e-ISSN 1135-9250

EDUTEC. Revista Electrónica de Tecnología Educativa

Issue 89 – September 2024

Section: Artificial intelligence in the evaluation and personalization of learning

An experimental analysis of the relationship between the evaluations of artificial intelligence and pre-service teachers

Un análisis experimental de la relación entre las evaluaciones proporcionadas por la inteligencia artificial y las proporcionadas por los docentes en formación

Héctor Galindo-Domínguez[1]; hector.galindo@ehu.eus;
Nahia Delgado[1]; nahia.delgado@ehu.eus;
Martín Sainz de la Maza[1]; martin.sainzdelamaza@ehu.eus;
Ernesto Expósito[2]; ernesto.exposito@univ-pau.fr;

Abstract

One of the potential benefits of artificial intelligence (AI) is its ability to optimize teachers' tasks. The aim of this study was to analyze the possible differences between assessments carried out by pre-service teachers and those performed by various AI systems. A total of 507 pre-service teachers participated, and they were provided with a rubric to evaluate 12 texts of different types and quality. The results showed that AI performance in evaluating written tasks closely replicated the functioning of pre-service teachers, with ChatGPT being the AI that most accurately mirrored the teachers' evaluations, achieving approximately 70% precision compared to human assessments. Similarly, there were minimal differences in the assessments made by pre-service teachers based on gender and academic year. Moreover, evaluations conducted by higher-performing pre-service teachers were more aligned with those provided by AI, compared to those from lower-performing students. These findings are valuable, highlighting how AI could serve as a supportive tool to guide the pedagogical knowledge of pre-service teachers in assessment tasks.

Keywords: Assessment, Artificial Intelligence, ChatGPT, Teacher Training

Resumen

Uno de los beneficios potenciales de la inteligencia artificial (IA) es que puede permitir la optimización de las tareas de los docentes. Este estudio tuvo como objetivo analizar las posibles diferencias entre las evaluaciones realizadas por docentes en formación y las realizadas por diferentes IA. Participaron un total de 507 docentes en formación, a quienes se les proporcionó una rúbrica para evaluar 12 textos de distintos tipos y calidades. Los resultados mostraron cómo el desempeño de las IA en la evaluación de tareas escritas replicó con bastante precisión el funcionamiento de los docentes en formación, siendo ChatGPT la IA que mejor replicó el comportamiento de los docentes en formación, con una precisión cercana al 70% de la evaluación proporcionada por humanos. Del mismo modo, hubo diferencias mínimas en las evaluaciones realizadas por los docentes en formación según su género y año académico. Asimismo, las evaluaciones realizadas por los docentes en formación con mejor desempeño estuvieron más alineadas con las proporcionadas por la IA en comparación con los estudiantes con menor desempeño. Estos resultados son útiles, al destacar cómo la IA podría ser una herramienta de apoyo que guíe el conocimiento pedagógico de los docentes en formación en tareas de evaluación.

Palabras clave: *Evaluación, Inteligencia Artificial, ChatGPT, Formación Docente*

[1] Euskal Herriko Unibertsitatea / University of the Basque Country (Spain)
[2] University of Pau and the Pays de l'Adour (France)

DOI: https://doi.org/10.21556/edutec.2024.89.3509

Received: 02-08-2024
Accepted: 06-09-2024

1. INTRODUCTION

1.1. Assessment in Teacher Education: Responsibilities, Challenges, and Influential Factors

In the Organic Law 3/2020, the primary educational law in Spain, Article 91 outlines the functions of teachers, including the assessment of student learning processes. Previous works such as that of Stiggins (2014) have demonstrated how teachers invest between one-third and up to half of their professional time in tasks associated with assessment and grading, demonstrating that it is a time-consuming task, especially with the increasing student-to-teacher ratio (Ramesh & Kumar, 2022). However, despite being a task that requires effort and dedication, it has been observed that increased teachers' assessment literacy has a direct impact on students' learning outcomes (e.g., Mellati & Khademi, 2018; Xu & Brown, 2016).

Considering this function that any teacher worldwide carries out with their students, teaching pre-service teachers new knowledge and skills to evaluate their future students could be essential for making the teaching task as efficient as possible (Atjonen et al., 2022). Alongside this idea, previous studies show significant limitations in how pre-service teachers are taught to evaluate, sometimes because it is excessively theoretical, and at times because it is disconnected from the daily tasks of a teacher (Atjonen, 2017; DeLuca et al., 2019; Salama & Subahi, 2020). As result, in many cases, pre-service teachers apply evaluation strategies that were used on them when they were students themselves (Hill et al., 2017).

Likewise, the amount of knowledge and skills of a pre-service teacher when assessing assignments could be conditioned based on a series of personal and academic variables. Although the existing evidence to date is excessively scant, some studies such as Salama & Subahi (2020) observed that the assessment literacy of pre-service teachers was relatively low and similar regardless of their gender, academic performance, or years of experience. Moreover, Lovorn & Reza (2011) observed how the training received can also influence how assessment is conducted through rubrics, and Deneen & Brown (2016) observed how academic performance of pre-service teachers plays a determining role in the depth of an assessment. Nevertheless, we must consider that the extent to which task evaluation is detailed may be influenced by academic performance, which in turn could be due to the impact of other socio-emotional factors, such as students' motivation towards the task or their current emotional state (Eklöf, 2010). For this reason, the degree of detail in task evaluation could be a multidimensional phenomenon.

1.2. The integration of artificial intelligence in digital assessment

With the technological advancements of recent years, some of the evaluation methods being employed to address issues such as high student-to-teacher ratios, personalized instruction, and reducing excessive time consumption involve the use of artificial intelligence-based systems (e.g. Vij et al., 2019).

Artificial intelligence is the capability of a machine to replicate intelligent human behavior (Wang, 2019). As of today, there is an abundance of tools based on AI. Within the field of AI, an expanding area is that of generative AI, which is understood as AI focused on creating content based on user input. Some of the most important generative AIs today are ChatGPT, from OpenAI, Bing's Copilot, from Microsoft, or Gemini, from Google, to name but a few.

Generative AI has potential applications in assessing student tasks, as various studies have demonstrated its accuracy in providing feedback. These AI systems, supported by natural language processing, offer tailored responses to complement teachers' efforts (Ocaña-Fernández et al., 2019; González-Calatayud et al., 2021). Jani et al. (2020) highlighted AI's role in formative evaluation, using machine learning to monitor student progress and improve clinical practices. Similarly, AI has been applied in medical training (Mirchi et al., 2020), engineering education (Samarakou et al., 2016; Liu et al., 2017), and programming assessments (Grivokostopoulou et al., 2017), providing automated feedback based on performance. Other studies (Rhienmora et al., 2011; Ouguengay et al., 2015; Ulum, 2020; Choi & McClenen, 2020) further demonstrate AI's capacity for grading and skill evaluation in various fields.

Similarly, certain studies have compared the effect of using AI-based systems to not using them. For instance, Grivokostopoulou et al. (2017) conducted a comparison between the results obtained by an AI and the hand-assessments made by teachers to verify the accuracy of this technology. The results showed a correlation between the two, with only slight differences observed in excellent works where teachers tended to overrate the task compared to the scores given by the AI. These results are coherent with those obtained by Houtao et al. (2022) who observed how generative AI's feedback might be as useful as teacher's feedback, albeit with some differences. Particularly, in text correction, while teacher feedback tended to focus on task structure and content, AI feedback was more detailed in vocabulary and grammar. These findings underscore the potential value of integrating both forms of feedback to ensure thoroughness.

As commented by Dillenbourg (2016) the transition from traditional to digital education doesn't signify the obsolescence of teachers in the future. Rather than deliberating on whether AI will supplant teachers, Hrastinski et al. (2019) propose acknowledging the potential benefits of AI and how these advantages could reshape their role in the classroom. Hence, in educational assessment, teachers continue to play a crucial role in ensuring the appropriate utilization of AI for measurement and evaluation objectives. Some of these responsibilities encompass crafting assessments and establishing learning objectives, contextualizing assessment queries to render them more pertinent and meaningful to students, interpreting outcomes to deliver personalized feedback tailored to students' strengths and weaknesses, monitoring student progress, among others (Owan et al., 2023).

1.3. Purpose of the study

As observed, the vast majority of studies commented before utilize (generative) AI-based systems to provide feedback to students, yet they do not compare experimentally such feedback with what could be provided by teachers. Likewise, the studies conducting comparative analyses between evaluations provided by AI and those provided by in-service teachers are minimal (e.g., Grivokostopoulou et al., 2017; Houtao et al., 2022), but to the authors' knowledge, studies evaluating the relationship between evaluations provided by generative AI and by pre-service teachers are nonexistent. Moreover, recent systematic reviews have demonstrated the increasing interest of the scientific community in using artificial intelligence and machine learning systems to automate essay scores to tackle the increasing student ratio and with time-consuming tasks, such as performing feedback and grading tasks (Ramesh & Kumar, 2022). For this very reason, and given the scant literature on the subject, it is imperative to ascertain whether such tools might be ready or not to serve as an additional aid in the pedagogical knowledge of the teacher during their digital assessment processes.

Based on these necessities, the objectives of this study are:

- O1: Analyze whether there are statistically significant differences between the assessments done by generative AI and the assessments done by pre-service teachers on different written texts.

- O2: Analyze whether the differences in the assessments between generative AI and pre-service teachers depend on their gender.

- O3: Analyze whether the differences in the assessments between generative AI and pre-service teachers depend on their training level.

- O4: Analyze whether the differences in the assessments between generative AI and pre-service teachers depend on their academic achievement.

2. METHOD

2.1. Participants

A total of 507 college students took part in the current study, with an average age of 20.56 years (SD = 5.42). Among them, there were 155 males, 348 females, and 4 individuals who did not affiliate with any specified gender category. Concerning the academic degree 130 university students came from Early Childhood Education, 327 from Primary Education, and 50 from related fields such as Pedagogy or Social Education. In terms of academic progression, 172 students were in their first year, 137 in their second, 168 in their third, 25 in their fourth, and 5 in their fifth year of study. A portion of the sample was selected based on proximity, consisting of students of the researchers involved in this study. Another portion of the sample was selected through the dissemination of an institutional message inviting participation from students enrolled in the degree programs offered by the three Faculties of Education at the University of the Basque Country. Despite the sample being non-probabilistic, there is prior evidence indicating that convenience samples can yield results similar to those obtained from randomized samples (e.g., Coppock et al., 2018).

2.2. Instruments

To collect all the data, two different instruments were employed. Firstly, participants were asked a series of personal variables such as gender, age, academic year, university degree they were studying and academic achievement. For $O_{3,}$ training level was codified based on the academic year of the student. Specifically, the first year students were placed in initial training, the second year students in intermediate training, and the third and fourth year students in final university degree training. The fourth year students were joined with the third year students due to the fact that the fourth year students have fewer subjects. Likewise, for O_4, academic achievement was measured through the arithmetic mean of the grades obtained by the students in the year prior to the one in which they were enrolled. This value is known from the academic transcript that is sent to them at the end of each year. Based on their scores, groups of low (percentile < 33), medium (percentile 34 to 66) and high (percentile > 67) academic performance were created.

Secondly, a series of ad-hoc texts written in Spanish were first generated by ChatGPT 3.5 and next, supervised by educational experts. These texts were later evaluated by in-service teachers (n = 3). The prompt used to generate the texts was as follows:

> *Write a [type of text] text of 5 to 10 lines written by a 10-year-old student on a free choice topic, with [type of quality] content, organization, vocabulary, coherence, and cohesion.*

In total, 12 texts of different types (3 descriptive texts, 3 narrative texts, 3 argumentative texts, and 3 instructional texts) and qualities (4 excellent, 4 normal, and 4 significantly improvable) were developed. Generated texts can be found in Appendix I. Additionally, an ad-hoc rubric was created with ChatGPT 3.5 and supervised also by in-service teachers (n = 3). When assessing an essay by means of AI several studies have analyzed which should be the main criteria teachers should use. In this sense, some criteria are joined with statistical features, like essay length with respect to the number of words, the essay length with respect to the sentence, the average sentence length or the average word length (Contreras et al., 2018: Kumar et al., 2019; Mathias & Bhattacharyya, 2018). Concerning the style- or syntax-based criteria, the most relevant criteria are the sentence structure, the punctuation, the grammar, the logical operators, and the vocabulary (Cummins et al., 2016; Darwish & Mohamed, 2020; Ke et al., 2019). Likewise, concerning content-based criteria, the most relevant criteria are the cohesion between sentences in a document, the relevance of information, the correctness, and the consistency (Dong et al., 2017).

Based on these criteria, the ad-hoc tool was built by 4 different criteria (Content, Organization, Vocabulary, Coherence and Cohesion) and 4-points achievement levels (excellent, good, fair and poor). Therefore, using the rubric, each text could have a maximum score of 16 points (4 criteria x 4 achievement levels), although for practical purposes to align with the Spanish educational system, these 16 points were weighted on a scale of 10, with 10 being the maximum grade for each text (equivalent to 16 unweighted points). Achievement levels were established by ChatGPT based on criteria derived from previous empirical work previously commented. These achievement levels were reviewed by researchers, showing high agreement between those provided by the AI and those found in other rubrics for assessing written works (e.g., Government of Newfoundland and Labrador, 2014).

This rubric, collected in Appendix II, was created to assist pre-service teachers in evaluating the generated texts. All participating students had received specific training on how to use assessment rubrics, as along their degree they have general didactics and specific didactics courses in which they are taught to design, interpret, and apply assessment rubrics. This aspect is important as previous studies have shown how a pre-service teacher without training on the use of rubrics issues an evaluation equally as subjective as a pre-service teacher without any assessment tools (Lovorn & Reza, 2011). The reliability of the different texts using the evaluation rubric is shown in Table 1. As can be seen, since all the values above the cut-off point of α > .70 (Tavakol & Dennick, 2011), it can be assumed that there is a good internal consistency of the criteria used for the evaluation of each text.

Table 1

Reliability values for each text

Order	Text	Quality	Reliability
1	Descriptive text	Significantly improvable	.792
2	Argumentative text	Normal	.801
3	Instructive text	Significantly improvable	.741
4	Narrative text	Excellent	.904
5	Descriptive text	Excellent	.904
6	Argumentative text	Significantly improvable	.781
7	Instructive text	Excellent	.922
8	Narrative text	Normal	.816
9	Descriptive text	Normal	.857
10	Argumentative text	Excellent	.875
11	Instructive text	Normal	.869
12	Narrative text	Significantly improvable	.814

Finally, each of the analyzed AIs was asked to evaluate the different texts using the following prompt:

> The following text has been produced by a 10-year-old student. Considering the criteria of content, organization, vocabulary, and coherence and cohesion, provide each text with a rating from 0 to 10: [Text crafted by ChatGPT].

2.3. Procedure

The process commenced with the creation of the different texts. These texts were created using ChatGPT, following the next structure: Write a [type of text] of 5 to 10 lines written by a 10-year-old student on a free choice topic, with [type of quality] content, organization, vocabulary, coherence, and cohesion. These texts were supervised by educational experts to detect coherence, lexical and grammatical mistakes. Subsequently, these texts were digitally converted to Google Forms and the rubric to assess each text was also added to each text to permit students to assess each criterion (content, organization, vocabulary, and coherence and cohesion).

DOI: https://doi.org/10.21556/edutec.2024.89.3509

Initially, a subset of the sample was drawn from the researchers' student cohort. Similarly, the research team sought permission from the dean's office to invite participation from all other students not under their instruction. On this process, all ethical measures accepted in the Declaration of Helsinki were strictly followed. In both instances, participants were required to review and agree to the procedures and terms of participation prior to contributing any data, ensuring adherence to ethical standards. This agreement included informing participants of the study's objectives, anticipated response time, the confidential and anonymous handling of data, and the voluntary nature of participation, which included the option to retract responses during the survey. Data from the researchers' student body were gathered during regular working hours, while responses from external students were accepted at any time. Finally, upon completing the analysis of the study, a report summarizing the key findings was sent to those who expressed a desire to receive the results.

2.4. Data analysis

The data analysis process was entirely conducted using the statistical software SPSS Statistics 27. Initially, considering all participants' responses, the scores for each text were calculated. As mentioned previously, although each text could receive a maximum of 16 points (4 criteria x 4 achievement levels), for practical purposes, these 16 points were weighted to 10 (a score of 10 equates to 16 points on the rubric) for easier interpretation, as in the Spanish educational system, a 10 represents the maximum grade. Then, to address the first objective, arithmetic means, and standard deviations were calculated for each group (AI vs Pre-service teachers). Subsequently, to identify potential significant differences, a Student's t-test supplemented with Cohen's d was performed. Additionally, the accuracy percentage, equal to the number of texts deviating by less than 1 point from the arithmetic means of the pre-service teachers, divided by the total number of texts was calculated. If the accuracy were 100%, it would mean that the evaluations provided by the AI would closely match (in all cases, deviating by less than 1 point) the evaluations given by the pre-service teachers. To address the second objective, arithmetic means and standard deviations were recalculated for each group, followed by a one-way ANOVA to examine potential differences between the groups. This analysis was accompanied by Tukey's post-hoc test to identify which groups exhibited significant differences. For the third objective, the same procedure as for the second objective was followed. Finally, for the fourth objective, given the academic performance of the pre-service teachers on a scale, three different groups were formed: lower performance (percentile 1 to 33), medium performance (percentile 34 to 66), and high performance (percentile 67 to 99). Once classified, another one-way ANOVA was performed to identify potential significant differences. This analysis was supplemented with a Post-Hoc analysis using Tukey's test to determine between which groups significant differences were found.

3. RESULTS

Firstly, concerning objective 1, to determine if there were significant differences between the scores assigned to the different texts using the evaluation rubric and the scores assigned by the AIs, various t-tests were conducted.

As can be observed in Table 2, out of 12 texts, statistically significant differences were found in only 4 of them between the ratings provided by the pre-service teachers and the AIs. However, it is worth noting that even though statistically significant differences were not present in the majority of cases, in those instances where there were significant differences, the effect size was large [Excellent Argumentative, p = .036, d = 1.25; Normal Argumentative, p = .029, d = 1.56; Excellent Narrative, p < .001, d = .70; Normal Narrative, p = .046, d = 1.43].

The accuracy of various AIs in replicating the evaluative ability of pre-service teachers was also analyzed. The method used to calculate the accuracy of the AI involved counting the number of texts evaluated by the AI with a deviation of less than 1 point from the assessment given by the pre-service teachers, and dividing it by the total number of texts (12). The 1-point deviation was arbitrarily selected, as it is understood that there may be good agreement between the assessment provided by the AI and the assessment provided by the pre-service teachers when there is a deviation of less than 1 point out of 10 between the two scores. The results revealed that ChatGPT was the AI among those analyzed that best replicated the ratings provided by the pre-service teachers (8/12 = 66.66% accuracy), with Gemini coming in second (7/12 = 58.33% accuracy), and Bing's Copilot in the worst position (6/12 = 50% accuracy).

Table 2

Main results from the t-test

Type	Quality	Artificial Intelligence				Pre-Service Teachers	p	d
		ChatGPT	Bing	Gemini	Total IA			
Descriptive	Excellent	8_+	9.25_+	9_+	$8.75 (.661)_+$	8.95 (1.46)	(ns)	-
	Normal	6_+	8.5	7_+	$7.16 (1.25)_+$	6.75 (1.77)	(ns)	-
	Low	4_+	4.5_+	6	$4.83 (1.04)_+$	4.95 (1.43)	(ns)	-
Argumentative	Excellent	9	9	8_+	8.66 (.577)	7.08 (1.68)	*	1.25
	Normal	7	8.5	7	7.5 (.866)	5.52 (1.56)	*	1.56
	Low	6	4_+	7	5.66 (1.52)	4.52 (1.32)	(ns)	-
Instructive	Excellent	9_+	8.75_+	9_+	$8.91 (.144)_+$	9.08 (1.37)	(ns)	-
	Normal	5	8.75	8_+	$7.25 (1.98)_+$	7.11 (1.63)	(ns)	-
	Low	4_+	3.5	5_+	$4.16 (.763)_+$	4.61 (1.44)	(ns)	-
Narrative	Excellent	9_+	8.75_+	9_+	$8.91 (.144)_+$	8.08 (1.65)	***	.70
	Normal	7_+	8.75	8	7.91 (.877)	6.07 (1.59)	*	1.43
	Low	3_+	2.75_+	5	$3.58 (1.23)_+$	3.18 (1.09)	(ns)	-
Accuracy			50%	58.33%	66.66%			

Note. +, the Generative AI's score is under 1-point deviation from the mean of the assessments provided by pre-service teachers; * p < .05; ** p < .01; *** p < .001; (ns), non-significant.

Subsequently, concerning objective 2, various one-way ANOVA tests were conducted, considering three different groups: evaluations provided by AI, evaluations provided by male pre-service teachers, and evaluations provided by female pre-service teachers. As observed in Table 3, the differences between genders were minimal, as only in 3 out of 12 texts significant differences based on gender were found. Of these 3 gender-based differences, in 2 of them, females had a score more similar to that provided by the AI, while in only one case did males have a score more similar to that provided by the AI. In the remaining cases, there were no statistically significant differences based on gender.

From this analysis onwards, accuracy was calculated as the number of texts from the pre-service teachers that had an arithmetic mean lower than 1 point compared to the arithmetic mean provided by the generative AIs. In this case, the accuracy for both genders concerning the evaluations provided by the generative AIs was 66.6% (8/12 texts).

Table 3

Main results from the one-way ANOVA based on pre-service teachers' gender

Type	Quality		Género		p	Post-Hoc
		Total IA (3)	Male (1)	Female (1)		
Descriptive	Excellent	8.75 (.661)	8.72 (1.54)$_+$	9.04 (1.42)$_+$	(ns)	-
	Normal	7.16 (1.25)	6.62 (1.64)$_+$	6.81 (1.82)$_+$	(ns)	-
	Low	4.83 (1.04)	5.06 (1.49)$_+$	4.90 (1.41)$_+$	(ns)	-
Argumentative	Excellent	8.66 (.577)	6.87 (1.50)	7.17 (1.75)	(ns)	-
	Normal	7.5 (.866)	5.78 (1.63)	5.41 (1.51)	**	2<1; 1<3; 2<3
	Low	5.66 (1.52)	4.56 (1.37)$_+$	4.50 (1.30)	(ns)	-
Instructive	Excellent	8.91 (.144)	8.98 (1.36)$_+$	9.13 (1.38)$_+$	(ns)	-
	Normal	7.25 (1.98)	6.89 (1.34)$_+$	7.21 (1.74)$_+$	(ns)	-
	Low	4.16 (.763)	4.56 (1.43)$_+$	4.63 (1.45)$_+$	(ns)	-
Narrative	Excellent	8.91 (.144)	7.81 (1.55)	8.20 (1.68)$_+$	*	1<2
	Normal	7.91 (.877)	5.89 (1.47)	6.15 (1.64)	*	(ns)
	Low	3.58 (1.23)	3.29 (1.21)$_+$	3.13 (1.04)$_+$	(ns)	-
	Precisión		8/12 (66.6%)	8/12 (66.6%)		

Note. +, the pre-service teachers' mean is under 1-point deviation from the mean of the assessments provided by the Generative AIs; * $p < .05$; ** $p < .01$; *** $p < .001$; (ns), non-significant. Post-hoc performed by Tukey's post-hoc.

Next, concerning objective 3, several one-way ANOVA tests were conducted, considering 4 groups: the AIs, pre-service teachers starting their university degree (year 1), pre-service teachers midway through their university degree (year 2), and pre-service teachers concluding their university degree (years 3 and 4). Broadly speaking, the analyses gathered in Table 4 reveal how the differences observed among the various groups are minimal, being in all cases 8 out of 12 assessments (66.6% accuracy) similar to those provided by Generative AIs. These results may confirm that regardless of the stage at which the pre-service teacher finds themselves, their ability to rate and evaluate will not be very different. The only exceptions were low-quality descriptive texts, where pre-service teachers concluding their studies overestimated the score provided by their peers and by the AI ($p < .001$), as well as excellent instructional texts, where they also overestimated the provided score ($p = .004$).

Table 4

Main results from the one-way ANOVA based on pre-service teachers' training level

Type	Quality	AI Total (4)	Training level			p	Post-Hoc
			Starting (1)	Medium (2)	Finishing (3)		
Desc	Excellent	8.75 (.661)	8.82 (1.53)$_+$	9.01 (1.20)$_+$	9.01 (1.56)$_+$	(ns)	-
	Normal	7.16 (1.25)	6.52 (1.73)$_+$	6.78 (1.75)$_+$	6.92 (1.81)$_+$	(ns)	-
	Low	4.83 (1.04)	4.76 (1.35)$_+$	4.61 (1.19)$_+$	5.34 (1.57)$_+$	***	1<3; 2<3
Argu	Excellent	8.66 (.577)	7.04 (1.65)	7.18 (1.63)	7.06 (1.76)	(ns)	-
	Normal	7.5 (.866)	5.69 (1.54)	5.48 (1.48)	5.41 (1.62)	*	(ns)
	Low	5.66 (1.52)	4.44 (1.28)	4.47 (1.23)	4.62 (1.40)	(ns)	-
Instr	Excellent	8.91 (.144)	8.79 (1.57)$_+$	9.34 (1.00)$_+$	9.16 (1.37)$_+$	**	1<2; 1<3
	Normal	7.25 (1.98)	7.16 (1.74)$_+$	7.07 (1.57)$_+$	7.11 (1.59)$_+$	(ns)	-
	Low	4.16 (.763)	4.81 (1.47)$_+$	4.42 (1.23)$_+$	4.57 (1.53)$_+$	(ns)	-
Narr	Excellent	8.91 (.144)	8.04 (1.61)$_+$	8.10 (1.70)$_+$	8.10 (1.70)$_+$	(ns)	-
	Normal	7.91 (.877)	6.16 (1.63)	5.96 (1.46)	6.07 (1.64)	(ns)	-
	Low	3.58 (1.23)	3.16 (1.10)$_+$	3.07 (1.05)$_+$	3.26 (1.11)$_+$	(ns)	-
	Precisión		8/12 (66.6%)	8/12 (66.6%)	8/12 (66.6%)		

Note. +, the pre-service teachers' mean is under 1-point deviation from the mean of the assessments provided by the Generative AIs; * p < .05; ** p < .01; *** p < .001; (ns), non-significant. Post-hoc performed by Tukey's post-hoc.

Finally, concerning objective 4, several one-way ANOVA tests were conducted, considering 4 groups: the AIs, pre-service teachers with lower academic performance (Percentile < 33), pre-service teachers with medium academic performance (Percentile 34 to 66), and pre-service teachers with high academic performance (Percentile > 67). The results, as observed in Table 5, showed how pre-service teachers with higher academic performance were more accurate evaluators (9 out of 12 texts, 75%) of the ratings provided by the AI, in contrast to pre-service teachers with medium (8 out of 12 texts, 66.6%) or lower academic performance (7 out of 12 texts, 58.3%). Similarly, as can be observed from the differences in means presented in Table 5, the number of texts with higher scores increases with higher academic performance. Therefore, it can be observed that students with higher performance tend to evaluate written texts more favorably compared to students with lower academic performance. Likewise, it can be seen that the analyzed AIs tended to overestimate scores in most texts, regardless of students' academic performance. Thus, the number of texts overestimated by the AI compared to student evaluations with low performance was 10 out of 12 texts (5 of them with more than 1-point deviation from the mean), in the case of students with medium performance was 8 out of 12 texts (4 of them with more than 1-point deviation from the mean), and in the case of students with high performance was 7 out of 12 texts (3 of them with more than 1-point deviation from the mean). These data demonstrate how academic performance could be an important factor in generating evaluations more closely aligned with those provided by AI.

Table 5

Main results from the one-way ANOVA based on pre-service teachers' academic achievement

Type	Quality	AI Total (4)	Academic Achievement						p	Post-Hoc
			Low P<33 (1)		Medium P 34-66 (2)		High P>67 (3)			
			M (SD)	Diff M^1	M (SD)	Diff M^1	M (SD)	Diff M^1		
Descriptive	Excellent	8.75 (.661)	8.72 (1.52)$_+$	.03	8.88 (1.50)$_+$	-.13	9.35 (1.24)$_+$	-.60	**	1<3; 2<3
	Normal	7.16 (1.25)	6.41 (1.73)$_+$	.75	6.84 (1.81)$_+$	.32	7.08 (1.69)$_+$	.08	**	1<3
	Low	4.83 (1.04)	4.74 (1.39)$_+$	.09	4.84 (1.39)$_+$	-.01	5.39 (1.48)$_+$	-.56	***	1<3; 2<3
Argumentative	Excellent	8.66 (.577)	6.87 (1.64)	1.79	7.11 (1.75)	1.55	7.34 (1.61)	1.32	*	(ns)
	Normal	7.5 (.866)	5.48 (1.54)	2.02	5.43 (1.56)	2.07	5.71 (1.56)	1.79	(ns)	-
	Low	5.66 (1.52)	4.40 (1.36)	1.26	4.44 (1.19)	1.22	4.80 (1.40)$_+$	.86	*	1<3
Instructive	Excellent	8.91 (.144)	8.94 (1.40)$_+$	-.03	9.06 (1.43)$_+$	-.15	9.31 (1.24)$_+$	-.40	(ns)	-
	Normal	7.25 (1.98)	7.05 (1.59)$_+$	.20	6.99 (1.68)$_+$	.26	7.37 (1.60)$_+$	-.12	(ns)	-
	Low	4.16 (.763)	4.48 (1.47)$_+$	-.32	4.60 (1.37)$_+$	-.44	4.79 (1.49)$_+$	-.63	(ns)	-
Narrative	Excellent	8.91 (.144)	7.68 (1.68)	1.23	8.07 (1.64)$_+$	.84	8.65 (1.46)$_+$	.26	***	1<3; 2<3
	Normal	7.91 (.877)	5.78 (1.50)	2.13	6.11 (1.64)	1.8	6.40 (1.57)	1.51	***	1<3
	Low	3.58 (1.23)	3.20 (1.11)$_+$	.38	3.10 (1.04)$_+$	.48	3.25 (1.15)$_+$	.33	(ns)	-
	Accuracy		7/12 (58.3%)		8/12 (66.6%)		9/12 (75%)			

Note. +, the pre-service teachers' mean is under 1-point deviation from the mean of the assessments provided by the Generative AIs; Diff M, The difference between the arithmetic mean of the AI and the arithmetic mean of the group. A higher positive value indicates a higher underestimation by the group compared to the AI's evaluation, while a higher negative value indicates a higher overestimation by the group compared to the AI's evaluation; * p < .05; ** p < .01; *** p < .001; (ns), non-significant. Post-hoc performed by Tukey's post-hoc.

4. DISCUSSION Y CONCLUSIONS

Assessing is among the main tasks that any teacher carries out in their professional duties (Organic Law 3/2020). This task, despite being time-consuming and complex to perform individually due to classroom ratios (e.g., Ramesh & Kumar, 2022), it may bring significant improvement to the student learning processes (Mellati & Khademi, 2018; Xu & Brown, 2016). Faced with this dilemma and with the introduction of new technologies in classrooms, a possible alternative being considered to address this problem is to employ tools based on AI that may allow for valid and reliable replication of assessments made by teachers. Taking this idea as a starting point, the main objective of this study has been to determine whether assessments provided by different generative AIs are faithful to those given by pre-service teachers, as well as to understand if there is any variable such as gender, training level, or academic performance that could influence to make an assessment more faithful to that provided by a generative AI.

The results showed that the different AIs analyzed are capable of replicating pre-service teachers' patterns fairly well when evaluating written tasks, with ChatGPT being the AI that yielded the highest accuracy (close to 70% agreement with pre-service teacher evaluation) and Bing's Copilot being the AI that yielded the lowest accuracy (50% agreement with pre-service teacher evaluation). These results are consistent with the limited previous literature on this topic, which generally reveals a concordance between the feedback provided by AI and the feedback given by in-service teachers, finding small differences between the two groups (Grivokostopoulou et al., 2017; Houtao et al., 2022).

Furthermore, the results have shown that this degree of agreement, broadly speaking, is identical with the limited literature on this topic regardless of the gender and training level of the pre-service teacher. As an exception, clear significant differences were found only in the evaluations given on the academic performance of the pre-service teachers, with those with higher academic performance providing assessments more closely aligned with those provided by the AI compared to those with lower academic performance. These results are also partially consistent with the limited literature on this topic. Specifically, consistent with the present study, Salama & Subahi (2020) also observed how gender and training level were variables that had little influence on evaluation knowledge and skills, while works such as Deneen & Brown (2016) showed how the academic performance of pre-service teachers significantly influenced the depth of evaluation conducted. However, the findings of the present study contradict those of Salama & Subahi (2020), who observed that academic performance was not an influential variable in the evaluation of knowledge and skills.

4.1. Theoretical and practical implications

These results have important theoretical and practical implications that need to be discussed. Firstly, the findings of this study may be valuable for the scientific community as they might contribute to expanding the current understanding of the degree of agreement between evaluations provided by educators and those provided by AI-based systems.

Secondly, these results may be relevant for university teachers in education-related areas as they underscore the potential interest on training future teachers in digital technologies, such as the use of artificial intelligence, big data, or learning analytics, to optimize temporal resources and provide as personalized assistance and monitoring as possible to their students.

In Europe, significant efforts have been made to develop a framework for the conceptualization and development of teachers' digital competence, with the DigCompEdu model being the primary model (Redecker, 2017). This framework highlights the relevance of employing digital technologies in teacher training for carrying out evaluative tasks within the fourth competence "Assessment". This competence is aimed at improving assessment strategies, analyzing learning evidence, and providing feedback by means of digital technologies. In this regard, using

AI-based technologies may contribute to partially addressing this competence. Similarly, previous studies show how an improvement in digital teaching competence might have important effects on improving time management and teacher self-efficacy, both of which are key variables in reducing stress and discomfort generated by professional tasks (Galindo-Domínguez & Bezanilla, 2021).

Finally, these results may be useful for in-service teachers as they highlight the growing potential of artificial intelligence-based systems as a means to assess students' written work. While previous studies have shown how primary and secondary education teachers currently use AI-based tools primarily for content creation purposes, such as texts or images (Galindo-Domínguez et al., 2024), it may be interesting to add specific training on how to employ AI for student evaluation and monitoring as part of the ongoing professional development process in educational institutions. As mentioned by different authors (e.g., Kasneci et al., 2023; Owan et al., 2023), the use of AI in the educational field requires that teachers and learners develop a set of necessary competencies to understand the technology itself, exploit its potential utilities, as well as acknowledge its limitations.

4.2. Limitations and prospective

The present study has several limitations that should be taken into account when interpreting the results. The first limitation concerns the sample, as although it is a relatively large sample of pre-service teachers, these results may vary from those that the same participants might yield in a few years when they are in active service. For this reason, future studies could replicate the methodology used in the present study, using in-service teachers of different educational stages as participants instead of pre-service teachers to observe if the obtained results are similar or not. Additionally, it could be interesting to compare novice teachers, mid-career educators, and veteran teachers’ evaluations, as it might reveal important insights into how teaching experience may influence the assessment procedures. This idea would be justified in that while some studies suggest that the amount of teaching experience could influence assessment literacy (e.g., Spear-Swerling et al., 2005), others point in the opposite direction (e.g., Bagsao & Peckley, 2020; Salama & Subahi, 2020).

Likewise, the present study only considered the evaluation of written works, making it impossible to determine if the level of agreement between the evaluations provided by AI and pre-service teachers could also occur in works provided in other formats such as audio, video, images, or mathematical equations, for example. Although it is more complex for the current state of AI systems, future studies could attempt to replicate the methodology used but evaluate tasks in different formats than the written format.

Finally, the written texts were generated by the AI instructed to put itself in the shoes of a 10-year-old student. However, differences in the results obtained could arise compared to those that might be obtained if real texts written by 10-year-old students were used. For this reason, future studies could replicate the methodology of the present work but using texts written by real students.

5. REFERENCES

Atjonen, P. (2017). Development of teacher assessment literacy in comprehensive schools – Views from the curriculum analysis. *Kriteerit Puntarissa, 74*, 132–169.

Atjonen, P., Pöntinen, S., Kontkanen, S., & Ruotsalainen, P. (2022). In Enhancing Preservice Teachers' Assessment Literacy: Focus on Knowledge Base, Conceptions of Assessment, and Teacher Learning. *Frontiers in Education, 7,* 1-12. https://doi.org/10.3389/feduc.2022.891391

Bagsao, J., & Peckley, M.K. (2020). Assessment Literacy of Public Elementary School Teachers in the Indigenous Communities in Northern Philippines. *Universal Journal of Educational Research, 8*(11b), 5693-5703. http://dx.doi.org/10.13189/ujer.2020.082203

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 8196. https://www.mdpi.com/2076-3417/10/22/8196#

Contreras, J.O., Hilles, S.M., & Abubakar, Z.B. (2018) Automated essay scoring with ontology based on text mining and NLTK tools. In I. Zen (Pres.), *2018 International Conference on Smart Computing and Electronic Enterprise* (pp. 1-6). IEEExplore.

Coppock, A., Leeper, T.J., Mullinix, K.J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *PNAS, 115*(49), 12441-12446. http://www.pnas.org/cgi/doi/10.1073/pnas.1808083115

Cummins, R., Zhang, M., & Briscoe, E. (2016). *Constrained multi-task learning for automated essay scoring.* Association for Computational Linguistics.

Darwish, S.M., & Mohamed, S.K. (2019) Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In A.E. Hassanien, A.T. Azar, T. Gaber, R. Bhatnagar, & M.F. Tolba (Eds.), *The International Conference on Advanced Machine Learning Technologies and Applications* (pp. 566-575). Springer.

DeLuca, D., Willis, J., Cowie, B., Harrison, C., Coombs, A., Gibson, A., et al. (2019). Policies, programs, and practices: exploring the complex dynamics of assessment education in teacher education across four countries. *Frontiers in Education, 4,* 1-19. https://doi.org/10.3389/feduc.2019.00132

Deneen, C.C., & Brown, G.T.L (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education, 3*(1), 1225380. https://doi.org/10.1080/2331186X.2016.1225380

Dillenbourg, P. (2016). The evolution of research on digital education. International *Journal of Artificial Intelligence in Education, 26*(2), 544-560. https://doi.org/10.1007/s40593-016-0106-z

Dong, F., Zhang, Y., Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In R. Levy & L. Specia (Eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 153–162). Association for Computational Linguistics.

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345-356. https://doi.org/10.1080/0969594X.2010.516569

Galindo-Domínguez, H., & Bezanilla, M.J. (2021). Promoting Time Management and Self-Efficacy Through Digital Competence in University Students: A Mediational Model. *Contemporary Educational Technology, 13*(2), ep294. https://doi.org/10.30935/cedtech/9607

Galindo-Domínguez, H., Delgado, N., Losada, D., & Etxabe, J.M. (2024). An analysis of the use of artificial intelligence in education in Spain: The in-service teacher's perspective. *Journal of Digital Learning in Teacher Education, 40*(1), 41-56. https://doi.org/10.1080/21532974.2023.2284726

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for student assessment: a systematic review. *Applied Sciences,* 11, 5467. https://doi.org/10.3390/app11125467

Government of Newfoundland and Labrador (2014). *English Language Arts Grade 6. Appendix D: Sample Elementary Classroom Rubrics and Checklists.* Department of Education of the Government of Newfoundland and Labrador. https://www.gov.nl.ca/education/files/k12_curriculum_guides_english_grade6_300614_g6_ela.pdf

Grivokostopoulou, F., Perikos, I., Hatzilygeroudis, I. (2017). An Educational System for Learning Search Algorithms and Automatically Assessing Student Performance. *International Journal of Artificial Intelligence in Education*, 27, 207–240. http://dx.doi.org/10.1007/s40593-016-0116-x

Hill, M., Ell, F., & Eyers, G. (2017). Assessment capability and student self-regulation: the challenge of preparing teachers. *Frontiers in Education, 2,* 1-15. https://doi.org/10.3389/feduc.2017.00021

Houtao, L., Wenjia, M., Tingting, W., & Chuanhua, X. (2022). The Study of Feedback in Writing from College English Teachers and Artificial Intelligence Platform Based on Mixed Method Teaching. *Pacific International Journal, 5*(4), 147-154. https://doi.org/10.55014/pij.v5i4.270

Hrastinski, S., Olofsson, A. D., Arkenback, C., Ekström, S., Ericsson, E., Fransson, G., Jaldemark, J., Ryberg, T., Öberg, L.-M., Fuentes, A., Gustafsson, U., Humble, N., Mozelius, P., Sundgren, M., & Utterberg, M. (2019). Critical imaginaries and reflections on artificial intelligence and robots in post-digital K-12 education. *Post-Digital Science and Education, 1*(2), 427-445. https://doi.org/10.1007/ s42438-019-00046-x

Jani, K.H., Jones, K.A., Jones, G.W., Amiel, J., Barron, B., & Elhadad, N. (2020). Machine learning to extract communication and historytaking skills in OSCE transcripts. *Medical Education, 54,* 1159–1170. https://doi.org/10.1111/medu.14347

Kasneci, E., Sessler, K., Küchemann, S., ..., Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences,* 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Ke, Z., Inamdar, H., Lin, H., & Ng, V. (2019). Give me more feedback II: Annotating thesis strength and related attributes in student essays. In A. Korhonen, D. Traum & L. Márquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3994-4004). Association for Computational Linguistics.

Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get it scored using autosas—an automated system for scoring short answers. In B. Williams, Y. Chen, & J. Neville (Eds.), *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 9662–9669). AAAI Press.

Liu, M., Wang, Y., Xu, W., & Liu, L. (2017). Automated Scoring of Chinese Engineering Students' English Essays. *International Journal of Distance Education Technologies, 15*(1), 52–68.

Lovorn, M.G., Reza, A. (2011). Assessing the Assessment: Rubrics Training for Pre-service and New In-service Teachers. *Practical Assessment, Research, and Evaluation, 16*(1), 16. https://doi.org/10.7275/sjt6-5k13

Mathias, S., & Bhattacharyya, P. (2018). Thank "Goodness"! A Way to Measure Style in Student Essays. In Y. Tseng, H. Chen, V. Ng. & M. Komachi (Eds.), *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 35–41). Association for Computational Linguistics.

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education, 43*(6), 1-18. http://dx.doi.org/10.14221/ajte.2018v43n6.1

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R.F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE 15,* e0229596. https://doi.org/10.1371/journal.pone.0229596

Ocaña-Fernández, Y., Valenzuela-Fernández, L.A., & Garro-Aburto, L.L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones, 7*(2), 536-568. https://doi.org/10.20511/pyr2019.v7n2.274

Organic Law 3/2020, of December 29th, amending Organic Law 2/2006, of May 3rd, on Education. *Official State Gazette,* 340, 122868-122953. https://www.boe.es/eli/es/lo/2020/12/29/3

Ouguengay, Y.A., El Faddouli, N.-E., & Bennani, S. (2015). A neuro-fuzzy inference system for the evaluation of reading/writing competencies acquisition in an e-learning environnement. *Journal of Theoretical and Applied Information Technology, 81*(3), 600–608.

Owan, V.J., Bekom, K., Emoji, D., Onor, E., & Asuquo, B. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. Modestum. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Ramesh, D., & Kumar, S. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review,* 55, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Redecker, C. (2017). *European Framework for the Digital Competence of Educators: DigCompEdu.* Joint Research Centre. http://dx.doi.org/10.2760/159770

Rhienmora, P., Haddawy, P., Suebnukarn, S., Dailey, M.N. (2011). Intelligent dental training simulator with objective skill assessment and feedback. *Artificial Intelligence in Medicine, 52*(2), 115–121. https://doi.org/10.1016/j.artmed.2011.04.003

Salama, S., & Subahi, A. M. (2020). The Impact of Specialty, Sex, Qualification, and Experience on Teachers' Assessment Literacy at Saudi Higher Education. *International Journal of Learning, Teaching and Educational Research, 19*(5), 200-216. https://doi.org/10.26803/ijlter.19.5.12

Samarakou, M., Fylladitakis, E.D., Karolidis, D., Früh, W.-G., Hatziapostolou, A., Athinaios, S.S., & Grigoriadou, M. (2016). Evaluation of an intelligent open learning system for engineering education. *Knowledge Management & E-Learning: An International Journal, 8*(3), 496–513.

Spear-Swerling, L., Owen, P., & Alfano, M.P. (2005). Teachers' literacy-related knowledge and self-perceptions in relation to preparation and experience. *Annals of Dyslexia, 55,* 266-296. https://doi.org/10.1007/s11881-005-0014-7

Stiggins, R. (2014). Improve assessment literacy outside of schools too. Phi Delta Kappan, 96, 65–72.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Ulum, Ö.G. (2020). A critical deconstruction of computer-based test application in Turkish State University. *Education and Information Technologies, 25,* 4883–4896. https://doi.org/10.1007/s10639-020-10199-z

Vij, S., Tayal, D., & Jain, A. (2020). A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. *Wireles*s Personal Communications, 111(2), 1271–1282. https://doi.org/10.1007/s11277-019-06913-x

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence, 10*(2), 1-37. https://doi.org/10.2478/jagi-2019-0002

Xu, Y., & Brown, G.T.L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teaching and Teacher Education, 58,* 149-162. http://dx.doi.org/10.1016/j.tate.2016.05.010

To cite this work:

Galindo-Domínguez, H., Delgado, N., Sainz de la Maza, M., & Expósito, E. An experimental analysis of the relationship between the evaluations of artificial intelligence and pre-service teachers [Un análisis experimental de la relación entre las evaluaciones proporcionadas por la inteligencia artificial y las proporcionadas por los docentes en formación]. *Edutec, Revista Electrónica de Tecnología Educativa*, (89), 84-104. https://doi.org/10.21556/edutec.2024.89.3509

6. APPENDICES

6.1. Appendix I: Generated texts

Tipo	Nivel	Texto
Texto descriptivo	Alto	En el tranquilo bosque detrás de mi casa, el sol de la tarde se filtra a través de las hojas de los altos árboles, creando un mágico juego de luces y sombras en el suelo. Los pájaros cantan melodías alegres mientras exploran entre las ramas, y el aire está lleno del aroma fresco de la naturaleza. A medida que camino por el sendero cubierto de hojas crujientes, el murmullo del arroyo cercano me guía hacia su orilla, donde el agua fluye suavemente sobre las piedras redondas y pulidas. Cada paso que doy me sumerge aún más en este misterioso mundo, donde la paz y la belleza se entrelazan en perfecta armonía."
	Medio	Me gusta mucho el parque cerca de mi casa. Cuando llego allí, siento una emoción especial. El parque es un lugar dibertido y emocionante para mí. Hay tantas cosas que hacer y explorar. Los columpios son mi parte favorita. Me balanceo alto y siento como si volara. También hay un tobogán grande y emocionante en el que deslizo rápidamente. Mis amigos y yo nos divertimos mucho compitiendo a ver quién llega primero. Además de los juegos, el parque está yeno de árboles altos que proporcionan sonbra en los días calurosos. Me encanta sentarme bajo un árbol con mis amigos y hablar de nuestras aventuras. También hay un pequeño estanque en el que a veces vemos patitos nadando. Cuando tenemos hambre, a veces traemos helado y nos sentamos en los bancos del parque para disfrutarlo. Es tan refrescante y delicioso.
	Bajo	Aier fui a un lugar bonito. Estaba afuera y hacia mucha calor. Me gustó mucho por qué era un día ermoso. Había columpios, un tovogán grande y un par de vancos para sentarse. Jugamos en los columpios, subimos al tovogán y reímos mucho. Yo y mis amigos pasamos un buen rato. Después de jugar, fuimos a una tienda cercana y comimos elado. Yo pedí un elado de fresa por qué es el más rico. La parte más dibertida fue ver cómo se deretía bajo el sol caliente. En general, el día estuvo guay, aunque no recuerdo el nombre esacto del lugar. Haber si volvemos otro dia.
Texto argumentativo	Alto	Creo que debemos cuidar a los animales. Los animales son nuestros amigos y necesitan amor y protección. Si cuidamos de ellos, podemos ayudar a que estén felices y saludables. También es importante no tirar basura en el suelo porque puede lastimar a los animales y contaminar su hogar. Debemos mantener limpio nuestro entorno para que los animales vivan en un lugar seguro. En resumen, debemos ser buenos amigos de los animales y cuidar de nuestro planeta para que todos podamos vivir en armonía.
	Medio	Creo que los videojuegos son divertidos. Muchos niños juegan a los videojuegos y se divierten mucho. A veces, es mejor jugar afuera, pero los videojuegos son buenos para cuando hace frío o llueve. También es importante hacer la tarea antes de jugar a los videojuegos. La tarea es aburrida, pero es importante hacerla para aprender cosas. En resumen, los videojuegos son divertidos, pero la tarea es importante. Los dos pueden ser buenos si los hacemos en el momento adecuado.
	Bajo	Los bideojuegos son guais. Me gusta jugar porque es dibertido. Tengo muchos juegos en mi consola. Algunos son de carreras y otros de abenturas. Me emociona mucho cuando gano. Mis amigos tanbién juegan a bideojuegos. Jugamos juntos y nos dibertimos mucho. Abeces, nos retamos haber quién gana más partidas. Son guais, pero tanvién se que es inportante acer otras cosas, como leer o acer los deveres. Asique, aunque me gusten mucho los bideojuegos, sé que no puedo jugar todo el tienpo.
Texto instructivo	Alto	Cómo Hacer una Cometa Materiales: Papel de seda o papel de envolver Palos de madera Tijeras Hilo fuerte

		Pegamento Cinta adhesiva Pinturas y pinceles (opcional) Instrucciones: Primero, corta cuatro palos de madera de igual longitud para hacer el armazón de la cometa. Dos de ellos serán los lados verticales y los otros dos, los lados horizontales. Une los palos para formar un marco en forma de cruz. Utiliza cinta adhesiva o hilo fuerte para asegurar las uniones. A continuación, corta el papel de seda o papel de envolver en forma de diamante, dejando unos centímetros extra en los bordes para doblarlos y pegarlos al armazón. Pega con cuidado el papel al marco de madera, asegurándote de que quede estirado y sin arrugas. Ahora, corta un trozo de hilo fuerte de aproximadamente un metro de largo. Amarra un extremo al centro del palo horizontal y deja el otro extremo libre. Decorar la cometa con pinturas y pinceles si lo deseas. Puedes hacer diseños creativos en el papel. Cuando hayas terminado de decorar la cometa y el pegamento esté seco, ¡está lista para volar! Encuentra un día con viento suave, ve a un campo abierto y suelta tu cometa al viento sujetando el extremo del hilo. ¡Disfruta viendo cómo se eleva y vuela alto en el cielo!”
	Medio	Cómo Hacer un Sándwich de PB&J Materiales: Pan de molde Mantequilla de maní Mermelada de fresa Cuchillo Instrucciones: Coge dos rebanadas de pan de molde. Abre la mantequilla de maní y la mermelada de fresa. Unta mantequilla de maní en una de las rebanadas de pan. Luego, pon mermelada de fresa en la otra rebanada. Presiona las dos rebanadas juntas para que la mantequilla de maní y la mermelada se mezclen.
	Bajo	Como Hacer Un Pastel: Conprar una mescla para pastel. Agregar uevos y leche. Mezclarlo todo. Poner en un molde. Meter el molde en el orno. Sacarlo cuando este listo.
Texto Narrativo	Alto	Ayer, junto a mis amigos, pasé un emocionante día en el parque. Juntos, construimos un inmenso castillo de arena y nos sumergimos en un emocionante juego de escondidas. Posteriormente, comimos deliciosos helados mientras admirábamos el colorido arco iris que se formó en el cielo. Sin duda, fue uno de los días más sorprendentes que he vivido.
	Medio	Un día soleado, fui al parque con mis amigos. Corrimos y jugamos en los columpios. Después, decidimos explorar el bosque cercano. Seguido, encontramos un arrollo y lanzamos piedras al agua. Después, nos sentamos bajo un árvol a comer sándwiches. Para acabar el día, regresamos a casa, cansados pero felices.
	Bajo	Un día, fui al parke. Jugamos mucho y comimos elado. Luego, fuimos a casa. Fin.

6.2. Appendix II: Rubric used by pre-service teachers to assess the different texts.

Criteria	Excellent (4)	Good (3)	Fair (2)	Poor (1)
Content	Demonstrates a thorough and nuanced understanding of the topic, covering all relevant aspects with insightful analysis and ample supporting details.	Shows a solid understanding of the topic, addressing most key points with adequate analysis and supporting details.	Presents a basic understanding of the topic, though may lack depth or overlook some key aspects.	Understanding of the topic is superficial or confused, with insufficient or irrelevant information provided.
Organization	Presents information in a clear and logical manner, with a well-structured flow of ideas and smooth transitions between paragraphs and sections.	Organizes information effectively, with a mostly coherent flow of ideas and transitions, although some sections may lack clarity or cohesion.	Organizes information adequately but may lack a clear structure or coherent flow, making it challenging for the reader to follow the text consistently.	Organization is poor, with disjointed or illogical sequencing of ideas, hindering comprehension and coherence throughout the text.
Vocabulary & Language	Utilizes a rich and varied vocabulary appropriate to the topic, demonstrating precise language and sophisticated expression, enhancing clarity and engagement.	Uses a diverse vocabulary with generally appropriate language, contributing to clarity and engagement, although there may be occasional inaccuracies or awkward phrasing.	Employs basic vocabulary and language, with occasional inaccuracies or imprecise expression that may detract from clarity or engagement.	Vocabulary and language are limited, repetitive, or inappropriate, hindering comprehension and detracting from the overall quality of expression.
Coherence & Cohesion	Maintains a strong sense of coherence and cohesion throughout the text, with seamless connections between ideas and paragraphs, ensuring a unified and fluid narrative or argument.	Demonstrates adequate coherence and cohesion, with generally clear connections between ideas and paragraphs, although transitions may occasionally be abrupt or unclear.	Exhibits some coherence and cohesion, but connections between ideas and paragraphs are often weak or disjointed, resulting in a fragmented or disjointed text.	Lacks coherence and cohesion, with disjointed or random connections between ideas and paragraphs, making it difficult for the reader to follow the text logically or smoothly.