E-ISSN: 2458-8342, P-ISSN: 1301-3718
Ankara University Journal of Faculty of Educational Sciences
2026, 59(2), 549-590
DOI: 10.30964/auebfd.1807153
Research Article

The Use of Artificial Intelligence Tools in Assessing Content Validity: A Comparative Study with Human Experts

Hatice Gürdil [1], Hatice Özlem Anadol [2], Yeşim Beril Soğuksu[3]

1 Turkish Ministry of National Education, Ankara, Türkiye
2 College of Foreign Languages, Gazi University, Ankara, Türkiye
3 Turkish Ministry of National Education, Ankara, Türkiye

ABSTRACT

This study investigated whether artificial intelligence (AI) systems could evaluate the content validity of B1-level English reading comprehension items comparably to human experts. A 25-item multiple-choice test was rated by four human experts and four AI models (GPT-4o, Gemini 2.0, Replika, and Chatfuel). Items were evaluated for their relevance to B1-level reading skills and representativeness of the targeted content using a 3-point scale: 1 = not representative, 2 = needs improvement, and 3 = representative. To support methodological transparency, a standardized few-shot prompting framework was used. Each model was accessed in a new session to minimize context drift, and identical prompts were provided across evaluations. Content Validity Ratio (CVR) and Item Content Validity Index (I-CVI) values were calculated from the ratings and compared using the Wilcoxon Signed-Rank Test. Results showed no statistically significant difference between AI- and human-generated scores, indicating similarity between the two evaluator groups under the study conditions. Although fine-tuning was not possible because the models were closed source, standardized prompting supported reliability and replicability. The findings suggest that AI tools can complement expert judgment in content validity assessment and support hybrid human–AI frameworks, provided that methodological control, model transparency, and context management are maintained.

Keywords: content validity, artificial intelligence, few-shot prompting, hybrid evaluation, educational measurement

1 Corresponding Author: Dr., **E-mail:** gurdilhatice@gmail.com **ORCID ID:** 0000-0002-0079-3202 **ROR ID:** https://ror.org/00jga9g46
2 Lec. Dr., **E-mail:** ozlemanadol@gazi.edu.tr **ORCID ID:** https://orcid.org/0000-0002-2881-9806
ROR ID: https://ror.org/054xkpr46
3 Dr., **E-mail:** berilsoguksu@gmail.com **ORCID ID:** 0009-0004-0870-4974 **ROR ID:** https://ror.org/00jga9g46
Received Date: 21.10.2025 **Accepted Date:** 24.06.2026 **Publication Date:** 06.08.2026
Citation: Gürdil, H., Anadol, H. Ö., & Soğuksu, Y. B. (2026). The use of artificial intelligence tools in assessing content validity: A comparative study with human experts. *Ankara University Journal of Faculty of Educational Sciences, 59*(2), 549-590. https://doi.org/10.30964/auebfd.1807153
All ethical declarations related to this article are provided on the declarations page (Page: 587).

The assessment of individuals' skills, attitudes, and behaviors is essential for both research and real-world applications in the domains of psychology and education (Cohen & Swerdlik, 2021). The development of tests and scales is crucial to achieving accurate and insightful evaluations. Based on an individual's skills, preferences, and psychological traits, these instruments form the basis for assessing cognitive, emotional, and psychomotor skills. Validity and reliability are two key ideas in test development, and they are carefully considered throughout the complex process of developing these assessment tools (Nunnally & Bernstein, 1994).

Assuring content validity is an essential part of test development. How well a test's or scale's items capture the concept they are meant to measure is known as content validity. In other words, the test must cover every facet of the construct without leaving out any crucial details. High content validity is essential for drawing reliable conclusions about individuals based on their test results. For instance, a well-designed reading comprehension test should include items that assess a range of reading skills, such as understanding vocabulary, analyzing texts, and identifying main ideas (Messick, 1989).

Assessing content validity is frequently accomplished through pilot studies, in which a small sample of people take the exam to see its efficacy. Researchers are able to see items that do not perform as predicted or that do not accurately reflect the construct during these pilot studies. The test is significantly enhanced by the insights gained from these studies before being administered to a broader population. However, pilot studies may not always be feasible, particularly in contexts where time or resources are limited. For this reason, content validity is not assessed through a single fixed procedure; rather, different approaches are used depending on the purpose of the test, the stage of development, and the available resources. One of the commonly used approaches in the literature is expert-based judgment, especially when the aim is to examine whether the items adequately represent the intended construct before wider administration. In such cases, the content validity of the test is assessed through expert consultation (Haynes et al., 1995).

Within expert-based content validity studies, researchers may use both qualitative judgments and quantitative indicators to summarize expert opinions. The Content Validity Index (CVI) and the Content Validity Ratio (CVR) are two essential instruments in this procedure. These statistical measures are employed to evaluate the validity of an item within a scale or test, ensuring that it accurately represents the construct being measured. Experts evaluate each item to determine whether it is essential for measuring the specific construct, useful but not critical, or unnecessary. Based on these assessments, the CVR is computed, yielding a numerical result that represents the content validity of every item. In contrast, the Content Validity Index (CVI) consolidates individual Content Validity Ratios (CVRs) to offer a comprehensive evaluation of a test's or scale's content validity (Lawshe, 1975). A significant factor affecting the validity and reliability of CVR and CVI computations is the number of experts engaged in the process. In general, evaluations are more

reliable when a greater number of experts are involved. This is because a broader range of professional perspectives can reduce the impact of individual biases and increase the likelihood that the final measures accurately represent the construct being examined. A large and varied panel of experts can offer a more thorough evaluation, guaranteeing that the test items are representative and pertinent (Mertler et al., 2025).

In practice, however, assembling a sufficiently large group of qualified experts can be challenging. Securing an adequate number of participants is often challenging due to limited resources and the availability of experts. The absence of sufficient expert input can compromise the validity and reliability of content validity assessments, potentially leading to biased or less accurate evaluations. In such cases, ensuring that scales and tests meet the standards for content validity can pose significant challenges for researchers.

Given these challenges, researchers have explored alternative approaches, including artificial intelligence (AI), which offers a scalable and data-driven method for evaluating test items. Leveraging advanced algorithms and large datasets, AI systems can evaluate test items, compare them against predefined standards, and calculate CVR and CVI values using machine learning and natural language processing techniques. While AI cannot fully replicate the nuanced insights of human experts, it can be particularly valuable when expert input is limited or unavailable. By integrating AI into the validation process, researchers can enhance efficiency and accuracy, with further improvements achieved through human expert evaluations. This study is significant for several reasons. Firstly, by examining the use of AI technologies in the content validity evaluation of English reading comprehension items, this study provides preliminary evidence on how AI tools may support expert-based item review processes, particularly in terms of efficiency and access to additional evaluative feedback. Validity is a critical aspect of test construction, as it ensures that assessments measure the constructs they are intended to evaluate. Additionally, AI techniques offer a novel approach to validating tests, particularly in scenarios where expert resources are limited or unavailable.

Upon reviewing the studies conducted on the potential of artificial intelligence tools in education, Sihite et al. (2023) evaluated ChatGPT's competence in generating reading comprehension questions for academic texts and emphasized the need for further development in this process. Additionally, Morjaria et al. (2023) examined ChatGPT's performance in short-answer assessment questions, highlighting potential risks related to the validity of such assessments. On the other hand, research conducted in the context of scale development has also made significant contributions to understanding the different dimensions of ChatGPT. Nemt-Allah et al. (2024) designed the "ChatGPT Usage Scale" for graduate students, examining dimensions such as academic writing, task support, and dependency. Lee and Park (2024) developed the ChatGPT Literacy Scale to assess its impact on knowledge transfer and creativity. Moreover, the effect of artificial intelligence tools on personalized learning processes has also been noteworthy. Siegle (2023) suggested that these tools could

enrich the learning processes of gifted students, while Lin (2024) proposed that ChatGPT could support self-regulated learning among adult learners. In the field of language learning, Kohnke et al. (2023) discussed the opportunities, controversial aspects, and disadvantages of ChatGPT in language teaching and learning, while Saraswati et al. (2023) found that AI Replika could help students practice with native speakers, though it remained limited in understanding cultural meanings. Furthermore, significant findings have emerged in studies comparing artificial intelligence tools with different AI models and human performance. Kim et al. (2024) revealed that, compared to human scores, ChatGPT showed limitations in terms of accuracy and reliability in second-language writing tasks. Hayawi et al. (2024) investigated the capacity of LLMs (BARD, GPT) to be indistinguishable from human writing, while ul haq Akhoon et al. (2024) examined the effectiveness of ChatGPT and Google Gemini in academic content generation and the associated ethical validation issues. Wachinger et al. (2025) found that ChatGPT's performance in qualitative data analysis largely aligned with that of an experienced researcher. Similarly, Gürdil et al. (2025) observed some inconsistencies in item parameters when examining ChatGPT's effectiveness in data generation, but they found results similar to human data. On the other hand, Tenhundfeld (2023) highlighted the need for further development in ChatGPT's human and AI interaction, whereas Onal and Kulavuz-Onal (2024) assessed ChatGPT's high performance in higher education assessment tasks, stating that these tools play a complementary role to human expertise. More recent studies have also discussed the role of generative AI in educational assessment with a more cautious perspective. Xia et al. (2024) reviewed empirical studies on the use of generative AI in higher education assessment and noted that these tools may change assessment practices, but their use requires careful attention to assessment design, teacher preparation, and institutional policies. Similarly, Weng et al. (2024) emphasized that generative AI has introduced new discussions about how student learning should be assessed and how assessment tasks may need to be redesigned. Kizilcec et al. (2024) compared educator and student perspectives and reported that generative AI is perceived to have a considerable impact on assessment quality, particularly in relation to authenticity, academic integrity, and higher-order thinking. In a recent study, Zacharis and Papadakis (2025) also found that AI-based grading should not be treated as interchangeable with human evaluation in high-stakes assessment contexts. Taken together, these studies suggest that AI tools may support assessment processes, but their use should be guided by human expertise and validity-related considerations. Upon reviewing the studies, it is evident that the potential of AI-based tools in education has been examined in terms of validity, reliability, and user experience, with comparisons to human performance. However, no research has been found that explores the potential of artificial intelligence tools in determining content validity. Content validity is a critical factor that assesses how accurately a measurement tool captures the entire domain it is intended to measure. Wynd et al. (2003) demonstrated that low inter-expert agreement in the Content Validity Index is crucial for scale revision. This topic is frequently addressed in studies (Ayre & Scally, 2014; Hinkin & Tracey, 1999; Newman et al., 2013), where the importance of

innovative approaches to evaluating content validity is emphasized. However, an examination of the literature reveals that no research has specifically explored the potential of artificial intelligence tools in determining content validity. This gap highlights the need for investigating the effectiveness of AI in content validity assessment.

Purpose of the Study

This study aims to explore the potential of artificial intelligence (AI) in assessing the content validity of reading skill items from a B1 level English exam. Specifically, the study seeks to examine the level of agreement among AI tools and calculate Content Validity Ratios (CVR) and Item Content Validity Indices (I-CVI) based on the ratings provided by four distinct AI tools and four human experts. In this context, the following research questions were formulated:

1. Are there statistically significant differences among the scores given by all evaluators for the items?
2. How consistent are the ratings among all evaluators?
3. Do the CVR values derived from AI evaluators differ significantly from those provided by human evaluators?
4. Do the I-CVI values derived from AI evaluators differ significantly from those provided by human evaluators?

Significance of the Study

This study explores whether AI can serve as a scalable alternative to expert-based content validity assessments, reducing time and resource constraints. If AI demonstrates effectiveness, it could provide a scalable and cost-efficient alternative, reducing resource demands and enhancing the accessibility of high-quality test validation processes. The study also aims to compare AI-generated content validity evaluations with those made by a human expert. This comparison is essential for evaluating the effectiveness and reliability of AI tools in educational assessment, identifying their strengths and limitations, and informing future test development practices.

Furthermore, the study ensures that its findings have practical implications for language instructors and item writers. Effective AI-based content validity evaluations could lead to more accurate and representative test designs, ultimately benefiting students by providing more reliable measures of their language abilities. Finally, by focusing on a specific application—content validity assessment—this study contributes to the growing body of research on AI applications in education. Successfully integrating AI into this domain could pave the way for further advancements and innovative uses of AI in educational settings, potentially transforming how tests are designed, administered, and validated. Overall, the study's findings could have a substantial impact on test development practices, resource

allocation, and the role of artificial intelligence in educational contexts. Additionally, the results may offer valuable insights into the intersection of technology and assessment, guiding future research and practice in the field.

Method

Research Design

This study was designed as a comparative descriptive study. The main purpose was to compare the content validity evaluations provided by human experts and AI tools for B1-level English reading comprehension items. Since the study focused on examining and comparing evaluator ratings rather than testing an intervention, a descriptive and comparative research approach was considered appropriate.

Data Development and Implementation of the Test

In this study, a 25-item test consisting of B1-level English reading texts and questions was used for the purposes of the study. In the preparation phase of the test, it was developed by four expert English language instructors. To ensure the validity and reliability of the test, the questions were reviewed by the coordinator of the testing unit, and necessary revisions were made. The items were developed within an item specification framework aligned with CEFR-based B1 reading sub-skills. In this framework, attention was paid to the targeted learning outcome, the relevant reading skill area, the suitability of the item for the B1 level, the alignment between the text and the item, the plausibility and functionality of distractors, and the linguistic clarity of the item. The same criteria were also reflected in the expert evaluation form used during the content validity review. Experts were asked to evaluate whether each item represented the intended reading skill, was appropriate for the B1 proficiency level, was directly related to the reading text, included functional distractors, and was linguistically clear and accurate. The test's achievements are compatible with the Common European Framework of Reference for Languages (CEFR). At this point, the respondents are expected to have reading skills corresponding to the B1 level. This means they should be able to understand the general meaning of different types of texts (newspaper articles, essays, stories, etc.), identify the main ideas and important details in the texts, and make inferences about information not explicitly stated. They should also have vocabulary knowledge appropriate to the B1 level and be able to deduce the meanings of words within the text, as well as understand the organization of the text and the connections between paragraphs.

The B1 level was chosen for the study because it represents an intermediate level of English proficiency and appeals to a wide audience. Individuals at this level can understand a variety of texts they may encounter in daily life and can use English for academic or professional purposes. In addition, the B1 level creates a more homogenous group compared to other levels, which ensures that the results of the study are more reliable.

Evaluators: Human and AI Tools

After the test was validated, the 25 items it contained were evaluated by 4 human experts and 4 AI tools. The human experts are experts in the field of English Language Teaching, have at least 5 years of experience and are proficient in CEFR assessment processes. These criteria were used to ensure that the evaluators had relevant subject-area expertise, practical experience in language assessment, and familiarity with CEFR-based reading skill descriptors. Since the items were developed for a B1-level English reading test, experts with experience in CEFR-aligned assessment were considered appropriate for evaluating the relevance and representativeness of the items. The AI tools used in the study were GPT-4o, Gemini 2.0, Replika, and Chatfuel. These tools were selected to include different types of accessible AI-based systems with text-processing and response-generation capacities. GPT-4o and Gemini 2.0 were included as advanced large language models, whereas Replika and Chatfuel were included as chatbot-based systems. This selection allowed the study to compare evaluations obtained from different AI-supported response environments. Developed by OpenAI, ChatGPT-4o is highly successful in analyzing and interpreting complex expressions naturally, thanks to its extensive language processing capacity. ChatGPT-4o, which stands out as a leading model among current artificial intelligence technologies, can be used to analyze texts from different disciplines. While the basic text processing features of GPT-4o can be used for free, some advanced features such as audio and image processing require a subscription (Ofgang, 2024; OpenAI, 2023).

Gemini 2.0 is a multimodal artificial intelligence model developed by Google DeepMind. This model strengthens the evaluation process by drawing not only on text analysis but also on visual and contextual information. The strong performance of Gemini 2.0 in text-based tasks was an important rationale for its use in the present study (Hassabis & Kavukcuoglu, 2024). Replika, on the other hand, is a social chatbot that employs advanced natural language processing (NLP) technologies and is built upon the Generative Pretrained Transformer 3 (GPT-3) model. Its technological infrastructure enables it to generate language responses based on a large dataset derived from user conversations, allowing for contextually rich and nuanced interactions (Pentina et al., 2023).

In addition to these tools, Chatfuel was also selected for this study. Chatfuel is an AI-based chatbot that uses natural language processing (NLP) techniques to interpret text and generate responses. While it is commonly used in customer service and automated messaging processes, it is also employed to provide feedback and create interactive learning environments. In their study, Ahmed & Hussein (2020) presented a chatbot design that uses the Chatfuel platform to help Kurdish-speaking individuals find answers through online conversations instead of direct contact with human agents. In this study, the use of AI-supported feedback mechanisms in evaluation processes is explored, and for this reason, Chatfuel has been selected.

Ethical Committee Approval

This study was reviewed by the Ethics Committee of Gazi University Rectorate at its meeting dated 23.09.2025 and numbered 14, and was approved as ethically appropriate with the document dated 09.10.2025 and numbered E.1352469. The research code number is 2025-1649.

Methodological Rigor

AI Prompting Procedure and Context Control

To ensure that the evaluations were comparable and replicable, all AI systems were accessed independently in new sessions. Before each evaluation, previous conversation histories were cleared to prevent context drift—the unintended influence of earlier prompts or responses on subsequent outputs. A standardized few-shot prompting framework was adopted across all models. Each AI system was first provided with two sample items containing correct labels (valid vs. invalid) to establish the rating logic. Following this short calibration step, the 25 items of the B1 reading comprehension test were entered sequentially, using the same input structure.

The main prompt was: "Rate the relevance of this item to the intended reading skill on a scale of 1 (not relevant) to 3 (highly relevant). Provide only a numerical score." This approach allowed the AI systems to emulate human evaluators' judgment criteria without introducing subjective variability across runs. The consistency of outputs across repeated sessions was checked manually to verify internal stability.

Reliability and Consistency Check

To determine the internal consistency of the AI evaluations, repeated trials were conducted at two different times using the same standardized prompts. The minor variations observed in the raw scores were found to be statistically non-significant. The inter-rater reliability among the eight evaluators (four human and four AI) was calculated using Fleiss' Kappa and found to be .431, indicating a moderate level of agreement among the evaluators. Differences between the evaluator groups were analyzed using the Kruskal–Wallis H Test, and no significant differences were observed ($p > .05$).

Data Analysis

In this study, human and AI evaluators assessed each test item by assigning scores ranging from 1 to 3 to measure how well the items reflected the intended content. The scoring system was straightforward: '1' meant the item did not represent the content well; '2' indicated the item needed improvement and '3' suggested the item represented the content effectively. Once all items were evaluated, the scores were organized into a table. These ratings were then used to calculate the Content Validity Ratio (CVR) and Item-Level Content Validity Index (I-CVI) for each item. To check whether there were significant differences in the scores given by the evaluators, a Kruskal-Wallis H test was used. This test is a non-parametric alternative

to One-Way ANOVA, ideal for analyzing ordinal data that does not follow a normal distribution (MacFarland & Yates, 2016). The analysis was conducted using the 'stats' package in R (R Core Team, 2024).

Each item was reviewed by 8 evaluators -4 humans and 4 AI tools. To assess the level of agreement between human and AI evaluators and among AI tools themselves, the Fleiss Kappa statistic was applied. Fleiss Kappa is a measure of agreement that works well for categorical data when multiple evaluators are involved (Fleiss, 1971; Hallgren, 2012). The calculations were carried out using the 'irr' package in R (Gamer et al., 2019). Fleiss Kappa values range from -1 to +1, with +1 indicating perfect agreement. In addition, the content validity of the test items was evaluated using two measures: the Content Validity Ratio (CVR) and the Item Content Validity Index (I-CVI). These measures were calculated based on the scores provided by both human and AI evaluators.

Content Validity Ratio (CVR)

CVR, introduced by Lawshe (1975), is a widely used method to assess whether test items adequately represent the intended content. Experts categorize each item as 'Essential', 'Useful but not essential' and 'Not necessary'. Items rated as Essential by a certain number of evaluators are kept, while others are discarded (Almanasreh et al., 2019). The CVR is calculated using the following formula:

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} \qquad 1$$

Where 'n_e' is the number of evaluators who rated the item as "Essential." and 'N' is the total number of evaluators. The CVR score ranges from -1 to +1: '+1' indicates perfect agreement; '-1' indicates complete disagreement and '0' means half of the evaluators rated the item as "Essential." Based on Lawshe's (1975) cut-off points, items with a CVR of at least 0.99 were included in the final version of the test.

Content Validity Index (I-CVI)

The Item-Level Content Validity Index (I-CVI) focuses on evaluating the relevance of individual items. Although I-CVI is often calculated using a 4-point relevance scale in the literature (Davis, 1992), the present study used a 3-point rating scale in line with the evaluation procedure applied to both human and AI evaluators. In this scale, "1" indicated that the item did not represent the intended content well, "2" indicated that the item needed improvement, and "3" indicated that the item represented the intended content effectively. For the I-CVI calculation, ratings of "3" were considered relevant, whereas ratings of "1" and "2" were not considered fully relevant. Therefore, the I-CVI value for each item was calculated as the proportion of

evaluators who rated the item as relevant. Lynn (1986) suggested that smaller panels (e.g., 3-5 evaluators) can use either 3- or 5-point scales. The I-CVI is calculated using this formula:

$$I - CVI = \frac{\text{Number of experts who rated the item as relevant}}{\text{Total number of experts}} \quad (2)$$

When there are fewer than 6 evaluators, Lynn (1986) recommended that the I-CVI should be 1 for the item to be considered valid. Similarly, Polit *et al.* (2007) noted that there should be 100% agreement among evaluators when the panel includes 3 or fewer experts. In this study, only items with an I-CVI of 1 were included in the final form. Items with an I-CVI below 1 were flagged for revision.

Results

The Kruskal-Wallis test was conducted to assess whether there was a significant difference between the mean ranks of the scores assigned by eight different evaluators (Human1, Human2, Human3, Human4, GPT-4o, Gemini 2.0, Replika, and Chatfuel) (Appendix 1). The results of the test and the corresponding effect size (ε^2) value are presented in Table 1.

Table 1

Kruskal-Wallis Test Results and Effect Size for Evaluators

Evaluators	Rank Mean	χ2	Df	*p*	Effect Size
8 raters*	13.00	10.18	7	0.180	0.05

**Human 1, Human2, Human3, Human4, GPT-4o, Gemini 2.0, Chatfuel and Replika*

When examining Table 1, the results of the Kruskal-Wallis test indicate that there is no significant difference between the evaluators ($\chi^2 = 10.182$, df = 7, $p > 0.05$, $\varepsilon^2 = 0.051$). This result suggests that all evaluators rated the test items in a similar manner. Additionally, the calculated effect size shows that the difference between evaluators is minimal, supporting the conclusion that there is no statistically significant difference between them. The results of the Fleiss Kappa analysis conducted to examine the consistency among all evaluators are presented in Table 2.

Table 2

Results of Consistency Analysis Among Evaluators

Evaluators	Kappa value	*p*
8 raters*	0.43	p < .001

**Human 1, Human2, Human3, Human4, GPT-4o, Gemini 2.0, Chatfuel and Replika*

As seen in Table 2, A Fleiss Kappa value of 0.431 indicates moderate agreement

among evaluators, suggesting some level of consistency, though differences exist in certain items. In other words, there are certain inconsistencies among the evaluators, but overall, they tended to score consistently. Additionally, the most frequently given score (mode) by all evaluators was calculated to be 3. These findings indicate that the evaluators scored consistently, and both human and AI evaluators showed similar performance in the evaluation processes. Based on the scores given by human and AI evaluators, the Content Validity Ratio (CVR) and Item Content Validity Index (I-CVI) values were calculated, and the obtained values are provided in Table 3.

Table 3

CVR and I-CVI Values for Human and Artificial Intelligence (AI) Scores

	CVR_AI	CVR_Human	I-CVI_AI	I-CVI_Human
1	0.50	0.50	0.75	0.75
2	0.50	-1.00	0.75	0.00
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	0.50	1.00	0.75
6	-0.50	-0.50	0.25	0.25
7	1.00	1.00	1.00	1.00
8	0.50	1.00	0.75	1.00
9	0.00	-1.00	0.50	0.00
10	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00
13	0.50	-0.50	0.75	0.25
14	1.00	1.00	1.00	1.00
15	1.00	0.00	1.00	0.50
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00

When examining the CVR and I-CVI values for Human and AI evaluators, it is observed that human evaluators validated 18 items, whereas AI evaluators validated 19 items (see Table 3). These findings indicate that both evaluator groups demonstrated similar performance in terms of content validity, with the AI evaluators approving one more item at a higher rate. However, both Human and AI evaluators decided not to finalize items 1, 2, 6, 9, and 13. These items were considered inadequate in terms of content validity by both groups and were not approved. The results show that both human and AI evaluators made parallel assessments on these items,

demonstrating a similar negative attitude towards them. Similarly, for items 3, 4, 7, 10, 11, 12, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, and 25, both groups decided to finalize the items. High approval was observed from both human and AI evaluators, and consistent and parallel results were obtained. This further reinforces the notion that AI can make human-like evaluations, yielding results that are consistent with human assessments.

However, item 8 (Appendix 2) reveals a disagreement among evaluators. Human evaluators found this item appropriate in terms of content validity, while AI evaluators preferred revision or removal. This difference may stem from the AI evaluations' inability to fully grasp the context of the text and insufficiently analyze key elements. AI evaluators tend to rely on syntactic and lexical patterns rather than contextual depth, which may explain why they recommended revision for item 8, while human evaluators likely conducted a more in-depth analysis, considering the societal impact and context of the text. Human evaluators considered cultural and societal context, whereas AI evaluators may have overlooked these aspects.

Additionally, for items 5 and 15 (Appendix 2), AI evaluators found both items appropriate in terms of content validity, while human evaluators adopted a more cautious approach, preferring to recommend revision or removal for these items. In the evaluation of item 5, AI evaluators, based on technical criteria, generally found the item sufficient and suggested its finalization, while human evaluators may have taken a more critical approach, assessing the text in terms of context, linguistic sensitivity, and content quality, and opted for a 'revise' or 'remove' decision. This difference may stem from several factors. First, human evaluators may have been more sensitive to the mismatch of the text with its context, while AI evaluators' standard-based evaluation process may overlook such nuances. Human evaluators may also have examined the linguistic quality of the text, the clarity of expressions, and the appropriateness for the target audience in more detail. Furthermore, the alignment of the text with pedagogical objectives and any deficiencies regarding students' learning outcomes may have influenced human evaluators' decisions. On the other hand, AI evaluators may have assessed the text primarily from a technical sufficiency perspective, while human evaluators focused more on content quality, context, and linguistic sensitivity. This discrepancy may be due to AI systems' tendency to overlook certain linguistic and pedagogical features.

Similarly, in the evaluation of item 15, AI evaluators recommended finalizing the question, while human evaluators decided on 'remove' or 'revise'. Several potential reasons for this difference can be considered. Human evaluators may have adopted a more critical approach, particularly regarding the context of the question and its relevance. For example, the statements about the festival's environmental impact and the connection with local producers in this question require a more complex and multifaceted interpretation. AI evaluators, focusing on the technical aspects of the text, concentrated on specific keywords, whereas human evaluators might have addressed the text's integrity and cultural context from a broader perspective. As a

result, the criticisms made by human evaluators may have required a more detailed evaluation of the way the question was phrased and the depth of meaning in the answer.

Differences Between the Calculated CVR and I-CVI for Evaluators' Scores

The differences between the Content Validity Ratio (CVR) values calculated for the items based on the scores of human and artificial intelligence (AI) evaluators were analyzed using the Wilcoxon Signed-Rank Test. The analysis revealed that the differences between human and AI evaluators were not statistically significant ($W = 335.5$, $p = 0.5708$). This result indicates that there was no significant change in the central tendency (median) between the two groups and, therefore, the CVR values are similar.

Although the results obtained from the CVR analysis show the general agreement on the decisions made regarding the items in terms of content validity, it is necessary to examine the suitability of each item in greater depth using the more detailed measurement, the Item Content Validity Index (I-CVI). In this context, the differences between the I-CVI values calculated from the scores of human and AI evaluators were analyzed using the Wilcoxon Signed-Rank Test. The analysis revealed that there was no statistically significant difference between the I-CVI values calculated by human and AI evaluators ($W = 335.5$, $p = 0.5708$). In other words, the decisions made by both evaluator groups regarding the suitability of the items were similar. The findings suggest that human and AI evaluators provided consistent results in content validity assessments and that AI systems offer reliability similar to that of human evaluations.

Discussion, Conclusion and Suggestions

This study investigated whether human and AI evaluators made similar evaluations in content validity assessments. The results indicated that for the vast majority of items, all evaluators made similar assessments. Furthermore, no evaluator was found to systematically score higher or lower than others. Additionally, the statistical analysis of the differences in the CVR and I-CVI values calculated for the ratings revealed no significant differences, suggesting that AI evaluators provided evaluations that were aligned and similar to those made by human evaluators. The absence of statistically significant differences between evaluators suggests that, within the scope of this study, AI evaluators may serve as a supportive tool in content validity analyses.

AI evaluators, based on technical criteria and adherence to standards, may produce rating patterns that are partly comparable to those of human evaluators in tasks that require objectivity and speed; however, they should be considered as a complementary tool rather than a replacement for human evaluators. In the future, the potential of AI evaluators to support faster and more systematic content validity analyses could be further explored, particularly within human-supervised or hybrid evaluation frameworks. In this study, overall, there was a moderate level of agreement

between evaluators in scoring an item, but differences were observed in some items between AI and human evaluators. These differences suggest that there may be different interpretations of evaluation criteria between the two groups. This indicates that certain items may be interpreted more subjectively, or they may require more detailed explanations. Notably, in some items, AI evaluators provided higher levels of approval, while human evaluators adopted a more cautious approach.

AI evaluators adopt a consistent and rule-based approach focused on technical competence but may overlook contextual and linguistic nuances. In contrast, human evaluators assess texts more critically and comprehensively, taking into account societal, cultural, and pedagogical contexts. It becomes clear that AI evaluation processes need improvement in areas such as contextual analysis, linguistic sensitivity, and alignment with pedagogical goals. The capabilities of human evaluators in these areas are particularly crucial when making critical pedagogical and cultural decisions. For example, in some items, AI evaluators, based on technical standards, considered the items sufficient and recommended their inclusion in the final form. On the other hand, human evaluators, taking into account more detailed aspects such as linguistic sensitivity, context, and content quality, made decisions of 'revision' or 'removal.' These differences may arise from the different prioritization of evaluation criteria and the inability of AI systems to fully grasp the subtleties and context of language. Especially in cases where context is critical, human evaluators may have conducted deeper analyses, considering the societal impact of the texts and their alignment with pedagogical objectives. In conclusion, it is suggested that the evaluation processes of AI evaluators should be improved in terms of contextual analysis, linguistic sensitivity, and alignment with pedagogical goals. These examples highlight that evaluators may adopt different perspectives on content validity, and these differences demonstrate that AI evaluators may not always fully reflect the decision-making processes of human evaluators. In this context, collaboration between AI and human evaluators could contribute to more consistent and reliable results in content validity assessments. While AI evaluators have the advantage of making quick and precise decisions based on linguistic cues, human evaluators may offer the potential for broader critical thinking and consideration of pedagogical concerns in a wider context. This situation suggests that while AI evaluators can perform similarly to human evaluators in certain assessment processes, it may not always be possible for them to fully replace human evaluators in all cases. Similarly, Kim et al. (2023) emphasize that human guidance is critical for AI to produce results with similar accuracy and reliability to humans. Additionally, Fjelland (2020) points out that AI cannot fully reflect human intelligence, while Shane (2019) suggests that AI may outperform human intelligence in certain tasks. In this context, it is important to provide AI evaluators with the necessary training, guidelines, and process improvements to work in harmony with human evaluations.

Taking into account the differences between both approaches in evaluation processes, emphasizing the importance of linguistic clarity and distinguishable options, is believed to contribute to AI evaluators performing similarly to human

evaluators. In this regard, defining the criteria more clearly and aligning evaluators' perceptions with these criteria could contribute to making the processes more consistent. The development of hybrid systems where AI-based assessment systems work alongside human evaluations could lead to more reliable and consistent results. Such systems could enable more efficient and accurate content validity analyses by leveraging the strengths of both parties. Furthermore, developing training programs and guidance systems that enhance collaboration between AI and human evaluators could help evaluators achieve more consistent results with a shared understanding. Similar to the results of this study, research conducted by Low et al. (2024), Tenhundfeld (2023), and Vassis et al. (2024) also highlights that AI-based systems do not fully replicate human-like evaluation performance. These studies emphasize that for these systems to align with human evaluations, further development is necessary.

Content validity is critically important, especially in situations where a thorough evaluation of the accuracy and validity of specific test items is required. In the literature, it is generally recommended to consult the opinions of at least 3 to 5 experts in content validity assessments. For example, Lawshe (1975) recommends at least three expert opinions, while Polit *et al.* (2007) suggest that the decision regarding whether an item should be finalized or not is typically made based on feedback from three experts. However, due to the workload of experts and other practical constraints, it is not always possible to reach these experts and receive feedback (Vassis *et al*., 2024). In such cases, the involvement of AI evaluators may allow for faster and more efficient content validity assessments. Similar to the studies by Low et al. (2024) and Abu Arqub *et al*. (2024), this study suggests that AI systems may provide supportive input in content validity reviews under controlled conditions. However, the findings do not imply that AI evaluators can replace human judgment, particularly in contexts requiring pedagogical, linguistic, or cultural interpretation.

However, this potential should be managed carefully, and possible methodological risks should be taken into account. In this regard, potential sources of bias such as context drift and user-specific histories were minimized in this study through the implementation of standardized prompting frameworks and the replication of evaluations across independent sessions. Nevertheless, these precautions alone may not be sufficient to achieve optimal reliability. This study utilized closed-source AI systems (e.g., GPT-4o, Gemini 2.0) without direct fine-tuning; however, it acknowledges that model retraining or fine-tuning on domain-specific validity data could further enhance performance. Therefore, future research should explore open-source large-language models such as LLaMA 2 or RoBERTa to enable systematic fine-tuning and reduce reliance on prompt-only optimization. Such methodological improvements could strengthen the consistency and generalizability of AI-based content validity assessments.

As a result, it can be said that AI evaluators could be an effective tool in technical analyses such as content validity, offering significant advantages, particularly in situations where the number of experts is high and the expertise requirements are

demanding. The partly comparable rating patterns observed between AI and human evaluators indicate that, in some cases, AI tools may support the evaluation process as an additional source of evidence rather than replace human intervention. This could allow for the use of AI evaluators' speed and efficiency advantages, especially in studies that require a large number of experts and have time constraints. The data processing speed of AI evaluators and their ability to make consistent decisions based on specific criteria can make the evaluation process more efficient. Future research may make progress in defining processes and evaluation criteria more clearly to make AI and human evaluations more aligned. This could help clarify how AI evaluators can be used more appropriately as supportive tools in content validity assessments. Additionally, conducting similar analyses with larger sample groups and different datasets would allow for testing these findings in a broader context. In conclusion, the strengths of AI evaluators can provide efficiency and effectiveness as a complement to human evaluations. However, it should still be acknowledged that human intervention and guidance are important. Future research is expected to improve the understanding of the balance between AI and human evaluations, making it possible for AI to be applied effectively in a wider range of areas.

E-ISSN: 2458-8342, P-ISSN: 1301-3718
Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi
2026, 59(2), 549-590
DOI: 10.30964/auebfd.1807153
Araştırma Makalesi

Kapsam Geçerliği Değerlendirmesinde Yapay Zekâ Araçlarının Kullanımı: İnsan Uzmanlarla Karşılaştırmalı Bir Çalışma

Hatice Gürdil [1], Hatice Özlem Anadol [2], Yeşim Beril Soğuksu [3]

1 Türkiye Cumhuriyeti Millî Eğitim Bakanlığı, Ankara, Türkiye
2 Gazi Üniversitesi, Yabancı Diller Yüksekokulu, Ankara, Türkiye
3 Türkiye Cumhuriyeti Millî Eğitim Bakanlığı, Ankara, Türkiye

ÖZ

Bu çalışmada, yapay zekâ (YZ) sistemlerinin B1 düzeyindeki İngilizce okuduğunu anlama maddelerinin kapsam geçerliğini insan uzmanlarla karşılaştırılabilir biçimde değerlendirip değerlendiremeyeceği incelenmiştir. Bu amaçla 25 maddelik çoktan seçmeli bir test geliştirilmiş ve maddeler dört insan uzman ile dört yapay zekâ modeli (GPT-4o, Gemini 2.0, Replika ve Chatfuel) tarafından değerlendirilmiştir. Maddeler, hedeflenen B1 düzeyi okuma becerileriyle ilişkileri ve hedef içeriği temsil etme düzeyleri açısından 3'lü bir derecelendirme ölçeğiyle puanlanmıştır: 1 = temsil etmiyor, 2 = geliştirilmesi gerekiyor ve 3 = temsil ediyor. Yöntemsel şeffaflığı sağlamak amacıyla standartlaştırılmış az örnekli yönlendirme çerçevesi kullanılmış; bağlam kaymasını önlemek için her modele yeni bir oturumda erişilmiş ve tüm değerlendirmelerde aynı yönlendirmeler uygulanmıştır. Verilerin analizinde Kapsam Geçerlik Oranı (KGO) ve Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ) hesaplanmış ve Wilcoxon İşaretli Sıralar Testi ile karşılaştırılmıştır. Bulgular, yapay zekâ ve insan uzman puanları arasında istatistiksel olarak anlamlı bir fark olmadığını göstermiştir. Bu sonuç, mevcut çalışmanın kapsamı ve koşulları içinde yapay zekâ ile insan değerlendirmeleri arasında belirli düzeyde benzerlik bulunduğunu düşündürmektedir. Modellerin kapalı kaynaklı yapısı nedeniyle ince ayar yapılamamış olsa da güvenirliği ve tekrarlanabilirliği desteklemek için standartlaştırılmış yönlendirme işlemleri uygulanmıştır. Bulgular, yöntemsel kontrol sağlandığında yapay zekâ araçlarının kapsam geçerliği değerlendirmesinde uzman yargısını tamamlayıcı bir rol üstlenebileceğini göstermektedir.

Anahtar sözcükler: Kapsam geçerliği, yapay zekâ, az örnekli yönlendirme, hibrit değerlendirme, eğitimde ölçme

1 Sorumlu Yazar: Dr., **E-posta:** gurdilhatice@gmail.com **ORCID ID:** 0000-0002-0079-3202 **ROR ID:** https://ror.org/00jga9g46
2 Öğr. Gör. Dr., **E-posta:** ozlemanadol@gazi.edu.tr **ORCID ID:** https://orcid.org/0000-0002-2881-9806
ROR ID: https://ror.org/054xkpr46
3 Dr., **E-posta:** berilsoguksu@gmail.com **ORCID ID:** 0009-0004-0870-4974 **ROR ID:** https://ror.org/00jga9g46
Gönderim Tarihi: 21.10.2025 **Kabul Tarihi**: 24.06.2026 **Yayım Tarihi**: 06.08.2026
Atıf: Gürdil, H., Anadol, H. Ö., & Soğuksu, Y. B. (2026). Kapsam geçerliği değerlendirmesinde yapay zekâ araçlarının kullanımı: İnsan uzmanlarla karşılaştırmalı bir çalışma. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 59*(2), 549-590. https://doi.org/10.30964/auebfd.1807153
Makaleye ilişkin tüm etik beyanlar, beyanlar sayfasında yer almaktadır (Sayfa: 587).

Bireylerin beceri, tutum ve davranışlarının değerlendirilmesi, psikoloji ve eğitim alanlarında hem araştırmalar hem de gerçek yaşam uygulamaları açısından büyük önem taşımaktadır. (Cohen ve Swerdlik, 2021). Doğru ve anlamlı değerlendirmelere ulaşabilmek için test ve ölçeklerin geliştirilmesi kritik bir role sahiptir. Bireylerin becerileri, tercihleri ve psikolojik özelliklerine dayalı olarak geliştirilen bu ölçme araçları; bilişsel, duyuşsal ve psikomotor becerilerin değerlendirilmesine temel oluşturur. Geçerlik ve güvenirlik, test geliştirme sürecinin iki temel kavramıdır ve bu ölçme araçlarının geliştirilmesine ilişkin karmaşık süreç boyunca dikkatle ele alınır (Nunnally ve Bernstein, 1994).

Kapsam geçerliğinin sağlanması, test geliştirme sürecinin temel unsurlarından biridir. Kapsam geçerliği, bir test ya da ölçeğin maddelerinin ölçülmesi amaçlanan kavramı ne ölçüde temsil ettiğini ifade eder. Başka bir deyişle, testin ölçülmek istenen yapının tüm yönlerini kapsaması ve önemli hiçbir boyutu dışarıda bırakmaması gerekir. Kapsam geçerliğinin yüksek olması, bireylerin test sonuçlarına dayalı olarak güvenilir çıkarımlar yapılabilmesi açısından önemlidir. Örneğin, iyi yapılandırılmış bir okuduğunu anlama testi; sözcük bilgisini anlama, metinleri analiz etme ve ana fikirleri belirleme gibi farklı okuma becerilerini değerlendiren maddeler içermelidir (Messick, 1989).

Kapsam geçerliğinin değerlendirilmesi, sıklıkla küçük bir örneklemin testi yanıtladığı ve testin etkililiğinin incelendiği pilot çalışmalar yoluyla gerçekleştirilmektedir. Araştırmacılar, bu pilot çalışmalar sırasında beklenen performansı göstermeyen ya da ölçülmek istenen yapıyı yeterince temsil etmeyen maddeleri belirleyebilir. Bu çalışmalardan elde edilen veriler, testin daha geniş bir gruba uygulanmadan önce önemli ölçüde geliştirilmesine katkı sağlar. Bununla birlikte, özellikle zamanın ya da kaynakların sınırlı olduğu durumlarda pilot çalışmalar her zaman uygulanabilir olmayabilir. Bu nedenle kapsam geçerliği tek ve sabit bir işlemle değerlendirilmez; bunun yerine testin amacına, geliştirme aşamasına ve mevcut kaynaklara bağlı olarak farklı yaklaşımlar kullanılabilir. Literatürde yaygın olarak kullanılan yaklaşımlardan biri, özellikle maddelerin daha geniş ölçekli uygulama öncesinde hedeflenen yapıyı yeterli düzeyde temsil edip etmediğinin incelenmek istendiği durumlarda, uzman yargısına dayalı değerlendirmedir. Bu tür durumlarda testin kapsam geçerliği uzman görüşleri aracılığıyla değerlendirilir (Haynes vd., 1995).

Uzman yargısına dayalı kapsam geçerliği çalışmalarında araştırmacılar, uzman görüşlerini özetlemek amacıyla hem nitel yargılardan hem de nicel göstergelerden yararlanabilir. Kapsam Geçerlik İndeksi (KGİ) ve Kapsam Geçerlik Oranı (KGO), bu süreçte kullanılan iki temel ölçüttür. Bu istatistiksel ölçütler, bir ölçek ya da testte yer alan bir maddenin ölçülmek istenen yapıyı doğru biçimde temsil edip etmediğini değerlendirmek amacıyla kullanılır. Uzmanlar, her bir maddeyi ilgili yapıyı ölçmede gerekli, yararlı ancak kritik olmayan ya da gereksiz olup olmadığı açısından değerlendirir. Bu değerlendirmelere dayalı olarak KGO hesaplanır ve her bir maddenin kapsam geçerliğini temsil eden sayısal bir değer elde edilir. Buna karşılık,

Madde veya Ölçek Düzeyinde hesaplanan Kapsam Geçerlik İndeksi (KGİ), uzmanların maddelere verdiği dereceleme puanlarını sistematik olarak bir araya getirerek testin ya da ölçeğin kapsam geçerliğine ilişkin daha bütüncül bir değerlendirme sunar (Lawshe, 1975). KGO ve KGİ hesaplamalarının geçerlik ve güvenirliğini etkileyen önemli faktörlerden biri, sürece dâhil edilen uzman sayısıdır. Genel olarak, daha fazla sayıda uzmanın yer aldığı değerlendirmeler daha güvenilir kabul edilmektedir. Bunun nedeni, daha geniş bir uzman görüşü çeşitliliğinin bireysel yanlılıkların etkisini azaltabilmesi ve elde edilen sonuçların incelenen yapıyı doğru biçimde temsil etme olasılığını artırabilmesidir. Geniş ve çeşitlilik gösteren bir uzman paneli, test maddelerinin temsil edici ve ilgili olmasını güvence altına alarak daha kapsamlı bir değerlendirme sunabilir (Mertler vd., 2025). Bununla birlikte, uygulamada yeterli sayıda nitelikli uzmandan oluşan bir grup oluşturmak güç olabilir. Sınırlı kaynaklar ve uzmanların erişilebilirlik düzeyleri nedeniyle yeterli sayıda değerlendiriciye ulaşmak çoğu zaman zorlaşmaktadır. Yeterli uzman görüşünün alınamaması, kapsam geçerliği değerlendirmelerinin geçerlik ve güvenirlik düzeylerini olumsuz etkileyebilir; bu durum yanlılığa ya da daha az doğru değerlendirmelere yol açabilir. Bu tür durumlarda, ölçek ve testlerin kapsam geçerliği standartlarını karşılamasını sağlamak araştırmacılar açısından önemli güçlükler doğurabilir. Bu güçlükler doğrultusunda araştırmacılar, test maddelerinin değerlendirilmesinde ölçeklenebilir ve veri temelli bir yöntem sunan yapay zekâ (YZ) dâhil olmak üzere alternatif yaklaşımları incelemeye başlamıştır. Gelişmiş algoritmalardan ve büyük veri kümelerinden yararlanan yapay zekâ sistemleri, makine öğrenmesi ve doğal dil işleme tekniklerini kullanarak test maddelerini değerlendirebilir, bu maddeleri önceden belirlenmiş ölçütlerle karşılaştırabilir ve KGO ile KGİ değerlerini hesaplayabilir. Yapay zekâ, insan uzmanların ayrıntılı ve bağlamsal yargılarını tam olarak karşılayamasa da uzman görüşünün sınırlı olduğu ya da uzman temininin güçleştiği durumlarda özellikle yararlı olabilir. Yapay zekânın geçerlik sürecine dâhil edilmesiyle araştırmacılar, insan uzman değerlendirmeleriyle desteklenmek üzere, sürecin verimliliğini ve doğruluğunu artırabilir. Bu çalışma birkaç açıdan önem taşımaktadır. İlk olarak, İngilizce okuduğunu anlama maddelerinin kapsam geçerliği değerlendirmesinde yapay zekâ teknolojilerinin kullanımını inceleyerek, yapay zekâ araçlarının özellikle verimlilik ve ek değerlendirme geri bildirimi sağlama açısından uzman temelli madde inceleme süreçlerini nasıl destekleyebileceğine ilişkin ön kanıt sunmaktadır. Geçerlik, testlerin ölçmeyi amaçladıkları yapıları ölçmesini sağlaması bakımından test geliştirme sürecinin kritik bir boyutudur. Ayrıca yapay zekâ teknikleri, özellikle uzman kaynaklarının sınırlı olduğu ya da uzmanlara ulaşılamadığı durumlarda testlerin geçerlenmesine yönelik yeni bir yaklaşım sunmaktadır.

Yapay zekâ araçlarının eğitimdeki potansiyeline ilişkin yürütülen çalışmalar incelendiğinde, Sihite vd. (2023) ChatGPT'nin akademik metinlere yönelik okuduğunu anlama soruları oluşturmadaki yeterliğini değerlendirmiş ve bu sürecin daha fazla geliştirilmesi gerektiğini vurgulamıştır. Ayrıca Morjaria vd. (2023), ChatGPT'nin kısa yanıtlı değerlendirme sorularındaki performansını incelemiş ve bu

tür değerlendirmelerin geçerliğiyle ilişkili olası risklere dikkat çekmiştir. Öte yandan, ölçek geliştirme bağlamında yürütülen araştırmalar da ChatGPT'nin farklı boyutlarının anlaşılmasına önemli katkılar sağlamıştır. Nemt-Allah vd. (2024), lisansüstü öğrenciler için "ChatGPT Kullanım Ölçeği"ni geliştirmiş; akademik yazma, görev desteği ve bağımlılık gibi boyutları incelemiştir. Lee ve Park (2024) ise bilgi aktarımı ve yaratıcılık üzerindeki etkisini değerlendirmek amacıyla ChatGPT Okuryazarlığı Ölçeği'ni geliştirmiştir. Bunun yanında, yapay zekâ araçlarının kişiselleştirilmiş öğrenme süreçleri üzerindeki etkisi de dikkat çekmektedir. Siegle (2023), bu araçların üstün yetenekli öğrencilerin öğrenme süreçlerini zenginleştirebileceğini ileri sürerken, Lin (2024) ChatGPT'nin yetişkin öğrenenlerde öz düzenlemeli öğrenmeyi destekleyebileceğini belirtmiştir. Dil öğrenimi alanında ise Kohnke vd. (2023), ChatGPT'nin dil öğretimi ve öğrenimindeki fırsatlarını, tartışmalı yönlerini ve dezavantajlarını ele almış; Saraswati vd. (2023) ise AI Replika'nın öğrencilerin ana dili konuşurlarıyla pratik yapmalarına yardımcı olabileceğini, ancak kültürel anlamları kavrama konusunda sınırlı kaldığını ortaya koymuştur. Bunun yanı sıra, yapay zekâ araçlarını farklı yapay zekâ modelleri ve insan performansıyla karşılaştıran çalışmalarda da önemli bulgular ortaya konmuştur. Kim vd. (2024), ikinci dilde yazma görevlerinde ChatGPT'nin insan puanlarıyla karşılaştırıldığında doğruluk ve güvenirlik açısından sınırlılıklar gösterdiğini ortaya koymuştur. Hayawi vd. (2024), büyük dil modellerinin (BARD, GPT) insan yazısından ayırt edilememe kapasitesini incelemiş; ul haq Akhoon vd. (2024) ise ChatGPT ve Google Gemini'nin akademik içerik üretimindeki etkililiğini ve buna ilişkin etik geçerlik sorunlarını ele almıştır. Wachinger vd. (2025), nitel veri analizinde ChatGPT'nin performansının büyük ölçüde deneyimli bir araştırmacının performansıyla örtüştüğünü bulmuştur. Benzer şekilde Gürdil vd. (2025), ChatGPT'nin veri üretimindeki etkililiğini incelerken madde parametrelerinde bazı tutarsızlıklar gözlemlemiş, ancak insan verisine benzer sonuçlar elde etmiştir. Öte yandan Tenhundfeld (2023), ChatGPT'nin insan ve yapay zekâ etkileşimi bağlamında daha fazla geliştirilmesi gerektiğini vurgulamış; Onal ve Kulavuz-Onal (2024) ise ChatGPT'nin yükseköğretimde değerlendirme görevlerindeki yüksek performansını inceleyerek bu araçların insan uzmanlığını tamamlayıcı bir rol üstlendiğini belirtmiştir.

Daha güncel çalışmalar, üretken yapay zekânın eğitimde değerlendirme süreçlerindeki rolünü daha temkinli bir bakış açısıyla ele almıştır. Xia vd. (2024), yükseköğretimde değerlendirme süreçlerinde üretken yapay zekâ kullanımına ilişkin ampirik çalışmaları incelemiş ve bu araçların değerlendirme uygulamalarını değiştirebileceğini, ancak kullanımlarının değerlendirme tasarımı, öğretmen hazırlığı ve kurumsal politikalar açısından dikkatle ele alınması gerektiğini belirtmiştir. Benzer şekilde Weng vd. (2024), üretken yapay zekânın öğrenci öğrenmesinin nasıl değerlendirilmesi gerektiğine ve değerlendirme görevlerinin nasıl yeniden tasarlanabileceğine ilişkin yeni tartışmaları gündeme getirdiğini vurgulamıştır. Kizilcec vd. (2024), eğitimcilerin ve öğrencilerin bakış açılarını karşılaştırmış ve üretken yapay zekânın özellikle özgünlük, akademik dürüstlük ve üst düzey düşünme becerileriyle ilişkili olarak değerlendirme kalitesi üzerinde önemli bir etkiye sahip

olduğunun düşünüldüğünü rapor etmiştir. Yakın tarihli bir çalışmada Zacharis ve Papadakis (2025) de yapay zekâ temelli puanlamanın yüksek riskli değerlendirme bağlamlarında insan değerlendirmesiyle birbirinin yerine kullanılabilir olarak görülmemesi gerektiğini ortaya koymuştur. Birlikte değerlendirildiğinde, bu çalışmalar yapay zekâ araçlarının değerlendirme süreçlerini destekleyebileceğini; ancak bu kullanımın insan uzmanlığı ve geçerlikle ilgili değerlendirmeler doğrultusunda yönlendirilmesi gerektiğini göstermektedir.

Çalışmalar incelendiğinde, yapay zekâ temelli araçların eğitimdeki potansiyelinin geçerlik, güvenirlik ve kullanıcı deneyimi açısından ele alındığı ve insan performansıyla karşılaştırıldığı görülmektedir. Bununla birlikte, ilgili literatürde yapay zekânın doğrudan kapsam geçerliği süreçlerindeki rolünü ve etkililiğini inceleyen bir araştırmaya rastlanmamıştır. Kapsam geçerliği, bir ölçme aracının ölçmeyi amaçladığı alanı ne derece doğru biçimde kapsadığını değerlendiren kritik bir unsurdur. Wynd vd. (2003), Kapsam Geçerlik İndeksi'nde uzmanlar arası düşük uyumun ölçek revizyonu açısından önemli olduğunu göstermiştir. Bu konu, kapsam geçerliğini değerlendirmeye yönelik yenilikçi yaklaşımların öneminin vurgulandığı çalışmalarda sıklıkla ele alınmaktadır (Ayre ve Scally, 2014; Hinkin ve Tracey, 1999; Newman vd., 2013). Ancak literatür incelendiğinde, yapay zekâ araçlarının kapsam geçerliğinin belirlenmesindeki potansiyelini özel olarak inceleyen bir araştırmanın bulunmadığı görülmektedir. Bu boşluk, kapsam geçerliği değerlendirmesinde yapay zekânın etkililiğinin araştırılması gerekliliğini ortaya koymaktadır.

Araştırmanın Amacı

Bu çalışma, B1 düzeyindeki bir İngilizce sınavında yer alan okuma becerisi maddelerinin kapsam geçerliğini değerlendirmede yapay zekânın (YZ) potansiyelini incelemeyi amaçlamaktadır. Daha özel olarak çalışmada, dört farklı yapay zekâ aracı ve dört insan uzman tarafından verilen puanlara dayalı olarak yapay zekâ araçları arasındaki uyum düzeyinin incelenmesi, Kapsam Geçerlik Oranlarının (KGO) ve Madde Düzeyinde Kapsam Geçerlik İndekslerinin (M-KGİ) hesaplanması amaçlanmıştır. Bu doğrultuda aşağıdaki araştırma soruları oluşturulmuştur:

1. Maddelere ilişkin tüm değerlendiriciler tarafından verilen puanlar arasında istatistiksel olarak anlamlı bir farklılık var mıdır?
2. Tüm değerlendiriciler arasındaki puanlamalar ne düzeyde tutarlıdır?
3. Yapay zekâ değerlendiricilerinden elde edilen KGO değerleri, insan değerlendiricilerden elde edilen KGO değerlerinden anlamlı düzeyde farklılaşmakta mıdır?
4. Yapay zekâ değerlendiricilerinden elde edilen M-KGİ değerleri, insan değerlendiricilerden elde edilen M-KGİ değerlerinden anlamlı düzeyde farklılaşmakta mıdır?

Araştırmanın Önemi

Bu çalışma, yapay zekânın zaman ve kaynak sınırlılıklarını azaltarak uzman temelli kapsam geçerliği değerlendirmelerine ölçeklenebilir bir alternatif sunup sunamayacağını incelemektedir. Yapay zekânın etkili olduğunun ortaya konması durumunda, yüksek nitelikli test geçerleme süreçlerinin erişilebilirliğini artıran, kaynak gereksinimlerini azaltan ve maliyet açısından daha verimli bir alternatif sağlayabileceği düşünülmektedir. Çalışmada ayrıca, yapay zekâ tarafından üretilen kapsam geçerliği değerlendirmelerinin insan uzman tarafından yapılan değerlendirmelerle karşılaştırılması amaçlanmaktadır. Bu karşılaştırma, yapay zekâ araçlarının eğitimde değerlendirme alanındaki etkililiğini ve güvenirliğini değerlendirmek, güçlü ve sınırlı yönlerini belirlemek ve gelecekteki test geliştirme uygulamalarına yön vermek açısından önem taşımaktadır.

Bunun yanı sıra, çalışmanın bulgularının dil öğretim elemanları ve madde yazarları için uygulamaya dönük çıkarımlar sunması beklenmektedir. Etkili yapay zekâ temelli kapsam geçerliği değerlendirmeleri, daha doğru ve temsil gücü yüksek test tasarımlarının geliştirilmesine sağlama potansiyeline sahiptir. Bu durum, öğrencilerin dil becerilerine ilişkin daha güvenilir ölçümler elde edilmesini kolaylaştırarak dolaylı olarak öğrencilere de yarar sağlayabilir. Son olarak, bu çalışma kapsam geçerliği değerlendirmesi gibi belirli bir uygulama alanına odaklanarak eğitimde yapay zekâ entegrasyonuna ilişkin gelişen araştırma alanına katkıda bulunmaktadır. Yapay zekânın bu alana başarılı biçimde entegre edilmesi, eğitim ortamlarında yapay zekânın daha ileri ve yenilikçi kullanımlarının önünü açabilir; testlerin tasarlanma, uygulanma ve geçerlenme süreçlerini dönüştürebilir. Genel olarak, çalışmanın bulguları test geliştirme uygulamaları, kaynak kullanımı ve eğitim bağlamlarında yapay zekânın rolü üzerinde önemli etkiler yaratabilir. Ayrıca sonuçlar, teknoloji ve değerlendirmenin kesişim noktasına ilişkin değerli bilgiler sunarak alandaki gelecekteki araştırma ve uygulamalara yön verebilir.

Yöntem

Araştırmanın Deseni

Bu çalışma, karşılaştırmalı betimsel bir araştırma olarak tasarlanmıştır. Çalışmanın temel amacı, B1 düzeyindeki İngilizce okuduğunu anlama maddelerine ilişkin insan uzmanlar ve yapay zekâ araçları tarafından yapılan kapsam geçerliği değerlendirmelerini karşılaştırmaktır. Çalışma, bir müdahalenin test edilmesinden ziyade değerlendirici puanlarının incelenmesine ve karşılaştırılmasına odaklandığı için betimsel ve karşılaştırmalı araştırma yaklaşımı uygun görülmüştür.

Testin Geliştirilmesi ve Uygulanması

Bu çalışmada, araştırmanın amacı doğrultusunda B1 düzeyindeki İngilizce okuma metinleri ve sorularından oluşan 25 maddelik bir test kullanılmıştır. Testin hazırlık aşamasında maddeler, dört uzman İngilizce öğretim elemanı tarafından geliştirilmiştir. Testin geçerlik ve güvenirliğini sağlamak amacıyla sorular, ölçme ve

değerlendirme birimi koordinatörü tarafından incelenmiş ve gerekli düzeltmeler yapılmıştır. Maddeler, CEFR temelli B1 okuma alt becerileriyle uyumlu bir madde belirtke çerçevesi kapsamında geliştirilmiştir. Bu çerçevede hedeflenen öğrenme kazanımı, ilgili okuma becerisi alanı, maddenin B1 düzeyine uygunluğu, metin ile madde arasındaki uyum, çeldiricilerin inandırıcılığı ve işlevselliği ile maddenin dilsel açıklığı dikkate alınmıştır. Aynı ölçütler, kapsam geçerliği incelemesi sırasında kullanılan uzman değerlendirme formuna da yansıtılmıştır. Uzmanlardan her bir maddenin hedeflenen okuma becerisini temsil edip etmediğini, B1 yeterlik düzeyine uygun olup olmadığını, okuma metniyle doğrudan ilişkili olup olmadığını, işlevsel çeldiriciler içerip içermediğini ve dilsel açıdan açık ve doğru olup olmadığını değerlendirmeleri istenmiştir. Testin kazanımları, Diller İçin Avrupa Ortak Başvuru Metni (CEFR) ile uyumludur. Bu noktada yanıtlayıcıların B1 düzeyine karşılık gelen okuma becerilerine sahip olmaları beklenmektedir. Bu durum, onların farklı metin türlerinin (gazete makaleleri, denemeler, hikâyeler vb.) genel anlamını anlayabilmelerini, metinlerdeki ana fikirleri ve önemli ayrıntıları belirleyebilmelerini ve açıkça ifade edilmeyen bilgilere ilişkin çıkarım yapabilmelerini gerektirir. Ayrıca B1 düzeyine uygun sözcük bilgisine sahip olmaları, metin içindeki sözcüklerin anlamlarını bağlamdan çıkarabilmeleri, metnin düzenini ve paragraflar arasındaki bağlantıları anlayabilmeleri beklenmektedir.

B1 düzeyi, orta düzeyde İngilizce yeterliğini temsil etmesi ve geniş bir hedef kitleye hitap etmesi nedeniyle bu çalışma için seçilmiştir. Bu düzeydeki bireyler, günlük yaşamda karşılaşabilecekleri çeşitli metinleri anlayabilir ve İngilizceyi akademik ya da mesleki amaçlarla kullanabilir. Ayrıca B1 düzeyi, diğer düzeylerle karşılaştırıldığında daha homojen bir grup oluşturduğundan, çalışma sonuçlarının daha güvenilir olmasına katkı sağlamaktadır.

Değerlendiriciler: İnsan Uzmanlar ve Yapay Zekâ Araçları

Test geçerlik süreci tamamlandıktan sonra testte yer alan 25 madde, 4 insan uzman ve 4 yapay zekâ aracı tarafından değerlendirilmiştir. İnsan uzmanlar, İngilizce Öğretimi alanında uzman, en az 5 yıllık deneyime sahip ve CEFR değerlendirme süreçlerine hâkim kişilerden oluşmaktadır. Bu ölçütler, değerlendiricilerin ilgili alan uzmanlığına, dil değerlendirmesi konusunda uygulama deneyimine ve CEFR temelli okuma becerisi tanımlayıcılarına aşinalığa sahip olmasını sağlamak amacıyla kullanılmıştır. Maddeler B1 düzeyinde bir İngilizce okuduğunu anlama testi için geliştirildiğinden, maddelerin ilgililiğini ve temsil ediciliğini değerlendirmek üzere CEFR ile uyumlu değerlendirme deneyimine sahip uzmanların uygun olduğu düşünülmüştür.

Çalışmada kullanılan yapay zekâ araçları GPT-4o, Gemini 2.0, Replika ve Chatfuel'dir. Bu araçlar, metin işleme ve yanıt üretme kapasitesine sahip, erişilebilir farklı yapay zekâ temelli sistem türlerini içerecek şekilde seçilmiştir. GPT-4o ve Gemini 2.0 gelişmiş büyük dil modelleri olarak çalışmaya dâhil edilirken, Replika ve Chatfuel sohbet robotu temelli sistemler olarak dâhil edilmiştir. Bu seçim, farklı yapay zekâ destekli yanıt ortamlarından elde edilen değerlendirmelerin karşılaştırılmasına

olanak sağlamıştır. OpenAI tarafından geliştirilen ChatGPT-4o, geniş dil işleme kapasitesi sayesinde karmaşık ifadeleri doğal biçimde analiz etme ve yorumlama konusunda oldukça başarılıdır. Güncel yapay zekâ teknolojileri arasında öne çıkan bir model olan ChatGPT-4o, farklı disiplinlerden metinleri analiz etmek amacıyla kullanılabilmektedir. GPT-4o'nun temel metin işleme özellikleri ücretsiz olarak kullanılabilirken, ses ve görüntü işleme gibi bazı gelişmiş özellikleri abonelik gerektirmektedir (Ofgang, 2024; OpenAI, 2023).

Gemini 2.0, Google DeepMind tarafından geliştirilen çok modlu bir yapay zekâ modelidir. Bu model, yalnızca metin analizinden değil, aynı zamanda görsel ve bağlamsal bilgilerden de yararlanarak değerlendirme sürecini güçlendirmektedir. Gemini 2.0'ın metin temelli görevlerdeki güçlü performansı, bu çalışma kapsamında kullanılmasında önemli bir gerekçe oluşturmuştur (Hassabis ve Kavukcuoglu, 2024). Replika ise gelişmiş doğal dil işleme (NLP) teknolojilerini kullanan ve Üretken Önceden Eğitilmiş Dönüştürücü 3 (GPT-3) modeli üzerine yapılandırılmış sosyal bir sohbet robotudur. Teknolojik altyapısı, kullanıcı konuşmalarından elde edilen geniş bir veri kümesine dayalı dil yanıtları üretmesine olanak tanımakta; böylece bağlamsal açıdan zengin ve nüanslı etkileşimler sağlayabilmektedir (Pentina vd., 2023).

Bu araçlara ek olarak, Chatfuel de bu çalışma için seçilmiştir. Chatfuel, metni yorumlamak ve yanıt üretmek amacıyla doğal dil işleme (NLP) tekniklerini kullanan yapay zekâ temelli bir sohbet robotudur. Genellikle müşteri hizmetleri ve otomatik mesajlaşma süreçlerinde kullanılmakla birlikte, geri bildirim sağlama ve etkileşimli öğrenme ortamları oluşturma amacıyla da kullanılmaktadır. Ahmed ve Hussein (2020), çalışmalarında Kürtçe konuşan bireylerin insan temsilcilerle doğrudan iletişime geçmek yerine çevrim içi konuşmalar yoluyla yanıt bulmalarına yardımcı olmak için Chatfuel platformunu kullanan bir sohbet robotu tasarımı sunmuştur. Bu çalışmada değerlendirme süreçlerinde yapay zekâ destekli geri bildirim mekanizmalarının kullanımı incelendiğinden, Chatfuel bu nedenle seçilmiştir.

Etik Kurul Kararı

Bu çalışma, Gazi Üniversitesi Rektörlüğü Etik Komisyonunun 23.09.2025 tarihli ve 14 sayılı toplantısında incelenmiş; 09.10.2025 tarihli ve E.1352469 sayılı belge ile etik açıdan uygun bulunmuştur. Araştırma kod numarası 2025-1649'dur.

Araştırmanın Bilimsel Niteliği

Yapay Zekâ Yönlendirme Süreci ve Bağlam Kontrolü

Değerlendirmelerin karşılaştırılabilir ve tekrarlanabilir olmasını sağlamak amacıyla tüm yapay zekâ sistemlerine bağımsız olarak yeni oturumlarda erişilmiştir. Her değerlendirme öncesinde, önceki yönlendirmelerin ya da yanıtların sonraki çıktılar üzerinde istenmeyen etkisini ifade eden bağlam kaymasını önlemek için önceki konuşma geçmişleri temizlenmiştir. Tüm modellerde standartlaştırılmış az örnekli yönlendirme çerçevesi benimsenmiştir. Puanlama mantığını oluşturmak amacıyla her bir yapay zekâ sistemine öncelikle doğru etiketleri içeren iki örnek

madde sunulmuştur (geçerli ve geçersiz). Bu kısa kalibrasyon adımının ardından, B1 düzeyindeki okuduğunu anlama testine ait 25 madde aynı girdi yapısı kullanılarak sıralı biçimde girilmiştir.

Temel yönlendirme şu şekilde yapılandırılmıştır: "Bu maddenin hedeflenen okuma becerisiyle ilgililiğini 1 (ilgili değil) ile 3 (yüksek düzeyde ilgili) arasında puanlayınız. Yalnızca sayısal bir puan veriniz." Bu yaklaşım, yapay zekâ sistemlerinin insan değerlendiricilerin yargı ölçütlerini taklit etmesine olanak sağlamış ve farklı uygulamalar arasında öznel değişkenlik oluşmasını önlemeyi amaçlamıştır. Tekrarlanan oturumlar boyunca elde edilen çıktıların tutarlılığı, içsel kararlılığı doğrulamak amacıyla manuel olarak kontrol edilmiştir.

Güvenirlik ve Tutarlılık Kontrolü

Yapay zekâ değerlendirmelerinin tutarlılığını belirlemek amacıyla aynı standartlaştırılmış yönlendirmeler kullanılarak iki farklı zamanda tekrarlı denemeler gerçekleştirilmiştir. Ham puanlarda gözlenen küçük farklılıkların istatistiksel olarak anlamlı olmadığı belirlenmiştir. Sekiz değerlendirici arasındaki değerlendiriciler arası uyum, Fleiss Kappa kullanılarak hesaplanmış ve .431 olarak bulunmuştur. Bu değer, değerlendiriciler arasında orta düzeyde bir uyum olduğunu göstermektedir. Değerlendirici grupları arasındaki farklılıklar Kruskal-Wallis H Testi ile analiz edilmiş ve anlamlı bir farklılık gözlenmemiştir ($p > .05$).

Verilerin Analizi

Bu çalışmada, insan ve yapay zekâ değerlendiricileri her bir test maddesini, maddelerin hedeflenen içeriği ne ölçüde yansıttığını belirlemek amacıyla 1 ile 3 arasında puanlayarak değerlendirmiştir. Puanlama sistemi açık bir biçimde yapılandırılmıştır: "1" maddenin içeriği yeterince temsil etmediğini, "2" maddenin geliştirilmesi gerektiğini ve "3" maddenin içeriği etkili biçimde temsil ettiğini göstermektedir. Tüm maddeler değerlendirildikten sonra puanlar bir tablo hâline getirilmiştir. Bu puanlar daha sonra her bir madde için Kapsam Geçerlik Oranı (KGO) ve Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ) hesaplamalarında kullanılmıştır. Değerlendiriciler tarafından verilen puanlar arasında anlamlı farklılıklar olup olmadığını belirlemek amacıyla Kruskal-Wallis H testi kullanılmıştır. Bu test, normal dağılım göstermeyen sıralı verilerin analizinde kullanılan Tek Yönlü ANOVA'nın parametrik olmayan bir alternatifidir (MacFarland ve Yates, 2016). Analizler, R programındaki "stats" paketi kullanılarak gerçekleştirilmiştir (R Core Team, 2024).

Her bir madde, 4 insan uzman ve 4 yapay zekâ aracı olmak üzere toplam 8 değerlendirici tarafından incelenmiştir. İnsan ve yapay zekâ değerlendiricileri arasındaki ve yapay zekâ araçlarının kendi aralarındaki uyum düzeyini değerlendirmek amacıyla Fleiss Kappa istatistiği uygulanmıştır. Fleiss Kappa, birden fazla değerlendiricinin yer aldığı kategorik veriler için uygun bir uyum ölçüsüdür (Fleiss, 1971; Hallgren, 2012). Hesaplamalar, R programındaki "irr" paketi kullanılarak gerçekleştirilmiştir (Matthias vd., 2019). Fleiss Kappa değerleri -1 ile +1

arasında değişmekte olup, +1 mükemmel uyumu göstermektedir. Ayrıca test maddelerinin kapsam geçerliği iki ölçüt kullanılarak değerlendirilmiştir: Kapsam Geçerlik Oranı (KGO) ve Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ). Bu ölçütler, hem insan hem de yapay zekâ değerlendiricileri tarafından verilen puanlara dayalı olarak hesaplanmıştır.

Kapsam Geçerlik Oranı (KGO)

Lawshe (1975) tarafından geliştirilen KGO, test maddelerinin hedeflenen içeriği yeterli düzeyde temsil edip etmediğini değerlendirmek amacıyla yaygın olarak kullanılan bir yöntemdir. Uzmanlar her bir maddeyi "gerekli", "yararlı ancak gerekli değil" ve "gerekli değil" biçiminde sınıflandırır. Belirli sayıda değerlendirici tarafından "gerekli" olarak puanlanan maddeler korunurken, diğer maddeler çıkarılır (Almanasreh vd., 2019). KGO aşağıdaki formül kullanılarak hesaplanır:

$$\boldsymbol{KGO = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}} \qquad \mathbf{1}$$

Burada (n_e), maddeyi "gerekli" olarak değerlendiren değerlendirici sayısını; (N) ise toplam değerlendirici sayısını ifade etmektedir. KGO değeri -1 ile +1 arasında değişmektedir. "+1" mükemmel uyumu, "-1" tam uyumsuzluğu, "0" ise değerlendiricilerin yarısının maddeyi "gerekli" olarak değerlendirdiğini göstermektedir. Lawshe'nin (1975) kesme noktalarına göre, KGO değeri en az 0.99 olan maddeler testin nihai formuna dâhil edilmiştir.

Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ)

Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ), bireysel maddelerin ilgililiğini değerlendirmeye odaklanmaktadır. Literatürde M-KGİ sıklıkla 4'lü bir uygunluk ölçeği kullanılarak hesaplanmasına rağmen (Davis, 1992), bu çalışmada hem insan hem de yapay zekâ değerlendiricilerine uygulanan değerlendirme süreciyle uyumlu olarak 3'lü bir derecelendirme ölçeği kullanılmıştır. Bu ölçekte "1" maddenin hedeflenen içeriği yeterince temsil etmediğini, "2" maddenin geliştirilmesi gerektiğini ve "3" maddenin hedeflenen içeriği etkili biçimde temsil ettiğini göstermektedir. M-KGİ hesaplamasında "3" puanları ilgili kabul edilirken, "1" ve "2" puanları tam olarak ilgili kabul edilmemiştir. Bu nedenle her bir maddeye ilişkin M-KGİ değeri, maddeyi ilgili olarak değerlendiren değerlendiricilerin oranı şeklinde hesaplanmıştır. Lynn (1986), daha küçük panellerde (örneğin 3-5 değerlendirici) 3'lü ya da 5'li ölçeklerin kullanılabileceğini belirtmiştir. M-KGİ aşağıdaki formül kullanılarak hesaplanır:

$$M - KG\dot{I} = \frac{\text{Maddeyi ilgili olarak değerlendiren uzman sayısı}}{\text{Toplam uzman sayısı}} \qquad 2$$

Değerlendirici sayısının 6'dan az olduğu durumlarda Lynn (1986), bir maddenin geçerli kabul edilebilmesi için M-KGİ değerinin 1 olması gerektiğini önermiştir. Benzer şekilde Polit vd. (2007), panelde 3 ya da daha az uzman bulunduğunda değerlendiriciler arasında %100 uyum olması gerektiğini belirtmiştir. Bu çalışmada yalnızca M-KGİ değeri 1 olan maddeler nihai forma dâhil edilmiştir. M-KGİ değeri 1'in altında olan maddeler ise revizyon için işaretlenmiştir.

Bulgular

Sekiz farklı değerlendiriciden (İnsan1, İnsan2, İnsan3, İnsan4, GPT-4o, Gemini 2.0, Replika ve Chatfuel) elde edilen puanların sıra ortalamaları arasında anlamlı bir farklılık olup olmadığını belirlemek amacıyla gerçekleştirilen Kruskal-Wallis testi sonuçları ve etki büyüklüğü (ε^2) değeri (Ek 1) Tablo 1'de sunulmuştur.

Tablo 1

Değerlendiricilere İlişkin Kruskal-Wallis Testi Sonuçları ve Etki Büyüklüğü

Değerlendiriciler	Sıra Ortalaması	$\chi 2$	Sd	*p*	Etki Büyüklüğü
8 değerlendirici*	13.00	10.18	7	0.180	0.05

***İnsan 1, İnsan 2, İnsan 3, İnsan 4, GPT-4o, Gemini 2.0, Chatfuel ve Replika*

Tablo 1 incelendiğinde, Kruskal-Wallis testi sonuçları değerlendiriciler arasında anlamlı bir farklılık olmadığını göstermektedir ($\chi^2 = 10.182$, sd = 7, $p > 0.05$, $\varepsilon^2 = 0.051$). Bu sonuç, tüm değerlendiricilerin test maddelerini benzer biçimde puanladığını düşündürmektedir. Bununla birlikte hesaplanan etki büyüklüğü, değerlendiriciler arasındaki farkın oldukça düşük olduğunu göstermekte ve aralarında istatistiksel olarak anlamlı bir farklılık olmadığı sonucunu desteklemektedir. Tüm değerlendiriciler arasındaki tutarlılığı incelemek amacıyla gerçekleştirilen Fleiss Kappa analizinin sonuçları Tablo 2'de sunulmuştur.

Tablo 2

Değerlendiriciler Arasındaki Tutarlılık Analizi Sonuçları

Değerlendiriciler	Kappa değeri	*p*
8 değerlendirici*	0.43	$p < .001$

**İnsan 1, İnsan 2, İnsan 3, İnsan 4, GPT-4o, Gemini 2.0, Chatfuel ve Replika*

Tablo 2'de görüldüğü üzere, Fleiss Kappa değerinin 0.431 olması değerlendiriciler arasında orta düzeyde bir uyum bulunduğunu göstermektedir. Bu sonuç, bazı maddelerde görüş ayrılıkları bulunmakla birlikte değerlendiriciler arasında belirli düzeyde bir tutarlılık olduğunu düşündürmektedir. Başka bir ifadeyle, değerlendiriciler arasında bazı tutarsızlıklar olsa da genel olarak puanlamaların tutarlı bir eğilim gösterdiği söylenebilir. Ayrıca tüm değerlendiriciler tarafından en sık verilen puan (mod) 3 olarak hesaplanmıştır. Bu bulgular, değerlendiricilerin tutarlı biçimde puanlama yaptığını ve hem insan hem de yapay zekâ değerlendiricilerinin değerlendirme süreçlerinde benzer performans gösterdiğini ortaya koymaktadır. İnsan ve yapay zekâ değerlendiricileri tarafından verilen puanlara dayalı olarak Kapsam

Geçerlik Oranı (KGO) ve Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ) değerleri hesaplanmış ve elde edilen değerler Tablo 3'te sunulmuştur.

Tablo 3

İnsan ve Yapay Zekâ (YZ) Puanlarına İlişkin KGO ve M-KGİ Değerleri

	KGO_YZ	KGO_İnsan	M-KGİ_YZ	M-KGİ_İnsan
1	0.50	0.50	0.75	0.75
2	0.50	-1.00	0.75	0.00
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	0.50	1.00	0.75
6	-0.50	-0.50	0.25	0.25
7	1.00	1.00	1.00	1.00
8	0.50	1.00	0.75	1.00
9	0.00	-1.00	0.50	0.00
10	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00
13	0.50	-0.50	0.75	0.25
14	1.00	1.00	1.00	1.00
15	1.00	0.00	1.00	0.50
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00

İnsan ve yapay zekâ değerlendiricilerine ilişkin KGO ve M-KGİ değerleri incelendiğinde, insan değerlendiricilerin 18 maddeyi, yapay zekâ değerlendiricilerinin ise 19 maddeyi geçerli kabul ettiği görülmektedir (bkz. Tablo 3). Bu bulgular, her iki değerlendirici grubunun kapsam geçerliği açısından benzer performans gösterdiğini; yapay zekâ değerlendiricilerinin ise bir maddeyi daha yüksek oranda onayladığını göstermektedir. Bununla birlikte, hem insan hem de yapay zekâ değerlendiricileri 1, 2, 6, 9 ve 13. maddelerin nihai forma alınmaması yönünde karar vermiştir. Bu maddeler, her iki grup tarafından kapsam geçerliği açısından yetersiz görülmüş ve onaylanmamıştır. Sonuçlar, insan ve yapay zekâ kararlarının bu maddeler üzerinde paralel örüntüler sergilediğini ve elenmesi gereken maddelerin tespitinde tam bir uzlaşı sağlandığını göstermektedir. Benzer şekilde 3, 4, 7, 10, 11, 12, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24 ve 25. maddeler için her iki grup da maddelerin nihai forma alınması yönünde karar vermiştir. Bu maddelerde hem insan hem de yapay zekâ değerlendiricilerinden yüksek düzeyde onay alınmış ve tutarlı, paralel sonuçlara

ulaşılmıştır. Bu durum, yapay zekânın insan benzeri değerlendirmeler yapabildiği ve insan değerlendirmeleriyle tutarlı sonuçlar üretebildiği görüşünü desteklemektedir.

Bununla birlikte, 8. madde (Ek 2) değerlendiriciler arasında bir görüş ayrılığı olduğunu göstermektedir. İnsan değerlendiriciler bu maddeyi kapsam geçerliği açısından uygun bulurken, yapay zekâ değerlendiricileri maddenin gözden geçirilmesini ya da çıkarılmasını önermiştir. Bu farklılık, yapay zekâ değerlendirmelerinin metnin bağlamını tam olarak kavrayamamasından ve temel unsurları yeterince analiz edememesinden kaynaklanmış olabilir. Yapay zekâ değerlendiricilerinin bağlamsal derinlikten ziyade sözdizimsel ve sözcüksel örüntülere dayanma eğiliminde olması, 8. madde için revizyon önermelerini açıklayabilir. Buna karşılık insan değerlendiriciler, metnin toplumsal etkisini ve bağlamını dikkate alarak daha derinlemesine bir analiz yapmış olabilir. İnsan değerlendiriciler kültürel ve toplumsal bağlamı göz önünde bulundururken, yapay zekâ değerlendiricileri bu yönleri gözden kaçırmış olabilir.

Buna ek olarak, 5. ve 15. maddeler (Ek 2) için yapay zekâ değerlendiricileri her iki maddeyi de kapsam geçerliği açısından uygun bulurken, insan değerlendiriciler daha temkinli bir yaklaşım benimsemiş ve bu maddeler için revizyon ya da çıkarma önerisinde bulunmayı tercih etmiştir. 5. maddenin değerlendirilmesinde yapay zekâ değerlendiricileri, teknik ölçütlere dayalı olarak maddeyi genel olarak yeterli bulmuş ve nihai forma alınmasını önermiştir. Buna karşılık insan değerlendiriciler; bağlam, dilsel duyarlılık ve içerik niteliği gibi ölçütleri devreye sokarak daha eleştirel bir süzgeç kullanmışlardır. Bu farklılık birkaç etmenden kaynaklanmış olabilir. İlk olarak, insan değerlendiriciler metnin bağlamıyla uyumsuzluğa karşı daha duyarlı davranmış olabilirken, yapay zekâ değerlendiricilerinin standartlara dayalı değerlendirme süreci bu tür nüansları gözden kaçırmış olabilir. İnsan değerlendiriciler ayrıca metnin dilsel niteliğini, ifadelerin açıklığını ve hedef kitleye uygunluğunu daha ayrıntılı biçimde incelemiş olabilir. Bunun yanı sıra, metnin pedagojik amaçlarla uyumu ve öğrencilerin öğrenme çıktıları açısından olası eksiklikleri de insan değerlendiricilerin kararlarını etkilemiş olabilir. Öte yandan yapay zekâ değerlendiricileri metni öncelikle teknik yeterlik açısından değerlendirmiş; insan değerlendiriciler ise içerik niteliğine, bağlama ve dilsel duyarlılığa daha fazla odaklanmış olabilir. Bu farklılık, yapay zekâ sistemlerinin belirli dilsel ve pedagojik özellikleri gözden kaçırma eğiliminden kaynaklanmış olabilir.

Benzer şekilde, 15. maddenin değerlendirilmesinde yapay zekâ değerlendiricileri sorunun nihai forma alınmasını önerirken, insan değerlendiriciler "çıkar" ya da "revize et" yönünde karar vermiştir. Bu farklılığın birkaç olası nedeni düşünülebilir. İnsan değerlendiriciler, özellikle sorunun bağlamı ve ilgililiği konusunda daha eleştirel bir yaklaşım benimsemiş olabilir. Örneğin, bu soruda festivalin çevresel etkisi ve yerel üreticilerle bağlantısına ilişkin ifadeler daha karmaşık ve çok yönlü bir yorum gerektirmektedir. Yapay zekâ değerlendiricileri, metnin teknik yönlerine odaklanarak belirli anahtar sözcüklere yoğunlaşmış olabilirken, insan değerlendiriciler metnin bütünlüğünü ve kültürel bağlamını daha

geniş bir perspektiften ele almış olabilir. Sonuç olarak, insan değerlendiriciler tarafından yapılan eleştiriler, sorunun ifade ediliş biçiminin ve yanıttaki anlam derinliğinin daha ayrıntılı biçimde değerlendirilmesini gerektirmiş olabilir.

Değerlendirici Puanlarına Göre Hesaplanan KGO ve M-KGİ Değerleri Arasındaki Farklılıklar

İnsan ve yapay zekâ (YZ) değerlendiricilerinin puanlarına dayalı olarak maddeler için hesaplanan Kapsam Geçerlik Oranı (KGO) değerleri arasındaki farklılıklar Wilcoxon İşaretli Sıralar Testi kullanılarak analiz edilmiştir. Analiz sonucunda, insan ve yapay zekâ değerlendiricileri arasındaki farklılıkların istatistiksel olarak anlamlı olmadığı belirlenmiştir ($W = 335.5$, $p = 0.5708$). Bu sonuç, iki grup arasında merkezi eğilimde (medyan) anlamlı bir değişim olmadığını ve dolayısıyla KGO değerlerinin benzer olduğunu göstermektedir.

KGO analizinden elde edilen sonuçlar, maddelere ilişkin kapsam geçerliği kararlarında genel bir uyum olduğunu gösterse de her bir maddenin uygunluğunu daha ayrıntılı bir ölçüt olan Madde Düzeyinde Kapsam Geçerlik İndeksi (M-KGİ) ile daha derinlemesine incelemek gerekmektedir. Bu bağlamda, insan ve yapay zekâ değerlendiricilerinin puanlarından hesaplanan M-KGİ değerleri arasındaki farklılıklar Wilcoxon İşaretli Sıralar Testi kullanılarak analiz edilmiştir. Analiz sonucunda, insan ve yapay zekâ değerlendiricileri tarafından hesaplanan M-KGİ değerleri arasında istatistiksel olarak anlamlı bir farklılık olmadığı belirlenmiştir ($W = 335.5$, $p = 0.5708$). Başka bir ifadeyle, her iki değerlendirici grubunun maddelerin uygunluğuna ilişkin verdiği kararlar benzerlik göstermektedir. Bulgular, insan ve yapay zekâ değerlendiricilerinin kapsam geçerliği değerlendirmelerinde tutarlı sonuçlar sunduğunu ve yapay zekâ sistemlerinin insan değerlendirmelerine benzer bir güvenirlik sağlayabileceğini düşündürmektedir.

Tartışma, Sonuç ve Öneriler

Bu çalışmada, insan ve yapay zekâ değerlendiricilerinin kapsam geçerliği değerlendirmelerinde benzer değerlendirmeler yapıp yapmadığı incelenmiştir. Sonuçlar, maddelerin büyük çoğunluğu için tüm değerlendiricilerin benzer değerlendirmeler yaptığını göstermiştir. Ayrıca hiçbir değerlendiricinin sistematik olarak diğerlerinden daha yüksek ya da daha düşük puan vermediği; tarafsız bir yönelim sergilediği belirlenmiştir. Bunun yanı sıra, puanlamalara dayalı olarak hesaplanan KGO ve M-KGİ değerleri arasındaki farklılıklara ilişkin istatistiksel analizler anlamlı bir farklılık ortaya koymamıştır. Bu durum, yapay zekâ değerlendiricilerinin insan değerlendiriciler tarafından yapılan değerlendirmelerle uyumlu ve benzer değerlendirmeler sunduğunu düşündürmektedir. Değerlendiriciler arasında istatistiksel olarak anlamlı farklılıkların bulunmaması, bu çalışmanın kapsamı içinde yapay zekâ değerlendiricilerinin kapsam geçerliği analizlerinde destekleyici bir araç olarak kullanılabileceğini göstermektedir.

Yapay zekâ değerlendiricileri, teknik ölçütlere ve standartlara uyuma dayalı olarak nesnellik ve hız gerektiren görevlerde insan değerlendiricilerle kısmen

karşılaştırılabilir puanlama örüntüleri üretebilir. Bununla birlikte, yapay zekâ değerlendiricileri insan değerlendiricilerin yerine geçen bir unsur olarak değil, tamamlayıcı bir araç olarak ele alınmalıdır. Gelecekte, yapay zekâ değerlendiricilerinin özellikle insan denetimli ya da hibrit değerlendirme çerçeveleri içinde daha hızlı ve daha sistematik kapsam geçerliği analizlerini destekleme potansiyeli daha ayrıntılı biçimde incelenebilir. Bu çalışmada genel olarak, bir maddenin puanlanmasında değerlendiriciler arasında orta düzeyde bir uyum bulunmuş; ancak bazı maddelerde yapay zekâ ve insan değerlendiriciler arasında farklılıklar gözlenmiştir. Bu farklılıklar, iki grup arasında değerlendirme ölçütlerinin farklı yorumlanabileceğini düşündürmektedir. Bu durum, belirli maddelerin daha öznel biçimde yorumlanabileceğini ya da daha ayrıntılı açıklamalara gereksinim duyabileceğini göstermektedir. Özellikle bazı maddelerde yapay zekâ değerlendiricileri daha yüksek düzeyde onay verirken, insan değerlendiriciler daha temkinli bir yaklaşım benimsemiştir.

Yapay zekâ değerlendiricileri, teknik yeterliğe odaklanan tutarlı ve kural temelli bir yaklaşım benimsemekte; ancak bağlamsal ve dilsel nüansları gözden kaçırabilmektedir. Buna karşılık insan değerlendiriciler, toplumsal, kültürel ve pedagojik bağlamları dikkate alarak metinleri daha eleştirel ve kapsamlı biçimde değerlendirmektedir. Bu durum, yapay zekâ değerlendirme süreçlerinin bağlamsal analiz, dilsel duyarlılık ve pedagojik hedeflerle uyum gibi alanlarda geliştirilmesi gerektiğini göstermektedir. İnsan değerlendiricilerin bu alanlardaki yeterlikleri, özellikle kritik pedagojik ve kültürel kararların verilmesinde büyük önem taşımaktadır. Örneğin, bazı maddelerde yapay zekâ değerlendiricileri teknik standartlara dayalı olarak maddeleri yeterli görmüş ve nihai forma alınmalarını önermiştir. Öte yandan insan değerlendiriciler, dilsel duyarlılık, bağlam ve içerik niteliği gibi daha ayrıntılı boyutları dikkate alarak "revizyon" ya da "çıkarma" kararları vermiştir. Bu farklılıklar, değerlendirme ölçütlerinin farklı biçimde önceliklendirilmesinden ve yapay zekâ sistemlerinin dilin inceliklerini ve bağlamını tam olarak kavrayamamasından kaynaklanmış olabilir. Özellikle bağlamın kritik olduğu durumlarda, insan değerlendiriciler metinlerin toplumsal etkisini ve pedagojik hedeflerle uyumunu dikkate alarak daha derinlemesine analizler yapmış olabilir. Sonuç olarak, yapay zekâ değerlendiricilerinin değerlendirme süreçlerinin bağlamsal analiz, dilsel duyarlılık ve pedagojik hedeflerle uyum açısından geliştirilmesi önerilmektedir.

Bu örnekler, değerlendiricilerin kapsam geçerliğine ilişkin farklı bakış açıları benimseyebileceğini göstermekte ve bu farklılıklar, yapay zekâ değerlendiricilerinin insan değerlendiricilerin karar verme süreçlerini her zaman tam olarak yansıtamayabileceğini ortaya koymaktadır. Bu bağlamda, yapay zekâ ve insan değerlendiriciler arasındaki iş birliği, kapsam geçerliği değerlendirmelerinde daha tutarlı ve güvenilir sonuçların elde edilmesine katkı sağlayabilir. Yapay zekâ değerlendiricileri, dilsel ipuçlarına dayalı olarak hızlı ve kesin kararlar verme avantajına sahipken, insan değerlendiriciler daha geniş bir bağlamda eleştirel düşünme ve pedagojik kaygıları dikkate alma potansiyeli sunabilir. Bu durum, yapay

zekâ değerlendiricilerinin belirli değerlendirme süreçlerinde insan değerlendiricilere benzer performans gösterebilmesine rağmen, her durumda insan değerlendiricilerin yerini tamamen almasının mümkün olmayabileceğini düşündürmektedir.

Benzer şekilde Kim vd. (2023), yapay zekânın insanlara benzer doğruluk ve güvenirlikte sonuçlar üretebilmesi için insan rehberliğinin kritik olduğunu vurgulamaktadır. Ayrıca Fjelland (2020), yapay zekânın insan zekâsını tam olarak yansıtamayacağına dikkat çekerken, Shane (2019) yapay zekânın belirli görevlerde insan zekâsından daha üstün performans gösterebileceğini ileri sürmektedir. Bu bağlamda, yapay zekâ değerlendiricilerinin insan değerlendirmeleriyle uyum içinde çalışabilmesi için gerekli eğitim, yönerge ve süreç iyileştirmelerinin sağlanması önem taşımaktadır.

Değerlendirme süreçlerinde her iki yaklaşım arasındaki farklılıklar dikkate alındığında, dilsel açıklığın ve ayırt edilebilir seçeneklerin öneminin vurgulanmasının, yapay zekâ değerlendiricilerinin insan değerlendiricilere benzer performans göstermesine katkı sağlayabileceği düşünülmektedir. Bu doğrultuda, ölçütlerin daha açık biçimde tanımlanması ve değerlendiricilerin bu ölçütlere ilişkin algılarının uyumlu hâle getirilmesi, süreçlerin daha tutarlı yürütülmesine katkıda bulunabilir. Yapay zekâ temelli değerlendirme sistemlerinin insan değerlendirmeleriyle birlikte çalıştığı hibrit sistemlerin geliştirilmesi, daha güvenilir ve tutarlı sonuçların elde edilmesini sağlayabilir. Bu tür sistemler, her iki tarafın güçlü yönlerinden yararlanarak daha verimli ve doğru kapsam geçerliği analizlerinin gerçekleştirilmesine olanak tanıyabilir. Ayrıca yapay zekâ ve insan değerlendiriciler arasındaki iş birliğini güçlendiren eğitim programları ve rehberlik sistemlerinin geliştirilmesi, değerlendiricilerin ortak bir anlayışla daha tutarlı sonuçlara ulaşmasına yardımcı olabilir. Bu çalışmanın sonuçlarına benzer şekilde, Low vd. (2024), Tenhundfeld (2023) ve Vassis vd. (2024) tarafından yürütülen çalışmalar da yapay zekâ temelli sistemlerin insan benzeri değerlendirme performansını tam olarak tekrar edemediğini göstermektedir. Bu çalışmalar, söz konusu sistemlerin insan değerlendirmeleriyle uyumlu hâle gelebilmesi için da ha fazla geliştirmeye gereksinim olduğunu vurgulamaktadır.

Kapsam geçerliği, özellikle belirli test maddelerinin doğruluğu ve geçerliğine ilişkin ayrıntılı bir değerlendirmenin gerekli olduğu durumlarda kritik öneme sahiptir. Literatürde, kapsam geçerliği değerlendirmelerinde genellikle en az 3 ila 5 uzmanın görüşüne başvurulması önerilmektedir. Örneğin Lawshe (1975) en az üç uzman görüşünü önerirken, Polit vd. (2007) bir maddenin nihai forma alınıp alınmamasına ilişkin kararın genellikle üç uzmandan alınan geri bildirime dayalı olarak verildiğini belirtmektedir. Bununla birlikte, uzmanların iş yükü ve diğer uygulamaya dönük sınırlılıklar nedeniyle bu uzmanlara ulaşmak ve geri bildirim almak her zaman mümkün olmayabilir (Vassis vd., 2024). Bu tür durumlarda, yapay zekâ değerlendiricilerinin sürece dâhil edilmesi daha hızlı ve daha verimli kapsam geçerliği değerlendirmelerinin yapılmasına olanak sağlayabilir. Low vd. (2024) ve Abu Arqub vd. (2024) tarafından yürütülen çalışmalara benzer şekilde, bu çalışma da kontrollü

koşullar altında yapay zekâ sistemlerinin kapsam geçerliği incelemelerinde destekleyici girdi sağlayabileceğini göstermektedir. Ancak bulgular, özellikle pedagojik, dilsel ya da kültürel yorum gerektiren bağlamlarda yapay zekâ değerlendiricilerinin insan yargısının yerini alabileceği anlamına gelmemektedir.

Bununla birlikte, bu potansiyel dikkatli biçimde yönetilmeli ve olası yöntemsel riskler göz önünde bulundurulmalıdır. Bu doğrultuda, bu çalışmada bağlam kayması ve kullanıcıya özgü geçmişler gibi olası yanlılık kaynakları, standartlaştırılmış yönlendirme çerçevelerinin uygulanması ve değerlendirmelerin bağımsız oturumlarda tekrarlanması yoluyla en aza indirilmiştir. Bununla birlikte, bu önlemler tek başına en uygun güvenirlik düzeyine ulaşmak için yeterli olmayabilir. Bu çalışmada, doğrudan ince ayar yapılmaksızın kapalı kaynaklı yapay zekâ sistemleri (örneğin GPT-4o, Gemini 2.0) kullanılmıştır. Ancak alana özgü geçerlik verileriyle modelin yeniden eğitilmesinin ya da ince ayar yapılmasının performansı daha da artırabileceği kabul edilmektedir. Bu nedenle gelecekteki araştırmalarda, sistematik ince ayara olanak sağlamak ve yalnızca yönlendirme temelli iyileştirmelere olan bağımlılığı azaltmak amacıyla LLaMA 2 ya da RoBERTa gibi açık kaynaklı büyük dil modelleri incelenmelidir. Bu tür yöntemsel geliştirmeler, yapay zekâ temelli kapsam geçerliği değerlendirmelerinin tutarlılığını ve genellenebilirliğini güçlendirebilir.

Sonuç olarak, yapay zekâ değerlendiricilerinin kapsam geçerliği gibi teknik analizlerde etkili bir araç olarak kullanılabileceği; özellikle uzman sayısının fazla olduğu ve uzmanlık gereksinimlerinin yoğun olduğu durumlarda önemli avantajlar sunabileceği söylenebilir. Yapay zekâ ve insan değerlendiriciler arasında gözlenen kısmen karşılaştırılabilir puanlama örüntüleri, bazı durumlarda yapay zekâ araçlarının insan müdahalesinin yerini almaktan ziyade değerlendirme sürecini ek bir kanıt kaynağı olarak destekleyebileceğini göstermektedir. Bu durum, özellikle çok sayıda uzmanın gerekli olduğu ve zaman sınırlılıklarının bulunduğu çalışmalarda yapay zekâ değerlendiricilerinin hız ve verimlilik avantajlarından yararlanılmasına olanak sağlayabilir. Yapay zekâ değerlendiricilerinin veri işleme hızı ve belirli ölçütlere dayalı olarak tutarlı kararlar verebilme kapasitesi, değerlendirme sürecini daha verimli hâle getirebilir. Gelecekteki araştırmalar, yapay zekâ ve insan değerlendirmelerinin daha uyumlu hâle gelmesi için süreçlerin ve değerlendirme ölçütlerinin daha açık biçimde tanımlanmasına odaklanabilir. Bu durum, yapay zekâ değerlendiricilerinin kapsam geçerliği değerlendirmelerinde destekleyici araçlar olarak nasıl daha uygun biçimde kullanılabileceğinin açıklığa kavuşturulmasına yardımcı olabilir. Ayrıca benzer analizlerin daha geniş örneklem grupları ve farklı veri setleriyle yürütülmesi, bu bulguların daha geniş bir bağlamda test edilmesine olanak sağlayacaktır. Sonuç olarak, yapay zekâ değerlendiricilerinin güçlü yönleri, insan değerlendirmelerini tamamlayıcı biçimde kullanıldığında verimlilik ve etkililik sağlayabilir. Bununla birlikte, insan müdahalesi ve rehberliğinin önemini koruduğu kabul edilmelidir. Gelecekteki araştırmaların, yapay zekâ ve insan değerlendirmeleri arasındaki dengeye ilişkin anlayışı geliştirmesi ve yapay zekânın daha geniş alanlarda etkili biçimde uygulanmasına katkı sağlaması beklenmektedir.

References

Ahmed, H. K., & Hussein, J. A. (2020). Design and implementation of a chatbot for Kurdish language speakers using Chatfuel platform. *Kurdistan Journal of Applied Research, 5*(2), 117–135. https://doi.org/10.24017/science.2020.2.10

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Arqub, S. A., Al-Moghrabi, D., Allareddy, V., Upadhyay, M., Vaid, N., & Yadav, S. (2024). Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. *The Angle Orthodontist, 94*(3), 263–272. https://doi.org/10.2319/071123-484.1

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (2021). *Psychological testing and assessment*. McGraw Hill.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Hassabis, D., & Kavukcuoglu, K. (2024, December 11). *Introducing Gemini 2.0: Our new AI model for the agentic era*. Google. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*(1), 1–9. https://doi.org/10.1057/s41599-020-0494-4

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement* (R package version 0.84.1). https://doi.org/10.32614/CRAN.package.irr

Gürdil, H., Soğuksu, Y. B., Salihoğlu, S., & Coşkun, F. (2025). Eğitimde ölçmede yapay zekânın entegrasyonu: Madde tepki kuramı kapsamında veri üretiminde ChatGPT'nin etkililiği. *Trakya Eğitim Dergisi, 15*(2), 887–918. https://doi.org/10.24315/tred.1509299

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

Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *Journal of Information Science*. Advance online publication. https://doi.org/10.1177/01655515241227531

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hinkin, T. R., & Tracey, J. B. (1999). An analysis-of-variance approach to content validation. *Organizational Research Methods, 2*(2), 175–186. https://doi.org/10.1177/109442819922004

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Kim, J., McFee, M., Fang, Q., Abdin, O., & Kim, P. M. (2023). Computational and artificial intelligence-based methods for antibody development. *Trends in Pharmacological Sciences, 44*(3), 175–189. https://doi.org/10.1016/j.tips.2022.12.005

Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Raduescu, C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence, 7*, 100269. https://doi.org/10.1016/j.caeai.2024.100269

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lee, S., & Park, G. (2024). Development and validation of ChatGPT literacy scale. *Current Psychology, 43*(21), 18992–19004. https://doi.org/10.1007/s12144-024-05723-0

Lin, X. (2024). Exploring the role of ChatGPT as a facilitator for motivating self-directed learning among adult learners. *Adult Learning, 35*(3), 156–166. https://doi.org/10.1177/10451595231184928

Low, D. M., Rankin, O., Coppersmith, D. D., Bentley, K. H., Nock, M. K., & Ghosh, S. S. (2024). *Using generative AI to create lexicons for interpretable text models with high content validity* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/vf2bc

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–386. https://doi.org/10.1097/00006199-198611000-00017

MacFarland, T. W., & Yates, J. M. (2016). Kruskal–Wallis H-test for oneway analysis of variance (ANOVA) by ranks. In *Introduction to nonparametric statistics for the biological sciences using R* (pp. 177–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-30634-6_6

Mertler, C. A., Vannatta, R. A., & LaVenia, K. N. (2025). *Advanced and multivariate statistical methods: Practical application and interpretation*. Routledge. https://doi.org/10.4324/9781003562436

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Morjaria, L., Burns, L., Bracken, K., Ngo, Q. N., Lee, M., Levinson, A. J., Smith, J., Thompson, P., & Sibbald, M. (2023). Examining the threat of ChatGPT to the validity of short answer assessments in an undergraduate medical program. *Journal of Medical Education and Curricular Development, 10*, 23821205231204178. https://doi.org/10.1177/23821205231204178

Nemt-Allah, M., Khalifa, W., Badawy, M., Elbably, Y., & Ibrahim, A. (2024). Validating the ChatGPT usage scale: Psychometric properties and factor structures among postgraduate students. *BMC Psychology, 12*, 497. https://doi.org/10.1186/s40359-024-01983-4

Newman, I., Lim, J., & Pineda, F. (2013). Content validity using a mixed methods approach: Its application and development through the use of a table of specifications methodology. *Journal of Mixed Methods Research, 7*(3), 243–260. https://doi.org/10.1177/1558689813476922

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Ofgang, E. (2024, May 14). GPT-4o: What educators need to know. *Tech & Learning*. https://www.techlearning.com/how-to/gpt-4o-how-to-use-it-to-teach

Onal, S., & Kulavuz-Onal, D. (2024). A cross-disciplinary examination of the instructional uses of ChatGPT in higher education. *Journal of Educational Technology Systems, 52*(3), 301–324. https://doi.org/10.1177/00472395231196532

OpenAI. (2023). *GPT-4 technical report* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior, 140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Saraswati, G. P. D., Farida, A. N., & Yuliati, Y. (2023). Implementing AI Replika in higher education speaking classes: Benefits and challenges. *ELT Forum: Journal of English Language Teaching, 12*(3), 207–215. https://doi.org/10.15294/elt.v12i3.76525

Shane, J. (2019). *You look like a thing and I love you*. Hachette UK.

Siegle, D. (2023). A role for ChatGPT and AI in gifted education. *Gifted Child Today, 46*(3), 211–219. https://doi.org/10.1177/10762175231168443

Sihite, M. R., Meisuri, M., & Sibarani, B. (2023). Examining the validity and reliability of ChatGPT 3.5-generated reading comprehension questions for academic texts. *Randwick International of Education and Linguistics Science Journal, 4*(4), 937–944. https://doi.org/10.47175/rielsj.v4i4.835

Tenhundfeld, N. L. (2023). Two birds with one stone: Writing a paper entitled "ChatGPT as a tool for studying human–AI interaction in the wild" with ChatGPT. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67*(1), 2007–2012. https://doi.org/10.1177/21695067231192916

ul haq Akhoon, I., Khan, M. Y., & Bhat, T. A. (2024). *Comparative analysis of AI-generated research content: Evaluating ChatGPT and Google Gemini* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-5265799/v1

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2025). Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research, 35*(9), 951–966. https://doi.org/10.1177/10497323241244669

Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology, 40*(6), 37–55. https://doi.org/10.14742/ajet.9540

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*(5), 508–518. https://doi.org/10.1177/0193945903252998

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, 40. https://doi.org/10.1186/s41239-024-00468-z

Zacharis, G., & Papadakis, S. (2025). Can AI grade like a human? Validity, reliability, and fairness in university coursework assessment. *Educational Process: International Journal, 19*, e2025591. https://doi.org/10.22521/edupij.2025.19.591

Ethical and Author Declarations | Etik ve Yazar Beyanları

Authors' Contributions

The first author contributed to developing the research idea, conducting the data analysis, interpreting the findings, and writing the manuscript. The second author contributed to data collection, writing the introduction, conducting the data analysis, and writing the manuscript. The third author contributed to planning the methodological process, conducting the data analysis, interpreting the findings, writing the conclusion, and carrying out the final review and editing of the manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Approve

This study was reviewed by the Ethics Committee of Gazi University Rectorate at its meeting dated 23.09.2025 and numbered 14, and was approved as ethically appropriate with the document dated 09.10.2025 and numbered E.1352469. The research code number is 2025-1649.

Use of Artificial Intelligence

Artificial intelligence tools were used to evaluate the content validity of the test items and support language editing. The authors are responsible for the scientific content and final version of the manuscript.

Yazarların Katkı Düzeyleri

Birinci yazar, araştırma fikrinin geliştirilmesi, veri analizi sürecinin yürütülmesi, bulguların yorumlanması ve makalenin genel yazımına katkı sağlamıştır. İkinci yazar, veri toplama sürecinin yürütülmesi, giriş bölümünün yazılması, veri analizinin gerçekleştirilmesi ve makalenin genel yazımına katkı sağlamıştır. Üçüncü yazar ise yöntemsel sürecin planlanması, veri analizine ve bulguların yorumlanmasına katkı sağlanması, sonuç bölümünün yazılması ile makalenin son okuma ve düzenleme sürecini yürütmüştür.

Çıkar Çatışması Beyanı

Yazarlar herhangi bir çıkar çatışması bulunmadığını beyan etmektedir.

Etik Onay

Bu çalışma, Gazi Üniversitesi Rektörlüğü Etik Komisyonunun 23.09.2025 tarihli ve 14 sayılı toplantısında incelenmiş; 09.10.2025 tarihli ve E.1352469 sayılı belge ile etik açıdan uygun bulunmuştur. Araştırma kod numarası 2025-1649'dur.

Yapay Zeka Kullanımı

Bu çalışmada yapay zekâ araçları, test maddelerinin kapsam geçerliğini değerlendirmek ve metnin dilsel düzenlemesini desteklemek amacıyla kullanılmıştır. Bilimsel içerik ve nihai sorumluluk yazarlara aittir.

This study has been evaluated under double-blind peer review and verified to be free of plagiarism using iThenticate software.
Bu çalışma, çift taraflı kör hakemlik kapsamında değerlendirilmiş ve iThenticate yazılımı kullanılarak intihal içermediği teyit edilmiştir.

Ethical disclosure | Etik bildirim: ebfd@ankara.edu.tr

Appendices

Appendix 1

Scores Obtained from Human and AI Evaluators

Gemini	ChatGPT	Replika	Chatfuel	Human1	Human2	Human3	Human4
3	2	3	3	3	2	3	3
2	3	3	3	2	2	2	2
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	2	3	3
2	3	2	2	2	2	2	3
3	3	3	3	3	3	3	3
3	3	2	3	3	3	3	3
2	3	2	3	2	2	2	2
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	2	3	2	2	2	3
3	3	3	3	3	3	3	3
3	3	3	3	2	2	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3

Appendix 2

Sample Items from the Test

PART I

Benefit for customers

The advantage of having a multicultural workforce is that it can teach people to look at things from a different point of view and solve problems more effectively. Employees have to learn to see things from a different point of view. As a result, this can often lead to a satisfactory solution that they were unable to see before. In addition, the different skills of cultural understanding and languages enable a company to give better customer service around the world.

5. How does a culturally diverse workforce benefit customers according to Paragraph 4?

a) By enhancing customer support services
b) By making innovative products
c) By maintaining a good reputation
d) By managing cultural differences
e) By offering problem solving skills for customers

PART II

Similarly, Oprah Winfrey, a favorite media entrepreneur, is well-known for her extensive network of charitable efforts. From funding poor children's education to supporting women's empowerment programs, Winfrey's behavior exemplifies the power of using one's platform for the wellbeing of positive change.

8. As stated in Paragraph 2, Oprah Winfrey supports _____.

a) learning organisations
b) children's psychological platforms
c) men's empowerment programs
d) women's employment training
e) public health education

PART III

Despite these recommendations, some obstacles such as moldy or stale food may arise. However, adopting some solutions such as freezing or preserving can help extend the shelf life of fresh items, minimising waste. Food festivals can continue to be a source of joy and inspiration while protecting the planet by considering the cost of food waste and embracing environmentally friendly approaches. They give people a chance to try delicious foods, learn about local traditions, and meet other food lovers.

15. According to Paragraph 4, food festivals _____.
a) affect the people's overall health
b) decline the number of imported foods
c) help people meet local food producers
d) make people prepare meals cooked in other countries
e) reduce the financial impact of wasteful consumption