Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 9, 103686 (2026)
https://doi.org/10.29140/lea.v9.103686

Introducing Sentinel Assessment: An AI-Powered Functionality for Evaluating Writing Competence in French as a Foreign Language

ABDELGHANI ES-SARGHINI[a]
ABDELAZIZ BOUMAHDI[b]

[a]*Faculty of Educational Sciences, Mohammed V University, Rabat, Morocco*
abdelghani_essarghini@um5.ac.ma

[b]*Faculty of Educational Sciences, Mohammed V University, Rabat, Morocco*
a.boumahdi@um5r.ac.ma

Abstract

The evaluation of writing competence in French as a Foreign Language remains a complex pedagogical challenge, traditionally constrained by high inter-rater variability and significant time burdens. The emergence of Artificial Intelligence (AI) offers an opportunity to shift the paradigm, yet empirical evidence regarding its reliability remains limited. This research addresses the tension between the limitations of traditional manual assessment and the potential of automated solutions to enhance formative practices.

Adopting a Design-Based Research approach, this research developed *"eCorrige,"* a mobile ecosystem powered by AI designed to analyze student-written compositions based on the criteria of relevance, coherence, cohesion, and linguistic proficiency. The performance of *eCorrige* was rigorously evaluated by comparing the assessment data from two groups of evaluators: three expert human teachers and three repeated iterations of the automated system, applied to a corpus of sixty-two compositions produced by students in the sixth grade, the final year of elementary education in Morocco, totaling 3,740 tokens and 1,118 unique types.

Paired-sample t-tests revealed a marked contrast in reliability. While human evaluators exhibited significant heterogeneity and high standard deviations, *eCorrige* demonstrated superior stability and consistency across repeated assessment trials. Statistical tests confirmed a significant concordance between human and automated scoring on objective criteria such as errors and structure, revealing that the automated system applies stricter linguistic standards in the assessment of qualitative language

Data Availability Statement: *All relevant data are within this paper.*

mastery. These findings suggest that the algorithm models a "conservative" expert rater, offering a standardized complement to human evaluation. Finally, this research proposes the *"sentinel functionality,"* a metaphorical concept where AI serves not just as a grader, but as a predictive tool, which forecasts potential learning trajectories and enables anticipatory remediation.

Keywords: Artificial Intelligence, eAssessment, formative assessment, writing competence, sentinel assessment, French as a Foreign language

Introduction

The evaluation of writing competence has traditionally been a substantial pedagogical challenge (Cumming, 2013). This challenge can be explained by the multiple formal, creative and communicative facets that writing involves (Hayes & Flower, 1980). The difficulty of evaluation amplifies when writing is considered not as a product to be measured, but as a process of active construction of meaning (Bereiter & Scardamalia, 2013; Hayes & Flower, 1980).

The ongoing pursuit of more equitable assessment systems (Bloom et al., 1956; Piéron, 1963) has gradually evolved toward a focus on pedagogical fairness. Within this evolution, socioconstructivist approaches have redefined evaluative actions as interactive and formative processes anchored in meaningful contexts (Perrenoud, 1991). This perspective prioritizes continuous construction of meaning over the traditional view of evaluation as a static measurement of performance (Allal, 1979; Scriven, 1966). However, logistical constraints inherent in manual grading often obstruct the full realization of the socioconstructivist benefits. Algorithmic technologies address this challenge by scaling these pedagogical principles through the implementation of automated real-time feedback loops (Shermis & Burstein, 2013; Zawacki-Richter et al., 2019). Admittedly, intelligent technological systems can offer novel solutions to traditional assessment constraints, such as the inter-corrector variability (Eckes, 2008) or the cumbersomeness related to the individualization of feedback (Bennett, 2015). Shermis & Burstein (2013) posit that these tools also reshape the evaluative process by questioning our ability to uphold constructivist principles (Zawacki-Richter et al., 2019), leading to the primary inquiry of how these systems can facilitate the construction of learning rather than merely serving as instruments for external measurement (Black et al., 2002).

The main problem of the present research paper arises from the conflict between the enduring constraints of traditional methodologies (with their subjectivity and time constraints) and the potential benefits of digital transformation. Such a problem is particularly acute in the Moroccan education system. Despite technological modernization initiatives, such as the "GENIE" program operating since 2005 (Nejjari & Bakkali, 2017), the assessment of complex skills remains challenging. At the same time, while AI and advances in language processing pave the way for instant, personalized feedback, their integration raises fundamental questions concerning pedagogical legitimacy, stakeholder buy-in, and socio-cultural appropriateness. Poorly designed automation could worsen existing inequalities or turn assessment into a purely technical process, which would hurt its important role in learning.

The focus of the present research is situated at the crossroads of these challenges; the research aimed at identifying how AI can help with the formative eAssessment of writing competence in FFL. This research was guided by the following question:

- What are the benefits of an AI-based system for formative eAssessment of writing competence in French as a foreign language in Moroccan primary schools?

To answer this question, the researchers developed an automated eAssessment system (*eCorrige*) for primary schools (Es-sarghini & Boumahdi, 2025). The conception of this system was based on an analysis of the educational contributions of AI models, linking the technical, educational, and communicational dimensions of automated assessment of writing competencies.

Literature Review

The Assessment of Writing Competence

Assessment is defined as a methodical and rigorous process of data collection. Its principal goal is comprehensively evaluating the spectrum of a learner's achievements, which encompasses factual knowledge and procedural skills as well as affective and motivational dispositions (Zawacki-Richter et al., 2019; Allal, 1979; Bloom et al., 1956). Within this framework, two primary assessment modalities are distinguished, each characterized by distinct objectives and temporal parameters. Formative assessment, the first category, is intrinsically integrated into the instructional cycle. It constitutes an iterative process, conducted continuously to monitor and support learner development. The main goal of formative assessment is to keep track of how well the learner is doing. This lets teachers change how they teach on the spot, which helps students stay on track with their learning. Formative assessment is very helpful because it can teach students how to think critically. It encourages individuals to think about what they did and ask questions by giving them regular and specific feedback. It teaches them how to find and rectify their own mistakes. Therefore, it facilitates deeper cognitive engagement and fosters learner autonomy (Black et al., 2002).

Summative assessment, on the contrary, occurs less frequently. It usually takes place at the end of the learning process to certify what has been learned. The main goal of this assessment is to give a general overview of what the student knows and can do at a certain point in time. Educators often do this to check or prove that students have learned something (Scriven, 1966).

The difficulty of evaluating writing competence makes it stand out in the field of learning assessment. While declarative knowledge can be quantified through objective metrics, the assessment of writing competence is inherently multidimensional. This is due to the intrinsic complexity of writing competence, which is manifested in the combination and the mobilization of various dimensions. First, we have linguistic dimensions that are related to mastery of the language system. Second, cognitive dimensions, which include planning, organization, and reasoning processes. Then, social dimensions relate to the consideration of the communication context and the audience characteristics. Finally, rhetorical dimensions that concern the ability to adapt the discourse to the desired objectives and effects (Cumming, 2013; Hayes & Flower, 1980). Hence, the assessment of this complex competence cannot be easily done objectively.

To adequately capture the multidimensional nature of writing competence, the assessment framework must be structured around specific criteria. In this regard, it is necessary to distinguish between two fundamental categories. Minimum criteria constitute the essential requirements for competency certification; these encompass the communicative relevance of the text, internal textual coherence, and the mastery of linguistic resources. Conversely, enrichment criteria are not mandatory for basic certification. These include the quality of formal presentation, and the originality of the ideas expressed. Consequently, while not pre-requisites for the minimum threshold of success, these attributes are considered desirable and valued indicators of competent performance (De Ketele & Gerard, 2005).

As documented by Knoch (2011) and Eckes (2012), the assessment of writing competence can be biased even when adopting standardized and explicit criteria. This is generally due to the subjectivity of teachers' assessment practices. There is always some subjectivity involved in how human judges use and understand the assessment criteria (Eckes, 2012). This possible subjectivity is always a problem for the reliability and validity of written work assessment.

The Advantages of eAssessment

Technology can be utilized to create evaluations that are interactive and customized to the individual (Bennett, 2015), redefining teacher assessment of writing. Assessment facilitated by technology can evaluate intricate competencies, such as problem-solving, creativity, and critical thinking, that are often challenging to measure through traditional assessments. Also, technology-facilitated assessment is capable of providing students with quick and accurate feedback. This feedback is regarded as a significant educational benefit, as it markedly enhances learners' active engagement and improves learning outcomes (Bennett, 2015). From this viewpoint, utilizing Information and Communication Technologies (ICT) in assessments enables ongoing, personalized tracking of each student's progress. This approach is operationalized by harvesting student digital traces, specifically linguistic performance metrics such as lexical frequency profiles and syntactic complexity indices, captured throughout their developmental trajectories. The systematic analysis of these indicators enables a precise evaluation of individual learning progress. Therefore, educational pathways are differentiated, facilitating an individualized and adaptive instructional experience for students (Bennett, 2015).

The Contributions of AI to eAssessment

Beyond scoring, AI systems can also generate qualitative formative feedback, providing educators with specific observations regarding learner achievements and actionable instructional guidance to facilitate pedagogical remediation (Es-sarghini & Boumahdi, 2025). Pupils can develop greater capacity for autonomous learning through iterative feedback loops (Black et al., 2002; Jiang et al., 2023). For instance, Intelligent Tutoring Systems (ITS) analyze student performance metrics to forecast future learning outcomes with increasing precision. By predicting these trajectories, educators can anticipate potential learning obstacles before they manifest. Consequently, real-time pedagogical interventions can be implemented to dynamically adjust instructional strategies and optimize the learning trajectory through personalized pathways. Within this framework, task difficulty is automatically calibrated, and targeted remedial resources are provided to the learner to preclude academic failure (Butz et al., 2006).

Additionally, AI-mediated assessments facilitate more precise and expedited grading, mitigating human bias and promoting greater equity in the grading process. The automation may also alleviate the administrative burden on educators, freeing up time that can be redirected toward more cognitively demanding pedagogical tasks (Shermis & Burstein, 2013). Moreover, AI possesses the capability to identify at-risk students to deliver targeted and adaptive educational support. Consequently, it facilitates proactive intervention to prevent the accumulation of learning deficits, fostering success for a broad spectrum of learners (Khan et al., 2021).

To enable such analysis and intervention, Educational Data Mining (EDM) and Learning Analytics (LA) provide frameworks that may enhance assessment practices. By analyzing digital traces, these techniques provide a clearer understanding of learning processes. They also help identify long-term patterns and the key factors driving academic success. Consequently, these methodologies can improve educational outcomes by allowing educators to adapt their teaching strategies and learning environments (Romero & Ventura, 2013).

Methodology

Research Objective

This research seeks to identify operational solutions to the concrete educational challenges delineated in the problem statement. As previously articulated, these challenges encompass the temporal constraints faced by educators, the inherent subjectivity of assessment, and the scarcity of formative feedback available to learners of FFL writing competence within Moroccan primary schools.

The methodological approach adopts a constructive Design-Based Research (DBR) framework (Design-Based Research Collective, 2003) driven by a dual objective. First, the study aims to develop an operational technological innovation – the *eCorrige* ecosystem – specifically engineered to meet the professional requirements of educators. Second, it seeks to generate empirical knowledge regarding the implementation of *eCorrige* and its pedagogical effectiveness in the formative eAssessment of writing competence. The ultimate goal is the deployment of this solution in a real-world context to transform identified assessment barriers into constructive learning opportunities.

Research Method

To achieve the research objectives, a Research and Development (R&D) methodology was adopted. The study employs a sequential mixed-methods design, an approach that, according to Creswell and Plano Clark (2017), ensures both practical relevance and scientific rigor within R&D contexts. Furthermore, this mixed-methods framework was selected to ensure robust data triangulation. The quantitative phase involved statistical performance analyses, enabling an objective assessment of the system's effectiveness via consistency measurements and revealing scoring trends in the evaluation of written compositions. Concurrently, qualitative data were processed via content analysis using a benchmarking grid, allowing for the identification of AI's specific contributions and opportunities within formative eAssessment. This triangulation strategy aimed to construct a formative eAssessment model and design a corresponding AI-based digital ecosystem.

The resulting system is constituted by two interconnected Android applications. First, *"eCorrige Prof"* functions as an educational management tool enabling teachers to design assessment activities. Through this interface, educators can define assessment criteria, assign relative weights, and monitor learner progress. Second, *"eCorrige Élève"* serves as a dedicated writing environment that provides immediate, personalized feedback to the learner. It facilitates the instant analysis of written work to scaffold the improvement of the assessed competencies.

Design Tools and Research Instruments

The Android Studio integrated development environment (IDE) was selected for its versatility and widespread adoption. To facilitate data collection, a comprehensive benchmarking grid was developed to systematically analyze and compare distinct AI models. Furthermore, the *eCorrige* applications served as the primary instruments for experimentation and the acquisition of usage data. To ensure scientific rigor, all developed instruments underwent validation by a panel of expert university professors and educational inspectors.

Design Process and Data Collection

The design was structured in three main phases:

> *Initial design:* Design and development of the *eCorrige* system, based on prior analysis of needs and teaching principles.

Implementation: Deployment of the system in a real-world context with 62 sixth-grade students (divided into three classes supervised by a teacher).
Performance analysis: Correction of work by the system, followed by descriptive statistical analysis conducted using the open statistical software JASP (JASP Team, 2025) in order to evaluate the effectiveness of the model and identify areas for improvement.

This approach made it possible to produce and test a functional prototype of the *eCorrige* system (Figure 1).

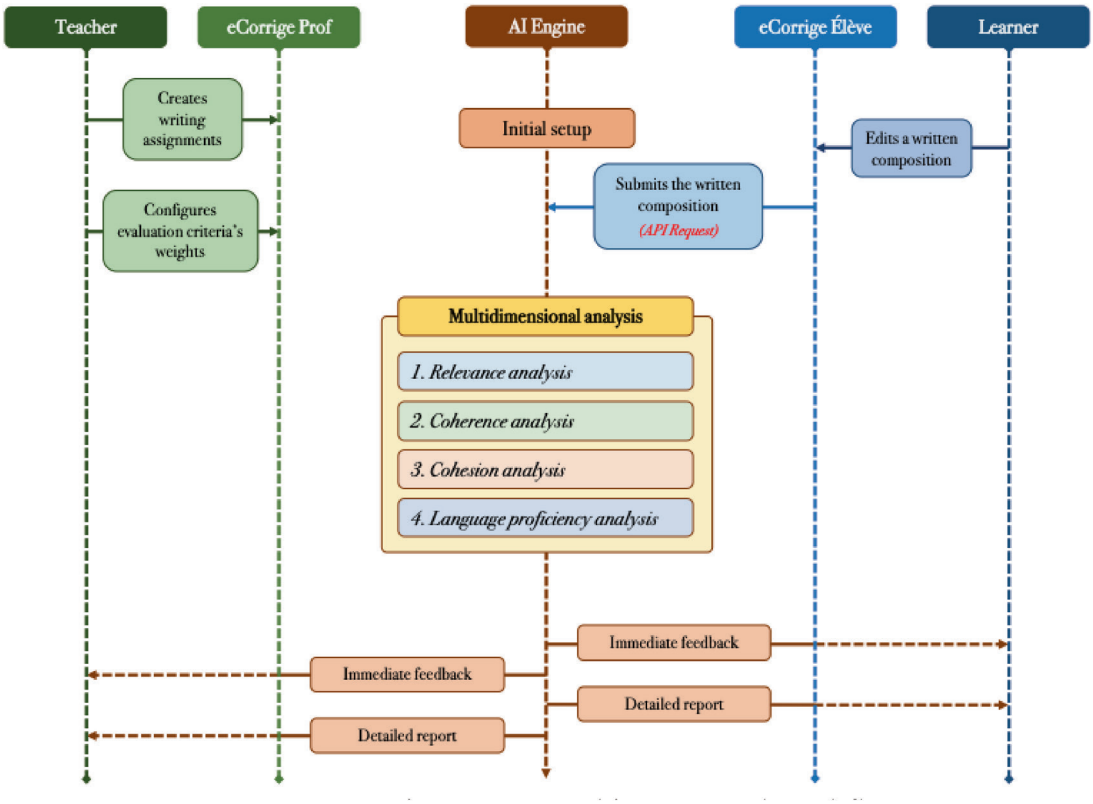


Figure 1 *eCorrige system architecture and workflow.*

The Scoring Procedure

The scoring procedure followed a standardized protocol for both human and automated evaluators. Each of the 62 compositions was independently assessed by three expert teachers and processed through three iterations of the *eCorrige* system. Scores were assigned on a scale of 0 to 10, using 0.25-point increments (e.g., 5.00, 5.25, 5.50). This granular scale was applied across the four criteria defined in the assessment grid: relevance, coherence, cohesion, and linguistic proficiency.

Results and Discussion

Turning Theory into a Functional eAssessment Model

AI tools, particularly those utilizing advanced language models, demonstrate increasing precision in analyzing the linguistic accuracy of written work. These systems can identify errors across lexical,

orthographic, syntactic, and grammatical dimensions. Furthermore, they provide rapid, relevant feedback on formal aspects of writing competence, such as sentence construction, idea structuring, and adherence to linguistic conventions. When integrated with human assessment, this automated feedback may contribute to the development of writing competence. Facilitating error awareness is pivotal for mastering the competencies required for effective writing (Jiang et al., 2023). Table 1 summarizes the contributions of ICT and AI to writing assessment, synthesized from an analysis of three prominent technological models: Intelligent Tutoring Systems (ITS) (Butz et al., 2006), Automated Essay Scoring Systems (AESS) (Hussein et al., 2019), and Educational Data Mining (EDM) (Romero & Ventura, 2013).

Table 1 *Summary of the contributions of some technological models to the assessment of writing competence*

Contributions	ITS	AESS	EDM
Immediate and targeted feedback	×		×
Continuous and individualized progress tracking	×	×	×
Intelligent automation of assessment	×	×	×
Rich and personalized feedback	×		×
Accurate prediction of learning trends			×
Instant anticipation of difficulties			×
Detailed linguistic analysis of writing		×	

The ITS model prioritizes personalized learner support through four core features: immediate and targeted feedback, continuous and individualized progress tracking, intelligent assessment automation, and comprehensive, tailored guidance.

The AESS exhibits a more specialized profile. While it shares the capabilities of continuous progress monitoring and intelligent automation with other models, its distinguishing feature is the capacity for detailed linguistic analysis of written work. This specialization renders it particularly effective for evaluating written compositions in scenarios necessitating comprehensive language analysis.

EDM represents the most comprehensive model, featuring six distinct capabilities. Beyond the standard features shared with other models, EDM is distinguished by its predictive analytics. It not only provides immediate feedback and personalized responses but also uniquely forecasts learning trends and anticipates difficulties. This configuration establishes it as an instrumental tool for proactive and preventive pedagogical support.

Our research focuses on synthesizing these three models (ITS, AESS, and EDM) to construct an innovative AI-driven eAssessment system (Figure 5). This integrative approach leverages the strengths of each model while mitigating its respective limitations.

The proposed framework is distinctive in its consolidation of diverse digital assessment functionalities into a unified system. It integrates the personalized feedback and monitoring of ITS, the in-depth linguistic analysis of AESS, and the predictive power of EDM. Each functionality has been optimized to ensure maximum efficiency in assessment. The primary innovation lies in the synergistic integration of seven key features: instant feedback, personalized tracking, intelligent automation, detailed assessment

reporting, learning trend prediction, difficulty anticipation, and comprehensive linguistic analysis. This combination enables a rigorous, accurate, and contextualized evaluation of writing competence. Therefore, this system offers a comprehensive assessment solution that addresses contemporary educational needs and contributes to the advancement of methodology in digital learning assessment.

The Assessment Automation Framework

De Ketele and Gerard (2005) established four fundamental minimum criteria that serve as the bedrock of the *eCorrige* assessment framework. The first criterion, relevance, evaluates the alignment of the written production with the assigned task and its specific communicative context. The second, coherence, assesses the logical organization of ideas and the thematic unity of discourse, ensuring a fluid transition between concepts. Cohesion, the third criterion, focuses on the structural links within the text, examining the effective use of logical connectors and referential mechanisms, such as pronominalization and anaphora. Finally, linguistic proficiency examines the student's adherence to orthographic, grammatical, lexical, and syntactic standards. This multidimensional approach ensures that the inherent complexity and multifaceted nature of writing competence are captured comprehensively.

Presentation of the Developed Techno-pedagogical Solution

Technical Architecture and Educational Objectives

The *eCorrige* system represents a novel mobile ecosystem leveraging AI to facilitate the formative assessment of writing competence within FFL elementary contexts. Designed as a techno-pedagogical intervention, this system mitigates the logistical and pedagogical challenges associated with manual evaluation by providing a unified framework comprising two interdependent mobile applications developed natively for Android using XML and Java. The primary interface, *"eCorrige Prof"* (Figure 2), empowers educators to design assessment activities, configure evaluation parameters, and systematically monitor student performance.

The second application, designated as *"eCorrige Élève"* (Figure 3), serves as a dedicated student interface designed to scaffold the writing process and facilitate the reception of formative feedback. The objectives of this technological solution are twofold: primarily, to automate the analysis and assessment of written productions, thereby mitigating the administrative burden on educators; and secondarily, to preserve and enhance the formative dimension critical to language acquisition. By delivering immediate, personalized, and actionable feedback on submissions, the system effectively operationalizes a learner-centric pedagogical approach.

The Automated Grading Framework

The application uses a process automated by AI to grade and provide scores related to each criterion after analyzing written work. As shown in Figure 4, the process of eAssessment is based on four minimum criteria that are clearly defined and known to be important for judging the quality of a piece of writing:

a. *Relevance criterion* refers to the evaluation of the extent to which the generated content aligns with the specified topic and the task of the assignment. This is done by examining whether the student had correctly executed the instructions and whether they had followed the formal and thematic aspects of the writing task.
b. *Coherence criterion* examines whether the ideas in the text are logically organized and whether the composition flows well.

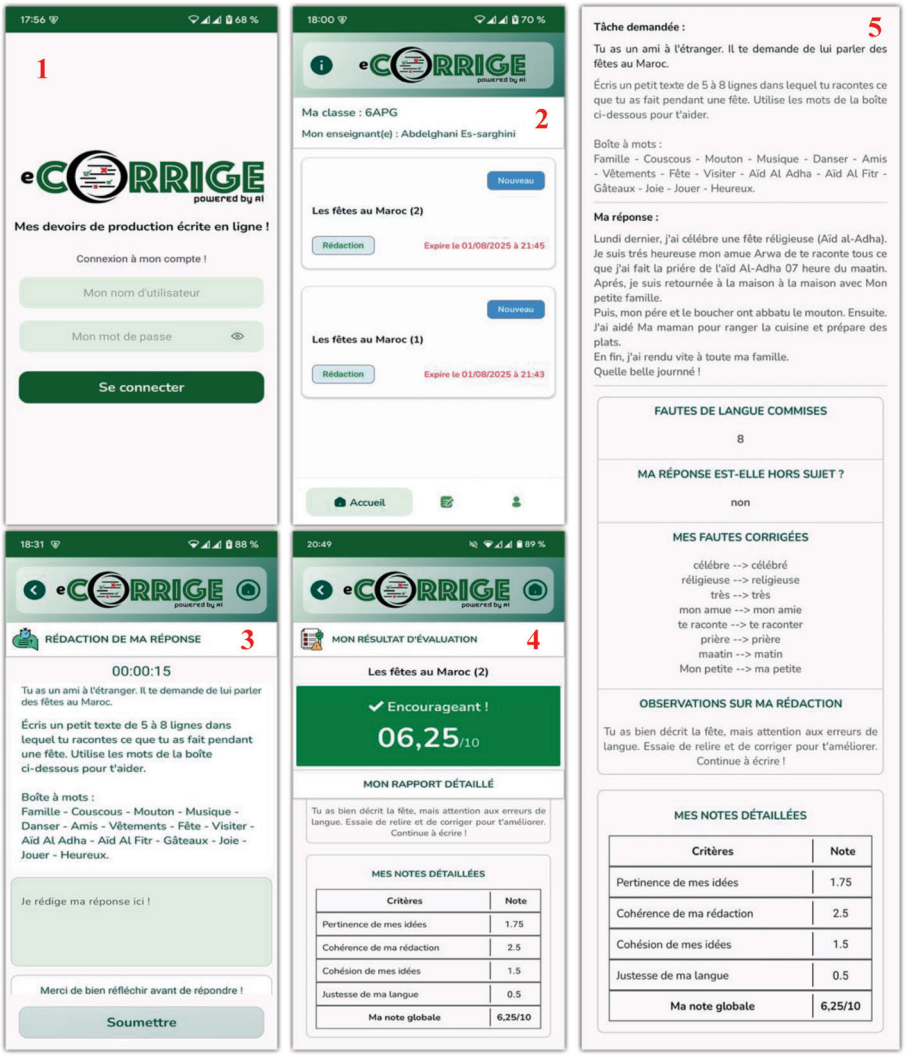
This collage illustrates the complete eCorrige ecosystem. It begins with secure teacher login (screenshot 1) and a statistical dashboard (screenshot 2), moves to the assignment configuration interface where grading rubrics are defined (screenshot 3), and concludes with the automated generation of detailed criterion-referenced feedback reports for students (screenshots 4 and 5).

Figure 2 *User interfaces for the "eCorrige Prof" application.*

c. *Cohesion criterion* checks whether the learner uses proper logical connectors and reference mechanisms (such as pronominalization and anaphora) to enhance the overall understanding of the writing.
d. *Linguistic proficiency* criterion is used to assess the correctness and the precision of written expression. It aims to analyze whether grammar, spelling, vocabulary, and syntax aspects are sufficiently proficient and appropriate.

The Mathematical Model Implemented

Using a mathematical model (Figure 5), the *eCorrige* system grades written work in a consistent way. It does this by using a vector to show how good a student is at writing, which is formulated as $S = [G, P, C_1, C_2, L]$. In this model, every aspect is put on a standard scale where scores can range from 0 to 10 in quarter-point increments (0.00, 0.25, 0.50, 0.75, etc.). According to De Ketele and Gerard (2005), a score of 5/10 is the lowest passing grade.



This sequence demonstrates the learner's experience within the eCorrige app. Students log in (screenshot 1) to view assigned tasks (screenshot 2), utilize a scaffolded writing interface with built-in vocabulary aids (screenshot 3), and receive instant granular feedback upon submission to facilitate self-regulated learning (screenshots 4 and 5).

Figure 3 *User interfaces for the "eCorrige Élève" application.*

There are five components of the vector: G stands for the overall score for the task at hand, P stands for the relevance of the content, C_1 corresponds to the coherence of the text, C_2 refers to the cohesion of the discourse, and L matches the proficiency of the language. The teacher assigns a weight to each of the four assessment criteria when designing an assignment: w for relevance, x for coherence, y for cohesion, and z for language proficiency. These weights must be numbers that total 10 ($w + x + y + z = 10$). To get the final score G, *eCorrige* uses the mathematical formula $G = (P*w + C_1*x + C_2*y + L*z)/10$ to find the weighted average of the four given scores. This way, the final grade reflects both the student's performance in different areas and the teacher's pedagogical priorities as shown by the criterion weightings.

The AI-driven automated assessment algorithm uses the above mathematical framework to compute pupils' scores based on linguistic measurements. These metrics encompass lexical analyses that concentrate on vocabulary richness and terminological precision, syntactic examinations that assess phrase complexity and structural variety, and semantic analyses that evaluate thematic relevance and

Critères	Barème
Pertinance des idées	0
Cohérence de la production	0
Cohésion de la production	0
Utilisation correcte de la langue	0
Total	0/10

This matrix allows teachers to define the specific weighting for each assessment criterion (relevance, coherence, cohesion, and language proficiency), providing a transparent scoring structure that sums to a final grade out of 10.

Figure 4 *Assessment grid used by the eCorrige system.*

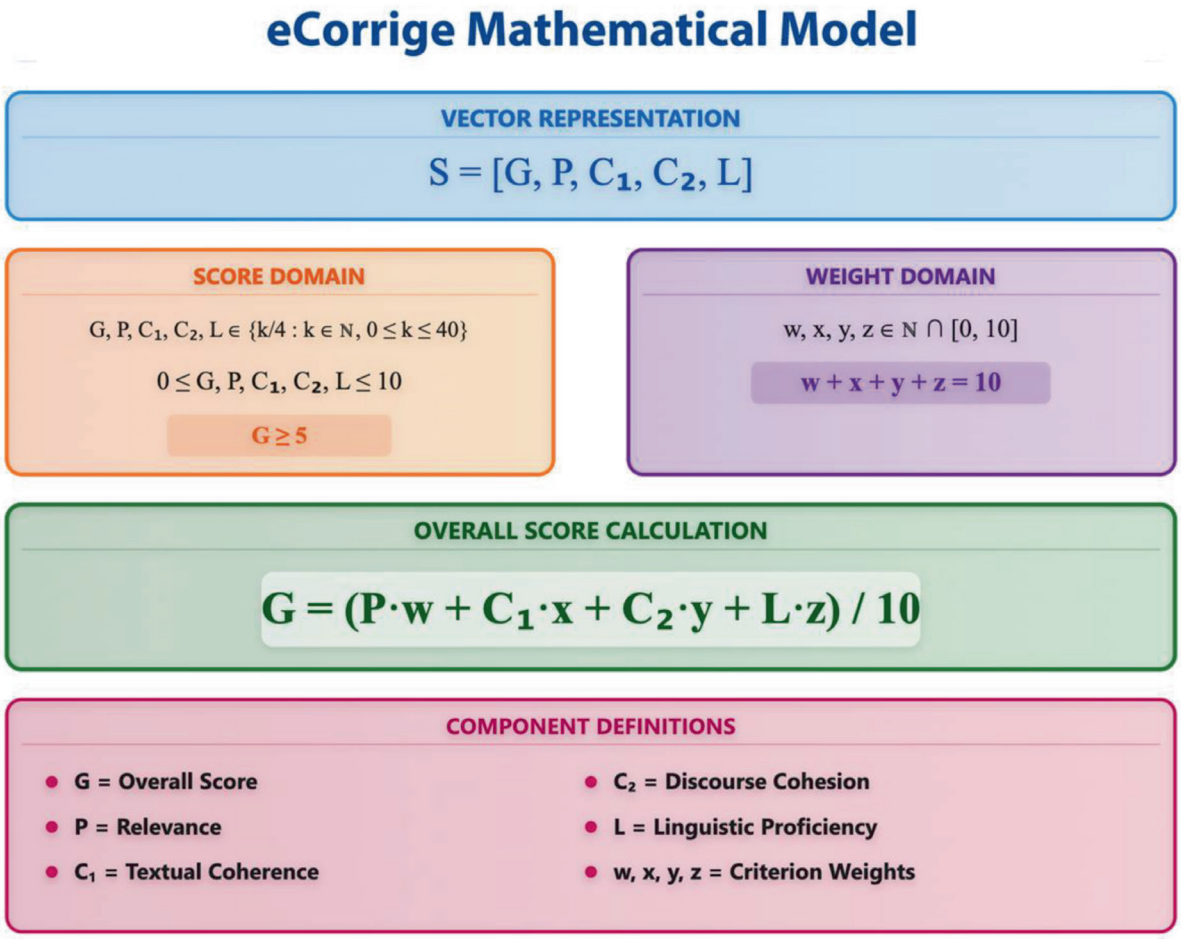


Figure 5 *eCorrige mathematical model.*

argumentative coherence. The AI-based algorithm uses weighted feature extraction to put a developmental framework into action, and it changes its analysis based on how well the learner is doing. It uses lexical metrics (e.g., Lexical Frequency Profile) to measure vocabulary richness, syntactic indices (e.g., Mean Length of T-unit) to measure structural complexity, and semantic measures (e.g., Latent Semantic Analysis) to measure coherence. The system automatically changes the weight of features, putting more emphasis on basic accuracy for beginners and more on linguistic sophistication for advanced learners. This is to better reflect the non-linear path of learning a second language (Crossley & McNamara, 2012; Laufer & Nation, 1995; Lu, 2010).

The Automated Assessment Process

The *"eCorrige Élève"* app gives students a structured digital writing space. There is a clear order regarding the assessment process:

1) *Receiving and executing instructions:* The student gets a writing assignment (Figure 6) that is appropriate for their level, along with specific grading criteria set by the teacher (relevance, coherence, cohesion, and language proficiency).

This dashboard identifies the student's class (6APG) and teacher. Pending tasks appear as cards; here, a new assignment titled 'La pollution de l'air' (Air Pollution) is highlighted with a 'Nouveau' (New) badge. The card provides essential information, including the expiration date, and allows for the launch of the writing environment.

Figure 6 *'eCorrige Élève' dashboard displaying a pending writing assignment.*

2) *Writing and keeping track of time:* A smart timer (Figure 7) accurately keeps track of how long the task took, which lets the writer examine how quickly he wrote later.

The student workspace features a clear prompt with context (in this case, 'Air pollution'), a built-in timer to track writing fluency, and a dedicated input area. It includes a pre-submission nudge ('Think carefully before writing your answer!') to encourage self-regulation before the work is sent for AI analysis.

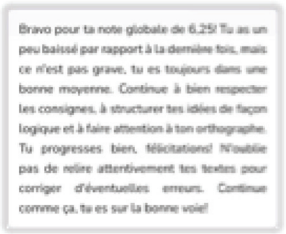
Figure 7 *Timer running when the student starts answering the writing task.*

3) Automated AI analysis: The output is sent to a server along with its evaluation criteria, and an AI engine analyzes the written work quality based on the four defined dimensions, which refer to the minimal criteria developed by Gerard and De Ketele (2005).

Display of Results and Formative Feedback

The result interface represents the formative feedback engine of *eCorrige*. It has a number of pedagogically pertinent features:

1) *Instant visual feedback:* The system rapidly displays overall performance through the use of icons and color coding (Figure 8).

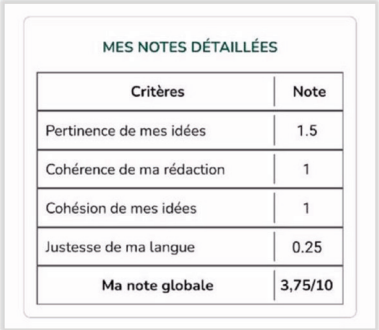


Feedback translated to English:

Well done on your overall grade of 6.25! You've dropped a little since last time, but that's okay, you're still within a good average. Keep following instructions carefully, structuring your ideas logically, and paying attention to your spelling. You're making good progress, congratulations! Don't forget to proofread your texts carefully to correct any potential mistakes. Keep it up, you're on the right track!

Figure 8 *Sample of feedback that students receive.*

2) *Analytical presentation of results:* The interface presents detailed scores for each criterion in written work, elucidating the strengths and weaknesses of the evaluated writing (Figure 9).



MES NOTES DÉTAILLÉES

Critères	Note
Pertinence de mes idées	1.5
Cohérence de ma rédaction	1
Cohésion de mes idées	1
Justesse de ma langue	0.25
Ma note globale	3,75/10

Translation:

MY DETAILED GRADES

Criteria	*Score*
Relevance of my ideas	*1.5*
Coherence of my writing	*1*
Cohesion of my ideas	*1*
Correctness of my language	*0.25*
My overall grade	*3.75/10*

Figure 9 *Example of displaying results for each evaluation criterion.*

3) *Detailed formative feedback:* The system offers comprehensive formative feedback by not only assigning scores but also delivering a detailed inventory of errors alongside suggested corrections, personalized feedback, and recommendations (Figure 10).

Other Features of eCorrige

Using AI for Text Analysis and Feedback Creation

The application is driven by an advanced AI engine (Figure 11) that employs Natural Language Processing (NLP) and Machine Learning (ML) techniques to process evaluation tasks. Upon submission

Translation:

LANGUAGE MISTAKES MADE

14

IS MY ANSWER OFF-TOPIC?

No

MY CORRECTED MISTAKES

tois --> toi
dérniorment --> dernièrement
muslimants --> musulmans
accasion --> occasion
Moutant --> mouton
C'est une grande accasion --> C'est une grande occasion
cette connu --> cette fête est connue
dans la matiné --> dans la matinée
aprés --> après
par la maniére musulmant --> de la manière musulmane
on a sacrifié le Moutant --> on a sacrifié le mouton
acheter --> acheter

OBSERVATIONS ON MY WRITING

It's good that you talk about Moroccan holidays. Pay attention to your spelling and some sentences that could be better phrased. Keep practicing!

Figure 10 *Sample of detailed formative feedback.*

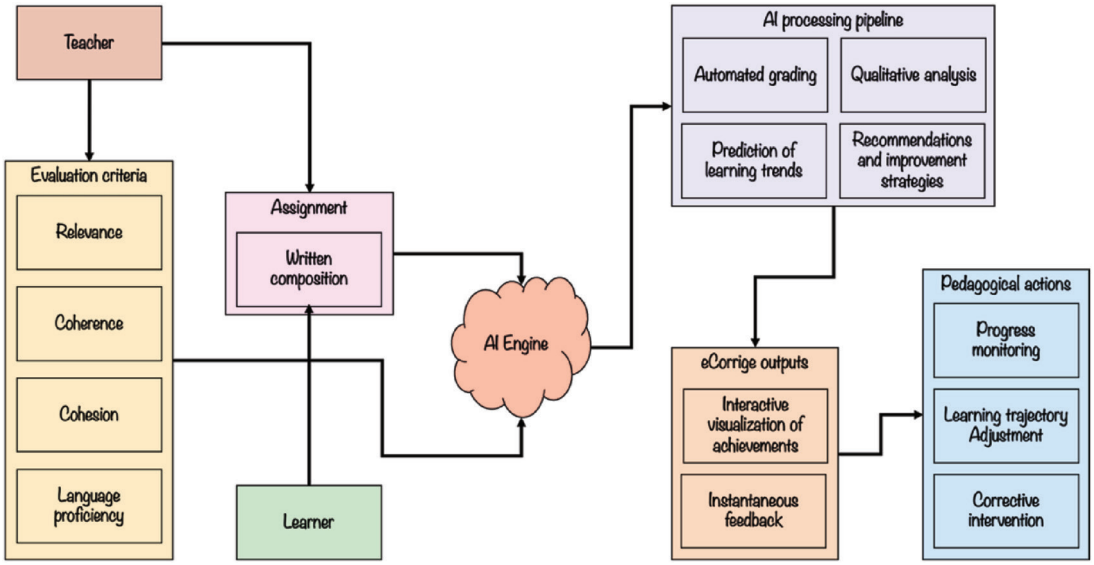


Figure 11 *eCorrige overall eAssessment process.*

of a student's work, the engine initiates a comprehensive textual analysis to diagnose both the strengths and weaknesses of the composition, evaluating it against the four established assessment criteria. Subsequently, the system automatically assigns a proficiency score to each dimension of writing competence. These quantitative assessments are derived from the algorithmic analysis of the text and the pre-configured weighting assigned to each criterion.

The integration of this quantitative assessment mechanism with qualitative analysis facilitates a robust and objective overall evaluation. Finally, the system delivers automated, personalized feedback to

both the learner and the educator. Transcending traditional numerical grading, this feedback comprises specific, actionable commentary provided in real-time. It offers explicit pedagogical guidance, precisely identifying areas requiring remediation and prescribing concrete steps for competency development.

Differentiated Feedback for Students and Teachers

Complementing the review report, which equips the student with immediate and actionable feedback (Figure 12), the assessment model implements differentiated feedback mechanisms tailored to the specific recipient (Figures 13, 14, and 15). This framework clearly distinguishes between student-oriented formative guidance and teacher-oriented diagnostic analytics, ensuring that each stakeholder receives data relevant to their specific role in the learning process.

The system's feedback mechanism distinctly highlights both individual and collective achievements while identifying specific areas for the development of competency. It provides learners with actionable and context-specific advice to enhance their writing proficiency. Conversely, instructor-oriented feedback is designed to be more comprehensive and contextually relevant; it incorporates individual student data alongside aggregated class-level analytics to offer a longitudinal view of progress. This dual-layered approach provides a granular profile of current performance and forecasts potential learning trajectories, facilitating the implementation of proactive strategies to mitigate future academic challenges.

Interactive Way to Track Learner and Class Progress

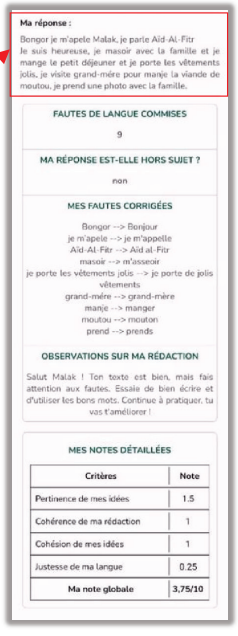
The application incorporates an Interactive Curves Graph (ICG) module (Es-sarghini & Boumahdi, 2025), designed to facilitate the interpretation of assessment data and enhance educational monitoring (Figure 16). This module provides a comprehensive longitudinal visualization of performance evolution, tracking changes for both the entire class and individual students across each assessment criterion. Through these interactive graphical representations, educators can rapidly discern general trends, forecast learning trajectories, diagnose recurrent difficulties, and monitor individual developmental paths. Furthermore, the system derives targeted pedagogical recommendations and instructional adjustments from these visual analyses, with the ultimate objective of enhancing the quality of written production.

Here is the literal translation of the student's written text, keeping the sentence structure and grammatical oddities as close to the original French as possible:

"Hello my name is Malak, I speak Aïd-Al-Fitr
I am happy, I sit myself with the family and I eat the breakfast and I wear the pretty clothes, I visit grandmother to eat the meat of sheep, I take a photo with the family."

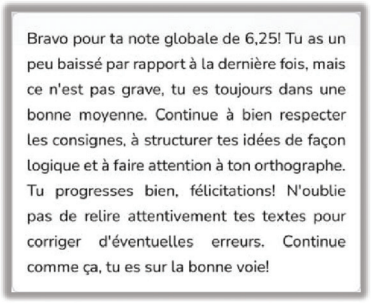
Notes on the student's errors (for context):

- ***a.*** *"Bongor": Misspelling of Bonjour (Hello).*
- ***b.*** *"Je parle Aïd-Al-Fitr": She wrote "I speak Aïd-Al-Fitr" instead of "I speak about (je parle de) Aïd-Al-Fitr."*
- ***c.*** *"je masoir": Phonetic misspelling of je m'assieds (I sit down).*
- ***d.*** *"manje": Phonetic misspelling of manger (to eat).*
- ***e.*** *"moutou": Misspelling of mouton (sheep/mutton).*
- **f.** *Contextual error: The student mentions eating sheep (viande de mouton), which is traditionally associated with Aïd-Al-Adha (The Festival of Sacrifice), not Aïd-Al-Fitr (The end of Ramadan), which she mentions in the first line.*



A comprehensive review screen provides the student with immediate and actionable feedback. It includes an error counter, a line-by-line correction list (pairing errors with their corrections), personalized qualitative advice, and a detailed grading rubric breaking down the overall score by pedagogical criteria.

Figure 12 *Detailed review report provided to the student and the teacher for each answered assignment.*



Translation:

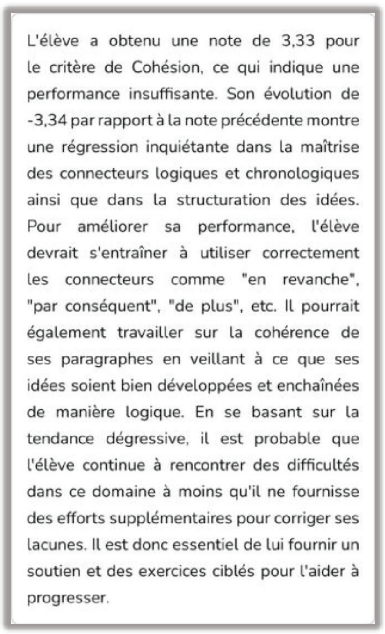
"Well done on your overall score of 6.25! You dropped slightly compared to last time, but it's not a big deal; you are still maintaining a good average. Keep following the instructions well, structuring your ideas logically, and paying attention to your spelling. You are making good progress, congratulations! Don't forget to proofread your texts carefully to fix any potential mistakes. Keep it up, you're on the right track!"

Figure 13 *Example of feedback delivered to a student about his own performance.*

Translation:

"The class has an average score of 5.15 for the 'Global Score' criterion, which is slightly higher than the passing average of 5. However, the average change of -0.01 compared to the previous average shows stagnation or even a slight regression in student performance. To improve this situation, it is important to emphasize the quality of writing by encouraging students to be more precise, to structure their ideas more clearly, and to work on the organization of their texts. It is also crucial to reinforce linguistic proficiency by insisting on grammatical correctness, conjugation, and spelling. Looking ahead, if students continue on this trend, it is possible that their performance will continue to stagnate or even decrease. It is therefore essential to intervene quickly to reverse this trend."

Figure 14 *Example of feedback delivered to a teacher about a whole class performance.*

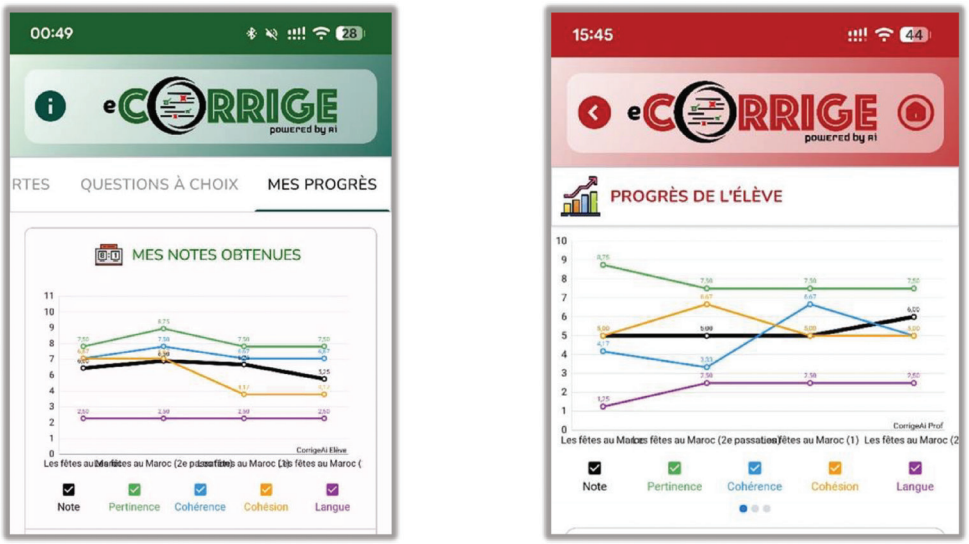


Translation:

"The student obtained a score of 3.33 for the Cohesion criterion, which indicates insufficient performance. Their change of -3.34 compared to the previous score shows a worrying regression in the mastery of logical and chronological connectors as well as in the structuring of ideas. To improve their performance, the student should practice using connectors correctly such as 'on the other hand,' 'consequently,' 'moreover,' etc. They could also work on the coherence of their paragraphs by ensuring that their ideas are well-developed and linked in a logical manner. Based on the downward trend, it is likely that the student will continue to encounter difficulties in this area unless they make extra efforts to correct their deficiencies. It is therefore essential to provide them with support and targeted exercises to help them progress."

Figure 15 *Example of feedback delivered to a teacher about a student's individual performance.*

The tool presents a dynamic graphical interface that provides immediate insight into writing performance trends for both students and instructors. Leveraging AI capabilities, users can access detailed progress or regression reports by interacting with specific data points on the graph (Figures 13, 14, and 15). Beyond historical data, these reports generate predictive analytics regarding future performance and prescribe targeted remedial strategies. For instance, the AI diagnoses specific areas for student improvement (e.g., syntactic refinement) and suggests corresponding pedagogical interventions for the teacher (e.g., reinforcing specific grammatical concepts during instruction).



The system provides mirrored dashboards for both teachers (Red) and students (Green). Both views utilize detailed line graphs to track longitudinal progress across five distinct assessment criteria: overall score, relevance, coherence, cohesion, and language proficiency. An AI-driven feedback is automatically generated when a point on the line is clicked.

Figure 16 *Interactive visualization of progress in eCorrige Élève and eCorrige Prof.*

The assessment process is inherently iterative, comprising a continuous cycle of text production, automated evaluation, and feedback. Aggregating these results into a longitudinal history enables the visualization of competency development over time for both individual learners and the entire class. Consequently, educators can implement data-driven pedagogical interventions, adjusting instructional parameters, such as writing prompts or assessment scales, and determining the appropriate level of scaffolding required (e.g., tutoring, targeted exercises, or personalized guidance).

Statistical Analysis of Inter-Rater Variability

Descriptive Statistics

The analysis was performed on a corpus of 62 written compositions, representing a total linguistic volume of 3,740 tokens and 1,118 unique types. Each composition was subjected to six ratings: three assessments by human evaluators (Teachers 1, 2, and 3) and three iterations by the automated *eCorrige* system (AI1, AI2, and AI3). This experimental design ensured that every text was assessed six separate times, generating a comprehensive dataset for comparative analysis. Table 2 summarizes the descriptive statistics for these assessments.

The analysis of central tendencies reveals a marked contrast in scoring patterns. The human evaluators exhibited significant heterogeneity, with mean scores ranging from a stringent 4.60 (Teacher 2) to a more lenient 5.48 (Teacher 3). Furthermore, the elevated standard deviations among the teachers (average $SD \approx 2.32$) indicate a wide dispersion in scoring, reflecting the inherent subjectivity and variability characteristic of traditional manual assessment. In contrast, the *eCorrige* system demonstrated remarkable stability across its three grading iterations ($M = 4.73$, 4.79, and 4.87), characterized by significantly lower dispersion (average $SD \approx 1.28$). This reduction in variance indicates that the automated system minimizes extreme scoring fluctuations, providing a more standardized evaluation. Notably, the automated system's scoring profile aligns most closely with that of Teacher 2, suggesting that the algorithm models a "stringent" human evaluator rather than a lenient one. Regarding the

Table 2 *Descriptive statistics of writing scores across six ratings (N = 62)*

	Teacher 1	Teacher 2	Teacher 3	AI1	AI2	AI3
N	62	62	62	62	62	62
Missing	0	0	0	0	0	0
Average	5.42	4.60	5.48	4.73	4.79	4.87
Median	5.00	4.50	5.75	4.50	4.88	5.00
Standard deviation	2.21	1.81	2.95	1.49	1.12	1.25
Minimum	1	1.00	0.500	0.25	1.75	2.00
Maximum	9	8.00	10.0	7.50	7.0	7.25
Asymmetry coefficient	–0.224	0.0586	–0.213	–0.340	–0.407	–0.356
Standard error asymmetry	0.304	0.304	0.304	0.304	0.304	0.304
Kurtosis	–0.607	–0.762	–1.27	–0.133	0.164	–0.293
Kurtosis of standard error	0.599	0.599	0.599	0.599	0.599	0.599
Shapiro-Wilk W	0.950	0.977	0.930	0.966	0.981	0.973
Shapiro-Wilk p-value	0.013	0.281	0.002	0.083	0.431	0.186

internal trends of the automated system, the slight but consistent rise in mean scores (from 4.73 to 4.87) suggests a gradual convergence towards the human consensus ($M = 5.16$) across iterations, rather than random fluctuation.

Analysis of Systematic Differences

To advance beyond purely descriptive comparisons and establish a rigorous statistical framework for evaluating *eCorrige*, paired-sample t-tests were employed. Before conducting this inferential analysis, verification of underlying statistical assumptions was imperative. As detailed in Table 3, Shapiro-Wilk tests confirmed that the distribution of differences adhered to the normality assumption ($p > 0.05$) for five of the six assessment criteria. This substantial compliance with the Gaussian distribution validates the application of parametric testing and ensures the statistical robustness of the findings presented in this section.

Table 3 *Normality test for paired differences between the means of human and AI scores (N = 62)*

Assessment criterion (Difference: $Mean_{Human} – Mean_{AI}$)	Shapiro-Wilk statistic (W)	Significance (p)	Distribution
Global Score	0.972	0.175	Normal
Pertinence	0.966	0.079	Normal
Coherence	0.967	0.091	Normal
Cohesion	0.989	0.837	Normal
Language proficiency	0.976	0.262	Normal
Error count	0.954	0.021	Slight deviation

This analysis seeks to quantify the precise level of agreement between the human consensus, operationalized as the aggregated mean of evaluations from Teachers 1, 2, and 3, and the automated scoring generated by the *eCorrige* system. Through the direct comparison of these paired data points, the research transcends simple averages to identify systematic discrepancies and assess the system's concurrent validity against the benchmark of expert human judgment. The detailed results of these comparative tests are summarized in Table 4.

Table 4 *Paired samples t-tests comparing human consensus and eCorrige evaluations (N=62)*

Criteria	$Mean_{Human}$	$Mean_{AI}$	t-value	p-value	Cohen's d	Interpretation
Global Score	5.17	4.79	1.953	0.055	0.25	Marginal difference
Pertinence	1.51	1.35	3.595	<.001	0.46	AI is stricter
Coherence	1.57	1.64	–1.034	0.305	–0.13	No significant difference
Cohesion	1.22	1.35	–1.636	0.107	–0.21	No significant difference
Language proficiency	0.87	0.46	9.586	<.001	1.22	AI is much stricter
Error count	11.90	12.03	–0.270	0.788	–0.03	No significant difference

Global Score Analysis

For the global writing score, the analysis reveals a marginally significant difference ($t(61) = 1.953$, $p = 0.055$) between the human consensus ($M = 5.17$) and the automated system ($M = 4.79$). While the p-value approaches the significance threshold, the effect size remains modest (Cohen's $d = 0.25$), indicating a satisfactory level of global convergence. However, the difference in standard deviations ($SD_{Human} = 2.19$ in contrast to $SD_{AI} = 1.22$) highlights a critical distinction: the *eCorrige* system exhibits a "centralizing tendency," avoiding the extremely high and low scores often assigned by human raters. This suggests the AI operates with a higher degree of evaluative caution, resulting in a narrower, more consistent distribution of grades.

Criterion-level Agreement and Bias

The analysis of specific assessment criteria reveals a clear dichotomy in the system's performance. First, *eCorrige* demonstrated a statistical alignment with human experts in criteria related to text structure and mechanics. No significant differences were found for coherence ($p = 0.305$), cohesion ($p = 0.107$), or error count ($p = 0.788$). The results for error count are particularly significant for validation. The automated error detection ($M = 12.03$) was statistically indistinguishable from the human count ($M = 11.90$). This substantiates that the system's core algorithms identify syntax and spelling errors with an accuracy level equivalent to that of expert human graders. Conversely, significant divergences were observed in the more subjective criteria of pertinence ($t(61) = 3.59$, $p < 0.001$) and language proficiency ($t(61) = 9.59$, $p < 0.001$). In both instances, the AI assigned consistently lower scores than the human consensus. The substantial effect size observed for language proficiency (Cohen's $d = 1.22$) suggests the presence of a "severity bias" in the automated system regarding qualitative dimensions. While human raters may differ in their sensitivity to linguistic nuance, that are often influenced by a "halo effect" (Thorndike, 1920), the algorithm adheres to a stricter set of lexical and grammatical standards. Therefore, the system penalizes linguistic complexity issues more heavily than the average teacher.

In summary, the statistical analysis characterizes *eCorrige* as a conservative and highly consistent evaluator. It replicates human judgment with high fidelity in objective tasks (error detection and

structure) but adopts a more rigorous standard for qualitative elements (language proficiency and pertinence). Notably, the system achieves this while significantly reducing the scoring variability ($SD_{AI} \approx 1.28$ compared to $SD_{Human} \approx 2.32$) inherent in human assessment. This suggests that the differences observed are not errors, but rather indications of a standardized, risk-averse algorithmic approach that complements the more variable nature of human evaluation.

Conclusion

The present research explored the integration of AI into the formative eAssessment of writing competence in Moroccan elementary schools. Based on a benchmarking of existing models, we developed *eCorrige*, a novel system comprising two interconnected Android applications (*"eCorrige Prof"* and *"eCorrige Élève"*). The system is founded on validated assessment criteria and an AI-powered core designed to address the inherent complexity of evaluating written compositions. By automating specific grading tasks and customizing assessment pathways, *eCorrige* aims to overcome the variability and logistical challenges associated with traditional pen-and-paper evaluation.

The research posits that AI can transcend the limitations of traditional evaluation through the integration of a *"sentinel functionality."* Serving as a predictive layer within established formative frameworks, this concept shifts the pedagogical focus from reactive correction to anticipatory remediation, empowering educators to address learning gaps before they manifest. Illustrated by the "Interactive Curves Graph" (ICG) module (Es-sarghini & Boumahdi, 2025), this function utilizes algorithmic analysis to identify recurrent patterns in student performance. Much like a sentinel alerting a unit to approaching danger, the system forecasts potential learning trajectories and warns teachers of likely future weaknesses.

The empirical evaluation of *eCorrige* was conducted on a corpus of 62 student compositions, assessed across six ratings. Inferential statistical analysis, validated by normality tests, provided evidence of the system's reliability. Paired t-tests confirmed a significant concordance between human and AI scoring in objective criteria such as error count and structure (coherence and cohesion) ($p > 0.05$). Conversely, significant divergences in qualitative dimensions revealed a "severity bias," indicating that the AI applies a stricter standard for pertinence and language proficiency than human raters. Notably, the analysis quantified the system's ability to significantly reduce scoring variability, offering a stable alternative to the fluctuations inherent in human judgment. These findings indicate that AI-enabled solutions can provide a reliable, standardized complement to teacher evaluation in primary education.

Despite our promising results, this research is subject to certain limitations. First, the sample size was restricted to 62 compositions from a specific demographic (sixth-grade students in Morocco), which may limit the generalizability of the findings to other educational contexts or proficiency levels. Second, the observed *severity bias* in the AI's assessment of qualitative dimensions suggests a need for further algorithmic calibration to better align with the nuances of human judgment. Future research should focus on validating the *eCorrige* system with larger, more diverse corpora and investigating the longitudinal impact of the *sentinel functionality* on student writing proficiency over an extended academic period. Additionally, qualitative inquiries into student and teacher perceptions of automated feedback would provide valuable insights for refining the user experience.

Use of Generative AI

This manuscript utilized *Gemini* to assist in the textual editing, paraphrasing, and refinement of the English language to ensure academic clarity and grammatical accuracy.

Furthermore, the *eCorrige* system described in this research utilizes artificial intelligence algorithms as the primary subject of investigation to generate comparative assessment data (AI1, AI2, and AI3 ratings).

No AI tools were used for the generation of the research hypotheses, the underlying conceptual framework, or the formulation of the conclusions. The intellectual content, data interpretation, and final decisions regarding the manuscript's text remain the sole responsibility of the authors. The final manuscript was reviewed and verified by the listed authors without additional AI input.

References

Allal, L. (1979). Stratégies d'évaluation formative : conceptions psycho-pédagogiques et modalités d'application. In L. Allal, J. Cardinet & P. Perrenoud (Éds.), L'évaluation formative dans un enseignement différencié (pp. 153–183). Peter Lang.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Bereiter, C., & Scardamalia, M. (2013). *The psychology of written composition*. Routledge. https://doi.org/10.4324/9780203812310

Black, P., Broadfoot, P., Daugherty, R., Gardner, J., Harlen, W., James, M., Stobart, G., & Wiliam, D. (2002). *Testing, motivation and learning*. Assessment Reform Group. http://hdl.handle.net/1893/32460

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay Company.

Butz, C. J., Hua, S., & Maguire, R. B. (2006). A web-based Bayesian intelligent tutoring system for computer programming. *Web Intelligence and Agent Systems*, *4*(1), 61–81.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. https://doi.org/10.1080/15434303.2011.622016

De Ketele, J.-M., & Gérard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences [Validation of assessment tests according to the competency-based approach]. *Mesure et évaluation en éducation*, *28*(3), 1–26. https://doi.org/10.7202/1087028ar

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5–8. https://doi.org/10.3102/0013189X032001005

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Es-sarghini, A., & Boumahdi, A. (2025). Interactive eAssessment of writing competency in French as a foreign language: Development and implementation of an AI-enhanced progress monitoring system. *Australian Journal of Applied Linguistics*, *8*(1), 102495. https://doi.org/10.29140/ajal.v8n1.102495

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Routledge. https://doi.org/10.4324/9781315630274-3

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

JASP Team. (2025). *JASP (Version 0.95.4)* [Computer software]. https://jasp-stats.org/

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, *8*(4), 247. https://doi.org/10.3390/languages8040247

Khan, I., Ahmad, A. R., Jabeur, N., & Mahadi, M. N. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, *8*(1). https://doi.org/10.1186/s40561-021-00161-y

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Nejjari, A., & Bakkali, I. (2017). L'usage des TIC à l'école marocaine : état des lieux et perspectives [A review of ICT use in Moroccan schools and future prospects]. *Hermès, La Revue*, *78*(2), 55–61. https://doi.org/10.3917/herm.078.0055

Perrenoud, P. (1991). Vers une approche pragmatique de l'évaluation formative [Towards a pragmatic approach to formative assessment]. *Mesure et évaluation en éducation*, *4*, 49–81.

Piéron, H. (1963). *Examens et docimologie* [Exams and docimology]. Presses Universitaires de France.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Scriven, M. (1966). *The methodology of evaluation*. Social Science Education Consortium.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation*. Routledge. https://doi.org/10.4324/9780203122761

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0