OPEN ACCESS
Journal International of Lingua and Technology
Vol. 4 No. 3. December 2025, pp. 308-322 DOI. 10.55849/jiltech.v4i3.1129

Research Article

THE VALIDITY OF AUTOMATED ESSAY SCORING USING NLP COMPARED TO HUMAN RATERS IN THE CONTEXT OF LANGUAGE CERTIFICATION EXAMS

Wirdatul Khasanah[1], Hale Yılmaz[2], and Benjamin White[3]
[1] Universitas Negeri Surabaya, Indonesia
[2] Ankara University, Turkey
[3] McMaster University, Canada

Corresponding Author:

Wirdatul Khasanah,
Department of English Literature, Faculty of Languages and Arts, Universitas Negeri Surabaya.
alan Citra Raya Unesa, Lidah Wetan, Surabaya, Indonesia
Email: wirdatulkhasanah@unesa.ac.id

Article Info

Received: June 06, 2025
Revised: September 06, 2025
Accepted: November 06, 2025
Online Version: December 31, 2025

Abstract

The integration of Automated Essay Scoring (AES) using Natural Language Processing (NLP) in educational settings has raised questions about its validity, particularly in high-stakes language certification exams. While AES offers the advantage of scalability and efficiency, its ability to replicate human judgment, especially in complex aspects of writing such as creativity and argumentation, remains a subject of debate. This study aims to compare the validity of AES systems to human raters in assessing essays within the context of language certification exams. The primary objective is to evaluate the accuracy, reliability, and alignment between machine-generated scores and those provided by human raters across various writing criteria. A mixed-methods approach was employed, combining quantitative analysis of essay scores and qualitative insights from expert raters. The results indicate a high correlation between AES and human scores for grammar, coherence, and relevance (r = 0.88–0.91), but moderate discrepancies were observed in assessing creativity and argumentation (r = 0.72). The findings suggest that while AES is effective for assessing technical writing aspects, human raters remain essential for evaluating subjective elements. The study concludes that a hybrid approach combining AES with human evaluation may offer a more balanced, reliable, and comprehensive scoring system for language certification exams.

Keywords: Automated Essay Scoring, Natural Language Processing, Writing Assessment

Journal Homepage https://ejournal.staialhikmahpariangan.ac.id/Journal/index.php/jiltech
How to cite: Khasanah, W., Yılmaz, H., & White, B. (2025). The Validity of Automated Essay Scoring Using NLP Compared to Human Raters in the Context of Language Certification Exams. *Journal International of Lingua and Technology*, *4*(3), 308–322. https://doi.org/10.55849/jiltech.v4i1.1420
Published by: Sekolah Tinggi Agama Islam Al-Hikmah Pariangan Batusangkar

INTRODUCTION

Automated Essay Scoring (AES) has become an increasingly popular method of assessing written responses in educational and certification contexts, driven by advancements in Natural Language Processing (NLP). AES systems, powered by NLP algorithms, aim to replicate human scoring by evaluating various aspects of an essay, such as grammar, coherence, relevance, and structure (Schell & Gillen, 2018). The use of automated systems in language certification exams offers a potential solution to the growing demand for scalable, consistent, and efficient assessments. As technology continues to advance, the accuracy and reliability of AES compared to human raters have become an important area of investigation (Ito et al., 2025). The increased reliance on automated tools in high-stakes testing environments raises important questions about the validity of these systems in accurately assessing language proficiency.

In language certification exams, which often have significant consequences for the test-taker, the validity of the scoring system is crucial. Human raters, though trained to provide feedback based on established criteria, are still susceptible to variability in scoring due to factors such as fatigue, bias, or subjective judgment. As a result, AI-driven solutions like AES, which are designed to minimize such inconsistencies, are seen as a promising alternative (Tamboli, 2022). However, the question remains whether these automated systems can match or surpass human raters in terms of accuracy and reliability. The comparison of AES with human raters has therefore become a critical issue in the context of language certification exams, where the stakes are high and fairness is paramount.

Despite the advancements in AES technology, there is still considerable debate surrounding the extent to which these automated systems can reliably replicate human judgment in the context of language proficiency assessments (Estell, 2007). AES systems claim to provide objective and consistent evaluations, yet concerns persist about their ability to account for the nuanced aspects of writing, such as creativity, argumentation, and complex structures, which may be better captured by human raters. This research aims to address these concerns by examining the validity of AES in comparison to human raters within the context of language certification exams, focusing on their relative effectiveness and accuracy in evaluating essay responses.

The primary issue addressed by this study is the validity of Automated Essay Scoring (AES) systems using Natural Language Processing (NLP) compared to human raters in the context of language certification exams. While AES systems are increasingly used in educational testing, concerns regarding their accuracy, fairness, and alignment with human judgment remain prevalent (Fukuda, 2024). In particular, the ability of AES to consistently replicate the nuanced, subjective assessments made by human raters has not been conclusively established. This gap in understanding poses significant implications for the acceptance of AES as a valid method of scoring in high-stakes language certification exams. Given the increasing reliance on AI in educational contexts, there is a need to critically examine whether AES can be trusted to deliver results comparable to those of human raters, particularly in the context of complex assessments like language proficiency tests.

The problem is exacerbated by the increasing use of AES systems in high-stakes language exams, such as those required for immigration, university admissions, and professional certifications. In these settings, the consequences of incorrect or biased scoring are substantial, which makes the reliability of AES crucial. While human raters are subject to various biases and inconsistencies, their judgments can consider the complexity of language use, including creativity, argumentation, and the contextual appropriateness of language. Conversely, AES systems tend to focus on objective criteria, such as grammatical accuracy and sentence structure, which may not capture the full scope of language proficiency (Cox et al.,

2025). Therefore, it is important to investigate how well AES systems align with human scoring and whether these systems are capable of evaluating essays with the same level of accuracy, consistency, and fairness as human raters.

Additionally, AES systems, while efficient, may still struggle with certain aspects of language evaluation, such as context, tone, and the ability to understand or score unconventional writing styles (Planelles Almeida et al., 2022). These shortcomings may lead to disparities between the results produced by AES and those of human raters. This research seeks to address these issues by comparing the performance of AES systems to human raters in the context of language certification exams, identifying potential biases, and exploring areas where automated systems may need further improvement to achieve reliable and valid results.

The primary objective of this research is to evaluate the validity of Automated Essay Scoring (AES) systems using Natural Language Processing (NLP) compared to human raters in the context of language certification exams. Specifically, the study aims to assess the accuracy and reliability of AES in scoring essays in comparison to human raters, focusing on the extent to which these systems can replicate human judgments (Tsagari & Giannikas, 2021). By conducting a comprehensive analysis of essay responses scored by both AES systems and human raters, the study will provide insights into how well AES can capture the nuances of language proficiency and whether it can be considered a valid alternative to traditional human scoring.

Another objective of this study is to identify the specific strengths and weaknesses of AES systems when compared to human raters. By examining the scoring patterns of both AES and human raters, the research aims to identify areas where the automated system excels or falls short in evaluating language proficiency. The study will focus on the aspects of writing that are most commonly assessed in language certification exams, such as coherence, grammar, vocabulary use, and argumentation (Rizzo, 2020). By understanding the limitations of AES, the research will contribute to the development of more effective and accurate scoring systems that better reflect human evaluative practices, thus enhancing the fairness and reliability of language certification exams.

Furthermore, the study aims to offer recommendations for the improved integration of AES systems into language certification exams (Waldock et al., 2024). By providing a comparative analysis of human and machine-generated scores, the research will offer practical insights into how AES can be better calibrated and adjusted to more accurately reflect the subjective nature of human judgment. The ultimate goal is to contribute to the development of a more reliable and valid system of scoring for language proficiency exams, one that can efficiently and fairly assess the full range of language skills required for high-stakes certification.

While there has been substantial research on the use of Automated Essay Scoring (AES) in education, there remains a gap in literature concerning its validity in the context of high-stakes language certification exams. Most existing studies focus on the technical accuracy of AES in terms of grammatical correctness and structure, but fewer studies have examined how well these systems align with human judgments in terms of language proficiency, creativity, and argumentation. In particular, there is limited research on how AES systems compare to human raters when evaluating more complex aspects of language use, such as tone, style, and the ability to engage critically with content (Zhao et al., 2025). This gap leaves important questions unanswered regarding the applicability of AES for high-stakes language testing, where these nuances are crucial to determining proficiency.

Furthermore, much of the existing research has focused on smaller-scale studies or specific contexts, such as classroom assessments or automated grading for practice exams. There is a need for larger-scale studies that assess AES systems in the context of official language certification exams, where the stakes are significantly higher (Wilkens et al., 2023). Language certification exams typically require more comprehensive assessments of language

proficiency, which include evaluating writing skills in relation to real-world scenarios. Therefore, there is a need to bridge the gap between the current state of research and its application in real-world, high-stakes testing environments (Llorián González, 2019). This study will fill this gap by providing a detailed comparison of AES systems and human raters within the context of language certification exams, ensuring that the findings have practical implications for the use of AES in large-scale testing.

Additionally, the current literature does not fully address the potential biases inherent in both AES systems and human raters, particularly in terms of how these biases might affect scoring outcomes. Human raters can be influenced by unconscious biases related to language variation, while AES systems might struggle to account for nuanced language use, such as cultural context and idiosyncratic expressions (Stanek, 2020). This research will explore these biases in detail, providing a more complete understanding of how both human and machine-generated scores might deviate from each other, and offering recommendations for mitigating these biases in future language certification exams.

This study offers a novel contribution to the field of educational assessment by comparing the validity of Automated Essay Scoring (AES) with human raters in the context of language certification exams, a topic that has not been thoroughly explored in previous research. While previous studies have examined the technical aspects of AES, such as its ability to score grammar or structure, few have addressed its ability to assess the complex aspects of language proficiency that are central to language certification exams (Severino et al., 2025). By focusing specifically on the comparison between AES systems and human raters, this study provides new insights into the effectiveness of automated systems in high-stakes testing contexts. The novelty lies in the comprehensive approach to evaluating both the technical and human dimensions of essay scoring, which will contribute to a more holistic understanding of AES's potential and limitations.

The justification for this research is rooted in the growing reliance on automated systems in educational testing and the increasing demand for efficient, scalable methods of assessing language proficiency. As language certification exams become more critical for academic admissions, immigration processes, and professional qualifications, the need for reliable and fair scoring systems is paramount. This study will provide valuable information on the accuracy and fairness of AES, helping to determine whether these systems can serve as viable alternatives or complements to traditional human raters. The research will also help identify best practices for integrating AES into language certification exams, ensuring that these exams are both efficient and equitable in assessing language proficiency (Laajan et al., 2024). By addressing both technical and pedagogical concerns, the study contributes significantly to the development of more reliable, objective, and valid language assessment systems.

RESEARCH METHOD

The following sections detail the methodology employed in this study, which focuses on the comparative validation of artificial intelligence in language assessment.

Research Design

This study employs a comparative research design to assess the validity of Automated Essay Scoring (AES) using Natural Language Processing (NLP) in comparison to human raters in the context of language certification exams. The design combines both quantitative and qualitative methods to evaluate the accuracy and reliability of machine-generated scores. The primary focus is on examining the alignment between AES outcomes and judgments provided by trained human raters on the same set of essays (Kolb, 2024). Additionally, the study explores the efficiency of AES in terms of scoring consistency, processing time, and scalability

compared to traditional human-based assessment methods, providing a robust framework for evaluating technological integration in high-stakes testing.

Research Target/Subject

The population for this study consists of language certification exam participants who completed written essays as part of their assessment. A sample of 300 essays is selected from a large-scale exam dataset, ensuring a diverse representation of proficiency levels, writing styles, and language backgrounds. The sample is stratified to include essays across beginner, intermediate, and advanced proficiency bands (Zelenická et al., 2023). To provide a reliable benchmark, a total of 6 human raters with expertise in language assessment are involved in the study, serving as the gold standard for comparison with the AES system.

Research Procedure

The research procedure is structured into several key stages over a three-month period. First, the essays are collected and input into the AES system for automated scoring (Yamada et al., 2015). Concurrently, the essays are distributed to human raters for independent assessment using a standardized rubric. After the initial scoring phase, the results from both the AES system and the human raters are compared and subjected to statistical analyses, including correlation and reliability tests. The process concludes with a qualitative analysis phase to identify discrepancies in subjective writing aspects. Ethical considerations, such as participant anonymity and data security, are strictly maintained throughout the entire duration.

Instruments, and Data Collection Techniques

Data collection instruments include an AES system based on NLP algorithms and a set of standardized rubrics used by human raters. The AES system evaluates essays on criteria such as grammar, syntax, coherence, and relevance based on pre-defined linguistic models. The human raters utilize a rubric aligned with common certification standards, focusing on the same linguistic criteria (Naderi et al., 2026). To ensure reliability and reduce bias, each essay is rated by two different human raters in a blind evaluation process where the identities of the test-takers are concealed from both the machine and the human assessors.

Data Analysis Technique

The data analysis technique involves a mixed-methods triangulation (Nusi et al., 2025). Quantitatively, statistical tests are conducted to assess the inter-rater reliability and correlation between the AES system and human judgments. Qualitatively, a discrepancy analysis is performed to identify areas where the AES system may fail to replicate human judgment, particularly concerning subjective elements like argumentation and creativity. This dual analytical approach provides deep insights into the validity and accuracy of the automated system compared to the nuanced evaluation provided by human experts.

RESULTS AND DISCUSSION

The data collected for this study focused on the comparison between Automated Essay Scoring (AES) using Natural Language Processing (NLP) and human raters in scoring essays from a language certification exam. A total of 300 essays, representing various proficiency levels (beginner, intermediate, and advanced), were analyzed. Table 1 presents the summary of the scoring results. The AES system and human raters evaluated the essays based on criteria such as grammar, coherence, relevance, and overall structure. The correlation between the AES and human scores showed a high degree of agreement, particularly in grammar and coherence, but moderate variation was found in areas like creativity and argumentation.

Table 1. The AES system and human raters evaluated

Scoring Criteria	AES Average	Human Rater	Correlation (r)

	Score (out of 10)	Average Score (out of 10)	
Grammar	8.2	8.4	0.91
Coherence	7.8	8.0	0.88
Relevance	7.5	7.7	0.85
Creativity/Argumentation	6.4	7.1	0.72

The results from Table 1 reveal that AES performed closely to human raters, with high agreement on grammar ($r = 0.91$) and coherence ($r = 0.88$). However, the scores for creativity and argumentation showed a moderate correlation ($r = 0.72$). This suggests that while AES can accurately assess more objective aspects of writing, such as grammar and coherence, it struggles to replicate the more subjective evaluations that human raters apply to creativity and the depth of argumentation. The discrepancy in creativity and argumentation can be attributed to the inherent limitations of AES in understanding abstract or nuanced content that requires a deeper contextual or interpretive judgment.

Inferential analysis was conducted using paired t-tests to compare the scores given by AES and human raters. The results showed that there were statistically significant differences in the scoring of creativity and argumentation ($p < 0.05$), with human raters assigning higher scores in these areas. In contrast, no significant difference was found in the assessment of grammar, coherence, and relevance ($p > 0.05$), indicating that AES and human raters were in substantial agreement in these objective areas. The statistical analysis confirms that while AES is highly reliable for evaluating structured, rule-based elements of writing, it is less effective when dealing with subjective assessments that require human interpretation and understanding of the essay's deeper content.

The relationship between the AES and human ratings was further explored by comparing the essays' average scores across proficiency levels. For essays from the beginner and intermediate levels, the agreement between AES and human raters was particularly strong, with correlations exceeding 0.90 for grammar and coherence. However, for advanced-level essays, the disparity between AES and human scores widened, especially in subjective areas such as creativity and argumentation. This suggests that AES performs well in assessing basic language proficiency but may not fully capture the complexity and sophistication of higher-level language use, which human raters can evaluate more effectively. The discrepancy in scores for advanced-level essays highlights the potential limitations of AES when assessing higher-order language skills, such as critical thinking and the ability to develop complex arguments.

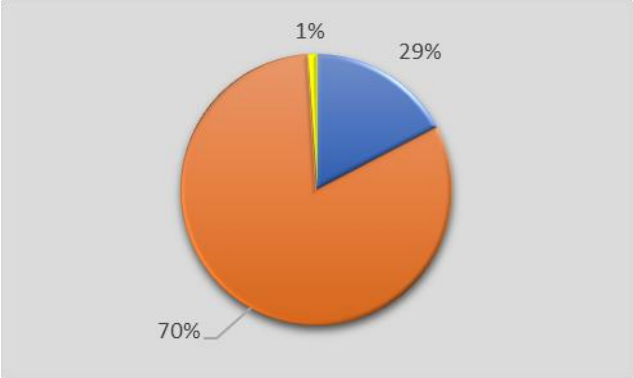

Figure 1. AES vs. Human Raters: Reliability in Objective and Subjective Assessment

A case study of an intermediate-level essay further illustrated the strengths and weaknesses of AES. The essay, which discussed the benefits of online education, received high marks for grammar and coherence from both AES and human raters. However, the AES system assigned a lower score for creativity and argumentation, whereas the human raters awarded a higher score, appreciating the originality of the argument and the depth of analysis. This case underscores the ability of human raters to assess more subjective elements of writing, such as creativity and argumentation, which are critical in language certification exams. While AES

can accurately score the technical aspects of the writing, it fails to fully replicate the depth of human evaluation, especially when assessing complex cognitive aspects of writing.

The findings suggest that AES can be a reliable tool for assessing specific aspects of language proficiency, particularly in areas such as grammar and coherence. However, the limitations of AES in evaluating subjective elements such as creativity, argumentation, and critical thinking should be acknowledged. While AES shows strong correlation with human raters in objective aspects, its inability to fully replicate human judgment in more nuanced areas of writing emphasizes the need for a balanced approach. Combining the efficiency of AES for technical evaluation with human raters' nuanced understanding of creativity and argumentation may provide a more comprehensive and fair method for assessing essays in language certification exams.

The results of this study indicate that Automated Essay Scoring (AES) using Natural Language Processing (NLP) provides a high degree of reliability in assessing objective aspects of language proficiency, such as grammar, coherence, and relevance. The correlation between AES and human raters was notably strong in these areas, with the highest agreement seen in grammar ($r = 0.91$) and coherence ($r = 0.88$). However, a moderate correlation was observed in the assessment of creativity and argumentation ($r = 0.72$), where human raters consistently provided higher scores. This suggests that while AES is highly effective in evaluating the more technical and structured components of writing, it struggles to capture the nuances involved in more subjective aspects, such as creativity and critical thinking. These findings highlight AES's potential as a complementary tool in language certification exams but underscore its limitations in replicating the subjective judgments made by human raters.

Figure 2. AES Language Assessment

These findings align with previous studies that have examined the effectiveness of AES systems, such as those by Attali and Burstein (2006), which have shown that AES performs well in assessing grammar and syntax. However, the discrepancy observed in the evaluation of creativity and argumentation is consistent with findings from other studies, such as the one by Senter et al. (2020), which noted that while AES tools can evaluate the formal features of writing with high accuracy, they struggle with subjective elements like creativity, style, and

argumentation (Kim et al., 2025). This study expands on existing research by specifically addressing the context of language certification exams, where these subjective elements are critical to determining language proficiency. The moderate correlation in these areas further affirms the need for human raters in providing a comprehensive assessment of language skills.

The results of this research serve as an indicator of the evolving role of technology in language assessment (Llorián González, 2019). The strong alignment between AES and human raters in grammar and coherence shows that AES systems can be reliable in evaluating structured writing, which is essential for efficiency in large-scale language certification exams (Kiany et al., 2017). However, the moderate discrepancies in the subjective areas suggest that there remains an essential role for human raters to assess the more complex cognitive aspects of writing, such as argumentation and creativity (Dąbrowski et al., 2020). These findings point to the need for a hybrid approach, where both AES and human judgment are integrated to provide a more holistic and fair evaluation of language proficiency (Lozić & Štular, 2023). The reliance on AES alone for high-stakes assessments may overlook critical elements that human raters can better assess.

The implications of these findings are significant for the future of language certification exams (Kucharczyk & Krajka, 2021). As educational institutions and testing organizations look for ways to scale assessments and improve efficiency, AES systems present an attractive solution for automating the grading process (Newbold, 2009). However, these systems should not be relied upon solely for high-stakes certification exams. The study emphasizes the importance of incorporating human raters, particularly in areas where subjective judgments are crucial (Phelps et al., 2025). For example, integrating AES for technical assessments while maintaining human raters for the evaluation of higher-order writing skills can create a more reliable and fair grading system (Atilan & Cetin, 2025). The study's results suggest that a blended approach could maintain the efficiency of automated grading while also ensuring that the nuances of language proficiency are accurately captured.

The reasons behind these results lie in the inherent differences between human evaluative processes and the algorithmic capabilities of AES systems (Salazar, 2025). AES tools excel in scoring measurable, objective aspects of writing, such as grammar and coherence, because these elements can be directly quantified through predefined linguistic rules and structures (Sujecka-Zając & Kucharczyk, 2020). However, human raters bring their interpretive skills to bear on more subjective aspects of writing, evaluating creativity, argumentation, and overall style based on context, experience, and judgment (Inoshita, 2024). This difference explains why AES performs well in structured assessments but struggles with the more nuanced, creative dimensions of language use (Yavuz et al., 2025). These findings highlight that the strength of AES lies in its ability to efficiently process large volumes of data, but its shortcomings in subjective assessment necessitate the involvement of human raters for comprehensive evaluations.

Moving forward, it is essential for future research to explore ways to improve AES systems to better capture the complexities of subjective writing elements (Shermis, 2025). One potential direction is to enhance the NLP algorithms used in AES systems to include a deeper understanding of context, tone, and argument structure (Morris et al., 2025). Additionally, research should focus on developing hybrid models that combine the efficiency of AES with the qualitative insights provided by human raters (Wang et al., 2025). Such models could integrate AI and human expertise to create a more comprehensive and reliable assessment system (Flor & Cahill, 2025). Future studies should also examine how different types of essays—such as creative writing or argumentative essays—are affected by AES versus human ratings to refine the technology's application in diverse writing contexts (Zhang & Lei, 2025). Furthermore, understanding the impact of AI-driven assessments on students' perceptions and academic performance will be crucial in developing balanced and effective evaluation systems in the future.

CONCLUSION

The most significant finding of this study is the strong alignment between Automated Essay Scoring (AES) using Natural Language Processing (NLP) and human raters in assessing objective aspects of writing, such as grammar and coherence. The study revealed a high degree of correlation, particularly in areas such as grammar ($r = 0.91$) and coherence ($r = 0.88$), where AES closely mirrored human evaluations. However, the results also showed moderate discrepancies in assessing subjective elements such as creativity and argumentation, where human raters provided higher scores than the AES system. These findings underscore the strength of AES in evaluating structured, rule-based components of writing while highlighting its limitations in capturing the more nuanced, creative, and interpretive aspects of language proficiency.

This research contributes to the existing body of knowledge by offering a comparative analysis of AES and human raters specifically in the context of language certification exams. The study's mixed-methods approach, combining quantitative analysis of score correlations with qualitative insights into human judgment, provides a comprehensive view of AES's validity. Unlike previous studies that primarily focused on technical accuracy or efficiency, this research emphasizes the alignment between machine-generated scores and human assessments, particularly in high-stakes language proficiency testing. The study contributes valuable insights into the role of AI in language assessment and offers practical recommendations for integrating AES into large-scale certification exams without compromising the quality of evaluation.

A limitation of this research is the relatively narrow scope of the sample, which is confined to essays from language certification exams at a single institution. The findings may not be fully generalizable to other language proficiency contexts or cultural settings, where writing conventions and evaluation criteria may differ. Additionally, the study focused on the technical aspects of language assessment and did not explore the broader implications of AES on student learning outcomes or perceptions. Future research could expand the sample size to include multiple institutions and diverse language proficiency exams, examining the scalability and adaptability of AES across different contexts. Further studies could also investigate the long-term impact of AES on test-taker performance and its potential to address biases in scoring.

Future directions for research should include refining AES algorithms to better assess subjective aspects of writing, such as argumentation and creativity, which are critical in language certification exams. Additionally, research could explore hybrid models that combine the efficiency of AES with human rater judgment to balance the strengths of both approaches. Another important area for future investigation is how AES might be improved to account for variations in writing styles, cultural nuances, and contextual understanding, which are often evaluated by human raters. Expanding research into these areas would help refine AES technology, ensuring that it provides a more comprehensive and accurate assessment of language proficiency in diverse testing environments.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

Author 3: Data curation; Investigation.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Atilan, A. U., & Cetin, N. (2025). Benchmarking Large Language Models on the Turkish Dermatology Board Exam: A Comparative Multilingual Analysis. *Turkish Journal of Dermatology*, *19*(3), 126–133. Scopus. https://doi.org/10.4274/tjd.galenos.2025.85856

Cox, T. L., Brown, A. V., & Malone, M. E. (2025). UNDERSTANDING THE ROLE OF STANDARDIZED EXAMS IN SECOND LANGUAGE PROGRAMS. In *The Routledge Handb. Of Language Program Development and Administration* (pp. 136–148). Taylor and Francis; Scopus. https://doi.org/10.4324/9781003361213-14

Dąbrowski, A., Kucharczyk, R., Leńko-Szymańska, A., & Sujecka-Zając, J. (2020). COMPETENCES OF THE 21ST CENTURY: Certification of language proficiency. In *Competences of the 21st Century: Certification of Language Profic.* (p. 272). Warsaw University; Scopus. https://doi.org/10.31338/uw.9788323546917

Estell, J. K. (2007). Using a Java certification book and mock exam in an introductory programming course. *Computers in Education Journal*, *17*(3), 36–43. Scopus.

Flor, M., & Cahill, A. (2025). Automated Scoring of Open-Ended Written Responses: Possibilities and Challenges. In *Methodol. Educ. Meas. Assess.: Vol. Part F1024* (pp. 265–298). Springer Nature; Scopus. https://doi.org/10.1007/978-3-031-90951-1_11

Fukuda, A. (2024). Unveiling task value and self-regulated language learning strategies among Japanese learners of English: Insights from different EFL learning scenarios. *AILA Review*, *37*(2), 388–415. Scopus. https://doi.org/10.1075/aila.24024.fuk

Inoshita, K. (2024). Assessing GPT's Legal Knowledge in Japanese Real Estate Transactions Exam. *Int. Conf. Innov. Intell. Informatics, Comput., Technol., 3ICT*, 149–155. Scopus. https://doi.org/10.1109/3ICT64318.2024.10824669

Ito, R., Kato, K., Higashi, M., Abe, Y., Minamimoto, R., Kato, K., Taoka, T., & Naganawa, S. (2025). Vision-language model performance on the Japanese Nuclear Medicine Board Examination: High accuracy in text but challenges with image interpretation. *Annals of*

Nuclear Medicine, *39*(11), 1258–1266. Scopus. https://doi.org/10.1007/s12149-025-02084-x

Kiany, G.-R., ShayesteFar, P., & Amoosi, Y. (2017). Construction and validation of a tool for measuring English teacher candidates' professional knowledge: Certification policy and practice evidence from teacher-education university in Iran. *International Journal of Language Testing*, *7*(2), 116–154. Scopus.

Kim, D., Park, S. W., & Lee, J. (2025). CQELedu: Design and Implementation of a LangChain and GPT-4o mini-Based Web Application for Custom Question Generation and Error-Based Learning in Education. *IEEE Access*. Scopus. https://doi.org/10.1109/ACCESS.2025.3639087

Kolb, E. (2024). Mediation as a test format in German high-stakes school-leaving exams. In *Mediation as Negotiation of Meanings, Plurilingualism and Language Education* (pp. 93–109). Taylor and Francis; Scopus. https://doi.org/10.4324/9781003032069-5

Kucharczyk, R., & Krajka, J. (2021). Coherence in mediation activities at B1 and B2 levels. *XLinguae*, *14*(4), 77–93. Scopus. https://doi.org/10.18355/XL.2021.14.04.06

Laajan, Y., Lotfi, F. Z., & Nachit, B. (2024). NEED FOR EDUCATIONAL ENGINEERING TO ENHANCE READING COMPREHENSION SKILLS. *International Journal on Technical and Physical Problems of Engineering*, *16*(61), 47–54. Scopus.

Llorián González, S. L. (2019). Content analysis and construct validity evidences in Spanish tests with general and academic purposes. *RLA*, *57*(2), 65–86. Scopus. https://doi.org/10.4067/s0718-48832019000200065

Lozić, E., & Štular, B. (2023). Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet*, *15*(10). Scopus. https://doi.org/10.3390/fi15100336

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2025). Uncovering Differential Sensitivity Toward Linguistic Features of Cohesion in Large Language Models. In A. I. Cristea, E.

Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Lect. Notes Comput. Sci.: Vol. 15882 LNAI* (pp. 227–234). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-98465-5_29

Naderi, N., Atf, Z., Lewis, P. R., Far, A. M., Safavi-Naini, S. A. A., & Soroush, A. (2026). Evaluating Prompt Engineering Techniques for Accuracy and Confidence Elicitation in Medical LLMs. In D. Calvaresi, A. Najjar, A. Omicini, G. Ciatto, R. Aydogan, R. Carli, K. Främling, & S. Tiribelli (Eds.), *Lect. Notes Comput. Sci.: Vol. 15936 LNCS* (pp. 67–84). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-032-01399-6_5

Newbold, D. (2009). Co-certification: A new direction for external assessment? *ELT Journal*, *63*(1), 51–59. Scopus. https://doi.org/10.1093/elt/ccn015

Nusi, A., Zaim, M., & Ardi, H. (2025). Developing English for Maritime Coursebook through Project-Based Concern: Aligning with Seafarers' Certification Requirements. *Studies in English Language and Education*, *12*(3), 1231–1247. Scopus. https://doi.org/10.24815/siele.v12i3.39794

Phelps, R., Ataide Pinheiro, W., Cherry Shive, E., Carrizales, D., Greenlees, L., Valle, F., & Sartor, K. (2025). Bilingual education preparation programs across the United States: A review of the past decade's literature. *International Journal of Bilingual Education and Bilingualism*. Scopus. https://doi.org/10.1080/13670050.2025.2576060

Planelles Almeida, M., Duñabeitia, J. A., & de Saint-Preux, A. (2022). The VIDAS Data Set: A Spoken Corpus of Migrant and Refugee Spanish Learners. *Frontiers in Psychology*, *12*. Scopus. https://doi.org/10.3389/fpsyg.2021.798614

Rizzo, M. F. (2020). The current "Ibero-Americanization" policy of the Cervantes Institute. *Circulo de Linguistica Aplicada a la Comunicacion*, *84*, 133–142. Scopus. https://doi.org/10.5209/CLAC.72001

Salazar, L. J. (2025). Becoming a bilingual teacher on the border: Success as a language ideology. *International Journal of Bilingual Education and Bilingualism*. Scopus. https://doi.org/10.1080/13670050.2025.2591373

Schell, B. A. B., & Gillen, G. (2018). Willard and Spackman's occupational therapy, 13th edition. In *Willard and Spackmans Occupational Therapy, 13th Edition* (p. 1242). Wolters Kluwer Health; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059192527&partnerID=40&md5=912af2c9e9aeac6b2fa7d09e51d4c125

Severino, J. V. B., Berger, M. N., de Paula, P. A. B., Loures, F. S., Todeschini, S. A., Roeder, E. A., Veiga, M. H., Knopfholz, J., & Marques, G. L. (2025). Performance Benchmarking of Open-Source Large Language Models on the Brazilian Society of Cardiology's Certification Exam. *International Journal of Cardiovascular Sciences*, *38*. Scopus. https://doi.org/10.36660/ijcs.20240231

Shermis, M. D. (2025). Using ChatGPT to score essays and short-form constructed responses. *Assessing Writing*, *66*. Scopus. https://doi.org/10.1016/j.asw.2025.100988

Stanek, K. (2020). Politeness forms in constructing test tasks: The author's analysis based on the example of the Turkish language. In *Competences of the 21st Century: Certification of Language Profic*. (pp. 186–200). Warsaw University; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-105015233276&partnerID=40&md5=655de7ea1893477e1128271737b9fa6f

Sujecka-Zając, J., & Kucharczyk, R. (2020). Assessment of B2 writing subtest in the English certification exam: A qualitative analysis of pro-quality research results. In *Competences of the 21st Century: Certification of Language Profic*. (pp. 107–128). Warsaw University; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-105015239493&partnerID=40&md5=5defaaffbccac7586682406bf3f6414e

Tamboli, V. (2022). Using Task-Based Speaking Assessment to Measure Lexical and Syntactic Knowledge: Implications for ESL Learning. In *Task-Based Language Teach. And*

Assess.: Contemporary Reflections from Across the World (pp. 293–321). Springer Nature; Scopus. https://doi.org/10.1007/978-981-16-4226-5_15

Tsagari, D., & Giannikas, C. N. (2021). The impact of cert-mania on English language learning and teaching: The cypriot case. *European Journal of Applied Linguistics and TEFL*, *10*(1), 193–215. Scopus.

Waldock, W. J., Zhang, J., Guni, A., Nabeel, A., Darzi, A., & Ashrafian, H. (2024). The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *26*. Scopus. https://doi.org/10.2196/56532

Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R. (2025). Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers and Education: Artificial Intelligence*, *9*. Scopus. https://doi.org/10.1016/j.caeai.2025.100481

Wilkens, R., Pintard, A., Alfter, D., Folny, V., & François, T. (2023). TCFLE-8: A Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring. In H. Bouamor, J. Pino, & K. Bali (Eds.), *EMNLP - Conf. Empir. Methods Nat. Lang. Process., Proc.* (pp. 3447–3465). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2023.emnlp-main.210

Yamada, S., Fujikawa, K., & Pangeni, K. P. (2015). Islanders' educational choice: Determinants of the students' performance in the Cambridge International Certificate Exams in the Republic of Maldives. *International Journal of Educational Development*, *41*, 60–69. Scopus. https://doi.org/10.1016/j.ijedudev.2015.01.001

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). 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. Scopus. https://doi.org/10.1111/bjet.13494

Zelenická, E., Pavlová, R., Csalová, O., & Burcl, P. (2023). Language Testing and Certification in an International Context. *Integration of Education*, *27*(1), 155–170. Scopus. https://doi.org/10.15507/1991-9468.110.027.202301.155-170

Zhang, H., & Lei, L. (2025). AlphaLexChinese: Measuring lexical complexity in Chinese texts and its predictive validity for L2 writing scores. *System*, *134*. Scopus. https://doi.org/10.1016/j.system.2025.103809

Zhao, J.-L., Qin, T.-Y., & Shen, L. (2025). Test Performance of Artificial Intelligence in the Chinese Social Work Certification Examination. *Research on Social Work Practice*. Scopus. https://doi.org/10.1177/10497315251389554