International Journal of Turkish Educational Studies,
2023, 11 (21), 1072-1099
https://doi.org/10.46778/goputeb.1345752

ISSN: 2148-2314

Investigation of ChatGPT and Real Raters in Scoring Open-Ended Items in Terms of Inter-Rater Reliability

Seda Demir

Assist. Prof. Dr., Tokat Gaziosmanpasa University, Tokat, Türkiye, seddadmr@gmail.com

ABSTRACT

The aim of this study is to examine the inter-rater reliability of the responses to open-ended items scored by ChatGPT, an artificial intelligence-based tool, and two real raters according to the scoring keys. The study group consists of 30 students, aged between 13 and 15, studying in Eskişehir province in the 2022-2023 academic year. The data of the study were collected face-to-face with the help of 16 open-ended items selected from the sample questions published in the International Student Assessment Program-PISA Reading Skills. Correlation, percentage of agreement and the Generalizability theory were used to determine inter-rater reliability. SPSS 25 was used for correlation analysis, Excel for percentage of agreement analysis, and EduG 6.1 for the Generalizability theory analysis. The results of the study showed that there was a positive and high level of correlation between the raters, the raters showed a high level of agreement, and the reliability (G) coefficients calculated using the Generalizability theory were lower than the correlation values and percentage of agreement. In addition, it was determined that all raters showed excellent positive correlation and full agreement with each other in the scoring of the answers given to the short-answer items whose answers were directly in the text. In addition, according to the results of the Generalizability theory, it was found out that the items (*i*) explained the total variance the most among the main effects and the student-item interaction (*sxi*) explained the most among the interaction effects. As a result, it can be suggested to educators to get support from artificial intelligence-based tools such as ChatGPT when scoring open-ended items that take a long time to score, especially in crowded classes or when time is limited.

Article Type
Research

Article Background
Received:
18.08.2023
Accepted:
22.09.2023

Keywords
Inter-rater Reliability,
Scoring Open-Ended Items,
Generalizability Theory,
ChatGPT

To cite this article: Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *International Journal of Turkish Educational Sciences, 11* (21), 1072-1099.

Corresponding Author: Seda Demir, e-mail: seddadmr@gmail.com

Introduction

The main purpose of measurement and evaluation processes is to determine the knowledge, skills and abilities of individuals accurately. For this reason, reliability is considered one of the cornerstones of this process (Doğan, 2021) and is generally defined as the degree of freedom of measurement results from random errors (Turgut, 1993). Errors that cause a decrease in reliability may differ according to the definition of reliability (stability, consistency, sensitivity) or the source of error (individual, item, rater, time, etc.) (Crocker & Algina, 1986; Lord & Novick, 1968). It is seen that raters, who are stated to be an important source of error, have been investigated within the scope of inter-rater reliability in many studies (Atılgan, 2005; Bilgen & Doğan, 2017; Güler & Teker, 2015; Hallgren, 2012; Kan, 2005; Lilford et al., 2007; Mancar, 2019; Park & Kim, 2015; Pekin et al., 2018). Rater reliability is defined as the degree of consistency between the scorings made by more than one rater (Aiken, 2000). Therefore, the difference between the scoring of the raters is a rater-induced error and causes a decrease in rater reliability. In cases where the measurement tool consists of open-ended items, it is expected that the scoring will move away from objectivity. For this reason, the consistency of item and test scores is examined over the items scored by different raters in order to evaluate how much the measurement results reflect the reality and how accurate the decisions made based on these measurement results are (Atılgan et al., 2011). Although there are many methods used to examine inter-rater reliability (Cohen's kappa, weighted kappa, Krippendorf alpha, Kendall's coefficient of concordance, many-facet Rasch measurement model, etc.), in the current study, inter-rater reliabilities were examined with the help of three different methods: Pearson's correlation coefficient, percentage of agreement and the Generalizability theory.

When the literature is examined, it is seen that the percentage of agreement (Güler & Teker, 2015; Gümüş & Arıkan, 2020; Hallgren, 2012; İlhan, 2016; Mancar, 2019) and correlation (Goodwin, 2001; Goodwin et al., 1991; Goodwin & Goodwin, 1991; Kan, 2005) are frequently used to determine inter-rater reliability. In addition, the Generalizability theory has also been used in many studies (Atılgan, 2005; Çakıcı Eser & Gelbal, 2012; Gage et al., 2014; Hill et al., 2012; Pekin et al., 2018).

Pearson correlation coefficient is a method that shows the linear relationship between the scores of two raters, in other words, the consistency of their scoring with each other, and is frequently used in the calculation of inter-rater reliability (Baykul, 2000). The percentage of agreement can be defined as the simple percentage of the number of items on which the raters agree by giving the same score to the same item (Meyer, 1999). In addition to these, in the Generalizability theory, the main effects of individual, item, and rater as well as their interaction effect are taken into account as error sources. Therefore, both potential errors from raters and other main effects and interaction effects can be examined with the Generalizability theory. This is seen as a great advantage of the Generalizability theory (Brennan, 2001).

Unlike the studies in the literature, ChatGPT, an artificial intelligence-based tool, was used as a rater in the current study. ChatGPT is one of the most recent developments belonging to the group of systems known as "chatbot". Chatbots are defined as intelligent systems developed using rule-based or self-learning (artificial intelligence) methods (OpenAI, 2015). After ChatGPT is available to people in late 2022, it is seen that studies have been carried out on its beneficial use in many areas from education to health. When the literature is examined, it is seen that the number of studies (Aktay et al., 2023; Broutin, 2023; Göktaş, 2023; Grassini, 2023; Lo, 2023; Opera et al., 2023; Zileli, 2023) on the

use of ChatGPT in the education process is increasing recent times. These studies generally investigated the role and use of ChatGPT in education, its use in distance education exams and language education. In addition to these, Mizumoto and Eguchi (2023) used ChatGPT to perform automatic text evaluation and to evaluate the reliability and accuracy of the evaluation. As a result of the study, it was emphasized that ChatGPT can provide significant support to real evaluators. In the current study, unlike these studies, open-ended items were scored by ChatGPT and inter-rater reliabilities were examined based on these scores and the scorings made by two Literature teachers. It can be said that this study will contribute to the literature in this respect. In addition, it is thought that the current study is important in terms of providing educators with ideas about both the potential of technology in the educational assessment process and the use of ChatGPT, an artificial intelligence-based tool.

The aim of the current study is to examine the inter-rater reliability of the responses to open-ended items scored by ChatGPT, an artificial intelligence-based tool, and two real raters according to the scoring keys. In line with this main purpose of the study, three sub-problems were identified.

For the scoring of open-ended items, for the scores given by ChatGPT and two real raters who are literature teachers,

1. What is the inter-rater reliability according to the correlation between raters?
2. How is inter-rater reliability according to the percentage of agreement between raters?
3. How is inter-rater reliability according to the Generalizability theory?

Method

Research Design

This study has a descriptive research characteristic since it is aimed at examining the inter-rater reliability coefficients calculated by various methods for open-ended items scored by two real raters and an artificial intelligence-based tool. In descriptive research, it is aimed to explain the situation as fully, detailed and carefully as possible (Büyüköztürk et al., 2011).

Study Group

The study group of this study consisted of 30 students who were selected on a voluntary basis among 13-15 year-old students studying in a public school in Eskişehir province in the 2022-2023 academic year. The responses of 30 students to 16 open-ended items were scored by three raters. The rater group consisted of the GPT-4.0 licensed version of ChatGPT, an artificial intelligence-based tool, and two volunteer Literature teachers who have been working in different schools for 8 and 10 years.

Data Collection

This study was approved by the Ethical Committee of Tokat Gaziosmanpasa University Social and Human Sciences Researches dated 13.06.2023 and numbered 10.15. The data of the study were collected face-to-face with the help of 16 open-ended items selected from the sample questions

published in the International Student Assessment Program-PISA Reading Skills and presented in Appendix-1. The maximum score that could be obtained from the items in total was determined as 100. Students' responses to the open-ended items in the questionnaire were scored by ChatGPT, an artificial intelligence-based tool, and two volunteer Literature teachers as Rater 1 (R1) and Rater 2 (R2). In scoring, a holistic rubric created by the researcher in line with the published scoring criteria for the selected PISA questions was used. Thus, the data set used in the study was obtained by the raters independently scoring each student's responses to the 16 items. The descriptive statistics calculated over the total scores given by the raters to each question are presented in Table 1.

Table 1

Descriptive Statistics of Scoring by ChatGPT, R1 and R2

Rater	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
ChatGPT	16	5	225	98.88	55.20	.429	.485
R1	16	5	150	82.19	41.41	-.115	-.581
R2	16	5	150	86.31	44.29	-.202	-.921

According to Table 1, in terms of the total scores given to each item, ChatGPT has a higher mean compared to the other two raters ($\bar{X}$= 98.88; SD= 55.20). Skewness and kurtosis values are in the range of [-1.5, +1.5]. Accordingly, it can be said that the scorings show a normal distribution (Tabachnick & Fidell, 2014).

Data Analysis

In the current study, correlation, percentage of agreement and the Generalizability theory were used to determine the scoring reliability between the raters who scored each student's responses to the 16 items. SPSS 25 was used for correlation analysis, Excel for percentage of agreement analysis and EduG 6.1 for the Generalizability theory analysis.

The Pearson correlation coefficient has assumptions such as the data being at least at an equal interval scale level and showing a normal distribution. In addition, correlation values less than .30 indicate a low relationship between two variables, values between .30 and .70 indicate a medium relationship, and values greater than .70 indicate a high relationship (Büyüköztürk et al., 2011). Since the research data were continuous, Pearson correlation coefficient was preferred, and inter-rater correlations were calculated in paired combinations. However, this method does not include information about the percentages of agreement of the raters and the variance between the raters (Güler & Teker, 2015; Şencan, 2005).

The percentage of agreement was calculated as the percentage of raters giving the same score to the same item in pairwise combinations. It is an easy calculation and interpretation method that can be used for all scale-level data. However, an important limitation of the method is that it ignores chance or coincidental agreements between raters' scorings (Goodwin, 2001). In order to determine the existence of inter-rater reliability, the percentage of agreement should be above 75% (Şencan, 2005).

In the Generalizability theory analyses, which is another method used to determine inter-rater reliability, the measurement object of the study was students (s) and the surfaces were items (i) and raters (r). Analyses were conducted using a completely crossed-random design (sxixr). In order to make comparisons, the analyses were repeated over the binary combinations of raters and all three

raters, and comparisons were made.

Findings

In the current study, the Pearson correlation coefficient values calculated over the scores given by ChatGPT, an artificial intelligence-based tool, and two Literature teachers to student responses to 16 open-ended questions are given in Table 2.

Table 2

Inter-rater Reliabilities through Correlation Coefficient

Item	Correlation between ChatGPT and R1	Correlation between ChatGPT and R2	Correlation between R1 and R2
1	.47*	.51*	.93*
2	.70*	.74*	.90*
3	1.00*	1.00*	1.000*
4	.62*	.62*	1.000*
5	[a]	[a]	[a]
6	.66*	.66	1.00*
7	1.00*	1.00*	1.00*
8	1.00*	1.00*	1.00*
9	.84*	.84*	1.00*
10	.72*	.74*	.92*
11	.84*	.84*	1.000*
12	.94*	.94*	1.000*
13	1.00*	1.00*	1.000*
14	.94*	.94*	1.000*
15	.81*	.94*	.76*
16	.87*	.94*	.93*
Test	.82*	.86*	.94*

Note. [a] Cannot be computed because at least one of the variables is constant.

*p< .001

When Table 2 is examined, it is seen that there is a perfect positive correlation (1.00) between ChatGPT and R1, ChatGPT and R2 and R1 and R2 for item 3, item 7, item 8 and item 13. In addition, the lowest correlations between ChatGPT and R1 and ChatGPT and R2 were calculated as .47 and .51 for item 1, respectively. The lowest correlation between R1 and R2 was observed for item 15 (.76). Therefore, it can be said that the correlation values calculated for inter-rater agreement vary from item to item. According to the calculated correlation values, it can be said that R1 and R2 showed a high correlation in all items in terms of their scorings. According to the correlation values between ChatGPT and R1 and ChatGPT and R2, it is seen that there is a moderate relationship only in item 1, item 4 and item 6, while the relationship in the other items is quite high. In addition, since all raters gave the same score to item 5, inter-rater correlations could not be calculated. It was also found that there were high correlations between the mean scores of the raters (.82, .86, and .94).

Inter-rater percentages of agreement are presented in Table 3.

Table 3

Inter-rater Reliabilities through Percentage of Agreement

Item	Percentage of Agreement between ChatGPT and R1	Percentage of Agreement between ChatGPT and R2	Percentage of Agreement between R1 and R2
1	50.00	53.33	90.00
2	36.67	26.67	90.00
3	100.00	100.00	100.00
4	80.00	80.00	100.00
5	100.00	100.00	100.00
6	80.00	80.00	100.00
7	100.00	100.00	100.00
8	100.00	100.00	100.00
9	63.33	63.33	100.00
10	73.33	76.67	96.67
11	93.33	93.33	100.00
12	96.67	96.67	100.00
13	100.00	100.00	100.00
14	96.67	96.67	100.00
15	53.33	93.33	53.33
16	93.33	96.67	96.67
Mean	82.29	84.79	95.42
Standard Deviation	21.04	21.08	11.73

The second, third, and fourth columns in Table 3 show the percentages of items that both raters gave the same score. Accordingly, in parallel with the data presented in Table 2, ChatGPT and R1, ChatGPT and R2, and R1 and R2 gave the same score to the same items in items 3, 5, 7, 8, and 13. As a remarkable finding obtained from the study, it is seen that the percentage of agreement with R1 for item 2 is approximately 37% and the percentage of agreement with R2 is approximately 27% when one of the raters is ChatGPT. In the same question, the agreement between R1 and R2 was calculated as 90%. Accordingly, it can be said that the percentage of agreement between raters varies from item to item. In addition, the mean percentage of agreement was calculated as 82.29 (SD= 21.04), 84.79 (SD= 21.08), and 95.42 (SD= 11.73) for ChatGPT and R1, ChatGPT and R2 and R1 and R2, respectively.

The Generalizability theory was also used to determine inter-rater reliability. While the source of error in other approaches is only the differentiation between the scores of the raters, in the Generalizability theory, different sources of error are considered at the same time. These error sources can be listed as item (*i*), rater (*r*), student-item interaction (*sxi*), student-rater interaction (*sxr*), item-rater interaction (*ixr*) and student-item-rater interaction (*sxixr*) for the current study. The results of the analysis of variance and the estimated variance components obtained from the analyses conducted for the binary combinations of raters and three raters are presented in Table 4.

Table 4

Analysis of Variance Results and Estimated Variance Components

Rater	Source of Variance	SS	df	MS	Estimated Variance Component	Percentage of Variance (%)	G Coefficient
ChatGPT and R1	student (*s*)	847.11	29	29.21	0.55	5.9	.61
	item (*i*)	2124.38	15	141.63	1.92	20.5	
	rater (*r*)	74.26	1	74.26	0.12	1.3	
	sxi	4735.71	435	10.89	4.69	50.3	
	sxr	60.90	29	2.10	0.04	0.4	
	ixr	256.36	15	17.09	0.52	5.6	
	sxixr, e	651.99	435	1.50	1.50	16.0	
ChatGPT and R2	student (*s*)	848.29	29	29.25	0.54	5.8	.59
	item (*i*)	2296.81	15	153.12	2.16	23.1	
	rater (*r*)	42.08	1	42.08	0.06	0.6	
	sxi	4780.22	435	10.99	4.86	51.9	
	sxr	64.88	29	2.24	0.06	0.6	
	ixr	207.50	15	13.83	0.42	4.5	
	sxixr, e	554.03	435	1.27	1.27	13.6	
R1 and R2	student (*s*)	827.92	29	28.55	0.52	6.1	.58
	item (*i*)	1779.90	15	118.66	1.72	20.1	
	rater (*r*)	4.54	1	4.54	0.00	0.0	
	sxi	5212.91	435	11.98	5.78	67.6	
	sxr	7.78	29	0.27	-0.01	0.0	
	ixr	58.10	15	3.87	0.12	1.3	
	sxixr, e	181.59	435	0.42	0.42	4.9	
ChatGPT, R1 and R2	student (*s*)	1239.40	29	42.74	0.54	5.9	.61
	item (*i*)	3013.55	15	200.90	1.93	21.3	
	rater (*r*)	80.59	2	40.29	0.06	0.6	
	sxi	7133.16	435	16.40	5.11	56.3	
	sxr	89.04	58	1.54	0.03	0.3	
	ixr	347.97	30	11.60	0.35	3.9	
	sxixr, e	925.07	870	1.06	1.06	11.7	

Note. SS: Sum of Squares, MS: Mean Square, df: degrees of freedom

According to Table 4, when the student, item, and rater main effects are examined when the raters are ChatGPT and R1, ChatGPT and R2, R1, and R2 or ChatGPT, R1, and R2, it is seen that the variability related to the item (*i*) has the highest value among the main effects and explains 20.5%, 23.1%, 20.1% and 23.1% of the total variance, respectively. This shows that the difficulty levels of the items differ from each other. The variances related to students (*s*) explain 5.9%, 5.8%, 6.1%, 6.1%, and 5.9% of the total variance, respectively. This value, which indicates that there is a difference between students, is expected to be as high as possible (Brennan, 2001). In addition, the variance related to the main effect of rater (*r*) explains 1.3% of the total variance when the raters are ChatGPT and R1, 0.6% when ChatGPT and R2, and 0.6% when ChatGPT, R1, and R2. Since this value expresses inter-rater variability, it is desired to be as close to zero as possible (Brennan, 2001). In this respect, when raters are R1 and R2, the variance related to the rater main effect is 0.0%, which meets the expectation and is a desired situation.

When Table 4 is analyzed in terms of student, item, and rater interactions when raters are ChatGPT

and R1, ChatGPT and R2, R1 and R2 or ChatGPT, R1 and R2, it is seen that student-item interaction (*sxi*) has the highest value among the interaction effects. Accordingly, the variance related to the student-item interaction (*sxi*) explains 50.3%, 51.9%, 67.6%, and 56.3% of the total variance, respectively. Therefore, the difficulty level of the items shows a significant variability according to the students. It is seen that the variance related to student-rater interaction (*sxr*) is quite low (0.4, 0.3, 0.6) in all rater groups and zero when the raters are R1 and R2. This indicates that the scoring of students' responses to the questions did not change from rater to rater. In addition, the variances related to item-rater interaction (*ixr*) correspond to 5.6%, 4.5%, 1.3%, and 3.9% of the total variance, respectively. It was determined that this interaction, which shows the change in the scores given to the items from rater to rater, had the highest value when the raters were ChatGPT and R1, and the lowest value when the raters were R1 and R2. Therefore, it can be said that the raters who differed the most from each other in terms of their scoring were ChatGPT and R1 raters, while the raters who differed the least were R1 and R2 raters. In addition, the variances related to the student-item-rater common effect (*sxixr,e*), which is also expressed as residual or error variance and which is desired to be as close to zero as possible, were calculated as 16%, 13.6%, 4.9%, and 11.7%, respectively. Here, the highest value was calculated when the raters were ChatGPT and R1, and the lowest value was calculated when the raters were R1 and R2.

The last column of Table 4 shows the generalizability (G) coefficient calculated when the raters were ChatGPT and R1, ChatGPT and R2, R1 and R2, or ChatGPT, R1 and R2. This coefficient, which is an indicator of the reliability or generalizability of the scores, takes values between 0.0 and 1.0 (Shavelson & Webb, 1991). For 30 students, 16 items, and two raters, the G coefficients were calculated as .61 (ChatGPT and R1 raters), .59 (ChatGPT and R2 raters), and .58 (R1 and R2 raters), respectively, and for three raters, .61 (ChatGPT, R1, and R2 raters).

Discussion and Conclusion

In the current study, inter-rater reliabilities were determined over the open-ended items scored by three raters, ChatGPT and two Literature teachers. At this stage, the raters were grouped in pairs and trios (ChatGPT and R1, ChatGPT and R2, R1, and R2, ChatGPT, R1, and R2) and inter-rater reliability was determined with the help of correlation, percentage of agreement and the Generalizability theory. The results obtained from the analyses were found to support each other in terms of the general conclusions reached.

The correlation values calculated between the raters show that there is a positive and high level relationship between the raters (.82 for ChatGPT and R1; .86 for ChatGPT and R2; .94 for R1 and R2). In support of this finding, it is seen that high correlations were obtained for inter-rater reliability in many studies in the literature (Goodwin, 2001; Goodwin & Goodwin, 1991; Goodwin et al., 1991; Güler & Teker, 2015; Öksüzoğlu, 2022; Özşavlı, 2023; Seheryeli, 2018; Wilson et al., 2022). In addition, as an important finding of the study, all raters showed excellent positive correlation (1.00) with each other in the scoring of the answers given to the short-answer items (item 3, item 7, item 8, and item 13) whose answers were directly in the text. Moreover, when one of the raters was ChatGPT (ChatGPT and R1 or ChatGPT and R2), the lowest inter-rater correlation was calculated for item 1. In this item, it is asked to write one of the sub-goals for the given text other than the main goal that is clearly stated in the text. Therefore, it can be interpreted that the scoring made by the real raters

(R1 and R2) was more compatible with the scoring of the responses for determining the sub-goals that were not as clearly stated as the main goal in the text. In addition, as a remarkable finding of the study, the correlation between ChatGPT and real raters was higher in item 15, where the lowest correlation between real raters (R1 and R2) was observed. When this item was analyzed, it was seen that the text used a complex language that reflected the character's inner monologues and thought processes and required the reader to think and use imagination. Based on this, it can be interpreted that while the differentiation between the real raters increased in the scoring of the items whose answers were based on imagination and creativity, ChatGPT's scoring was close to the average of the real raters. In addition, all raters gave one and the same score to the 15th item in which the answer to the question was stated in the text in a remarkable, short, and clear way. Therefore, it is possible to say that the raters showed complete agreement in the short-answer and open-ended items whose answer was clearly emphasized in the text. However, it should be noted that the correlation used in the calculation of inter-rater reliability does not show the similarities between the scorings of two raters or their strictness/generosity since it is calculated independently of the mean. Therefore, it may be insufficient in determining inter-rater reliability (Goodwin, 2001).

As a remarkable finding obtained from the study, it was observed that the scoring made by the real raters showed a higher correlation with each other and the percentages of agreement were also higher. In addition, it was observed that the lowest percentage of agreement between ChatGPT and real raters was calculated for item 2. However, the percentage of agreement between the real raters for this item is quite high (90%). When item 2, which was based on the given text, was examined, it was seen that the item was two-stage, and it was asked to write information in the text as an answer and then to support the answer with the expressions in the text. Therefore, it can be interpreted that the agreement between real raters is higher in the scoring of open-ended items that can be characterized as two-stage, while the scoring made by ChatGPT differs from real raters. The most significant limitation of the percentage of agreement method, which provides convenience in term0s of calculation and interpretation, is that it does not take into account the agreement of the raters that occurs by chance (Güler & Teker, 2015). At this point, it can be suggested to use Cohen's Kappa statistic (Cohen, 1960), which also takes into account the artificial agreement between raters due to chance, in the calculation of the reliability coefficient.

The Generalizability theory addresses different error sources at the same time and provides detailed information. This feature makes the Generalizability theory superior to other methods used in the study. In the present study, it is seen that the G coefficients obtained by using the Generalizability theory are lower than the correlation values and percentages of agreement (.61 for ChatGPT and R1; .59 for ChatGPT and R2; .58 for R1 and R2; .61 for ChatGPT, R1 and R2). However, the fact that no significant difference was observed between different rater groups supports the findings obtained from other methods used to determine inter-rater reliability. When the analyses conducted within the scope of the Generalizability theory are examined, it is seen that important findings were reached. When the main effects of student (*s*), item (*i*) and rater (*r*) were analyzed, it was determined that the variability related to the item had the highest value among the main effects. Therefore, the difficulty levels of the items used in data collection differed from each other. In addition, in line with the expectation, it can be said that the main effect for students indicates that students differ from each other, albeit slightly, and the main effect for raters indicates that there is no variability among raters. It was observed that the variance related to student-item interaction (*sxi*) explained a large portion of the total variance in all rater groups. Accordingly, the difficulty level of the items shows

a significant variability according to the students. According to Güler and Teker (2015), this situation is more likely in areas such as mathematics and statistics where students' past learning experiences are effective. However, as another striking finding of the study, it was determined that the difficulty level of the items in the questions related to students' reading skills also showed a significant variability according to the students. This may have been due to the fact that the questions in the PISA Reading Skills area are aimed at different cognitive processes such as fluent reading, reading comprehension, evaluation, and reflection. The fact that the variance related to student-rater interaction (*sxr*) was quite low in all rater groups and even zero when the raters were R1 and R2 indicates that the scoring of students' answers to the questions did not change from rater to rater. Therefore, it can be concluded that the scores of ChatGPT are similar to the real raters. On the other hand, the item-rater interaction (*ixr*) values obtained can be considered as an indicator that ChatGPT as a rater differs from real raters, albeit with small differences. The fact that the highest value of this interaction, which shows the change in the scores given to the items from rater to rater, was obtained when the raters were ChatGPT and R1, shows that ChatGPT and R1 differed the most in terms of scoring the items. In addition, the fact that the item-rater interaction (*ixr*) took its lowest value when the raters were R1 and R2 can be interpreted that the raters who were the most similar in terms of scoring the items were the real raters. The findings regarding the variance (residual, error variance) related to the student-item-rater common effect (*sxixr,e*) show that the error variance is higher when one of the raters is ChatGPT. This may be due to a non-systematic change between student-item-rater and/or the interference of unknown factors that do not systematically affect scoring (Güler & Teker, 2015). In contrast to this finding, the lowest value of error variance was calculated when the raters were R1 and R2. This indicates that it is more appropriate for real raters to score open-ended questions compared to ChatGPT. When the calculated G coefficients are examined, it can be said that the relative evaluations (e.g. achievement ranking) to be made in line with the scores made will have moderate reliability or generalizability.

Based on all these findings, it can be concluded that the use of more than one method in estimating inter-rater reliability can provide more information to the reader about the real situation. The results of the current study reveal that there is a similarity and generally high reliability between real raters and between ChatGPT and real raters. Accordingly, it can be stated that artificial intelligence-based tools can play a potential role in educational assessments. Mizumoto and Eguchi (2023) used ChatGPT for automatic text evaluation and found that ChatGPT has a certain level of accuracy and reliability, can provide significant support for real raters, and linguistic features can improve the accuracy of scoring. The slightly higher reliability among real raters may be an indication that real raters are better able to detect some subtle details or subtexts in student responses.

In practice, educators may be advised to get support from artificial intelligence-based tools such as ChatGPT when scoring open-ended items that take a long time to score, especially in crowded classes or when time is limited. In this way, faster feedback can be provided and the teaching process can be made more effective. In addition, it may be recommended to conduct further research on how ChatGPT or similar artificial intelligence-based tools evaluate more complex or abstract responses. In addition, larger-scale studies can be conducted to determine how such tools affect the learning process and teacher feedback.

Ethics Committee Approval: This study was conducted by the approval of Ethics Committee in Tokat

Gaziosmanpaşa University dated 13.06.2023 and numbered 10.15.

Author Contributions: The whole process was carried out by the author.

Conflict of Interest: The author declares that they have no conflict of interest.

Uluslararası Türk Eğitim Bilimleri Dergisi,
2023, 11 (21), 1072-1099
https://doi.org/10.46778/goputeb.1345752

ISSN: 2148-2314

Açık Uçlu Maddelerin Puanlanmasında ChatGPT ve Gerçek Puanlayıcıların Puanlayıcılar Arası Güvenirlik Bakımından İncelenmesi

Seda Demir[a]

[a] *Dr. Öğr. Üyesi, Tokat Gaziosmanpaşa Üniversitesi, Tokat, Türkiye, seddadmr@gmail.com*

ÖZET

Bu araştırmanın amacı, açık uçlu maddelere verilen yanıtlar için yapay zekâ tabanlı bir araç olan ChatGPT ve iki gerçek puanlayıcı tarafından puanlama anahtarlarına göre yapılan puanlamanın puanlayıcılar arası güvenirlik bakımından incelenmesidir. Araştırmanın çalışma grubunu, 2022-2023 eğitim öğretim yılında Eskişehir ilinde öğrenim gören 13-15 yaş grubundan 30 öğrenci oluşturmaktadır. Araştırmanın verileri, Uluslararası Öğrenci Değerlendirme Programı-PISA Okuma Becerileri alanında yayımlanmış örnek sorular arasından seçilen 16 açık uçlu madde yardımıyla yüz yüze toplanmıştır. Puanlayıcılar arası güvenirliği belirlemek amacıyla korelasyon, uyuşma yüzdesi ve Genellenebilirlik kuramından yararlanılmıştır. Korelasyon analizlerinde SPSS 25, uyuşma yüzdesinin analizlerinde Excel ve genellenebilirlik kuramı analizlerinde EduG 6.1 programları kullanılmıştır. Araştırma sonuçları, puanlayıcılar arasında pozitif yönlü ve yüksek düzeyde bir ilişki olduğunu, puanlayıcıların yüksek oranda uyuşma gösterdiğini ve Genellenebilirlik kuramı kullanılarak hesaplanan güvenirlik (G) katsayılarının, korelasyon değerleri ve uyuşma yüzdelerine kıyasla daha düşük olduğunu göstermiştir. Bunun yanı sıra cevabı doğrudan metnin içinde geçen ve kısa cevaplı olan maddelere verilen yanıtların puanlanmasında tüm puanlayıcıların birbirleriyle mükemmel pozitif korelasyon ve tam uyuşma gösterdiği belirlenmiştir. Ayrıca Genellenebilirlik kuramı sonuçlarına göre toplam varyansı ana etkiler arasından en çok maddelerin (*m*), etkileşim etkileri arasından ise en çok öğrenci-madde etkileşiminin (*öxm*) açıkladığı görülmüştür. Sonuçta, uygulamaya dönük olarak eğitimcilere, kalabalık sınıflarda veya zamanın kısıtlı olduğu durumlarda özellikle puanlaması uzun zaman alan açık uçlu maddeler puanlanırken ChatGPT gibi yapay zekâ tabanlı araçlardan destek almaları önerilebilir.

MAKALE BİLGİSİ

Makale Türü
Araştırma

Makale Geçmişi
Gönderim tarihi:
18.08.2023
Kabul tarihi:
22.09.2023

Anahtar Kelimeler
Puanlayıcılar Arası Güvenirlik,
Açık Uçlu Maddelerin Puanlanması,
Genellenebilirlik Kuramı,
ChatGPT

Atıf Bilgisi: Demir, S. (2023). Açık uçlu maddelerin puanlanmasında ChatGPT ve gerçek puanlayıcıların puanlayıcılar arası güvenirlik bakımından incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi, 11* (21), 1072-1099.

Sorumlu yazar: Seda Demir, e-posta: seddadmr@gmail.com

Giriş

Ölçme ve değerlendirme süreçlerinin temel amacı, bireylerin bilgi, beceri ve yeteneklerini doğru bir şekilde tespit etmektir. Bu nedenle güvenirlik, bu sürecin temel taşlarından biri olarak kabul edilmektedir (Doğan, 2021) ve genel olarak, ölçme sonuçlarının tesadüfî hatalardan arınıklık derecesi olarak tanımlanmaktadır (Turgut, 1993). Güvenirliğin düşmesine neden olan hatalar, güvenirlik tanımına (kararlılık, tutarlılık, duyarlılık) ya da hata kaynağına (birey, madde, puanlayıcı, zaman vb.) göre farklılık gösterebilir (Crocker ve Algina, 1986; Lord ve Novick, 1968). Önemli bir hata kaynağı olduğu belirtilen puanlayıcıların, çok sayıda çalışmada puanlayıcılar arası güvenirlik kapsamında araştırıldığı görülmektedir (Atılgan, 2005; Bilgen ve Doğan, 2017; Güler ve Teker, 2015; Hallgren, 2012; Kan, 2005; Lilford ve diğerleri, 2007; Mancar, 2019; Park ve Kim, 2015; Pekin ve diğerleri, 2018). Puanlayıcı güvenirliği, birden fazla puanlayıcının yaptığı puanlamalar arasındaki tutarlılığın derecesi olarak tanımlanmaktadır (Aiken, 2000). Dolayısıyla puanlayıcıların puanlamaları arasındaki farklılık puanlayıcı kaynaklı bir hatadır ve puanlayıcı güvenirliğinin düşmesine neden olmaktadır. Ölçme aracının açık uçlu maddelerden oluştuğu durumlarda, puanlamanın objektiflikten uzaklaşması beklenen bir durumdur. Bu nedenle ölçme sonuçlarının gerçeği ne kadar yansıttığını ve bu ölçme sonuçlarına dayalı olarak verilen kararların ne kadar doğru olduğunu değerlendirmek için farklı puanlayıcıların puanladığı maddeler üzerinden madde ve test puanlarının tutarlılığı incelenir (Atılgan ve diğerleri, 2011). Puanlayıcılar arası güvenirliğin incelenmesinde kullanılan çok sayıda yöntem (Cohen's kappa, ağırlıklandırılmış kappa, Krippendorf alfa, Kendall uyuşma katsayısı, çok yüzeyli Rasch ölçme modeli vb.) bulunmakla birlikte mevcut araştırmada Pearson korelasyon katsayısı, uyuşma yüzdesi ve Genellenebilirlik kuramı olmak üzere üç farklı yöntem yardımıyla puanlayıcılar arası güvenirlikler incelenmiştir.

Alanyazın incelendiğinde puanlayıcılar arası güvenirliği belirlemede uyuşma yüzdesinin (Güler ve Teker, 2015; Gümüş ve Arıkan, 2020; Hallgren, 2012; İlhan, 2016; Mancar, 2019) ve korelasyonun (Goodwin, 2001; Goodwin ve diğerleri, 1991; Goodwin ve Goodwin, 1991; Kan, 2005) sıklıkla kullanıldığı görülmektedir. Bunun yanı sıra Genellenebilirlik kuramı da çok sayıda araştırmada (Atılgan, 2005; Çakıcı Eser ve Gelbal, 2012; Gage ve diğerleri, 2014; Hill ve diğerleri, 2012; Pekin ve diğerleri, 2018) kullanılmıştır.

Pearson korelasyon katsayısı, iki puanlayıcının puanlarının doğrusal ilişkisini başka bir ifadeyle, yaptıkları puanlamanın birbirleriyle tutarlılığını gösteren ve puanlayıcılar arası güvenirliğin hesaplanmasında sıklıkla kullanılan bir yöntemdir (Baykul, 2000). Uyuşma yüzdesi ise puanlayıcıların aynı maddeye aynı puanı vererek uyuşma sağladıkları madde sayısının basit yüzdesi olarak tanımlanabilir (Meyer, 1999). Bunların yanı sıra Genellenebilirlik kuramında, hata kaynağı olarak birey, madde ve puanlayıcı ana etkilerinin yanı sıra bunların etkileşim etkisi de dikkate alınmaktadır. Dolayısıyla hem puanlayıcılardan gelen potansiyel hatalar hem de diğer ana etkiler ve etkileşim etkileri Genellenebilirlik kuramıyla incelenebilir. Bu, Genellenebilirlik kuramının sağladığı büyük bir avantaj olarak görülmektedir (Brennan, 2001).

Alanyazında yer alan çalışmalardan farklı olarak mevcut araştırmada, yapay zekâ tabanlı bir araç olan ChatGPT, puanlayıcı olarak kullanılmıştır. ChatGPT, "chatbot" olarak bilinen sistemler grubuna ait en son gelişmelerdendir. Chatbot'lar, kural tabanlı veya kendi kendine öğrenme (yapay zekâ) yöntemleri kullanılarak geliştirilen akıllı sistemler olarak tanımlanmaktadır (OpenAI, 2015). ChatGPT'nin, 2022 yılının sonlarında insanların kullanımına sunulmasının ardından eğitimden sağlığa kadar pek çok alanda faydalı kullanımı üzerine çalışmalar yapıldığı görülmektedir.

Alanyazın incelendiğinde ChatGPT'nin eğitim öğretim sürecinde kullanımına ilişkin çalışmaların (Aktay ve diğerleri, 2023; Broutin, 2023; Göktaş, 2023; Grassini, 2023; Lo, 2023; Opera ve diğerleri, 2023; Zileli, 2023) sayısının gün geçtikçe arttığı görülmektedir. Bu çalışmalarda genel olarak ChatGPT'nin eğitimdeki rolü ve kullanımı, uzaktan eğitim sınavlarında kullanımı ve dil eğitiminde kullanımı araştırılmıştır. Bunların yanı sıra Mizumoto ve Eguchi (2023) tarafından yapılan araştırmada ChatGPT, otomatik metin değerlendirmesi yapmak ve yapılan değerlendirmenin güvenirliğini ve doğruluğunu değerlendirmek için kullanılmıştır. Araştırmanın sonucunda, ChatGPT'nin gerçek değerlendiricilere önemli ölçüde destek sağlayabileceğine vurgu yapılmıştır. Mevcut araştırmada ise bu araştırmalardan farklı olarak ChatGPT ile açık uçlu maddeler puanlanmış ve bu puanlamalar ile iki Edebiyat öğretmeninin yaptığı puanlamalar üzerinden puanlayıcılar arası güvenirlikler incelenmiştir. Araştırmanın bu yönüyle alanyazına katkı sağlayacağı söylenebilir. Bunun yanı sıra mevcut araştırmanın hem teknolojinin eğitsel değerlendirme sürecindeki potansiyeline hem de yapay zekâ tabanlı bir araç olan ChatGPT'nin kullanımına yönelik eğitimcilere fikir verebilmesi bakımından önemli olduğu düşünülmektedir.

Mevcut araştırmanın amacı, açık uçlu maddelere verilen yanıtlar için yapay zekâ tabanlı bir araç olan ChatGPT ve iki gerçek puanlayıcı tarafından puanlama anahtarlarına göre yapılan puanlamanın puanlayıcılar arası güvenirlik bakımından incelenmesidir. Araştırmanın bu temel amacı doğrultusunda üç alt problem belirlenmiştir.

Açık uçlu maddelerin puanlanmasında, ChatGPT ve Edebiyat öğretmeni olan iki gerçek puanlayıcının verdiği puanlar için,

1. Puanlayıcılar arasındaki korelasyona göre puanlayıcılar arası güvenirlik nasıldır?
2. Puanlayıcılar arasındaki uyuşma yüzdesine göre puanlayıcılar arası güvenirlik nasıldır?
3. Genellenebilirlik kuramına göre puanlayıcılar arası güvenirlik nasıldır?

Yöntem

Araştırmanın Modeli

Bu araştırma, iki gerçek puanlayıcı ve bir yapay zekâ tabanlı aracın puanladığı açık uçlu maddeler için çeşitli yöntemlerle hesaplanan puanlayıcılar arası güvenirlik katsayılarının incelenmesine yönelik olduğundan betimsel araştırma özelliği taşımaktadır. Betimsel araştırmalarda ele alınan durumun olabildiğince tam, ayrıntılı ve dikkatli bir şekilde açıklanması amaçlanmaktadır (Büyüköztürk ve diğerleri, 2011).

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2022-2023 eğitim öğretim yılında Eskişehir ilinde bir devlet okulunda öğrenim gören 13-15 yaş grubundaki öğrenciler arasından tamamen gönüllülük esasına göre seçilen 30 öğrenci oluşturmaktadır. 30 öğrencinin, 16 açık uçlu maddeye verdikleri yanıtlar üç puanlayıcı tarafından puanlanmıştır. Puanlayıcı grubunu ise yapay zekâ tabanlı bir araç olan ChatGPT'nin GPT-4.0 lisanslı versiyonu ve farklı kurumlarda 8 ve 10 yıldır görev yapmakta olan gönüllü iki Edebiyat öğretmeni oluşturmaktadır.

Verilerin Toplanması

Bu çalışma, Tokat Gaziosmanpaşa Üniversitesi Sosyal ve Beşerî Bilimler Araştırmaları Etik Kurulu'nun 13.06.2023 tarih ve 10.15 sayılı onayı ile gerçekleştirilmiştir. Araştırmanın verileri, Uluslararası Öğrenci Değerlendirme Programı-PISA Okuma Becerileri alanında yayımlanmış örnek sorular arasından seçilen ve Ek-1'de sunulan 16 açık uçlu madde yardımıyla yüz yüze toplanmıştır. Maddelerden toplamda alınabilecek en yüksek puan 100 olarak belirlenmiştir. Oluşturulan soru formunda yer alan açık uçlu maddelere öğrencilerin verdikleri yanıtlar, yapay zekâ tabanlı bir araç olan ChatGPT ve Puanlayıcı 1 (P1), Puanlayıcı 2 (P2) olarak iki gönüllü Edebiyat öğretmeni tarafından puanlanmıştır. Puanlamada, seçilen PISA soruları için yayımlanmış olan puanlama kriterleri doğrultusunda araştırmacı tarafından oluşturulan bir bütüncül dereceli puanlama anahtarı kullanılmıştır. Böylece araştırmada kullanılan veri seti, puanlayıcıların birbirinden bağımsız şekilde her bir öğrencinin 16 maddeye verdiği yanıtları puanlamasıyla elde edilmiştir. Puanlayıcıların her bir soruya verdikleri toplam puanlar üzerinden hesaplanan betimsel istatistikler Tablo 1'de sunulmuştur.

Tablo 1

ChatGPT, P1 ve P2 Tarafından Yapılan Puanlamaların Betimsel İstatistikleri

Puanlayıcı	N	Minimum	Maksimum	Ortalama	Standart Sapma	Çarpıklık	Basıklık
ChatGPT	16	5	225	98.88	55.20	.429	.485
P1	16	5	150	82.19	41.41	-.115	-.581
P2	16	5	150	86.31	44.29	-.202	-.921

Tablo 1'e göre her bir maddeye verilen toplam puanlar bakımından ChatGPT, diğer iki puanlayıcıya kıyasla daha yüksek bir ortalamaya sahiptir ($\bar{X}$= 98.88; SS= 55.20). Çarpıklık ve basıklık değerlerinin ise [-1.5, +1.5] aralığında olduğu görülmektedir. Buna göre yapılan puanlamaların normal dağılım gösterdiği (Tabachnick ve Fidell, 2014) söylenebilir.

Verilerin Analizi

Mevcut araştırmada, her bir öğrencinin 16 maddeye verdiği yanıtları puanlayan puanlayıcılar arasındaki puanlama güvenirliğini belirlemek amacıyla korelasyon, uyuşma yüzdesi ve Genellenebilirlik kuramından yararlanılmıştır. Bunlardan korelasyon analizlerinde SPSS 25, uyuşma yüzdesi analizlerinde Excel ve Genellenebilirlik kuramı analizlerinde EduG 6.1 programları kullanılmıştır.

Pearson korelasyon katsayısının, verilerin en az eşit aralık ölçek düzeyinde olması ve normal dağılım göstermesi gibi varsayımları bulunmaktadır. Ayrıca .30'dan küçük olan korelasyon değerleri iki değişken arasındaki düşük ilişkiyi, .30 ile .70 arasındaki değerler orta ve .70'ten büyük olan değerler ise yüksek ilişkiyi göstermektedir (Büyüköztürk ve diğerleri, 2011). Araştırma verileri sürekli olduğu için Pearson korelasyon katsayısı tercih edilmiş ve puanlayıcılar arası korelasyonlar ikili kombinasyonlar halinde hesaplanmıştır. Ancak bu yöntem puanlayıcıların uyuşma yüzdeleri ve puanlayıcılar arasındaki varyansla ilgili bilgi içermemektedir (Güler ve Teker, 2015; Şencan, 2005).

Uyuşma yüzdesi, ikili kombinasyonlar halinde puanlayıcıların aynı maddeye aynı puanı verme yüzdeleri olarak hesaplanmıştır. Tüm ölçek düzeyindeki veriler için kullanılabilen hesaplama ve

yorumlama kolaylığı sunan bir yöntemdir. Ancak puanlayıcıların yaptıkları puanlamalar arasındaki şansa bağlı veya tesadüfi uyuşmaları göz ardı ediyor olması yöntemin önemli bir sınırlılığıdır (Goodwin, 2001). Puanlayıcılar arası güvenirliğin var olduğundan bahsedebilmek için uyuşma yüzdesinin %75'in üzerinde olması gerekmektedir. (Şencan, 2005).

Puanlayıcılar arası güvenirliğin belirlenmesinde kullanılan bir diğer yöntem olan Genellenebilirlik kuramı analizlerinde ise çalışmanın ölçme objesi öğrenciler (*ö*) olarak yüzeyler ise maddeler (*m*) ve puanlayıcılar (*p*) olarak belirlenmiştir. Analizler tümüyle çaprazlanmış tesadüfi desen (*öxmxp*) üzerinden yürütülmüştür. Karşılaştırma yapabilmek amacıyla analizler, puanlayıcıların ikili kombinasyonları ve üç puanlayıcının tamamı üzerinden tekrarlanmış ve karşılaştırmalar yapılmıştır.

Bulgular

Mevcut araştırmada yapay zekâ tabanlı bir araç olan ChatGPT ve iki Edebiyat öğretmenin 16 açık uçlu soruya ilişkin öğrenci yanıtlarına verdikleri puanlar üzerinden hesaplanan Pearson korelasyon katsayısı değerleri Tablo 2'de verilmiştir.

Tablo 2

Korelasyon Katsayısı Aracılığıyla Puanlayıcılar Arası Güvenirlikler

Madde	ChatGPT ile P1 arası korelasyon	ChatGPT ile P2 arası korelasyon	P1 ile P2 arası korelasyon
1	.47*	.51*	.93*
2	.70*	.74*	.90*
3	1.00*	1.00*	1.000*
4	.62*	.62*	1.000*
5	[a]	[a]	[a]
6	.66*	.66	1.00*
7	1.00*	1.00*	1.00*
8	1.00*	1.00*	1.00*
9	.84*	.84*	1.00*
10	.72*	.74*	.92*
11	.84*	.84*	1.000*
12	.94*	.94*	1.000*
13	1.00*	1.00*	1.000*
14	.94*	.94*	1.000*
15	.81*	.94*	.76*
16	.87*	.94*	.93*
Test	.82*	.86*	.94*

Not. [a] Değişkenlerden en az biri sabit olduğu için hesaplanamaz.

*p< .001

Tablo 2 incelendiğinde, ChatGPT ile P1, ChatGPT ile P2 ve P1 ile P2 arasında 3. madde, 7. madde, 8. madde ve 13. maddede mükemmel pozitif korelasyon (1.00) olduğu görülmektedir. Ayrıca sırasıyla ChatGPT ile P1 ve ChatGPT ile P2 arasındaki en düşük korelasyonlar .47 ve .51 olarak 1. madde için hesaplanmıştır. P1 ile P2 arasındaki en düşük korelasyon 15. maddede (.76) gözlenmektedir.

Dolayısıyla puanlayıcılar arası uyuma yönelik olarak hesaplanan korelasyon değerlerinin maddeden maddeye değiştiği söylenebilir. Hesaplanan korelasyon değerlerine göre P1 ile P2'nin yaptıkları puanlamalar bakımından tüm maddelerde yüksek korelasyon gösterdikleri söylenebilir. ChatGPT ile P1 ve ChatGPT ile P2 arasındaki korelasyon değerlerine göre ise sadece 1. madde, 4. madde ve 6. maddede orta düzeyde bir ilişkinin olduğu, diğer maddelerdeki ilişkinin ise oldukça yüksek olduğu görülmektedir. Bunun yanı sıra 5. maddeye tüm puanlayıcılar aynı puanı verdiği için puanlayıcılar arası korelasyonlar hesaplanamamıştır. Ayrıca puanlayıcıların maddelere verdiği puanların ortalamaları arasında da (.82, .86 ve .94) yüksek ilişkiler olduğu tespit edilmiştir.

Puanlayıcılar arası uyuşma yüzdeleri Tablo 3'te sunulmuştur.

Tablo 3

Uyuşma Yüzdesi Aracılığıyla Puanlayıcılar Arası Güvenirlikler

Madde	ChatGPT ile P1 arası uyuşma yüzdesi	ChatGPT ile P2 arası uyuşma yüzdesi	P1 ile P2 arası uyuşma yüzdesi
1	50.00	53.33	90.00
2	36.67	26.67	90.00
3	100.00	100.00	100.00
4	80.00	80.00	100.00
5	100.00	100.00	100.00
6	80.00	80.00	100.00
7	100.00	100.00	100.00
8	100.00	100.00	100.00
9	63.33	63.33	100.00
10	73.33	76.67	96.67
11	93.33	93.33	100.00
12	96.67	96.67	100.00
13	100.00	100.00	100.00
14	96.67	96.67	100.00
15	53.33	93.33	53.33
16	93.33	96.67	96.67
Ortalama	82.29	84.79	95.42
Standart Sapma	21.04	21.08	11.73

Tablo 3'te yer alan ikinci, üçüncü ve dördüncü sütunlarda, her iki puanlayıcının da aynı puanı verdiği maddelerin yüzdeleri görülmektedir. Buna göre Tablo 2'de sunulan verilere paralel olarak 3. madde, 5. madde, 7. madde, 8. madde ve 13. maddede ChatGPT ile P1, ChatGPT ile P2 ve P1 ile P2 aynı maddelere aynı puanı vermiştir. Araştırmadan elde edilen dikkat çekici bir bulgu olarak, puanlayıcılardan birinin ChatGPT olduğu durumlarda 2. madde için P1 ile uyuşma yüzdesinin yaklaşık %37 olduğu, P2 ile uyuşma yüzdesi ise yaklaşık %27 olduğu görülmektedir. Aynı soruda P1 ile P2'nin uyuşması ise %90 olarak hesaplanmıştır. Buna göre puanlayıcılar arası uyuşma yüzdelerinin maddeden maddeye değiştiği söylenebilir. Ayrıca, uyuşma yüzdesi ortalaması ChatGPT ile P1, ChatGPT ile P2 ve P1 ile P2 için sırasıyla 82.29 (SS= 21.04), 84.79 (SS= 21.08) ve 95.42 (SS= 11.73) olarak hesaplanmıştır.

Puanlayıcılar arası güvenirliğin belirlenmesinde Genellenebilirlik kuramından da yararlanılmıştır. Diğer yaklaşımlardaki hata kaynağı sadece puanlayıcıların puanları arasındaki farklılaşma iken Genellenebilirlik kuramında farklı hata kaynakları aynı anda ele alınmaktadır. Bu hata kaynakları mevcut araştırma için madde (*m*), puanlayıcı (*p*), öğrenci-madde etkileşimi (*öxm*), öğrenci-puanlayıcı

etkileşimi (*öxp*), madde-puanlayıcı etkileşimi (*mxp*) ve öğrenci-madde-puanlayıcı etkileşimi (*öxmxp*) olarak sıralanabilir. Puanlayıcıların ikili kombinasyonları ve üç puanlayıcı için yapılan analizlerden elde edilen varyans analizi sonuçları ve kestirilen varyans bileşenleri Tablo 4'te sunulmuştur.

Tablo 4

Varyans Analizi Sonuçları ve Kestirilen Varyans Bileşenleri

Puanlayıcılar	Varyans Kaynağı	KT	sd	KO	Kestirilen Varyans Bileşeni	Varyans Yüzdesi (%)	G Katsayısı
ChatGPT ve P1	öğrenci (*ö*)	847.11	29	29.21	0.55	5.9	.61
	madde (*m*)	2124.38	15	141.63	1.92	20.5	
	puanlayıcı (*p*)	74.26	1	74.26	0.12	1.3	
	öxm	4735.71	435	10.89	4.69	50.3	
	öxp	60.90	29	2.10	0.04	0.4	
	mxp	256.36	15	17.09	0.52	5.6	
	öxmxp, e	651.99	435	1.50	1.50	16.0	
ChatGPT ve P2	öğrenci (*ö*)	848.29	29	29.25	0.54	5.8	.59
	madde (*m*)	2296.81	15	153.12	2.16	23.1	
	puanlayıcı (*p*)	42.08	1	42.08	0.06	0.6	
	öxm	4780.22	435	10.99	4.86	51.9	
	öxp	64.88	29	2.24	0.06	0.6	
	mxp	207.50	15	13.83	0.42	4.5	
	öxmxp, e	554.03	435	1.27	1.27	13.6	
P1 ve P2	öğrenci (*ö*)	827.92	29	28.55	0.52	6.1	.58
	madde (*m*)	1779.90	15	118.66	1.72	20.1	
	puanlayıcı (*p*)	4.54	1	4.54	0.00	0.0	
	öxm	5212.91	435	11.98	5.78	67.6	
	öxp	7.78	29	0.27	-0.01	0.0	
	mxp	58.10	15	3.87	0.12	1.3	
	öxmxp, e	181.59	435	0.42	0.42	4.9	
ChatGPT, P1 ve P2	öğrenci (*ö*)	1239.40	29	42.74	0.54	5.9	.61
	madde (*m*)	3013.55	15	200.90	1.93	21.3	
	puanlayıcı (*p*)	80.59	2	40.29	0.06	0.6	
	öxm	7133.16	435	16.40	5.11	56.3	
	öxp	89.04	58	1.54	0.03	0.3	
	mxp	347.97	30	11.60	0.35	3.9	
	öxmxp, e	925.07	870	1.06	1.06	11.7	

Not. KT: kareler toplamı, KO: kareler ortalaması, sd: serbestlik derecesi

Tablo 4'e göre puanlayıcılar ChatGPT ve P1 olduğunda, ChatGPT ve P2 olduğunda, P1 ve P2 olduğunda veya ChatGPT, P1 ve P2 olduğunda öğrenci, madde ve puanlayıcı ana etkileri incelendiğinde maddeye (*m*) ilişkin değişkenliğin ana etkiler içinde en yüksek değere sahip olduğu ve toplam varyansın sırasıyla %20.5'ini, %23.1'ini, %20.1'ini ve %23.1'ini açıkladığı görülmektedir. Bu durum maddelerin güçlük düzeylerinin birbirinden farklılaştığını göstermektedir. Öğrencilere (*ö)* ilişkin varyanslar ise toplam varyansın sırasıyla %5.9'unu %5.8'ini, %6.1'ini %5.9'unu açıklamaktadır. Öğrenciler arasında farklılığın bulunduğunu gösteren bu değerin olabildiğince yüksek olması beklenmektedir (Brennan, 2001). Bunun yanı sıra puanlayıcı (*p*) ana etkisine ilişkin varyans ise puanlayıcılar ChatGPT ve P1 olduğunda toplam varyansın %1.3'ünü , ChatGPT ve P2 olduğunda %0.6'sını, ChatGPT, P1 ve P2 olduğunda yine %0.6'sını açıklamaktadır. Bu değer

puanlayıcılar arası değişkenliği ifade ettiğinden mümkün olduğu kadar sıfıra yakın çıkması istenmektedir (Brennan, 2001). Bu bakımdan puanlayıcılar P1 ve P2 olduğunda puanlayıcı ana etkisine ilişkin varyansın %0.0 olması beklentiyi karşılayan, istenen bir durumdur.

Tablo 4, puanlayıcılar ChatGPT ve P1 olduğunda, ChatGPT ve P2 olduğunda, P1 ve P2 olduğunda veya ChatGPT, P1 ve P2 olduğunda öğrenci, madde ve puanlayıcı etkileşimleri bakımından incelendiğinde, öğrenci-madde etkileşiminin (*öxm*) etkileşim etkileri içinde en yüksek değere sahip olduğu görülmektedir. Buna göre öğrenci-madde etkileşimine (*öxm*) ilişkin varyans, toplam varyansın sırasıyla %50.3'ünü, %51.9'unu, %67.6'sını ve %56.3'ünü açıklamaktadır. Dolayısıyla maddelerin güçlük düzeyi öğrencilere göre önemli bir değişkenlik göstermektedir. Öğrenci-puanlayıcı etkileşimine (*öxp*) ilişkin varyansın ise tüm puanlayıcı gruplarında oldukça düşük (0.4, 0.3, 0.6) olduğu, puanlayıcılar P1 ve P2 olduğunda ise sıfır olduğu görülmektedir. Bu durum, öğrencilerin sorulara verdikleri yanıtlar için yapılan puanlamaların puanlayıcıdan puanlayıcıya değişmediğine işaret etmektedir. Bunun yanı sıra madde-puanlayıcı etkileşimine (*mxp*) ilişkin varyanslar toplam varyansın sırasıyla %5.6'lık, %4.5'lik, %1.3'lük ve %3.9'luk kısmına karşılık gelmektedir. Maddelere verilen puanların puanlayıcıdan puanlayıcıya değişimini gösteren bu etkileşimin en yüksek değerini puanlayıcılar ChatGPT ve P1 olduğunda aldığı, en düşük değerini ise puanlayıcılar P1 ve P2 olduğunda aldığı belirlenmiştir. Dolayısıyla yaptıkları puanlama bakımından birbirinden en çok farklılaşan puanlayıcıların ChatGPT ve P1 puanlayıcıları olduğu, en az farklılaşan puanlayıcıların ise P1 ve P2 puanlayıcıları olduğu söylenebilir. Bunların yanı sıra artık (residual) veya hata varyansı olarak da ifade edilen ve mümkün olduğunca sıfıra yakın çıkması istenen öğrenci-madde-puanlayıcı ortak etkisine (*öxmxp,e*) ilişkin varyanslar sırasıyla %16, %13,6, %4.9 ve %11.7 olarak hesaplanmıştır. Burada en yüksek değerin puanlayıcılar ChatGPT ve P1 olduğunda, en düşük değerin ise puanlayıcılar P1 ve P2 olduğunda hesaplandığı görülmektedir.

Tablo 4'ün son sütununda puanlayıcılar ChatGPT ve P1 iken, ChatGPT ve P2 iken, P1 ve P2 iken veya ChatGPT, P1 ve P2 iken hesaplanmış olan genellenebilirlik (G) katsayısı yer almaktadır. Puanların güvenirlik veya genellenebilirlik düzeyinin göstergesi olan bu katsayı 0.0 ile 1.0 arasında değerler almaktadır (Shavelson ve Webb, 1991). 30 öğrenci, 16 madde ve iki puanlayıcı için G katsayıları sırasıyla .61 (ChatGPT ve P1 puanlayıcıları), .59 (ChatGPT ve P2 puanlayıcıları) ve .58 (P1 ve P2 puanlayıcıları) olarak, üç puanlayıcı içinse .61 (ChatGPT, P1 ve P2 puanlayıcıları) gibi diğerlerine yakın bir değer olarak hesaplanmıştır.

Tartışma ve Sonuç

Mevcut araştırmada biri yapay zekâ tabanlı bir araç olan ChatGPT diğer ikisi Edebiyat öğretmeni olan üç puanlayıcının puanladıkları açık uçlu maddeler üzerinden puanlayıcılar arası güvenirlikler belirlenmiştir. Bu aşamada, puanlayıcılar ikili ve üçlü olarak gruplanmış (ChatGPT ve P1 olarak, ChatGPT ve P2 olarak, P1 ve P2 olarak, ChatGPT, P1 ve P2 olarak) ve puanlayıcılar arası güvenirlik; korelasyon, uyuşma yüzdesi ve Genellenebilirlik kuramı yardımıyla belirlenmiştir. Analizlerden elde edilen sonuçların, ulaşılan genel çıkarımlar noktasında birbirini destekler nitelikte olduğu görülmüştür.

Puanlayıcılar arasında hesaplanan korelasyon değerleri, puanlayıcılar arasında pozitif yönlü ve yüksek düzeyde (ChatGPT ve P1 için .82; ChatGPT ve P2 için .86; P1 ve P2 için .94) bir ilişki olduğunu göstermektedir. Bu bulguyu destekler nitelikte, alanyazındaki pek çok araştırmada da

puanlayıcılar arası güvenirliğe yönelik olarak yüksek korelasyonlar elde edildiği görülmektedir (Goodwin, 2001; Goodwin ve Goodwin, 1991; Goodwin ve diğerleri, 1991; Güler ve Teker, 2015; Öksüzoğlu, 2022; Özşavlı, 2023; Seheryeli, 2018; Wilson ve diğerleri, 2022). Bunun yanı sıra araştırmadan elde edilen önemli bir bulgu olarak, cevabı doğrudan metnin içinde geçen ve kısa cevaplı olan maddelere (3. madde, 7. madde, 8. madde ve 13. madde) verilen yanıtların puanlanmasında tüm puanlayıcılar birbirleriyle mükemmel pozitif korelasyon (1.00) göstermiştir. Ayrıca, puanlayıcılardan biri ChatGPT olduğunda (ChatGPT ve P1 veya ChatGPT ve P2), puanlayıcılar arası en düşük korelasyon 1. madde için hesaplanmıştır. Bu maddede, verilen metne yönelik olarak metinde net olarak belirtilen ana amaç dışındaki alt amaçlardan birinin yazılması istenmektedir. Dolayısıyla metinde ana amaç kadar net belirtilmeyen alt amaçların belirlenmesine yönelik yanıtların puanlanmasında gerçek puanlayıcıların (P1 ve P2) yaptığı puanlamaların daha uyumlu olduğu yorumu yapılabilir. Bunun yanı sıra araştırmanın dikkat çekici bir bulgusu olarak, gerçek puanlayıcılar (P1 ve P2) arasındaki en düşük korelasyonun gözlendiği 15. maddede, ChatGPT ile gerçek puanlayıcılar arasındaki korelasyon daha yüksek hesaplanmıştır. Bu madde incelendiğinde, verilen metinde, karakterin içsel monologlarını ve düşünce süreçlerini yansıtan, okuyucunun düşünmesini ve hayal gücünü kullanmasını gerektiren karmaşık bir dil kullanıldığı görülmüştür. Buradan yola çıkarak, yanıtı hayal gücü ve yaratıcılığa dayanan maddelerin puanlanmasında gerçek puanlayıcılar arasındaki farklılaşma artarken ChatGPT'nin yaptığı puanlamanın gerçek puanlayıcıların ortalamasına yakın bir puanlama yaptığı yorumu yapılabilir. Ayrıca sorulan sorunun yanıtının verilen metinde dikkat çekici, kısa ve net olarak belirtildiği görülen 15. maddeye tüm puanlayıcılar tek ve aynı puanı vermiştir. Dolayısıyla kısa cevaplı ve cevabı metinde net olarak vurgulanan açık uçlu maddelerde puanlayıcıların tam uyum gösterdiğini de söylemek mümkündür. Ancak şunu belirtmek gerekir ki, puanlayıcılar arasındaki güvenirliğin hesaplanmasında kullanılan korelasyon, ortalamadan bağımsız olarak hesaplandığından iki puanlayıcının puanlamaları arasındaki benzerlikleri veya katılık/cömertlik durumlarını göstermez. Bu nedenle puanlayıcılar arası güvenirliğin belirlenmesinde yetersiz kalabilir (Goodwin, 2001).

Araştırmadan elde edilen dikkat çekici bir bulgu olarak, gerçek puanlayıcıların yaptığı puanlamaların birbiriyle daha yüksek ilişki gösterdiği ve uyuşma yüzdelerinin de daha yüksek olduğu görülmüştür. Bunun yanı sıra ChatGPT ile gerçek puanlayıcılar arasındaki en düşük uyuşma yüzdesinin 2. madde için hesaplandığı görülmüştür. Ancak bu madde için gerçek puanlayıcılar arası uyuşma yüzdesi (%90) oldukça yüksektir. Verilen metne dayalı olan 2. madde incelendiğinde, maddenin iki aşamalı olduğu, yanıt olarak metinde yer alan bir bilginin yazılmasının ardından verilen yanıtın metindeki ifadelerle desteklenmesinin istendiği görülmüştür. Dolayısıyla iki aşamalı olarak nitelendirilebilecek açık uçlu maddelerin puanlamasında gerçek puanlayıcılar arasındaki uyumun daha yüksek olduğu, ChatGPT'nin yaptığı puanlamanın ise gerçek puanlayıcılardan farklılaştığı yorumu yapılabilir. Hesaplama ve yorumlama bakımından kolaylık sağlayan uyuşma yüzdesi yöntemin en büyük sınırlılığı, puanlayıcıların tesadüfen/şansla ortaya çıkan uyumunu dikkate almamasıdır (Güler ve Teker, 2015). Bu noktada güvenirlik katsayısının hesaplanmasında, puanlayıcılar arası şansa bağlı yapay uyumu da dikkate alan Cohen's Kappa istatistiğinin (Cohen, 1960) kullanılması önerilebilir.

Genellenebilirlik kuramı, farklı hata kaynaklarını aynı anda ele almakta ve ayrıntılı bilgiler sunmaktadır. Bu özelliği, Genellenebilirlik kuramını araştırmada kullanılan diğer yöntemlerden daha üstün kılmaktadır. Mevcut araştırmada Genellenebilirlik kuramı kullanılarak elde edilen G katsayılarının, korelasyon değerleri ve uyuşma yüzdelerine kıyasla daha düşük (ChatGPT ve P1 için

.61; ChatGPT ve P2 için .59; P1 ve P2 için .58; ChatGPT, P1 ve P2 için .61) olduğu görülmektedir. Ancak farklı puanlayıcı grupları arasında ciddi bir farklılığın gözlenmemesi, puanlayıcılar arası güvenirliği belirlemek için kullanılan diğer yöntemlerden elde edilen bulguları destekler niteliktedir. Genellenebilirlik kuramı kapsamında yapılan analizler incelendiğinde önemli bulgulara ulaşıldığı görülmektedir. Öğrenci (*ö*), madde (*m*) ve puanlayıcı (*p*) ana etkileri incelendiğinde, maddeye ilişkin değişkenliğin, ana etkiler içinde en yüksek değere sahip olduğu ve dolayısıyla da veri toplamada kullanılan maddelerin güçlük düzeylerinin birbirinden farklılaştığı belirlenmiştir. Bunun yanı sıra beklentiye uygun olarak, öğrencilere ilişkin ana etkinin öğrencilerin birbirinden az miktarda da olsa farklılaştığına, puanlayıcılara ilişkin ana etkinin ise puanlayıcılar arasında değişkenliğin bulunmadığına işaret ettiği söylenebilir. Öğrenci-madde etkileşimine (*öxm*) ilişkin varyansın tüm puanlayıcı gruplarında toplam varyansın büyük bir kısmını açıkladığı görülmüştür. Buna göre maddelerin güçlük düzeyi öğrencilere göre önemli bir değişkenlik göstermektedir. Güler ve Teker'e (2015) göre bu durum, matematik ve istatistik gibi öğrencilerin geçmiş öğrenme yaşantılarının etkili olduğu alanlarda daha olasıdır. Ancak araştırmanın dikkat çeken bir diğer bulgusu olarak, öğrencilerin okuma becerilerine yönelik olan sorularda da maddelerin güçlük düzeyinin öğrencilere göre önemli bir değişkenlik gösterdiği belirlenmiştir. Bu durum, PISA Okuma Becerileri alanındaki soruların akıcı okuma, okuduğunu anlama, değerlendirme ve derinlemesine düşünme gibi farklı bilişsel süreçlere yönelik olmasından kaynaklanmış olabilir. Öğrenci-puanlayıcı etkileşimine (*öxp*) ilişkin varyansın tüm puanlayıcı gruplarında oldukça düşük olması ve hatta, puanlayıcılar P1 ve P2 olduğunda sıfır olması öğrencilerin sorulara verdikleri yanıtlar için yapılan puanlamaların puanlayıcıdan puanlayıcıya değişmediğine işaret etmektedir. Dolayısıyla ChatGPT'nin de gerçek puanlayıcılara benzer özelliklerde puanlamalar yaptığı sonucuna varılabilir. Buna karşılık elde edilen madde-puanlayıcı etkileşimi (*mxp*) değerleri, ChatGPT'nin puanlayıcı olarak gerçek puanlayıcılardan küçük farklarla da olsa ayrıştığının bir göstergesi olarak değerlendirilebilir. Maddelere verilen puanların puanlayıcıdan puanlayıcıya değişimini gösteren bu etkileşimin en yüksek değerini puanlayıcılar ChatGPT ve P1 olduğunda almış olması, maddeleri puanlama bakımından en çok ChatGPT ve P1'in birbirinden farklılaştığını göstermektedir. Bunun yanı sıra madde-puanlayıcı etkileşiminin (*mxp*) en düşük değerini puanlayıcılar P1 ve P2 olduğunda almış olması ise maddeleri puanlama bakımından en çok benzeşen puanlayıcıların gerçek puanlayıcılar olduğu şeklinde yorumlanabilir. Öğrenci-madde-puanlayıcı ortak etkisine (*öxmxp,e*) ilişkin varyansa (artık, hata varyansı) yönelik bulgular, puanlayıcılardan birinin ChatGPT olması durumunda hata varyansının daha yüksek olduğunu göstermektedir. Bunun nedeni, öğrenci-madde-puanlayıcı arasında sistematik olmayan bir değişim ve/veya ölçmeye bilinmeyen ve puanlamayı sistematik olarak etkilemeyen faktörlerin karışması olabilir (Güler ve Teker, 2015). Elde edilen bu bulguya karşılık hata varyansının en düşük değeri puanlayıcılar P1 ve P2 olduğunda hesaplanmıştır. Bu durum açık uçlu soruların puanlanmasında, puanlamayı ChatGPT'ye kıyasla gerçek puanlayıcıların yapmasının daha uygun olduğuna işaret etmektedir. Hesaplanan G katsayıları incelendiğine ise yapılan puanlamalar doğrultusunda gerçekleştirilecek bağıl değerlendirmelerin (örneğin başarı sıralaması), orta düzeyde güvenirlik veya genellenebilirliğe sahip olacağı söylenebilir.

Araştırmadan elde edilen tüm bu bulgulardan yola çıkarak, puanlayıcılar arası güvenirlik kestirilirken birden fazla yöntemin kullanılmasının, gerçek durumla ilgili okuyucuya daha fazla bilgi verebileceği yorumu yapılabilir. Mevcut araştırmanın sonuçları, gerçek puanlayıcılar arasında ve ChatGPT ile gerçek puanlayıcılar arasında benzerlik ve genellikle yüksek bir güvenirlik olduğunu ortaya koymaktadır. Buna göre yapay zekâ tabanlı araçların eğitsel değerlendirmelerde

potansiyel bir rol oynayabileceği söylenebilir. Mizumoto ve Eguchi (2023) tarafından yapılan araştırmada otomatik metin değerlendirmesi için ChatGPT kullanılmış ve araştırma sonucunda, mevcut araştırmanın sonuçlarını destekler nitelikte, ChatGPT'nin belirli bir düzeyde doğruluk ve güvenirliğe sahip olduğu, gerçek değerlendiriciler için önemli bir destek sağlayabileceği ve dilbilimsel özelliklerin puanlamanın doğruluğunu artırabileceği ifade edilmiştir. Gerçek puanlayıcılar arasındaki güvenirliğin biraz daha yüksek olması, gerçek puanlayıcıların öğrenci yanıtlarına ilişkin bazı ince ayrıntıları veya alt metinleri daha iyi tespit edebildiklerinin bir göstergesi olabilir.

Uygulamaya dönük olarak eğitimcilere, kalabalık sınıflarda veya zamanın kısıtlı olduğu durumlarda özellikle puanlaması uzun zaman alan açık uçlu maddeler puanlanırken ChatGPT gibi yapay zekâ tabanlı araçlardan destek almaları önerilebilir. Bu sayede daha hızlı geri bildirimler sağlanıp öğretim süreci daha etkili hale getirilebilir. Bunun yanı sıra ChatGPT'nin veya benzeri yapay zekâ tabanlı araçların daha karmaşık veya soyut ifadeler içeren yanıtları nasıl değerlendirdiğine yönelik yeni araştırmaların yapılması önerilebilir. Ayrıca, bu tür araçların öğrenme sürecine ve öğretmen geri bildirimine nasıl bir etkisi olduğunu belirlemeye yönelik daha geniş ölçekli çalışmalar gerçekleştirilebilir.

Etik Kurul Onayı: Bu çalışma Tokat Gaziosmanpaşa Üniversitesi, Sosyal ve Beşeri Araştırmalar Etik Kurulu'ndan 13.06.2023 tarihinde 10.15 sayılı etik kurul izni alınarak gerçekleştirilmiştir.

Araştırmacıların Katkı Oranı: Araştırma sürecinin tüm aşamaları yazar tarafından gerçekleştirilmiştir.

Çatışma Beyanı: Yazar herhangi bir çıkar çatışması olmadığını beyan eder.

References

Aiken, L. R. (2000). *Psychological testing and assessment*. Allyn and Bacon.

Aktay, S., Seçkin, G. Ö. K., & Uzunoğlu, D. (2023). ChatGPT in education. *TAY Journal, 7*(2), 378-406. https://doi.org/10.29329/tayjournal.2023.543.03

Atılgan, H. (2005). Generalizability theory and a sample application for inter-rater reliability. *Educational Sciences and Practice, 4*(7), 95-108. http://www.ebuline.com/pdfs/7Sayi/7_6.pdf

Atılgan, H., Kan, A., & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. (5th ed.) Anı Yayıncılık.

Baykul, Y. (2000) *Eğitimde ve psikolojide ölçme: Klasik Test Teorisi ve uygulaması [Measurement in education and psychology: Classical Test Theory and its application]*. ÖSYM Yayınları.

Bilgen, Ö. B., & Doğan, N. (2017). The comparison of interrater reliability estimating techniques. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 63-78. https://doi.org/10.21031/epod.294847

Brennan, R. L. (2001). *Generalizability Theory*. Springer-Verlag.

Büyüköztürk, Ş., Çakmak, E. Kılıç, A., Özcan, E., Karadeniz, Ş., & Demirel, F. (2011). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Pegem Akademi.

Doğan, N. (Ed.). (2021). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Pegem Akademi.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winson.

Çakıcı Eser, D., & Gelbal, S. (2012). Comparison of interrater agreement Calculated with generalizability theory and logistic regression. *Kastamonu Education Journal, 21*(2), 423-438. https://acikerisim.kku.edu.tr/xmlui/handle/20.500.12587/1380

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches in emotional and/or behavioral disorders research using generalizability theory. *Behavioral Disorders, 39*(4), 228-244. https://doi.org/10.1177/019874291303900407

Goodwin, L. D., & Goodwin, W. L. (1991). Using generalizability theory in early childhood special education. *Journal of Early Intervention, 15*(2), 193-204. https://doi.org/10.1177/105381519101500208

Goodwin, L. D., Sands, D. J., & Kozleski, E. B. (1991). Estimating interinterviewer reliability for interview schedules used in special education research. *The Journal of Special Education, 25*(1), 73-89. https://doi.org/10.1177/002246699102500105

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2

Göktaş, L. S. (2023). Can ChatGPT succeed in distance education exams? A research on accuracy and verification in tourism. *Journal of Tourism & Gastronomy Studies, 11*(2), 892-905. https://doi.org/10.21325/jotags.2023.1224

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Güler, N., & Teker, G. T. (2015). The evaluation of rater reliability of open ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 12-24. https://doi.org/10.21031/epod.63041

Gümüş, F. Ö., & Arıkan, Ç. A. (2020). Investigation of solutions of mathematical problems using multiple representations in terms of inter-rater reliability. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 14*(1), 606-628. https://doi.org/10.17522/balikesirnef.687639

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64. https://doi.org/10.3102/0013189X12437203

İlhan, M. (2016). A comparison of the ability estimations of classical test theory and the many facet Rasch model in measurements with open-ended questions. *Hacettepe University Journal of Education, 31*(2), 346-368. https://doi.org/10.16986/HUJE.2016015182

Kan, A. (2005). *The effect of using grading scale and answer key to grader's reliability*. Eurasian Journal of Educational Research, 20. 166-177. https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=50df9fc0-9dbc-43f8-a338-7a1110d5ce44%40redis

Lilford, R., Edwards, A., Girling, A., Hofer, T., Di Tanna, G. L., Petty, J., & Nicholl, J. (2007). Inter-rater reliability of case-note audit: A systematic review. *Journal of Health Services Research & Policy, 12*(3), 173-180. https://doi.org/10.1258/135581907781543012

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Lord, F. M., & Novick, M. R. (1968) *Statistical theory of mental test scores*. Addison-Wesley.

Mancar, S. A. (2019). *The comparison of inter rater reliability estimating Techniques in performance based assessment*. [Unpublished Master Thesis]. Ankara University.

Meyer, G. J. (1999). Simple procedures to estimate chance agreement and kappa for the interrater reliability of response segments using the Rorschach Comprehensive System. *Journal of Personality Assessment, 72*(2), 230-255. https://doi.org/10.1207/S15327752JP720209

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. https://doi.org/10.1016/j.rmal.2023.100050

OpenAI. (2015). OpenAI. https://openai.com/about

Opara, E., Mfon-Ette Theresa, A., & Aduke, T. C. (2023). ChatGPT for teaching, learning and research: Prospects and challenges. *Global Academic Journal of Humanities and Social Sciences, 5*(2), 33-40. https://doi.org/10.36348/gajhss.2023.v05i02.001

Öksüzoğlu, M. (2022). *The investigation of items measuring high-level thinking skills in terms of student score and score reliability*. [Unpublished Doctoral Dissertation]. Hacettepe University.

Özşavlı, M. (2023). The effect of peer feedback on the writing skills of students learning Turkish as a foreign language. *International Journal of Turkish Literature Culture Education, 12*(1), 253-273. https://doi.org/10.7884/teke.5638

Park, C. U., & Kim, H. J. (2015). Measurement of inter-rater reliability in systematic review. *Hanyang Medical Reviews, 35*(1), 44-49. https://doi.org/10.7599/hmr.2015.35.1.44

Pekin, Z., Çetin, S., & Güler, N. (2018). Comparison of Interrater Reliability Based on Different Theories for Autism Social Skills Profile. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 202-215. https://doi.org/10.21031/epod.388590

Seheryeli, M. Y. (2018). *An examination of the reliability estimates of a scoring rubric of a writing skill examination using the classical test theory, generalizability theory and the item response theory models*. [Unpublished Master Thesis]. Gazi University.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Sage Publications.

Şencan, H. (2005). *Sosyal ve davranışsal ölçmelerde güvenirlik ve geçerlik [Reliability and validity in social and behavioural measurements]*. Sözkesen Matbaacılık.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using Multivariate Statistics*. Pearson.

Tapan Broutin, M. S. (2023). Examination of questions asked by pre-service mathematics teachers in their initial experiences with ChatGPT. *Journal of Uludag University Faculty of Education, 36*(2), 1-26. https://doi.org/10.19171/uefad.1299680

Turgut, M. F. (1993). *Eğitimde ölçme ve değerlendirme metotları*. Saydam Matbaacılık.

Wilson, M. H., Ashworth, E., Hutchinson, P. J., & British Neurotrauma Group. (2022). A proposed novel traumatic brain injury classification system–an overview and inter-rater reliability validation on behalf of the Society of British Neurological Surgeons. *British Journal of Neurosurgery, 36*(5), 633-638. https://doi.org/10.1080/02688697.2022.2090509

Zileli, E. N. (2023). ChatGPT example in learning Turkish as a foreign language. *International Journal of Karamanoğlu Mehmetbey Educatioanal Research,* *5*(1), 42-51. https://doi.org/10.47770/ukmead.1296013

Appendix-1: OPEN-ENDED QUESTION FORM

ATİNA'DA DEMOKRASİ

A BÖLÜMÜ

Tukidids, Klasik Yunan döneminde milattan önce beşinci yüzyılda yaşamış bir tarihçi ve aynı zamanda bir askerdir. Atina'da doğmuştur. Tukidids, Atina ve Sparta arasındaki Peloponnes Savaşı boyunca (M.Ö. 431 – 404) Trakya'daki Amfipolis şehrini korumakla görevlendirilen donanmanın komutanıydı. Ancak şehre zamanında ulaşmayı başaramadı. Şehir, Spartalı general Brasidas'ın eline geçti ve bundan dolayı Tukidids yirmi yıllık bir sürgüne gitmek zorunda kaldı. Bu sürgün, Tukidids'e savaşan iki taraf hakkında ayrıntılı bilgi toplama ve *Peloponnes Savaşı'nın Tarihi* adlı çalışması için araştırma yapma imkânı verdi.

Tukidids, tarih öncesi zamanların en önemli tarihçileri arasında kabul edilir. O, tarihinin gelişimini açıklarken kader veya ilahi güçlerin müdahaleleri yerine, bireysel tutum ve davranışlar ve doğal sebeplere odaklanır. Onun çalışmalarında gerçekler, sadece kuru hikâyeler olarak sunulmazlar. Gerçekler, ana karakterlerin o veya bu şekilde davranmalarına neden olan sebepleri bulmak adına araştırılırlar. Tukidids'in bazen hayali konuşmalara yer vermesinin nedeni, bireylerin davranışlarını vurgulamaktır. Bu hayali konuşmalar, Tukidids'in, tarihi karakterlerin davranışlarının arkasındaki nedenleri açıklamasına yardımcı olur.

B BÖLÜMÜ

Tukidids, aşağıdaki konuşmayı Peloponnes Savaşının ilk yılında ölen askerlerin onuruna Atinalı hükûmdar Perikles'e (M.Ö. 5. yüzyıl) atfeder.

Bizim devlet yönetimi sistemimiz komşu devletlerin kanunlarını taklit etmez; tam tersine taklitçi olmak yerine biz başkaları için bir örnek oluşturuyoruz. Yönetim az kişiye değil, çok kişiye bağlı olduğundan dolayı bizim sistemimize demokrasi denir. Bizim kanunlarımız insanlara özel yaşamlarında eşit haklar verir. Bununla birlikte toplum yaşamında saygınlık, sosyal sınıfa değil kişinin erdemine bağlıdır.

Sosyal sınıf, bir kişiyi herhangi bir toplumsal mevkiye sahip olmaktan alıkoymaz (…). Özel yaşama müdahale etmememize karşın toplumsal konuları ilgilenden kuralları benimseyiz. Biz yetkili konuma getirdiğimiz kişilere ibadi edebiz ve konuları kurallara, özellikle de ezilenleri korumak için konulmuş kurallara ve uymamanın utanç verici olarak kabul edildiği yazılı olmayan kurallara uyarız.

Dahası, zihinlerimizi dinlendirmek için birçok imkân sunarız. Tüm yıl boyunca düzenlenen oyunlar ve kurban şenlikleri, dinlenmek amacıyla tasarlanmış özel konutlarımızın zarafeti, bütün üzüntüleri uzaklaştıran keyif kaynaklarıdır. Ayrıca Atina'da ikamet edenlerin birçoğu, tüm dünyada üretilen mahsullerden alıp Atina'ya getirmez ki Atinalılar diğer ülkelerin mahsullerini kendi mahsulleri gibi bilirler.

Tukidids, Peloponnes Savaşı'nın Tarihi (uyarlamadır)

1

Aşağıdaki soruları yanıtlamak için önceki sayfada yer alan "Atina'da Demokrasi" adlı metinden yararlanınız.

Soru 1: ATİNA'DA DEMOKRASİ

B bölümündeki konuşmanın amaçlarından biri Peloponnes Savaşı'nın ilk yılında ölen askerleri onurlandırmaktı.

Bu konuşmanın DİĞER bir amacı nedir?

.....
.....
.....
.....
.....
.....

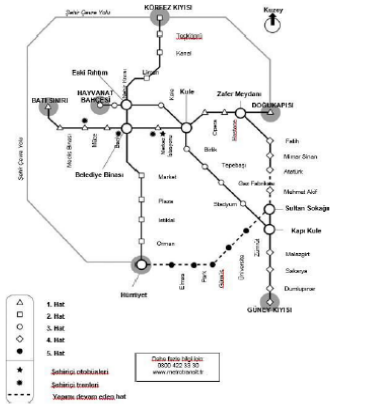
Soru 2: ATİNA'DA DEMOKRASİ

B bölümündeki konuşmayı kim yazmıştır? Yanıtınızı parçadan ifadelerle destekleyiniz.

.....
.....
.....
.....
.....
.....

2

METROTRANSİT



3

Bir önceki sayfada yer alan "Metrotransit" adlı metin bir metro sistemi hakkında bilgi vermektedir.

Aşağıdaki soruları yanıtlamak için "Metrotransit" adlı metinden yararlanınız.

Soru 3: METROTRANSİT

Hangi Metrotransit istasyonundan hem şehir içi otobüslere hem de şehir içi treniere binilebilir?

.....
.....
.....
.....

Soru 4: METROTRANSİT

Batı Girişi, Hayvanat Bahçesi, Hürriyet gibi bazı istasyonların etrafı gri gölgelendirme ile gösterilmiştir. Bu gölgelendirme bu istasyonlarla ilgili hangi bilgiyi sağlamaktadır?

.....
.....
.....
.....

4

SÜPERMARKET DUYURUSU

Yerfıstığı Alerjisi Uyarısı

Limon Kremalı Bisküviler

Uyarı Tarihi: 04 Şubat
Üretici Firma: Duru Gıda Ltd.
Ürün Bilgisi: 125 g Limon Kremalı Bisküviler
(Son kullanma tarihi: 18 Haziran ve 01 Temmuz)
Ayrıntılar: Bu gruplarda yer alan bazı bisküviler, içindekiler listesinde bulunmadığı halde yerfıstığı parçacıkları içerebilir. Yerfıstığı alerjisi olan kişiler bu bisküvileri **yememelidir**.
Tüketicinin Yapması Gerekenler: Bu bisküvilerden satın almışsanız, ürünü satın aldığınız **yere iade ederek** parasını **geri alabilirsiniz. Daha fazla** bilgi için 0 800 666 44 22 numaralı **telefonu** arayınız.

5

Önceki sayfada yer alan duyuru bir süpermarkete asılmıştır.

Aşağıda yer alan soruları yanıtlamak için bu duyurudan yararlanınız.

Soru 5: SÜPERMARKET DUYURUSU

Bisküvileri yapan firmanın adı nedir?

.....
.....
.....
.....

Soru 6: SÜPERMARKET DUYURUSU

Eğer siz bu bisküvilerden satın almış olsaydınız ne yapardınız?

.....
.....
.....

Neden bunu yapardınız?
Yanıtınızı desteklemek için metindeki bilgilerden yararlanınız.

.....
.....
.....
.....

Soru 7: SÜPERMARKET DUYURUSU

Neden duyuruda "Son kullanma tarihleri" belirtilmektedir?

.....
.....
.....
.....

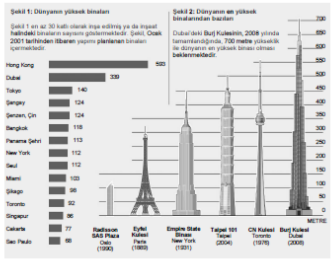
6

YÜKSEK BİNALAR

"Yüksek Binalar" 2006 yılında bir Norveç dergisinde yayımlanmış bir makaledir.

Şekil 1: Dünyanın yüksek binaları

Şekil 1 en az 30 katlı olarak inşa edilmiş ya da inşaat halindeki binaların sayısını göstermektedir. Şekil, Ocak 2001 tarihinden itibaren yapımı planlanan binaları içermektedir.



Şekil 2: Dünyanın en yüksek binalarından bazıları

Dubai'deki Burj Kulesinin, 2008 yılında tamamlandığında, 700 metre yüksekliğe ile dünyanın en yüksek binası olması beklenmektedir.

7

Aşağıda verilen soruları yanıtlamak için "Yüksek Binalar" makalesinden yararlanınız.

Soru 8: YÜKSEK BİNALAR

Dergideki makale yayımlandığında, şekil 2'de yer alan binalardan hangisi tamamlanmış en yüksek binaydı?

.....
.....
.....
.....

Soru 9: YÜKSEK BİNALAR

Norveç'te, Oslo'da bulunan Radisson SAS Plaza binası sadece 117 metredir. Bu bina neden şekil 2'de yer almıştır?

.....
.....
.....
.....

8

VARIŞ YERİ BUENOS AİRES

Patagonya[1], Şili ve Paraguay' dan gelen üç posta uçağı güney, batı ve kuzeyden Buenos Aires' e geri dönüyordu. Gece yarısına doğru havalanıp Avrupa'ya gidecek olan uçak orada bu uçakların taşıdıkları yükü **bekliyordu.**

Üç **pilotun her biri,** ağır motor kısmının **hemen** arkasında, uçuşa bucaksız şekre yaklaşırken gecenin kararlığında uçuşlarını planlıyorlardı. Bunu sonra, dağdan inen garip köylüler gibi, bulundukları konuma göre farkalı veya sakin olan gökyüzünden yavaş yavaş şehre **inecekleri.**

Tüm işleyişten sorumlu olan Riviar, Buenos Aires havalanı pisti üzerinde gidip geliyordu. Sessizdi, çünkü üç uçak da gelene kadar Riviar bir **huzursuzluk** duygusu içindeydi. Her geçen **dakika telgrafla elde** ulaştığı, **kendini kaderde** yanlı olanlardan **bir** şeyler koparıp alıyormuş **gibi hissediyor, belirsizlikleri** yavaş yavaş **yok edip akibini gecenin** karanlığının içinden koyuya doğru **getiriyordu.**

Adamlarından **biri gelip tahsis** mesajını Riviar'e **verdi:**

Şili' den **gelen** posta uçağı Buenos Aires' in ışıklarını görebildiğini bildiriyor.

Güzel.

Riviar kısa bir süre sonra bu uçağın sesini duyacaktı. Gece, gelgitlerle ve ızlarla dolu bir denizin uzun süre sakladığı hazineleri kıyıya bırakışı gibi, uçaklardan birini salıveriyordu. **Daha sonra da** diğer ikisini geri verecekti.

Böylece bugünün işi bitmiş olacaktı. Yorgunluktan bitkin düşmüş olan ekip uyumaya gidecek, yerine yeni bir ekip gelecekti. Fakat Riviar'in hiç dinlenmeye vakti olmayacaktı: Şimdi de sırası **gelen** Arrupa postası onu karşılayacaktı. Bu **her** zaman böyle olacaktı. Sonsuza **dek.**

Antoine de Saint-Exupéry. *Vol de Nuit.* © Editions Gallimard

[1] Şili ve Arjantin'in güney bölgesi

9

Önceki sayfada yer alan "Varış Yeri Buenos Aires" adlı metin 1931 yılında yazılan bir romandan bir alıntıdır. Bu roman, yazarın Güney Amerika'da posta pilotu olarak çalıştığı zamanlardaki tecrübelerine dayanmaktadır.

Aşağıdaki soruları yanıtlamak için bu alıntıdan yararlanınız.

Soru 10: VARIŞ YERİ BUENOS AİRES

Riviar mesleği hakkında ne hissetmektedir? Yanıtınızı destekleyecek bir neden belirtmek için metinden yararlanınız.

.....
.....
.....
.....

Soru 11: VARIŞ YERİ BUENOS AİRES

"Varış Yeri Buenos Aires" 1931 yılında yazılmıştır. Sizce Riviar'in bugün de benzer kaygıları **olur muydu?** Yanıtınızı açıklamak için **bir neden belirtiniz.**

.....
.....
.....
.....

10

ÖĞRENCİ GÖRÜŞLERİ

Dışarıda bir yerlerde birçok insan açık ve hastalıktan ölürken, **biz gelecekteki** gelişmelere **çok daha fazla ilgilenmekteyiz. Biz** bu insanları geride bırakıyor ve unutup yolumuza devam ediyoruz. Büyük firmalarca, her yıl milyarlarca dolar uzay araştırmalarına aktılmaktadır. Uzayın keşfi için harcanan bu para **aşağıdakilere** değil de yoksulların yararına kullanılmış olsaydı, milyonlarca insanın çektiği acılar **hafifletilebilirdi.**

Ayşe

Uzayı keşfetme mücadelesi birçok insanın **ilham** kaynağıdır. Biz binlerce yıldır gökyüzünün nasıl olduğunu merak ediyor, yıldızlara ulaşmanın ve onlara dokunmanın özlemini duyar ve var olabildiğini sadece düşlediğimiz bir şeyle iletişim kurmayı bekler ve bilmek isteriz... **Biz** yalnız mıyız?

Uzay araştırmaları bir tür öğrenmedir ve öğrenme ise dünyamızın yönlendirilmesidir. Gerçekler bize dünyanın var olan sorunlarını hatırlatırken, hayaller zihnimizi açarlar. Bizi geleceğe taşıyacak olan **hayallerin** öngörüleri, umutları ve arzularıdır.

Betül

Biz yer altındaki petrol için yağmur ormanlarını tahrip ediyor, uranyum uğruna kutup topraklarına mayın döşüyoruz. Kendi yarattığımız sorunların çözümü uğruna başka bir gezegeni de **tahrip eder miydik? Kesinlikle!**

Uzay araştırmaları, insanlığın sorunlarının çevre üzerindeki egemenliğimizin artması ile çözülebileceğine dair tehlikeli inancı güçlendirmektedir. İnsanlık, nehirler ve yağmur ormanları gibi doğal kaynakları, sanki köşeyi dönünce yararlanılmayı bekleyen başka bir gezegen varmış gibi kötüye kullanma özgürlüğü olduğunu **hissetmeye devam edecektir.**

Dünyaya yetersiz zarar vermiş durumdayız. Uzayı rahat bırakmalıyız.

Dursun

11

Dünyanın kaynakları hızla yok olmaktadır. Dünya nüfusu çarpıcı bir hızla artmaktadır. Eğer bu şekilde yaşamaya devam edersek, yaşam sürdürülemeyebilir. Kırılık ozon tabakasının delinmesine neden olmuştur. Verimli topraklar elden gitmektedir ve yakında besin kaynaklarımız azalacaktır. Şimdiden aşırı nüfus artışının neden olduğu kıtlık ve sağın hastalık olayları görülmektedir.

Uzay kendi yararımıza kullanabileceğimiz çok büyük bir boş alandır. Uzay araştırmalarını destekleyerek, bir gün üzerinde yaşayabileceğimiz bir gezegen bulabiliriz. Şu anda bu hayal bile edilemez gibi görünüyör amcak uzay yolculuğu da bir zamanlar imkânsız olduğu düşünülen bir kavramdı. Acil sorunların çözümü adına uzay araştırmalarını durdurmak büyük bir dar kafalılık ve kısa vadeli bir yaklaşımdır. Sadece bu kuşak için değil gelecek kuşaklar için de düşünmeyi öğrenmek zorundayız.

Fatih

Uzay araştırmalarının sağlayacağı şeyleri göz ardı etmek tüm insanlık için büyük bir kayıp olurdu. Evreni ve onun sağlığını daha iyi bir şekilde anlama olasılığı boşa harcanamayacak kadar değerlidir. Diğer çok otomleri ile ilgili yapılan çalışmalar çevre sorunlarımız ve etkinliklerimizi yönetmeyi öğrenmediğimizde dünyaya neler olabileceği hakkında **bilgilerimiz** artırmıştır.

Uzay yolculuğu araştırmalarının ayrıca olaylı yararları vardır. Laser teknolojisi ve diğer tıbbi tedavilerin yaratılışını uzay araştırmalarına borçluyuz. Teflon gibi maddeler insanlığın uzaya **yolculuk** arayışıyla ortaya çıkmıştır. Bu nedenle uzay araştırmaları için yaratılan teknolojilerin herkes için kısa vadede yararları **olabilir.**

Kadriye

12

*Önceki **iki sayfada yer alan metinler**, **okulun son sınıfına devam eden** öğrenciler tarafından yazılmıştır. Aşağıdaki soruları yanıtlarken bu bölümlerden yararlanınız.*

Soru 12: ÖĞRENCİ GÖRÜŞLERİ

Beş öğrenci tarafından ifade edilen fikirleri düşündüğünüzde, en çok hangisinin görüşüne katılırsınız?

Öğrencinin adı: .....

Seçiminizi, kendi görüşünüze **ve** öğrenci tarafından sunulan görüşlere dayanarak kendi sözcüklerinizle açıklayınız.

.....

.....

.....

13

AFRİKA YÜRÜYÜŞÜ

Kuzey Drakensberg Yürüyüşü: Güney Afrika / Lesoto

Bilgi

GENEL AÇIKLAMA

- **Kuzey Drakensberg** Yürüyüşü, Kuzey Drakensberg'in doruklarındaki dik kayalıklarından geçişi de kapsamaktadır. Güzergâh, yaklaşık olarak 40 mil (65km) uzunluğundadır. Güney Afrika ve Lesoto sınırı üzerinde, tamamlanması 5 **yorucu** gün sürer. Yürüyüş **unutulmayacak yerlerle doludur.** Demirmerdiven'e giderken, **amfi tiyatronun** üzerinden şeytan Dişine **kadar nefes kesen bir manzara ve saat** alarmınızı **kurmaya değecek** Mponjwane'den gün doğuşu **bunlardan** birkaçıdır.
- Başlangıç: Başlangıç oto parkı, **Kraliyet Ulusal** Parkı.
- Bitiş: **Katedral Zirve Otell.**
- Zorluk ve Yükseklik: Bu, Drakensberg sıradağarının üora köşelerinin birinde yapılan bir yüksek dağ yürüyüşüdür. Gidiş olduıça zor olabilir ve günler sürebilir. İyi yön belirleme **becerileri** güvenli bir geçiş için şarttır.

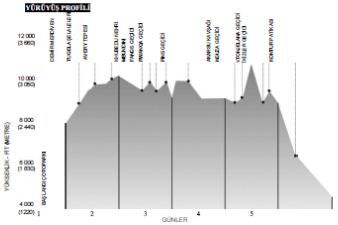
ZAMANLAMA VE MEVSİM

- **Ziyaret için en uygun aylar: Nisan,** Mayıs, **Haziran ya da** Eylül, **Ekim,** Kasım.
- İklim: Drakensberg'de yazlar çok sıcak ve çok yağışlı olabilir. Kış daha kuru geçer, fakat yüksek yerlerde kar olmak üzere, her an yağmur yağma ihtimali vardır. İlkbaharda ve sonbaharda, gün içinde sıcaklık idealdir. (60°F/15°C ile 70°F/20°C arasında), fakat gece sıcaklık çoğu **zaman** donma noktasının altına düşer.

SICAKLIK VE YAĞIŞ

Günlük **ortalama en** yüksek sıcaklık												
(°F)	72	70	70	66	63	60	60	63	66	68	70	70
(°C)	22	21	21	19	17	15	15	17	19	20	21	21
Günlük **ortalama en** düşük sıcaklık												
(°F)	55	55	54	48	46	41	41	43	46	48	52	54
(°C)	13	13	12	9	8	5	5	6	8	9	11	12
Aylık **ortalama** yağış												
(Inç)	9,3	8,5	7,7	3,1	1,1	0,6	0,5	1,3	2,4	4,0	6,5	7,9
(Milimetre)	237	216	196	78	29	14	12	33	62	101	165	201
	OCA	**SUB**	**MAR**	**NİS**	**MAY**	**HAZ**	**TEM**	**AĞU**	**EYL**	**EKİ**	**KAS**	**ARA**

14



*Bir **önceki sayfada** ve yukarıda **yer alan** "Afrika Yürüyüşü" **metni, Klasik** Yürüyüşler adlı **kitaptan** bir alıntıdır.*

*Aşağıda **yer alan** soruları yanıtlamak için "Afrika Yürüyüşü" adlı **metinden** yararlanınız.*

15

Soru 13: AFRİKA YÜRÜYÜŞÜ

Başlangıç otoparkı **hangi** yükseltide bulunmaktadır? Yanıtınızı **fit ve metre cinsinden veriniz.**

..... fit

.....metre

Soru 14: AFRİKA YÜRÜYÜŞÜ

Yürüyüş süresince, hangi günün en zor gün olacağını düşünüyorsunuz? Metinde verilen **bilgileri kullanarak,** yanıtınızı **destekleyiniz** için **bir neden** yazınız.

.....

.....

.....

.....

.....

.....

.....

16

MACONDO

Sayısız ve akıl almaz buluşlarla başarı dönen Macondolular şaşkınlıklarının nerede başladığını bilemediler. Bütün gece oturuyorlar, Aureliano Triste'nin ikinci tren seferinde getirdiği santralden enerji alan elektrik ampullerinin öğün ışığını seyrediyorlardı. Santralin görüntüsüne alışmaları oldukça uzun sürdü. Yokluğu tüm bir tüccar olan Bruno Crespi'nin asılan ağzını çiğleri olan tiyatroda oynattığı canlı resimler ise, Macondoluları çok öfkelendiriyordu. Çünkü bir filmde ölüp gömülen ve ardından seyircilerin gözyaşı döktüğü biri, bir sonraki filmde yeniden canlanıyor ve bu kez Arap kılığında ortaya çıkıyordu. Oyuncuların başarılı geçen felaketleri paylaşmak için adam başına iki sentavos verip bilet alan **seyirciler,** bu sahtekârlık karşısında galeyana geldiler ve sandalyeleri kırdılar. Bruna Crespi'nin zoralması üzerine, belediye başkanı bir bildiri yayınlayarak, sinemanın seyircilerin duygusal patlamalarını gerektirmeyen bir görüntü makinesi olduğunu açıkladı. Bu **cesaretli** kınama açıklamadan sonra, çoğu kişi kendilerini yeni ve gösterişli bir çingene numarasının kurbanı sayıdılar ve kendi derilerinin kendilerine yetiştiğine, bir de hayali kişilerin düzmece felaketlerine gözyaşı dökmenin **gereksiz** olduğuna **karar verip** sinemayı **boykot ettiler.**

17

*Önceki sayfadaki metin bir romandan alınmıştır. **Hikayenin bu** bölümünde **tren yolu** ve **elektrik** Macondo **ilimi hayali kasabaya yeni** ulaşmış ve **ilk sinema** açılmıştır.*

***Metni** göz önünde **bulundurarak** aşağıdaki soruları cevaplayınız.*

Soru 15: MACONDO

Filmlerin hangi özelliği Macondo'luların öfkelenmesine **neden** olmuştur?

.....

.....

.....

.....

.....

.....

Soru 16: MACONDO

Macondo'luların sinemanın değeriyle ile ilgili aldıkları son karara katılıyor musunuz? Filmler hakkında **kendi** düşüncenizi, onların düşünceleri ile karşılaştırarak yanıtınızı yazınız.

.....

.....

.....

.....

.....

.....

18