Education and Science

Original Article

2025 Supplement 1 1-18

Examining the Performance of Artificial Intelligence in Scoring Students' Handwritten Responses to Open-Ended Items *

Mahmut Sami Yiğiter [1], Erdem Boduroğlu [2]

Abstract

Open-ended items, which have been used as a measurement method for centuries in the evaluation of student achievement, have many advantages, such as measuring high-level skills, providing rich diagnostic information about the student, and not having chance success. However, today, open-ended items cannot be used in exams with a large number of students due to the potential for errors in the scoring process and disadvantages in terms of labour, time, and cost. At this point, Artificial Intelligence (AI) has an important potential in scoring open-ended items. The aim of this study is to examine the scoring performance of AI in scoring students' handwritten responses to open-ended items. In the study, an achievement test consisting of 3 open-ended and 10 multiple-choice items was developed within the scope of the Measurement and Assessment in Education course at a state university. Open-ended items were scored in a structured way (0-1-2), while multiple-choice items were scored as true-false (0-1). 84 participants took part in the study, and the open-ended items were scored by the expert group and the AI tool (ChatGPT-4o). The visual responses written by the students in their handwriting were scored by the AI tool in two different scenarios. In the first scenario, the AI tool was asked to score without giving any scoring criteria to the AI, whereas in the second scenario, the AI was asked to score according to the standard scoring criteria. The findings of the study showed that there were low agreement and correlation coefficients between the AI scores without criteria and expert scores, while there were high agreement and correlation coefficients between the AI scores with standard scoring criteria and expert scores. Similar to these findings, while the item discriminations of the AI scoring without criteria were quite low, the item discriminations of the AI scores with standard scoring criteria were high. In the study, the reasons for the discrepancies between expert scores and AI scores with standard criteria were also investigated and reported. The results show that AI can score handwritten open-ended items with standardized scoring criteria at a good level. In the future, with the development and transformation of AI, it is thought that it can reach scoring accuracy comparable to expert raters in terms of consistency.

Keywords

Open-ended item
Artificial intelligence
AI
ChatGPT
Automated scoring
Handwritten responses
Constructed response item

Article Info

Received: 10.16.2024
Accepted: 01.07.2025
Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14119

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] Social Sciences University of Ankara, Distance Education Application and Research Center, Türkiye, mahmutsamiyigiter@gmail.com

[2] Ministry of National Education, Niğde Measurement and Evaluation Center, Türkiye, erdemboduroglu@gmail.com

Introduction

Measurement and assessment in education has critical roles in educational systems, providing feedback on student learning, offering evidence for students' placement in higher education and guiding educational policies (Lohman, 1993). There are various assessment methods to determine students' achievements. Open-ended items, which enable students to answer independently, have been used for many years and have an important place (Freedman, 1994). Unlike multiple-choice items that provide predetermined answers, open-ended items require students to create their own unique responses, which provides a deeper understanding of students' knowledge, cognitive processes and abilities (Agustianingsih & Mahmudi, 2019; Doğan, 2019). However, open-ended items cannot be used in country-wide exams and large-scale assessments because they are very difficult to assess fairly and require time and effort for scoring (Karadag, Boz Yuksekdag, Akyildiz, & Ibileme, 2020). In open-ended items used in classroom assessments, teachers spend time and effort and are sometimes criticised for subjective scoring (Baykul & Turgut, 2012). Thanks to the scoring support provided by Artificial Intelligence (AI), it is thought to overcome these two problems (Gao, Merzdorf, Anwar, Hipwell, & Srinivasa, 2024).

Today, digital technologies lead to radical changes in many areas of educational processes and require restructuring of these processes (Beksultanova, Vatyukova, & Yalmaeva, 2020; Senkivska, 2022). This change also has an impact on assessment methods, and especially the use of AI-based tools in education is spreading (Owan, Abang, Idika, Etta, & Bassey, 2023). AI, which contributes to the rapid and reliable evaluation of student performances, has the potential to transform measurement and assessment methods in education. In order to adequately understand the potential of AI in this field, it is important to compare it with expert raters (Chen, Chen, & Lin, 2020).

Open-ended items are an important measurement tool that allows students to demonstrate their ability not only to recall information but also to apply, analyse and transfer this information in their own words (Badger & Thomas, 2019; Geer, 1988). Since the evaluation of such questions requires more time and expertise compared to multiple-choice items, it creates a great workload for teachers. In order to reduce this workload of teachers, the use of AI-based tools is becoming increasingly widespread, especially in environments with large groups of students. However, the performance and reliability of AI in scoring open-ended items is still an issue that continues to be argued in the literature (Fernandez et al., 2022; Yaneva, Baldwin, Jurich, Swygert, & Clauser, 2023).

Although the effectiveness of AI in scoring multiple-choice items is widely accepted, the difficulties faced in scoring open-ended items are more complex. Since open-ended items vary in terms of the way students express their thoughts, it may be difficult to develop a standardised criterion for the scoring of such questions (Sychev, Anikin, & Prokudin, 2020). At this point, it is a critical question to what extent AI can accurately and consistently score open-ended items, especially those answered in handwriting. Whether AI can provide results as consistent as expert raters when evaluating these questions is an important research problem (Lin et al., 2020).

Open-Ended Items

Open-ended items are a type of question that allows participants to respond in their own words in line with their own ideas and thoughts without being limited to predefined options or a specific format (Karadag et al., 2020). The outstanding features of open-ended items are to encourage elaboration, exploration, and reflection by allowing participants to express their thoughts, opinions, or experiences in depth (Sarwanto, Fajari, & Chumdari, 2021; Suherman & Vidákovich, 2022). Open-ended items have many advantages over other assessment and evaluation methods in identifying students' learning and enable the measurement of higher-order thinking skills such as evaluate and create (Brookhart, 2010). It forces students to express their thoughts, justify their reasoning and show the extent of their understanding. This depth of understanding provided by open-ended items is often lacking in multiple-choice formats where students may choose the correct answer by guessing or elimination rather than true understanding. On the other hand, the structure of open-ended items encourages

creativity and critical thinking (Fitriyah, Wahyudin, Suhendra, Nurhayati, & Febrianti, 2024; Monrat, Phaksunchai, & Chonchaiya, 2022; Winarso & Hardyanti 2019). In addition, open-ended items give students the freedom to approach the question from various perspectives, develop original answers and offer innovative solutions, and support students' intellectual exploration (Kartikasari, Usodo, & Riyadi, 2022; Septiani, Retnawati, & Arliani, 2022).

Open-ended items provide rich diagnostic information about the student. Teachers can learn about students' misconceptions, problem-solving strategies, and the logical flow of their thinking and provide detailed feedback (Karakaya, 2022). In contrast, multiple-choice items often fail to reveal the logic underlying students' choices, limiting their diagnostic utility. Moreover, open-ended items are more compatible with authentic assessment practices. They simulate real-world tasks in which individuals are required to generate responses, solutions or explanations without being constrained by predefined options. This makes open-ended items useful in assessing students' preparedness for real-life challenges and their ability to apply knowledge in practical situations. Since open-ended items can be partially scored (e.g. 0-1-2), they provide more information than multiple-choice items (0-1) in assessing student achievement and increase the validity of the exam (Karimi, 2014). Since there is no chance success in open-ended items, students cannot obtain unfair scores. In addition, the cheating rate is lower in open-ended items (Abdolreza Gharehbagh, Mansourzadeh, Montazeri Khadem, & Saeidi, 2022). The fact that open-ended items do not contain options eliminates the memory effect. In multiple-choice items with options, students can remember the correct answer with the associations in the options or reach the correct answer by trial and error with the help of the options. Beyond the important advantages of open-ended items listed here, they have many advantages in discovering students' real performances.

Disadvantages and Scoring Difficulties of Open-ended Items

The scoring of open-ended items has several challenges and disadvantages that can significantly affect the reliability and efficiency of assessments. These challenges can be caused by the inherent complexity of open-ended responses and the subjective nature of expert raters. The first of these challenges and disadvantages is subjectivity and rater bias (Hogan & Murphy, 2007; Karakaya, 2022). Human raters are usually needed to score the responses of open-ended items, and these raters may be influenced by their personal tendencies, which may lead to inconsistencies in scoring. Due to this subjectivity, the scores given by different raters may differ and inconsistencies may occur. This may reduce the reliability of scoring (Güler, 2014; Maris & Bechger, 2006). The expertise and subjective judgements of the raters may also affect the scores given to student responses (Baburajan, de Abreu e Silva, & Pereira, 2022). Secondly, scoring open-ended items is not economical in terms of labour, time and cost. Expert raters need to carefully read and score each response, which can be both costly and labour intensive, especially in large-scale assessments. This not only increases the workload of educators, but may also delay the feedback process (Aznar-Mas, Atarés Huerta, & Marin-Garcia, 2023; Sychev et al., 2020). The third is the complexity of scoring the responses of open-ended items. Students' responses to open-ended items require students to answer the items with their own thoughts and ideas. Therefore, students answer the items in line with the thoughts and ideas that they make sense of in their cognitive schemes, and therefore such responses may be more complex. Unlike multiple-choice items, which are scored by automated systems, open-ended items require more detailed analyses to capture the meaning, nuance and context of students' responses. In this case, standardised scoring forms need to be prepared to accurately score the accuracy and quality of responses. The difficulties in the preparation of standardised scoring forms is another challenge in the assessment of open-ended items. In summary, while open-ended items provide a deeper insight into students' comprehension and cognitive skills, they have potential disadvantages in terms of scorer variability, time and effort. These difficulties and disadvantages of open-ended items can be minimised by developing productive AI systems (Mizumoto & Eguchi, 2023; Pinto et al., 2023).

Automatic Scoring of Open-ended Items with AI

AI scores open-ended items with advanced natural language processing (NLP) and machine learning (ML) techniques (Beiting-Parrish & Whitmer, 2023). The scoring process involves several stages combining image processing and predictive analytics. If the responses are handwritten, optical character recognition (OCR) technology is used to convert them into text. OCR systems use deep learning algorithms to achieve high accuracy, especially when dealing with complex handwriting styles. Errors in character recognition are corrected at this stage to ensure accurate and readable text output. After reading, the text is analysed semantically through advanced NLP techniques (Jescovitch et al., 2021). Language models such as Word2Vec, GloVe and BERT analyse responses both at the lexical level and within their contextual framework. These models enable AI systems to assess not only the surface content but also the deeper meaning of the responses (Zhang & Yuan, 2022). Next, the AI scores the responses according to predefined rubrics or scoring tools that outline the scoring criteria. At this stage, classification or regression models are used (Jamil & Hameed, 2023).

Literature Review

In this section, some remarkable findings of some studies in the literature on scoring with AI are presented. Alers, Malinowska, Meghoe, and Apfel (2024) examined the performance of the GPT-4 model for scoring student responses and compared the responses of 105 students with AI and human scores. It states that there is a strong correlation between the two scoring, although there are some inconsistencies. The researchers also report that AI technologies can significantly speed up scoring. Jukiewicz (2024) compared ChatGPT and human scoring in scoring students' programming tasks. This study reports that human scores are higher than ChatGPT scores, but there is a high correlation between the scores. Poole and Coss (2024) investigated the effectiveness of ChatGPT in the assessment of second language essays by comparing the expert and AI tool. The researchers also investigated the effectiveness of different prompts. The findings of this study showed that the scoring quality improved as scoring criteria and sample responses were provided to ChatGPT. Demir (2023) examined the consistency between expert scores and AI scores in scoring students' responses to open-ended items. The findings of this study indicate that there is a high level of agreement and correlation between ChatGPT and expert scores in the scoring of open-ended items. Quah, Zheng, Sng, Yong, and Islam (2024) evaluated the exams of undergraduate dentistry students with three different experts and the AI tool (ChatGPT). The findings of this study indicate that the AI tool has a moderate correlation compared to the experts, and that the AI tends to score more strictly and does not have the ability to give zero points to irrelevant or incorrect content (Quah et al., 2024). Another study states that the variability of handwriting styles and the lack of standardised answer formats further complicate the scoring process and lead to potential errors (Lu, Zhou, & Ji, 2021). In conclusion, it is seen that there are few studies on examining the performance of AI in scoring open-ended items answered in handwriting and the issue is still not fully clarified.

Aim and Importance of the Research

The aim of this study is to investigate the effectiveness of AI in scoring students' handwritten responses to open-ended items. In this context, the performance of the AI in scoring the handwritten open-ended items was analysed under two different scenarios. In the first scenario, the AI was not given any scoring criteria and was only asked to score the student responses according to its own algorithm. In the second scenario, the AI was presented with standard scoring criteria and asked to score based on these criteria. In both cases, the scores given by the AI were compared with the scores given by the expert raters. The findings of the present study will reveal to what extent the AI can provide consistent results, especially when standardised scoring criteria are used. At the same time, by analysing the differences in the scoring of AI compared to expert raters and whether these differences are significant, it will shed light on the areas of improvement needed for a more widespread and reliable use of AI in education. The contributions of AI in this field may enable faster and more effective evaluation processes in education.

Open-ended items have been used for many years due to their advantages such as assessing high-level cognitive skills and providing comprehensive diagnostic feedback. Despite these advantages, they are not used in large-scale assessments because they have disadvantages such as subjective judgements may be involved in scoring and the scoring process requires labour, time and cost. By investigating the potential of AI to provide efficient and consistent scoring, this study may shed light on the wider application of open-ended items in large-scale assessments. In addition, this study focussed on students' handwritten responses to the items. The majority of examinations administered to large samples in Turkey are still administered in paper-and-pencil format as in this study (e.g. LGS, ALES, YKS, etc.). Therefore, the findings of this study may be instructive for Turkey to integrate open-ended items into large-scale assessments in the future.

Research Questions

The questions sought to be answered in the research are presented below:

1. When scoring criteria are not given to the AI tool, what is the level of scoring agreement between AI and expert raters?
2. When standard scoring criteria are given to the AI tool, what is the level of scoring agreement between AI and expert raters?
3. Is there a statistically significant difference between expert raters and AI scoring without criteria?
4. Is there a significant difference between the expert raters' and AI's scoring based on standard scoring criteria?
5. What are the item statistics obtained from the expert raters' scoring according to AI's standard scoring criteria and AI's uncriteria scoring?
6. What are the reasons for incorrect scoring according to the expert scores of AI?

Method

Type of Research

This study was designed as a descriptive research to examine the performance of AI in scoring open-ended items given in handwriting. Descriptive research aims to determine a case or situation as it is without any intervention (Karasar, 2012).

Sample

The sample of the study consisted of 84 undergraduate students taking Measurement and Assessment in Education course at a state university. The sample was determined by the convenience sampling method, which is a method in which individuals with appropriate information for the purpose of the research are selected (Patton, 2002). The handwritten responses of the participant students in this sample to the open-ended items were used as data sources for the analyses.

Data Collection Tool

In order to obtain the research data, an achievement test consisting of a total of 13 items, three open-ended and ten multiple-choice items, was developed. The open-ended items in the achievement test were developed with a structured response (0=False, 1=Partially correct, 2=Fully correct) scoring system, while the multiple-choice items were developed to be scored with a two-category (0=False, 1=Correct) scoring system. In order to ensure the content validity of the achievement test, firstly, a specification table was created and the distribution of the questions was provided according to the weights of the seven subjects determined. Open-ended items were prepared to take place in different subjects. In addition, in the preparation of open-ended items, the item type was structured as a computational or explanation item. A computational item can be defined as a item that requires the student to perform mathematical operations in the solution of the problem, while an explanation item

can be defined as a item that requires the student to explain the solution of the problem with verbal expressions. The first of the open-ended items was prepared as a computational item, the second as a computational + explanation item (one of the two options of the question requires calculation and the other requires explanation), and the third as an explanation item. The cognitive levels of the open-ended items according to the revised Bloom's taxonomy were apply, apply and analyse for questions 1, 2 and 3 respectively. In addition, standard scoring instructions were prepared for open-ended items. The achievement test and standard scoring instructions were sent to three experts to be evaluated in terms of scientific accuracy, readability and content validity of the questions and expert opinions were obtained. In line with the expert opinions, revisions were made on the questions and standard scoring instructions and the final form of the Achievement Test was created. After the form development process, the final form of the Achievement Test was administered to the participant students in paper-and-pencil format. The expert raters consisted of two field experts with a PhD degree in measurement and assessment. After the application of the achievement test, the reliability of the scores made by the two expert raters was tested with agreement and consistency analyses and it was seen that there was a high level of agreement (see Table 1), this finding supports the reliability of the study.

Data Analysis

Two different scoring scenarios were applied in the study. In the first scenario (Scenario-1: AI Scoring Without Criteria), the AI tool was asked to score student responses without any scoring criteria. In the second scenario (Scenario-2: AI Scoring with Standard Scoring Criteria), the AI was presented with standardized scoring instructions prepared in advance and asked to score the responses according to the criteria in these instructions. All AI scoring was conducted between September 17, 2024 and September 20, 2024 with the "gpt-4o-2024-08-06" version of the ChatGPT-4o model. The AI scores in these two scenarios were compared with the scores given by the experts. Percentage of Agreement, Cohen's Kappa, Weighted Quadratic Kappa, Gwet's AC1 statistic, and correlation coefficient were used to assess the agreement and consistency between experts and between experts and AI scores. In addition, an independent sample t-test was applied to determine the difference between the AI scores and expert scores obtained from the two scenarios. If the assumption of homogeneity of variances is violated in the t test, the Welch t test without the assumption of variance equivalence was applied (Aydın, Algina, Leite, & Atılgan, 2018). In addition, Cohen's d effect sizes were calculated and reported to determine the significance of the difference in t-test results. Cohen's d effect size was interpreted as small effect (d = .20), medium effect (d = .50) and large effect (d = .80) (Cohen, 1992). Then, the item discrimination and item difficulty coefficients of the scores obtained from the expert scores and AI scores obtained from two different scenarios were calculated. Item discrimination of multiple-choice items coded as 0-1 was calculated with the Point Biserial Correlation coefficient, while item discrimination of open-ended items scored as 0-1-2 was calculated with the corrected item-total correlation. Item difficulty was calculated by the average correct response rate. Finally, the scoring differences between the expert scores and the AI scores obtained from Scenario-2 were determined and to investigate the reason for these differences, the participants were asked to explain why they assigned this score to the AI tool. The reasons for the differences between the expert and Scenario-2 scores were categorized and reported.

Results

This section contains the findings of the study. Firstly, the agreement between the two experts who scored the open-ended items was examined, and then the relationships between the expert scores and AI scores were analyzed by means of agreement coefficients, t-test and item parameters. Finally, the main reasons for the differences between expert and AI scores were determined.

Agreement between Experts

In the study, two expert raters scored the open-ended items independently. Table 1 shows the agreement coefficients between the two experts.

Table 1. Inter-Expert Agreement Indices

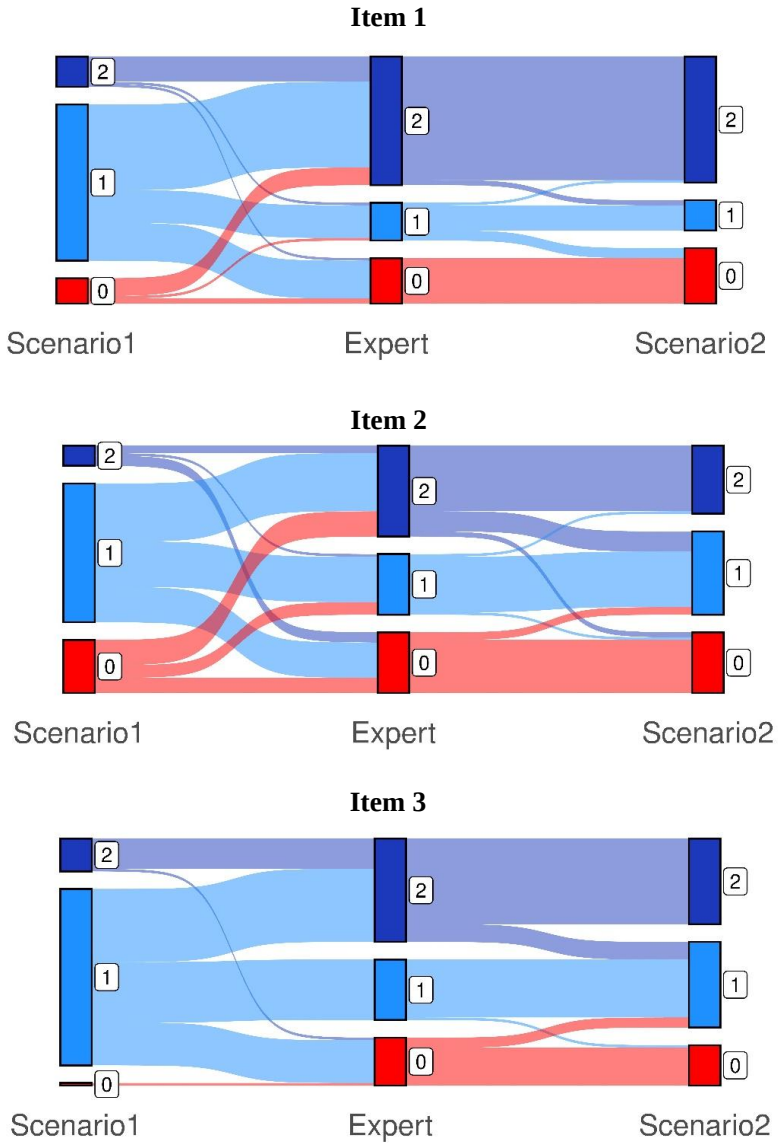
Item	Agreement	Cohen's Kappa	Weighted Quadratic Kappa	Gwet AC1	Correlation
Item 1	0.98	0.98	0.99	0.98	0.99
Item 2	0.95	0.92	0.93	0.93	0.97
Item 3	0.94	0.91	0.91	0.91	0.93
Mean	0.96	0.94	0.94	0.94	0.96

Item 1: Computational, Item 2: Computational + Explanation, Item 3: Explanation

When Table 1 is examined, it is seen that there is a high level of agreement between the two experts according to both agreement and Cohen's kappa coefficients (Landis & Koch, 1977). High agreement coefficients indicate both that the questions can be scored similarly by two different experts and that the scoring criteria are well designed. Each response that showed disagreement between the experts was examined one by one in a meeting organized by the two experts and a common decision was reached. The scores agreed upon by the two experts will be referred to as "Expert Scores" in the remainder of the study.

Agreement between Expert Scores and AI

A total of 252 student responses were obtained from 84 students who answered the three open-ended items in the study. The 252 student responses were scored by the AI tool (ChatGPT-4o model) under 2 different scenarios (252*2= 504 image responses were analyzed). In this section, the relationships between expert scores and AI scores in two different scenarios are analyzed. First, the relationships between expert and AI scores in two different scenarios are presented in Figure 1 with Sankey diagram.



Item 1: Computational, Item 2: Computational + Explanation, Item 3: Explanation, Scenario-1: AI Scoring without Criteria, Scenario-2: AI Scoring with Standard Scoring Criteria.

Figure 1. Agreement between Expert and AI Scores (Sankey Diagram)

Figure 1 shows the scores given by the rater to the open-ended item. The red color indicates that the rater gave a score of "0 = False"; the light blue color indicates that the rater gave a score of "1 = Partially correct"; and the dark blue color indicates that the rater gave a score of "2 = Fully correct". At the bottom of the graphs are the abbreviations of the names of the raters. When the scores for the computational item type named as Item 1 were analyzed, the scores of Expert and Scenario-2 showed a more similar distribution among the scoring categories, while Scenario-1 scores followed a different distribution. These findings indicate that Expert and Scenario-2 scores were more consistent, while Scenario-1 scores differed. Similar to this question, in the item with computational and explanation named as Item 2 and in the explanational item named as Item 3, the scores of Expert and Scenario-2 were similar, while Scenario-1 diverged. When the findings are analyzed in general, in all three items, the category of 2 (Fully correct) scores is more common in the expert scores than in Scenario-1 and Scenario-2 scores. This finding shows that the experts gave more full points according to the two different AI scenarios. On the other hand, in Scenario-1, which is called as AI scoring without criteria, it is seen that the category of 1 (Partially correct) score is more and the category of 0 (False) is less than

the other two scorers. This shows that AI scoring without criteria are quite different from the other two raters.

Table 2 shows the agreement indices of the scores obtained by the experts and two different AIs scoring the open-ended items.

Table 2. Agreement Indices between AI and Expert Scores

Item	Groups	Agreement	Cohen's Kappa	Weighted Quadratic Kappa	Gwet AC1	Correlation
Item 1	Expert -Scenario-1	0.30	0.07	0.07	-0.02	0.09
	Expert-Scenario-2	0.92	0.85	0.94	0.88	0.95
Item 2	Expert-Scenario-1	0.32	0.03	-0.06	0.00	-0.08
	Expert-Scenario-2	0.82	0.73	0.81	0.73	0.82
Item 3	Expert-Scenario-1	0.44	0.18	0.26	0.22	0.34
	Expert-Scenario-2	0.86	0.78	0.88	0.79	0.89
Mean	Expert-Scenario-	0.35	0.09	0.09	0.07	0.12
	Expert-Scenario-2	0.87	0.79	0.88	0.80	0.89

Item 1: Computational, Item 2: Computational + Explanation, Item 3: Explanation, Scenario-1: AI Scoring without Criteria, Scenario-2: AI Scoring with Standard Scoring Criteria.

When Table 2 is examined, it is seen that the agreement indices of the scores between Expert and Scenario-1 are quite low for all three items, while the agreement indices between Expert and Scenario-2 are at medium or high level. This finding indicates that Scenario 1, which is named as Scenario-1, shows that the AI scoring without criteria differed significantly from the expert scores and showed a very low agreement. In other words, it can be said that the reliability of the scores made with AI without criteria is quite low. On the other hand, it is seen that the AI Scoring with Standard Scoring Criteria, which is named as Scenario-2, shows a moderate or high level of agreement with the expert scores. In other words, it can be said that the reliability of the scores made with standard criteria with AI is at a medium or high level.

Differences between Scenario 1, Scenario 2 and Expert Scores

Regarding another research question, the significance of the difference between expert scores and AI scores with two different scenarios was examined with an independent samples t-test. Before applying the t-test, the normality of the data was examined. Hair, Black, Babin, & Anderson (2010) and Byrne (2010) stated that data can be considered normal if skewness is between -2 and +2 and kurtosis is between -7 and +7. Since the skewness and kurtosis values of all variables in the research data were within the specified ranges, it was decided that all data were normally distributed (Hair et al., 2010; Byrne, 2010). When the assumption of homogeneity of variances was examined, it was seen that the assumption of homogeneity of variances was violated between expert and Scenario-1 scores in all items (Levene Test, $p<.05$). The variances were homogeneous between expert and Scenario-2 scores (Levene Test, $p>.05$). In cases where the homogeneity of variances was violated, the Welch t-test, which does not assume homogeneity of variances, was used. t-test findings are presented in Table 3.

Table 3. Differences between Expert and AI Scores

Item	Grup	Mean	SS	t	p	Effect Size
Item 1	Expert	1.393	0.822	3.48	0.00*	0.54
	Scenario-1	1.024	0.514			
	Expert	1.393	0.822	0.45	0.65	0.07
	Scenario-2	1.333	0.869			
Item 2	Expert	1.143	0.838	2.69	0.00*	0.41
	Scenario-1	0.845	0.57			
	Expert	1.143	0.838	0.85	0.39	0.13
	Scenario -2	1.036	0.783			
Item 3	Expert	1.262	0.808	1.22	0.23	0.19
	Scenario-1	1.143	0.385			
	Expert	1.262	0.808	0.39	0.69	0.06
	Scenario-2	1.214	0.746			

Item 1: Computational, Item 2: Computational + Explanation, Item 3: Explanation, Scenario-1: AI Scoring without Criteria, Scenario-2: AI Scoring with Standard Scoring Criteria.

When Table 3 is examined, there is a significant difference between the expert and Scenario-1 scores in two questions, while there is no significant difference in one item. In all three open-ended items, the mean scores obtained from Scenario-1 were lower than the mean expert scores. There is a significant difference in two open-ended items. When the effect sizes are analyzed, there are medium, medium and small effect sizes between Scenario-1 and expert scores, respectively. These findings indicate that AI's scores without criteria differed according to expert ecores. On the other hand, when Scenario-2 and expert scores are compared, it is seen that there is no significant difference between the two score groups in all three items. In addition, the effect sizes between Scenario-2 and expert scores are very close to zero. These findings indicate that scoring with AI's scores with standardized criteria does not differ significantly from expert scores, and therefore, they are similar.

Item Parameters Obtained from AI and Expert Scores

The validity of the scores obtained from two different AI scenarios and experts were examined with the parameters of Classical Test Theory due to the number of samples. The item discrimination and item difficulty parameters of Scenario-1, Scenario-2 and expert scores are presented in Table 4.

Table 4. Item Discrimination and Item Difficulty of AI and Expert Scores

Item		Item Discrimination	Item Difficulty
Item 1	Expert	0.37	0.69
	Scenario-1	-0.01	0.51
	Scenario-2	0.39	0.67
Item 2	Expert	0.47	0.57
	Scenario-1	0.02	0.43
	Scenario-2	0.43	0.52
Item 3	Expert	0.34	0.63
	Scenario-1	0.31	0.57
	Scenario-2	0.42	0.61
Item 4		0.47	0.68
Item 5		0.46	0.80
Item 6		0.61	0.83
Item 7		0.45	0.85
Item 8		0.25	0.88
Item 9		0.45	0.76
Item 10		0.48	0.73
Item 11		0.47	0.67
Item 12		0.49	0.86
Item 13		0.57	0.57

Item 1: Computational, Item 2: Computational + Explanation, Item 3: Explanation, Scenario-1: AI Scoring without Criteria, Scenario-2: AI Scoring with Standard Scoring Criteria

When Table 4 is examined, the item discrimination coefficients of Item 1, Item 2 and Item 3 are -0.01, 0.02, 0.31 and 0.39, 0.43, 0.42 respectively for Scenario-1 and Scenario-2, while the expert scores are 0.37, 0.47 and 0.43 respectively. These findings show that the item discrimination coefficients of Scenario-1 are quite low compared to Scenario-2 and experts. Therefore, it is seen that the validity of the scoring procedures performed with AI's score without criteria is quite low. On the other hand, when the item discriminations of Scenario 2 and expert scores are compared, while the discrimination of Scenario-2 is slightly higher in Item 1 and Item 3, the item discrimination of expert scores is slightly higher in Item 2. This finding shows that the discriminations of Scenario-2 and expert scores are similar to each other and item validity is high.

When item difficulties were analyzed, Scenario-1 had the lowest item difficulty in all three items, followed by Scenario-2 and expert scores. This finding indicates that Scenario-1 gives lower scores than Scenario-2 and experts, while Scenario-2 and experts tend to give higher scores.

Reasons for Differences between AI and Expert Scores

In this part of the study, the reasons for the difference between Scenario 2 and expert scores were investigated. In the previous sections, it was mentioned that AI's performance in scoring without criteria (Scenario 1) was quite inconsistent with expert scores and had low reliability and validity. Therefore, in this section, only 32 inconsistent responses between Scenario 2 and expert scores were analyzed in order to find out the main reason for the differences between AI scoring with Standard Criteria (Scenario 2) and expert scores, and in these responses, it was asked to explain why the AI assigned this score to the instrument. In line with the explanations obtained, the scoring differences were combined under the categories formed and analyzed. The findings are presented in Table 5.

Table 5. Reasons for Differences between AI and Expert Scores

S.N.	Scoring Differences between AI and Experts	f	%
1	Inability to read bad handwriting accurately	4	12.5
2	Not fully understanding the context of the sentence	4	12.5
3	Inability to predict the accuracy of inflated responses	8	25.0
4	In cases where the pencil writes faintly, the answer cannot be read clearly	2	6.3
5	AI's sharper scoring of simple errors	8	25.0
6	Not being able to catch and fully score the meaning in short answers	5	15.6
7	Not understanding the answer indicated by the arrow or symbol	1	3.1

Two of the most important reasons for the difference between AI and expert scores are “Inability to predict the accuracy of inflated responses” (f = 8, % = 25.0) and “AI's sharper scoring of simple errors” (f = 8, % = 25.0). It is seen that there is a difference between AI and expert scores in the scoring of students' inflated answers (arguments that are correct but irrelevant to the content) to open-ended items. For example, in the first option of Item 3, students were asked about the advantages of multiple-choice items in terms of reliability, and Student_62 answered this option as “Reliability is high because when we repeatedly apply multiple-choice tests consisting of multiple-choice items to the participants, we are likely to obtain the same results". This response is similar to the definition of the test-retest reliability type and is an argument that has nothing to do with the question. While AI gave full points to this response, experts evaluated it as false. In the other option, “AI's sharper scoring of simple errors”, while experts are more tolerant to simple mistakes made by students in mathematical operations, AI is more sharp and strict. For example, in Item 1, while the correct answer should be 0/10=0 when calculating the Z score, Student_23 calculated 0/10= 0.10. While the experts gave full points for this answer, considering that the processing error made by the student was not critical for the cognitively measured feature, the AI evaluated this answer as false. In another option, “Not being able to catch and fully score the meaning in short answers”, the AI was not able to fully catch the meaning in the student's short answer to the open-ended item and therefore gave an incomplete score. For example, in one option of Item 2, Student_23 correctly calculated the item difficulty as 0.24 and interpreted it as “The item difficulty is high”. In this response, what the student actually means is that the item difficulty is difficult even

though the item difficulty is numerically low. However, while AI evaluated this answer as false, experts evaluated it as correct. These options are followed by "Inability to read bad handwriting accurately" (f = 4, % = 12.5) and "Not fully understanding the context of the sentence" (f = 4, % = 12.5). In these options, it was observed that the AI could not understand some responses with bad handwriting and sentences with low expression or sentences that were not well expressed semantically were not well understood by the AI. In addition, in the option "In cases where the pencil writes faintly, the answer cannot be read clearly", it was observed that the AI could not read two answers that were written without pressing the pencil enough (f = 2, % = 6.3). Finally, in one response, the student showed his/her answer by drawing an arrow, and while the student should have received a partial correct score with this answer, the AI tool did not interpret the arrow sign and evaluated this answer as incorrect.

Discussion, Conclusion and Suggestions

Recently, there has been an increasing number of studies on the performance of AI in scoring open-ended items. Especially in the context of large-scale assessments, it has gained much attention due to its potential to reduce workload and increase scoring consistency. The literature shows significant findings regarding the reliability, accuracy and limitations of AI-based scoring systems compared to expert raters.

Studies in the literature have shown that AI tools such as ChatGPT exhibit moderate or high levels of correlation and agreement with human raters when scoring open-ended items. This situation shows that AI has the potential to be a reliable tool for scoring, especially in large-scale assessments where human resources are limited (Demir, 2023; Uysal & Doğan, 2021). Xiao, Ma, Song, Xu, Zhang, Wang, and Fu (2024) examined GPT-4 and GPT-3.5-turbo under various approaches to score compositions. It is seen that the weighted kappa values for different configurations are between 0.67 and 0.80, indicating that AI is compatible with expert raters. von Davier, Tyack, & Khorramdel (2022), in their research on automatic scoring of graphical open-ended item responses in TIMSS 2019 using artificial neural networks, stated that these tools can effectively process complex structured item responses and potentially eliminate the need for second human raters. In this study, in line with the literature, it was observed that the scoring by ChatGPT showed a high level of agreement with human raters when standard scoring criteria and detailed rubrics were used. However, in the scoring done by ChatGPT without any criteria, the level of agreement is extremely low. This shows that AI can be a useful tool, but it needs to be used carefully and with human supervision and training.

Differences between expert and AI scores were analyzed by t-test. It was observed that there was a significant difference between expert scores and AI scores without scoring criteria in two of the three items. On the other hand, there was no significant difference between expert scores and AI scores with standardized scoring criteria on any of the three items. This finding shows that AI scores without scoring criteria differ according to expert scores. On the other hand, there is no significant difference between expert scores and AI scores with standardized scoring criteria. When the averages are analyzed, it is seen that expert scores are higher than the AI scores with standardized criteria and without criteria. Jukiewicz (2024) stated that similar to the findings of the current study, many studies in the literature indicated that ChatGPT scores were lower than human scores (Almusharraf & Alotaibi, 2023; Bui & Barrot, 2024; Jukiewicz, 2024).

When the item parameters of the AI scores in the expert and two scenarios were examined, it was seen that this scoring was not valid because the item discrimination parameters obtained from AI scoring without criteria were low. In other words, it was concluded that any scoring without training AI tools with standard scoring tools would produce non-valid results.

This study also investigated the reasons for the differences between the experts' scores and the AI and standardized criterion scores. This analysis revealed seven different specific reasons. The two of these reasons with the highest frequency are that AI scores inflated responses and AI scores simple errors more precisely and clearly. In the studies in the literature, some problems that may be encountered in the process of scoring open-ended items with AI were mentioned. One of these is the "scalability" problem. Current AI models usually require training a separate model for each item, which is not scalable. This may prevent the use of AI as a practical method in situations where the number of items is high and a large amount of computation is required. In addition, AI models may fail to understand contextual relationships in contextualized questions where multiple items are associated with a common reading text. This leads to scoring errors. The types of errors and bias involved in the scoring process are another problem. AI models may exhibit error types and bias that affect the reliability of scoring. This may be due to the training data or the model may not fully understand the context of the responses (Fernandez et al., 2022). Another limitation is that different AI algorithms show different levels of performance in automatic scoring (Uysal & Doğan, 2021). Yaneva et al. (2023) reported that the repeated responses of large language models to the same items differed significantly. He stated that expert validation is needed for the AI tools used. On the other hand, many authors who have conducted research on the scoring performance of AI, similar to the authors of the current study, state that AI tends to score more strictly in scoring performance and does not have the ability to penalize irrelevant, inflated or false content (low scoring) (Bui & Barrot, 2024; Parker, Becker, & Carroca, 2023; Quah et al., 2024).

It is stated that the scoring performance of AI tools may vary depending on the difficulty level of the items and the specific knowledge domain (Zesch, Horbach, & Zehner, 2023). It has been reported that ChatGPT is more likely to give correct answers to items found easier by test takers and performs significantly worse on practice-based items (Yaneva et al., 2023). Demir (2023) reported that AI tools generally showed high correlation and agreement with human raters, but reliability coefficients calculated with more sensitive methods such as generalizability theory were found to be lower. He stated that AI may not always match expert rater standards. In this study, it was observed that the percentage of agreement of AI with expert raters varied according to the characteristics of the items. Especially for items requiring direct calculation, the percentage of agreement and correlation values were found to be higher. On the other hand, the percentage of agreement was relatively lower for items requiring explanation.

The current research has four limitations. First, a limitation of this research is that the researchers used the GPT model instead of developing a scoring systematic using machine learning algorithms and natural language processing methods. There are two main technological requirements for scoring students' handwritten responses. These are optical character recognition and scoring the responses with a natural language processing model. The GPT model was used because it performs these two functions simultaneously. The second limitation of the current study is that the questions in the study were developed to be scored in three categories as 0-1-2. Therefore, the AI's performance in scoring with four or more categories may vary. Therefore, the current study does not provide an idea about the AI's performance in scoring with four or more categories. Third, the use of GPT-4o model in the current research is another limitation of this research. It is thought that the scoring performance of open-ended items may change with the development of new GPT models in the future. The fourth limitation is the small sample size of the current study. Although the item discrimination coefficients calculated with the point two-series correlation coefficient and the corrected item-total correlation provide information about the validity of the achievement test, the lack of factor analytic methods is a limitation of the current study.

AI-based scoring systems for open-ended items show great promise in terms of efficiency and reduction of human effort. However, significant limitations include variability in algorithm performance, scalability issues, bias, and lower reliability compared to human raters. Future work is needed to further improve the performance and applicability of AI-based scoring systems. It is believed that optimizing AI performance would be beneficial to expand its use in educational and psychological measurements. In conclusion, the authors of this study, similar to previous authors (Poole & Coss, 2024; Ramineni & Williamson, 2018), believe that AI tools can be used as "second raters" alongside human raters in educational assessments.

Suggestions for Practitioners

In line with the results of the study, six recommendations for practitioners are presented. First, when scoring with AI, it is very important to establish a clear framework on which criteria of AI will be used in scoring. As seen in the results of this study, scoring with standard criteria (Scenario-2) yielded results close to expert scoring. Secondly, ChatGPT gave points to some answers that were not related to the context of the question or were inflated. In this case, students can exploit this feature of the AI tool. Therefore, it is recommended that practitioners should take precautions and be careful about inflated answers. Third, ChatGPT scored some answers that contained small errors, especially in mathematical expressions, more sharply and clearly and gave the student a low score. Therefore, practitioners should take into account that AI tools reduce the grade for small errors. Fourth, ChatGPT cannot clearly understand the students' handwriting if it is bad or faint, which may cause problems in scoring. In this regard, practitioners can increase the contrast of the writing if the handwriting is faint, warn students if the handwriting is bad, or ask the AI tools to reorganize the writing in a meaningful way before the writing is evaluated. Fifth, it may be useful to have the AI's scores checked by human raters. This would provide an opportunity to review fine details that the AI may have missed. As observed in this study, response inconsistencies due to factors such as inflated responses, faint and unreadable handwriting, etc. should be eliminated. It can be stated that the use of a hybrid system (human + artificial intelligence) in scoring processes will increase scoring accuracy. Sixth, before final scoring of open-ended items with AI, it would be useful to test the accuracy and reliability of the system by conducting small pilot studies and to determine the strengths and weaknesses of the system. Improvements can be made on the scoring key after the pilot study.

Suggestions for Researchers

In this section, eight suggestions are presented to the researchers on the topics that researchers can work on in the future regarding this research topic. First, only ChatGPT was used in this study. Researchers can compare different AI tools such as Google Bard, Microsoft Copilot, Gemini in scoring open-ended items. Secondly, researchers can conduct studies on scoring reliability by having the same response re-scored by the same AI tool. Third, this study examined the performance of AI in scoring open-ended responses in a final exam. In the future, researchers can consider approaches to formative assessment by having the AI provide personalized feedback to students. Fourth, since AI has a learning structure that evolves over time, studies can be conducted in which the same answers are scored repeatedly in certain periods and the consistency between them can be examined. Fifth, the responses where the AI shows inconsistencies with human raters can be examined comprehensively and research can be conducted on the causes and solutions of these inconsistencies. Sixth, studies on the scoring performance of AI in large-scale national and international educational researches (PISA, TIMSS, ABIDE, etc.) in which open-ended items are used and which are not within the scope of high-stakes exams can be conducted. Seventh, in this study, open-ended items were scored in three categories (0-1-2). In the future, researchers can conduct similar studies with more categories. Eighth, in this study, ChatGPT's optical character recognition technology was used to read handwritten responses by the AI tool. Future research can be designed to compare the effectiveness of different OCR tools in terms of scoring.

References

Abdolreza Gharehbagh, Z., Mansourzadeh, A., Montazeri Khadem, A., & Saeidi, M. (2022). Reflections on using open-ended questions. *Medical Education Bulletin, 3*(2), 475-482.

Agustianingsih, R., & Mahmudi, A. (2019). How to design open-ended questions?: Literature review. *Journal of Physics: Conference Series, 1320*(1). doi:10.1088/1742-6596/1320/1/012003

Alers, H., Malinowska, A., Meghoe, G., & Apfel, E. (2024). Using ChatGPT-4 to grade open question exams. In K. Arai (Ed.), *Advances in information and communication* (pp. 1-9). Switzerland: Springer Nature. doi:10.1007/978-3-031-53960-2_1.

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology Knowledge and Learning, 28*(3), 1015-1031.

Aydın, B., Algina, J., Leite, W. L., & Atılgan, H. (2018). *Sosyal bilimler için r'a giriş.* Ankara: Anı Yayıncılık.

Aznar-Mas, L. E., Atarés Huerta, L., & Marin-Garcia, J. A. (2023). Effectiveness of the use of open-ended questions in student evaluation of teaching in an engineering degree. *Journal of Industrial Engineering and Management, 16*(3), 521. doi:10.3926/jiem.5620

Baburajan, V., de Abreu e Silva, J., & Pereira, F. C. (2022). Open vs closed-ended questions in attitudinal surveys-comparing, combining, and interpreting using natural language processing. *Transportation Research. Part C, Emerging Technologies, 137*(12), 103589. doi:10.1016/j.trc.2022.103589

Badger, E., & Thomas, B. (2019). Open-ended questions in reading. *Practical Assessment, Research, and Evaluation, 3*(1), 4.

Baykul, Y. & Turgut, M. F. (2012). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Beiting-Parrish, M., & Whitmer, J. (2023). Lessons learned about evaluating fairness from a data challenge to automatically score NAEP reading items. *Chinese/English Journal of Educational Measurement and Evaluation*, *4*(3). doi:10.59863/nkcj9608

Beksultanova, A. I., Vatyukova, O. Y., & Yalmaeva, M. A. (2020). *Application of digital technologies in the educational process*. Proceedings of the 2nd International Scientific and Practical Conference on Digital Economy (ISCDE 2020).

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*. doi:10.1007/s10639-024-12891-w

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8, 75264-75278.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. doi:10.1111/1467-8721.ep10768783

Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2023*(21), 1072-1099. doi:10.46778/goputeb.1345752

Doğan, N. (2019). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınevi.

Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., & Lan, A. (2022). Automated scoring for reading comprehension via in-context bert tuning. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *International Conference on Artificial Intelligence in Education* (pp. 691-697). Cham: Springer International Publishing.

Fitriyah, Y., Wahyudin, Suhendra, Nurhayati, H., & Febrianti, T. S. (2024). Open-ended approach for critical thinking skills in mathematics education: A meta-analysis. *EduMatSains: Jurnal Pendidikan, Matematika Dan Sains, 9*(1), 156-174. doi:10.33541/edumatsains.v9i1.5975

Freedman, R. L. H. (1994). *Open-ended questioning: A handbook for educators*. Boston: Addison-Wesley.

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, 100206. doi:10.1016/j.caeai.2024.100206

Geer, J. G. (1988). What do open-ended questions measure?. *Public Opinion Quarterly, 52*(3), 365-367.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research, 55*, 73-90. doi:10.14689/ejer.2014.55.5

Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Educational International.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20*(4), 427-441. doi:10.1080/08957340701580736

Jamil, F., & Hameed, I. A. (2023). Toward intelligent open-ended questions evaluation based on predictive optimization. *Expert Systems with Applications*, 231, 120640. doi:10.1016/j.eswa.2023.120640

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150-167. doi:10.1007/s10956-020-09858-0

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, 52, 101522. doi:10.1016/j.tsc.2024.101522

Karadag, N., Boz Yuksekdag, B., Akyildiz, M., & Ibileme, A. I. (2020). Assessment and evaluation in open education system: Students' opinions about Open-Ended Question (OEQ) practice. *Turkish Online Journal of Distance Education*, *22*(1), 179-193. doi:10.17718/tojde.849903

Karakaya, İ. (2022). *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi*. Ankara: Pegem Yayınları.

Karasar, N. (2012). *Bilimsel araştırma yöntemi* (24. bs.). Ankara: Nobel Yayın Dağıtım.

Karimi, L. (2014). The effect of constructed-responses and multiple-choice tests on students' course content mastery. *Southern African Linguistics and Applied Language Studies, 32*(3), 365-372. doi:10.2989/16073614.2014.997067

Kartikasari, S. A., Usodo B., & Riyadi (2022). The effectiveness open-ended learning and creative problem solving models to teach creative thinking skills. *Pegem Journal of Education and Instruction,* 12(4), 29-38. doi:10.47750/pegegog.12.04.04

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lin, Y., Zheng, L., Chen, F., Sun, S., Lin, Z., & Chen, P. (2020). *Design and implementation of intelligent scoring system for handwritten short answer based on deep learning*. IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China. doi:10.1109/ICAIIS49377.2020.9194943

Lohman, D. F. (1993). Learning and the nature of educational measurement. *NASSP Bulletin*, *77*(555), 41-53. doi:10.1177/019263659307755506

Lu, M., Zhou, W., & Ji, R. (2021). Automatic scoring system for handwritten examination papers based on YOLO algorithm. *Journal of Physics: Conference Series,* 2026. doi:10.1088/1742-6596/2026/1/012030

Maris, G., & Bechger, T. (2006). Scoring open ended questions. In *Handbook of statistics* (pp. 663-681). Hollanda: Elsevier.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. doi:10.1016/j.rmal.2023.100050

Monrat, N., Phaksunchai, M., & Chonchaiya, R. (2022). Developing students' mathematical critical thinking skills using open-ended questions and activities based on student learning preferences. *Education Research International,* 2022, 1-11. doi:10.1155/2022/3300363

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. doi:10.29333/ejmste/13428

Parker, J. L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721-727.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks: CA: Sage Publications.

Pinto, G., Cardoso-Pereira, I., Ribeiro, D. M., Lucena, D., de Souza, A., & Gama, K. (2023). *Large language models for rducation: Grading open-ended questions using ChatGPT.* arXiv. doi:10.48550/ARXIV.2307.16696

Poole, F. J., & Coss, M. D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). *Journal of Technology & Chinese Language Teaching*, *15*(1).

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® General Test. *ETS Research Report Series, 2018*(1), 1-31. doi:10.1002/ets2.12192

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, *24*(1). doi:10.1186/s12909-024-05881-6

Sarwanto, F., Widi, L. E., & Chumdari. (2021). Open-Ended Questions to Assess CriticalThinking Skills in Indonesian Elementary School. *International Journal of Instruction, 14*(1), 615-630. doi:10.29333/iji.2021.14137a

Senkivska, L. (2022). The role of digital technologies in education. *Journal of Education, Health and Sport*, *12*(1), 419-423. doi:10.12775/jehs.2022.12.01.036

Septiani, S., Retnawati, H., & Arliani, E. (2022). Designing closed-ended questions into open-ended questions to support student's creative thinking skills and mathematical communication skills. *JTAM (Jurnal Teori Dan Aplikasi Matematika), 6*(3), 616. doi:10.31764/jtam.v6i3.8517

Suherman, S., & Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, 44, 101019. doi:10.1016/j.tsc.2022.101019

Sychev, O., Anikin, A., & Prokudin, A. (2020). Automatic grading and hinting in open-ended text questions. *Cognitive Systems Research,* 59, 264-272. doi:10.1016/j.cogsys.2019.09.025

Uysal, İ., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Egitimde ve Psikolojide Olçme ve Degerlendirme Dergisi, 12*(1), 28-53. doi:10.21031/epod.817396

von Davier, M., Tyack, L., & Khorramdel, L. (2022). *Automated scoring of graphical open-ended responses using artificial neural networks.* arXiv. doi:10.48550/arXiv.2201.01783

Winarso, W., & Hardyanti, P. (2019). Using the learning of reciprocal teaching based on open ended to improve mathematical critical thinking ability. *EduMa: Mathematics Education Learning and Teaching, 8*(1). doi:10.24235/eduma.v8i1.4632

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). *Human-AI collaborative essay scoring: A dual-process framework with LLMs*. arXiv. doi:10.48550/arXiv.2401.06431

Yaneva, V., Baldwin, P., Jurich, D. P., Swygert, K., & Clauser, B. E. (2023). Examining ChatGPT performance on USMLE sample items and implications for assessment. *Academic Medicine, 99*(2), 192-197.

Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement Issues and Practice, 42*(1), 44-58. doi:10.1111/emip.12544

Zhang, D., & Yuan, X. (2022). Intelligent scoring of English composition by machine learning from the perspective of natural language processing. *Mathematical Problems in Engineering*, 2022, 1-9. doi:10.1155/2022/9070272