Using Artificial Intelligence (AI) to create language exam questions: A case study

Joanna Kic-Drgas – Ferit Kılıçkaya

DOI: 10.18355/XL.2024.17.01.02

Abstract
This paper investigates how artificial intelligence (AI) can be used to generate classroom foreign language tests and aims at determining challenges and opportunities during item creation. The study benefited from a case study, a form of qualitative research, using data from a single subject to examine the issue of generating classroom assessment items using an AI system in its natural environment. The findings revealed that using AI posed several opportunities and challenges for the participant. The analysis of the comments indicated three significant opportunities/benefits of using AI: practicality, customization, efficiency, and four challenges: readability, validity, lack of data for other languages, and ownership.
Key words: artificial intelligence, assessment, exam questions, foreign language teaching, readability

Introduction

Assessment is an indispensable component of the training or selection system in the education and business sectors. It provides teachers and students with valuable feedback about the effectiveness of the language instruction, helps to pinpoint areas of difficulty, and is the basis for making decisions about the placement of students in different language courses (Ferrara et al., 2017; Freddi, 2021). It is also an essential part of developing and refining language curricula. Regarding additional/foreign language assessment, it can also be stated that assessment provides means for teachers not only to determine the current language proficiency of their learners but also helps them to identify where further instruction is needed by determining their strengths and weaknesses (Brown & Abeywickrama, 2010; Purpura, 2016). As such, teachers gain invaluable insight into which language skills their students have already acquired and which areas they need to focus on to enhance their learners' language abilities. However, while language assessment provides critical information, teachers are often required to spend significant time creating language items based on the syllabus and content determined by their institutions or the Ministry of Education in their country. The items used to assess learners' language skills might include multiple choice questions, reading comprehension tasks, or even spoken language activities (Hughes & Hughes, 2020), and creating these items is a labor-intensive process considering the analysis of the current and target needs of learners, which is essential for the most accurate and efficient evaluation of student language proficiency. However, the time and efforts spent by teachers can be considered an investment in their learners' future success and are mandatory to accurately assess teaching and learning to ensure appropriate instruction and support.

Technology has rapidly transformed in recent years, and mobile-based technologies have become a significant part of our lives (Çakmak, 2019; Kukulska-Hulme, 2020; Selwyn et al., 2021). From online shopping to the internet of things, the advancement of technology has made our lives easier and more efficient in several ways. In educational institutions, technology has become an integral part of the learning process, especially the use of artificial intelligence (AI) provided the new opportunities in teaching and learning practices (Kessler, 2023; Nazaretsky et al., 2022; Van Moere & Downey, 2016). Assessment is one of the important uses of

recent technology in teaching and learning (Russell & Murphy-Judy, 2021; Voss, 2018). Traditionally, assessment was used to determine learner performance, but technology has enabled more sophisticated and automated use of assessment (García-Peñalvo et al., 2021; Gardner et al., 2021). For example, with the help of computers and software, teachers can quickly assess student performance and measure their progress as well as determining weaknesses and strengths. However, despite the potential benefits of AI-powered language testing and assessment, there is currently a lack of empirical research that examines the potential risks and opportunities. Therefore, to fill this gap, this paper investigates the challenges and opportunities which educators might face when using AI to create questions for foreign language assessment.

Literature review

Artificial Intelligence

'Artificial Intelligence' can be defined as the "science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy et al., 2006: 2). AI can be defined as machine intelligence that can perform human-like tasks and activities (Eaton et al., 2021). In other words, AI is a platform programmed to think, reason, and take actions as or more than humans would do. Recent technology offers a variety of new uses in the education sector, from task design to their atomatised assessment (Gardner et al., 2021; González-Calatayud et al., 2021; Levy & Stockwell, 2006; Stephenson & Harvey, 2022). Students, teachers, and researchers are using/will use technological tools to enhance writing (Godwin-Jones, 2022; Zhang & Zou, 2022), such as Writing assistants (i.e., Grammarly), Paraphrasing tools (i.e., Quilbot), Research assistants (i.e., Elicit) and Reference and citations checkers (i.e., Reciteworks). In addition to these uses, AI is also used for automated essay scoring systems and adaptive tests, which consider test-taker abilities and arrange the number and the order of the items based on the responses given (Gardner et al., 2021), which are two core applications of AI in educational assessment. Moreover, artificial intelligence can also be of help to learners and teachers to personalize learning via personalized instruction, which might help them master content more quickly and effectively (Baker et al., 2019). The possibilities of recent technology in education are endless, and it will continue to significantly impact how we learn and teach (Kılıçkaya & Kic-Drgas, 2023).

AI and Language Test Creation

AI can create automated item creation without human intervention. This automated process can be used to create a variety of items, including but not limited to blog posts, writing summaries, and providing feedback, and has led to the rise in the use of AI in language teaching and learning (UNESCO, 2021; Yanhua, 2020). Teachers can now carry out more sophisticated tasks, such as automated grading of essays and oral proficiency for their writing and speaking classrooms (Borade & Netak, 2021; Kessler, 2023; Langenfeld et al., 2022; Yu et al., 2022; Yunjiu et al., 2022). This technology goes beyond being merely intelligent language tutors and feedback providers to offer interactions between the learner and the computer in the text and the spoken form. The field of language teaching, as with other various fields and businesses (Chatterjee et al., 2023), has the opportunity to benefit from recent technological advancements, including AI, which enables the field to utilize chatbots as well as collect and analyze large corpora not only to determine the current language use but also to investigate variations in language patterns (e.g., Çakmak, 2022; Kılıçkaya, 2020; Vermeer et al., 2019). According to Cope et al. (2020), "Assessment is perhaps the most significant area of opportunity offered by artificial intelligence for

transformative change in education" (p. 1233). This has been motivated by the need for more efficient and effective ways to assess language proficiency since these technologies can help language teachers and test creators save time and effort while creating language questions (Hinkelman, 2018).

Several studies have investigated the use of AI in generating questions automatically without or with little human intervention for assessment purposes (Killawala et al., 2018; Odilinye et al., 2015; Settles et al., 2020; Wang, 2018; Yunjiu et al., 2022; Zawacki-Richter et al., 2019; Zhai et al., 2021). Odilinye et al. (2015) aimed to develop algorithms that would automatically generate - questions for assessment purposes from the test provided as input and assess and align the questions created based on the pedagogical goals. Similar to this study, Killawala et al. (2018) developed a computational intelligence framework to benefit from AI and neural network models to generate exam questions in several formats, such as multiple-choice questions. The framework developed at the end of the study allowed authors to create true/false, Wh-type questions in addition to the multiple-choice questions based on the input provided. The framework was promising as it enabled the researchers to generate selected and productive response item questions. Wang (2018), on the other hand, investigated how the use of AI could help prepare questions. The study investigated several advantages of using artificial intelligence based on genetic algorithms and fuzzy and close matching, such as ensuring the fairness of the examinations, benefiting from more practical exams in terms of practicality, and providing a more accurate representation of the test-takers' ability. The systematic review conducted by Zawacki-Richter et al. (2019) focused on the final list of 148 articles between 2007 and 2018 that investigated the use of artificial intelligence applications. In addition to the findings on the use of AI in institutional and administrative levels, such as admission and library services, the analysis of the studies led to the finding that AI could be used in assessment and evaluation tasks with high accuracy and efficiency. However, it was noted that since the use of AI required lengthy training and machine learning processes, it was suggested that these applications based on AI be used at institutions where courses include a large number of learners.

Similarly, Settles et al. (2020) developed an online proficiency test entitled "Duolingo English Test" using the Computer Adaptive Test (CAT) administration algorithm, machine learning, and natural language processing. The content of this test was determined and created via automated item generation. The application of this test to learners indicated that the scores align with other English tests, and the questions generated by CAT were found efficient. Zhai et al.'s review study (2021) did a content analysis of a hundred manuscripts published between 2010 and 2020 investigating the use of AI in educational contexts. The findings of the study indicated that in addition to several opportunities, the use of AI led to several challenges due to inappropriate expectations or uses and overreliance on AI-integrated tools by teachers. Regarding the performance of AI and human experts in creating language assessment questions, Yunjiu et al. (2022) compared vocabulary tests AI and human experts created. The findings revealed that the items created by human experts had more discrimination power than those created by AI. In addition, AI-created items were found to be assessing different constructs by eliciting bottom-up strategies to be used, while human expert items focused on rote memorization ability.

Previous studies investigated the development and the use of algorithms and AI to generate language items as well as comparing and contrasting the items and questions created by humans and AI; however, to the best knowledge of the author, only a few studies have investigated the use of AI in creating questions for classroom assessment.

In line with this aim, the following research questions were proposed:

1. How can an application of AI support the design of foreign language exam questions in education sector?
2. What challenges did the participant encounter when using AI to create exam questions?

Methodology

Research Design

The current study is a case study, a form of qualitative research that involves an in-depth, detailed examination of a specific instance or phenomenon within its real-life context. It is aimed to examine a problem "in situ" in its natural environment, the language classroom (Creswell, 2007). The presented study has an ethnographical character (Jones & Smith, 2017) using observation as a tool. The conducted research aimed to provide an in-depth analysis of an individual's experience, feelings, and perceptions to explore, analyze, and explain complex social phenomena. Case studies allow researchers to comprehensively understand a single person, group, or event by collecting and analyzing data from multiple sources. Crucial for the study design was the meticulous analysis of collected data and the researcher's involvement in the group activities during research (Dumont, 2023). By looking at a single case in detail, researchers can uncover patterns, trends, and insights that can be applied to other cases and contexts (Griffee, 2012; Mackey & Gass, 2022). In the study described, the role of the method used is vital and innovative, as it allows a new phenomenon to be traced from the teacher's perspective. It, therefore, looks at the authentic use of AI in education.

Participant and setting

The participant for this study was also the author and the lecturer of the elective course,' Classroom Assessment,' who worked at a state university in Burdur, Turkey. This course was offered as a departmental elective course to junior students trained to be teachers of English in the English Language Teaching Programme at the Department of Foreign Language Education in the Fall semester of the 2022-2023 Academic Year. The course aimed to equip students to comprehend evaluation principles and methods related to assessing classroom learning, including traditional and alternative testing techniques. The participant with a Ph.D. in ELT was a lecturer at this department for more than 20 years and has ample experience using technology in language teaching and learning, with research projects and publications on the beneficial and harmful effects of technology in teaching and learning languages. The participant used the AI tools for several months before the study to integrate these tools into the language classroom.

Teaching materials and grading policy

Two coursebooks were determined as the reading materials for this departmental elective course: *Language assessment: Principles and classroom practices* by Brown and Abeywickrama (2010), and *Testing for language teachers* by Hughes and Hughes (2020), the second of which was determined as the required reading. The lectures included assessment topics, including the introduction of testing, assessment, measurement, and evaluation, kinds of tests and testing, validity, reliability, standard test techniques, beyond testing: other means of assessment, washback (backwash), and websites and applications for online classroom assessment. Students were required to attend 70% of the classes. All the students were required to sit for the midterm exam, which accounted for 40% of their final grades respectively. The

midterm comprised selected-response items such as multiple-choice questions and limited constructed response items such as short-answer questions, which aimed to test students' knowledge of the lectures covered until the midterm.

However, the final exam comprised two distinct tasks. Firstly, students were instructed to watch an English lesson on EBA TV geared towards either secondary or high school levels. Subsequently, based on the content and objectives presented in the lesson, students were required to prepare two sets of questions to evaluate both the content and objectives: Traditional Assessment Question: This type of question took various forms, including multiple-choice and fill-in-the-blanks. Students were asked to create questions aligned with the content they observed during the English lesson. Alternative Assessment Question: In this case, students were expected to formulate a question in the form of alternative assessment such as creating a poster or writing an essay. Following the midterm exam, students presented their proposed questions, explaining the rationale behind their choices via 15-minute presentations. During these presentations, students discussed their questions, received feedback, and engaged in constructive dialogue with their lecturer and classmates. In response to the feedback received during these presentations, students undertook a revision process for their questions. Ultimately, the students submitted the revised questions, incorporating the valuable input received as a partial requirement for the final examination in the course.

Procedure

The participant first registered for the free use of ChatGPT, the AI based on the GPT-3 (Generative Pre-trained Transformer) language model developed by OpenAI (https://chat.openai.com/chat), which is a neural network-based language model that can chat-based on the responses or the questions provided by a human. In addition to having natural language conversations with humans based on the input it receives, ChatGPT can create texts such as paragraphs and longer texts, write poems and stories based on the instruction or prompt provided without the need for human intervention except for the prompt or instruction (Hern, 2022; Perkins, 2023).

Before the exam period, this AI was used to create questions for the course content and classroom discussions. During this process, the participant took notes in his research diary regarding his experience using AI to create assessment items. Before creating the questions, the AI was asked to create short passages on the topics covered in the classroom, such as reliability, validity, and washback. For example, the following paragraph was created by OpenAI, using the prompt, "Write a paragraph on the importance of test reliability using formal language."

> *Test reliability is the degree to which a test accurately measures what it is supposed to measure. It is an important concept in psychometrics, as it is necessary for a test to be reliable in order for the results to be considered valid. Test reliability is determined by the consistency of the test results from one administration to another, meaning that if a test is reliable, then individuals should get similar results when taking the same test multiple times. It is important for test reliability to be established in order for the results to be reliable and meaningful.*

For the selected response assessment items, several prompts (instructions) were used to create the following questions:

(a) Fill-in-the-blank (Gap filling) questions:

Prompt / instruction: Create a gap-filling activity based on the following text. Remove the following words: accurately, valid, consistency, multiple, meaningful.

Test reliability is the degree to which a test measures what it is supposed to measure. It is an important concept in psychometrics, as it is necessary for a test to be reliable in order for the results to be considered ----- . Test reliability is determined by the ----- of the test results from one administration to another, meaning that if a test is reliable, then individuals should get similar results when taking the same test ----- times. It is important for test reliability to be established in order for the results to be ----- and ----- . accurately, valid, consistency, multiple, meaningful

(b) Short-answer / open-ended questions:
Prompt / instruction: Create five open-ended questions based test reliability.

1. What are the consequences of administering an unreliable test?
2. How can test reliability be established and improved?
3. What factors can impact the reliability of a test?
4. Are there any methods to determine if a test is reliable before administering it?
5. Why is it important for test reliability to be established?

(c) Multiple-choice questions:
Prompt / instruction: Create two multiple-choice questions on the reliability of a language test.

1. Which of the following is the most important factor for determining the reliability of a language test?
a. Grammatical accuracy
b. Scoring procedures
c. Length of the test
d. Number of test takers
e. Type of language assessment

2. What is the term used to describe the consistency of a language test?
a. Validity
b. Reliability
c. Efficiency
d. Precision
e. Accuracy

(d) Matching activity/questions:
Prompt / instruction: Create a matching activity for the following terms/concepts: validity, reliability, construct validity, internal validity, and test-retest reliability.

Terms / Concepts
1. Validity
2. Reliability
3. Construct Validity
4. Internal Validity
5. Test-Retest Reliability

Definitions:
A. The degree to which a measure accurately reflects the concept it is intended to measure

B. The consistency of the results of a measure across different settings, times, and test administrators
C. The degree to which a test measures what it is intended to measure
D. The extent to which research results are not affected by other factors
E. The degree to which a measure yields the same results each time it is used

Matching Activity:
1. Validity - C
2. Reliability - B
3. Construct Validity - A
4. Internal Validity - D
5. Test-Retest Reliability - E

Data collection and analysis
Data were collected through the journal kept by the author while creating exam questions for the course. The journal included the following questions:

a) What are the learning objectives of this course, and how can AI be utilized to create questions that align with these objectives?
b) How did the use of AI allow you to tailor your assessment items to the specific needs of your course and curriculum?
c) What opportunities or benefits did you encounter using AI to create the questions?
d) What are the potential limitations or challenges of using AI for creating assessment questions, and how can these be mitigated or addressed?
e) Any other comments or suggestions?

The participant started to write his responses to these questions in the journal two weeks before each assessment period, including the midterm and the final exams. The content analysis method was used to analyze the data collected from the journal (Coffey & Atkinson, 1996). The researcher's responses were coded and categorized based on the responses to the journal questions. A total of 78 entries were analyzed, and these entries were collected for four weeks. Regarding word count, it was determined that 4,006 words were written in these entries. These categories were then analyzed to identify any patterns or trends from the data. The results of the analysis were then used to draw conclusions and make recommendations.

Findings and discussion
Research Question 1. How AI application in education sector can support the design of foreign language exam questions?

The responses in the journal reveal that there are three major opportunities/benefits of using AI in creating language assessment items: practicality, customization, and efficiency.

Practicality. In terms of time and effort, it can be stated that using AI to create assessment items was practical in that in less than a minute, depending on the instruction given to AI, questions were automatically created with no human intervention. In other words, almost no effort was required regarding the language used in the output and the content editing. AI-based systems can generate unlimited questions, which can help reduce the time and effort required to create and maintain large assessment items and scoring, which is critical regarding reliability. Using AI to

generate questions could help reduce the time spent researching and preparing for the test, allowing more time to focus on understanding the material. This finding is also consistent with the one conducted by Wang (2018), which indicated the practical aspects of using AI in creating tests.

Customization. The use of AI enabled the participant to create assessment items based on the objectives and the assessment needs regarding the curriculum. In other words, the questions created represented the objectives and the course content. Considering this, it can be alleged that AI-based assessments can be tailored to the needs of the course lecturer and the curriculum. For example, the participant in the study changed the instructions or questions to suit the needs of the course lecturer and the curriculum. Also, AI helped to address potential needs more efficiently, for example, by creating authentic texts (cf. Kılıçkaya & Kic-Drgas, 2023). This flexibility ensured the assessments aligned with the learning objectives and effectively measured the desired outcomes.

Efficiency. Regarding the quality of the assessment items created, it was determined that the majority of the questions were of quality enough to assess the student's knowledge about the course content, including the fundamental principles of assessment and the assessment formats that could be used in classroom assessment. The notes in the journal also indicated that the participant had little intervention to edit or revise the questions, as they were well-written in terms of grammar and mechanics. These findings align with those of the studies by Zawacki-Richter et al. (2019) and Settles et al. (2020), indicating that the questions created by the AI systems were accurate and efficient. However, though it was conducted to a limited extent, a manual review of the output was always necessary, as indicated by the review study by Zhang and Li (2021).

Research Question 2. What challenges did the participant encounter when using AI to create exam questions?

While the use of AI in creating language assessment provided several opportunities and benefits, as discussed in the section, several challenges were determined to be considered. The comments and the explanations in the journals reveal three major challenges of using AI in creating language assessment items: Readability, validity, and ownership.

Readability. It was determined in the journals that AI could create questions based on the instruction provided to the AI system. In other words, the questions might not be relevant to the student's current knowledge level and understanding of the material. Moreover, since AI created questions based on the data or the instruction that it was provided with, and if the data or the instruction was not clear or not well-written, the questions could also be of low quality and relevance. The notes taken during the item creation process showed that the participant had to revise his instruction or explanation to AI as some created were not found to be relevant to the objective or the course content. Additionally, as the notes indicated, sometimes the questions were found to be too basic or complex for the intended audience. As indicated by Yunjiu et al. (2022), the questions created by human experts might have a higher level of item discrimination power than those created by AI.

Validity. Although AI could be used to create items that are easier to score, such as multiple-choice and gap-filling activities with only one answer, short responses, and questions with straightforward right and wrong answers, which contributes to

reliability, it was also seen in the journal that AI-created items might also be limited in their ability to generate culturally appropriate language times. In other words, AI systems may be unable to consider the cultural and linguistic context of the questions they generate, and they can often produce questions that might be biased towards specific groups, which require intervention by teachers or human experts. This can be particularly problematic if the questions generated are used in an assessment context, as the bias can lead to unfair results, especially in high-stakes nationwide or worldwide examinations.

Ownership. Although AI (OpenAI) assigns all the rights regarding the use of the output generated by the AI to the user and states that it will not claim copyright over content generated by the API (OpenAI, 2023), it is not certain who owns the rights to the work. The use of AI-created questions in the classroom assessment, according to the notes and comments, might not be an issue in terms of copyright; however, when they are published or used publicly, copyright might appear as an essential concern for both the user of AI as well as the creator of this artificial system. Further discussions and ethical issues might also exist even when human experts revise or alter these questions, which was also voiced in other studies (e.g., Adams et al., 2022).

Implications

Considering all these opportunities/benefits and the challenges, it can be put forward that AI-based systems have great potential to change the future of language assessment both in the language classroom and in high-stakes examinations. Moreover, the following implications regarding using AI in language assessment practices can be suggested.

- There has been an increased focus on the impact of technology on teaching and learning. In addition to a large number of tools and websites being amassed, it has now become imperative for language teachers to embrace artificial intelligence (AI) rather than ignore it. This is mainly due to the fact that AI will have a significant impact on teaching and learning practices.
- Although further research is mandatory regarding the potential role and use of AI in language assessment and there is not enough evidence regarding how it can be utilized, in addition to other uses in teaching and learning, basics of AI should be introduced in pre-and in-service teacher education programs (Hubbard, 2022; Levi & Inbar-Lourie, 2020) to introduce the pedagogical use of AI. This might appear critical, especially when AI is expected to change the roles of teachers in the classroom (Shen & Su, 2020).
- AI can be introduced to students in the classroom so that they can also create their own assessment items and implement various types of uses that meet their needs and preferences. In this vein, learners can move forward toward the learner as an agent and participant rather than the learner as a knowledge consumer (Cope & Kalantzis, 2019; Darvishi et al., 2022). This can also benefit those preparing for exams, as it allows them to practice similar questions they are likely to face in the actual exams or tests.

Conclusion

Language assessment helps teachers and learners to evaluate teaching and learning practices and to identify any areas of difficulty and tailor instruction and learning to meet the needs of individual learners. This study aimed to determine the challenges and opportunities of using AI to create language assessment questions. The study was designed as a case study, having a single participant whose experience and perceptions were explored. The findings revealed that AI offered several opportunities

such as practicality, customization, and efficiency in creating assessment questions. However, using AI also posed several challenges, such as readability, validity, lack of data for other languages, and ownership. Despite the benefits of using AI in creating assessment items, human intervention was still considered necessary to ensure the validity and quality of the questions generated.

Limitations of the study and suggestions for further research

The current study was limited to a single case, which limited the generalizability of findings. Additionally, case studies may be subject to researcher bias, as the researcher may be influenced by his opinions and perspectives of the situation. Furthermore, the study was limited to the data provided by the participant and the available journals, which may not represent the whole picture. Finally, the study was also limited by the time frame of the research, which may not have fully captured the complex dynamics between the different actors, such as using AI in preparing questions for reading tasks such as automatic creation of pre-reading questions as well as possible answers (Attali et al., 2022; Henrickson, 2021; Killawala et al., 2018; Taylor, 2022). Since AI-generated content is becoming increasingly prevalent, from music to artwork, it raises important legal questions about who owns the rights to the work. As such, regarding the ownership or copyright protection (Zurth, 2021) of the questions or items for assessment purposes, further research can investigate who owns the intellectual property rights to question language items produced by artificial intelligence. Moreover, it will be crucial to investigate to what extent the revisions and changes introduced by teachers on the produced questions and items can affect this ownership in addition to other issues related to the productive, disruptive, or destructive roles that AI might have, such as equity (Stephenson & Harvey, 2022) and AI-based cheating (Fyfe, 2022).

Bibliographic references

Adams, C., Pente, P., Lemermeyer, G., Turville, G. J., & Rockwell, G. (2022). Artificial Intelligence and Teachers' New Ethical Obligations. International Review of Information Ethics, 31, 1-18. https://doi.org/10.29173/irie483

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The Interactive Reading Task: Transformer-Based Automatic Item Generation. Frontiers in Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.903077

Baker, T., Smith, L., & Anissa, N. (2019). Educ-AI-tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges. London: NESTA. https://www.nesta.org.uk/report/education-rebooted

Borade, J. G., & Netak, L. D. (2021). Automated Grading of Essays: A Review. Intelligent Human-Computer Interaction (IHCI 2020): Lecture Notes in Computer Science. New York: Springer. 238-249. https://doi.org/10.1007/978-3-030-68449-5_25

Brown, H. D., & Abeywickrama, P. (2010). Language Assessment: Principles and Classroom Practices (2nd ed.). New York: Pearson Education. ISBN 978-0138149314

Çakmak, F. (2019). Mobile Learning and Mobile Assisted Language Learning in Focus. Language and Technology, 1 (1), 30-48. https://dergipark.org.tr/tr/download/article-file/665969

Çakmak, F. (2022). Chatbot-Human Interaction and Its Effects on EFL Students' L2 Speaking Performance and Speaking Anxiety. Novitas-ROYAL (Research on Youth and Language), 16 (2), 113-131. https://novitasroyal.org/current-issue/?wpdmc=volume-16-issue-2

Chatterjee, J. M., Garg, H., & Thakur, R. N. (Eds.). (2023). A Roadmap for Enabling Industry 4.0 by Artificial Intelligence. Hoboken: Wiley and Sons. ISBN 978-1-119-90512-7.
Coffey, A., & Atkinson, P. (1996). Making Sense of Qualitative Data: Complementary Research Strategies. Newbury Park: Sage.
Cope, B., & Kalantzis, M. (2019). Education 2.0: Artificial Intelligence and the End of the Test. Beijing International Review of Education, 1, 528-543. https://brill.com/view/journals/bire/1/2-3/article-p528_528.xml
Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial Intelligence for Education: Knowledge and Its Assessment in AI-Enabled Learning Ecologies. Educational Philosophy and Theory, 53 (12), 1229-1245. https://doi.org/10.1080/00131857.2020.1728732
Creswell, J. W. (2007). Qualitative Inquiry and Research Design: Choosing Among Five Approaches (2nd ed.). Newbury Park: Sage.
Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and Learning Analytics to Build Trustworthy Peer Assessment Systems. British Journal of Educational Technology, 53 (4), 844-875. https://doi.org/10.1111/bjet.13233
Dumont, G. (2023). Immersion in Organizational Ethnography: Four Methodological Requirements to Immerse Oneself in the Field. Organizational Research Methods, 26 (3), 441-458. https://doi.org/10.1177/10944281221075365
Eaton, S. E., Mindzak, M., & Morrison, R. (2021, May 29 - June 3). Artificial Intelligence, Algorithmic Writing & Educational Ethics [Paper Presentation]. Canadian Society for the Study of Education Société canadienne pour l'étude de l'éducation, Edmonton, AB, Canada. http://hdl.handle.net/1880/113569
Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled Approaches to Assessment Design, Development, and Implementation in A. A. Rupp & J. P. Leighton (Eds.), The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications. 41-72. Hoboken: Wiley and Sons. https://doi.org/10.1002/9781118956588.ch3
Freddi, M. (2021). Reflection on Digital Language Teaching, Learning, and Assessment in Times of Crisis: A View from Italy. In N. Radić, A. Atabekova, M. Freddi & J. Schmied (Eds.), The World Universities' Response to COVID-19: Remote Online Language Teaching, 279-293. Research-publishing.net. https://doi.org/10.14705/rpnet.2021.52.1278
Fyfe, P. (2022). How to Cheat on Your Final Paper: Assigning AI for Student Writing. AI & Society. https://doi.org/10.1007/s00146-022-01397-z
García-Peñalvo, F. J., Corell, A., Abella-Garcia, V., & Grande-de-Prado, M. (2021). Recommendations for Mandatory Online Assessment in Higher Education During the COVID-19 Pandemic. In D. Burgos, A. Tlili, & A. Tabacco (Eds.), Radical Solutions for Education in a Crisis Context: COVID-19 as an Opportunity for Global Learning (pp. 85-98). Berlin: Springer. https://doi.org/10.1007/978-981-15-7864-3_7
Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? or buncombe and ballyhoo?" Journal of Computer Assisted Learning, 37(5), 1207-1216. https://doi.org/10.1111/jcal.12577
Griffee, D. T. (2012). An introduction to second language research methods: Design and data. Dale T. Griffee. TESL-EJ Publications. http://www.tesl-ej.org/pdf/ej60/sl_research_methods.pdf
Godwin-Jones, R. (2022). Partnering with AI: Intelligent Writing Assistance and Instructed Language Learning. Language Learning & Technology, 26(2), 5-24. http://doi.org/10125/73474
Mackey, A., & Gass, S. M. (2022). Second Language Research: Methodology and Design (3rd ed.). Oxfordshire: Routledge. ISBN 9781032036632

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. Applied Sciences, 11(12), 54-67. https://doi.org/10.3390/app11125467
Henrickson, L. (2021). Reading Computer-Generated Texts. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108906463
Hern, A. (2022, December 4). AI Bot ChatGPT Stuns Academics with Essay-Writing Skills and Usability. The Guardian. https://www.theguardian.com/technology/2022/dec/04/ai-botchatgpt-stuns-academics-with-essay-writing-skills-and-usability
Hinkelman, D. (2018). Blending Technologies in Second Language Classrooms (2nd ed.). Hampshire: Palgrave Macmillan. ISBN 978-0-230-23261-7
Hubbard, P. (2022). Bridging the Gap Between Theory and Practice: Technology and Teacher Education. In N. Ziegler & M. González-Lloret (Eds.), Routledge Handbook of Second Language Acquisition and Technology, 21-35. Oxfordshire: Routledge. https://doi.org/10.4324/9781351117586
Hughes, A., & Hughes, J. (2020). Testing for Language Teachers (3rd ed.). Cambridge: Cambridge University Press. ISBN 978-1108714822.
Jones, J., & Smith, J. (2017). Ethnography: Challenges and Opportunities. Evidence-Based Nursing, 20, 98-100. https://doi: 10.1136/eb-2017-102786
Kessler, G. (2023). Computer Assisted Language Learning. E. Hinkel (Ed.), Handbook of Practical Second Language Teaching and Learning, 173-183. Oxfordshire: Routledge. https://doi.org/10.4324/9781003106609
Killawala, A., Khokhlov, I., & Reznik, L. (2018). Computational Intelligence Framework for Automatic Quiz Question Generation. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1-8. https://doi.org/10.1109/FUZZ-IEEE.2018.8491624
Kılıçkaya, F. (2020). Using a Chatbot, Replika, to Practice Writing Through Conversations in L2 English: A Case Study. M. Kruk, & M. Peterson (Eds.), New Technological Applications for Foreign and Second Language Learning and Teaching, 222-239. Hershey: IGI Global. https://doi.org/10.4018/978-1-7998-2591-3.ch011
Kılıçkaya, F., & Kic-Drgas, J. (2023). Misuse of AI (Artificial Intelligence) in Assignments: Can AI-Written Content Be Detected?. R. E. Ferdig, R. Hartshorne, E. Baumgartner, R. Kaplan-Rakowski & Ch. Mouza (Eds.), What PreK-12 Teachers Should Know about Educational Technology in 2023: A Research-to-Practice Anthology, 145-153. AACE – Association for the Advancement of Computing in Education. https://www.learntechlib.org/p/222690/, ISBN: 978-1-939797-72-8.
Kukulska-Hulme, A. (2020). Mobile Assisted Language Learning. C. A. Chapelle (Ed.), The Concise Encyclopedia of Applied Linguistics. Hoboken: Wiley and Sons. https://doi.org/10.1002/9781405198431.wbeal0768.pub2
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
Levi, T., & Inbar-Lourie, O. (2020). Assessment Literacy or Language Assessment Literacy: Learning from the Teachers. Language Assessment Quarterly, 17(2), 168-180. https://doi.org/10.1080/15434303.2019.1692347
Levy, M., & Stockwell, G. (2006). CALL Dimensions: Options and Issues in Computer-Assisted Language Learning. NJ: Routledge. ISBN 0-8058-5634-X.
McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27 (4), 12-14. https://doi.org/10.1609/aimag.v27i4.1904
Nazaretsky, T., Ariely, M., & Cukurova, M., & Alexandron, G. (2022). Teachers' Trust in AI-Powered Educational Technology and a Professional Development

Program to Improve It. British Journal of Educational Technology, 53 (4), 914-930. https://doi.org/10.1111/bjet.13232
Odilinye, L., Popowich, F., Zhang, E., Nesbit, J., & Winne, P. H. (2015). Aligning Automatically Generated Questions to Instructor Goals and Learner Behavior. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). 216-223. https://doi.org/10.1109/ICOSC.2015.7050809
OpenAI. (2023). Will OpenAI Claim Copyright Over What Outputs I Generate with the API? https://help.openai.com/en/articles/5008634-will-openai-claim-copyright-over-what-outputs-i-generate-with-the-api
Perkins, M. (2023). Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond. Journal of University Teaching & Learning Practice, 20(2). https://doi.org/10.53761/1.20.02.07
Purpura, J. E. (2016). Second and Foreign Language Assessment. The Modern Language Journal, 100 (Supplement). https://doi.org/10.1111/modl.12308
Russell, V., & Murphy-Judy, K. (2021). Teaching Language Online: A Guide for Designing, Developing, and Delivering Online, Blended, and Flipped Language Courses. Oxfordshire: Routledge. https://doi.org/10.4324/9780429426483
Selwyn, N., Hillman, T., Rensfeldt, A. B., & Perrotta, C. (2021). Digital Technologies and the Automation of Education- Key Questions and Concerns. Postdigital Science and Education, 5. https://doi.org/10.1007/s42438-021-00263-3
Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine Learning – Driven Language Assessment. Transactions of the Association for Computational Linguistics, 8. https://doi.org/10.1162/tacl_a_00310
Shen, L., & Su, A. (2020). The Changing Roles of Teachers with AI. M. K. Habib (Ed.), Revolutionizing Education in the Age of AI and Machine Learning. 1-25. Hershey: IGI Global. https://doi.org/10.4018/978-1-5225-7793-5
Stephenson, B., & Harvey, E. (2022). Student Equity in the Age of AI-Enabled Assessment: Towards a Politics of Inclusign. In R. Ajjawi, J. Tai, D. Boud, T. J. de St Jorre (Eds.), Assessment for Inclusion in Higher Education: Promoting Equity and Social Justice in Assessment. Oxfordshire: Routledge. 120-130. https://doi.org/10.4324/9781003293101
Taylor, A. (2022). Technology and L2 Reading: Current Research and Application. In N. Ziegler & M. González-Lloret (Eds.), Routledge Handbook of Second Language Acquisition and Technology. Oxfordshire: Routledge. 174-186. https://doi.org/10.4324/9781351117586
UNESCO. (2021). AI and Education: Guidance for Policy Makers. https://unesdoc.unesco.org/ark:/48223/pf0000376709
Van Moere, A., & Downey, R. (2016). Technology and Artificial Intelligence in Language Assessment. In D. Tsagari & J. Banerjee (Eds.), Handbook of Second Language Assessment. Berlin: De Gruyter Mouton. 341-358. https://doi.org/10.1515/9781614513827-023
Vermeer, S. A. M., Araujo, T., Bernritter, S. F., & van Noort, G. (2019). Seeing the Wood for the Trees: How Machine Learning Can Help Firms in Identifying Relevant Electronic Word-of-Mouth in Social Media. International Journal of Research in Marketing, 6(3). https://doi.org/10.1016/j.ijresmar.2019.01.010
Voss, E. (2018). Technology and Assessment. In J. Liontas (Ed.), The TESOL Encyclopedia of English Language Teaching Hoboken: Wiley and Sons. 1-7. https://doi.org/10.1002/9781118784235.eelt0388
Yu, Y., Han, L., Du, X., & Yu, J. (2022). An Oral English Evaluation Model Using Artificial Intelligence Method. Mobile Information Systems. Article ID 3998886. https://doi.org/10.1155/2022/3998886
Yunjiu, L., Wei, W., & Zheng, Y. (2022). Artificial Intelligence-Generated and Human Expert-Designed Vocabulary Tests: A Comparative Study. SAGE Open, 12(1). https://doi.org/10.1177/21582440221082130

Wang, H. (2018). Research on Intelligent Standardized English Test Systems with Artificial Intelligence. J. Mizera-Pietraszko & P. Pichappan (Eds.), Lecture Notes in Real-Time Intelligent Systems: Advances in Intelligent Systems and Computing 613, 33-40. London, New York: Springer. https://doi.org/10.1007/978-3-319-60744-3_4
Yanhua, Z. (2020). The Application of Artificial Intelligence in Foreign Language Teaching. Proceedings of the 2020 International Conference on Artificial Intelligence and Education (ICAIE) 40-42. IEEE. https://doi.org/10.1109/ICAIE50891.2020.00017
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? International Journal of Educational Technology in Higher Education, 16, Article 39. https://doi.org/10.1186/s41239-019-0171-0
Zhang, M., & Li, J. (2021). A Commentary of GPT-3 in MIT Technology Review 2021. Fundamental Research, 1(6), 831-833. https://doi.org/10.1016/j.fmre.2021.11.011
Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. Hindawi. https://doi.org/10.1155/2021/8812542
Zhang, R., & Zou, D. (2022). Types, Purposes, and Effectiveness of State-of-the-Art Technologies for Second and Foreign Language Learning. Computer Assisted Language Learning, 35(4), 696-742. https://doi.org/10.1080/09588221.2020.1744666
Zurth, P. (2021). Artificial Creativity? A Case Against Copyright Protection for AI Generated Works. UCLA Journal of Law & Technology, 25(2), 1-20. https://ssrn.com/abstract=3707651

Words: 6759
Characters: 47 332 (26,3 standard pages)

Joanna Kic-Drgas
Institute of Applied Linguistics,
Adam Mickiewicz University in Poznań,
Poland
https://orcid.org/0000-0002-8133-9190
joanna.kic-drgas@amu.edu.pl

Prof. Ferit Kılıçkaya, Dr.
Department of Foreign Language Education,
Burdur Mehmet Akif Ersoy University,
Türkiye
https://orcid.org/0000-0002-3534-0924
ferit.kilickaya@gmail.com