VII Conferencia Internacional de Estudios Humanísticos (CIESHUM)

Taller IV: Enseñanza de lenguas y desarrollo de competencias en el ámbito educativo

How Automated Essay Scoring Prompts Support Assessment for Learning to Improve Essay Writing

Como las rúbricas automatizadas apoyan la evaluación para mejorar la redacción de ensayos

Humberto Miñoso Machado[1], Universidad Central Marta Abreu de Las Villas, Cuba. E-mail: humberto@uclv.edu.cu

Bertha Elena Romero Molina[2], Universidad Central Marta Abreu de Las Villas, Cuba. E-mail: bromero@uclv.cu

Tania Machado Armas[3], Universidad Central Marta Abreu de Las Villas, Cuba. E-mail: taniama@uclv.edu.cu

Abstract: This study investigates the use of Automated Essay Scoring (AES) prompts to enhance essay writing among first-year special English studies students at Universidad Central Marta Abreu de Las Villas. While students have shown progress in essay writing, there is limited data on how they utilize teacher feedback and engage in proofreading. Literature highlights the role of Assessment for Learning (AfL) strategies, including peer and self-assessment, in improving writing through effective feedback. This research explores the potential of AES prompts to support AfL in essay writing.

An action research approach was employed, with data collected through surveys on peer-assessment, self-assessment, and the impact of AES feedback on writing improvement. Initially, feedback was provided using a prompt based on the Automated Student Assessment Prize (ASAP) model. This model was followed by a refined multiple-shot prompt that offered a more detailed quantitative feedback. The study

involved 18 participants enrolled in a course aimed at achieving a B2 level according to the Common European Framework of Reference for Languages (CEFR).

Results show that a well-designed strategy combining ASAP-based prompts and enriched variant prompts positively influenced students' perceptions of AES feedback, particularly within an AfL framework. Students valued peer assessment during proofreading and acknowledged the benefits of diverse feedback elements. This study underscores the importance of AfL strategies for teachers and encourages students to engage in thorough proofreading with enhanced feedback beyond the classroom.

Keywords: Written production; Assessment; Feedback, Proofreading

Resumen: Este estudio investiga el uso de rúbricas para la Evaluación Automatizada de Ensayos (AES) con el fin de mejorar la escritura de estudiantes de primer año de la Carrera de Lengua Inglesa en la Universidad Central "Marta Abreu" de Las Villas. Aunque los estudiantes muestran progreso en la escritura de ensayos, existen datos limitados sobre cómo utilizan la retroalimentación docente y se involucran en la revisión. La literatura destaca el papel de estrategias de Evaluación para el Aprendizaje (AfL), como la coevaluación y autoevaluación, en la mejora de la escritura mediante retroalimentación efectiva. Esta investigación explora el potencial de las rúbricas de AES para apoyar la estrategia AfL.

Se empleó un enfoque de investigación-acción, recolectando datos mediante encuestas sobre coevaluación, autoevaluación y el impacto de la retroalimentación AES. Inicialmente, se proporcionó retroalimentación usando una rúbrica basada en el modelo del Premio ASAP, seguido de otra rúbrica mejorada con solicitudes múltiples para ofrecer retroalimentación cuantitativa más detallada. Participaron 18 estudiantes de un curso orientado al nivel B2 del MCER.

Los resultados muestran que una estrategia bien diseñada, combinando rúbricas basadas en ASAP y variantes enriquecidas, influyó positivamente en las percepciones estudiantiles sobre la retroalimentación AES, especialmente dentro del marco AfL. Los estudiantes valoraron la coevaluación durante la revisión y reconocieron los beneficios de los diversos elementos de retroalimentación. El estudio subraya la importancia de las estrategias AfL para los profesores y fomenta la participación estudiantil en revisiones exhaustivas con retroalimentación mejorada más allá del aula.

Palabras Claves: Producción escrita; Evaluación; Retroalimentación, Corrección de textos

1. Introduction

Traditionally, summative assessment is one of the most followed methods by teachers. Its goal is to evaluate students' learning at the end of an instructional unit by comparing it against some standard or benchmark. However, formative assessment when understood as Assessment for Learning (AFL) can lead to enhance students' learning through the systematic resulting feedback, and without the psychological stressful conditions, summative assessment implies.

1.1 Statement of the problem,

One of the greatest problems teachers have faced in terms of grading the students' essays is the cost of revising lots of writing output during a course. The solution has been the use of engineering platforms where teachers and students can grade their assignments quickly and have immediate feedback. However, the best-automated essay scoring systems require a premium if one wants to dispose of all resources such as plagiarism. The implementations of Large Language Models like ChatGPT 3.5, Perplexity, and DeepSeek have opened the door for the democratization of automated essay scoring for both teachers and students. The Automated Student Assessment Prize (ASAP) project has opened the door to use automated essay scoring departing from the study of its dataset which allowed us to design effective prompts. The implementation of such prompts through an Assessment for Learning (AfL) strategy led to the formulation of the following:

1.2 Research question,

How does feedback generated by Automated Essay Scoring (AES) support the improvement of written production when used as part of Assessment for Learning (AFL) in the first-year course of English language studies at UCLV?

1, 3 Overall objective

To investigate how AES-generated feedback can enhance essay-writing outcomes through AFL strategies in the first-year course of English Language Studies at UCLV.

2. Assessment for Learning

Assessment for learning is an approach, integrated into teaching and learning, which creates feedback for students and teachers to improve learning and guide their next steps. AfL is concerned with maximizing the feedback process to optimize students' learning (Hansen, 2024).

2.1 Feedback

Feedback ranges from the informal (e.g. oral comments given immediately to learners as they think through problems), to more formal (e.g. written feedback given after an end-of-topic test). Feedback requires being informative, goal-oriented, and delivered in a friendly manner to guarantee positive results (Hansen, 2024).

2.3 Peer and self-assessment

Peer and self-assessment are an effective approach to enhance the learning of students. Explicitly teaching students how to assess their own work, and the work of their peers, has many benefits. It promotes student understanding of their learning, and provides opportunities for critical analysis of their own efforts encouraging them to become learners that are more autonomous (Florez & Sammons, 2013).

2.4 Automated essay scoring

As Nathan Thompson explains, automated essay scoring (AES) is an important application of machine learning and artificial intelligence in psychometrics and assessment. In fact, AES predates the widespread use of the terms "machine learning" and "artificial intelligence" as buzzwords. For decades, the field of psychometrics has pioneered such groundbreaking work (Thompson, 2023).

2. Methodology

2.1. Population and sample

First year cohorts are generally composed of approximately 20 students. We have chosen 19 students (13 females and 6 males) this course.

2.2 Instruments for data collection

The study followed a structured methodological approach, beginning with the collection of a minimum of 19 essays for analysis. These essays were then filtered based on text length (approximately ±50 tokens) and their relevance to the assigned writing prompt. To ensure the authenticity of the dataset, essays suspected of being generated by AI large language models (LLMs) were excluded using a Retrieval-Augmented Generation (RAG) detection prompt (Li, Wang, Wang, Hung, Xie, & Wang, 2025).

The assessment phase consisted of two key components: a human-based evaluation and an AI-assisted evaluation. First, participants conducted self-assessments and peer-assessments without AI intervention. Subsequently, they repeated the assessment process using AI-generated prompts to compare outcomes. Following these assessments, a post-study survey was administered to gauge participants' engagement with feedback. The results revealed that 11 out of 18 respondents rarely incorporated peer feedback into their revisions, while 16 out of 18 reported used feedback in some way. However, only 8 out of 18 considered peer assessment a standard practice.

In the final phase, participants employed a structured JSON prompting template to analyze essays based on three linguistic indices: the SYNLE Index (measuring syntactic left-word embedding before the main verb), CELEX (assessing lexical density), and the MTLD (evaluating textual lexical diversity). These metrics served as key differentiators between high- and low-quality essays, providing quantifiable insights into writing proficiency (McNamara, Crossley, & McCarthy, 2010).

2.3 Data analysis process

A dataset of 25 essays, written by 19 students, was filtered based on text length and adherence to the writing prompts. This process resulted in the selection of 18 essays for the study.

The survey aimed to assess the state of AfL (Assessment for Learning) by examining self-assessment, peer assessment, and the role of feedback from automated essay scoring systems. Additionally, the essays were analyzed using computational tools such as Coh-Metrix and other text analyzers to measure linguistic features, including MTLD (Measure of Textual Lexical Diversity), SYNLE (Syntactic Complexity), and CELEX-based metrics.

3. Results and Discussion

The study shows that using automated essay scoring for enhancing the students' essay writing skills has proved a fruitful strategy because both teachers and students save time (Mansour, Albatarni, & Elsayed, 2024). They both can have immediate feedback in different rubrics. For example, Task achievement, organization and coherence, accuracy, etc.

This study addresses key knowledge gaps in self- and peer-assessment as a strategy for organizing assessment for learning in and beyond the classroom. Considering the students' rubric rankings, the writing process requires greater teacher guidance. For instance, after drafting their essays, students should be guided not only in self-revision but also in sharing their work for peer assessment. Effective organization of peer assessment is essential to ensure better utilization of immediate feedback from automated essay scoring systems, facilitated by well-structured prompts.

Summarizing results

From this, it can be concluded that the data collected show moderate results. For instance, more than half of the students say that they sometimes proofread their essays with their classmates. The students also reported how they ranked the automated essay scoring rubrics. The rubrics were ranked in the following descendant order: first,

grammar and mechanics, vocabulary and word choice, prompt and essay content relationship, organization and structure, and last, evidence and examples.

They also reported that they check the feedback provided by the automated essay scoring system; however, not all the students responded they always rewrite their essays following the feedback hints. The students did not deny the importance of peer correction; nevertheless, they insist that this is not a common practice.

Although rewriting essays after receiving feedback was uncommon among students, the immediate feedback generated by the zero-shot prompt (an AI tool providing rubric-based scores without prior training) closely aligned with the rankings students assigned to each rubric criterion (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2023). Let us see two examples for the penultimate rubric in the students' ranking: organization and coherence.

Organization and Coherence (Student A)

- **Evaluation: Very Good.** The essay follows a clear structure: Introduction → Emotional State/Energy → Impact on Behavior → Duration → Conclusion. Each paragraph has a focused topic sentence, and transitions ("In contrast," "On the other hand," "Conversely") effectively signal shifts.
- **Areas for Improvement:** The transition between the introduction and first body paragraph could be smoother (e.g., "To begin, their emotional states and energy levels differ markedly"). The duration paragraph might better contrast the long-term effects of boredom vs. the short-term intensity of excitement.
- **Suggestion:** Use subheadings to enhance readability, e.g., **Emotional State and Energy**, **Behavioral Impact**, **Duration and Longevity**.

Organization and Coherence: (Student B)

0. **Evaluation: Very Good.** The essay has a clear structure: Introduction → Holmes (Work/Personality) → Dupin (Work/Personality) → Holmes as Copy (Conclusion). Headings enhance readability. The comparative structure (discussing the same points for each detective) creates inherent coherence. Transitions between sections are logical ("Regarding," "With respect to," "Thus").
1. **Areas for Improvement:** Some transitions *within* the Holmes/Dupin sections could be smoother (e.g., the shift between discussing work conditions and

personality within each section feels slightly abrupt). Repetition occurs (e.g., "withdraws again when his case is over," "remains mentally separate," "somehow controls his scorn" appear almost identically for both characters). The "Holmes as a Copy" section effectively summarizes; however, it could have synthesized the parallels more originally.

2. **Suggestion:** Use stronger transitional phrases when shifting focus within sections (e.g., "Beyond their working conditions, Holmes' personality is also defined by..."). Vary the language describing identical traits to avoid verbatim repetition. Consider a brief comparative analysis *before* the conclusion to link the points explicitly (DeepSeek, 2024).

Although the ASAP project trained Large Language Models (LLMs) using a dataset based on rubrics such as Prompt/Essay Content Relationship, Organization and Coherence, Accuracy, and Lexical and Grammatical Range, we introduced additional quantitative rubrics to better distinguish between high- and low-quality essays (McNamara, Crossley, & McCarthy, 2010). See the following JSON prompting template for eliciting such complementary quantitative data:

```
{
  "\ufeff\n\"Please score the provided essay ("essay":[insert essay here]) according to the following prompt("prompt":[insert prompt]using automated essay scoring techniques. Evaluate the essay based on the following rubrics: 1. Prompt:[insert prompt] and essay content relationship: [Insert evaluation] 2. Organization and coherence: [Insert evaluation] 3. Accuracy: [Insert evaluation] 4. Lexical and grammatical range: [Insert evaluation] Additionally, provide both quantitative and qualitative feedback on the following aspects of the essay: * Average sentence length: [Insert feedback] * Lexical diversity: [Insert feedback] * Formality: [Insert feedback] * SYNLE:[Insert SYNLE feedback] Please provide detailed and specific feedback for each rubric and aspect, including any areas for improvement and suggestions for enhancement.\"": null
}
```

Figure 1 JSON prompting template for eliciting complementary quantitative data (Vertopal).

A practical implementation of teacher assessment includes a JSON prompt template designed to generate accurate quantitative feedback for students' essays. This template distinguishes between high-quality and low-quality essays. Below, readers can see the prompting template in JSON code to obtain feedback for CELEX, SYNLE and MTLD indices:

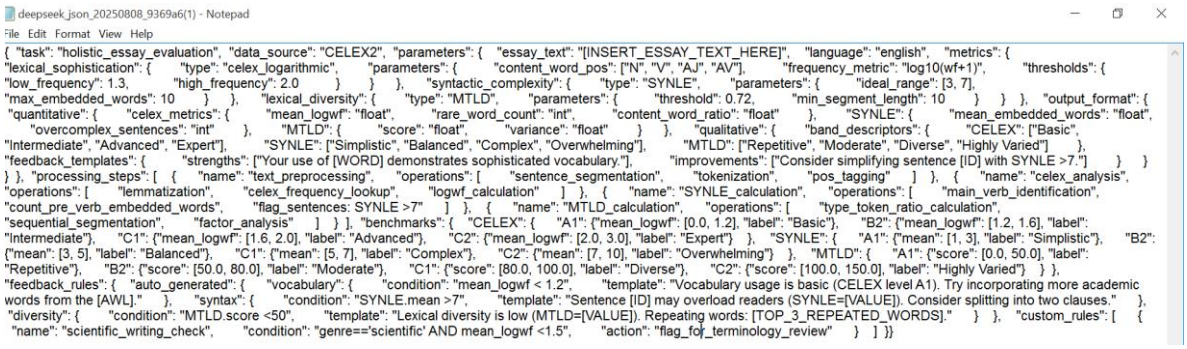

Figure 2 JSON prompting template for calculating CELEX, SYNLE and MTLD (DeepSeek, 2024)

Readers can see the results that the structured prompting template provides as quantitative feedback across three rubrics simultaneously. Using these rubrics—CELEX, SYNLE, and MTLD—teachers can demonstrate students' progress quantitatively, as illustrated in the following example:

Essay evaluation: Boredom vs. Excitement

Key Findings:

Lexical Sophistication

Advanced vocabulary (CELEX 1.63) with minor spelling errors 7 rare words including domain terms ("obsession", "vicious cycle") Recommendation: Use "compulsive behavior" (logwf=1.92) as alternate for "obsession" Syntactic Complexity

Two sentences exceed ideal SYNLE range:

- Sentence 4: 8 embedded words (comparison structure)
- Sentence 9: 7 embedded words (fulfillment contrast)

Lexical Diversity:

Moderate score (76.4) with expected repetition of key terms

Suggestion: Add synonyms for "passion" (e.g., "ardor", "dedication")

Sentence-Level Revision Example:

Original (SYNLE=8)
"While passion **is** an internal drive born from joy or purpose, obsession **is** an irrational and automatic urge...'

Revised (SYNLE=4/4):

"Passion stems from joyful purpose. Obsession, however, manifests as irrational compulsion."

Benchmark Comparison:

Metric	**Your Score**	B2 **Range**	**C1 Threshold**
CELEX	1.63	1.2-1.6	1.6
SYNLE	5.2	3-5	5
MTLD	76.4	50- 80	80

Figure 3 Key findings for CELEX, SYNLE, and MTLD (DeepSeek, 2024).

This is highly consistent with McNamara et al. (2010) views and their indication that these three indices are the most predictive features of essay quality. They concluded that if proficient writing, and thereby essay quality, is judged largely by the sophistication of the writing rather than on the ease of processing, then essays judged to be of higher quality are likely to be characterized by more complex sentences, less frequent words, and a greater diversity of words (McNamara, Crossley, & McCarthy, 2010).

This study has uncovered significant characteristics of the writing process that require implementing new strategies in our context to enhance assessment for learning in and beyond the classroom; for instance, redrafting essays after self and peer assessment practice. This implies organizing the writing activities in a way that peer assessment will be possible: Workshops in the Moodle platforms are the ideal actions to encourage students' peer assessment. As Julian M. Cox et al (2012) proposed for peer review so have, we followed the same methodology for peer assessment (Cox, Posada, & Waldron, 2012).

It is acknowledged that the focus of this study is limited by the small sample used characterized by small cohort of only 19 students. The authors of this work are willing to continue studying the role of AfL in the writing process, especially, to control the use of feedback provided by automated essay scoring systems. It is also important to recognize the role of a JSON structured prompts based on Chain of Thought (CoT) prompting (Wei, et al., 2023). This facilitates teachers offer the students feedback that identify the use of retrieval augmented generation (RAG) in the writing process (Li, Wang, Wang, Hung, Xie, & Wang, 2025). For example,

Analysis Report for One of the Essays

RAG Detected: `False`

Confidence Score: 20 (Low confidence, minimal RAG indicators)

Primary Indicators:

a. Consistent authorial voice

b. Gradual knowledge buildup

c. No abrupt citations or factual precision anomalies

Key Evidence:

Excerpt: "Boredom is an unpleasant, draining emotion where time seems to crawl, and nothing feels engaging. In contrast, excitement is invigorating..."

Paragraph: 2

Indicator Type: consistent tone (Subjective description without external sourcing cues)

Excerpt: "A writer might feel bored until a sudden burst of inspiration excites them into crafting a story..."

Paragraph: 4

Indicator Type: conceptual connection (Organic narrative progression)

Anomaly Summary:

No structural or stylistic hallmarks of RAG. Human-like flow with emotional introspection and no modular data dumps.

Explanation:

Non-RAG Signals Dominate: The essay uses original analogies (e.g., "sluggish drag of boredom"), maintains a uniform reflective tone, and lacks hyper-specific external references.

Absence of RAG Red Flags: No abrupt citations, terminology mismatches, or disjointed transitions. Claims about emotional states are generalized, not fact-heavy.

False Positive Safeguard: While some conceptual comparisons (e.g., adrenaline effects) could theoretically derive from retrieval, their integration is seamless and contextually justified.

Note: This assessment distinguishes RAG from legitimate human research by prioritizing narrative cohesion over isolated factual accuracy (DeepSeek, 2024).

Figure 4 Rag Detected (DeepSeek, 2024)

Teachers must emphasize the use of peer assessment in asynchronous workshops to encourage students in the proper use of feedback for enhanced essay writing outputs. The literature, particularly, McNamara, Crossley, & Roscoe, 2013 insisted on the reliability of feedback provided by AES. They reported that human and computer-based

scores correlate from .60 to .85 (McNamara D. J., Crossley, Roscoe, Allena, & Daia, 2015). However, teachers should be cautious because the validity of the AES depends on the specific dataset of essays, which may not fully represent the entire scope of writing styles (Li, Xuefeng, Cui, Li, & Zeng, 2023).

4. Conclusions

The implementation of an Assessment for Learning (AfL) strategy that involves students in both self-assessment and active peer assessment has been a priority for enhancing essay-writing skills. Integrating Automated Essay Scoring (AES) at key stages of the course has yielded positive results. Our goal is to maintain systematic pedagogical use of AES, employing not only zero-shot prompts for targeted feedback but also few-shot prompting templates to generate additional rubrics. This approach will further strengthen students' self- and peer-assessment capabilities.

5. Bibliographic references

1. Cox, J. M., Posada, J. P., & Waldron, R. (2012). Moodle Workshop activities support peer review in Year 1 Science: present and future. (M. Brown, M. Harnett, & T. Stewart, Eds.) *Ascilite* .

2. DeepSeek. (2024). *DeepSeek Chat (version 2) [Large Language Models]*. Retrieved 2025, from https://www.deepseek.com

3. Florez, M. T., & Sammons, P. (2013). Assessment for learning: impact: effects and impact. *CfBT Education Trust* , 32.

4. Hansen, C. T. (2024). Assessment for Learning: a summary of concepts, tactics and strategies. (S. Institute, Ed.) *ResearchGate* .

5. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large Language Models are Zero-Shot Reasoners. *Arxiv:2205.11916v4[cs.CL]* , 42.

6. Li, F., Xuefeng, X., Cui, Z., Li, D., & Zeng, W. (2023). Automatic Essay Scoring Method Based on MUlti-Scale Features. *Applied Sciences* , 18.

7. Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. (ELSVIER, Ed.) *Science Direct* , 14.

8. Mansour, W., Albatarni, S., & Elsayed, T. (2024). Can Large Language Models Automatically Score Proficiency of Written Essays? *LREC-COLING 2024* , 2777–2786.

9. McNamara, D. J., Crossley, S. A., Roscoea, R. D., Allena, L. K., & Daia, J. (2015). A Hierarchical Classification Approach to Automated Essay Scoring. (ScienceDirect, Ed.) 59.

10. McNamara, D., Crossley, S. A., & Mccarthy, P. M. (2010). Linguistic Features of Writing Quality. (SAGE, Ed.) *ResearchGate* , 63.

11. Vertopal. (n.d.). *Vertopal [Free Online Converter]*. Retrieved 2025, from https://www.vertopal.com

12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv: 2201.11903v6* , 43.