System 133 (2025) 103784

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays☆

Burak Aydın [a,b,*], Tarık Kışla [b], Nursel Tan Elmas [c], Okan Bulut [d]

[a] Educational Sciences, Leuphana University, Germany
[b] Faculty of Education, Ege University, Türkiye
[c] Gazi University, Türkiye
[d] Faculty of Education, University of Alberta, Canada

ARTICLE INFO

Keywords:
Automated scoring
Large language models
Zero-shot with rubric
Rater reliability
Turkish essays
Multilevel models

ABSTRACT

Automated scoring (AS) has gained significant attention as a tool to enhance the efficiency and reliability of assessment processes. Yet, its application in under-represented languages, such as Turkish, remains limited. This study addresses this gap by empirically evaluating AS for Turkish using a zero-shot approach with a rubric powered by OpenAI's GPT-4o. A dataset of 590 essays written by learners of Turkish as a second language was scored by professional human raters and an artificial intelligence (AI) model integrated via a custom-built interface. The scoring rubric, grounded in the Common European Framework of Reference for Languages, assessed six dimensions of writing quality. Results revealed a strong alignment between human and AI scores with a Quadratic Weighted Kappa of 0.72, Pearson correlation of 0.73, and an overlap measure of 83.5 %. Analysis of rater effects showed minimal influence on score discrepancies, though factors such as experience and gender exhibited modest effects. These findings demonstrate the potential of AI-driven scoring in Turkish, offering valuable insights for broader implementation in under-represented languages, such as the possible source of disagreements between human and AI scores. Conclusions from a specific writing task with a single human rater underscore the need for future research to explore diverse inputs and multiple raters.

1. Introduction

Automated scoring (AS) can be broadly defined as using computers to characterize the quality of individuals' performance (Foltz et al., 2020). Tracing back to the late 1960s (e.g., Page, 1966), research activity on AS for student essays has gained momentum in the last two decades. In their recent introduction to *The Routledge International Handbook of Automated Essay Evaluation*, Shermis and Wilson (2024) summarized these developments under three definitions: automated essay scoring (AES), automated essay evaluation (AEE), and automated writing evaluation (AWE). AES involves assigning scores to essays for summative assessment purposes, AWE refers to formative assessment to create scores and feedback on texts of varying lengths, and AEE encompasses both assigning scores

☆ Preliminary results of this study were presented at the 9th International Conference on Measurement and Evaluation in Education and Psychology.

* Corresponding author.
E-mail addresses: burak.aydin@leuphana.de, burak.aydin@ege.edu.tr (B. Aydın), tarik.kisla@ege.edu.tr (T. Kışla), nursel.elmas@gazi.edu.tr (N.T. Elmas), bulut@ualberta.ca (O. Bulut).

https://doi.org/10.1016/j.system.2025.103784
Received 6 March 2025; Received in revised form 1 July 2025; Accepted 14 July 2025
Available online 21 July 2025
0346-251X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

and generating feedback for both summative and formative assessment purposes. AES, by definition, does not require the generation of qualitative feedback and is arguably a more straightforward task than both types of evaluation. These three definitions can be considered indicators of a vast body of literature on AS of student essays. Reviews on the subject include those by Yavuz et al. (2024), Gierl et al. (2014), and Ifenthaler (2022). However, as Firoozi and Gierl (2024) pointed out, most of this existing research has focused on essays or texts written in English, probably due to the significant demand for English language learning and evaluation (McNamara & Potter, 2024).

There is a growing interest in utilizing AES for languages other than English, however. In his recent review, Shermis (2020) provided examples of AES for Chinese, Hebrew, Bahai Malaysian, Japanese, Filipino, Finnish, German, Thai, French, Korean, and Swedish. Attempts to score Arabic (Abdul Salam et al., 2022; Gomaa & Fahmy, 2014), Brazilian Portuguese (Ribeiro et al., 2024), and Indonesian inputs (Ratna et al., 2019; Salim et al., 2022) have also been emerging. It is not surprising to witness such an effort, because, despite some validity-related concerns, AES is imminent due to human raters' inconsistency in scoring and the high costs of rater training, especially when acceptable reliability is desired. Nevertheless, many other languages, including Turkish, have yet to see such high research activity in this area.

1.1. The Turkish language

Turkish, the seventh most used language worldwide with approximately 200 million users (MCTRT, 2024), belongs to the Turkic family under Altaic language phylum along with Mongolic, Tungusic, Koreanic and Japonic language families (Janhunen, 2023). There has also been an interest to learn Turkish as a second language. For example, the number of international students seeking higher education in Türkiye has been steadily increasing (Council of Higher Education [YÖK], 2023). This growth can be largely attributed to the strategic efforts of YÖK, particularly its internationalization policies and forward-looking planning (YÖK, 2025). Additionally, scholarship programs play a significant role in promoting international student mobility in Türkiye (Türkiye Scholarships, 2025). International students wishing to enroll in undergraduate and graduate programs taught in Turkish must demonstrate their proficiency in reading, listening, speaking, and writing in Turkish (YÖK, 2024). Another example is that adult learners of Turkish as a second language who aim to work in specific professional fields in Türkiye are required to certify their Turkish proficiency at a minimum B2 level, as mandated by relevant regulations (Board for Medical Specialties [TUK], 2024). The importance of learning Turkish has grown, not only for academic and professional reasons but also for social integration purposes. Türkiye's ongoing migration dynamics and the increasing number of migrants have amplified the necessity of learning Turkish (see UNICEF, 2023). Given the rising demands influenced by these circumstances, assessing Turkish language skills, especially through standardized, valid, and reliable measurement tools, has become crucial for maintaining the quality of second language instruction in Turkish.

The rising demand has also resulted in institutional developments in Türkiye. The field of teaching Turkish as a second language is supported by a robust institutional framework, both domestically and internationally. Key organizations, such as the Yunus Emre Institute, the Turkish Maarif Foundation, the Turkish Cooperation and Coordination Agency (TIKA), and the Ministry of National Education, play a leading role in promoting Turkish language education abroad. Within Türkiye, Turkish Language Teaching Centers (TÖMER), which are operated by universities take the main responsibility. Additionally, public education centers and private language courses contribute to this effort. The increasing significance of teaching Turkish as a foreign language has led to the establishment of standardized language proficiency exams, such as the Turkish Proficiency Exam (TYS), developed by the Yunus Emre Institute. Overall, these multifaceted institutional efforts in Turkish language teaching address the diverse needs of individuals wishing to learn Turkish as a foreign language and help document their proficiency on an international scale.

For those studying Turkish for academic and professional purposes, the ability to produce various written texts—such as articles, reports, projects, petitions, and emails—is crucial. Demir and Genç (2019) noted that international students learning Turkish as a foreign language face significant challenges in writing. Assessing writing skills through standardized exams is necessary for helping learners achieve their academic and professional goals. As the number of learners continues to grow, the scalable nature of AS systems offers a significant advantage by increasing assessment capacity and supporting the sustainability of evaluation processes. However, Turkish has very specific language processing challenges due to its agglutinative word structure, consonant/vowel harmony, practically infinite vocabulary, and phonological rules (Tohma & Kutlu, 2020). As reviewed by Firoozi et al. (2023), very few studies have focused on AS for the Turkish language, and no publicly available Turkish essays exist to study AES. Hence, the present study aims to fill this gap by empirically testing AES for Turkish using artificial intelligence (AI) and sharing both the research results and the labeled essays for use in further studies. Ultimately, this study has the potential to drive increased AES research in Turkish and to enhance global understanding of AI in non-English languages, thereby facilitating cross-linguistic AES development.

2. Related work

2.1. A Brief history of AES systems

Traditionally, AES systems relied on rule- and feature-based predictive models that required predefined linguistic features or scoring rules to predict essay quality. Rule-based predictive systems for essay scoring operate by extracting specific textual features or applying explicit scoring rules, then using those features to assign a score. Typically, domain experts define a set of rules or features that correlate with writing quality, such as the number of grammar errors, vocabulary complexity, essay length, or the presence of relevant keywords (Bexte et al., 2024). Classic systems like ETS's e-rater® follow this approach: they analyze an essay for multiple hand-crafted features (e.g., grammar errors, lexical complexity, and organization) which are validated to ensure they logically relate to

the writing prompt, and then combine these features in a statistical model to produce a final score (Burstein et al., 2013). Some rule-based AES systems use simple if-then rules (e.g., "if keyword X appears, award points") for short answers or specific content, while others use regression or classification models trained on a feature vector representation of essays (Bexte et al., 2024). In all cases, these systems require labeled training data (essays graded by humans) to either inform the rule design or to fit the predictive model's parameters. Rule-based AES models are relatively transparent in how they make decisions because scoring rules or feature weights can be inspected to understand why a certain score was given. Furthermore, these models are often deterministic and thus produce the same result for the same input every time, ensuring high consistency. However, since manually crafted rules often struggle to capture the full complexity of good writing, rule-based AES models mostly evaluate form over meaning, thereby misjudging essays that are conceptually strong but do not necessarily fit expected patterns. In addition, rule-based models often must be redeveloped or retrained for each new prompt or writing task due to their prompt-specific nature. This rigidity makes rule-based AES less practical in settings where essay topics vary widely.

Recently, advances in large language models (LLMs) and generative AI have introduced new approaches to essay scoring that leverage deep neural networks trained on massive text corpora (Bulut et al., 2024; Kaldaras et al., 2024). It is possible to utilize AI not only for AES, but in each step of testing (Burstein et al., 2024; Langenfeld et al., 2022; von Davier & Burstein, 2024). Among several LLMs, Chatbot Generative Pre-trained Transformer (ChatGPT) has evidently influenced the field (Allam et al., 2023). The use of ChatGPT for AES has attracted researchers' interest, demonstrating the potential for scoring efficiency, as Shermis (2024) stated. Regarding agreement between AES and human raters, research has shown both discouraging (e.g., Bui & Barrot, 2025; Kim et al., 2024) and encouraging (e.g., Bucol & Sangkawong, 2024) results. These contradictory results are not unanticipated, as shown by Yavuz et al. (2024); the consistency and performance of AI depend on how it is utilized for AES. The authors utilized the same LMM, the ChatGPT, in two different ways, first with the default settings and then with additional settings by providing a detailed prompt and asking for more deterministic outputs; the latter produced consistent scores when scoring the same essay in different sessions and provided scores closer to human scores when scoring different student essays.

Prompting strategy is a critical factor for AI utilization for AES. Xiao et al. (2024) categorized these strategies as: zero-shot without rubrics, zero-shot with rubrics, and few-shot with rubrics. Rubrics are essentially scoring instructions and, therefore, play a vital role in rater consistency in language assessment. Zero-shot refers to feeding LLMs with materials but not expected results (e.g., human scores), with no prompt adjustment after the initiation of scoring; in other words, using LLMs for an AES task with no additional training. Few shots, on the other hand, involve feeding LLMs with materials and also with expected results (i.e., k samples). The main advantage of the zero-shot approach is its convenience, but Xiao et al. (2024) reported discouraging results for the zero-shot approach for AES in English. However, Chamieh et al. (2024) reported mixed findings depending on the context, whereas Seßler et al. (2024) reported highly optimistic results for AES in German.

Another key consideration in the utilization of machines to evaluate essays is based on the purpose and genre of the essays. Shermis and Wilson (2024) devised six categories of automated essay evaluations: long essays (i.e., 150 words or more), short-form constructed responses (i.e., 1 to <150 words), content-intensive responses, content-superficial responses, summatively scored essays, and formative assessments. Among these categories, arguably, the least challenging for AI is the summatively scored essays which are generally written as part of a summative assessment, where students demonstrate knowledge and understanding of a topic and It involves synthesizing information, critically analyzing the subject, and presenting a well-structured essay with an introduction, body, and conclusion (Shermis & Wilson, 2024). These essays are typically employed to assess a student's overall performance, often under time constraints and they are generally preferred in high-stakes testing in which evidence for validity is important; hence, challenges for AES (DiCerbo et al., 2020) are generally addressed in this area, such as the need for good writing prompts and good rubrics.

2.2. AES validation

State-of-the-art methods for validating scores from AES systems typically focus on aligning automated scores with human-rater scores while addressing broader validity measures. Most AES validation approaches benchmark against scores assigned by a human rater because a human-machine agreement is considered the gold standard (Powers et al., 2015). Metrics like Pearson correlation, Quadratic Weighted Kappa (QWK), and distributional overlap are employed to measure the agreement between human and automated scores. These methods have shown that AES systems can approximate human scoring accuracy under specific conditions (Hamner & Shermis, 2012; Powers et al., 2002).

Pearson correlation (r) measures the consistency in the rankings of two sets of scores and quantifies the degree to which AES scores and human scores move together in a linear fashion. The value of r ranges from −1 to 1, where values closer to 1 indicate a high degree of alignment between AES and human scores. A study by Powers et al. (2002) used Pearson correlation to evaluate the relationship between AES scores and human scores, reporting correlations in the range of 0.7–0.9 as strong evidence of system reliability. Although Pearson correlation enables a straightforward evaluation of trends, it fails to measure how close the AES scores are to human scores. In addition, Pearson correlation may fail to reflect the true agreement between AES and human scores when the relationship is not linear (e.g., AES performs better on mid-range scores).

QWK quantifies the level of agreement between the AES system and human raters while considering the possibility of agreement occurring by chance. Unlike the Pearson correlation, QWK considers the size of discrepancies between ratings as it assigns higher penalties for larger discrepancies (e.g., a difference of three points is penalized more than a difference of one point). Since the Hewlett Foundation's famous Automated Student Assessment Prize (ASAP) competition in 2012, QWK has been used extensively to evaluate AES models' performance across datasets (Taghipour & Ng, 2016). QWK values higher than 0.65 are reported as validity evidence for AES (Palermo & Wibowo, 2024); however, QWK also has several limitations, such as being influenced by the range of scores and

providing less reliable scores in the case of uneven distributions of scores. A recent study by Doewes et al. (2023) critiques QWK's limitations and advocates for additional measures to complement this metric in comparing AES and human scores.

Distributional overlap quantifies the extent to which the score distributions (e.g., frequency or density of scores across a range) from the AES system match those from human raters. This approach focuses specifically on the aggregate behavior of the scoring system, rather than on individual essay-level comparisons. In addition to the visual inspection of mismatches in frequency or patterns, the area under the overlapping curves can also be used to measure how closely the AES and human score distributions align. This method focuses on the overall scoring behavior of the system, complementing essay-level metrics like Pearson correlation and QWK. However, this advantage turns into a weakness due to the lack of granularity. Powers et al. (2015) suggested using distributional comparisons to validate AES systems, especially for fairness and demographic equity.

In addition to the metrics summarized above, there are also alternative methods that do not necessarily rely on the availability of human scores. For example, tertium quid is a third variable or criterion used to evaluate the alignment of AES scores with true writing ability, independent of human rater scores. This metric addresses the limitations of directly comparing AES scores with human raters, who may themselves have biases or inconsistencies. In this method, external criteria such as standardized test scores, course grades, and performance in writing-related activities can be used to compute correlations or build a regression model to quantify the alignment between AES scores and the selected criterion. Alternatively, the aggregated scores from multiple human raters can also be utilized as a criterion that reflects the "true" scores (Cohen et al., 2018). The effectiveness of this method depends highly on the choice of tertium quid (i.e., finding a reliable, unbiased, and universally accepted criterion). Tertium quid can also provide a resolution method when two raters are not in agreement, and a third rater is needed for a final score (Ahmadi, 2019; Elosua, 2022), which is a real possibility since, despite rater training and quality rubrics, score discrepancies might still exist among the raters. For example, the "hawk and dove effect" refers to the systematic differences in how raters score, often resulting in leniency or stringency errors (McManus et al., 2006). The importance of human rater expertise in validating AS has been discussed in detail by several studies. Wind et al. (2017) for example, discussed that rater severity in human assigned scores might affect the fairness and validity of automated scoring outcomes whereas Powers et al. (2015) reported that AS aligns more closely with scores assigned by experienced raters. Kumar and Boulanger (2020) utilized the level of agreement between two human raters as a benchmark to evaluate AS performance; hence demonstrated that human rater expertise is curial for setting a standard for AS.

2.3. AI-AES for Non-English languages

Shermis (2020, pp. 126–128) reviewed AS systems for 13 different non-English languages; however, LLMs were not mentioned, which are the prevailing technology used for scoring at present. Further, in their systematic literature review, Xu et al. (2024) claimed that LLMs such as ChatGPT have the potential to impact AES in several ways, including precision enhancement for scoring texts based on coherence, argumentation, and grammar. Recent articles have been optimistic about the future use of LLMs for AS (e.g., Jamieson et al., 2024; Karatay & Karatay, 2024). For example, Seßler et al. (2024) compared human vs. LLM-generated scores for 20 essays written in German by 7th- and 8th-grade students. A total of 10 pre-defined scoring criteria were utilized both by 37 teachers and 4 LLM zero-shot approaches, and the authors reported high alignment. Feng et al. (2024) used a holistic rubric with only two criteria to score Chinese essays and selected 189 samples based on scores from two human raters. The chosen essays were then scored by LLMs with a similar categorization as Xiao et al. (2024); however, the authors reported that zero-shot or few-shot approaches with or without a rubric performed poorly, whereas additional training for GPT resulted in promising results. Ribeiro et al. (2024) reported similar conclusions after training GPT to score high school-level essays in Portuguese, scoring four different criteria from 1 to 5 by two human raters and five LLMs, and finding substantial agreement.

To our knowledge, there has been no published research on using LLMs to score essays written in Turkish. Arslan and Özdamar (2019) attempted to score computer programming assignments using relatively simple programming approaches. Similarly, Çetin and Ismailova (2019) studied the statistical features of Turkish essays (e.g., total words used, total number of sentences). Uysal and Doğan (2021) reported promising results for trained neural network approaches to score short student answers in a Turkish test. Ince and Kutlu (2021) also reported promising results in scoring short-answer student responses to a computer networks test using latent semantic analysis. Firoozi et al. (2023) provided programming insights on how to utilize LLMs for Turkish language research, such as program installations, data pre-processing, and model training, but did not report any empirical results.

The current study aims to address the gap between state-of-the-art AES approaches and scoring attempts for Turkish essays. Utilizing a rubric with substantial validity evidence and comparing it to scores by professional human raters, this study aims to address the following research questions:

1. How reliably does zero-shot Chat GPT-4o with rubric (OpenAI, 2023, 2024) perform in automated Turkish essay scoring (ATES)?
2. Does the agreement between professional raters and ATES change based on raters' characteristics?

3. Materials and methods

3.1. Data source and the scoring rubric

The dataset for the present study consisted of 590 essays, each written by a participant aged 18 to 40, including both undergraduate and graduate students, as well as professionals. They were motivated to learn Turkish for various reasons, such as fulfilling academic or professional requirements or integrating into environments where Turkish is spoken as a native language. Most participants have

studied Turkish for one to three years, and their language proficiency levels range from beginner to advanced. The participants were invited via e-mail to voluntarily provide their answers to a prompt that required writing an e-mail complaining about a tour (see supplementary). The task had to be completed in 20 min with at least 125 words. These essays can be classified as "guided" or "semi-guided" writing tasks (Hyland, 2003; Read, 1990) or in a broader category as "independent" writing tasks (Guo et al., 2013). Each essay was anonymized by replacing participant names with the words *name* and *surname*. In addition to student essays, we informed human raters about the study's purpose and collected data on rater characteristics using a short self-report form. Specifically, we collected information on raters' gender and years of experience in scoring productive skills (i.e., writing and speaking), in addition to asking if they tended towards leniency (i.e., dove) or stringency (i.e., hawk) when they felt that an essay fell in between two ratings. These characteristics, along with all 590 essays and scores by the human raters and Chat GPT-4o zero-shot, are shared as supplementary files.

Each essay was scored by a single rater using a scoring rubric developed based on the Common European Framework of Reference for Languages (Council of Europe, Europe Modern Languages Division, 2001). The raters were experienced educators who specialize in teaching Turkish as a foreign language. Each rater holds at least a bachelor's degree in Turkish language and literature, Turkish language education, or a related field, with the majority having a master's degree in applied linguistics. On average, these raters have over ten years of experience teaching Turkish as a second or foreign language. For at least the past five years, they have been officially designated to assess writing performances for a standardized Turkish proficiency test and received regular training on how to apply an analytic scoring rubric. Our study involved a total of 20 raters, their average scoring experience in years was 9.5, with a minimum value of 6 and a maximum of 13; 7 out of 20 raters were female (35 %), and 13 were male (65 %). Further, 5 out of 7 female raters and 8 out of 13 male raters selected dove as their tendency. Raters followed a rubric with substantial validity evidence that had been previously investigated during an international audit (Elmas, 2025). The rubric included six criteria with different scoring weights: task completion, coherence/cohesion, grammatical accuracy, lexis, spelling/punctuation, and format. Both human raters and Chat GPT-4o applied the rubric. The present study focused only on the overall essay scores, which ranged between 0 and 10. A total of 24 essays (4 %) were marked as unusual (i.e., extremely good or completely irrelevant) by the human raters, but were not excluded from the data set.

3.2. AES procedure

We utilized the ChatGPT-4o model developed by OpenAI because of its strong performance in multilingual tasks, including those involving under-resourced languages like Turkish (Robinson et al., 2023). Compared to other commercially available and open-source large language models, ChatGPT-4o offers a robust application programming interface (API), advanced reasoning capabilities, ease of integration for bulk processing, and dependable zero-shot performance. We accessed the model via OpenAI's official API between July and August 2024 through a custom-built web interface developed using PHP, HTML, CSS, and JavaScript. Although the widespread adoption of ChatGPT-4o in Türkiye is still in its early stages (i.e., Arslan, 2025), it is garnering attention in both academic and educational contexts (Chapelle, 2025; Yang & Li, 2024). The version of the model we used corresponds to the GPT-4o release from May 2024, as specified in OpenAI's system documentation (OpenAI, 2024). Each essay was individually scored using the zero-shot with rubric prompt, ensuring an independent evaluation without prior fine-tuning or exposure to the data set. The scoring process followed a structured workflow to ensure data reliability and consistency, involving the following steps:

1. Data Preparation: Student essays were formatted to be compatible with the language model's data structure. This step included removing extra spaces and information, leaving the essay itself as the input.
2. Rubric Definition: A comprehensive rubric was utilized, which encompassed the six writing quality dimensions.
3. Prompt Design: Each essay was presented to the model using a zero-shot prompt (instruction) aligned with the rubric's criteria.
4. Essay Evaluation: Each essay was independently evaluated in separate sessions to prevent data leakage and ensure the independence of each assessment.
5. Data Collection and Quality Control: In fewer than 20 out of 590 cases, the ATES interface failed to provide a score. For these cases, to enhance consistency and reduce the model's natural variability, each essay was evaluated multiple times, and the average score was calculated. The evaluation results were stored in JSON format, facilitating data analysis and statistical review.

Our interface enabled the automated assessment of 500 student essays via the OpenAI API. Each essay was scored independently, and the results were output in JSON format, making data processing and analysis more systematic and accessible. The prompt specifically instructed the AI to act as a certified language assessor and evaluate anonymized student essays written in Turkish using a defined rubric. It requested that the model provide a numerical score for each criterion, as well as a total score between 0 and 10. Furthermore, the AI was directed to present its output in JSON format to enable automated processing. This standardized structure ensured consistent evaluation across all essays. In supplementary files, we partially shared the prompt and hence the rubric.

3.3. Data analysis

To answer the first research question, we compared the ChatGPT-4o zero-shot ATES with rubric and human scores using R (R Core Team, 2025). We investigated (a) distributional overlap (Makowski et al., 2019), (b) Pearson correlation coefficient (Wei & Simko, 2024), (c) quadratic weighted kappa (QWK) using the *metrics* package (Hamner & Frasco, 2018), and (d) proportion of tertium quid computed as the proportion of cases where the difference between ATES and human score (HS) was at least 2.5 out of 10 points.

To answer the second research question, we examined if the rater characteristics affected score differences between ATES and HS. The dependent variable was the difference between the scores, where scores close to 0 indicate better agreement. The independent variables at the student level were whether the essay was considered usual or not; while the independent variables at the rater level were gender, amount of scoring experience in years, and whether the raters considered themselves as a dove or a hawk. These analyses were completed using multilevel models (Bates et al., 2015; Kuznetsova et al., 2017; Pinheiro & Bates, 2000; Snijders & Bosker, 2012) given that the dependent variable was nested in 20 raters; see Model 1a below, which combines the level 1 and level 2 equations:

$$Level\ 1:\ (ATES - HS)_{ij} = \beta_{0j} + \beta_{1j}UT_{ij} + R_{ij}, \tag{1}$$

$$Level\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01}G_j + \gamma_{02}E_j + \gamma_{03}D_j + U_{0j}\ ;\ \beta_{1j} = \gamma_{10} \tag{2}$$

$$Model\ 1a: (ATES - HS)_{ij} = \gamma_{00} + \gamma_{10}UT_{ij} + \gamma_{01}G_j + \gamma_{02}E_j + \gamma_{03}D_j + U_{0j} + R_{ij}, \tag{3}$$

where $(ATES - HS)_{ij}$ is the score difference between the ATES and human scorers (HS) for essay i scored by rater j, γ_s are the regression coefficients, UT is effect coded as 1 if rater j considers essay i as a usual text (i.e., random letters, suspiciously good essays) and -1 if the rater does not consider it usual, G is also effect coded as 1 for female and -1 for male raters, E is a grand mean centered continuous variable where 0 refers to average experience in years, D is also effect coded as 1 if rater j thinks they lean towards leniency or -1 for stringency. U_{0j} is the rater level, and R_{ij} is the essay level residual. Effect coding and grand mean centering were chosen to have a meaningful intercept γ_{00}.

For Model 1b, we re-ran the same model after standardizing $(ATES - HS)_{ij}$ and E to have a mean of 0 and standard deviation of 1 and dummy coding (i.e., 0 and 1) our categorical variables to compute partially standardized regression coefficients (Lorah, 2018).

4. Results

A single human rater scored each essay, and these scores ranged between 0 and 10, with a mean value of 6.1, median value of 6.3, and standard deviation of 2.4. The range for ATES scores was also between 0 and 10, with a mean value of 6.2, median value of 6.3, and standard deviation of 2.2. The difference score (i.e., $ATES - HS$) ranged between -6.3 and 5, with a mean value of 0.09, median value of 0, and standard deviation of 1.7. Additional descriptive statistics are provided in the supplementary tables.

Fig. 1 shows the overlap between HS and ATES. The overlap measure computed based on Kernel density and composite trapezoid rule was 83.5 %. The QWK value was 0.72. In 89 out of 590 cases, the absolute difference between scores (i.e., $|ATES - HS|$) was larger than 2.5, corresponding to a tertium quid proportion of 15.01 %. Table 1 shows the correlation between HS and ATES for the overall scores, along with the correlations between scores assigned for each criterion. These correlations ranged between 0.39 and 0.92. The Pearson correlation coefficient for the overall scores by HS and ATES was 0.73. Low correlations were generally observed with the sixth criterion, which corresponded with the format of the text.

The empty multilevel model resulted in an intercept of 0.09 $(SE = 0.23, t_{570} = 0.38, p = .70)$, between variance of 0.97 and within level variance of value of 2.04 (see Table 2), hence an unconditional ICC value of 0.32 (see supplementary for a plot of random intercepts). As reported in Table 3, Model 1a resulted in an intercept value of 0.21 $(SE = 0.27, t_{569} = 0.76, p = .45,)$ indicating that the difference between ATES and HS was 0.21 points for a rater with 9.5 years of experience (i.e., the average for the study sample). The coefficient for unusual text indicator was 0.03 $(SE = 0.15, t_{569} = -0.17, p = .87)$ and self-reported rater tendency was 0.05 $(SE = 0.22, t_{16} = 0.81, p = .81)$ indicating that these predictors had a negligible effect on the difference score. However, gender and experience had relatively large partially standardized coefficients, as reported in Table 4. The coefficient for gender indicator was 0.37 $\left(SE = 0.23, t_{16} = 1.62, p = .13, \beta_{partial} = 0.43\right)$ indicating that the average score difference for female raters was 0.58 (i.e., 0.21 +

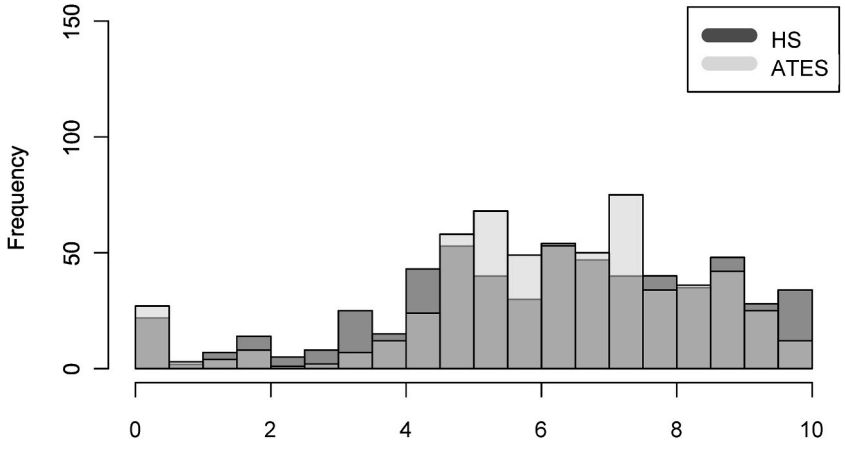

Fig. 1. Distribution of HS and ATES

Table 1
Correlations between criterion and sum scores.

	C1H	C2H	C3H	C4H	C5H	C6H	HS	C1A	C2A	C3A	C4A	C5A	C6A	ATES
C1H	1													
C2H	0.82	1												
C3H	0.76	0.73	1											
C4H	0.72	0.68	0.72	1										
C5H	0.64	0.65	0.72	0.59	1									
C6H	0.59	0.65	0.50	0.50	0.48	1								
HS	0.92	0.89	0.89	0.86	0.77	0.68	1							
C1A	0.65	0.63	0.57	0.58	0.50	0.52	0.68	1						
C2A	0.64	0.61	0.56	0.54	0.50	0.52	0.67	0.82	1					
C3A	0.59	0.60	0.61	0.54	0.58	0.46	0.66	0.77	0.75	1				
C4A	0.56	0.54	0.53	0.52	0.46	0.46	0.61	0.78	0.69	0.81	1			
C5A	0.48	0.52	0.51	0.41	0.53	0.39	0.55	0.57	0.63	0.69	0.50	1		
C6A	0.49	0.51	0.46	0.45	0.39	0.44	0.54	0.66	0.71	0.67	0.63	0.56	1	
ATES	0.67	0.67	0.63	0.60	0.57	0.54	0.73	0.92	0.90	0.92	0.88	0.73	0.78	1

Note: C = Criterion, H = Human, A = Automated, HS = Human Score, ATES = Automated Turkish Essay Scoring.

Table 2
Empty model results.

	Value	Std. Error	*df*	*t*-value	*p*-value
(Intercept)	0.09	0.23	570	0.38	0.704

Table 3
Model 1a fixed effect results.

	Value	Std. Error	*df*	*t*-value	*p*-value
(Intercept)	0.21	0.27	569	0.76	0.449
UT	−0.03	0.15	569	−0.17	0.869
G	0.37	0.23	16	1.62	0.125
E	−0.14	0.09	16	−1.47	0.160
D	0.05	0.22	16	0.24	0.812

Note. UT = unusual text, G = gender, E = experience, D = Dove.

0.37×1) and for male raters was -0.16 (i.e., $0.21 + 0.37 \times -1$). The coefficient for experience in years was -0.14 $\left(SE = 0.09, t_{16} = -1.47, p = .16, \beta_{partial} = -0.18\right)$ indicating that a one-year change in experience was predicted to correspond to a decrease of -0.14 points in the score difference. For example, the score difference for a relatively novice rater (e.g., 6.5 years of experience) was predicted to be 0.63 points on average (i.e., $0.21 - (-3 \times 0.14)$), whereas it was predicted to be -0.21 points (i.e., $0.21 - (3 \times 0.14)$) for a relatively experienced rater (e.g., 12.5 years). Last but not least, for Model 1a, the conditional ICC was 0.28; the explained variance computed as the ratio of total variance difference between the null and Model 1a to the total variance of the null was 4.9 % (i.e., $pseudo - R^2$), indicating that study predictors explained approximately 5 % of the variance in difference scores.

5. Discussion and conclusion

In this study, we aimed to empirically test AES for the Turkish language using AI, specifically ChatGPT-4o. Our results indicated that LLMs can approach human-level scoring performance under the right conditions. Specifically, using a zero-shot prompting strategy with an explicit rubric, ChatGPT-4o evaluated guided Turkish writing tasks with acceptable accuracy and consistency. The model's scores showed agreement with expert human ratings (QWK = 0.72) and a strong positive correlation ($r = 0.73$) with human scores. Such metrics indicate convergent validity of the AES scores with human judgment, approaching the reliability one would expect

Table 4
Model 1b fixed effect results.

	Value	Std. Error	*df*	*t*-value	*p*-value
(Intercept)	−0.16	0.28	569	−0.58	0.561
UT	−0.03	0.18	569	−0.17	0.869
G	0.43	0.26	16	1.62	0.125
E	−0.18	0.13	16	−1.47	0.160
D	0.06	0.26	16	0.24	0.812

Note. UT = unusual text, G = gender, E = experience, D = Dove.

between two trained human raters (Palermo & Wibowo, 2024; Powers et al., 2002).

Notably, we observed that a handful of outlier cases (i.e., the AES score diverging markedly from the human score) had a disproportionate effect on consistency measures. After removing these anomalous cases, the alignment improved further. This suggests that, aside from a few exceptional instances, ChatGPT-4o′s scoring behavior was generally stable and in line with human evaluators. In fact, recent work has shown that ChatGPT-4o can produce highly self-consistent ratings across iterations, such as intra-class correlations of 0.94–0.99 (Hackl et al., 2023), reinforcing the notion that most variability stems from isolated misunderstandings rather than random inconsistency. Overall, the effectiveness of the zero-shot rubric-guided approach underscores the promise of LLM-based AES for languages like Turkish, even without task-specific fine-tuning.

In addition to the above metrics, we analyzed the disagreement in scores from a qualitative perspective. Ratings that resulted in a difference of 4 or more points out of 10 between ATES and HS were examined further (i.e., $|\text{ATES} - \text{HS}| \geq 4$). We first examined the cases with a large negative difference score (i.e., HS > ATES), observing that ChatGPT tended to assign a score of 0 to texts unaligned with the task. In contrast, human raters were more likely to assign non-zero scores when the performance was partially aligned with the given task. ChatGPT might be more sensitive to morphological and lexical errors as it tends to assign lower scores than human raters, who might be more tolerant when the text's intended meaning is still understandable.

Regarding the *format* criterion, we found that ChatGPT struggled to assess the formal characteristics of the text type compared to human raters. This suggests that ChatGPT might not fully understand the *format* criterion. On the other hand, human raters were inclined to assign a more lenient score, focusing on the overall consistency and coherence of the text, which might explain why human raters tend to score higher than ChatGPT in such cases. We then focused on the cases with a large positive difference score (i.e., ATES > HS). This investigation indicated that, in some cases, some parts of the essays were suspiciously well-written due to stylistic differences in the text, various language level elements, or disparities in pragmatics and discourse competence that native speakers of the target language might have. These texts seemed to have components borrowed from the Internet, and human raters consequently assigned lower scores to the texts. Overall, these large differences between HS and ATES negatively affected the evaluation metrics. Specifically, when we removed cases with $|\text{ATES} - \text{HS}| \geq 4$, the overlap measure increased from 83.5 % to 84.1 %, QWK from 0.72 to 0.76, and correlation from 0.73 to 0.76; the proportion of tertium quid decreased from 15 % to 13 %.

Performances flagged as unusual by the human raters (i.e., 24 out of 590 essays) were also studied. Both human raters and ChatGPT assigned 0 points to texts produced in languages other than Turkish. Similarly, texts produced by random key strikes that did not meet the textuality criteria were also assigned 0 points by both human raters and ChatGPT. Human raters flagged 5 out of 24 unusual texts to share their concerns about the authenticity of essays that they believed might have been entirely taken from the internet. Nevertheless, they scored these essays and generally assigned high scores similar to ChatGPT. In one exception, both the human rater and ChatGPT assigned a score of zero to a suspiciously well-written essay due to the lack of textuality, such as weak cohesion, unclear references, and poor narrative flow (see text id 165 in the data set and additional explanation in the supplementary materials). Nevertheless, performances flagged as unusual by raters did not result in substantial disagreement between ATES and HS.

Our second research question was mainly focused on the impact of rater effects. Our results agreed with those of previous studies by Kumar and Boulanger (2020), Powers et al. (2015), and Wind et al. (2017). For example, after controlling for variables in our Model 1a, the raters' experience with rating had a relatively large, partially standardized regression coefficient, with our results indicating that novice raters tended to assign lower scores than ChatGPT, whereas experienced raters assigned scores that were similar to or slightly higher than ChatGPT. The second relatively large partially standardized coefficient was observed for the raters' gender indicator; our results showed that, after controlling for variables in the model, female raters assigned on average 0.58 points lower than ChatGPT, whereas male raters assigned 0.16 points higher on average. Self-reported rater tendency, on the other hand, did not result in substantial differences after controlling for the remaining variables. Overall, our null model indicated that there was a substantial variation in score differences (i.e., ATES - HS) due to rater differences, which resulted in an ICC value of 0.32, but our predictors in the model could explain only 5 % of the total variation in score differences.

Notwithstanding the encouraging performance of ChatGPT-4o, our analysis also revealed important limitations and challenges associated with AES systems for Turkish. One concern is the model's handling of linguistic errors and non-standard writing features. We observed that ChatGPT-4o, following the rubric criteria, tended to over-penalize certain language errors, particularly morphological and lexical mistakes common to non-native Turkish writers. Whereas a human rater might recognize a minor inflectional error or an unconventional word choice and still give credit for the overall content, the AES often treated such deviations more strictly. This finding aligns with Parker et al. (2023), who found that ChatGPT's grading of student writing was generally stricter than human grading on similar tasks. In other words, the model may focus heavily on surface-level correctness (grammar, spelling, word form) and accordingly assign lower scores for essays containing non-native patterns, even if the essay's ideas and structure are solid. This finding suggests that human raters retain certain strengths, such as interpreting partially correct or creative answers and understanding cultural and linguistic nuances, that AES cannot yet replicate in the scoring process.

In summary, the zero-shot ChatGPT-4o scoring approach shows promise for evaluating Turkish essays quickly and with a decent degree of agreement with human standards. It can substantially improve scoring consistency and alleviate the workload and variability issues associated with human rating, especially in large-scale or formative assessment settings. At the same time, our findings underscore the irreplaceable value of human evaluators in language assessment. Until AI systems can reliably handle those aspects, a prudent path forward is to use AES as a complementary tool rather than a wholesale replacement. By leveraging ChatGPT-4o′s strengths (speed, consistency, standardization) and coupling them with human expertise (judgment, cultural insight, ethical oversight), educational institutions can move toward more reliable and fair writing assessments. To our knowledge, this study is one of the first attempts to utilize AI for ATES in Turkish; thus, we made our data set and analysis syntax available to interested readers to support growth in the field.

5.1. Limitations and future directions

This study has several limitations that should be considered when interpreting its results. First, we considered a single human-rater score as the gold standard to investigate the performance of ATES simply due to reducing the cost. Although the human raters involved in this study were well-trained professionals, previous studies showed that individual raters, despite extensive training, may exhibit significant variations in their scoring patterns due to personal biases, inconsistent application of scoring criteria, and cognitive limitations in processing complex written responses (Cohen, 2017; Kayapınar, 2014). Therefore, further studies should employ more than one human-rater score to inspect reliability in scoring Turkish essays. Also, our study involved a limited number of indicators for quantifying self-reported rater tendency. Future studies should also consider model-based tendency indicators (e.g., DeCarlo, 2005).

Second, this study focused on a single, semi-guided writing task in Turkish. However, AES systems are capable of evaluating a broader spectrum of writing tasks, which can provide a more comprehensive assessment of their effectiveness. For instance, AES can be applied to holistic writing tasks, where essays are scored based on their overall quality (e.g., Doi et al., 2024). Additionally, AES is widely used for analytic writing tasks, which involve scoring essays across multiple dimensions such as grammar, organization, and content (e.g., Sun & Wang, 2024). Beyond these, AES models have also been developed for other genres and formats, including argumentative, narrative, and expository essays, as well as for tasks that require the integration of discourse features or the assessment of grammatical variety and errors. The adaptability of AES methods to different essay genres and educational levels further underscores the importance of expanding research beyond a single task type. Therefore, future studies should investigate the application of AES to a variety of writing tasks to obtain a more robust evaluation of AES models in Turkish.

Lastly, while this study incorporated all collected texts into its performance analysis, it did not specifically investigate the cases flagged as unusual. These outliers warrant dedicated scrutiny to assess potential plagiarism or anomalies in writing patterns. Notably, both HS and ATES systems assigned comparable scores to these atypical texts, suggesting either consistent detection of shared irregularities or limitations in both systems' ability to identify unconventional content. We acknowledge a methodological trade-off: including these unusual texts preserves ecological validity by reflecting real-world assessment conditions, where anomalous submissions may occur. Nevertheless, future research should prioritize (1) developing robust plagiarism-detection frameworks for AES systems (Amirzhanov et al., 2025) and (2) examining how atypical writing patterns affect algorithmic scoring reliability. Such investigations could refine anomaly-handling protocols while maintaining the integrity of automated assessment in authentic educational contexts (Hussein et al., 2019).

5.2. Conclusion

To take the AES tools for Turkish and other languages to the next level, we invite researchers interested in scoring Turkish inputs to test our findings' reproducibility and take advantage of our insights regarding the possible explanations for the disagreement between the ATES and HS. We also invite researchers to investigate LLMs' performance for AES in standardized settings for different languages similar to Jung et al. (2025) and Firoozi et al. (2024) as these authors recently reported that LLMs might perform differently for less complex languages or for well-resourced languages due to differences in training opportunities and translation quality.

We want to conclude our study with a cautionary note. While AI-powered AES tools offer promising opportunities for streamlining assessment, these technologies are not yet fully mature for classroom use. The effective and responsible deployment of AES tools requires a thorough understanding of the underlying AI models, their functionalities, and their inherent limitations. Currently, there are significant variations among available AI models and tools, each with different capabilities, prompt requirements, and performance characteristics. Many teachers may not have the expertise needed to select, configure, or interpret the outputs of these systems appropriately. Inadequate understanding or misuse of AES tools may lead to unreliable or biased evaluations, which can have unintended consequences for learners. Therefore, we caution practitioners against adopting AI-based automated scoring as a straightforward, ready-to-use classroom solution without sufficient training and consideration of these complexities.

CRediT authorship contribution statement

Burak Aydın: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Tarık Kışla:** Writing – original draft, Software, Methodology. **Nursel Tan Elmas:** Writing – original draft, Resources, Project administration, Data curation. **Okan Bulut:** Writing – review & editing, Validation.

Ethical considerations

The data used in this study were collected by the Yunus Emre Institute. The institute confirms that (a) all procedures were performed in compliance with relevant laws and institutional guidelines; (b) the study was approved by the appropriate the institutional committee, (c) informed consent was obtained from human participants. The Institute assumes full responsibility for all ethical procedures related to the collection, use, and sharing of these data.

Consent for publication

All authors have reviewed and approved the final manuscript and consent to its publication in the System.

Availability of data and materials

We share all materials as supplementary.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly in order to improve readibility. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding information

This work is supported by Tübitak, 123K835. This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg.

Competing interests

We do not have any competing interests.

Acknowledgments

N/A.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2025.103784.

References

Abdul Salam, M., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *PLoS One, 17*(8), Article e0272269. https://doi.org/10.1371/journal.pone.0272269

Ahmadi, A. (2019). A study of raters' behavior in scoring L2 speaking performance: Using rater discussion as a training tool. *Issues in Language Teaching, 8*(1), 195–224. https://doi.org/10.22054/ilt.2020.49511.461

Allam, H., Dempere, J., Akre, V. L., Parakash, D., Mazher, N., & Ahamed, J. (2023). Artificial intelligence in education: An argument of Chat-GPT use in education. In *9th international conference on information technology trends (ITT), Dubai, United Arab Emirates* (pp. 151–156). https://doi.org/10.1109/ITT59889.2023.10184267, 2023.

Amirzhanov, A., Turan, C., & Makhmutova, A. (2025). Plagiarism types and detection methods: A systematic survey of algorithms in text analysis. *Frontiers of Computer Science, 7*, Article 1504725. https://doi.org/10.3389/fcomp.2025.1504725

Arslan, S. (2025). English-as-a-foreign language university instructors' perceptions of integrating artificial intelligence: A Turkish perspective. *System, 131*, Article 103680. https://doi.org/10.1016/j.system.2025.103680

Arslan, A., & Özdamar, N. (2019). Yükseköğretimde programlama derslerine Yönelik Bir Otomatik Ödev Notlandırma Sistemi Önerisi. *Ege Eğitim Teknolojileri Dergisi, 3*(2), 42–51.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bexte, M., Horbach, A., & Zesch, T. (2024). Strengths and weaknesses of automated scoring of free-text student answers. *Informatik-Spektrum, 47*, 78–86. https://doi.org/10.1007/s00287-024-01573-z

Board for Medical Specialties. (2024). Yabancı uyruklu uzmanlık eğitimi öğrencilerinin girmesi gereken Türkçe dilbilgisi sınavı hakkında. Retrieved from https://tuk.saglik.gov.tr/TR-30657/.

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education & Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2363901

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies, 30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wison, J., Yildirim-Erbalsi, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation, 5*(3). https://doi.org/10.59863/MIQL7785

Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2024). Responsible AI for test equity and quality: The duolingo English test as a case study. *arXiv Preprint arXiv:2409.07476*.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In *Handbook of automated essay evaluation* (pp. 55–67). Routledge.

Cetin, M. A., & Ismailova, R. (2019). Assisting tool for essay grading for Turkish language instructors. *MANAS Journal of Engineering, 7*(2), 141–146.

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024)* (pp. 309–315).

Chapelle, C. A. (2025). Generative AI as game changer: Implications for language education. *System, 132*, Article 103672. https://doi.org/10.1016/j.system.2025.103672

Cohen, Y. (2017). Estimating the intra-rater reliability of essay raters. *Frontiers in Education, 2*, Article 279593. https://doi.org/10.3389/feduc.2017.00049

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "true" scores. *Applied Measurement in Education, 31*(3), 241–250. https://doi.org/10.1080/08957347.2018.1464450

Council of Europe, Modern Languages Division. (2001). Common European framework of reference for languages. *Learning, teaching, assessment*. Cambridge University Press.

Council of Higher Education. (2023). Yükseköğretimde uluslararasılaşma ve Türkiye'deki üniversitelerin uluslararası görünürlüğü çalıştay raporu. Retrieved from https://www.yok.gov.tr/Documents.
Council of Higher Education. (2024). Yurt dışından öğrenci kabulüne ilişkin esaslar. Retrieved from https://www.yok.gov.tr/ogrenci.
Council of Higher Education. (2025). Yükseköğretimde uluslararasılaşma strateji belgesi (2024–2028). Retrieved from https://www.yok.gov.tr/Documents.
DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53–76. https://doi.org/10.1111/j.0022-0655.2005.00004.x
Demir, D., & Genç, A. (2019). Academic Turkish for international students: Problems and suggestions. *Journal of Language and Linguistic Studies, 15*(1), 34–47. https://doi.org/10.17263/jlls.547601
DiCerbo, K., Lai, E., & Matthew, V. (2020). Assessment design with automated scoring in mind. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 29–48). Chapman and Hall/CRC.
Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th international conference on educational data mining* (pp. 103–113). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115784 (IEDMS).
Doi, K., Sudoh, K., & Nakamura, S. (2024). Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. *ArXiv*. https://arxiv.org/abs/2406.08817.
Elmas, T. N. (2025). *Setting standards for measurement and evaluation of language competencies: A case study on the Turkish as a foreign language examination (doctoral dissertation)*. Ankara: Gazi University, Graduate School of Educational Sciences.
Elosua, P. (2022). Validity evidence for scoring procedures of a writing assessment task. A case study on consistency, reliability, unidimensionality and prediction accuracy. *Assessing Writing, 54*, Article 100669. https://doi.org/10.1016/j.asw.2022.100669
Feng, H., Du, S., Zhu, G., Zou, Y., Phua, P. B., Feng, Y., Zhong, H., Shen, Z., & Liu, S. (2024). Leveraging large language models for automated Chinese essay scoring. In A. M. Olney, I. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 454–467). Nature Switzerland: Springer.
Firoozi, T., Bulut, O., & Gierl, M. (2023). Language models in automated essay scoring: Insights for the Turkish language. *International Journal of Assessment Tools in Education, 10*(Special Issue), 149–163. https://doi.org/10.21449/ijate.1394194
Firoozi, T., & Gierl, M. (2024). Scoring essays written in Persian using a transformer-based model: Implications for multilingual AES. In M. D. Shermis, & J. Wilson (Eds.), *The routledge international handbook of automated essay evaluation* (pp. 55–77). Routledge.
Firoozi, T., Mohammadi, H., & Gierl, M. (2024). Using automated procedures to score educational essays written in three languages. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12406
Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 1–10). Chapman and Hall/CRC.
Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*(10), 950–962. https://doi.org/10.1111/medu.12517
Gomaa, W. H., & Fahmy, A. (2014). Arabic short answer scoring with effective feedback for students. *International Journal of Computer Application, 86*(2), 35–41.
Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002
Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education, 8*, Article 1272229. https://doi.org/10.3389/feduc.2023.1272229
Hamner, B., & Frasco, M. (2018). Metrics: Evaluation metrics for machine learning. https://CRAN.R-project.org/package=Metrics.
Hamner, B., & Shermis, M. D. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. https://doi.org/10.4324/9780203122761.CH19.
Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, Article e208. https://doi.org/10.7717/peerj-cs.208
Hyland, K. (2003). Second language writing. *Cambridge language education*. Cambridge University Press.
Ifenthaler, D. (2022). Automated essay scoring systems. In O. Zawacki-Richter, & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1–15). Springer.
Ince, E. Y., & Kutlu, A. (2021). Web-based Turkish automatic short-answer grading system. *Natural Language Processing Research, 1*(3–4), 46–55. https://doi.org/10.2991/nlpr.d.210212.001
Jamieson, A. R., Holcomb, M. J., Dalton, T. O., Campbell, K. K., Vedovato, S., Shakur, A. H., Kang, S., Hein, D., Lawson, J., Danuser, G., & Scott, D. J. (2024). Rubrics to prompts: Assessing medical student post-encounter notes with AI. *NEJM AI, 1*(12), Article AIcs2400631. https://doi.org/10.1056/AIcs2400631
Janhunen, J. A. (2023). The Unity and diversity of altaic. *Annual review of linguistics, 9*(1), 135–154. https://doi.org/10.1146/annurev-linguistics-030521-042356
Jung, J., Tyack, L., & Von Davier, M. (2025). Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2025.100375
Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*. https://doi.org/10.3389/feduc.2024.1399377
Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System, 123*, Article 103332. https://doi.org/10.1016/j.system.2024.103332
Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *European Journal of Educational Research, 57*, 113–136. https://doi.org/10.14689/ejer.2014.57.2
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.
Kumar, V. S., & Boulanger, D. (2020). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education, 31*(3), 538–584. https://doi.org/10.1007/s40593-020-00211-5
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13
Langenfeld, T., Burstein, J., & von Davier, A. A. (2022). Digital-first learning and assessment systems for the 21st century. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.857604
Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-scale Assessments in Education, 6*(8). https://doi.org/10.1186/s40536-018-0061-2
Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence, and significance within the bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541
McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet rasch modelling. *BMC Medical Education, 6*(42). https://doi.org/10.1186/1472-6920-6-42
McNamara, D. S., & Potter, A. (2024). The two u's in the future of automated essay evaluation: Universal access and user-centered design. In M. D. Shermis, & J. Wilson (Eds.), *The routledge international handbook of automated essay evaluation* (pp. 590–608). Routledge.
MCTRT. (2024). Ministry of culture and tourism of the republic of Türkiye — Language. https://www.ktb.gov.tr.
OpenAI. (2023). Gpt-4 technical report. *arXiv Preprint arXiv:2303.08774*.
OpenAI. (2024). OpenAI O1 system card. *Open*. https://cdn.openai.com/o1-system-card-20241205.pdf.
Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243. http://www.jstor.org/stable/20371545.
Palermo, C., & Wibowo, A. (2024). Automated essay evaluation at scale: Hybrid automated scoring/hand scoring in the summative assessment program. In M. D. Shermis, & J. Wilson (Eds.), *The routledge international handbook of automated essay evaluation* (pp. 23–39). Routledge.

Parker, J. L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721–727. https://doi.org/10.3928/01484834-20231006-02

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. New York: Springer. https://doi.org/10.1007/b98882

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard.". *Applied Measurement in Education, 28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ratna, A. A. P., Wulandri, N. A., Kaltsum, A., Ibrahim, I., & Purnamasari, P. D. (2019). Answer categorization method using k-Means for Indonesian language automatic short answer grading system based on latent semantic analysis. In *16th international conference on Quality in Research (QIR): International symposium on electrical and computer engineering* (pp. 1–5). IEEE.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*(2), 109–121. https://doi.org/10.1016/0889-4906(90)90002-T

Ribeiro, E., Mamede, N., & Baptista, J. (2024). Exploring the automated scoring of narrative essays in Brazilian Portuguese using transformer models.". In *Proceedings of the 16th international conference on computational processing of Portuguese, 2* pp. 14–17).

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high-(but not low-) resource languages. arXiv 2309.07423 https://doi.org/10.48550/arXiv.2309.07423.

Salim, H. R., De, C., Pratamaputra, N. D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. *Bulletin of Electrical Engineering and Informatics, 11*(3), 1586–1603. https://doi.org/10.11591/eei.v11i3.3531

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2024). Can AI grade your essays? A comparative analysis of large Language models and teacher ratings in multidimensional essay scoring. *arXiv Preprint arXiv:2411.16337*.

Shermis, M. (2020). International applications of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 113–132). Chapman and Hall/CRC.

Shermis, M. (2024). Using ChatGPT to score essays and short-form constructed responses. https://arxiv.org/abs/2408.09540.

Shermis, M., & Wilson, J. (2024). *The routledge international handbook of automated essay evaluation*. Taylor & Francis.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage.

Sun, K., & Wang, R. (2024). Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. *ArXiv*. https://arxiv.org/abs/2406.01198.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 conference on empirical methods in natural language processing*. https://doi.org/10.18653/v1/d16-1193

Tohma, K., & Kutlu, Y. (2020). Challenges encountered in Turkish natural language processing studies. *Natural and Engineering Sciences, 5*(3), 204–211. https://doi.org/10.28978/nesciences.833188

Türkiye Scholarships. (2025). Türkiye scholarships. Retrieved from https://www.turkiyeburslari.gov.tr.

UNICEF. (2023). Türkiye'de geçici koruma altında Olan suriyeli çocuklara yönelik eğitim müdahalesinin belgelenmesi-nihai rapor. Retrieved from https://www.unicef.org/turkiye/raporlar.

Uysal, İ., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 28–53. https://doi.org/10.21031/epod.817396

von Davier, A. A., & Burstein, J. (2024). AI in the assessment ecosystem: A human-centered AI perspective. In P. Ilic, I. Casebourne, & R. Wegerif (Eds.), *Intelligent systems reference library: 261. Artificial intelligence in education: The intersection of technology and pedagogy* (pp. 93–109). Springer. https://doi.org/10.1007/978-3-031-71232-6_6.

Wei, T., & Simko, V. (2024). R package' corrplot': Visualization of a correlation matrix. https://github.com/taiyun/corrplot.

Wind, S. A., Wolfe, E. W., Engelhard, G., Foltz, P., & Rosenstein, M. (2017). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing, 18*(1), 27–49. https://doi.org/10.1080/15305058.2017.1361426

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

Xu, W., Mahmud, R., & Lam Hoo, W. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access, 12*, 77639–77657. https://doi.org/10.1109/access.2024.3399163

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System, 124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494