OmniaScience

Journal of Technology and Science Education

JOTSE, 2025 – 15(2): 479-494 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349

https://doi.org/10.3926/jotse.3135

RASCH-BASED COMPARISON OF ITEMS CREATED WITH AND WITHOUT GENERATIVE AI

Karla Karina Ruiz-Mendoza[1*], Luis Horacio Pedroza-Zúñiga[1]

[1]IIDE-UABC (Mexico)

[2]Doctor, Instituto de Investigación y Desarrollo Educativo (IIDE) (México) & PTC del IIDE de la Universidad Autónoma de Baja California (Mexico)

[*]*Corresponding author: ruiz.karla32@uabc.edu.mx*
horacio.pedroza@uabc.edu.mx

Received October 2024
Accepted April 2025

Abstract

This study explores the evolving interaction between Generative Artificial Intelligence (AI) and education, focusing on how technologies such as Natural Language Processing and specific models like OpenAI's ChatGPT can be used on high-stakes examinations. The main objective is to evaluate the ability of ChatGPT version 4.0 to generate written language assessment items and compare them to those created by human experts. The pilot items were developed for the Higher Education Entrance Examination (ExIES, according to its Spanish initials) administered at the Autonomous University of Baja California. Item Response Theory (IRT) analyses were performed on responses from 2,263 test-takers. Results show that although ChatGPT-generated items tend to be more challenging, both sets exhibit a comparable Rasch model fit and discriminatory power across varying levels of student ability. This finding suggests that Generative AI can effectively complement exam developers in creating large-scale assessments. Furthermore, ChatGPT 4.0 demonstrates a slightly higher capacity to differentiate among students of varying skill levels. In conclusion, the study underscores the importance of continually exploring AI-driven item generation as a potential means to enhance educational assessment practices and improve pedagogical outcomes.

Keywords – Artificial intelligence, ChatGPT, Educational evaluation, Test, Digital technology.

To cite this article:

Ruiz-Mendoza, K.K., & Pedroza-Zúñiga, L.H. (2025). Rasch-based comparison of items created with and without generative AI. *Journal of Technology and Science Education,* 15(2), 479-494. https://doi.org/10.3926/jotse.3135

---

1. Introduction

The emergence of Generative Artificial Intelligence (GAI), as indicated by Bozkurt, Karadeniz, Baneres, Guerrero-Roldán and Rodríguez (2021) and Dimitriadou and Lanitis (2023), is reaching new heights, and in some way or another, both consciously and unconsciously, people are already interacting with these technologies. In the field of education, since 2021 an increase has been seen in the publication of articles on the relationship between AI and education (Bozkurt et al., 2021). The main focus of this work is the

usefulness of AI for developing test items, i.e., comparing those generated with the support of GAI to those designed without the assistance of technology. Although there are few studies on this topic, a review has shown that some language models, such as ChatGPT (from OpenAI) can generate test items with levels of complexity and metric quality similar to those designed by experts, albeit with differences in difficulty and the required cognitive level (Kasneci, Sessler, Küchemann, Bannert, Dementieva, Fischer et al., 2023; Nasution, 2023; Russel-Lasalandra, Christensen & Golino, 2024).

This has demonstrated that Natural Language Processing (NLP), which forms part of computer science and is driven by automatic learning (and now with GAI), offers advanced interaction capable of generating human responses to questions in natural language. ChatGPT, as a Large Language Model (LLM), is trained with large volumes of data, which allows it to understand and answer questions according to the application's policies and the available data (Susnjak, 2022; OpenAI, 2023). This process is accompanied by tokenization, a crucial step in NLP to organize unstructured information in a text form that is suitable for computer processing (Hosseini, Rasmussen & Resnik, 2023). Furthermore, ChatGPT differs from traditional search engines (for example, Google) in that it can understand requests and generate specific responses. In this sense, LLMs could even serve as assistants in first-aid situations (providing Basic Life Support or BLS), although as stated by Aqavil-Jahromi, Eftekhari, Akbari and Aligholi-Zahraie (2025), for the moment they still require the supervision of qualified personnel.

Based on the main premise, ChatGPT can be used in different ways in an educational setting, from asking informational questions to generating multiple-choice exam options. Nasution (2023) mentions that while ChatGPT currently requires explicit instructions to create precise evaluation instruments it cannot be ruled out that someday, with enough data and training, it will be able to generate complex questions autonomously. In his study, Nasution (2023) evaluated the validity and reliability of 21 questions generated by ChatGPT, administered to 272 university students in Indonesia. Following the analysis (a Pearson correlation for validity and Cronbach's alpha for reliability), only one item was discarded, with a Cronbach's alpha of 0.655 and adequate to good discriminatory power. According to Nasution (2023), possible inaccuracies and biases may depend on either the prompt used or the specific version of ChatGPT, since the prompt (Ruiz, 2023) must be formulated based on the same design recommendations for the tests as the standards (American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME), 2014) or other specific criteria (Jornet, González & Suárez, 2010; Instituto Nacional para la Evaluación de la Educación (INEE), 2017).

At the scale level, the AI-GENIE (Rusell-Lasalandra et al., 2024) project was based on simplifying the usually laborious and costly task of scale development by creating a procedure that elaborates thousands of items with LLMs. Discarding redundancies through Unique Variable Analysis, it selects only the most stable ones through Exploratory Graph Analysis (EGA) and bootstrap EGA. In this methodology, five models were used (e.g., GPT 3.5, GPT 4.0 and Llama 3) with different temperature configurations to generate items of the five personality traits (OCEAN), applying carefully designed prompts to improve the quality and diversity of the proposals. The results of a simulation with 75 conditions showed that AI-GENIE produces high-quality items, eliminating a large part of the work required by experts in the initial writing and filtering stage.

On the other hand, other studies have explored the performance of ChatGPT in taking professional exams (Kung, Cheatham, Medenilla, Sillos, De Leon, Elepaño et al., 2023; Choi, Hickman, Monahan & Schwarcz, 2023; Bommarito & Katz, 2023; Liu, Zheng, Du, Ding & Qi, 2023), finding that the model can pass significant sections of these exams. However, these studies are focused on the execution of AI instead of its support in designing the test items, although they serve as a theoretical reference for the growing impact of AI in education. In the context of Item Response Theory (IRT) and the Rasch model, this contribution is especially relevant, as the joint measurement of participants and items on the same latent scale (Prieto & Delgado, 2003) makes it possible to precisely analyze aspects such as

unidimensionality, construct validity and specific objectivity (Ghio, Bruzzone, Rojas-Torres & Cupani, 2020).

These findings suggest that advanced language models are reaching levels of competence similar to that of professionals in training, raising new questions about the future of education and assessment. For these reasons, concerns are also raised about academic integrity, since GAI use for generating (and potentially responding to) test items requires oversight and verification protocols by specialists (Cotton, Cotton & Shipway, 2023). In this sense, Kasneci et al. (2023) discuss the opportunities offered by language models such as ChatGPT in the field of education, including the personalization of learning and support for academic tasks. They also consider the ethical and practical challenges, especially related to its use in creating and taking exams, due to the possible inclusion of cultural and ideological biases. They thus recommend constant human oversight or making opportune adjustments to the prompts (Ruiz, 2023).

Barrot (2023), in turn, offers suggestions for English teachers in the area of writing in the L2. The recommendations include how to integrate ChatGPT into the instructional practice to capitalize on its use. At the same time, he warns about ethical considerations and procedures that are not addressed, for example, in the Standards for Educational and Psychological Testing from the AERA, APA and NCME (2014). These suggestions include emphasizing the value of the writing process, fostering a distinctive voice and identity in writing and using ChatGPT's editing and correction capacities to teach appropriate forms and styles of language.

Advantages	Disadvantages
Adaptive and opportune feedback: ChatGPT can offer personalized practical feedback on writing at any time.	**Inaccurate responses:** The chatbot may produce inaccurate or irrelevant responses to the query.
It is an informational database: With access to a vast knowledge base, ChatGPT can be a valuable source of language input.	**Dependence by students:** There is a concern that students may depend too much on ChatGPT, potentially affecting their creativity and critical thinking.
Generation of coherent and grammatically correct content: ChatGPT can help users refine their writing and improve their language use in various forms.	**Difficulty distinguishing between the student's work and text generated by ChatGPT:** This could complicate the evaluation of writing by instructors.
Assistance in generating topics and organizing ideas: It can generate essay topics and create diagrams in various formats.	**Rigid templates and limited plagiarism verification:** ChatGPT follows specific structures and can have difficulties in adjusting text to a specific audience group or detecting plagiarism.
Correction tool for automated writing: It offers useful functions related to writing assessment, including automatic grading and specific feedback.	**Ethical matters and academic integrity:** The use of ChatGPT poses challenges to academic integrity and writing instruction.

Table 1. Advantages and disadvantages of using Chat GPT in teaching writing (Barrot, 2023)

These recommendations are useful for creating items, since ChatGPT offers opportune adaptive feedback with a large database of information that can help to generate coherent, grammatically correct content, assisting in generating topics and ideas and serving as a correction tool (Barrot, 2023). However, as previously mentioned, there are also disadvantages, such as possible inaccurate answers, student dependence and uncertainty regarding authorship and ethical considerations. Table 1 summarizes the advantages and disadvantages of using ChatGPT for teaching writing, according to Barrot (2023), indicating aspects that also apply to the generation of test items.

The aim of the present work is thus to analyze the metric properties of a set of written language items created with ChatGPT as compared to traditional items, using the Rasch approach as a framework for measurement and psychometric validation. The objective is to provide empirical evidence regarding the potential of AI to assist in developing assessment materials of a quality comparable to the usual ones (those without AI assistance) and even present possible advantages, such as the greater characteristic difficulty reported in certain scenarios (Law, So, Lui, Choi, Cheung, Hung et al., 2025). Finally, we return

to the relevance of generalization inference (Kane, 2006) in this type of study, since if the AI-generated items could retain their metric properties in different applications, this would improve the efficiency of the tests and expand the test item bank without compromising the validity and reliability of the results. The following section reviews the methods used to design, pilot and statistically analyze these items, highlighting the pertinence of the Rasch model in the joint comparison and evaluation of both groups of test items in a specified test (ExIES).

2. Methodology

2.1. Methodological Focus

This study uses an *ex post facto* quantitative design focused on comparing two groups of items: those generated with AI assistance and those designed according to traditional methods. The theoretical framework is based on Item Response Theory (IRT), a probabilistic approach that describes how the probability of correctly answering an item depends on both the latent ability of the examinee and on the characteristics of the test item itself. Selected from within this family of models is the Rasch model, also referred to as 1PL. It assumes that the probability of a correct response is determined by the difference between the skill of the examinee and the difficulty of the item (Wright & Stone, 1979). According to Bond and Fox (2015), this model provides invariant estimates of both the person's ability and the difficulty of the test item, facilitating the comparability among different groups and different items. On a similar note, Tristán (1998) stresses that according to this non-deterministic approach, the probability of answering an item correctly is defined based on the distance between the mean of the personal trait and the difficulty of the item, which is especially relevant for the validity and objectivity of the assessment interpretations.

2.2. Participants and Samples

The sample was comprised of 2,263 examinees who were taking the Higher Education Entrance Exam (ExIES), with an equal distribution of gender (50.06% female and 49.93% male) (See Table 2). This exam was administered in November 2023 as part of a special pilot program at the Institute of Educational Research and Development (IIDE, according to its Spanish initials) at the Autonomous University of Baja California.

Variable	Response	N	%	Mean	SD	p-value
Gender	Female	1.133	50.06	997.8	58.23	0.250
	Male	1.130	49.93	1,000.7	60.46	
	Total	2.263				

Table 2. t-test for independent samples of the variable Gender to confirm the nonexistence of biases. Taken from the Technical Report of the ExIES (2024).

For this study, the items being compared included 28 items created with ChatGPT 4.0 that were added to the traditional bank of test items, which thus consisted of two groups:

1. Items with GAI (n=28).
2. Items without GAI (from a previous bank).

It is important to stress that this version was applied in November 2023, where it had a total of six sub-versions (See Figure 1), with 36 anchor items and 14 pilot items per version (according to the ExIES Technical Report (ExIES, 2024) the "Pilot items are recently designed items for which the evidence of content validity has been demonstrated, but not the metric properties. They are not taken into account when measuring the examinees' performance, rather the sole objective is to field test them under conditions equal or similar to a common application, to assess whether their characteristics support their inclusion in the general test item bank of the instrument") in all areas, and in which 4 or 5 items created with ChatGPT 4.0 were included as part of the pilot items.

Form A	Subversion 1 (5 items with ChatGPT 4.0)	**Form C**	Subversion 4 (5 items with ChatGPT 4.0)
	Subversion 2 (4 items with ChatGPT 4.0)		Subversion 5 (5 items with ChatGPT 4.0)
	Subversion 3 (5 items with ChatGPT 4.0)		Subversion 6 (4 items with ChatGPT 4.0)

Figure 1. Subversions by consolidated form of the ExIES. Technical report from the ExIES (2024)

2.3. Instruments

- ExIES: This exam evaluates Written Language competencies and is designed for large-scale application. It also includes Reading Comprehension and Mathematics sections, however, these were not created with ChatGPT. It includes anchor and pilot items divided into various subversions of the exam. It should be noted that this is a High Impact Exam (HIE) or High-Stakes Exam (AERA, APA & NCME, 2018; Instituto Nacional para la Evaluación de la Educación (INEE), 2017) developed by the Institute of Educational Research and Development (IIDE) of the Autonomous University of Baja California (UABC). This implies constant review in light of the implications and consequences for decision-making (Shepard, 2006).

- Generation of items with GAI: Version 4.0 of ChatGPT was used to generate 28 pilot items, using detailed specifications detailed in the ExIES Written Language manual (ExIES, 2023). Criteria based on the Taxonomy of Anderson and Krathwohl (2001) were followed to ensure that the items reflected the appropriate level of cognitive demand.

2.4. Variables Subject to Study

The dependent variables consist of the following metric indicators, all related to Written Language competence:

- **Difficulty (Rasch parameter)** (Jurado-Núñez, Flores-Hernández, Delgado-Maldonado, Sommer-Cervantes, Martínez-González & Sánchez-Mendiola, 2013; Ghio et al., 2020): This defines the position of the item on the scale; values close to 0.50 suggest a medium level of difficulty.

- **Fit indexes (*Infit* and *Outfit*, MNSQ and ZSTD)** (Jurado-Núñez et al., 2013; Ghio et al., 2020): They check how well the data fit the Rasch model; they are considered adequate if MNSQ ≈ 1.0 and ZSTD falls within the range of ± 2.

- **Point biserial correlation (Ptbis)** (Jurado-Núñez et al., 2013; Ghio et al., 2020): It measures the coherence between the response to the item and the total score; an ideal threshold is > 0.20.

- **Discrimination Index** (Jurado-Núñez et al., 2013; Ghio et al., 2020): This evaluates the capacity of the item to differentiate among examinees of different levels; ≥ 0.40 is excellent, 0.30-0.39 good, 0.20-0.29 marginal and < 0.20 poor.

2.5. Procedure

The examination process is rigorous, so there is a solid basis for its process design, both for the part applying GAI and the human-designed part (Jornet et al., 2010; Kolen & Brennan, 2014; Lane, Raymond & Haladyna, 2016). In light of this, a hybrid process was carried out, where ChatGPT 4.0 was integrated as a designer and judge. As observed in Figure 2, it still went through 10 fundamental steps for its implementation.

After the call for participants and the training of designers and judges, the process continued with the design of the test items. The paid 4.0 version of the GAI ChatGPT (chat.openai.com) was used for item development. Separate conversations were conducted for each item, due to the problems and mistakes that can occur when the chat gets saturated. During this month, the option to create your own chat with specific characteristics was unavailable. In this sense, each item was requested with the characteristics from the Written Language manual (ExIES, 2023), starting with Anderson and Krathwohl's Taxonomy (2001)

for the level of cognitive demand according to the specifications table. For each prompt created, the following was specified:

1. Identification of the content to be evaluated.
2. Description of the content to be evaluated:
   a) Interpretation.
   b) Examples.
   c) Delimitation of contents.
   d) Prior knowledge and skills.
   e) Cognitive activities.
3. Item template:
   a) Base structure of the item.
   b) Characteristics of the text.
   c) Structure and description of the correct response and distractors.
4. Peculiarities of the template:
   a) Item base.
   b) Vocabulary used.
   c) Publishing.
   d) Peculiarities of the distractors.
5. Bibliography consulted.

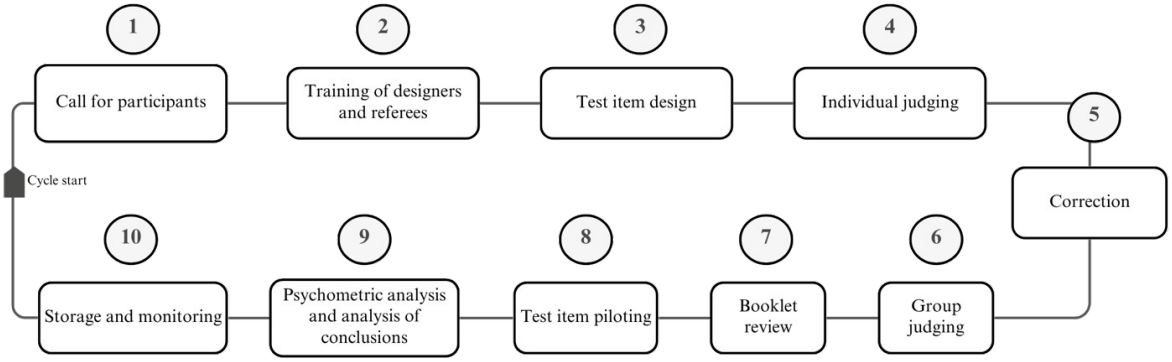

Figure 2. ExIES test development process

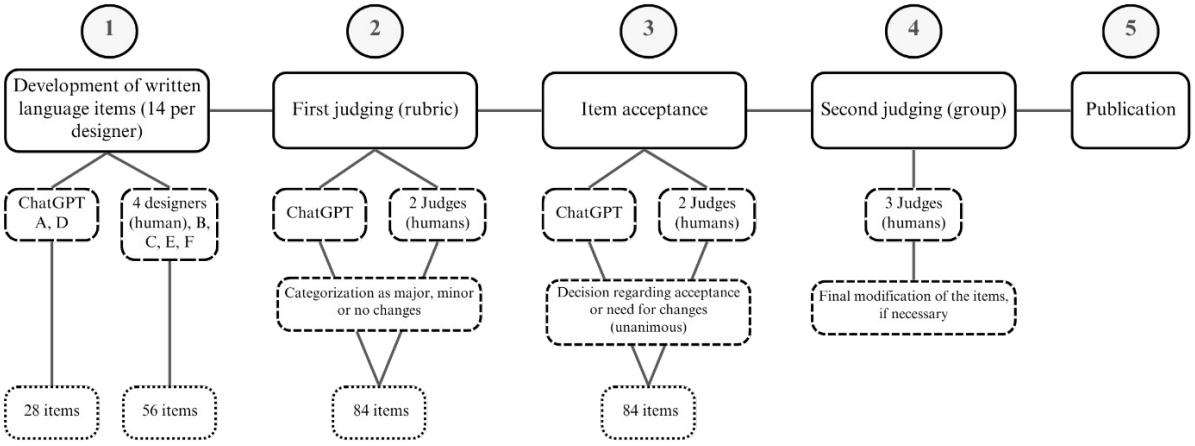

Figure 3. Process of evaluating and reviewing the items

The item creation was followed by an independent judging process, in which ChatGPT 4.0 reviewed and corrected the items. Next came a group judging process, where the item was submitted to two human judges and ChatGPT 4.0, finding no significant differences between humans and the use of ChatGPT in the judging. However, all changes were made by the chat itself, and so the results of this study consider what was said by Bozkurt et al. (2021) regarding symbiosis between humans and GAI (see Figures 2 and 3).

The topics addressed by ChatGPT 4.0 are shown in Table 4. To compare items from the same topic, two topics were randomly chosen to compare humans and GAI. Altogether, 18 items were compared, 9 from humans and 9 created with ChatGPT, using the Student's t-test technique (Field, 2013) and the Rasch analysis described in the ExIES Technical Report (2024).

Topic	Items created with ChatGPT	Items created by humans for comparison
Effective use of semantics: synonyms	5	
Subject-verb agreement	5	5
Punctuation conventions: commas	5	
Punctuation conventions: questions	5	
Language economy	4	
Use of phrases or words in sentences	4	4
Total	28	9

Table 4. Items used for the study. Author's own work

2.6. Data Analysis

The study of the metrics of each application was based on the Rasch model. This non-deterministic probabilistic model predicts the likelihood that a person will select the appropriate response to an item, depending on the discrepancy between the applied stimulus and the individual's attribute level (Tristán, 1998). The t-test and Mann-Whitney U test were run on SPSS for comparisons between the items created by ChatGPT vs. humans.

Below are the results of the items generated by ChatGPT 4.0, following a comparison between those created by humans vs. ChatGPT, and ending with the results of the Student's t-test to confirm whether there is any significant difference.

3. Results

3.1. Results Of The Items Created by ChatGPT 4.0

According to the results in Table 6 for the 28 items generated by ChatGPT 4.0 to evaluate different topics in the area of Written Language, it has been shown that most have a level of difficulty ranging from medium to difficult, which is indicative of an appropriate challenge for the evaluation of the corresponding educational level, in this case, as an entrance exam for higher education. A mean difficulty value of 0.5 is desirable on evaluations (Jurado-Núñez et al., 2013), and several items come close to this score, with the exceptions of 1 and 10. This provides a balance between questions all students can answer and those that can only be answered by those with high ability levels.

In terms of fit, most of the items remained close to the optimal value of 1.0 for *Infit* and *Outfit* MNSQ, suggesting that the responses by the examinees were in line with the expectations of the statistical model. The ZSTD values within the range of -2 and 2 for most items indicate a good fit. 92.86% of the items created by Chat GPT fall within the established parameters. This includes the *Infit* and *Outfit* metrics (MNSQ and ZSTD) (See Table 5).

The items generated by ChatGPT showed compliance with the Infit MNSQ parameters, where 100% of the items fell within the optimal range (0.8 to 1.2), indicating a very adequate fit to the expected model for

these evaluations. For Outfit MNSQ, even though the majority (82.14%) of the items also showed a good fit, 17.86% did not comply with this criterion, which could suggest certain variability in how the items behave with regard to atypical responses from the examinees. As for the ZSTD measures, for both Infit (92.86%) and Outfit (96.43%), the vast majority of the items were within the accepted limits of -2 to 2, indicating a normality in the dispersion of the responses. However, a small percentage fell outside this range, which could reflect potential problems of over- or under-fitting in certain items.

No.	Topic	Cognitive demand	Difficulty	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point-biserial corr.	Discrimination
1	Use of words in sentences	Evaluation	0.39	0.93	-0.8	0.85	-1.1	0.35	1.16
2	Use of words in sentences	Evaluation	0.65	1.13	0.9	1.27	1.4	0.01	0.81
3	Use of words in sentences	Evaluation	0.68	1.07	0.5	1.26	1.1	0.04	0.9
4	Use of words in sentences	Evaluation	0.56	0.96	-0.6	0.96	-0.4	0.32	1.12
5	Language economy	Evaluation	0.56	1.05	0.7	1.07	0.8	0.19	0.82
6	Language economy	Evaluation	0.58	1.07	0.9	1.1	0.9	0.14	0.79
7	Language economy	Evaluation	0.55	1.12	1.8	1.17	1.9	0.05	0.52
8	Language economy	Evaluation	0.53	0.95	-0.9	0.93	-1	0.34	1.23
9	Effective use of semantics	Evaluation	0.6	1.08	1	1.1	0.7	0.13	0.82
10	Effective use of semantics	Evaluation	0.28	0.96	-0.2	0.82	-0.6	0.25	1.05
11	Effective use of semantics	Evaluation	0.77	1.07	0.3	1.3	0.8	0	0.93
12	Effective use of semantics	Evaluation	0.81	1.05	0.2	1.87	1.6	-0.06	0.92
13	Effective use of semantics	Evaluation	0.6	1	0	1.04	0.3	0.23	0.98
14	Subject-verb agreement	Application	0.68	1.05	0.3	1.19	0.8	0.1	0.93
15	Subject-verb agreement	Application	0.51	1.05	0.9	1.06	0.9	0.18	0.75
16	Subject-verb agreement	Application	0.51	0.98	-0.4	0.96	-0.5	0.28	1.13
17	Subject-verb agreement	Application	0.49	0.89	-2.3	0.87	-2	0.42	1.64
18	Subject-verb agreement	Application	0.48	0.87	-2.4	0.84	-2.1	0.45	1.63
19	Punctuation conventions: commas	Application	0.62	1.01	0.1	1.1	0.7	0.2	0.95
20	Punctuation conventions: commas	Application	0.42	0.94	-0.8	0.9	-1	0.33	1.17
21	Punctuation conventions: commas	Application	0.51	0.94	-1.1	0.93	-1	0.34	1.31

No.	Topic	Cognitive demand	Difficulty	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point-biserial corr.	Discrimination
22	Punctuation conventions: commas	Application	0.57	1.07	0.9	1.11	1	0.16	0.79
23	Punctuation conventions: commas	Application	0.58	1.03	0.4	1.08	0.7	0.19	0.88
24	Punctuation conventions: questions	Application	0.43	0.92	-1.1	0.87	-1.2	0.37	1.26
25	Punctuation conventions: questions	Application	0.51	1.01	0.2	1	0.1	0.24	0.96
26	Punctuation conventions: questions	Application	0.44	0.88	-1.9	0.85	-1.8	0.42	1.44
27	Punctuation conventions: questions	Application	0.42	0.88	-1.6	0.84	-1.5	0.42	1.32
28	Punctuation conventions: questions	Application	0.63	1.13	1.2	1.26	1.4	0.03	0.77
Overall mean			**0.55**	**1.00**	**-0.14**	**1.06**	**0.03**	**0.22**	**1.04**

Table 6. Results of the 28 items created by ChatGPT 4.0

Furthermore, means were also calculated for each topic and according to cognitive level to provide a more detailed analysis (see Table 7). The variability in the difficulty of the items points to the complexity of fitting them to a wide spectrum of skills among the participants. According to the Rasch analysis (1960), we understand that the difficulty of an item reflects the interaction between the skill of the participant and the item itself, which highlights the importance of precise calibration for valid and reliable evaluations. For example, the topic "Questions" has a lower difficulty (mean of 0.49) as compared to "Semantics" (mean of 0.61). This shows how the content and focus of the items influence the perceived difficulty, underscoring the need for the items to have a balanced design.

Means	Difficulty	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point-biserial corr.	Discrimination
By topic: Use of words	0.57	1.02	0.00	1.09	0.25	0.18	1.00
By topic: Language economy	0.56	1.05	0.63	1.07	0.65	0.18	0.84
By topic: Semantics	0.61	1.03	0.26	1.23	0.56	0.11	0.94
By topic: Subject and verb	0.53	0.97	-0.78	0.98	-0.58	0.29	1.22
By topic: Comma use	0.54	1.00	-0.10	1.02	0.08	0.24	1.02
By topic: Question marks	0.49	0.96	-0.64	0.96	-0.60	0.30	1.15
Cognitive demand: Evaluation	0.52	0.99	-0.32	1.02	-0.20	0.25	1.08
Cognitive demand: Application	0.52	0.98	-0.51	0.99	-0.37	0.28	1.13
Overall mean	**0.55**	**1.00**	**-0.13**	**1.06**	**0.04**	**0.22**	**1.03**

Table 7. Results by topic and cognitive demand

By considering the fit of the items through the *Infit* and *Outfit* metrics, it is seen that these values reflect the alignment of the items to the theoretical expectations. Items close to 1.00 on *Infit MNSQ* suggest a good fit, indicating predictability and coherence with the participants' skills. However, high values on these metrics, such as the *Infit MNSQ* of 0.00 for "Use of words", imply the need for a detailed analysis to identify fit or replacement needs, thus maintaining the precision of the evaluations.

Likewise, the capacity of the items to discriminate between different levels of skill, evidenced by both the point-biserial correlation and the discrimination indexes, is crucial. This variability in the discrimination, with examples ranging from 0.11 in the case of "Semantics" to 0.30 in the case of "Questions", illustrates the importance of aligning items with clear and relevant educational objectives, as well as with solid psychometric principles.

3.2. Comparative Results between ChatGPT and Humans

As observed in Table 8, the difference in difficulty between the items generated by humans and by ChatGPT 4.0 is noticeable, with the ChatGPT items on average being more difficult. This could suggest that ChatGPT tends to generate questions requiring a higher level of comprehension or skill to answer them correctly, which could be desirable, depending on the objective of the evaluation. The overall results per item can be consulted in Appendix 1. In addition, a *t*-test was conducted to compare the means of *InfitMNSQ* among the items created by humans and by ChatGPT 4.0. The results of this test reveal a t-value of 0.550 and a p-value of 0.590, which indicates that there is no statistically significant difference between the groups as far as the quality of the fit of the items is concerned, according to the *Infit MNSQ*, t=0.550, p=0.590 (there is no significant difference).

The *Infit ZSTD* values reflect the standard deviation of the item's fit to the model. The human items show a slight overfitting, while the ChatGPT items show an underfitting. Ideally, the values should be close to 0. The difference suggests variations in how the items fit the expected model, but neither group shows a very unpredictable or predictable deviation. As with *Infit MNSQ*, the values for *Outfit MNSQ* close to 1.0 are desirable and here we see that both groups are almost equally fit, indicating that the items in both groups show a good overall fit to the model. Likewise, it can also be seen that 86.11% (Table 8) of the items created by humans and those created by ChatGPT are found within the indicated parameters.

Metric	Humans (Overall mean)	ChatGPT 4.0 (Overall mean)	Mean
Difficulty	0.458	0.550	0.5039
Point-biserial correlation	0.229	0.233	0.2339
Discrimination	0.919	1.107	1.0189
% of Items within the parameters	86.11%	86.11%	86.11%

Table 8. Comparative results between humans and ChatGPT 4.0

With regard to the Point-biserial correlation that indicates how the items discriminate among participants of different ability levels, Table 8 indicates that the results are similar between humans and ChatGPT 4.0. That is to say, the items from both groups have a similar capacity to discriminate among participants of different skill levels. This is a key metric in terms of the quality of the items, indicating that both groups produce effective items. In general, the items generated by ChatGPT 4.0 show the greatest capacity to discriminate among these groups, which suggests that they could be especially useful in evaluations.

We also opted to compare the results of humans vs. ChatGPT, with a total of 9 items for each. By comparing the difficulty of the items in the "Word use" and "Subject and Verb" categories, it was discovered that the items generated by ChatGPT 4.0 have a greater level of difficulty than those generated by humans. This suggests that from an item design perspective, ChatGPT 4.0 tends to produce questions that require greater levels of mastery in Written Language by those being evaluated on these two topics, keeping in mind the process the items go through before being published.

In terms of the quality of the items based on the Rasch model fit parameters, such as Infit and Outfit MNSQ, the items "Subject and verb" generated by humans show a fit that is closer to the ideal, indicating that these items are aligned more efficiently with the expectations of the model. This contrasts with the "Word use" items, where the difference in fit between human and ChatGPT items is less pronounced, suggesting a similarity in item quality between the two sources.

Metric	Humans (Word use) | Evaluation	ChatGPT 4.0 (Word use) | Evaluation	Humans (Subject and verb) | Application	ChatGPT 4.0 (Subject and verb) | Application
Difficulty	0.5000	0.5700	0.4240	0.5340
Infit MNSQ	1.0150	1.0225	1.0220	0.9680
Infit ZSTD	-0.2000	0.0000	0.6000	-0.7800
Outfit MNSQ	1.0725	1.0850	1.0000	0.9840
Outfit ZSTD	0.0750	0.2500	0.3000	-0.5800
Point-biserial correlation	0.2250	0.1800	0.2320	0.2860
Discrimination	1.0725	0.9975	0.7960	1.2160

Table 9. Comparison of topics: humans vs. ChatGPT

Another aspect is the variability in the fit, measured through Infit and Outfit ZSTD, which is greater for items generated by humans, especially in the "Subject and verb" category. This would indicate that while the ChatGPT items generally fit the model well, there is a greater inconsistency in how these items behave among different groups of students.

The capacity for discrimination, evaluated through the point-biserial correlation and the discrimination coefficient, is generally higher on human-generated items, with a notable advantage on the "Subject and verb" topic. This means that the items developed by humans are more effective in differentiating among students of different skill levels in this specific category.

3.3. Results of the Student's T-test

In the comparative study of the items generated by humans compared to those generated by ChatGPT 4.0, a Student's t-test was used, complemented by the Mann-Whitney U test to examine differences in several item quality metrics. The results indicate that the only metric that showed a statistically significant difference was difficulty, where the items generated by ChatGPT 4.0 proved to be more difficult as compared to those created by humans ($t = -2.144$, $U = 0.019$, $p = 0.037$), suggesting ChatGPT has a greater capacity for creating questions that pose a greater challenge for the examinees (see Table 10).

Metric	Mean (Humans)	Mean (ChatGPT 4.0)	T statistic	Mann-Whitney U	p-value (Levene)	p-value (two-tailed)	Interpretation
Difficulty	0.458	0.550	-2.144	0.019	0.363	0.037	ChatGPT generates more difficult items.
Infit MNSQ	1.019	0.992	-0.618	0.489	0.323	0.273	There is no significant difference.
Infit ZSTD	0.244	-0.433	1.024	0.436	0.395	0.161	There is no significant difference.
Outfit MNSQ	1.032	1.029	-1.086	1.00	0.185	0.147	There is no significant difference.
Outfit ZSTD	0.200	-0.211	0.632	0.489	0.791	0.268	There is no significant difference.
Point-biserial corr.	0.229	0.239	-0.126	0.931	0.953	0.451	There is no significant difference.
Discrimination	0.919	1.119	-1.145	0.436	0.362	0.135	Minimal tendency towards better discrimination in ChatGPT

Table 10. Results of the Student's t-test

With regard to the other metrics evaluated —Infit MNSQ, Infit ZSTD, Outfit MNSQ, Outfit ZSTD, Point-biserial correlation and Discrimination— the results do not show any statistically significant differences. This suggests that items generated by humans and those generated by ChatGPT 4.0 are comparable in terms of fit to the model and capacity to discriminate among different skill levels in examinees. Specifically, in the case of discrimination, even though no significant difference was found ($t = -1.145$, $U = 0.436$, $p = 0.135$), the results point to a slight trend towards better discrimination by the ChatGPT items, which could imply a marginally better potential for ChatGPT in differentiating among responses by examinees with different levels of competence.

4. Discussion

The results obtained in this study reveal significant aspects concerning the use of GAI, and more specifically, ChatGPT 4.0, in the creation of items for high-impact educational evaluations. First, it provides evidence that the items generated by ChatGPT show greater difficulty as compared to those created by humans. This coincides with the findings by Kung et al. (2023), Choi et al. (2023) and Bommarito and Katz (2023), who reported that ChatGPT can reach or surpass approval thresholds on complex exams, suggesting its capacity to generate highly complex contents that are also challenging. However, it should be kept in mind that these are merely early approximations and it may be that over time we will encounter additional results that help elucidate this type of discussions.

Furthermore, the high rate of items within the optimal fit parameters for the Rasch model indicates that ChatGPT 4.0 is competent in generating items that are aligned with established psychometric standards. This is consistent with the indications of Nasution (2023), who highlights the potential of AI in creating reliable multiple-choice questions. In light of this, these advances are also raising concerns about academic integrity and the validity of evaluations (Cotton et al., 2023). If students can use tools like ChatGPT to get answers on tests, it challenges the ability of traditional assessments to measure learning and competences in an authentic manner. This brings us back to the ethical dilemma, where the main concern will be the need to develop strategies that promote academic integrity and minimize the potential misuse of these tools, as stated by Kasneci et al. (2023).

On the other hand, Liu et al.(2023) indicated that while ChatGPT shows remarkable competencies, it still faces limitations in terms of specific and contextual knowledge. This is in agreement with the findings in the present study. Although the items generated by ChatGPT have a greater level of difficulty and a good Rasch model fit, it is crucial to ensure that they adequately reflect the learning objectives and the cultural and educational context of the students. Therefore, as reflected in the results, the items generated by ChatGPT 4.0 demonstrated a capacity similar to that of human items in terms of Rasch model fit and discriminative capacity. This suggests that GAI may be a useful tool to support the process of developing evaluations, especially when the goal is efficiency and consistency in item generation.

However, it is important to consider the role of expert human judgment in reviewing and validating AI-generated items. As indicated by Ruiz (2023) and Nasution (2023), the quality of the items may depend on the design of the prompts and the version of ChatGPT used. Furthermore, the judging and revision process by human experts continues to be fundamental to ensure the relevance and instructional appropriateness of the items.

5. Conclusions

The incorporation of GAI in education, as indicated by Bozkurt et al. (2021) and Dimitriadou and Lanitis (2023) represents a growing phenomenon that suggests a symbiotic interaction between humans and technology. By analyzing items created by ChatGPT 4.0 on the topic of Written Language, we intend to contribute to a promising field for educational assessment, such as the inclusion of GAI for High Impact Exams. As observed in Tables 9 and 10, the following findings have resulted from this work:

1. We can find significant differences among topics, e.g., some topics may be simpler for ChatGPT and others for humans, but the combination of the two could dissipate these differences.

2. In this study, the ChatGPT 4.0 items showed a level of difficulty generally greater than those created by humans, challenging the perception that ChatGPT makes many mistakes (Barrot, 2023). This is notwithstanding the constant evolution of the Open AI model itself.

3. Rasch model fit: The items generated by ChatGPT 4.0 demonstrate a fit that is closer to ideal, indicative of precise alignment with the theoretical expectations of the model and high-quality items. However, it should be noted that they were subjected to a judging process, and although the final editing was done by ChatGPT itself, it cannot be ruled out that this improved the items.

4. Discrimination capacity: GAI, though ChatGPT 4.0, shows a superior capacity to discriminate among different levels of student ability, thus underscoring its usefulness in designing effective evaluations; with the same consideration as in point 3.

The studies by Nasution (2023) and Russell-Lasalandra et al. (2024) complement these findings by demonstrating the potential of ChatGPT in the creation of multiple-choice items for educational purposes. They also point to the importance of prompt design (Ruiz, 2023) and the specific version of ChatGPT used. The overall results of this study emphasize the value and relevance of using ChatGPT 4.0 as a tool for generating items for educational assessment, demonstrating and reflecting the need for a balanced focus that takes advantage of both the human capacity to create challenging items and the accuracy and effectiveness of GAI to adjust to and discriminate adequately among the students' skills. This highlights the importance of continuing to explore and optimize the use of ChatGPT and other GAI technologies in the educational context, both to improve the quality of the evaluations, and also to enrich educational practices through the effective integration of innovative tools. Finally, later studies will be necessary on the difficulty of items by gender. Similarly, this gives rise to the development of complete tests to then carry out a fully validation process from the Argument-Based Approach based on the seven inferences according to Chapelle (2021; some shared with Kane, 2013): Domain Definition, Evaluation, Generalization, Explanation, Extrapolation, Utilization and Implication of Consequences. Also pending are the development of a system similar to the AI-GENIE project (Rusell-Lasalandra et al., 2024) and the possibility of adequate systemization. Adding GAI to the process could mean changes in each part where a human is involved, leading to improvements in time optimization in test design and administration, as well as in item generation and accuracy. Nonetheless, according to this study, human beings remain necessary to oversee the process and assume responsibility for the prompt design and the results.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Anderson, L.W., & Krathwohl, D.R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition.* New York: Longman.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing. American Educational Research Association.*

Aqavil-Jahromi, S., Eftekhari, M., Akbari, H., & Aligholi-Zahraie, M. (2025). Evaluation of correctness and reliability of GPT, Bard, and Bing chatbots' responses in basic life support scenarios. *Scientific Reports,* 15(11429). https://doi.org/10.1038/s41598-024-82948-w

Barrot, J.S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bommarito, M.J., & Katz, D.M. (2023). GPT Takes the Bar Exam. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4314839

Bond, T.G., & Fox, C.M. (2015*). Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Bozkurt, A., Karadeniz, A., Baneres, D., Guerrero-Roldán, A.E., & Rodríguez, M.E. (2021). Artificial intelligence and reflections from educational landscape: A review of AI studies in half a century. *Sustainability*, 13(2), 800. https://doi.org/10.3390/su13020800

Chapelle, C. (2021). *Argument-Based Validation in Testing and Assessment*. SAGE. https://doi.org/10.4135/9781071878811

Choi, J.H., Hickman, K.E., Monahan, A., & Schwarcz, D. (2023). ChatGPT goes to law school. *Journal of Legal Education*, 71, 387. https://doi.org/10.2139/ssrn.4335905

Cotton, D., Cotton, P.A., & Shipway, J.R. (2023). Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Dimitriadou, E., & Lanitis, A. (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, 10(12). https://doi.org/10.1186/s40561-023-00231-3

ExIES (2023). *Manual de Desarrollo de ítems de escritura para el Examen de Ingreso a la Educación Superior (ExIES)*. IIDE-UABC.

ExIES (2024). *Reporte técnico del Examen de Ingreso a la Educación Superior (ExIES)*. IIDE-UABC.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Ghio, F.B., Bruzzone, M., Rojas-Torres, L., & Cupani, M. (2020). Calibración de un banco de ítems mediante el modelo de Rasch para medir razonamiento numérico, verbal y espacial. *Avances en Psicología Latinoamericana*, 38(1), 123-137. https://doi.org/10.12804/revistas.urosario.edu.co/apl/a.7760

Hosseini, M., Rasmussen, L.M., & Resnik, D.B. (2023). Using AI to write scholarly publications. *Accountability in Research*, 31(7), 715-723. https://doi.org/10.1080/08989621.2023.2168535

Instituto Nacional para la Evaluación de la Educación (INEE) (2017). *Criterios Técnicos para el Desarrollo y Uso de Instrumentos de Evaluación Educativa 2014-2015*. INEE.

Jornet, J.M., González, J., & Suárez, J.M. (2010). Validación de los procesos de determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios Sobre Educación*, 19, 11-29. https://doi.org/10.15581/004.19.4578

Jurado-Núñez, A., Flores-Hernández, F., Delgado-Maldonado, L., Sommer-Cervantes, H., Martínez-González, A., & Sánchez-Mendiola, M. (2013). Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias?. *Investigación en educación médica*, 2(8), 202-210. https://doi.org/10.1016/S2007-5057(13)72713-3

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In Downing, S.M., & Haladyna, T.M. (Eds.), *Handbook of test development* (131-153). Lawrence Erlbaum Associates Publishers.

Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review*, 42(4), 448-457. https://doi.org/10.1080/02796015.2013.12087465

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (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

Kolen, M.J., & Brennan, R.L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). SSBS. https://doi.org/10.1007/978-1-4939-0317-7

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C. et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health,* 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lane, S., Raymond, M.R., Haladyna, T.M. (2016). *Handbook of test development*. Routledge.

Law, A. K.K., So, J., Lui, C.T., Choi, Y.F., Cheung, K.H., Hung, K.K.-C. et al. (2025). AI versus human-generated multiple-choice questions for medical education: A cohort study in a high-stakes examination. *BMC Medical Education,* 25, Article 208. https://doi.org/10.1186/s12909-025-06796-6

Liu, J., Zheng, Y., Du, Z., Ding, R., & Qi, H. (2023). Can ChatGPT Pass the Chinese Civil Service and College Entrance Exams? *arXiv preprint*. Available at: https://arxiv.org/abs/2302.06476

Nasution, N.E.A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, 2(1), em002. https://doi.org/10.29333/agrenvedu/13071

OpenAI (2023). *ChatGPT (versión del 14 de marzo) [Modelo de Lenguaje Grande]*. Available at: https://chat.openai.com/chat

Prieto, G., & Delgado, A.R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, 15(1), 94-100. https://www.psicothema.com

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Ruiz, K. (2023). El uso de ChatGPT 4.0 para la elaboración de exámenes: crear el prompt adecuado : The Use of ChatGPT 4.0 for Test Development: Creating the Right Prompt. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 4(2), 6142-6157. https://doi.org/10.56712/latam.v4i2.1040

Russell-Lasalandra, L.L., Christensen, A.P., & Golino, H. (2024). Generative psychometrics via AI-GENIE: Automatic item generation and validation via network-integrated evaluation. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/fgbj4

Shepard, L. (2006). *La evaluación en el aula*. Instituto Nacional para la Evaluación de la Educación.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? *arXiv.* https://doi.org/10.48550/arXiv.2212.09292

Tristán, A. (1998). *Análisis de Rasch para todos. Una guía Simplificada para evaluadores educativos*. Instituto de Evaluación e Ingeniería Avanzada.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. MESA Press.

Appendix 1

Creation	No.	Difficulty	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Point-biserial corr.	Discrimination
Human	2	0.4200	1.0600	0.7000	1.0800	0.7000	0.1800	0.8400
Human	13	0.4800	0.9400	-1.0000	0.9300	-0.8000	0.3500	1.2900
Human	14	0.4700	0.8900	-1.9000	0.8600	-1.7000	0.4300	1.5000
Human	25	0.6300	1.1700	1.4000	1.4200	2.1000	-0.0600	0.6600
Human	12	0.4900	1.1400	2.3000	1.1500	1.7000	0.0800	0.3400
Human	16	0.3800	0.9900	-0.1000	0.9600	-0.2000	0.2700	1.0300
Human	20	0.3800	0.8500	-1.4000	0.7800	-1.5000	0.4600	1.2400
Human	12	0.4900	1.1400	2.3000	1.1500	1.7000	0.0800	0.3400
Human	16	0.3800	0.9900	-0.1000	0.9600	-0.2000	0.2700	1.0300
Mean (Humans)		**0.458**	**1.019**	**0.244**	**1.032**	**0.200**	**0.229**	**0.919**
ChatGPT 4.0	37	0.39	0.93	-0.8	0.85	-1.1	0.35	1.16
ChatGPT 4.0	37	0.65	1.13	0.9	1.27	1.4	0.01	0.81
ChatGPT 4.0	37	0.68	1.07	0.5	1.26	1.1	0.04	0.9
ChatGPT 4.0	37	0.56	0.96	-0.6	0.96	-0.4	0.32	1.12
ChatGPT 4.0	42	0.68	1.05	0.3	1.19	0.8	0.1	0.93
ChatGPT 4.0	42	0.51	1.05	0.9	1.06	0.9	0.18	0.75
ChatGPT 4.0	45	0.51	0.98	-0.4	0.96	-0.5	0.28	1.13
ChatGPT 4.0	46	0.49	0.89	-2.3	0.87	-2	0.42	1.64
ChatGPT 4.0	42	0.48	0.87	-2.4	0.84	-2.1	0.45	1.63
Mean (ChatGPT)		**0.550**	**0.992**	**-0.433**	**1.029**	**-0.211**	**0.239**	**1.119**
Overall mean		**0.5039**	**1.0056**	**-0.0944**	**1.0306**	**-0.0056**	**0.2339**	**1.0189**

Table 11. Comparison between humans and ChatGPT 4.0

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2025 (www.jotse.org)