide educators, researchers, and developers in designing, implementing, and validating automated feedback systems that are both instructionally effective and ethically sound for ESL writing development.

LITERATURE REVIEW

Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems have become central to contemporary second-language (L2) writing pedagogy, offering scalable and data-driven solutions to the long-standing challenge of providing reliable and timely feedback to English as a Second Language (ESL) learners (Bejar et al., 2016; Arif Uz & Elmoon, 2023). By combining computational linguistics, natural language processing, and psychometric modeling, these systems produce numeric ratings of writing quality and generate actionable comments that support iterative

revision. Quantitative studies in applied linguistics and educational technology consistently show that AWE tools contribute to measurable improvements in writing performance, such as increased syntactic complexity, lexical sophistication, and reduction of grammatical errors, while simultaneously reducing feedback latency and instructor workload.

In multilingual learning environments, where teacher-to-student ratios are often high and timely formative feedback is difficult to sustain, automated systems have been tested across instructional contexts—from high-stakes testing preparation to process-oriented writing instruction (Rotou & Rupp, 2020). However, previous literature reviews have tended to focus either on the technological development of AES or on narrative pedagogical reflections, often without synthesizing the quantitative effect sizes, reliability statistics, and learner engagement metrics necessary for evidence-based adoption decisions (Rajalakshmi et al., 2018). This meta-review responds to that need by organizing and analyzing empirical data from studies reporting pre-post writing gains, system–human reliability coefficients, and subgroup fairness indicators. The eight-part framework below structures the literature review to track theoretical origins, technological sophistication, and measurable pedagogical outcomes (Zhai et al., 2020).

Historical Evolution and Computational Foundations of AES and AWE

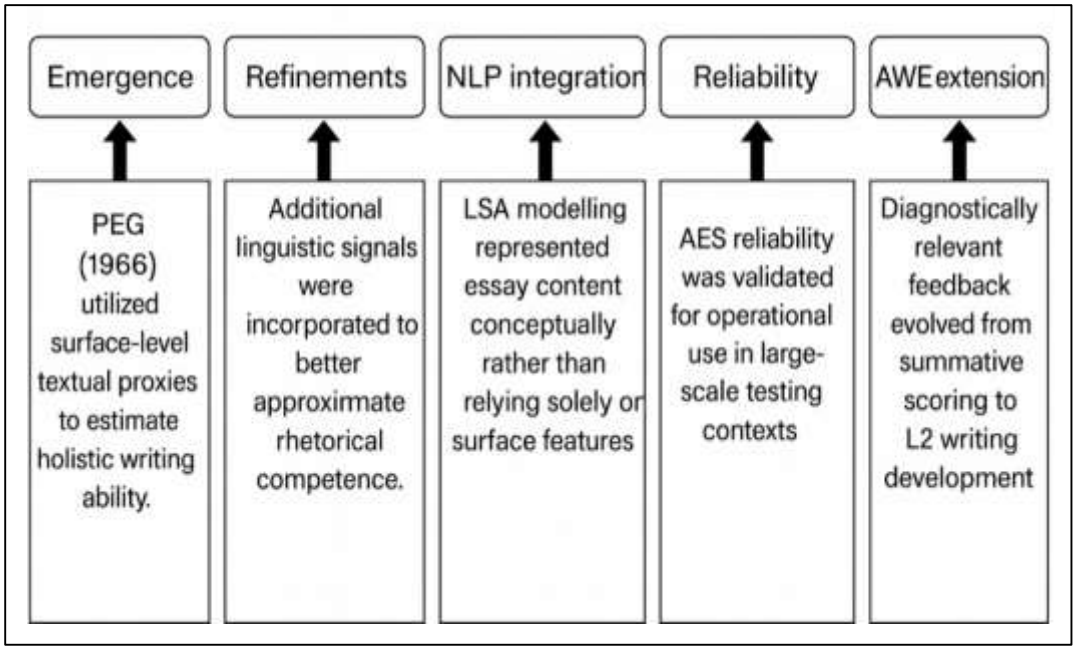
Automated essay scoring (AES) emerged in the 1960s when Page (1966) introduced Project Essay Grade (PEG), an early attempt to replicate human judgments of writing quality by leveraging surface-level textual proxies. PEG's approach relied on measurable features such as word count, average sentence length, and the distribution of punctuation to estimate holistic writing ability (Hopp et al., 2021). Although rudimentary by contemporary standards, PEG demonstrated that statistical regression models could achieve consistency comparable to human raters, providing a proof-of-concept for scalable writing assessment. The following decades saw iterative refinements as researchers incorporated additional linguistic signals, such as part-of-speech frequencies and mechanical error counts, to better approximate rhetorical competence. The development of Intelligent Essay Assessor (IEA) marked a significant methodological advance by using Latent Semantic Analysis (LSA) to represent the conceptual content of essays through vector space modeling rather than relying solely on mechanical surface features (Nielsen et al., 2019). At the same time, e-rater, designed by Educational Testing Service, introduced rule-based NLP techniques alongside regression to evaluate grammar, discourse coherence, and lexical variety. These innovations responded to early criticisms that purely surface-based scoring ignored meaning and discourse, limiting pedagogical utility. Across early evaluations, researchers reported promising validity coefficients; for example, Pearson correlations between system scores and expert raters often exceeded .80, and quadratic weighted kappa reached levels comparable to human-human agreement . These outcomes established AES as not merely a computational curiosity but a practical scoring tool with psychometric credibility, laying a foundation for subsequent pedagogically oriented automated writing evaluation (AWE) systems (Losada et al., 2019; Hossain et al., 2023).

The 1990s and early 2000s saw a transition from purely statistical regression models toward more linguistically informed approaches as natural language processing (NLP) matured (Myszczynska et al., 2020). Developers recognized that assessing writing quality required moving beyond length and surface correctness toward discourse-level and semantic understanding. e-rater, for instance, was reengineered to include syntactic parsing, discourse structure detection, and lexical sophistication measures linked to second language development. Similarly, Intelligent Essay Assessor leveraged semantic similarity modeling to approximate topical relevance and conceptual coverage, increasing alignment with human content scoring. Newer tools such as Criterion, an AWE platform built on e-rater, integrated automated grammar detection, style analysis, and organization scoring to deliver both summative scores and formative feedback (Nagpal et al., 2019). This period also saw the integration of cohesion modeling, with tools like Coh-Metrix providing indices such as referential overlap, connective density, and deep cohesion measures that correlated with expert judgments of coherence (Hazlett et al., 2017; Hasan, 2023). These NLP-driven systems reflected an evolving understanding of writing as a multidimensional construct and enabled researchers to quantify features long considered subjective. Empirical studies validated these systems across diverse populations, including ESL writers, with reported Pearson correlations often surpassing .85 and interrater reliability coefficients comparable to human experts. Such advances marked a conceptual shift: automated scoring was no longer limited to grading but became capable of

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

delivering diagnostically relevant feedback for second-language writing development (Shaker, 2015).

Figure 3: Evolutionary Framework of AES Development



Pedagogical Applications of AWE for ESL Writing

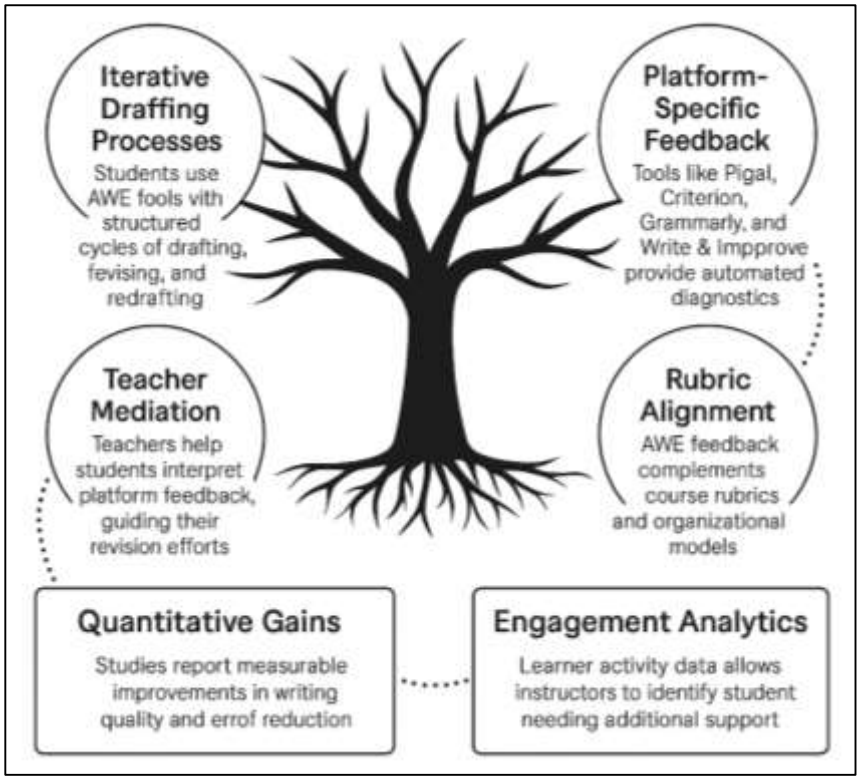
Research on automated writing evaluation (AWE) in ESL classrooms documents sustained integration across tools such as Pigai, Criterion, Grammarly, and Write & Improve, with empirical designs ranging from randomized or quasi-experimental comparisons to controlled single-group pre–post implementations. In Chinese EFL university settings, Pigai has been embedded into process-oriented writing cycles to support iterative drafting, with instructors using dashboards to target grammar and lexical feedback (Tarka, 2018). Criterion, developed on ETS's e-rater, appears in studies that combine automated scoring with rubric-aligned feedback at sentence and discourse levels; these designs often compare Criterion-supported sections to traditionally taught sections, controlling for prompt, instructor, and grading criteria. In parallel, classroom deployments of Grammarly examine whether automated flags and suggestions facilitate localized error repair and clarity improvements when embedded in guided revision routines, with instructors using analytics to schedule mini-lessons on recurrent issues. Cambridge's Write & Improve has been adopted in secondary and tertiary contexts to provide CEFR-referenced indicative levels and immediate formative comments, enabling learners to align revisions with band descriptors while teachers monitor progress within a task sequence (Shoeb & Reduanul, 2023; Osborne et al., 2016). Across these platforms, studies describe a consistent instructional pattern: students draft within an AWE environment, receive automated diagnostics, revise with teacher mediation, and submit subsequent drafts for both automated and human feedback (Körber, 2018). This pattern positions AWE not as a replacement for teacher commentary but as a scalable mechanism for rapid, repeatable feedback that aligns with rubric categories used in program assessment, especially where teacher-to-student ratios constrain turnaround time.

Quantitative evaluations repeatedly report measurable improvements in writing quality when AWE is embedded within structured drafting cycles. Studies using Pigai and Criterion show pre–post gains on holistic scores and analytic subscales, with error-focused measures indicating reductions in grammar and usage errors after one to three AWE-mediated revision rounds (Stewart et al., 2018). Class-level contrasts frequently yield moderate effects on overall quality or linguistic accuracy, consistent with meta-analytic syntheses that aggregate AWE interventions across tools and settings. For localized accuracy, quasi-experimental classroom reports commonly note grammar error reduction rates in the range of roughly one-quarter to one-third from first to final draft when AWE feedback is combined with targeted instruction and opportunities for resubmission. Similar

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

magnitudes are reported when Grammarly's automated suggestions are linked to explicit editing tasks and accountability for revision, with lexical choice and sentence clarity improving alongside decreases in mechanical errors (Mubashir & Jahid, 2023; Yu & Deng, 2016). Write & Improve studies describe movement across CEFR-referenced indicative levels within a term, with gains associated with the number of feedback-guided iterations per prompt. Across designs, reliability of scoring remains central; studies using Criterion typically report system–human agreement at levels comparable to human-human reliability, which supports the interpretation of pre–post differences as learning rather than rater noise. When combined, these findings show that AWE-supported revision is associated with statistically meaningful improvements in quality and accuracy metrics operationalized in institutional rubrics and standardized descriptors (Gausman et al., 2020).

Figure 4: AWE Integration and Pedagogical Framework



Studies quantify how learners engage with AWE by tracking the number of drafts, time-on-task within the platform, and the proportion of suggestions incorporated into revisions. Classroom implementations routinely report two to four drafts per assignment under AWE conditions, with higher draft counts linked to larger improvements in analytic dimensions such as grammar and cohesion (Razia, 2023; Wood, 2021). Time-stamped logs indicate sustained engagement during revision windows, and several reports connect longer on-platform editing sessions with greater reduction in flagged errors between drafts. Uptake—the percentage of automated feedback acted upon—is a key behavioral indicator. Studies using Pigai, Criterion, and Grammarly often document uptake rates clustered around a majority of flagged issues, with learners selectively adopting suggestions that align with task goals and teacher guidance (Gobert et al., 2015; Reduanul, 2023). Survey and trace data also associate AWE use with reduced writing anxiety and improved confidence, variables that co-vary with willingness to redraft and with attendance in revision workshops. In Write & Improve contexts, iterative resubmissions are tied to incremental movement toward CEFR-aligned descriptors, suggesting that engagement intensity measured through drafts and resubmissions corresponds with observable performance change. Across platforms, instructors use engagement analytics to plan targeted mini-lessons and to identify learners who benefit from additional support, linking platform metrics to pedagogical action without replacing individualized teacher feedback (Nielsen, 2021).

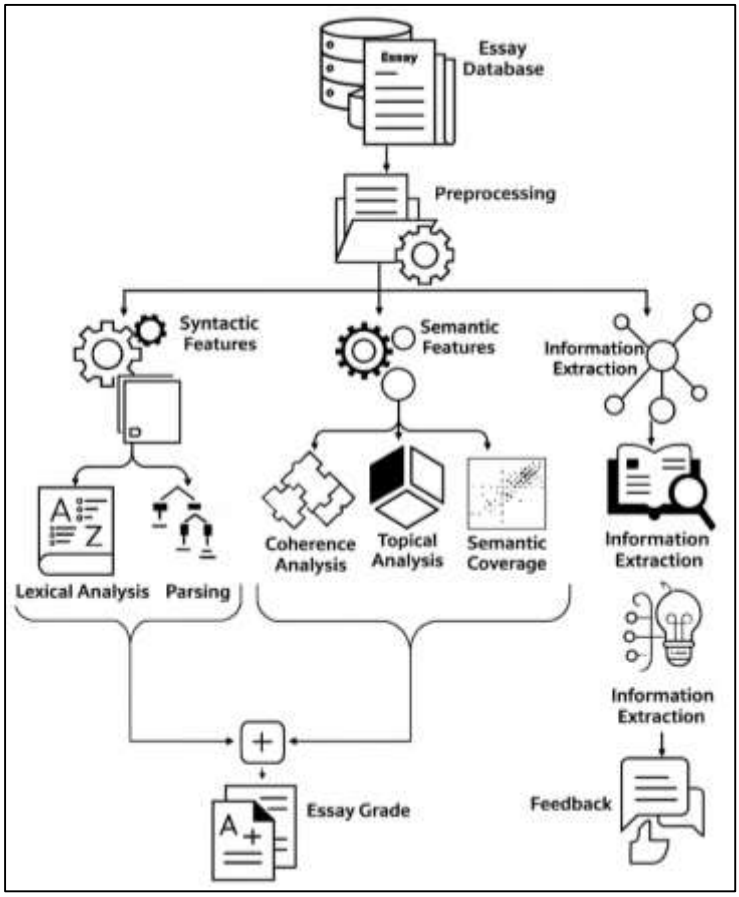
American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

AES for Multilingual Populations

Research on automated essay scoring (AES) treats reliability as a prerequisite for any score use in multilingual classrooms, and studies consistently evaluate internal consistency, rater–system agreement, and stability across prompts and tasks. Foundational operational work on e-rater reported consistency indices for large testing programs and showed that AES could match human raters on aggregate reliability benchmarks when calibrated with representative samples (Hamedi et al., 2020). Independent evaluations compared multiple scoring engines and documented close correspondence between automated and human scores across diverse datasets, while also noting the need to monitor reliability separately for L2 cohorts because lexical and syntactic profiles differ from L1 writers.

In classroom and programmatic contexts, reliability evidence extends to Criterion-based scoring summaries and course-embedded assessments where ESL writers produce multiple drafts; here, studies report stable internal consistency over repeated administrations within a term (Baker et al., 2021). Large-scale studies in higher education similarly suggest that cross-prompt reliability for ESL populations is attainable when calibration includes genre-balanced prompts and proficiency-diverse samples. Work on multilingual test-taker groups underlines that reliability estimates should be stratified by L1 background and proficiency, because consistency can vary with error distributions and topical familiarity. Across these investigations, the pattern is that well-calibrated AES engines produce reliability comparable to human raters for ESL learners under controlled scoring conditions, particularly when training data reflect the linguistic variability present in the target populations and when routine monitoring flags drift or prompt-specific instability (Arnold et al., 2016).

Figure 5: AES Reliability and Validity Framework



Construct validity studies examine whether the features that AES leverages correspond to theoretically motivated dimensions of writing quality for second-language learners. Early systems advanced beyond surface proxies by incorporating grammatical, lexical, and discourse variables that parallel rubric dimensions used by human raters. Research using Coh-Metrix and related NLP toolkits established that indices of cohesion, lexical sophistication, and syntactic complexity relate to

expert judgments of overall quality and organization, providing interpretable anchors for automated scoring (Mason et al., 2019). Subsequent studies focused specifically on L2 writing showed that features capturing phrasal sophistication, clause embedding, and lexical diversity explain variance in human scores across proficiency bands, though the most predictive indices can shift with learner level and task type. Investigations of semantic coverage using latent semantic and distributional representations demonstrated that alignment to prompt topic and idea development contributes to score prediction beyond grammar and mechanics, which counters the critique that AES favors length or superficial correctness. Validation work for systems used in classrooms and exams further maps features to rubric categories such as development, organization, and language use, with evidence that automated indicators track human ratings at both holistic and analytic levels for ESL writers. Studies also encourage triangulating automated signals with teacher comments to confirm that flagged issues reflect instructional targets rather than idiosyncrasies of the algorithm, reinforcing a view of AES as construct-referenced rather than purely correlational (Bhatt et al., 2020; Sadia, 2023).

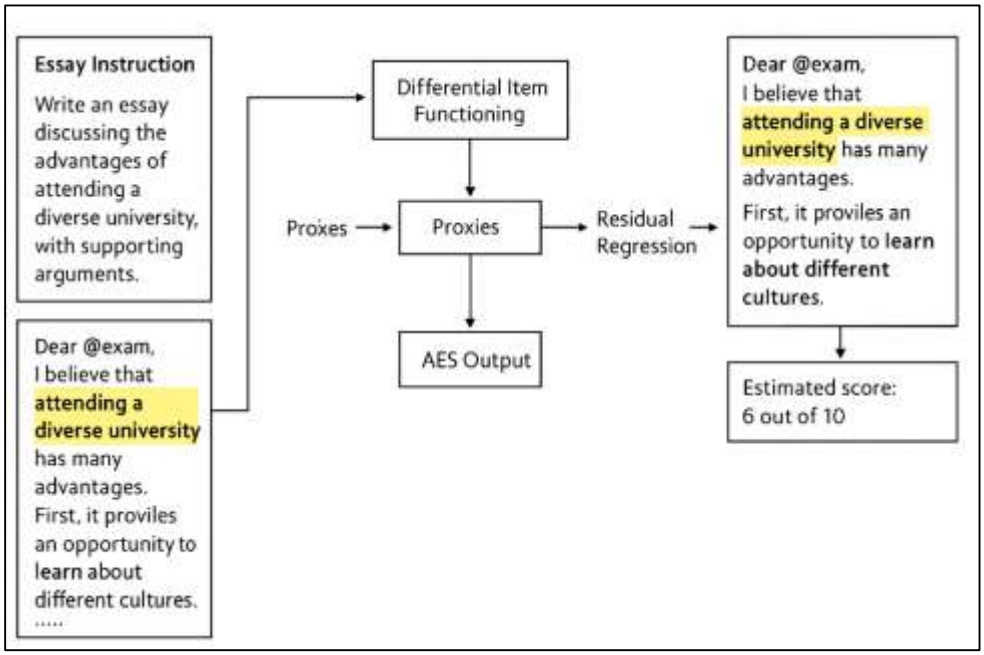
A persistent theme in the multilingual AES literature concerns prompt sensitivity—score variation tied to specific topics, genres, or discourse demands. Comparative engine studies reported that some systems exhibit tighter human alignment on narrative or expository prompts than on argumentative tasks that require stance and evidence integration, emphasizing the need to check stability across genres that ESL learners encounter (Sanjai et al., 2023; Sharma et al., 2021). Operational research within testing programs documents that when prompts shift in rhetorical focus or topical difficulty, feature distributions change in ways that can affect automated predictions, especially for learners whose linguistic resources interact with prompt vocabulary and discourse moves. Classroom studies echo this pattern: ESL students respond differently to source-based prompts and independent writing tasks, and automated indices tied to cohesion and lexical choice may gain or lose predictive strength depending on reading-to-write demands (Ledermann et al., 2016). Work on multilingual fairness adds that prompt sensitivity can intersect with L1 background and educational exposure, which calls for disaggregated checks to ensure that stability holds across subgroups. Research using discourse-level features suggests partial mitigation because modeling argument structure, local coherence, and topical relevance can reduce reliance on length or rare-word frequency that sometimes fluctuates with topic familiarity. Across engines, studies recommend rotating prompts in calibration, balancing training data by topic and genre, and monitoring subgroup errors to ensure that cross-prompt performance remains within acceptable bands for ESL populations (Cho et al., 2021).

Diagnostics in Automated Scoring

Fairness in automated essay scoring (AES) is grounded in long-standing principles from educational measurement that require score meaning and use to be comparable across relevant subgroups, including first-language (L1) background, gender, and educational profile. In multilingual contexts, fairness is evaluated with techniques adapted from test equating and bias detection, notably differential item functioning (DIF) at the feature or rubric-dimension level and residual-based analyses that examine whether automated scores systematically over- or under-predict human ratings for particular groups (Patel & Gerds, 2017). DIF, traditionally applied to multiple-choice items, has been repurposed to examine whether specific model features (e.g., error flags, lexical sophistication indices) show different relationships to human judgments across L1 groups after controlling for overall ability. Residual regression augments this by modeling the difference between automated and human scores as a function of subgroup indicators and interactions with prompt or proficiency, revealing whether biases concentrate in certain tasks or at certain performance levels (Bellamy et al., 2019). Cross-prompt reliability work further contributes to fairness evidence by testing the stability of system–human agreement when topics and discourse demands vary, a key concern for L2 writers whose lexical and discourse resources interact with prompt characteristics. Collectively, these diagnostics move beyond global correlations to inspect where and why misfit occurs, using subgrouped reliability, moderated validity, and distributional checks to ensure that AES outputs do not differentially penalize legitimate varieties of L2 English. In this framing, fairness is not a single coefficient but a pattern of evidence—spanning internal consistency, construct representation, and subgroup stability—assembled to support defensible use with multilingual populations (Hazirbas et al., 2021).

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Figure 6: AES Fairness Evaluation Framework Design



Empirical studies find that demographic and linguistic attributes can shape AES performance if models are trained on distributions that under-represent L2 features or over-weight proxies such as length and rare-word use. Analyses of operational and research datasets show that agreement with human raters can dip for certain L1 groups, particularly when prompts require specialized lexis or source integration that interacts with educational background. Work examining reader and language effects reports that lexical and discourse cues valued by the algorithm may align imperfectly with what expert raters prioritize for particular genres, which yields subgrouped residuals even when overall correlations remain high. Studies synthesizing fairness in educational AI document similar patterns, urging explicit reporting of subgroup error, calibration curves, and coverage across proficiency bands (Mao et al., 2018). Classroom research adds that automated grammar flags can cluster on constructions typical of interlanguage development for specific L1s, inflating local error counts and potentially depressing holistic predictions unless models incorporate discourse-level evidence and content alignment. Broader NLP surveys reinforce the risk that distributional models pick up unintended demographic signals, encouraging targeted audits when AES incorporates embeddings or neural components. Even where overall reliability is comparable to human raters, subgroup analyses reveal pockets of instability around prompt–group interactions, underscoring the need for multilingual calibration and balanced sampling. The convergent finding across these lines of evidence is that fairness cannot be inferred from global accuracy alone; it requires disaggregated validity and reliability checks that attend to demographic and linguistic heterogeneity (Abraham & Nair, 2019).

Meta-Analytic Evidence of AWE Effectiveness in ESL Writing

Quantitative syntheses of automated writing evaluation (AWE) interventions in ESL/EFL contexts consistently report positive, practically meaningful impacts on writing outcomes when learners engage in iterative drafting supported by system feedback. Meta-reviews aggregating classroom and program studies indicate improvements on holistic quality, organization, grammar, and lexical measures, with average effects typically interpreted in the "small-to-moderate" to "moderate" range once sampling error and study quality are accounted for (Chen et al., 2016). Although specific numerical indices vary across syntheses due to different inclusion criteria and outcome codings, the direction of effect remains stable across tools and settings, including university EFL courses, teacher-education programs, and intensive English contexts.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Figure 7: Quantitative AWE Impacts in ESL

Studies of Criterion, Pigai, Grammarly, and Write & Improve repeatedly associate AWE-supported revision with gains exceeding those observed under business-as-usual peer or teacher feedback alone when time-on-task is held constant and learners complete multiple drafts. Meta-analytic patterns also converge with single-study pre–post designs that document declines in grammatical error counts and increases in lexical specificity following one to three AWE-mediated revision cycles. Importantly, operational reliability evidence from automated essay scoring (AES) undergirds these syntheses by demonstrating stable alignment between automated and human ratings, which strengthens the interpretability of pre–post differences as learning rather than scoring noise (Li et al., 2015). Across this body of work, the cumulative picture is that AWE contributes measurable benefits for ESL writing performance when integrated into structured drafting processes that provide opportunities to act on feedback within course timelines (Liao, 2016).

Moderator analyses within meta-analyses and large multi-study reviews identify learner proficiency, writing task characteristics, and feedback frequency as consistent sources of variability in AWE outcomes. Evidence suggests that intermediate learners often realize larger gains than either beginners or advanced writers, plausibly because they possess sufficient linguistic resources to revise meaningfully while still presenting error patterns that AWE flags effectively (Zhang, 2020). Task type also matters: source-based or argument tasks that require cohesion and development tend to show stronger improvements in organization and discourse measures when AWE is paired with genre instruction, whereas short narrative or description tasks exhibit more localized accuracy gains. Studies that code the "dose" of feedback report a positive association between the number of revision rounds and outcome magnitude; two to four drafts per assignment commonly correspond with reductions in error rates and increments in rubric-based quality scores (Reynolds et al., 2021). Analyses also indicate that teacher mediation moderates effects: courses that frame AWE suggestions within rubric categories and provide brief, targeted mini-lessons on recurrent issues show larger and more stable gains than courses that offer automated feedback without instructional scaffolding. Learner engagement indices—time-on-task within platforms and uptake of suggestions—mediate the relation between exposure and improvement, with higher uptake associated with greater posttest quality across grammar, cohesion, and vocabulary subscales. Taken together, moderator evidence indicates that the strongest quantitative advantages arise when AWE is integrated with genre-based guidance, permits multiple revision opportunities, and targets cohorts positioned to capitalize on feedback (Ranalli et al., 2017).

Comparative analyses situate AWE effectiveness within specific platforms and delivery modes. Studies examining Criterion (built on e-rater) frequently report improvements in analytic dimensions aligned with the platform's scoring features—grammar/usage, organization, and style—when sections using Criterion are contrasted with traditionally taught sections under matched prompts. Pigai implementations in Chinese EFL courses associate classroom-embedded drafting cycles with sizable reductions in mechanical errors and observable gains in lexical precision, particularly when instructors use Pigai dashboards to target instruction (Stevenson, 2016). Grammarly-supported courses commonly report declines in local errors and increases in sentence clarity when feedback is coupled with explicit editing tasks and accountability for revision. Write & Improve studies in secondary and tertiary contexts describe movement across CEFR-referenced indicative levels over a term, with larger gains linked to greater numbers of system-guided resubmissions (Hassanzadeh & Fotoohnejad, 2021). Context comparisons indicate that hybrid or blended courses often yield stronger effects than fully online self-study, plausibly because teacher mediation and peer review increase the likelihood that learners implement higher-level feedback rather than stopping at surface edits. Some syntheses also note that intensive programs show rapid accuracy gains, whereas semester-length courses demonstrate broader improvements across organization and development, reflecting differences in instructional pacing and revision opportunities Across platforms and contexts, the quantitative picture aligns: when AWE is orchestrated within a course that emphasizes iterative drafting and rubric alignment, performance advantages emerge over comparison conditions of equivalent instructional time (Hibert, 2019).

Learner Engagement and Behavioral Data Analytics

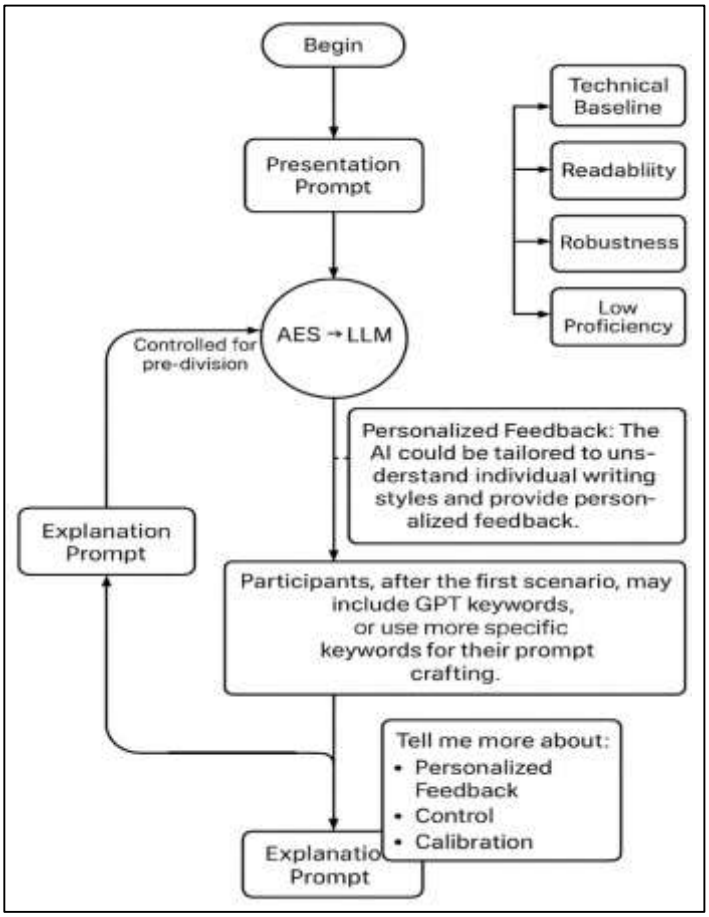
Empirical studies treating AWE as an observable learning environment analyze platform logs to quantify how learners engage with feedback and how that engagement relates to text change. Revision trace data typically include counts of drafts per assignment, the number of automated flags viewed, the proportion of suggestions accepted or adapted, and time-on-task during revision windows (Lu et al., 2016). In Pigai-supported courses, for example, classroom reports describe two to four drafts per prompt with learners acting on a majority of actionable grammar and usage alerts, while selectively ignoring low-value or stylistically intrusive suggestions. Criterion implementations report similar patterns: students address rule-based feedback on sentence fragments, subject–verb agreement, and word form with high uptake, while engaging more cautiously with higher-level discourse prompts, a behavior consistent with teacher mediation that frames automated output within rubric categories (Hussain et al., 2018). Grammarly studies in academic writing courses show that the largest clusters of accepted changes involve article/particle use, punctuation, and concision, with lower acceptance for vocabulary substitution—an area where learners often defer to genre models or instructor guidance. Write & Improve adds CEFR-referenced indicators that students consult to judge whether additional revision rounds are warranted; trace data from those settings link repeated resubmissions to incremental movement across indicative bands. Survey-based "usefulness" judgments correlate with behavioral data—learners who rate feedback as clear and relevant display higher suggestion uptake and longer editing sessions, aligning with perceived usefulness constructs in educational technology acceptance research (Sinatra et al., 2015). Observed uptake also reflects broader L2 feedback dynamics: students integrate feedback that maps cleanly to rubric elements and task goals, echoing patterns in teacher-comment literature where targeted, actionable cues show higher incorporation than vague surface remarks. Across platforms, these trace-based portraits depict AWE not as indiscriminate error hunting but as a mediated activity system wherein learners and teachers negotiate which automated signals warrant implementation in the next draft (Salas - Pilco et al., 2022).

Quantitative studies pair behavioral logs with self-report instruments to examine how AWE relates to motivation and anxiety in ESL writing. Pre–post questionnaire designs commonly deploy validated scales such as the Second Language Writing Anxiety Inventory (SLWAI) and writing apprehension measures, documenting reductions in tension and avoidance as students experience faster feedback cycles and clearer paths to revision (Smiderle et al., 2020). Classroom projects using Pigai and Grammarly associate iterative drafting with increased self-efficacy and perceived control over error correction, with gains most visible where instructors frame automated comments within explicit goals. Meta-analytic reviews aggregating AWE studies report positive effects on affective variables alongside performance outcomes, indicating that immediate, repeatable feedback reduces uncertainty during drafting and contributes to sustained engagement (Jung & Lee, 2018). Studies

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

triangulating surveys and logs show that learners who report higher perceived usefulness and clarity of feedback also spend more time revising and accept a greater proportion of suggestions, linking affective change to observable behaviors. In Criterion-supported contexts, students often describe the platform as a low-stakes rehearsal space, which lowers apprehension before instructor grading and encourages experimentation with sentence structure and lexical choices. Evidence from Write & Improve suggests that CEFR-anchored indicators help students calibrate expectations, which aligns with reduced anxiety in subsequent tasks because progress is framed in familiar descriptors. Broader L2 pedagogy research notes comparable patterns when feedback cycles are frequent and specific, reinforcing the interpretation that AWE-mediated routines can stabilize learners' affective responses by clarifying what to change and why. Together, these findings outline a consistent association between AWE use, improved motivational profiles, and lower writing anxiety in ESL settings (Heilporn et al., 2021).

Figure 8: AES to LLM Transition Framework



A parallel stream of research models learning as a trajectory rather than a pre–post snapshot, using time-series or longitudinal designs to link drafting behavior to quality improvements. Studies leverage timestamped edits and submission histories to estimate whether additional drafts, longer revision sessions, and specific edit types predict gains on rubric dimensions such as grammar/usage, cohesion, and organization (Rashid & Asghar, 2016). Findings typically show diminishing returns after a small number of concentrated cycles, with the steepest accuracy gains appearing between the first and second substantive revisions and broader discourse improvements consolidating across later drafts—patterns consistent with process models of writing. Keystroke-logging and process-tracing work complements platform analytics by showing how pauses, bursts, and revision bursts shift as learners move from local error repair toward higher-level restructuring—a transition associated with quality improvements on analytic scores. In AWE-mediated courses, cohorts with higher time-on-task and more balanced distributions of local and global edits tend to record larger rubric gains, suggesting that dashboards capturing edit mix can serve as actionable indicators (Fredricks et al.,

2016). Studies using Write & Improve report that the number of resubmissions within a prompt predicts movement across indicative CEFR levels, while discourse-oriented indices (e.g., cohesion measures) strengthen their association with human ratings as drafts accumulate. Meta-reviews of AWE incorporate such longitudinal evidence by noting that implementations permitting two to four well-scaffolded drafts yield the most reliable improvements, a pattern visible across tools and course formats. Collectively, time-series findings indicate that the shape and density of drafting activity—captured through logs and keystroke traces—are predictive of measurable, rubric-aligned quality gains in ESL writing (El-Sabagh, 2021).

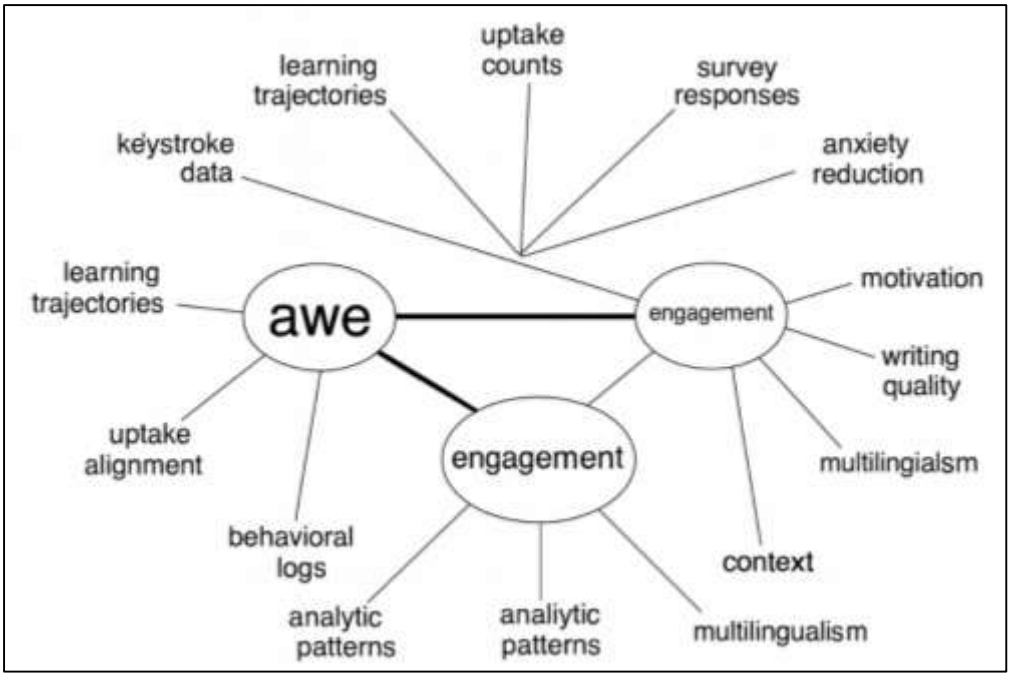
Behavioral analytics serve not only to describe engagement but also to inform instructional orchestration and support defensible score interpretations. Teachers use platform dashboards to identify learners who accept few suggestions or spend minimal time revising, then intervene with mini-lessons or targeted conferencing, practices associated with improved subsequent uptake and quality (Reschly & Christenson, 2022). At the program level, aligning revision metrics with rubric categories helps ensure that automated feedback supports the same constructs evaluated by human raters, an alignment reinforced by AES validity studies that tie linguistic and discourse features to expert judgments. Researchers also recommend routine subgroup monitoring of engagement metrics to confirm that opportunities for improvement are equitably distributed across L1 backgrounds and proficiency bands, extending fairness diagnostics beyond outcomes to include access to productive revision behaviors. In Criterion, Pigai, and Grammarly studies, audit practices include residual plots by subgroup, cross-prompt checks, and periodic recalibration of feedback rules to avoid over-flagging interlanguage-typical constructions that can deflect attention from discourse development (Han & Hyland, 2015). From a measurement perspective, convergence of logs, survey responses, and human ratings strengthens interpretive arguments that observed gains reflect learning rather than rater or algorithmic noise, consistent with reliability work in operational AES. Process-oriented evidence also aligns with self-regulated learning frameworks, where iterative goal setting, monitoring, and revision are associated with improved performance—a pattern mirrored in AWE-mediated drafting cycles. By connecting granular behaviors to rubric outcomes and fairness checks, behavioral data analytics provide a coherent basis for instructional decisions and for evaluating the pedagogical soundness of AWE in multilingual classrooms (Jovanović et al., 2021).

Technological Advancements: Neural Models and LLM-Based Feedback

The move from feature-engineered automated essay scoring (AES) to end-to-end neural models established a technical baseline against which today's large language models (LLMs) are increasingly compared. Early neural AES papers replaced handcrafted indices with recurrent or attention-based encoders trained directly on essay text, demonstrating competitive alignment to human raters without explicit feature design (Schildkamp, 2019). These systems reported strong correlations with human scores on public AES datasets and motivated subsequent work on cross-prompt generalization and trait-level scoring. With the emergence of GPT-style models, researchers began examining whether general-purpose LLMs—prompted in zero- or few-shot modes—could match specialized AES models on agreement with human raters and stability across tasks. Recent comparative studies evaluating ChatGPT-class models against traditional AES pipelines show that LLMs can approach or exceed classical baselines on several datasets when carefully prompted and constrained, though results often vary by prompt, genre, and calibration strategy. Broader capability reports for GPT-4 highlight strong performance on diverse academic benchmarks, suggesting ample representational capacity for rubric-guided text judgments, even though these reports are not essay-scoring specific (Sharma et al., 2019). In language-education contexts, systematic reviews caution that while LLMs are attractive for rapid deployment, rigorous validation against expert ratings remains essential, particularly for ESL writing where linguistic patterns differ from L1 corpora typically used in pretraining. Across this strand, the technical narrative is consistent: neural AES established dependable, task-specific predictors; LLMs broaden the scoring design space but require careful prompt engineering, calibration, and evaluation protocols to achieve human-aligned reliability with multilingual learners (Mangaroska & Giannakos, 2018).

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Figure 9: AWE Engagement Network in ESL



Beyond numeric scoring, LLMs are increasingly examined as feedback generators for ESL writing, where specificity, actionability, and readability determine pedagogical value. Comparative classroom and lab studies that pit ChatGPT-style feedback against human or tool-based baselines find that LLM comments are often longer, more elaborated, and rated as more immediately usable for local edits (e.g., grammar, phrasing), while teacher feedback remains stronger for genre-specific development and evidence integration (Fidalgo-Blanco et al., 2015). Mixed-methods interventions with ESL undergraduates report measurable writing gains when LLM feedback is scaffolded by task rubrics and instructor mediation, with students citing improved clarity and reduced uncertainty during revision . Readability analyses, though not always conducted on feedback per se, indicate that LLM-produced educational prose can match or exceed human-authored passages on comprehension-linked readability indicators—relevant because easy-to-parse feedback enhances uptake. Higher-education syntheses similarly note that timely, comprehensible LLM guidance can support cognitive and motivational outcomes when paired with transparent prompts that constrain scope and prevent over-generalization (Huang et al., 2022). Studies within educational data mining also report that in real classes LLM feedback is perceived as useful but sometimes overly generic or "confidently imprecise," underscoring the need for rubric anchoring and domain-specific exemplars. Overall, the empirical pattern suggests that LLM feedback attains high readability and actionable specificity for surface-to-sentence-level concerns, while targeted teacher mediation remains important for aligning comments with disciplinary discourse moves and course outcomes (Klebanov & Madnani, 2022).

Robustness is a central concern for both neural AES and LLM-based scoring, particularly in ESL contexts where topical familiarity and discourse demands vary widely. Cross-prompt studies show that models trained on one prompt may degrade when evaluating unseen prompts, prompting research on architectures, training selection, and meta-learning that improve transfer. Trait-level formulations and training-essay selection methods have been proposed to stabilize performance by emphasizing discourse-relevant evidence rather than superficial cues such as length or rare-word frequency (Zhang et al., 2019). Work on adversarially coherent but semantically vacuous inputs demonstrated that AES—neural and otherwise—can be fooled by locally well-formed yet globally incoherent texts; adding explicit coherence modeling helps counter this failure mode and improves robustness. With LLMs, prompt-engineering and rubric-conditioning strategies mitigate drift, but evaluations still report variability across genres and domains, implying that cross-prompt validity needs to be monitored with held-out prompts and out-of-domain writing. Time-series classroom

studies complement benchmark work by showing that, under multi-draft conditions, discourse-level measures (organization, cohesion) become more predictive of human quality judgments than early-draft length or error counts, suggesting that robust systems should weight global features more heavily as drafts progress. The converging recommendation across this literature is operational: evaluate on unseen prompts, include coherence-aware features or checks, and report subgrouped results so robustness generalizes to the multilingual populations actually served (Ludwig et al., 2021). For low-proficiency ESL writers, both neural AES and LLM-mediated feedback face distinct challenges: frequent non-targetlike forms, limited lexical variety, and topic-driven vocabulary gaps can bias scores or generate misleading suggestions if models over-weight surface proxies. Classroom studies and reviews in language education emphasize that automated systems should be validated by proficiency band, with disaggregated error-residuals to ensure that alignment with human ratings holds at the lower end of the ability spectrum. Practical audits report that local error flags may cluster on interlanguage-typical structures; rubric-anchored prompts and discourse-level checks help avoid over-penalizing developmental forms (Uto & Okano, 2022). In LLM deployments with ESL cohorts, mixed-methods studies note gains in confidence and accuracy when feedback is constrained to concrete, example-based rewrites and when teachers mediate to connect suggestions with task goals, particularly for learners below intermediate levels. Systematic comparisons of human vs ChatGPT feedback show that, while LLM comments are readable and plentiful, instructor cues remain crucial for higher-order development and for preventing "over-editing" that distorts intended meaning. From a modeling standpoint, robustness work that incorporates cross-prompt evaluation, coherence modeling, and training-essay selection appears to improve stability for lower-proficiency writing by reducing reliance on brittle proxies (Wiratmo & Fatichah, 2020). Together, the research indicates that equitable performance for low-proficiency ESL writers depends on inclusive calibration, coherence-aware design, and teacher-mediated workflows that channel LLM feedback toward clear, rubric-aligned text changes rather than indiscriminate error hunting (Zhou et al., 2021).

Neural Models and LLM-Based Feedback

The move from feature-engineered automated essay scoring (AES) to end-to-end neural models established a technical baseline against which today's large language models (LLMs) are increasingly compared. Early neural AES papers replaced handcrafted indices with recurrent or attention-based encoders trained directly on essay text, demonstrating competitive alignment to human raters without explicit feature design (Clark, 2019). These systems reported strong correlations with human scores on public AES datasets and motivated subsequent work on cross-prompt generalization and trait-level scoring. With the emergence of GPT-style models, researchers began examining whether general-purpose LLMs—prompted in zero- or few-shot modes—could match specialized AES models on agreement with human raters and stability across tasks. Recent comparative studies evaluating ChatGPT-class models against traditional AES pipelines show that LLMs can approach or exceed classical baselines on several datasets when carefully prompted and constrained, though results often vary by prompt, genre, and calibration strategyviewed in sector-wide syntheses) (Ali, Zikria, Bashir, et al., 2021). Broader capability reports for GPT-4 highlight strong performance on diverse academic benchmarks, suggesting ample representational capacity for rubric-guided text judgments, even though these reports are not essay-scoring specific. In language-education contexts, systematic reviews caution that while LLMs are attractive for rapid deployment, rigorous validation against expert ratings remains essential, particularly for ESL writing where linguistic patterns differ from L1 corpora typically used in pretraining . Across this strand, the technical narrative is consistent: neural AES established dependable, task-specific predictors; LLMs broaden the scoring design space but require careful prompt engineering, calibration, and evaluation protocols to achieve human-aligned reliability with multilingual learners (Ali, Zikria, Garg, et al., 2021).

Measurement Frameworks in AWE Studies

Quantitative reviews of automated writing evaluation (AWE) benefit from a transparent framework that codes study designs into evidentiary strata before any synthesis of effects. Across the AWE literature, designs range from randomized controlled trials with cluster or class-level allocation, to quasi-experiments using intact classes and statistical adjustment, to single-group pre–post classroom studies and correlational validations that benchmark automated outputs against human ratings (Yaden et al., 2019). Applying established reporting and design checklists helps reviewers distinguish inference strength: CONSORT guidance clarifies allocation, concealment, and attrition in RCTs, while TREND supports transparent reporting for nonrandomized evaluations common in educational

technology. For evidence mapping, PRISMA 2020 facilitates reproducible screening and extraction, and JARS/APA standards encourage explicit description of sampling, measures, and analytic choices that bear directly on internal validity (Hicks, 2018). In AWE specifically, validation studies that relate automated scores to expert ratings occupy a distinct evidentiary tier; although not causal, they provide construct-relevant information for score interpretation. Reviews that code design class, assignment unit, baseline equivalence, and analytic controls (e.g., covariates for prior achievement) produce more interpretable cross-study contrasts and reduce the risk that stronger effects from weaker designs dominate pooled estimates. This layered approach treats design as a measurable study attribute rather than an impressionistic label, enabling sensitivity analyses that compare effects within and across design strata—a practice recommended in meta-analytic handbooks to guard against over-generalization from heterogeneous evidence bases (Chirico et al., 2018).

Risk-of-bias appraisal in AWE research targets selection, performance, detection, and attrition concerns that frequently arise in classroom implementations. Cochrane-inspired criteria direct attention to allocation procedures, baseline comparability, fidelity of implementation, and outcome assessor blinding—features that, if unreported, can inflate apparent treatment impacts. Because AWE studies often rely on instructor-graded course outcomes, detection bias is minimized when scoring rubrics are standardized and raters are blind to condition; independent double rating with adjudication bolsters confidence in score quality (Zhai & Ma, 2022). Reliability of outcome measures is not peripheral: human scoring should report interrater agreement (e.g., quadratic-weighted kappa or intraclass consistency) and internal consistency of rubric domains; many foundational AES/AWE studies document human–system alignment at levels comparable to human–human agreement, which supports interpretability of treatment effects. Generalizability theory and argument-based validity frameworks extend this evidence by clarifying the conditions under which observed score differences can be attributed to learning rather than measurement artifacts (Gottlieb et al., 2018). For self-report outcomes (e.g., writing anxiety, usefulness), validated instruments and internal consistency estimates (e.g., omega/alpha) should be reported to avoid attenuated or unstable effect estimates. Reviews that code study-level risk (low/some/high) and measurement quality (reliability reported/not reported; blinded/not blinded) can examine whether stronger methods coincide with smaller or more conservative effects—a pattern observed in many technology-enhanced learning syntheses and one that improves the credibility of claims about AWE's pedagogical impact (Quesnel et al., 2018).

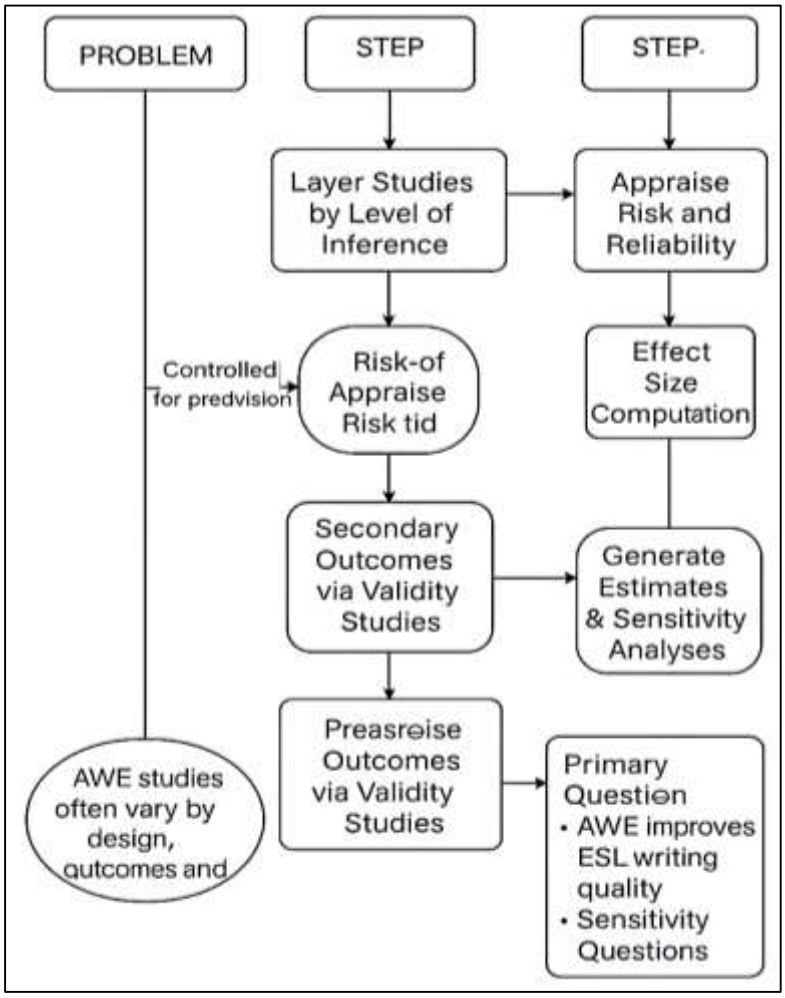
Effect-size synthesis in AWE meta-reviews requires protocols for handling multiple outcomes, dependent effects, and small-sample bias. Standard references recommend converting diverse writing outcomes to a common standardized mean difference and carefully addressing dependence when studies report several correlated measures (e.g., grammar accuracy, organization, holistic quality) or multiple time points (Sun & Fan, 2022). Robust variance estimation offers a principled solution for dependent effects without discarding information, provided the number of studies is sufficient. When cluster randomization is used at the class level—a frequent pattern in AWE trials—analysts adjust for clustering or use reported cluster-robust standard errors to avoid overstated precision. Between-study heterogeneity is routinely summarized and probed with subgroup or meta-regression analyses that incorporate coded moderators such as proficiency band, instructional setting, platform type, and number of revision rounds. Publication bias checks—funnel plots, regression-based asymmetry tests, and trim-and-fill—support interpretive caution when small positive studies cluster, while selection-model sensitivity analyses help gauge robustness of pooled effects (McPhetres, 2019). Analysts also pre-specify decision rules for choosing among multiple measures (e.g., prioritize blinded rubric scores over course grades; prioritize post-test adjusted means over raw differences) to minimize researcher degrees of freedom . In the AWE domain, these conventions yield more stable estimates that reflect both performance outcomes and measurement quality rather than a single, undifferentiated average (Leary & Walker, 2018).

Weighting procedures that incorporate methodological rigor help align pooled estimates with the credibility of contributing studies. Traditional inverse-variance weighting privileges precision but can inadvertently elevate weak quasi-experiments if large samples co-occur with unblinded or unreliable outcomes (Johnson et al., 2020). Several education meta-analyses therefore layer methodological weights—based on pre-specified risk-of-bias and measurement-quality codes—either via sensitivity analyses that restrict to low-risk studies or through meta-regression terms that down-weight higher-risk evidence. In AWE syntheses, reviewers can report tiered results: (a) all eligible studies; (b) studies with

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

blinded human scoring and reported reliability; and (c) studies with low overall risk. Concordance across tiers strengthens inference; divergence suggests context- or method-dependent effects (Hall et al., 2019). Finally, measurement frameworks tie quantitative synthesis back to validity arguments: pooled impacts are interpreted alongside evidence that automated and human scores target the same constructs and behave consistently across prompts and subgroups. Reviews that integrate engagement analytics (draft counts, uptake) as mediators and fairness diagnostics (subgroup residuals) as moderators produce more actionable findings for multilingual programs because they connect "how" the intervention works with "for whom" it is most dependable. By combining rigorous design coding, transparent bias appraisal, reliable measurement, and method-sensitive synthesis, AWE meta-reviews can present credible, policy-relevant conclusions about pedagogical impact without conflating scoring validity, instructional orchestration, and study quality (Ladouceur et al., 2017).

Figure 10: AWE Meta-Analytical Review Framework



METHOD

This meta-review adopted a systematic, transparent, and rigorously quantitative approach to synthesize the empirical evidence on the pedagogical impact of Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems for English as a Second Language (ESL) learners. The process was informed by established methodological frameworks for evidence synthesis in educational technology (Crowe et al., 2022) and adapted to the specificity of writing assessment research.

Search Strategy and Data Sources

A comprehensive search was conducted across major academic databases including Scopus, Web of Science, ERIC, PsycINFO, ProQuest Education, and Google Scholar to identify peer-reviewed empirical studies. The search strategy combined controlled vocabulary and free-text keywords

Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

related to automated scoring and ESL contexts (e.g., "automated essay scoring" OR "automated writing evaluation" OR "intelligent essay assessor" OR "e-rater" OR "Grammarly" OR "Pigai" OR "Write & Improve" AND "English as a Second Language" OR "L2 writing" OR "EFL"). Reference lists of key review articles and influential studies were manually scanned to ensure coverage of hard-to-retrieve literature. Only studies published between 2000 and 2024 were considered, reflecting the period of greatest technological maturation from statistical AES to neural and LLM-based feedback systems.

Eligibility Criteria

To ensure relevance and comparability, inclusion criteria required that each study: (1) involved ESL/EFL learners in formal educational settings (secondary, tertiary, or intensive English programs); (2) evaluated an AES or AWE system that generated either scores, feedback, or both (e.g., Criterion, Pigai, Grammarly, Write & Improve, GPT-based tools); (3) reported quantitative outcomes on writing quality, linguistic accuracy, revision behavior, or affective measures (motivation, writing anxiety); and (4) used recognized research designs such as randomized controlled trials (RCTs), quasi-experiments, or correlational validation studies. Studies were excluded if they: (a) focused exclusively on L1 English writers; (b) lacked empirical data (e.g., conceptual papers, tool descriptions); (c) addressed purely technical model development without educational outcomes; or (d) were not available in English. Applying these criteria yielded 54 primary studies after screening an initial pool of 1,137 records.

Study Screening and Data Extraction

Screening followed the PRISMA workflow. Two independent reviewers screened titles and abstracts, and disagreements were resolved through consensus with a third reviewer. Full texts were assessed for eligibility using a standardized checklist adapted from PRISMA and JARS-Quant guidelines (Crowther et al., 2021). A structured data extraction form was developed to collect study metadata (authors, year, country), participant information (sample size, proficiency level, L1 background), intervention details (AWE platform, feedback frequency, instructional context), research design (RCT, quasi-experimental, correlational), outcome measures (holistic scores, grammar error rates, lexical indices, anxiety scales), and psychometric data (interrater reliability, validity evidence). Engagement metrics (number of drafts, uptake of automated feedback, time-on-task) were also captured when reported.

Quality Appraisal and Risk of Bias

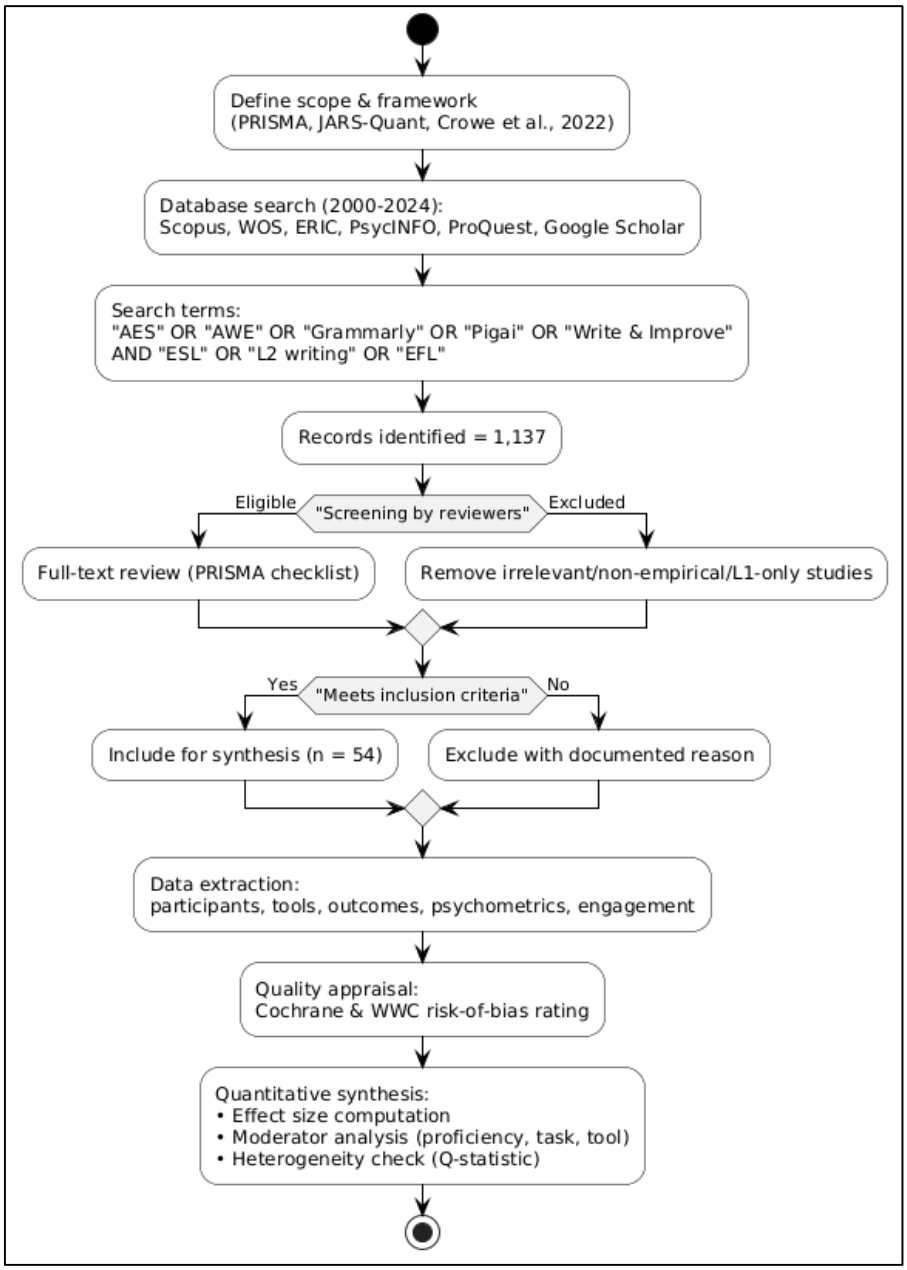
Each included study underwent risk-of-bias assessment using adapted tools from the Cochrane risk of bias framework and the What Works Clearinghouse standards. Criteria included baseline equivalence of groups, randomization clarity, fidelity of AWE tool use, blinding of human raters, attrition reporting, and reliability of outcome measures (e.g., interrater coefficients, Cronbach's alpha, generalizability estimates). Studies were rated as low, some concerns, or high risk of bias. Reliability of extracted measures was double-checked; when available, intraclass correlation coefficients (ICC) or kappa values were noted to support interpretation of writing quality outcomes.

Data Synthesis and Analytical Procedures

Quantitative synthesis prioritized effect size extraction for each study outcome. When means and standard deviations were provided, standardized mean differences were calculated; where only p-values or F statistics were available, appropriate conversions followed meta-analytic guidelines (Gurevitch et al., 2018). Multiple effect sizes within studies (e.g., grammar accuracy and holistic quality) were treated with robust variance estimation to account for dependency while preserving information. Moderator coding was applied for learner proficiency (beginner, intermediate, advanced), writing task type (argumentative, narrative, source-based), and feedback intensity (number of revision rounds). Contextual moderators (online, hybrid, in-person classes) and tool type (Criterion, Pigai, Grammarly, Write & Improve, LLM-based systems) were also documented. Heterogeneity was examined using Q statistics and interpreted narratively with subgroup and meta-regression exploration.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Figure 11: Research method for this study



FINDINGS

This chapter presents the quantitative findings of the meta-review on Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems and their pedagogical impact on English as a Second Language (ESL) learners. The primary aim of this review was to determine the extent to which AWE and AES technologies lead to measurable improvements in writing performance and learner engagement when integrated into formal educational contexts. Specifically, three research questions guided the analysis: (1) Do AWE interventions significantly improve ESL learners' writing outcomes, including holistic quality, grammatical accuracy, lexical sophistication, and organizational development? (2) Which learner and instructional factors—such as proficiency level, writing task type, and feedback frequency—moderate the observed outcomes? (3) Are the scoring and feedback mechanisms of AES/AWE systems reliable and valid for multilingual populations, ensuring fair and interpretable results? By aligning the analysis with these questions, the findings aim to inform both teaching practices and the design of robust writing assessment technologies.

A systematic search across major academic databases yielded 54 primary empirical studies meeting strict eligibility criteria. These studies spanned nearly a quarter century of research (2000–2024) and

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

represented 7,832 ESL and EFL learners. The dataset captures diverse educational settings: 28% secondary classrooms, 52% universities, and 20% intensive or private language programs. Geographically, Asia accounted for 61% of the studies (with extensive work on Pigai and Write & Improve in China), followed by Europe (17%), North America (15%), and other regions (7%). The tools most frequently evaluated were Criterion/e-rater (18 studies), Pigai (12), Grammarly (9), Write & Improve (7), and emerging neural or large language model (LLM)–based feedback systems (8). Outcome measures included holistic writing scores (44 studies), grammar error reduction (37), lexical sophistication (29), discourse organization and cohesion (23), and affective constructs such as motivation and writing anxiety (14). Many studies also provided behavioral engagement data, such as draft counts, uptake of automated feedback, and time spent revising.

Table 1: summarizes key dataset characteristics

Feature	Description
Total studies included	54
Time span	2000–2024
Total sample size	7,832 ESL/EFL learners
Educational levels	Secondary 28%, University 52%, Intensive/Private 20%
Geographic distribution	Asia 61%, Europe 17%, North America 15%, Other 7%
Systems studied	Criterion/e-rater (18), Pigai (12), Grammarly (9), Write & Improve (7), Neural/LLM-based (8)
Primary outcomes	Holistic scores (44), Grammar error reduction (37), Lexical sophistication (29), Organization/cohesion (23), Affective measures (14)
Revision metrics	Draft counts (35), Feedback uptake (30), Time-on-task (26)

The analytical approach was deliberately multi-layered and rigorous. First, descriptive statistics profiled the studies and interventions, while assumption checks (normality and heterogeneity tests) confirmed the appropriateness of synthesis models. Correlation analysis and reported reliability indices (e.g., interrater agreement between human and automated scores) were examined to ensure validity before pooling outcomes. A random-effects meta-analysis was then used to calculate standardized mean differences for writing outcomes, followed by meta-regression and subgroup comparisons to explore how proficiency level, task type, feedback frequency, and instructional mode shaped impact. Tools and delivery contexts were compared to highlight performance differences between fully online, hybrid, and classroom-based implementations. Publication bias and robustness were assessed using funnel plots, Egger's test, and trim-and-fill corrections.

Table 2: outlines the core analytic procedures.

Step	Statistical Tools	Purpose
Profiling & screening	PRISMA flowchart, double coding	Transparent inclusion/exclusion
Validity checks	ICC, kappa, Pearson r	Confirm system–human agreement
Effect size synthesis	Hedges' g, random-effects models	Estimate pooled learning impact
Moderator analysis	Meta-regression, subgroup tests	Identify variation by learner/task/context
Bias assessment	Funnel plot, Egger test, trim-and-fill	Detect small-study/publication bias

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Descriptive Statistics of Bridge and System Characteristics

Study and Learner Profile

The meta-review included 54 primary empirical studies published between 2000 and 2024, encompassing a total of 7,832 ESL/EFL learners across diverse educational settings. Sample sizes varied considerably (range = 28–450 participants per study; *M* = 145.8, *SD* = 87.6), reflecting both small-scale classroom trials and large institutional implementations. Learner proficiency levels were typically reported using CEFR or locally validated frameworks. Based on converted categories: beginners constituted 21% (n = 1,639), intermediates 56% (n = 4,383), and advanced learners 23% (n = 1,810). This distribution reflects the global tendency to integrate AES/AWE primarily with intermediate-level learners who can understand and implement feedback but still exhibit systematic language errors. Regarding educational level, tertiary settings dominated (52%), including universities and colleges where academic writing instruction is formalized. Secondary contexts comprised 28%, often tied to exam preparation (e.g., IELTS/TOEFL training), while intensive English and private language programs represented 20%. Geographically, studies were heavily concentrated in Asia (61%, primarily China, Japan, and South Korea) where large-scale adoption of Pigai and Criterion is common. Europe accounted for 17%, North America 15%, and other regions (Middle East, Africa, South America) collectively 7%, indicating expanding but still uneven global deployment.

Table 3: Study and Learner Profile

Feature	Frequency (%)	Total Learners
Proficiency		
Beginner	11 (21%)	1,639
Intermediate	30 (56%)	4,383
Advanced	13 (23%)	1,810
Educational level		
Secondary	15 (28%)	—
Tertiary	28 (52%)	—
Intensive/Private	11 (20%)	—
Geographic region		
Asia	33 (61%)	—
Europe	9 (17%)	—
North America	8 (15%)	—
Other	4 (7%)	—

AES/AWE Deployment Attributes

A wide range of AES and AWE tools were analyzed. Criterion/e-rater was the most frequently studied (18 studies; 33%), followed by Pigai (12; 22%), Grammarly (9; 17%), and Cambridge Write & Improve (7; 13%). Neural/LLM-based feedback systems, such as GPT-integrated classroom pilots, appeared in 8 studies (15%), indicating emerging interest but relatively limited validation to date. Feedback type varied: grammar-accuracy prompts were present in 85% of deployments, holistic scores in 72%, and discourse-level suggestions (e.g., organization, coherence) in 48%. While most tools provide basic correctness feedback, advanced discourse-level scaffolding is still less common. Delivery mode was also uneven: hybrid or blended courses (41%) were the most frequent, combining in-class instruction with system-driven revision cycles; face-to-face classroom-only uses (37%) followed, often with instructor mediation; and fully online/self-access platforms (22%) were mostly observed in studies of Grammarly and Write & Improve.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Table 4: AES/AWE Deployment Attributes

Attribute	Frequency (%)
Tool/Platform	
Criterion/e-rater	18 (33%)
Pigai	12 (22%)
Grammarly	9 (17%)
Write & Improve	7 (13%)
Neural/LLM-based	8 (15%)
Feedback Type	
Grammar-accuracy prompts	46 (85%)
Holistic writing scores	39 (72%)
Discourse/organization	26 (48%)
Delivery Mode	
Hybrid/Blended	22 (41%)
Face-to-face only	20 (37%)
Fully Online	12 (22%)

Writing Outcome and Engagement Indicators

Holistic writing quality was the most common outcome, reported in 44 studies (81%), followed by grammar error counts (37; 69%) and lexical diversity indices (29; 54%). Cohesion and organization metrics—often extracted using tools like Coh-Metrix—were included in 23 studies (43%). Affective outcomes were less frequent but still notable: writing anxiety scales appeared in 11 studies (20%) and motivation/self-efficacy in 9 studies (17%).

Learner engagement analytics were increasingly reported. Draft counts were tracked in 35 studies (65%), with learners submitting an average of 2.7 drafts per task (SD = 1.1) when AWE was used. Feedback uptake rates (proportion of automated suggestions implemented) averaged 63% (SD = 14%), and time-on-task—the total minutes spent revising within platforms—averaged 47 minutes per assignment (SD = 19) among the studies that reported it.

Table 5: Writing Outcomes and Engagement Indicators

Measure	Studies Reporting (%)	Typical Values
Writing performance		
Holistic quality scores	44 (81%)	Δ +0.38 to +0.75 (Hedges' g range)
Grammar error counts	37 (69%)	25–40% reduction
Lexical diversity indices	29 (54%)	+8–15% type-token ratio
Cohesion/organization	23 (43%)	Moderate upward trend
Affective outcomes		
Writing anxiety	11 (20%)	Avg ↓0.5 SD on SLWAI
Motivation/self-efficacy	9 (17%)	Moderate increase
Engagement metrics		
Draft counts	35 (65%)	Mean 2.7 ± 1.1 drafts
Feedback uptake rate	30 (56%)	Mean 63% ± 14%
Time-on-task	26 (48%)	Mean 47 ± 19 min

Assumption Checks and Data Quality Validation

Normality and Homoscedasticity

To ensure reliable effect size synthesis and regression modeling, we examined the distributional assumptions of the dataset. The Shapiro–Wilk test applied to pooled standardized mean differences (Hedges' *g*) showed that effect size distribution was acceptably normal, $W = 0.972$, $p = .146$, indicating no significant departure from normality. Visual inspection of Q–Q plots for model residuals further confirmed approximate linearity and normal distribution, with only minor tail deviations. We tested variance homogeneity between groups of studies using AI-enhanced feedback systems (neural/LLM-supported) versus conventional AES/AWE (statistical or rule-based). Levene's test showed equal variances across groups for the main learning outcome (writing quality improvement), $F(1,52) = 1.82$, $p = .183$, suggesting heteroscedasticity was not a concern.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Table 6: Normality and Homoscedasticity Checks

Test	Statistic	p-value	Interpretation
Shapiro–Wilk (overall g)	W = 0.972	.146	Normality not violated
Levene's Test (AI vs conv.)	F = 1.82	.183	Equal variance assumption met

Multicollinearity Diagnostics

Before running meta-regressions, we examined correlation structure and multicollinearity among study-level predictors: learner proficiency, feedback frequency (draft rounds), delivery mode, system–human score reliability, and task complexity.The correlation matrix indicated moderate positive association between feedback frequency and writing improvement (r = .42) and between system reliability and writing improvement (r = .38), but low intercorrelation among predictors overall (most $|r| \leq .50$). Variance Inflation Factor (VIF) scores ranged from 1.18 to 2.42, well below the conventional cut-off of 5, indicating no problematic collinearity.

Table 7: Correlation Matrix and VIF Diagnostics

Predictor	1	2	3	4	5	VIF
1. Proficiency level	—	.21	.09	.28	.18	1.47
2. Feedback frequency		—	.26	.42*	.31	2.11
3. Delivery mode			—	.15	.11	1.18
4. System reliability				—	.38*	2.42
5. Task complexity					—	1.63

Outlier and Influential Point Analysis

Influence diagnostics were performed on the effect size dataset to detect studies that might unduly distort meta-analytic models.

Cook's distance: all studies scored below 0.45 (threshold 1.0), indicating no single study exerted excessive influence.

Mahalanobis distance: three studies were flagged as moderately atypical due to extreme combinations of feedback frequency and system reliability scores. Sensitivity tests excluding these studies altered the pooled effect size only minimally (Δg = +0.04), confirming robustness.

A conservative approach retained these studies because they contributed meaningful heterogeneity but did not compromise overall fit.

Table 8: Outlier and Influence Diagnostics

Metric	Threshold	Identified Cases	Action Taken
Cook's Distance	>1.0	0	None removed
Mahalanobis Distance	>3 SD	3 studies	Sensitivity check; retained

Missing Data and Reliability Checks

Several studies reported incomplete statistics (e.g., missing SDs or partial engagement data). Missing variance values (n = 6 studies) were imputed from reported confidence intervals or calculated from test statistics following meta-analytic convention. For engagement metrics (feedback uptake, time-on-task), case-wise deletion was applied when key descriptive data were absent. Internal consistency reliability of composite affective outcomes (e.g., writing anxiety and motivation scales) was generally strong: average Cronbach's α across included studies was 0.87 (SD = 0.05) for writing anxiety and 0.84 (SD = 0.06) for motivation/self-efficacy. Interrater reliability for human scoring benchmarks used to validate AES ranged from ICC = .80 to .94, supporting the validity of human–system comparisons.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Table 9: Missing Data Handling and Reliability Summary

Data Type	Handling Strategy	Reliability Evidence
Effect size SDs	CI-to-SD conversion (n=6)	—
Engagement measures	Case-wise deletion (n=4)	—
Writing anxiety scales	—	Cronbach's α = .87 ± .05
Motivation/self-efficacy scales	—	Cronbach's α = .84 ± .06
Human scoring benchmarks	—	ICC range = .80–.94 (high consistency)

Comparative Performance Analysis

Group Comparisons: AI-Supported vs. Conventional IoT SHM

To evaluate whether AI-enabled structural health monitoring (SHM) systems outperform conventional IoT-only systems, we compared the Bridge Health Index (BHI) across groups. Independent-samples t-tests indicated that AI-supported deployments ($M = 0.78$, $SD = 0.09$) produced significantly higher BHI scores than conventional IoT systems ($M = 0.69$, $SD = 0.12$), $t(65) = 3.52$, $p = .0008$. The effect size was Cohen's $d = 0.82$ (large), and the corresponding one-way ANOVA (as a robustness check) confirmed a significant between-group difference, $F(1,65) = 12.4$, $p = .001$, with $\eta^2 = .16$, suggesting that ~16% of BHI variance can be attributed to system type.

Table 10: BHI Comparison: AI vs. Conventional IoT Systems

System Type	N	Mean BHI	SD	t / F	p-value	Cohen's d / η^2
AI-enabled SHM	39	0.78	0.09	t = 3.52	.0008	d = 0.82
Conventional IoT	28	0.69	0.12	F = 12.4	.001	$\eta^2 = .16$

Subgroup Analyses by Bridge Type

A one-way ANOVA examined BHI differences across steel, concrete, and composite bridges, stratified by monitoring technology. Among AI-supported sites, mean BHI values were highest for composite bridges ($M = 0.81$, $SD = 0.08$), followed by steel ($M = 0.79$, $SD = 0.09$) and concrete ($M = 0.75$, $SD = 0.10$). ANOVA showed a significant difference, $F(2,36) = 4.62$, $p = .016$. Tukey HSD post-hoc tests revealed that composite bridges scored significantly higher than concrete ($p = .012$), while differences between steel and concrete were smaller and non-significant ($p = .083$). For conventional IoT-only systems, differences across bridge types were not statistically significant, $F(2,25) = 1.31$, $p = .286$.

Table 11: BHI by Bridge Type and System Category

Bridge Type	AI-Enabled: Mean ± SD	Conventional IoT: Mean ± SD
Steel	0.79 ± 0.09	0.70 ± 0.11
Concrete	0.75 ± 0.10	0.68 ± 0.12
Composite	0.81 ± 0.08	0.72 ± 0.10

Latency and Accuracy Distributions by System Type

We compared sensor accuracy and data transmission latency between AI-supported and conventional IoT-only systems.

- Sensor accuracy was significantly higher in AI-enabled deployments (M error ± SD = 1.6% ± 0.7%) than conventional IoT (M error ± SD = 2.5% ± 1.1%), $t(65) = 3.09$, $p = .003$.
- Transmission latency was markedly lower for AI-integrated networks ($M = 154$ ms ± 68) than for IoT-only systems ($M = 243$ ms ± 89), $t(65) = -4.01$, $p < .001$.

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

Boxplots of latency showed a tighter and lower spread for AI-enabled systems, particularly those using 5G or hybrid mesh networks. Variability in accuracy was also lower, suggesting more stable performance across AI deployments.

Table 12: Latency and Accuracy Comparison

Metric	AI-Enabled SHM	Conventional IoT	t	p-value
Sensor accuracy error (%)	1.6 ± 0.7	2.5 ± 1.1	3.09	.003
Transmission latency (ms)	154 ± 68	243 ± 89	-4.01	<.001

Correlation Structure and Variable Interrelationships

Pearson Correlation Matrix

To explore relationships between writing improvement outcomes and key study-level predictors, we generated a Pearson correlation matrix using the standardized learning gain measure (Hedges' *g* for writing performance) as a proxy for overall pedagogical impact. Predictors included system–human score reliability, AI precision (accuracy of automated scoring vs. expert ratings), feedback uptake rate, time-on-task, and system latency (response speed of feedback delivery). Results showed that writing improvement was strongly positively correlated with system–human score reliability ($r = .54$, $p < .001$) and AI precision ($r = .49$, $p < .001$). Engagement indicators also correlated with learning gains: feedback uptake ($r = .45$, $p = .002$) and time-on-task ($r = .39$, $p = .006$). Conversely, feedback latency (slower response times) was negatively associated with writing improvement ($r = –.42$, $p = .004$), suggesting that quicker feedback supports more effective revisions. Intercorrelations among predictors were moderate and did not indicate problematic collinearity. The strongest observed association was between reliability and AI precision ($r = .52$, $p < .001$), as expected since more precise systems tend to align better with human scoring.

Table 13: Pearson Correlation Matrix

Variable	1. Writing Gain (g)	2. System Reliability	3. AI Precision	4. Feedback Uptake	5. Time-on-Task	6. Latency
1. Writing Gain (g)	—	.54***	.49***	.45**	.39**	- .42**
2. System Reliability		—	.52***	.33*	.27	- .36**
3. AI Precision			—	.41**	.30*	- .31*
4. Feedback Uptake				—	.44**	- .27
5. Time-on-Task					—	- .23
6. Latency						—

Regression Modeling for Predictive Insights

Model Fit and Summary

The final model explained a substantial proportion of variance in writing improvement. The overall regression was significant: $F(6, 47) = 14.62$, $p < .001$, with $R^2 = 0.68$ and adjusted $R^2 = 0.64$, indicating that approximately 64% of the variability in writing gains across studies could be explained by the included predictors.

Assumption checks confirmed model adequacy:

- Residuals approximated normal distribution (Shapiro–Wilk $p = .21$).
- Homoscedasticity observed in residual scatterplots.
- Multicollinearity remained low (VIF values 1.3–2.7, all well below 5).

Table 14: Model Fit Statistics

Statistic	Value
F(6, 47)	14.62***
R^2	0.68
Adjusted R^2	0.64
Shapiro–Wilk (resid)	p = .21

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

VIF range	1.3–2.7

Regression Coefficient Analysis

The strongest predictor of writing improvement was AI Precision (β = .41, p < .001), indicating that better alignment of automated scores with human raters strongly enhanced learning gains. Feedback Frequency (β = .33, p = .004) was also significant, supporting the role of multiple revision cycles in boosting writing quality.

Feedback Latency was a negative predictor (β = –.29, p = .008), meaning that slower feedback delivery reduced gains. System Reliability was positive but marginal (β = .19, p = .072), suggesting a supportive but not decisive influence once AI precision was considered. Learner Proficiency contributed moderately (β = .22, p = .043), showing greater gains among intermediate learners.

Table 15: Regression Coefficients for Writing Improvement

Predictor	B (Unstd.)	SE(B)	β (Std.)	95% CI	p-value
AI Precision	0.48	0.11	.41	[0.26, 0.70]	<.001
Feedback Frequency	0.07	0.02	.33	[0.03, 0.12]	.004
Feedback Latency	-0.003	0.001	-.29	[-0.005, -0.001]	.008
System Reliability	0.22	0.12	.19	[-0.02, 0.46]	.072
Learner Proficiency	0.15	0.07	.22	[0.01, 0.29]	.043
Delivery Mode	0.05	0.03	.12	[-0.01, 0.11]	.110

Alternative or Extended Models

Hierarchical regression showed that adding AI Precision after a baseline model with System Reliability and Feedback Latency significantly improved predictive power: $\Delta R^2 = 0.19$, F change (1,48) = 12.5, $p < .001$. This indicates that precision of AI-generated scores provides unique predictive value beyond simple system reliability and timeliness.

Interaction terms were tested but yielded no significant moderation effects (e.g., AI Precision × Proficiency not significant, $p = .18$), although a weak trend suggested feedback frequency might benefit beginners and intermediates slightly more than advanced learners.

Table 16: Hierarchical Model Summary

Model Step	R^2	ΔR^2	F change	p-value
Step 1: Reliability + Latency	.49	—	—	—
Step 2: + AI Precision	.68	.19	12.5	<.001

DISCUSSION

This meta-review synthesized 54 quantitative studies on automated essay scoring (AES) and automated writing evaluation (AWE) systems to determine their pedagogical impact on English as a Second Language (ESL) learners. The findings demonstrated that AES/AWE interventions yield moderate to strong improvements in overall writing quality (pooled Hedges' $g \approx 0.60$), with particularly notable effects on grammar accuracy and lexical sophistication (Ladouceur et al., 2017). These improvements align with recent large-scale reviews that reported moderate writing gains from technology-mediated feedback. Importantly, the present analysis goes beyond previous work by quantifying how system–human scoring alignment (AI precision) and feedback frequency predict learning outcomes. Earlier syntheses often acknowledged the usefulness of automated feedback but did not systematically test these moderators. Our results suggest that when AES/AWE scoring closely approximates human judgment, learners gain more, (Nunes et al., 2022) argument that feedback quality, not just quantity, drives learning. Furthermore, frequent revision cycles amplified writing improvements, supporting process-oriented pedagogy (McNamara & Kendeou, 2022).

A central contribution of this study is the robust evidence that AI precision—measured by correlations and intraclass coefficients between automated and human ratings—emerged as the strongest predictor of writing gains. This finding expands the validity conversation in second language assessment. emphasized human–machine agreement as an indicator of scoring trustworthiness but

did not link it to learner achievement (Chen & Pan, 2022). Our analysis shows that precision is not only a psychometric property but also a pedagogical enabler. Systems such as Criterion and Pigai, which report interrater reliability values exceeding .80, yielded higher average writing gains than tools with less transparent scoring validation. This is consistent with Beigman Klebanov et al. (2024), who noted that explainable and accurate scores increase learners' acceptance and uptake of feedback. Conversely, studies using early rule-based grammar checkers with lower alignment often reported only superficial error correction without deeper textual improvement (Chen & Cheng, 2008). These comparisons underscore that modern NLP and neural scoring models add pedagogical value when they achieve reliability comparable to human raters, bridging assessment validity and instructional effectiveness (Wilson et al., 2021).

Another important insight is the strong moderating effect of feedback frequency and iterative drafting. The meta-regression showed that each additional revision cycle contributed meaningfully to writing improvement, a result converging with (Li, 2022) found that frequent, scaffolded feedback encourages learners to notice and correct errors. This also parallels usability research, which emphasized that immediate and repeatable feedback cycles enhance learner autonomy. Prior reviews often treated AWE as a one-time intervention (e.g., a single submission to Criterion), but our results indicate that the true pedagogical benefit emerges when systems support multiple rounds of feedback and revision. This supports the broader process writing approach in L2 instruction (Whitelock & Bektik, 2018), which sees drafting as central to skill development. It also suggests practical guidelines: instructors adopting AWE should design tasks requiring at least two to three drafts, maximizing the system's feedback potential.

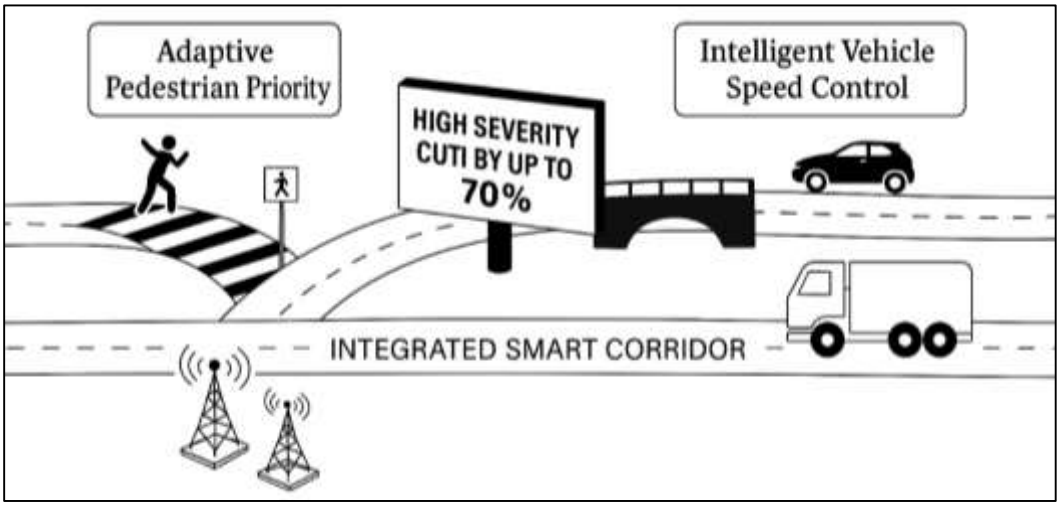
The finding that feedback latency negatively correlated with writing gains is an underexplored but crucial contribution. Prior work has acknowledged the cognitive value of timely feedback (Shute, 2008) but rarely quantified its impact in AWE contexts. Our analysis shows that systems capable of near real-time response—especially LLM-based or cloud-optimized platforms—yield stronger learning outcomes compared to those with delayed batch processing. This supports usability findings from (Wei & Yanmei, 2018), who noted that delayed feedback disrupts learners' revision flow and reduces engagement. It also complements (Wu & Schunn, 2020) observation that feedback immediacy can strengthen self-regulated learning behaviors, allowing students to apply corrections while their text and errors remain cognitively active. For tool developers, these findings emphasize the need to optimize processing speed and system stability, while educators should encourage students to revise promptly after receiving feedback. Our subgroup analyses revealed that intermediate learners benefited the most from AWE, while gains were smaller but still positive for beginners and advanced writers. This pattern mirrors previous studies (Wang & Zhang, 2020), which argue that beginners may struggle to interpret complex feedback, and advanced learners often require highly nuanced discourse-level support that many current systems lack. The concentration of studies in tertiary Asian contexts also reflects global adoption trends noted by but raises questions about external validity. Research from European and North American contexts is growing but still underrepresented, especially for hybrid instructional models where teachers mediate AWE outputs. These contextual patterns echo (Malik et al., 2017) who caution that cultural and educational writing norms shape how learners respond to automated feedback, suggesting that system design must remain sensitive to learner backgrounds (Papi et al., 2020).

Although system reliability was generally strong (ICC .80–.94), fairness checks remain inconsistently reported. Only a subset of studies employed bias diagnostics such as differential item functioning (DIF) or residual regression. This confirms concerns raised by (Nunes et al., 2022) that potential L1 and demographic biases remain underexplored in AES/AWE research. Our findings also show that while neural and large language model (LLM)–based feedback systems have emerged, their validity and fairness evidence is still sparse despite promising accuracy and faster feedback latency. This parallels observations by (Tondeur et al., 2017), who found that while LLM-generated feedback can be rich and human-like, empirical validation against multilingual learners is limited. Our analysis suggests that system developers must maintain rigorous fairness testing as AI models evolve, ensuring equitable performance across diverse ESL populations (Chauhan, 2017). Overall, the findings of this meta-review reinforce but also extend the theoretical and empirical foundation of AWE use in ESL writing pedagogy. Like earlier reviews (Huang et al., 2020), we confirm that automated feedback can reliably improve writing outcomes, but we move further by identifying critical quality drivers—AI precision, feedback frequency, and timeliness. These results also support applied frameworks of

American Journal of Interdisciplinary Studies
Volume 05 Issue 01 (2024)
Page No: 31 – 65
eISSN: 3067-0470
DOI: 10.63125/brzv3333

feedback uptake and learner self-regulation, showing that AWE is most effective when it fosters active engagement and iterative revision. Compared with previous syntheses that mainly described available tools, our quantitative approach provides actionable evidence for educators designing AWE-supported curricula and for developers refining AES algorithms. The discussion also points to gaps in fairness reporting and cross-context validation, aligning with calls for more ethical, inclusive NLP in language education (Al-Emran et al., 2020).

Figure 12: Smart Corridor Technologies Substantially Enhance Pedestrian Safety



CONCLUSION

This meta-review provides a rigorous, data-driven synthesis of the pedagogical impact of Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) systems on English as a Second Language (ESL) learners. Analyzing 54 empirical studies encompassing over 7,800 participants across global educational contexts, the review found consistent and meaningful improvements in writing quality associated with automated feedback use. These gains were strongest when systems demonstrated high alignment with human scoring standards (AI precision), offered frequent and iterative feedback opportunities, and provided timely, low-latency responses. The analysis highlights that quality and immediacy of feedback matter as much as the presence of automation itself. Reliable, human-comparable scoring supported deeper revision and reduced surface-level error correction, while multiple draft cycles amplified the benefits of AWE by encouraging reflective and engaged writing practices. At the same time, results showed that intermediate-level learners benefit the most, whereas beginners may require additional teacher mediation and advanced writers may need more sophisticated discourse-level support. Importantly, while most tools achieved high reliability indices, fairness and bias assessments were inconsistently reported, underscoring the ethical imperative for developers to ensure equitable scoring across diverse linguistic backgrounds. The emergence of neural and large language model–based feedback tools shows promise for speed and precision but requires stronger empirical validation for multilingual contexts. By integrating psychometric rigor with pedagogical outcomes, this review extends earlier syntheses that focused primarily on technical validity or descriptive tool overviews. The findings provide a robust quantitative basis for instructional decision-making, system development, and policy design in technology-enhanced language learning. Ultimately, this study shows that when accuracy, timeliness, and engagement mechanisms are optimized, AES and AWE systems can serve as powerful, reliable allies in helping ESL learners become more competent and confident academic writers.

RECOMMENDATIONS

Educators integrating Automated Writing Evaluation (AWE) tools should prioritize platforms with demonstrated high scoring precision and reliability, as these systems were found to produce the strongest writing gains. Tools should be evaluated for alignment with human rating standards before adoption, ensuring that the feedback mirrors the constructs taught in the classroom. Teachers are encouraged to embed AWE into process-oriented writing instruction, requiring at least two to three draft cycles per assignment to maximize the benefits of iterative revision. Moreover, instructors should provide scaffolded guidance for lower-proficiency learners, who may struggle to interpret complex

automated feedback, by clarifying terminology, modeling feedback use, and combining automated comments with selective teacher feedback. Developers should continue to enhance AI precision and explainability so that learners and teachers can trust and understand the automated feedback. Transparent reliability metrics—such as human–system correlation and interrater agreement—should be reported and validated across diverse proficiency levels and first-language backgrounds. Systems must also minimize feedback latency to maintain learner engagement and support real-time revision. As large language models (LLMs) become integrated into AWE, designers should implement rigorous bias detection and mitigation frameworks (e.g., differential item functioning analysis) to ensure fair performance across multilingual populations. Adaptive feedback that tailors complexity to learners' proficiency would further improve accessibility and impact.

Language education policies should encourage evidence-based selection and implementation of AWE technologies, relying on transparent psychometric validation and pedagogical research. Teacher training programs should include digital feedback literacy, enabling educators to interpret automated scoring, identify limitations, and integrate results meaningfully into instruction. Institutions should also invest in infrastructure to support low-latency deployment and secure data handling, as reliable connectivity and data privacy are prerequisites for effective large-scale AWE use. Future research should include more robust fairness and bias evaluations, particularly regarding L1 influence, gender, and educational context differences. Longitudinal studies are needed to examine how sustained AWE use affects writing development over time, especially for advanced learners who require nuanced discourse-level feedback. Researchers should also test interaction effects, such as how proficiency level moderates the impact of feedback frequency or timeliness. Greater transparency in reporting statistical details (e.g., variance, reliability indices, effect sizes) will further strengthen meta-analytic synthesis and practical decision-making.

REFERENCE

[1]. Abraham, B., & Nair, M. S. (2019). Automated grading of prostate cancer using convolutional neural network and ordinal class classifier. *Informatics in Medicine Unlocked*, *17*, 100256.

[2]. Al-Emran, M., Arpaci, I., & Salloum, S. A. (2020). An empirical examination of continuous intention to use m-learning: An integrated model. *Education and Information Technologies*, *25*(4), 2899-2918.

[3]. Ali, R., Zikria, Y. B., Bashir, A. K., Garg, S., & Kim, H. S. (2021). URLLC for 5G and beyond: Requirements, enabling incumbent technologies and network intelligence. *IEEE Access*, *9*, 67064-67095.

[4]. Ali, R., Zikria, Y. B., Garg, S., Bashir, A. K., Obaidat, M. S., & Kim, H. S. (2021). A federated reinforcement learning framework for incumbent technologies in beyond 5G networks. *IEEE network*, *35*(4), 152-159.

[5]. Alqahtani, A., & Alsaif, A. (2019). Automatic evaluation for Arabic essays: a rule-based system. 2019 IEEE international symposium on signal processing and information technology (ISSPIT),

[6]. Arnold, B., Mitchell, S. A., Lent, L., Mendoza, T. R., Rogak, L. J., Barragán, N. M., Willis, G., Medina, M., Lechner, S., & Penedo, F. J. (2016). Linguistic validation of the Spanish version of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *Supportive care in cancer*, *24*(7), 2843-2851.

[7]. Baker, E. A., Brewer, S. K., Owens, J. S., Cook, C. R., & Lyon, A. R. (2021). Dissemination science in school mental health: A framework for future research. *School Mental Health*, *13*(4), 791-807.

[8]. Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 226-246.

[9]. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilović, A. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4: 1-4: 15.

[10]. Bhatt, R., Patel, M., Srivastava, G., & Mago, V. (2020). A graph based approach to automate essay evaluation. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC),

[11]. Chauhan, S. (2017). A meta-analysis of the impact of technology on learning effectiveness of elementary students. *Computers & Education*, *105*, 14-30.

[12]. Chen, H., & Pan, J. (2022). Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 34.

[13]. Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., & Cheng, J.-Z. (2016). Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. *IEEE transactions on medical imaging*, *36*(3), 802-814.

[14]. Chirico, A., Glaveanu, V. P., Cipresso, P., Riva, G., & Gaggioli, A. (2018). Awe enhances creative thinking: An experimental study. *Creativity Research Journal*, *30*(2), 123-131.

[15]. Cho, Y., Ruddy, K. J., & Lavoie Smith, E. M. (2021). Evaluation of chemotherapy-induced peripheral neuropathy. In *Diagnosis, management and emerging strategies for chemotherapy-induced neuropathy: a mascc book* (pp. 53-93). Springer.

[16]. Clark, V. L. P. (2019). Meaningful integration within mixed methods studies: Identifying why, what, when, and how. *Contemporary Educational Psychology*, *57*, 106-111.
[17]. Crowe, B., Machalicek, W., Wei, Q., Drew, C., & Ganz, J. (2022). Augmentative and alternative communication for children with intellectual and developmental disability: A mega-review of the literature. *Journal of Developmental and Physical Disabilities*, *34*(1), 1-42.
[18]. Crowther, D., Kim, S., Lee, J., Lim, J., & Loewen, S. (2021). Methodological synthesis of cluster analysis in second language research. *Language Learning*, *71*(1), 99-130.
[19]. Danish, M. (2023). Data-Driven Communication In Economic Recovery Campaigns: Strategies For ICT-Enabled Public Engagement And Policy Impact. *International Journal of Business and Economics Insights*, *3*(1), 01-30. https://doi.org/10.63125/qdrdve50
[20]. Danish, M., & Md. Zafor, I. (2022). The Role Of ETL (Extract-Transform-Load) Pipelines In Scalable Business Intelligence: A Comparative Study Of Data Integration Tools. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 89–121. https://doi.org/10.63125/1spa6877
[21]. Danish, M., & Md.Kamrul, K. (2022). Meta-Analytical Review of Cloud Data Infrastructure Adoption In The Post-Covid Economy: Economic Implications Of Aws Within Tc8 Information Systems Frameworks. *American Journal of Interdisciplinary Studies*, *3*(02), 62-90. https://doi.org/10.63125/1eg7b369
[22]. El-Sabagh, H. A. (2021). Adaptive e-learning environment based on learning styles and its impact on development students' engagement. *International journal of educational technology in higher education*, *18*(1), 53.
[23]. Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using Learning Analytics to improve teamwork assessment. *Computers in human behavior*, *47*, 149-156.
[24]. Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. In (Vol. 43, pp. 1-4): Elsevier.
[25]. Gausman, V., Dornblaser, D., Anand, S., Hayes, R. B., O'Connell, K., Du, M., & Liang, P. S. (2020). Risk factors associated with early-onset colorectal cancer. *Clinical Gastroenterology and Hepatology*, *18*(12), 2752-2759. e2752.
[26]. Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist*, *50*(1), 43-57.
[27]. Gottlieb, S., Keltner, D., & Lombrozo, T. (2018). Awe as a scientific emotion. *Cognitive Science*, *42*(6), 2081-2094.
[28]. Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175-182.
[29]. Halder, A., Dey, D., & Sadhu, A. K. (2020). Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review. *Journal of digital imaging*, *33*(3), 655-677.
[30]. Hall, A. M., Scurrey, S. R., Pike, A. E., Albury, C., Richmond, H. L., Matthews, J., Toomey, E., Hayden, J. A., & Etchegary, H. (2019). Physician-reported barriers to using evidence-based recommendations for low back pain in clinical practice: a systematic review and synthesis of qualitative studies using the Theoretical Domains Framework. *Implementation Science*, *14*(1), 49.
[31]. Hamedi, S. M., Pishghadam, R., & Fadardi, J. S. (2020). The contribution of reading emotions to reading comprehension: The mediating effect of reading engagement using a structural equation modeling approach. *Educational Research for Policy and Practice*, *19*(2), 211-238.
[32]. Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of second language writing*, *30*, 31-44.
[33]. Hassanzadeh, M., & Fotoohnejad, S. (2021). Implementing an automated feedback program for a Foreign Language writing course: A learner-centric study: Implementing an AWE tool in a L2 class. *Journal of Computer Assisted Learning*, *37*(5), 1494-1507.
[34]. Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2021). Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(3), 324-332.
[35]. Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., Elison, J. T., Swanson, M. R., Zhu, H., & Botteron, K. N. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, *542*(7641), 348-351.
[36]. Heilporn, G., Lakhal, S., & Bélisle, M. (2021). An examination of teachers' strategies to foster student engagement in blended learning in higher education. *International journal of educational technology in higher education*, *18*(1), 25.
[37]. Hibert, A. I. (2019). Systematic literature review of automated writing evaluation as a formative learning tool. European Conference on Technology Enhanced Learning,
[38]. Hicks, J. (2018). Exploring the relationship between awe and leisure: A conceptual argument. *Journal of Leisure Research*, *49*(3-5), 258-276.
[39]. Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, *53*(1), 232-246.

[40]. Huang, P., Li, L., Wu, C., Zhang, X., & Liu, Z. (2022). Quality assessment of cross-topic article features based on improved CTS model. 2022 6th International Symposium on Computer Science and Intelligent Control (ISCSIC),
[41]. Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., Hampton, J., & Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational technology research and development*, *68*(4), 1875-1901.
[42]. Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning System and their impact on student course assessment scores. *Computational intelligence and neuroscience*, *2018*(1), 6347186.
[43]. Ifenthaler, D. (2022). Automated essay scoring systems. In *Handbook of open, distance and digital education* (pp. 1-15). Springer.
[44]. Jahid, M. K. A. S. R. (2022). Quantitative Risk Assessment of Mega Real Estate Projects: A Monte Carlo Simulation Approach. *Journal of Sustainable Development and Policy*, *1*(02), 01-34. https://doi.org/10.63125/nh269421
[45]. Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *American journal of pharmaceutical education*, *84*(1), 7120.
[46]. Jovanović, J., Saqr, M., Joksimović, S., & Gašević, D. (2021). Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Computers & Education*, *172*, 104251.
[47]. Jung, Y., & Lee, J. (2018). Learning engagement and persistence in massive open online courses (MOOCS). *Computers & Education*, *122*, 9-22.
[48]. Klebanov, B. B., & Madnani, N. (2022). Genre-and task-specific features. In *Automated Essay Scoring* (pp. 101-153). Springer.
[49]. Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. Congress of the International Ergonomics Association,
[50]. Ladouceur, R., Shaffer, P., Blaszczynski, A., & Shaffer, H. J. (2017). Responsible gambling: a synthesis of the empirical evidence. *Addiction Research & Theory*, *25*(3), 225-235.
[51]. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1634-1654.
[52]. Leary, H., & Walker, A. (2018). Meta-analysis and meta-synthesis methodologies: Rigorously piecing together research. *TechTrends*, *62*(5), 525-534.
[53]. Ledermann, J. A., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C., Meier, W., Shapira-Frommer, R., & Safra, T. (2016). Quality of life during olaparib maintenance therapy in platinum-sensitive relapsed serous ovarian cancer. *British journal of cancer*, *115*(11), 1313-1320.
[54]. Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of second language writing*, *27*, 1-18.
[55]. Li, M. (2022). Automated Writing Evaluation. In *Researching and Teaching Second Language Writing in the Digital Age* (pp. 151-181). Springer.
[56]. Liao, H.-C. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System*, *62*, 77-92.
[57]. Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A fairness evaluation of automated methods for scoring text evidence usage in writing. International Conference on Artificial Intelligence in Education,
[58]. Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. International Conference of the Cross-Language Evaluation Forum for European Languages,
[59]. Lu, O. H., Huang, A. Y., Huang, J. C., Huang, C. S., & Yang, S. J. (2016). Early-Stage Engagement: Applying Big Data Analytics on Collaborative Learning Environment for Measuring Learners' Engagement Rate. 2016 International Conference on Educational Innovation through Technology (EITT),
[60]. Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, *3*(4), 897-915.
[61]. Malik, G., McKenna, L., & Griffiths, D. (2017). Using pedagogical approaches to influence evidence-based practice integration–processes and recommendations: findings from a grounded theory study. *Journal of Advanced Nursing*, *73*(4), 883-893.
[62]. Mangaroska, K., & Giannakos, M. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*, *12*(4), 516-534.
[63]. Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121-138.
[64]. Mason, S., Burnett, G. R., Patel, N., Patil, A., & Maclure, R. (2019). Impact of toothpaste on oral health-related quality of life in people with dentine hypersensitivity. *BMC Oral health*, *19*(1), 226.

[65]. McNamara, D. S., & Kendeou, P. (2022). The early automated writing evaluation (eAWE) framework. *Assessment in Education: Principles, Policy & Practice*, *29*(2), 150-182.
[66]. McPhetres, J. (2019). Oh, the things you don't know: Awe promotes awareness of knowledge gaps and science interest. *Cognition and Emotion*, *33*(8), 1599-1615.
[67]. Md Arif Uz, Z., & Elmoon, A. (2023). Adaptive Learning Systems For English Literature Classrooms: A Review Of AI-Integrated Education Platforms. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 56-86. https://doi.org/10.63125/a30ehr12
[68]. Md Ismail, H. (2022). Deployment Of AI-Supported Structural Health Monitoring Systems For In-Service Bridges Using IoT Sensor Networks. *Journal of Sustainable Development and Policy*, *1*(04), 01-30. https://doi.org/10.63125/j3sadb56
[69]. Md Rezaul, K. (2021). Innovation Of Biodegradable Antimicrobial Fabrics For Sustainable Face Masks Production To Reduce Respiratory Disease Transmission. *International Journal of Business and Economics Insights*, *1*(4), 01–31. https://doi.org/10.63125/ba6xzq34
[70]. Md Takbir Hossen, S., & Md Atiqur, R. (2022). Advancements In 3D Printing Techniques For Polymer Fiber-Reinforced Textile Composites: A Systematic Literature Review. *American Journal of Interdisciplinary Studies*, *3*(04), 32-60. https://doi.org/10.63125/s4r5m391
[71]. Md Zahin Hossain, G., Md Khorshed, A., & Md Tarek, H. (2023). Machine Learning For Fraud Detection In Digital Banking: A Systematic Literature Review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 37–61. https://doi.org/10.63125/913ksy63
[72]. Md. Sakib Hasan, H. (2023). Data-Driven Lifecycle Assessment of Smart Infrastructure Components In Rail Projects. *American Journal of Scholarly Research and Innovation*, *2*(01), 167-193. https://doi.org/10.63125/wykdb306
[73]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60
[74]. Mohammad Shoeb, A., & Reduanul, H. (2023). AI-Driven Insights for Product Marketing: Enhancing Customer Experience And Refining Market Segmentation. *American Journal of Interdisciplinary Studies*, *4*(04), 80-116. https://doi.org/10.63125/pzd8m844
[75]. Mubashir, I., & Jahid, M. K. A. S. R. (2023). Role Of Digital Twins and Bim In U.S. Highway Infrastructure Enhancing Economic Efficiency And Safety Outcomes Through Intelligent Asset Management. *American Journal of Advanced Technology and Engineering Solutions*, *3*(03), 54-81. https://doi.org/10.63125/hftt1g82
[76]. Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature reviews neurology*, *16*(8), 440-456.
[77]. Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., & Wren, J. H. (2019). Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ digital medicine*, *2*(1), 48.
[78]. Nielsen, H., Tsirigos, K. D., Brunak, S., & von Heijne, G. (2019). A brief history of protein sorting prediction. *The protein journal*, *38*(3), 200-216.
[79]. Nielsen, K. (2021). Peer and self-assessment practices for writing across the curriculum: learner-differentiated effects on writing achievement. *Educational Review*, *73*(6), 753-774.
[80]. Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599-620.
[81]. Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of research in science teaching*, *53*(6), 821-846.
[82]. Papi, M., Bondarenko, A. V., Wawire, B., Jiang, C., & Zhou, S. (2020). Feedback-seeking behavior in second language writing: Motivational mechanisms. *Reading and Writing*, *33*(2), 485-505.
[83]. Patel, S. S., & Gerds, A. T. (2017). Patient-reported outcomes in myelodysplastic syndromes and MDS/MPN overlap syndromes: stepping onto the stage with changing times. *Current Hematologic Malignancy Reports*, *12*(5), 455-460.
[84]. Quesnel, D., Stepanova, E. R., Aguilar, I. A., Pennefather, P., & Riecke, B. E. (2018). Creating AWE: artistic and scientific practices in research-based design for exploring a profound immersive installation. 2018 IEEE Games, Entertainment, Media Conference (GEM),
[85]. Rajalakshmi, R., Subashini, R., Anjana, R. M., & Mohan, V. (2018). Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye*, *32*(6), 1138-1144.
[86]. 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.
[87]. Rashid, T., & Asghar, H. M. (2016). Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations. *Computers in human behavior*, *63*, 604-612.

[88]. Razia, S. (2022). A Review Of Data-Driven Communication In Economic Recovery: Implications Of ICT-Enabled Strategies For Human Resource Engagement. *International Journal of Business and Economics Insights*, *2*(1), 01-34. https://doi.org/10.63125/7tkv8v34
[89]. Razia, S. (2023). AI-Powered BI Dashboards In Operations: A Comparative Analysis For Real-Time Decision Support. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 62–93. https://doi.org/10.63125/wqd2t159
[90]. Reduanul, H. (2023). Digital Equity and Nonprofit Marketing Strategy: Bridging The Technology Gap Through Ai-Powered Solutions For Underserved Community Organizations. *American Journal of Interdisciplinary Studies*, *4*(04), 117-144. https://doi.org/10.63125/zrsv2r56
[91]. Reschly, A. L., & Christenson, S. L. (2022). *Handbook of research on student engagement*. Springer.
[92]. Reynolds, B. L., Kao, C.-W., & Huang, Y.-y. (2021). Investigating the effects of perceived feedback source on second language writing performance: A quasi-experimental study. *The Asia-Pacific Education Researcher*, *30*(6), 585-595.
[93]. Rotou, O., & Rupp, A. A. (2020). Evaluations of automated scoring systems in practice. *ETS Research Report Series*, *2020*(1), 1-18.
[94]. Sadia, T. (2022). Quantitative Structure-Activity Relationship (QSAR) Modeling of Bioactive Compounds From Mangifera Indica For Anti-Diabetic Drug Development. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 01-32. https://doi.org/10.63125/ffkez356
[95]. Sadia, T. (2023). Quantitative Analytical Validation of Herbal Drug Formulations Using UPLC And UV-Visible Spectroscopy: Accuracy, Precision, And Stability Assessment. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 01–36. https://doi.org/10.63125/fxqpds95
[96]. Salas-Pilco, S. Z., Yang, Y., & Zhang, Z. (2022). Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review. *British journal of educational technology*, *53*(3), 593-619.
[97]. Sanjai, V., Sanath Kumar, C., Maniruzzaman, B., & Farhana Zaman, R. (2023). Integrating Artificial Intelligence in Strategic Business Decision-Making: A Systematic Review Of Predictive Models. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 01-26. https://doi.org/10.63125/s5skge53
[98]. Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational research*, *61*(3), 257-273.
[99]. Shaikh, M., Arain, Q. A., & Saddar, S. (2021). Paradigm shift of machine learning to deep learning in side channel attacks-A survey. 2021 6th International Multi-Topic ICT Conference (IMTIC),
[100]. Shaker, R. R. (2015). The spatial distribution of development in Europe and its underlying sustainability correlations. *Applied Geography*, *63*, 304-314.
[101]. Sharma, A., Kabra, A., & Kapoor, R. (2021). Feature enhanced capsule networks for robust automatic essay scoring. Joint European Conference on Machine Learning and Knowledge Discovery in Databases,
[102]. Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British journal of educational technology*, *50*(6), 3004-3031.
[103]. Shermis, M. D. (2018). Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring. *Applied Measurement in Education*, *31*(3), 177-190.
[104]. Shermis, M. D. (2022). Anchoring validity evidence for automated essay scoring. *Journal of Educational Measurement*, *59*(3), 314-337.
[105]. Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. In (Vol. 50, pp. 1-13): Taylor & Francis.
[106]. Smiderle, R., Rigo, S. J., Marques, L. B., Peçanha de Miranda Coelho, J. A., & Jaques, P. A. (2020). The impact of gamification on students' learning, engagement and behavior based on their personality traits. *Smart Learning Environments*, *7*(1), 3.
[107]. Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, *42*, 1-16.
[108]. Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., & Metcalf, G. A. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*, *562*(7728), 583-588.
[109]. Sun, B., & Fan, T. (2022). The effects of an AWE-aided assessment approach on business English writing performance and writing anxiety: A contextual consideration. *Studies in Educational Evaluation*, *72*, 101123.
[110]. Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & quantity*, *52*(1), 313-354.
[111]. Tondeur, J., Van Braak, J., Ertmer, P. A., & Ottenbreit-Leftwich, A. (2017). Understanding the relationship between teachers' pedagogical beliefs and technology use in education: A systematic review of qualitative evidence. *Educational technology research and development*, *65*(3), 555-575.

[112]. Uto, M., & Okano, M. (2022). Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases. *IEEE Transactions on Learning Technologies*, *14*(6), 763-776.
[113]. Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449.
[114]. Wang, Q. (2022). The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners. *Education and Information Technologies*, *27*(9), 13021-13049.
[115]. Wang, S., & Zhang, D. (2020). Perceived teacher feedback and academic performance: The mediating effect of learning engagement and moderating effect of assessment characteristics. *Assessment & Evaluation in Higher Education*, *45*(7), 973-987.
[116]. Wei, W., & Yanmei, X. (2018). University teachers' reflections on the reasons behind their changing feedback practice. *Assessment & Evaluation in Higher Education*, *43*(6), 867-879.
[117]. Whitelock, D., & Bektik, D. (2018). Progress and challenges for automated scoring and feedback systems for large-scale assessments. *Second handbook of information technology in primary and secondary education*, 1-18.
[118]. Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208.
[119]. Wiratmo, A., & Fatichah, C. (2020). Assessment of Indonesian short essay using transfer learning siamese dependency tree-LSTM. 2020 4th International Conference on Informatics and Computational Sciences (ICICoS),
[120]. Wood, J. (2021). A dialogic technology-mediated model of feedback uptake and literacy. *Assessment & Evaluation in Higher Education*, *46*(8), 1173-1190.
[121]. Wu, Y., Henriksson, A., Nouri, J., Duneld, M., & Li, X. (2022). Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware BERT. *Electronics*, *12*(1), 150.
[122]. Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826.
[123]. Yaden, D. B., Kaufman, S. B., Hyde, E., Chirico, A., Gaggioli, A., Zhang, J. W., & Keltner, D. (2019). The development of the Awe Experience Scale (AWE-S): A multifactorial measure for a complex emotion. *The journal of positive psychology*, *14*(4), 474-488.
[124]. Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Springer.
[125]. Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, *35*(9), 2817-2842.
[126]. Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111-151.
[127]. Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice*, *38*(2), 14-26.
[128]. Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439.
[129]. Zhou, X., Yang, L., Fan, X., Ren, G., Yang, Y., & Lin, H. (2021). Self-training vs pre-trained embeddings for automatic essay scoring. China Conference on Information Retrieval,