UNIVERSIDAD DE COSTA RICA

INIE Instituto de Investigación en Educación

Revista Actualidades Investigativas en Educación

ISSN 1409-4703

Assessing artificial intelligence and professors' calibration in English as a foreign language writing courses at a Costa Rican public university

Evaluación de la inteligencia artificial y de la calibración de docentes en los cursos de escritura de inglés como lengua extranjera en una universidad pública costarricense

Volumen 24, Número 1
Enero - Abril
pp. 1-25

William Charpentier-Jiménez

Citar este documento según modelo APA

Charpentier-Jiménez, William. (2024). Assessing artificial intelligence and professors' calibration in English as a foreign language writing courses at a Costa Rica public university. *Revista Actualidades Investigativas en Educación, 24*(1), 1-25. https://doi.org/10.15517/aie.v24i1.55612

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Assessing artificial intelligence and professors' calibration in English as a foreign language writing courses at a Costa Rican public university

Evaluación de la inteligencia artificial y de la calibración de docentes en los cursos de escritura de inglés como lengua extranjera en una universidad pública costarricense

William Charpentier-Jiménez[1]

***Abstract**: This article paper explores the evaluation of artificial intelligence (AI) in English as a Foreign Language (EFL) writing courses and the importance of calibration in writing evaluations. The role of calibration has received little attention in language contexts, while the role of artificial intelligence has gained increased attention in the last couple of years. This investigation, conducted from August 2022 to March 2023, involved eight TESOL students enrolled in an English as a Foreign Language (EFL) major at a Costa Rican public university, ten TESOL university professors, and one AI piece of software. It used a quantitative, quasi-experimental design, and a language elicitation data collection process. Data was collected by means of a rubric-based writing assessment. Quantitative data were analyzed using descriptive statistics. Data analyses indicate that: 1) human-created paragraphs ($\bar{X}$ = 7,56) and AI writing ($\bar{X}$ = 7,61) yield similar results when evaluated; 2) some criteria may favor human creativity or computer, rule-oriented writing; and 3) professors' ratings reveal inconsistencies when grading human writing in particular. These findings demonstrate that AI matches, at least to a basic level, human writing skills. Furthermore, data show that students may be falling behind in aspects such as grammar, vocabulary, and mechanics. Finally, the analysis indicates that professors' grading lacks consistency, and a calibration model should be incorporated as part of regular training workshops.*

***Keywords**: artificial intelligence, assessment, higher education, second language instruction, writing (composition)*

***Resumen**: Este artículo explora la evaluación de la inteligencia artificial (IA) en cursos de escritura en inglés como lengua extranjera (ILE) y la importancia de la calibración en las evaluaciones de escritura. El papel de la calibración ha recibido poca atención en contextos lingüísticos, mientras que la inteligencia artificial ha ganado mayor reconocimiento en los últimos años. La investigación se realizó desde agosto de 2022 hasta marzo de 2023, e involucró a ocho estudiantes de TESOL en un bachillerato en inglés como lengua extranjera (ILE) en una universidad pública de Costa Rica: diez docentes de TESOL a nivel universitario y un software de IA. Se utilizó un diseño cuasiexperimental cuantitativo y una recopilación de datos de elicitación de lenguaje. Los datos fueron recopilados mediante una rúbrica que midió la producción escrita. Los datos cuantitativos se analizaron utilizando estadística descriptiva. El análisis de datos indica que: 1) los párrafos creados por humanos ($\bar{X}$ = 7,56) y la escritura de IA ($\bar{X}$ = 7,61) producen resultados similares; 2) algunos criterios pueden favorecer la creatividad humana o la escritura orientada a reglas; y 3) el profesorado presenta inconsistencias al calificar la escritura humana en particular. Estos hallazgos demuestran que la IA se equipara, al menos a nivel básico, con las habilidades de escritura humana. Además, los datos muestran que el estudiantado puede estar quedándose atrás en aspectos como gramática, vocabulario y puntuación. Finalmente, el análisis indica que la calificación de docentes carece de consistencia, y un modelo de calibración debería ser incorporado como parte de su formación.*

***Palabras clave**: inteligencia artificial, evaluación, educación superior, enseñanza de una lengua extranjera, expresión escrita*

[1] *Docente e Investigador de la Universidad de Costa Rica, en la Escuela de Lenguas Modernas, San José, Costa Rica. Magister en Enseñanza del Inglés y en Administración Universitaria, ambos de la Universidad de Costa Rica. Orcid https://orcid.org/0000-0002-8554-7819*

Dirección electrónica: william.charpentier@ucr.ac.cr

***Artículo recibido**: 27 de junio, 2023*
***Enviado a corrección**: 8 de setiembre, 2023*
***Aprobado**: 20 de noviembre, 2023*

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

1. Introduction

1.1 Background

The advent of novel educational technologies frequently elicits powerful emotions ranging from fear and apprehension to irrepressible joy. However, the possible implications of writing and using Artificial Intelligence (AI) in English as a Second Language (ESL) settings remain unclear. First, learning to write is viewed as a creative process that should resemble real-world tasks. As a result, students should be allowed to use the same technological tools typically found in out-of-class contexts (Oh, 2020). Not only does AI assist writers, but it can also replace them. Artificial intelligence now writes sentences, paragraphs, or even essays in seconds and is virtually error-free regarding grammar and spelling. In addition, many professors are unaware of the possible advantages and disadvantages of using AI as a writing aid. This lack of awareness may hinder students' writing development. For example, students could utilize writing models and tailored materials. However, they could, intentionally or unintentionally, commit plagiarism by passing off AI-generated text as their own. Finally, students' writing development may also be hindered by inconsistent marking. Heterogeneous grading across courses or academic levels may confuse students, impact their grades, and diminish their writing motivation.

Currently, the international community has devoted increasing attention to AI and its capabilities. Some of this attention has been directed toward academic writing, often labeling the use of AI as academic dishonesty. However, the use of AI-generated texts in educational settings has received little to no attention, especially in ESL settings where learning to write requires practice, time, and patience. Many professors are unaware of the possible implications of unrestricted access to AI technologies (Martín-Marchante, 2022). In contrast to AI, calibration has received growing attention in the past few years. Evaluators have attempted to increase the reliability of graded activities through better test designs and calibration. However, despite decades of research on calibration, its implementation in the core writing courses of the English and English teaching majors has been less than satisfactory.

The research findings will directly benefit two populations. On the one hand, the university and professors will incorporate calibration training to improve clarity and reliability in writing assessments, thereby adding objectivity to an often subjective evaluation process (Gunnell et al., 2016; Ricker-Pedley, 2011; Sundqvist et al., 2020). As a result, adequate and ongoing writing assessment training benefits both students and the curriculum. On the other hand, students will have access to a fairer, less subjective, and more consistent evaluation throughout

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

their major. This will make students less anxious about writing assessments and establish more precise goals for them to achieve.

The specific objectives of this article are to compare ESL student-created writing to AI-generated writing at a general level and to examine the role of calibration in ESL writing courses at a Costa Rican public university.

2. Literature Review

In recent years, writing instruction and assessment have been the subject of extensive research. However, to the best of the researcher's knowledge, experiments that compare and contrast ESL students' writing to AI writing are scarce. This review of the literature is not intended to be exhaustive. Nevertheless, it explores some basic concepts to support the arguments derived from this study.

2.1 Definition and Uses of AI

Artificial intelligence is the simulation of human intelligence in machines. These machines are programmed to think and learn like humans or to perform tasks that typically require human-like intelligence, such as understanding language, recognizing patterns, learning, and problem-solving tasks (Arora, 2022; Cameron, 2019; Kent, 2022). This can also involve comprehending natural language, recognizing objects and sounds, making decisions, and reproducing information in various forms (Cameron, 2019; Zimmerman, 2018). Artificial intelligence can be subdivided into two main categories: narrow or weak AI. Weak AI is designed to perform a specific task. General or strong AI, on the other hand, can perform any intellectual task that a human can (Cameron, 2019).

As previously stated, narrow or weak AI, also known as "applied AI," refers to artificial intelligence systems designed to perform a particular task or set of tasks. These systems are trained to perform specific tasks by being fed large volumes of data, which they then utilize to make judgments, perform actions, or accomplish other tasks (Gulson et al., 2022). Examples of narrow AI include virtual assistants like Apple's Siri and Amazon's Alexa, which are designed to understand and respond to voice commands. Other examples include self-driving cars, which are designed to navigate roads and avoid obstacles. Narrow AI can effectively complete specific tasks; however, it is limited in its ability to adapt to new situations or perform tasks outside of its specific area of expertise. It is also not self-aware; it does not have consciousness or a sense of self.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Depending on their basic structure, AI can also be subdivided into three different types:

Rule-based AI: It follows predetermined rules to perform a task. For example, a rule-based AI system might be programmed to recognize and classify objects in an image based on specific features (Bourg and Seemann, 2004; Kochmar, 2022; Raynor, 2009).

Machine learning: It allows a system to learn and improve over time without being explicitly programmed. Machine learning algorithms use data to "train" the system to recognize patterns and make decisions based on those patterns (Bernard, 2021; Jones, 2018; Raynor, 2009).

Deep learning: It is a subset of machine learning that involves training artificial neural networks on a large dataset. These networks are inspired by the structure and function of the human brain and are able to learn and make decisions on their own (Jones, 2018; Roberts, 2022).

These types of AI are representative of what AI can currently do. As can be seen, the progression from a rule-based AI that is more dependent on humans to a deep learning stage in which machines can act and use language more like humans should generally raise awareness in society, as well as in academic and educational settings.

2.2 Natural Language Processing

When discussing computer language recognition or production, the user deals with Natural Language Processing (NLP). It is a branch of AI that deals with the interaction between computers and human language and seeks to guarantee a proper interaction among all the parties involved (Kochmar, 2022; McRoy, 2021). The goal of NLP is to enable computers to understand, interpret, and generate human language so that it is both meaningful and useful. Natural Language Processing can be used for many applications, including speech recognition, text-to-speech synthesis, machine translation, sentiment analysis, and text summarization.

Natural Language Processing relies on linguistics, computer science, and machine learning techniques to analyze and understand human language (Raaijmakers, 2022). It involves several steps, including tokenization, which breaks down text into individual words or phrases, and parsing, which analyzes the grammatical structure of the text. Natural Language Processing also involves using algorithms and models to determine the meaning of texts, such as word embeddings or neural networks (Kochmar, 2022; McRoy, 2021).

Recently, NLP has advanced to the point where it can quickly generate original, largely trustworthy content. For example, chatbots are trained on a massive amount of text data and can generate responses to a wide range of topics, including answering questions, generating

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

text, and carrying out conversations. The model is designed to be flexible, so it can be fine-tuned for specific use cases and domains, making it useful for various applications, such as customer service chatbots and personal assistants (Adamopoulou and Moussiades, 2020; Luo et al., 2022). Content generators, or AI-powered writing generators, are computer programs or tools that automatically produce written, visual or multimedia content for websites, advertisements, and other forms of media (Giansiracusa, 2021; Sharples and Pérez y Pérez, 2022). Although both systems are designed to generate responses to a prompt, the role of content generators focuses on producing text or variations of it based on a single prompt, while a chatbot may need to remember a previous input to decide on a possible answer.

One of the critical goals of AI research is to create systems that can learn and adapt on their own rather than being specifically programmed for every task. Artificial intelligence has the potential to revolutionize many industries, including healthcare, finance, transportation, and education (Clark, 2020; Hamdan et al., 2021; Kent, 2022; Lasry and Kobayashi, 2018; Popenici, 2023; Yu and Yu, 2021). However, it also raises ethical concerns, such as the possibility of job loss, the need for appropriate safeguards to prevent unintended consequences, plagiarism, and a lack of human creativity (Holmes and Porayska-Pomsta, 2023; Johnston, 2023; Lasry and Kobayashi, 2018; Roumate, 2023; Srinivasan, 2018; Tzen and Moquet, 2018).

2.3 Writing Assessment

Writing assessment is the process of evaluating and scoring an individual's written work in terms of its quality, content, style, grammar, and other language skills (Nation, 2009). These categories may vary from evaluation to evaluation, and professors may weigh them differently according to their specific requirements or students' needs (Brown and Lee, 2015). For some raters, not all errors deserve the same attention (Reid, 2006), and language programs should emphasize what students need to know to communicate effectively in written form (Adler-Kassner and O'Neill, 2010).

Since writing is one of the most challenging skills for students to master (Campbell, 2019; Dunn, 2021; Nosratinia and Razavi, 2016; Tillema, 2012), evaluators should consider several factors when designing writing assessment tasks. For example, according to Coombe et al. (2007) and Hyland (2019), writing assessment tasks should include a well-designed rubric, a prompt, an expected response, and a post-task evaluation where the professor reflects upon the writing task. This task design promotes a fair and orderly valuation of students' writing. Furthermore, other authors emphasize writing construct dimensions in which students should also be aware of extra-linguistic aspects such as context and purpose, audience awareness,

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

and genre conventions (Sparks et al., 2014). Some of these aspects are not usually attributed to machines or AI, as will be explained later. A third line of thought views writing as a process. From this perspective, writing is often the result of a process of thought, design, and revision that requires special skills that not all speakers develop naturally (Brown and Lee, 2015). This viewpoint stresses behaviors that are typically associated with humans and, to some extent, excludes writing created by AI.

Another factor that deserves special attention is the use of rubrics. Rubrics have become a staple in writing assessments, and their use is widely recommended (Ferris and Hedgcock, 2023; Harmer, 2011; Li, 2022; Shabani and Panahi, 2020). They lower students' anxiety (Arindra and Ardi, 2020), increase reliability, and avoid bias (Glass, 2005). Rubrics are often classified as analytic or holistic. Analytic rubrics assign markings to separate descriptors, such as grammar, vocabulary, or punctuation. The weight each descriptor receives may be equal or variable. On the other hand, when using holistic marking, raters assign an overall grade to students' written work (Carr, 2000; Coombe et al., 2007; Nation, 2009; Reid, 2006). Although holistic rubrics are time-saving and straightforward, an analytic rubric is usually preferred since it provides specific feedback to students (Ghalib and Al-Hattami, 2015; Ma, 2022; Peaci, 2020).

Finally, machine scoring programs and Intelligent Tutoring Systems have been developed and used by educators over the last few decades (Ericsson and Haswell, 2006). According to McAllister and White (2006), computerized writing assessment has its beginnings in Ellis Page's Project Essay Grade. In this study (Page, 1966) reported that the results provided by the computer could be used to predict human ratings. Similar results were later obtained from subsequent studies where some human scoring disadvantages, such as slowness and marking inconsistency, were also pointed out (Page and Dieter, 1968). By 1995, computers started to demonstrate a better judgement than their human counterparts (Page and Petersen, 1995). Despite this growing evidence, the academic community, particularly writing instructors and composition experts, has largely remained silent on the topic of machine scoring (Ericsson and Haswell, 2006) and the benefits it may bring over human raters.

2.4 Calibration and reliability

Assessment takes into account various factors to guarantee a fair and dependable evaluation process. These factors include validity, reliability, usefulness, practicality, and transparency (Brown and Abeywickrama, 2019; Coombe et al., 2007; Fulcher, 2010). For the purposes of the present study, only reliability will be explored. Reliability deals with the

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

consistency of test scores (Brown and Abeywickrama, 2019; Coombe et al., 2007), and it can be divided into two subcategories: intra-rater and inter-rater reliability.

Intra-rater reliability refers to the consistency or stability of measurement results obtained by a single rater when assessing a phenomenon multiple times (Scheel et al., 2018). In other words, it measures the degree to which a single rater's assessments are consistent over time. For example, if the same rater assesses similar groups in different places or over time, neither time nor place should affect the results. Intra-rater reliability deals more with the evaluators and their environment and mental or physical state, among other factors.

On the other hand, the results of two professors rating the same group of students should be consistent. When this does not happen, inter-rater reliability might be problematic. Inter-rater reliability assessment is a method used to evaluate the consistency of ratings or scores assigned by multiple evaluators, such as human raters or algorithms (Gwet, 2014). It measures the degree of agreement between raters on the scores assigned to a set of items or tasks, such as test questions, job performance evaluations, or oral or written productions. Inter-rater reliability assessment aims to ensure that the ratings or scores are consistent and accurately reflect the performance being evaluated, even when different evaluators are involved. Despite its importance in evaluation settings, inter-rater reliability has received little attention (Wilhelm et al., 2018).

Calibration in assessment refers to the process of ensuring that scores or ratings assigned to test items, tasks, or candidates are consistent and accurate across different raters or evaluators (Congdon and McQueen, 2000; Wendler et al., 2019). Although calibration is not one of the cornerstones of evaluation, it seeks to attain some of them, especially during grading. Calibration improves scoring accuracy and the reliability of ratings (Gunnell et al., 2016; Ricker-Pedley, 2011). This process helps to reduce the impact of subjective biases and ensure that ratings or scores are reliable and valid indicators of performance (Gunnell et al., 2016; Sundqvist et al., 2020). Calibration may involve training raters to use consistent standards, providing them with clear scoring guidelines, and regularly evaluating their performance through inter-rater reliability assessments (Ferris and Hedgcock, 2023). Calibration is a multi-phase process often requiring multiple rounds of recalibration to ensure consistency. Additionally, it plays a crucial role in test validation. The goal of test validation is to establish the credibility of test scores by confirming that the test is well-founded (Weir, 2005) and accurately measures what it is designed to assess (Coombe, 2007). Both calibration and test validation are essential for achieving fairness and generating reliable data.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

3. Methodology

The following section describes the study's methodological framework.

3.1 Approach

This study used a quantitative approach since it involves the collection and analysis of data that can be analyzed statistically. Quantitative research uses structured methods such as surveys, tests, or experiments to collect data, which are then usually analyzed through statistical methods (Creswell, 2019; Mackey and Gass, 2016; Mertler, 2019). The results of this type of research are typically presented in the form of statistical models, graphs, or tables.

Its design is quasi-experimental since it bypasses the need for random assignment. Instead of random assignment, quasi-experimental designs generally enable researchers to manage the allotment to the treatment condition. This is typically followed by the application of a specific matching technique to construct a control group that closely aligns with the characteristics of the treatment group (Creswell, 2019). The ultimate objective is to fabricate two groups that, barring the treatment condition, are analogous on all pertinent attributes (Mackey and Gass, 2016; Mertler, 2019).

It also uses an exploratory type of research because this type of research is utilized when the study topic is very novel and there is not much prior information available (Hernández Sampieri et al., 2010). Its primary objective is to familiarize oneself with the phenomenon or concept and gain a better understanding of it (Hernández Sampieri et al., 2010). Exploratory research, in general, is more open and flexible because its purpose is to explore the terrain, discover new variables and problem formulations, and formulate more precise hypotheses.

3.2 Unit of analysis

This study includes eight Costa Rican university students currently taking their last writing course in English or English teaching major at a public university. The researcher visited students' writing classes to invite them to participate. The participants were selected because they were currently enrolled in an advanced writing course. The final list included eight students who agreed to participate in the study. Participants received monetary compensation of $10 and a small bag of snacks for their collaboration. All participants speak Spanish as their first language. Of the eight students who wrote their paragraph, four (50%) were female and four (50%) were male. No student chose to identify as non-binary or did not respond to the question. The eight students (100%) reported being between the ages of 18 and 24. All students are

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

native Spanish speakers who are studying ESL. Regarding studies, students enrolled in the BA in English (n = 6, 75%) or English teaching (n = 2, 25%). Both majors share the same core language courses, including writing courses.

Raters were Costa Rican university professors who taught writing skills to ESL students at the same public university. All professors had more than ten years of experience teaching writing courses. The researcher sent eleven professors an electronic invitation. The final list of professors included ten raters who agreed to participate in the study; one professor did not reply to the invitation. The professors did not receive any compensation for the collaboration. Furthermore, of the ten professors who collaborated in the grading process, six (60%) were female and four (40%) were male. All professors (100%) work in an ESL BA program and have over ten years of experience teaching English to university students. In terms of their academic pursuits, two professors have a Ph.D. (20%), while the remaining eight professors (80%) have a Master's degree related to language teaching.

3.3 Data collection

3.3.1 Materials

The materials include written consent, a writing prompt created by the researcher (see Appendix 1), a rubric, and sixteen paragraphs, eight of which were created by students and eight generated by AI. The written consent was sent to students electronically. A checkbox labeled "agree to the terms and conditions" was included to certify voluntary participation in the study. The prompt (see Appendix 1) was created considering criteria such as not requiring specialist background knowledge, being accessible to students, and being clear and unambiguous (Coombe et al., 2007). The rubric was adapted from Booth (n.d.) and then revised by two professors with experience designing and using writing rubrics. The rubric included five criteria (content, organization, sentence structure, mechanics, and vocabulary), a set of numerical and descriptive labels (Masterful, 5 points; Skilled, 4 points; Able, 3 points; Developing, 2 points; Novice, 1 point; and Unacceptable, 0 points), and their corresponding descriptors. The study used sixteen paragraphs to simulate a small composition class. Paragraphs were evenly divided into human-created and AI-generated. Each paragraph was around ten sentences long. All paragraphs were standardized in terms of format (font type, font size, alignment, and spacing). However, they were not modified in any other manner.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

3.3.2 Procedure

This study used a language elicitation data collection process. The researcher created a simple yet informative prompt for students to write a paragraph and adapted the rubric to match five general criteria. The rubric did not include aspects modified by the researcher (e.g., format) or not generated by AI (e.g., title). Afterward, the rubric was revised by two professors experienced in writing courses. In the case of the students, each student wrote one paragraph of around ten sentences following the prompt given and in a place with minimal noise and no distractions. For this task, students were given 40 minutes. No student asked for extra time or reached the time limit. Students did not have access to any printed or electronic resources to aid them during their writing. To minimize the researcher's interference, they typed their paragraph directly on the computer. The AI-written texts were created using copy.ai. At the time of writing, Copy.ai was one of the newest and most reliable writing generators. In addition, Copy.ai is a free AI-powered writing generator, which allowed anyone who use it. Paragraphs created with this application were randomly chosen based on their length and not on content or other characteristics. Therefore, students' and AI-written paragraphs were uniform and did not bias raters. Eight AI-generated paragraphs were chosen to match the number of human-created paragraphs. All paragraphs were standardized in terms of format (font type, font size, alignment, and spacing) and were based on the prompt provided by the researcher (see Appendix 1). However, they were not modified in any other manner.

The ten professors were asked for their collaboration via email. In this stage, they were able to see a sample paragraph (not included to be evaluated) and the rubric. Once they accepted to participate, they received the paragraphs, rubric, and detailed instructions. They were given two months to grade the paragraphs. Paragraphs were given an alphanumeric ID and randomly organized. Before and during this stage, the researcher asked professors to grade the sixteen paragraphs; however, the inclusion of AI-generated paragraphs was not disclosed. The study was conducted from August 2022 to March 2023.

3.4 Data Processing and Analysis

The specific methods of analysis employed were chosen to best address the research questions posited at the outset of this study. Descriptive statistics were selected since they provide a snapshot of the data at hand through measures of central tendency and measures of variability (or dispersion). Some measures of central tendency included in this paper comprise the mean (average), median (the middle value when data is ordered from lowest to

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

highest), and mode (the most frequently occurring value). These measures give a center point of data distribution. Each paragraph was graded following a five point criteria which included content, organization, grammar, mechanics, and vocabulary. These criteria were selected since they are widely used in composition courses and are familiar to the evaluating professors.

The data set was subjected to computational analysis using Microsoft Excel 2016. The data was derived from professors' markings on students' paragraphs. Grades were organized in various data sheets according to the intended analysis (grades for AI-created paragraphs, grades for human-created paragraphs, and sub-category comparisons). As previously mentioned, the analysis included descriptive statistics, where nominal data, percentages, and the standard deviation, among other basic statistics, were performed to compare students' results using the data analysis tools add-in.

4. Analysis of the Results

The following description presents the study's results in three distinct sections. The first section includes scores assigned to each paragraph and contrasts the five analytic criteria used to assess the paragraphs. The second section compares the consistency of the raters' marks.

4.1 Analysis of Participants' Obtained Scores

In order to examine participants' writing, professors graded 16 paragraphs. Half of the paragraphs were written by advanced ESL students. The other eight paragraphs were written using Copy.ai (2022), an artificial intelligence. For the purposes of this study, human writers and AI are referred to as "participants." No paragraph was retained from the analysis, and no artificially generated paragraph was modified. To provide professors with sufficient context, they were instructed to grade the paragraphs using a rubric and to imagine this was a diagnostic test. Table 1 summarizes the main findings of this section.

Table 1

Summary of participants' paragraph scores: mean and standard deviation, Costa Rica, 2023

Paragraph Source	Min.	Max.	$\bar{X}$	SD
Human	6,22	9,09	7,56	0,79
Artificial Intelligence	7,28	8,01	7,61	0,21

Note. N = 16. Min. = Minimum; Max. = Maximum; $\bar{X}$ = arithmetic mean; SD = Standard Deviation
Source: Compiled by the author based on survey responses.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

As can be seen, AI participants outperformed human participants by a small margin. These results suggest that, at least in short writing samples, AI-generated paragraphs can easily compete with those created by humans. The standard deviation of students' scores indicates a moderate level of variability in the dataset. Although the scores are not all identical, they are not evenly distributed across the entire 1 to 10 scale, and they tend to cluster somewhat close to the mean. However, AI scores present a lower level of variability. These scores are more similar and are not evenly distributed across the full range of the scale. Finally, humans also obtained the highest and lowest scores of the 16 samples. This may also suggest that AI-generated paragraphs are more standard and share common features, while human abilities vary from individual to individual.

Additionally, some important information can be extracted by comparing participants' grades. First, only two students outperformed at least one AI sample. Second, only one student scored above 9; however, only one student scored below 7, which is a non-passing grade in this context. Most grades (n = 13) were in the 7 to 8 band, which, as seen before, indicates a low level of dispersion. Table 2 compares human and AI scores from the highest to the lowest.

Table 2
Participants' average scores: points and grades, Costa Rica, 2023

Sample	Human		Artificial Intelligence	
	Average Points	Average Grade	Average Points	Average Grade
Sample 1	22,73	9,09	20,02	8,01
Sample 2	19,52	7,81	19,29	7,71
Sample 3	19,03	7,61	19,06	7,62
Sample 4	18,98	7,59	19,06	7,62
Sample 5	18,78	7,51	18,95	7,58
Sample 6	18,49	7,40	18,84	7,53
Sample 7	18,09	7,24	18,74	7,49
Sample 8	15,56	6,22	18,19	7,28

Note. Each set of participants contributed eight paragraphs. Numbers depict the average grade from the criteria, where 0 was the minimum number of points, and 25 was the highest. Grades appear on a scale of 0 - 10.
Source: Compiled by the author based on professors' ratings

Some data indicates variations in outcomes based on the criteria used. For example, students achieved better results in content and organization. On the other hand, AI exhibited superior performance in grammar, mechanics, and vocabulary. This may imply that when written by humans, content and organization remain more appealing. Usually, content and organization do not follow specific rules; therefore, humans may incorporate more varied elements that appeal more to them. In contrast, elements such as grammar, mechanics, and

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

vocabulary follow stricter rules that can be objectively verified and, therefore, programmed for AI to use. Yet, AI may still use wordy structures or exhibit a restricted range of vocabulary, which may affect professors' perceptions when grading each paragraph. Table 3 summarizes human and AI criteria scores.

Table 3
Participants' average scores per criterion, Costa Rica, 2023

Paragraph Source	Content	Organization	Grammar	Mechanics	Vocabulary
Human	3,76	3,88	3,76	3,84	3,66
Artificial Intelligence	3,64	3,54	3,92	4,12	3,78

Note. Each criterion presents the average score obtained on a 0 - 5 scale.
Source: Compiled by the author based on professors' ratings

As previously stated, human and AI writing share many characteristics and are usually indistinguishable. However, when analyzing written production, specific patterns in dispersion, grade average, and scores per criterion show significant differences.

4.2 Analysis of Raters' Scores

When examining a particular piece of writing, professors have to take into account and are subject to several variables. Although professors across disciplines share some common traits, they also have unique qualities based on the subject they teach, their personal philosophies, or their teaching methods, among other characteristics. However, professors should also strive toward fairness and objectives when assessing students. In this sense, any institution should strive to train its staff to be consistent and accurate when grading. Taking this into account, this section focused on raters' average grades. Table 4 summarizes the most important findings of this section.

Table 4
Summary of raters' average given scores, Costa Rica, 2023

Source	Min.	Max.	Mode	$\bar{X}$	SD
Rater 1	6,00	9.60	7.20	7.55	0.95
Rater 2	8.16	9.92	9.88	9.56	0.48
Rater 3	4.40	8.20	7.60	7.18	0.99
Rater 4	6,00	10,00	9.20	8.25	1.19
Rater 5	5.20	7.60	5.20	6.43	0.93
Rater 6	6,00	10,00	9.40	8.63	1.13
Rater 7	4.80	8.80	7.20	6.70	1.18
Rater 8	6.40	9.20	6.80	7.60	0.89
Rater 9	6.40	9.60	6.40	7.59	1.05
Rater 10	4.40	9.20	5.60	6.35	1.48

Note. Each rater graded 16 paragraphs and contributed eight paragraphs. Min. = Minimum; Max. = Maximum; $\bar{X}$ = arithmetic mean; SD = Standard Deviation
Source: Compiled by the author based on professors' ratings

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

The analysis of these data reveals some important grading behaviors or practices. Except for one professor, all raters graded at least one student with a non-passing mark. However, some professors scored students severely while others were more lenient, considering that a difference of up to two points remains important. Equally important are some differences found in the maximum grades, where the greatest gap surpasses the two points. Several professors consider some participants excellent or outstanding, while others consider them average.

This lack of inter-rater reliability can also be observed when analyzing the mode and the mean. On average, three raters tend to assign non-passing grades to students. Four raters graded students in the range of seven to eight marks. The other three raters graded students' writing above eight. This includes one case where marks were comparatively high (mode = 9.88, $\bar{X}$ = 9.56).

In addition, the standard deviation, coupled with the other data, reveals crucial information. For example, Rater 2 has the lowest standard deviation and the high scores. This trend suggests that this professor consistently tends to grade students less severely than other colleagues. In contrast, Rater 5 presents consistent but more severe behavior when grading. Finally, Rater 10 has the highest level of dispersion when grading; however, the given grades tend to score low but acknowledge certain students' written work.

To exemplify this further, Table 5 presents the nominal results of the four largest grade differences.

Table 5
Summary of raters' average given scores, Costa Rica, 2023

Sample	Min.	Max.	Range
Sample 1	4.44	9.88	5.48
Sample 2	4.44	9.72	5.32
Sample 3	4.48	9.72	5.24
Sample 4	4.8	9.56	4.76

Note. Each rater graded 16 paragraphs. Min. = Minimum; Max. = Maximum; $\bar{X}$ = arithmetic mean; SD = Standard Deviation
Source: Compiled by the author based on professors' ratings

As these data show, depending on the professor, a student may completely fail an assignment or get an almost perfect mark on another. These four samples are the highest, according to the range. In addition, all of them belong to human participants. Considering the range of these grades, this information evidences the need to calibrate professors and offer them the necessary guidance to increase their accuracy and objectiveness when grading students' writing.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

The analysis provided valuable insights into how professors perceive students' writing and the possible similarities and differences between it and AI writing, shedding light on educational areas that demand greater attention and a more direct approach. These findings also highlight specific aspects of evaluating written works and the need for calibration to provide more accurate and objective feedback when grading writing.

5. Conclusions

Artificial intelligence has gained popularity in recent years, and its capabilities have, for good or bad, already permeated educational settings. The present analysis concludes that paragraphs generated by artificial intelligence performed slightly better than those written by human participants in the study. The results indicate that AI-generated paragraphs can effectively compete with human-created ones in short writing samples. While there was a moderate level of variability in the students' scores, the scores tended to cluster close to the mean, suggesting a consistent level of performance among the human participants. On the other hand, the AI scores exhibited less variability, indicating a more standardized and consistent performance. Moreover, the AI-generated paragraphs displayed fewer extremes in scores compared to human writing, suggesting a more standardized and predictable outcome.

Furthermore, the analysis revealed differences in scores based on the grading criteria used. Human participants achieved higher scores in content and organization, indicating that human-written paragraphs may possess more varied elements that appeal to readers. In contrast, AI performed better in grammar, mechanics, and vocabulary, which are areas governed by stricter rules. However, AI-generated paragraphs may still exhibit wordy structures and limited vocabulary, potentially influencing professors' perceptions when grading. These findings suggest that content and organization may be more appealing in human writing, while AI excels in rule-based aspects of writing. This opens up new possibilities for assessing student writing. In line with Nation's (2009) suggestion that multiple markers should evaluate student work to ensure reliability, a hybrid approach combining both human and machine assessments could offer a viable solution for enhancing the quality and reliability of grading and feedback from professors.

The analysis of raters' scores highlighted variations in grading practices among professors. While some professors graded more leniently, others scored more strictly, resulting in discrepancies in students' final grades. The results obtained by students vary greatly depending on the professor who evaluated each paragraph. As pointed out by Gunnell et al.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

(2016) and Ricker-Pedley (2011), this indicates the need for calibration in language courses. Although incorporating human-created writing is crucial, AI-written samples could serve to model what is expected from students according to their proficiency level and guide professors toward a more objective, efficient, and fair grading of students' writing.

Another important factor to consider is plagiarism (Smith, 2022). Although several techniques can be employed to detect computer-generated writing (Abd-Elaal et al., 2022), professors may be unaware of their use or have difficulty asserting that a piece of writing was completely or partially generated with AI. As this study shows, both sets of participants yielded similar results. Professors do not always have the tools, training, time, or definitive proof to label a piece of writing as plagiarism. Students often present their writing printed, making identifying longer writing pieces more demanding and inefficient. Other researchers (Beyduz, 2023; Salas-Pilco and Yang, 2022; Schiff, 2022; Smith, 2022) have claimed that AI may merge with educational practices and help teach students. However, the results demonstrate that professors may not be ready to cope with this integration or guide students to use AI educationally and ethically.

The analysis concluded that AI-generated paragraphs could compete with human-written paragraphs in short writing samples. While human writing may excel in content and organization, AI outperforms in grammar, mechanics, and vocabulary. Variations in grading practices among professors emphasized the need for calibration and training to enhance consistency and objectivity. In addition, professors tend to grade at slower rates, be influenced by external factors, or even intrinsic factors that may alter scorings (Page and Dieter, 1968). These insights shed light on the evaluation of written works and the importance of addressing specific areas for improvement in writing assessment.

This study has three main limitations that deserve attention. First, the sample size was relatively small. It is impossible to generalize the performance of the entire population based on eight paragraphs. It is also impossible to generalize the potential of AI based on eight samples. However, the number of paragraphs adequately simulates a writing class. In addition, a larger set of paragraphs may also hinder professors from collaborating or dedicating enough time to each sample.

Second, as previously stated, writing includes prewriting strategies, drafting, revising, and revising, among other strategies and activities. Producing a good piece of writing takes time and requires revisiting and improving one's writing (Cheung, 2016; Murray and Christison, 2011; Sethuraman and Radhakrishnan, 2020). In this study, students had a limited time frame

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

to complete the task. Although no student reached the time limit, they did not have enough time to prewrite or self-edit their work. To minimize this limitation, professors were instructed to consider this a diagnostic or placement test in which students would have similar conditions.

Third, research has suggested that one piece of writing may be insufficient to properly assess a student's writing ability (Nation, 2009). A more reliable setting would include several markers (at least two or three) and several pieces of writing (at least two or three) (Nation, 2009). Multiple evaluations were not practical for practical reasons. In particular, the contrast with AI-generated texts would not have been possible since, unlike humans, machines do not interact differently and are not susceptible to being influenced by environmental, social, or daily situations. An improvement or decline in these types of paragraphs cannot be measured over short periods of time.

This study should be replicated in other ESL settings or with different types of AI. For example, not all populations write general English but instead focus more on it for academic or specific purposes. In addition, not all AI applications write alike; some might be able to develop cohesion among paragraphs to create essays or more extended pieces of writing. They may also adapt to the user or become more proficient over time. Further research should also be undertaken to determine to what extent the limitations described here may affect the results of the present study.

6. Acknowledgements

The author wishes to express sincere gratitude to Mag. Shazia Alfaro Magnan, Mag. Hector Alvarado Picado, Dr. Marisela Bonilla López, Dr. Alonso Canales Viquez, Mag. Zúñiga Coudin Randolph, Dr. Alberto Delgado Álvarez, Mag. Carolina Gonzales Ramírez, Mag. César Navas Brenes, Mag. Rosalba Rojas Víquez, Mag. Mayra Solís Hernández, and Mag. Netzi Valdelomar Miranda for their careful revision of the paragraphs as well as their insightful assistance during the research process.

7. References

Abd-Elaal, El-Sayed., Gamage, Sithara., and Mills, Julie. (2022). Assisting academics to identify computer generated writing. *European Journal of Engineering Education*, *47*(5), 725-745. https://doi.org/10.1080/03043797.2022.2046709

Adamopoulou, Eleni., and Moussiades, Lefteris. (2020). An Overview of Chatbot Technology. In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (Vol. 584, pp. 373–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4_31

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Adler-Kassner, Linda., and O'Neill, Peggy. (2010). *Reframing writing assessment to improve teaching and learning*. Utah State University Press.

Arindra, Margaretha Yola., and Ardi, Priyatno. (2020). The Correlation between Students' Writing Anxiety and the Use of Writing Assessment Rubrics. *LEARN Journal: Language Education and Acquisition Research Network*, *13*(1), 76–93. https://eric.ed.gov/?id=EJ1242955

Arora, Varun. (2022). *Artificial intelligence in schools: a guide for teachers, administrators, and technology leaders*. Routledge.

Bernard, Etienne. (2021). *Introduction to machine learning*. Wolfram Media.

Beyduz, Baris. (2023). *The Parent`s Guide to Artificial Intelligence and Education: Helping your Child Adapt and Succeed in a Rapidly Changing World: How A.I. Will Shape Our Kids*. Independently published.

Booth, Melanie. (n.d.). *College-Level Writing Rubric*. Saint Mary's College. https://my.smccme.edu/ICS/icsfs/College_Writing_Rubric.pdf?target=7037f7b6-6809-4d28-86a5-f9ed01f0acf0

Bourg, David M., and Seemann, Glenn. (2004). *AI for game developers*. O'Reilly.

Brown, H. Douglas., and Abeywickrama, Priyanvada. (2019). *Language assessment: principles and classroom practices* (3th ed.). Pearson Education.

Brown, H. Douglas., and Lee, Heekyeong. (2015). *Teaching by principles: an interactive approach to language pedagogy* (4th ed.). Pearson Education.

Cameron, Ryan M. (2019). *A.I. - 101: a primer on using artifical intelligence in education*. Exceedly Press.

Campbell, Madelaine. (2019). Teaching Academic Writing in Higher Education. *Education Quarterly Reviews*, *2*(3). https://doi.org/10.31014/aior.1993.02.03.92

Carr, Nathan T. (2000). A Comparison of the Effects of Analytic and Holistic Rating Scale Types in the Context of Composition Tests. *Issues in Applied Linguistics*, *11*(2). https://doi.org/10.5070/L4112005035

Cheung, Yin Ling. (2016). Teaching Writing. In Willy A. Renandya and Handoyo Puji Widodo (Eds.), *English Language Teaching Today: Linking Theory and Practice* (1st ed. 2016). Springer International Publishing: Imprint: Springer.

Clark, Donald. (2020). *Artificial intelligence for learning: how to use AI to support employee development*. Kogan Page Limited.

Congdon, Peter J., and McQueen, Joy. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163-178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Coombe, Christine A., Folse, Keith S., and Hubley, Nancy J. (2007). *A practical guide to assessing English laugage learners*. University of Michigan.

CopyAI, Inc. (2022). Copy.ai (July 14 version) [Large language model]. https://copy.ai

Creswell, John. (2019). *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). Pearson.

Dunn, Michael. (2021). The Challenges of Struggling Writers: Strategies That Can Help. *Education Sciences*, *11*(12), 795. https://doi.org/10.3390/educsci11120795

Ericsson, Patricia., and Haswell, Richard. (2006). *Machine Scoring of Student Essays: Truth and Consequences*. USU Press Publications. https://digitalcommons.usu.edu/usupress_pubs/139

Ferris, Dana., and Hedgcock, John S. (2023). *Teaching L2 composition: purpose, process, and practice* (4th ed.). Routledge.

Fulcher, Glenn. (2010). *Practical language testing*. Hodder Education.

Ghalib, Thikra., and Al-Hattami, Abdulghani. (2015). Holistic versus Analytic Evaluation of EFL Writing: A Case Study. *English Language Teaching*, *8*(7), p225. https://doi.org/10.5539/elt.v8n7p225

Giansiracusa, Noah. (2021). Crafted by Computer: Artificial Intelligence Now Generates Headlines, Articles, and Journalists. In Noah Giansiracusa, *How Algorithms Create and Prevent Fake News* (pp. 17–39). Apress. https://doi.org/10.1007/978-1-4842-7155-1_2

Glass, Kathy Tuchman. (2005). *Curriculum design for writing instruction: creating standards-based lesson plans and rubrics*. Corwin Press.

Gulson, Kalervo N., Sellar, Sam., and Webb, P. Taylor. (2022). *Algorithms of education: how datafication and artificial intelligence shape policy*. University of Minnesota Press.

Gunnell, K. L., Fowler, D., and Colaizzi, K. (2016). Inter-rater reliability calibration program: critical components for competency-based education. *The Journal of Competency-Based Education*, *1*(1), 36-41. https://doi.org/10.1002/cbe2.1010

Gwet, Kilem Li. (2014). *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters* (Fourth edition). Advances Analytics, LLC.

Hamdan, Allam Mohammed Mousa., Hassanien, Aboul Ella., Khamis, Reem., Alareeni, Bahaaeddin., Razzaque, Ajum., and Awwad, Bahaa Sobhi Abde Latif. (Eds.). (2021). *Applications of artificial intelligence in business, education and healthcare*. Springer.

Harmer, Jeremy. (2011). *How to teach writing* (9a. impr). Longman, Pearson Education.

Hernández Sampieri, Roberto., Fernández Collado, Carlos., and Baptista Lucio, Pilar. (2010). *Metodología de la investigación* (5a. ed). McGraw-Hill.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Holmes, Wayne., and Porayska-Pomsta, Kaska. (Eds.). (2023). *The ethics of artificial intelligence in education: practices, challenges, and debates*. Routledge, Taylor and Francis Group.

Hyland, Ken. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Johnston, Michael. (2023). *The Artificial Intelligence Disruption: How to Adapt and Succeed in the Age of Intelligent Machines*. Self Published.

Jones, Herbert. (2018). *Deep Learning: An Essential Guide to Deep Learning for Beginners Who Want to Understand How Deep Neural Networks Work and Relate to Machine Learning and Artificial Intelligence*. CreateSpace Independent Publishing Platform.

Kent, David. (2022). *Artificial intelligence in education: fundamentals for educators*. Kotesol DDC.

Kochmar, Ekaterine. (2022). *Getting started with Natural Language Processing*. Manning Publications.

Lasry, Brigitte., and Kobayashi, Hael. (Eds.). (2018). *Human decisions: thoughts on AI*. United Nations Educational, Scientific and Cultural Organization.

Li, Wentao. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing*, *35*(10), 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Luo, Bei., Lau, Raymond Y. K., Li, Chunping., and Si, Yain-Whar. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1). https://doi.org/10.1002/widm.1434

Ma, Wenyue. (2022). What the analytic versus holistic scoring of international teaching assistants can reveal: Lexical grammar matters. *Language Testing*, *39*(2), 239–264. https://doi.org/10.1177/02655322211040020

Mackey, Alison, and Gass, Susan. (2016). *Second language research: methodology and design* (2nd ed.). Routledge.

Martín-Marchante, Beatriz. (2022). The use of ICTs and artificial intelligence in the revision of the writing process in Valencian public universities. *Research in Education and Learning Innovation Archives*, (28), 16-31. https://doi.org/10.7203/realia.28.20622

McAllister, Ken., and White, Edward. (2006). Interested Complicities: The Dialectic of Computer-Assisted Writing Assessment. In Patricia Ericsson and Richard Haswell, *Machine Scoring of Student Essays: Truth and Consequences* (pp. 8-27). USU Press Publications. https://digitalcommons.usu.edu/usupress_pubs/139

McRoy, Susan. (2021). *Principles of natural language processing*. Susan McRoy.

Mertler, Craig. (2019). *Introduction to educational research* (2nd ed.). SAGE Publications, Inc.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Murray, Denise E., and Christison, MaryAnn. (2011). *What English language teachers need to know*. Routledge.

Nation, Paul. (2009). *Teaching ESL/EFL reading and writing*. Routledge.

Nosratinia, Mania., and Razavi, Faezeh. (2016). Writing Complexity, Accuracy, and Fluency among EFL Learners: Inspecting Their Interaction with Learners' Degree of Creativity. *Theory and Practice in Language Studies*, *6*(5), 1043-1052. https://doi.org/10.17507/tpls.0605.19

Oh, Saerhim. (2020). Second Language Learners' Use of Writing Resources in Writing Assessment. *Language Assessment Quarterly*, *17*(1), 60–84. https://doi.org/10.1080/15434303.2019.1674854

Page, Ellis. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, *47*(5), 238-243. http://www.jstor.org/stable/20371545

Page, Ellis., and Dieter, Paulus. (1968). *The Analysis of Essays by Computer* (Final Report of U.S. Office of Education Project No. 6-1318). Washington, DC: Department of Health, Education, and Welfare. ERIC Document Reproduction Service, ED 028 633. https://archive.org/details/ERIC_ED028633/mode/2up

Page, Ellis., and Petersen, Nancy. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561. https://www.proquest.com/docview/218533317?pq-origsite=gscholar&fromopenview=true

Peaci, Davut. (2020). Writing evaluation in university English preparatory programs: Two universities of Turkey and Saudi Arabia. *Dil ve Dilbilimi Çalışmaları Dergisi*, *16*(1), 253–264. https://doi.org/10.17263/jlls.712798

Popenici, Stefan. (2023). *Artificial intelligence and learning futures: critical narratives of technology and imagination in higher education*. Routledge.

Raaijmakers, Stephan. (2022). *Deep learning for natural language processing*. Manning Publications Co.

Raynor, William J. (2009). *International dictionary of artificial intelligence* (2. ed., new ed). Global Professional Publ.

Reid, Joy M. (2006). *Essentials of teaching academic writing*. Houghton Mifflin.

Ricker-Pedley, Kathryn L. (2011). An examination of the link between rater calibration performance and subsequent scoring accuracy in graduate record examinations® (GRE ®) writing. *ETS Research Report Series*, *2011*(1), i–22. https://doi.org/10.1002/j.2333-8504.2011.tb02239.x

Roberts, Daniel A. (2022). *The principles of deep learning theory: an effective theory approach to understanding neural networks*. Cambridge University Press.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Roumate, Fatima. (2023). *Artificial intelligence in higher education and scientific research: future development*. SPRINGER VERLAG, SINGAPOR.

Salas-Pilco, Sdenka Zobeida., and Yang, Yuqin. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. International Journal of Educational Technology in Higher Education, 19(1), 21. https://doi.org/10.1186/s41239-022-00326-w

Scheel, Carrie., Mecham, Jim., Zuccarello, Vic., and Mattes, Ryan. (2018). An evaluation of the inter-rater and intra-rater reliability of OccuPro's functional capacity evaluation. *Work*, *60*(3), 465-473. https://doi.org/10.3233/WOR-182754

Sethuraman, Mekala., and Radhakrishnan, Geetha. (2020). Promoting Cognitive Strategies in Second Language Writing. *Eurasian Journal of Educational Research*, (88), 1–17. https://doi.org/10.14689/ejer.2020.88.5

Shabani, Enayat A., and Panahi, Jaleh. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia*, *10*(1), 12. https://doi.org/10.1186/s40468-020-00111-4

Sharples, Mike., and Pérez y Pérez, Rafael. (2022). *Story Machines: How Computers Have Become Creative Writers*. Routledge. https://doi.org/10.4324/9781003161431

Smith, Adam. (2022). *Revolutionizing Education with Artificial Intelligence*. Independently published.

Schiff, Daniel. (2022). Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies. *International Journal of Artificial Intelligence in Education, 32*(3), 527–563. https://doi.org/10.1007/s40593-021-00270-2

Sparks, Jesse R., Song, Yi., Brantley, Wyman., and Liu, Ou Lydia. (2014). Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment: Assessing Written Communication. *ETS Research Report Series*, *2014*(2), 1-52. https://doi.org/10.1002/ets2.12035

Srinivasan, Rajeev. (2018). The Ethical Dilemmas of Artificial Intelligence. In Brigitte Lasry and Hael Kobayashi (Eds.), *Human decisions: thoughts on AI* (pp. 103-107). United Nations Educational, Scientific and Cultural Organization.

Sundqvist, Pia., Sandlund, Erica., Skar, Gustaf B., and Tengberg, Michael. (2020). Effects of Rater Training on the Assessment of L2 English Oral Proficiency. *Nordic Journal of Modern Language Methodology*, *8*(1), 3-29. https://doi.org/10.46364/njmlm.v8i1.605

Tillema, Marion. (2012). *Writing in first and second language: empirical studies on text quality and writing processes*. Netherlands Graduate School of Linguistics.

Tzen, MonZen., and Moquet, Xavier. (2018). A.I and big data: what kind of education and what kind of place is there for the citizen? In Brigitte Lasry and Hael Kobayashi (Eds.), *Human decisions: thoughts on AI* (pp. 108-111). United Nations Educational, Scientific and Cultural Organization.

Doi https://doi.org/10.15517/aie.v24i1.55612
Volumen 24, Número 1, Art. Cient., ene-abr 2024

Wendler, Cathy., Glazer, Nancy., and Cline, Frederick. (2019). Examining the Calibration Process for Raters of the *GRE* ® General Test. *ETS Research Report Series*, *2019*(1), 1–19. https://doi.org/10.1002/ets2.12245

Weir, Cyril. (2005). *Language testing and validation: An evidence-based approach*. Houndmills UK: Palgrave Macmillan. https://ztcprep.com/library/tesol/Language_Testing_and_Validation/Language_Testing_and_Validation_(www.ztcprep.com).pdf

Wilhelm, Anne Garrison., Rouse, Amy Gillespie., and Jones, Francesca. (2018). Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research, and Evaluation*, *23*. https://doi.org/10.7275/AT67-MD25

Yu, Shengquan., and Yu, Lu. (2021). *An introduction to artificial intelligence in education*. Springer Nature.

Zimmerman, Michelle Renée. (2018). *Teaching AI: exploring new frontiers for learning*. International Society for Technology in Education.