

Aleksa Đ. Stošić¹

Belgrade Metropolitan University
Faculty of Foreign Languages
English in Business
<https://orcid.org/0009-0005-4601-0401>

Marija V. Milojković²

Belgrade Metropolitan University
Faculty of Foreign Languages
English in Business
<https://orcid.org/0000-0003-4787-1017>

ESL ASSESSMENT IN THE AGE OF AI: SERBIAN ESL TEACHERS' ATTITUDES TOWARDS THE USE OF CHATGPT

Following the introduction of publicly accessible AI and LLMs, they have gradually started to be adopted in the ESL classroom. Recent studies have revealed that these tools have had a wider impact on ESL assessment, allowing for enhanced accuracy, efficiency, and personalization in language evaluation. However, how widely these technologies are understood and used by teachers is still being investigated. This study explores how Serbian ESL teachers use and perceive ChatGPT as a tool for assessing student work. The study employs a mixed-methods approach, combining survey data from 53 participants with follow-up interviews from 10 educators to examine their tendencies, habits, and opinions regarding ChatGPT for assessment. Findings revealed that grammar and vocabulary are the language skills which Serbian ESL teachers assessed most often using ChatGPT, while skills such as speaking, listening, and critical thinking were rarely assessed using AI. Teachers also expressed moderate trust in the tool's ability to detect grammar and syntax errors, but were concerned about its limitations. These limitations included ChatGPT's ability to provide accurate assessment, its tendency to give vague feedback, and its lack of contextual understanding. While many teachers recognized its ability to save time for assessment tasks, they also recognized their own limited experiences with AI tools and stressed the need for human oversight, formal training, and ethical safeguards. This study indicated that even though ChatGPT does not yet have widespread use in the Serbian ESL classroom, interviewees frequently emphasized that it should be used as a supplementary tool to assist teachers, but that guidelines and professional training on its use need to be provided to the teachers before they would readily implement it in the classroom.

Keywords: ChatGPT, Artificial Intelligence, EFL, assessment, teacher attitudes

Introduction

Artificial intelligence (AI) and large language models (LLMs) have had a wide impact on several fields in the last few years, particularly on those which focus on com-

¹ aleksa.stosic@metropolitan.ac.rs

² marija.milojkovic@metropolitan.ac.rs

munication. Radford et al. (2019) explained how AI has the potential to process and generate text which is similar to text produced by humans and provided practical examples of how LLMs might be integrated into human society. Because of this, many articles have been published testing these wider applications of artificial intelligence. Similarly, in ESL teaching, many language instructors have begun using artificial intelligence in the classroom to assist with various teaching-related activities such as classroom material creation and preparation (Jeon et al., 2023), assignment grading and assessment (MIZUMOTO, EGUCHI 2023), as well as providing feedback for students (KORAISHI 2023).

In terms of ESL assessment, ChatGPT has primarily been used in assessing and providing feedback for student writing (ALBUQUERQUE DA SILVA et al. 2024), but skills such as reading, listening, and speaking have also successfully been assessed using this tool (LO et al. 2024). However, a study conducted by Zaiarna et al. (2024) showed that the number of ESL teachers who were not familiar with how to use ChatGPT's functions or were only moderately familiar with the technology was greater than the number of teachers who were confident in their abilities to use AI tools proficiently. This lack of AI literacy suggests that there is a need for targeted training to help educators integrate AI tools effectively into their teaching and assessment practices.

Therefore, the goal of this research was to perform an analysis of Serbian ESL teachers' habits and attitudes regarding their use of ChatGPT in the classroom. This study sought to answer how often Serbian ESL teachers used AI to assess student assignments, what type of skills they assessed with it, and their general attitude towards its use as a tool for student assessment.

Previous research

AES and AWE

The use of automated tools to assist with ESL assessment is not a new concept, as computer algorithms such as Automated Essay Scoring (AES) have been used in the past to assist examiners in evaluating and grading student essays (DIKLI 2006). AES tools typically utilize natural language processing to focus on analyzing the quality of the written essay and assigning it a score. However, in recent years, they have adapted deep linguistic features to provide students with feedback and more accurate scoring (HUSSEIN et al. 2019). These AES tools consist of two distinct engines, which are together called Automated Writing Evaluation (AWE) and are used for calculating scores and providing automatic feedback (KOLTOVSKAIA 2020). While research has shown that AWE can enhance student writing accuracy and revision practices, some researchers warn against the use of such tools, stating that they are incapable of capturing the nuances of classroom-based writing instruction (CONDON 2013).

The emergence of LLMs such as Bert and ChatGPT has further expanded the capabilities of automated assessment. Unlike earlier systems, LLMs use deep learning to process and generate human-like text, which allows for more nuanced evaluations. For instance, BERT's bidirectional architecture allows it to analyze context more effectively, leading to improved scoring accuracy when combined with traditional linguistic features (LAGAKIS, DEMETRIADIS 2021). Meanwhile, ChatGPT's generative abilities have opened new possibilities for prompt-based scoring and feedback, making AES more acces-

sible to educators without technical expertise (MIZUMOTO, EGUCHI 2023). Because of these factors, the Education Testing Service (ETS) has already begun using AI tools in scoring and assessing students undertaking tests such as TOEFL (MARSH RUNYON et al. 2022).

ChatGPT in Writing and Speaking Assessment

There have been several studies that examined how ChatGPT can be used to assess ESL writing. One example study is a paper by Koraishi (2023) which investigated how ESL teachers could use ChatGPT for assessment purposes. Koraishi explained that ChatGPT could be used for diagnostic assessment in placement tests, as the tool is familiar with the CEFR descriptors and can explain why it assigned a specific language level to a student based on the input that it received. Additionally, the paper mentioned that the AI tool can be used to quickly assess student writing in proficiency exams such as IELTS, TOEFL, GRE, etc., and even provide feedback and reasoning for its assessment. However, the author noted that there were risks of ChatGPT making mistakes if it was not provided with exact rubrics by which to mark papers, and stressed that teachers needed to review the output manually. Similar concerns were raised by Mizumoto and Eguchi (2023), who conducted an analysis on the potential of ChatGPT for evaluating student writing compared to human benchmarks. Their study involved ChatGPT performing an analysis of 12,100 essays with pre-assigned proficiency levels taken from the TOEFL11 Corpus and scoring them based on IELTS writing band descriptors (0-9 scale). This resulted in ChatGPT showing moderate agreement with the previous human assessment, and it was discovered that it excelled at lexical prediction, but did not perform well when assessing writing structure due to a lack of a deeper understanding of the text. Due to only a moderate success rate, this paper came to a similar conclusion as Koraishi (2023) – that ChatGPT requires human oversight and ought to primarily be used as a supplementary tool for assessment purposes.

A further study by Yoon et al. (2023) gave a more critical perspective on ChatGPT's use for writing assessment. This study suggested that ChatGPT struggles with finding the controlling idea in lower-scoring essays and would therefore provide students with inaccurate or overly positive feedback. Additionally, it was discovered that ChatGPT could not detect repetitions if they were expressed in different words, and that it was overly sensitive to how transitional words were used and often did not know when they were used correctly. On the other hand, Mahapatra (2024) offers a more positive outlook, stating that students who used ChatGPT to assist them in learning ESL writing reported improvements in their ESL writing performance when compared to their peers who did not use it. However, this study also raised a new concern regarding students becoming overreliant on using AI and thus slowing the development of their critical thinking skills.

While most studies have focused on how ChatGPT could be used to assess student writing, there are a few studies that have investigated its use in speaking assessment. An example study by Wang et al. (2023) examined the use of ChatGPT and BreakBERT for detecting phase breaks in the speech of ESL learners. The study showed that if given enough training, models such as BreakBERT helped with the evaluation of phase breaks significantly. On the other hand, ChatGPT showed inconsistent results and had more limitations when assessing subtle pauses between semantic groups. Another study that

examined speaking assessment by Muniandy and Selvanathan (2024) encouraged students to use ChatGPT as a conversation partner in order to practice persuasive dialogue. Due to ChatGPT lacking an option to have a voice conversation at the time the research was undertaken, students were asked to use an extension called ChatGPT Voice Master to be able to communicate with it orally. Both studies came to a similar conclusion – that ChatGPT could be a useful tool to help improve student speaking, but that it is held back by the inconsistent quality of its feedback and its tendency to provide incorrect answers. These concerns coincide with previous research into how ChatGPT was used for writing assessment, as those studies mentioned similar concerns about feedback accuracy.

Research by Reza Abedi et al. (2025) also draws attention to the limitations of ChatGPT for speaking assessment, noting that AI tools struggle with detecting nuances and understanding diverse accents. However, this paper also provides further evidence that AI tools can have a positive impact on student writing and reading skills due to their ability to provide instant feedback on text-based inputs. Additionally, the paper highlights further potential issues with using AI for language learning assessment, as many AI tools offer premium versions that have fewer limitations and can offer more nuanced and well-thought-out responses. This poses a potential problem as students from lower-income backgrounds who use free versions would receive less improvement in their language skills when compared to students that have access to these premium versions.

The findings of these research papers provide context for how ChatGPT may fit into the Sociocultural theory and Second language acquisition (SLA) (LANTOLF, BECKETT 2009). Traditionally, teachers are the ones who help learners by pushing them to try and do independent work and provide feedback afterwards. In the context of the SLA framework, AI tools such as ChatGPT can be considered mediational tools, as they provide instant feedback to students using them and can provide guidelines on how students can improve further or even provide more exercises for further practice. However, previous research has revealed that AI tools are not foolproof or error-free, and further improvements to their accuracy need to be made before they can replace teachers in the assessment process.

Teachers' Opinions on GPT and Technology Acceptance

The Technology Acceptance Model (TAM) (DAVIS 1989) and the Unified Theory of Acceptance and Use of Technology (UTAUT) (VENKATESH et al. 2003) can assist in understanding the factors that will lead to ESL teachers adopting and using ChatGPT in the classroom. TAM explains that two factors govern whether a new technology will be adopted: the perceived usefulness of a tool (how much a person believes a tool will enhance their performance) and the perceived ease of use (how easy it is to use and adopt the tool for various tasks). UTAUT builds on the postulates of TAM and expands them with more factors such as performance and effort expectancy (how easy it is to use a new technology and the perceived benefits of using that technology), social influence (an individual's perception of how others believe a tool should be used) and facilitating conditions (the degree to which an individual believes that there is technical infrastructure to support the use of new technology). Considering that ChatGPT and AI tools have only been publicly available for a few years, research on their adoption by teachers is still ongoing. Papers investigating their adoption, such as those by Dahri et al. (2024), have

revealed that there is a high acceptance of ChatGPT among academics and that their willingness to adopt AI tools is primarily influenced by personal competencies, perceived usefulness of AI tools, and their ease of use. This study also discovered that trust in AI, its perceived intelligence, and even joy of use were also significant factors for its adoption, and that factors such as social influence and institutional practices played a significant role in their acceptance too. Similar results were also reported by Balaskas et al. (2025), who discovered that age and prior experience with AI were moderately significant factors for its adoption, implying that younger teachers, or teachers with more technical experience, are more likely to adopt AI tools like ChatGPT.

Regarding ESL teacher's tendencies and willingness to use ChatGPT and other AI tools, research by Zaiarna et al. (2024) indicated that teachers generally recognized their usefulness for tasks such as lesson planning, content creation, grammar and vocabulary assessment, but that they were hesitant to fully adopt the technology in the classroom due to ethical concerns (primarily its tendency to give incorrect feedback and imagine mistakes) and lack of training on how to use AI tools. The lack of training and institutional support, and the need for human oversight, could lead to reduced adoption rates despite the interest in AI tools, thus reducing the probability based on TAM and UTAUT. Hossain and Younus (2025) add further evidence to these findings, as they also discovered that teachers had a generally positive view of AI tools, but called attention to a need for a specific framework to be developed that would allow teachers to use AI tools with more confidence.

Gaps in the Literature

Even though the number of papers investigating the use of ChatGPT for language learning is growing quickly, there are still several gaps in the literature. First, most studies focus on writing assessment (ALBUQUERQUE DA SILVA et al. 2024; BUCOL, SANGKAWONG 2024; MAHAPATRA 2024), which means that it is still not entirely clear how it can be used effectively to assess listening, reading, and speaking. While speaking assessment has been explored to a degree (WANG et al. 2023; MUNIANDY, SELVANATHAN 2024), there are few or no papers discussing its impact on reading and listening (LO et al. 2024). Second, around two-thirds of studies were conducted at the university level or adult learning contexts, meaning that there is not enough information on how ChatGPT might be used in an elementary or secondary school setting. Since the focus of this paper is on Serbian ESL teachers, many of whom work in public schools at the elementary or high school level, this paper could provide more information on the experiences of these teachers. Thirdly, the geographical gap is also significant, as according to Lo et al. (2024), nearly half of all published studies on this topic have come from East Asia, and another quarter from the Middle East, while European studies accounted for around 14% of all studies. By focusing on Serbian ESL teachers, this paper will focus on an underrepresented area and potentially discover unique challenges or advantages, such as how the national curriculum affects teachers' use of AI. Therefore, this study aims to contribute both empirical data and practical insights to the discourse of AI's use in education.

Methodology

This paper set out to discover Serbian ESL teachers' tendencies, habits, and opinions regarding ChatGPT when used for student assessment. The data for this paper were collected from two sources using a mixed-methods approach to gather both quantitative and qualitative data. The first source was a survey investigating Serbian ESL teachers' habits and attitudes toward using ChatGPT in the classroom. This survey was carried out in Google Forms and contained 19 items comprising of four types of questions. Multiple-response items asked teachers to select all relevant ways in which they used ChatGPT (e.g. checking grammar and vocabulary, generating test items, assessing reading or listening tasks, etc.) thereby capturing the breadth of functions and skill for which the tool was employed. Open-ended questions asked participants to elaborate on their answers in their own words, describe specific assessment scenarios in which they used ChatGPT, and explain perceived benefits, risks, or limitations that could not be captured through fixed-choice options. Yes/no questions focused on key categorical indicators, such as whether teachers had ever used ChatGPT for assessment, whether their institution officially allowed or supported AI-assisted assessment, and whether they had received any form of training in AI tools. Finally, several 5-point Likert scale items (never, rarely, sometimes, often, always) measured the frequency with which teachers used ChatGPT for different assessment-related purposes, as well as their self-reported satisfaction and perceived usefulness of ChatGPT in supporting assessment practices. In addition to these questions, the introductory segment of the survey included questions pertaining to demographic data such as age, gender, and teaching experience.

The second source was a series of interviews with Serbian ESL teachers. The purpose of the interviews was to obtain more detailed accounts of how teachers used ChatGPT to assess student assignments and how satisfied they were with its accuracy, fairness, and practicality. A semi-structured interview format was employed, consisting of 13 open-ended questions focusing on concrete examples of ChatGPT-based assessment, perceived reliability across different task types, ethical and institutional concerns, and the extent to which teachers verified or edited ChatGPT's suggestions. Additional ad hoc follow-up questions asked when further explanation or specification was required. The survey included 53 participants, all of whom were currently teaching in a school, university, or similar educational institution. This active teaching role served as the inclusion criterion for participation. Out of this group of 53 participants, 10 teachers who expressed a willingness to elaborate on their experiences volunteered for follow-up semi-structured interviews focusing on their use of ChatGPT in the classroom.

Survey Results

Survey Demographic Information

The survey included a total of 53 participants. The following table shows demographic information such as age, teaching experience, teaching level, teacher technical skills, familiarity with ChatGPT, and formal training with AI tools.

Table 1. *Survey participant demographic and technical skill data*

Category	Response Option	Percentage (%)
Gender	Female	84.9%
	Male	15.1%
Age	Under 25	1.9%
	25–34	22.6%
	35–44	34%
	45–54	30.2%
	55 and over	11.3%
Teaching Experience	Less than 2 years	3.8%
	2–5 years	7.5%
	6–10 years	20.8%
	11–20 years	32.1%
	More than 20 years	35.8%
Teaching Level	Primary school	43.4%
	Secondary school	37.7%
	University/College	35.8%
	Language institute (adult learners)	1.9%
Technical Skills	Very high (help others with tools)	20.8%
	Comfortable with digital tools	54.7%
	Moderate skills	22.6%
	Low skills (need frequent help)	1.9%
Familiarity with ChatGPT	Very familiar (understand features well)	18.9%
	Familiar (use regularly)	28.3%
	Somewhat familiar (used a few times)	47.2%
	Slightly familiar	5.6%
Formal Training in AI Tools	Yes (workshop/webinar attended)	5.66%
	No (no formal training)	94.34%

Most survey participants were female teachers aged between 35 and 54 with high self-reported technical skills and some familiarity with ChatGPT. Only a small number had received formal training on how to use AI tools for student assessment.

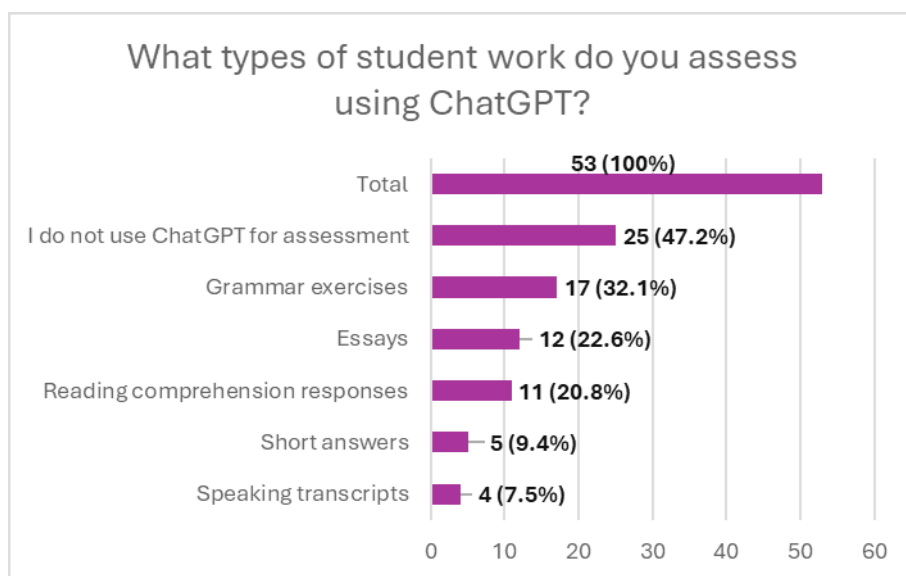
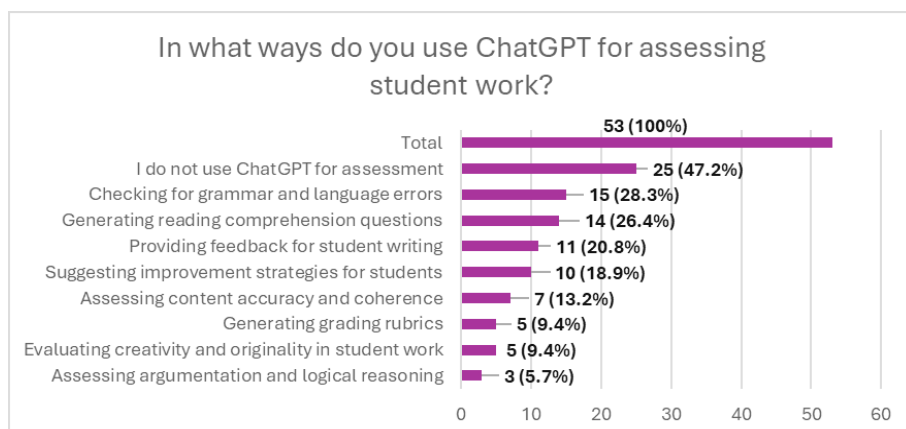
ChatGPT in Assessment

ChatGPT Type and Frequency of Use

The first section of the survey investigated how frequently Serbian ESL teachers used ChatGPT to assess various tasks and what kind of assessments they performed using it. The teachers were asked to rate how often they use ChatGPT to assess several different language skills.

Table 2. *Frequency of skill assessment*

Skill Assessed	Never	Rarely	Sometimes	Often	Always	% Total Usage
Grammar	62.3%	15.1%	9.4%	13.2%	0%	37.7%
Vocabulary	66%	9.4%	13.2%	11.3%	0%	34%
Reading Comprehension	67.9%	20.8%	5.7%	5.7%	0%	32.1%
Essay Writing	67.9%	17%	9.4%	5.7%	0%	32.1%
Argumentation / Persuasion	67.9%	18.9%	7.5%	3.8%	1.9%	32.1%
Creative Writing	75.5%	13.2%	3.8%	5.7%	1.9%	24.5%
Higher-Order Thinking Skills	77.4%	7.5%	9.4%	3.8%	1.9%	22.6%
Speaking	79.2%	11.3%	5.7%	3.8%	0%	20.8%
Listening Comprehension	81.1%	13.2%	1.9%	3.8%	0%	18.9%

Figure 1. *Types of student work assessed with ChatGPT*Figure 2. *Ways in which ChatGPT is used for assessment*

The data revealed that most of the surveyed teachers did not use ChatGPT for assessment, while those who did primarily used it to assess grammar and vocabulary. Speaking and listening skills were the least assessed skills using ChatGPT. In order to better understand the variables that influence how teachers use ChatGPT for different assessment tasks, multiple regression analysis between age, teaching experience, technical skills, and familiarity with ChatGPT was conducted. Additionally, a Pearson correlation was also conducted on the combined data of assessment frequency and usage patterns.

Table 3. *ChatGPT usage logistic regression*

Variable	B	S.E.	Wald	df	p-value	Exp(B)	Lower 95% CI	Upper 95% CI
const	-3.676	3.873	0.9	1	0.342	0.025	-11.268	3.916
Technical skills	0.598	0.622	0.925	1	0.336	1.819	-0.621	1.819
Familiarity with ChatGPT	1.835	0.84	4.766	1	0.029	6.267	0.187	3.483
Age	-0.181	0.682	0.07	1	0.79	0.833	-1.519	1.156
Teaching experience	-0.59	0.789	0.559	1	0.454	0.554	-2.136	0.956

B – Estimated regression coefficient (log-odds coefficient) for each predictor.

S.E. – Standard error of the coefficient B.

df – Degrees of freedom

p-value – Probability value

Exp(B) – Exponentiated B

CI – Coefficient interval

Table 4. *Most relevant Pearson correlation data on use and familiarity of ChatGPT*

Respondent Factor	Category	Pearson r	p-value	Mean	Standard Deviation
Uses ChatGPT	Grammar	0.327	0.017	1.736	1.095
Uses ChatGPT	Vocabulary	0.313	0.022	1.698	1.085
Uses ChatGPT	Reading comprehension	0.282	0.041	1.491	0.846
Uses ChatGPT	Composite	0.273	0.048	1.501	0.774
Familiarity with ChatGPT	Grammar	0.356	0.009	1.736	1.095
Familiarity with ChatGPT	Vocabulary	0.301	0.028	1.698	1.085

Statistical analysis revealed that age, teaching experience, and technical skill had no significant impact on how ChatGPT was used, and that the primary determiner for how teachers used it was their previous experience with it. Furthermore, the results indicated that those teachers who already used ChatGPT and had some level of familiarity

with it were most likely to use it for grammar, vocabulary, and reading comprehension assessment. Other relationships did not produce significant results.

ChatGPT Reliability and Assessment Modification

The second section of the survey investigated how reliable teachers found the assessments that ChatGPT generated to be and in what ways they modified them. Tables 3, 4, and 5 show the data related to these topics.

Table 5. *ChatGPT's reliability*

Statement	% of Respondents Agreeing
ChatGPT provides reliable assessment of grammar and syntax errors	24.5%
ChatGPT provides fair and unbiased opinions	17%
ChatGPT feedback is too generic and lacks depth	15.1%
ChatGPT provides reliable feedback for student writing	7.5%

Table 6. *Mistakes made by ChatGPT*

Common Mistake by ChatGPT	% of Respondents
Provides overly generic or vague feedback	28%
Misunderstands context or meaning	24%
Struggles to assess creativity and original thinking	18%
Overlooks nuanced errors in grammar and syntax	11%
Incorrectly evaluates content accuracy	10%
Misjudges argumentation and logical reasoning	9%

Table 7. *Modification to ChatGPT's assessment*

Modification Approach	% of Respondents
Only use ChatGPT for suggestions, create own assessments	26.4%
Edit ChatGPT assessments slightly for accuracy	18.9%
Make significant modifications to ChatGPT assessments	13.2%
Use ChatGPT assessments as they are	5.7%

The highest number of teachers stated that ChatGPT is reliable for assessing grammar and syntax errors, which coincides with the data on the frequency and type of work assessed by ChatGPT. This implies that teachers who find ChatGPT to be reliable for the assessment of a certain task are more likely to use it to assess that task. Additionally, teachers reported that the most common mistakes made by ChatGPT are that it provides overly generic or vague feedback and that it misunderstands context or meaning. Regarding their modification of ChatGPT's output, most teachers responded that they only used it for suggestions and that they created their own assessments, or that they edited its

output for accuracy.

Because of how the results on ChatGPT's reliability coincided with the results of use frequency and use type, this relationship was further explored through correlation and regression analysis.

Table 8. *Reliability correlations*

Assessment Task	R ²	B (Regression Coefficient)	p-value
Grammar and syntax error identification	0.48	1.67	< .001
Feedback on student writing	0.26	1.37	< .001
Checking content accuracy and coherence	0.31	1.79	< .001

The results of these tests confirmed that teachers who perceived ChatGPT as more reliable are significantly more likely to use it frequently for assessment purposes. There was a positive correlation between teachers' thinking that ChatGPT effectively identifies grammar and syntax errors and the frequency of them using it for this type of assessment. A similar result was found for how often ChatGPT was used to provide feedback on student writing, and checking content accuracy and coherence.

ChatGPT Challenges, Fairness, and Ethical Concerns

As part of the survey, ESL teachers were also asked to give an estimate on their own technical skills and ability to use AI tools. Then, they were asked to share their experience with ChatGPT and its limitations, the challenges they faced when using it, and their beliefs on whether it was fair to use AI for student assessment.

Table 9. *Self-evaluation of technical skills relating to AI use*

Response	% of Respondents
Not very confident or not confident at all	66%
Somewhat confident	24.5%
Very confident	9.5%
Would like to receive formal training on AI tools	67.9%

Table 10. *Perception of ChatGPT's limitations*

Response	% of Respondents
Did not notice any limitations when using ChatGPT	90.6%
Noticed limitations when using ChatGPT	9.4%

Table 11. *Challenges when using ChatGPT for assessment*

Challenge Faced	% of Respondents
Lack of training or knowledge on how to use ChatGPT efficiently	41.5%
Ethical concerns (e.g., fairness and bias)	35.8%
Inaccuracy of AI-generated feedback	24.5%
Institutional policies as a challenge	13.2%
Institutions provided unclear guidelines	9.4%
Institutions provided clear guidelines	3.8%

Table 12. *ChatGPT assessment fairness*

Belief about Fairness	% of Respondents
Somewhat fair, but should be combined with human assessment	71.7%
Lacks ability to fully understand student work	24.5%
Fair and provides consistent feedback	3.8%

Survey results showed that the biggest challenge that Serbian ESL teachers face when using AI is the lack of training or knowledge on how to use ChatGPT efficiently. This data correlates with the fact that the vast majority also stated that they are not very confident in their AI technical skills and may explain their willingness to receive formal training on how to use AI tools. Regarding their opinions on ChatGPT's fairness, most believed that it produced somewhat fair assessments, but that it needed to be combined with human oversight.

Future Use of AI and ChatGPT

The final section of the survey asked participants to state what they thought the future of AI in ESL would be like. Teachers were asked to state what role they think it should and will have in the future, as well as to comment on whether they would recommend it to other teachers.

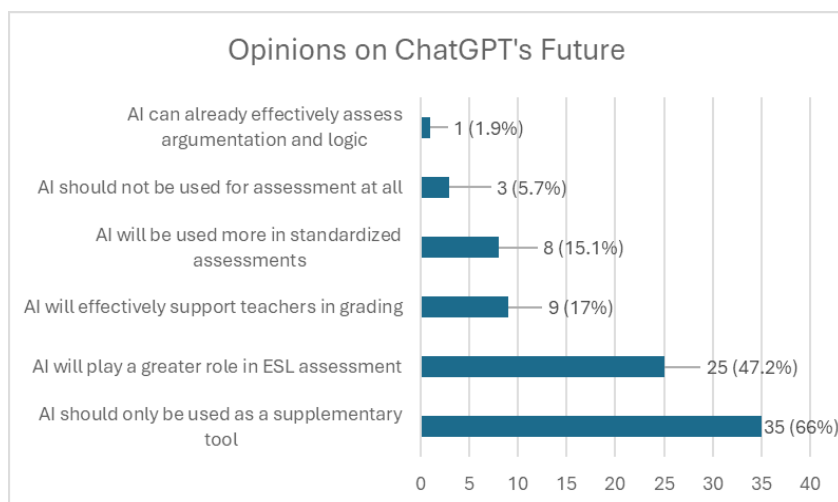
Figure 3. *Participant opinions of ChatGPT's future*

Table 13. *Chance of recommending ChatGPT to other teachers*

Recommendation Response	% of Respondents
Yes, but with caution	64.2%
No	26.4%
Yes, definitely	9.4%

Around two-thirds of survey respondents thought that AI should only be used as a supplementary tool in ESL assessment, with almost half also stating that they believed it would play a greater role in the future. Most respondents also claimed that they would recommend it to their colleagues, but would urge caution with its use.

Interview Findings

The follow-up interviews provided more insight into how Serbian ESL teachers used ChatGPT for assessment. Many of the key findings mirror the data from the surveys.

Grammar and Vocabulary

First, follow-up interviews revealed that vocabulary and grammar were the most assessed skills due to ChatGPT easily being able to check textual input and compare it with the exercise key if provided with one. Another reason was its ability to provide instant feedback and correct grammar and vocabulary mistakes. Throughout the interviews, several teachers revealed that they trusted ChatGPT to assess simple grammatical exercises such as fill-in-the-blanks and cloze tests. When asked to provide an explanation for these observations, one teacher stated:

“Using ChatGPT for the initial scan of any potential grammatical or lexical mistakes saves a lot of time when I have to grade multiple assignments per week. I can just copy and paste the students’ work directly into the program and have it produce a detailed list of mistakes and provide explanations on how to correct them. Of course, I still have to check both ChatGPT’s output and the students’ work manually, but rather than going over the exercises multiple times, now I just check them once or twice. Additionally, I can often just use the automated feedback ChatGPT provides instead of writing my own comments. This is especially helpful for exercises where I can provide a key to the program and have it check against it.”

However, it was noted that it was more difficult to assess exercises which relied on open-ended questions, rephrasing sentences, and connecting sentences with linking words as ChatGPT often made mistakes when assessing these kinds of tasks. A high school teacher provided her experience in this regard:

“Yes, ChatGPT can be of great help in finding grammatical mistakes, but I have noticed that it struggles to assess exercises that require something to be rephrased, or exercises that require two sentences to be linked together. It often ignores the original intent of the exercise and generates nonsensical responses or answers that include linking words or phrases that weren’t even targeted in the exercises.”

This mirrors the observations of Koraishi (2023), who noted AI’s ability to reliably assess structured language exercises, while acknowledging its shortcomings with open-ended tasks that required understanding context.

Reading and Listening Comprehension

A particularly interesting finding is that the interviewed teachers employed ChatGPT to assess reading and listening in nearly identical ways. If provided with a key to the reading and listening exercises (by giving ChatGPT a transcript of the text or giving it a preexisting key to the exercises), ChatGPT could check one-word, fill-in-the-blanks, and multiple-choice exercises with some degree of accuracy. However, long-form and open-ended questions still needed to be checked manually. Additionally, exercises that required students to summarize what they had heard or read were also difficult to assess, as ChatGPT could not effectively decide which parts of the text were important and which could be omitted in the summaries. A university-level professor teaching ESP (English for Specific Purposes) had the following to say about assessing reading and listening comprehension:

“When I make listening and reading comprehension exercises for my students, I almost always provide them with multiple-choice answers and create a key. That way, I can give ChatGPT the key and have it check the answers for me. ChatGPT follows the key accurately for the most part, but it’s not much faster than checking manually, since I already have the key ready. What it can’t do is assess tasks that require open-ended responses, as it does not have enough context from the text to evaluate if the provided answer was correct. Maybe if I started giving it a transcript of the text first, it would be able to do it, but I have not tried this yet.”

Another teacher shared an observation about the type of answers it produced for listening and reading comprehension, stating that it often selected answer b) as the correct response for multiple-choice questions.

“An interesting observation that I have made during my testing is that ChatGPT seems heavily biased towards selecting one letter as the correct answer. In my case, it almost always generated questions where answer b) was the correct response several times in a row, and only occasionally generated questions where a) or c) were correct responses. When I instructed it to not choose b) as often, it started generating questions where the answer was c) instead.”

This information on reading and listening comprehension is quite valuable because there was a notable gap in previous research on how teachers assessed these skills with ChatGPT (LO et al. 2024).

Writing and Critical Thinking

Most interviewees confirmed that they had used ChatGPT to assess writing to some degree in the past, but explained they did not use it to assess the creativity or logic of the texts. They mostly relied on ChatGPT to help them catch grammatical or lexical mistakes they had missed. The primary reason given for this was that they doubted AI could accurately interpret these aspects. One female university teacher clarified by saying:

“For writing, I think it’s great that it can call attention to the mistakes that I missed initially, and sometimes it considers factors that I wouldn’t think about assessing at all. But it tends to value all parts of writing equally, whereas I would personally award fewer points for vocabulary and grammar and give more points for the type of argument the student wrote.”

The hesitation to use AI to assess creativity and critical thinking in many ways echoes previous findings by Yoon et al. (2023), which highlighted that AI lacked the nu-

ance to provide adequate feedback on these skills.

Speaking Skills

Speaking skills were the least frequently assessed using ChatGPT. The interviewed teachers explained that they were unsure if ChatGPT could process voice input and assess it correctly due to the students' accents. Moreover, the interviewees mentioned that speaking assessment is something that they can assess and provide immediate feedback on in the classroom, which, to a degree, negates much of ChatGPT's appeal. However, one older university lecturer commented:

"I think AI could be a great tool for students that struggle with shyness or anxiety, particularly those who freeze up whenever we ask them to speak in front of the class. They can use AI to practice their speaking skills in a safe and comfortable environment where they won't feel pressured by everyone looking at them."

Unfortunately, due to having to use third-party software to get ChatGPT to process voice input (MUNIANDY, SELVANATHAN 2024), the interviewees did not provide additional information that could help address the gap in how AI can be used to assess speaking skills.

Reliability and Limitations

Regarding ChatGPT's reliability and limitations, the biggest concerns were that ChatGPT provided too generic or vague feedback and that it often misunderstood context or meaning. Additionally, ChatGPT would sometimes hallucinate mistakes and often confuse the teachers, as they could not find the mistakes it was stating the students had made. Additionally, nearly a third of respondents reported that they needed to adjust and correct ChatGPT's assessments. One teacher shared the following anecdote:

"I remember one occasion where I tried to provide ChatGPT with a model test and asked it to generate a new one following the model. I provided it with very detailed instructions on what to do, but it ended up just duplicating the test I had just given it. I then told it that it had provided me with a duplicate test and ignored my instructions and asked it to regenerate the test. The second time it had successfully changed half the test, but the latter half of the test was still a duplicate of the original. Finally, it succeeded on the third attempt. I then asked ChatGPT why it had failed to generate the content accurately the first time, to which it responded that its memory was limited and could not handle the entire prompt at once."

Lack of Technical Knowledge and Ethical Concerns

Another concern expressed during the interviews was that Serbian ESL teachers do not feel confident in their own technical ability when using AI tools. It was emphasized that the biggest challenge teachers faced was their own lack of training and knowledge on how to use them effectively. Moreover, many teachers expressed that they had ethical concerns about using AI, as they did not think it was fair or ethical to use it to assess their students' work. There were also concerns that many educational institutions did not provide clear guidelines for integrating AI or whether it could be used at all. One teacher shared her experiences regarding this issue:

"Our university encourages us to use AI tools to help us with material creation and student assessment when possible. The university wants to integrate AI to assist with

student learning in the future. The goal is to use AI to help the students find gaps in their knowledge and help them find out what skills or areas they still need to improve in. However, we have not been given any training or specific guidelines on how to use AI, and it feels like we are expected to grasp how to use it by ourselves. It would be helpful if I were at least given some training on these tools if I am expected to use them.”

Discussion

The combined findings from the survey and the interviews indicate that Serbian ESL teachers are cautiously selective in their use of ChatGPT for assessment. Quantitative results from the survey showed that most respondents either did not use ChatGPT for assessment or used it infrequently. However, when ChatGPT was utilized, it was primarily used for structured, rule-based tasks such as grammar and vocabulary checking. Considerably lower usage was noted for reading, listening, writing, speaking, and higher-order thinking skills. Logistic regression and correlation analyses further indicated that demographic factors such as age, teaching experience, and self-reported technical skill did not significantly influence the use of ChatGPT, whereas prior familiarity and experience with ChatGPT were associated with greater adoption, particularly for grammar, vocabulary, and reading comprehension tasks. This suggests that the main driver of ChatGPT’s integration into assessment practices is engagement with the tool itself, rather than background variables.

Data from interviews further reinforce this pattern of targeted, low-risk use. Teachers reported relying on ChatGPT to support the assessment of simple, text-based tasks such as fill-in-the-blanks, cloze tests, and short-answer tasks, especially those they could provide a key for. They emphasized that ChatGPT was useful when scanning for grammatical and lexical errors and generating feedback, but most still reserved final judgment for themselves. At the same time, interviewees described systematic difficulties when ChatGPT was applied to open-ended responses, paraphrasing tasks, sentence-linking tasks, or exercises that required interpretation of intent, as ChatGPT often misinterpreted the task, lacked sufficient context to provide an adequate answer, or provided fabricated information. These findings are consistent with earlier studies by Koraishi (2023), Mizumoto and Eguchi (2023), and Reza Abedi et al. (2025). Similar concerns arose regarding the assessment of reading and listening, as teachers found that ChatGPT could check one-word or multiple-choice questions if provided with a key, but was not dependable for evaluating open-ended comprehension tasks. The data on ChatGPT’s use for reading and listening tasks were particularly noteworthy, as it directly addresses the gap in the literature noted by Lo et al. (2024). Speaking was rarely assessed with ChatGPT, with teachers either unsure of its ability to handle spoken input or viewing in-person evaluation as more appropriate.

Perceptions of reliability, fairness, ethicality, and limitations provide a further explanation for this constrained use. Only a minority of survey participants agreed that ChatGPT offers reliable assessment of grammar and syntax errors, and very few trusted it to evaluate student writing beyond the surface level. Respondents frequently identified overly generic feedback, misinterpretation of context, and hallucinated errors as recurring issues, which are echoed in interview anecdotes describing duplicated tasks and misunderstood instructions. Most teachers reported modifying ChatGPT’s assessments

(either using them only as a starting point or editing them for increased accuracy), while only a small proportion accepted its assessments without applying changes.

Finally, both data sources highlight an awareness of personal and institutional constraints. A substantial number of survey respondents reported low confidence in their AI-related skills and identified lack of training as the main challenge, with many explicitly expressing a desire for formal guidance. Interviews corroborated this sense of underpreparedness and added concerns about unclear or absent institutional policies, leaving teachers to feel uncertain about how far they are allowed or expected to integrate AI into assessment. At the same time, most teachers viewed AI-based assessment as “somewhat fair” only when combined with human oversight and indicated that they would recommend ChatGPT to colleagues with caution, emphasizing the need to protect students from potential inaccuracies.

Summary and Conclusion

This study explored Serbian ESL teachers’ experiences, habits, and attitudes regarding the use of ChatGPT as a tool for assessment in the classroom. By using a mixed-methods approach that combined both quantitative data from the survey and qualitative data from several interviews, this paper revealed that most Serbian ESL teachers are still cautious about fully integrating ChatGPT into their assessment practices. However, there is an interest in its potential, especially for assessing structured and rule-based tasks such as grammar and vocabulary. The unique contribution of this paper lies in its exploration of how AI is used in the Serbian educational context, therefore addressing a notable geographical gap in previous literature.

The survey indicated that grammar and vocabulary were the most assessed skills using ChatGPT, while higher-order thinking skills, speaking, and listening comprehension were the least assessed. When asked about the reason for this in follow-up interviews, many participants explained that ChatGPT was effective in assessing textual input and rule-based tasks, but that it was less effective in assessing skills that required creativity and critical thinking, as it was unable to establish reliable criteria for assessing those abilities.

The findings of this paper have practical implications and could offer specific recommendations to stakeholders. Namely, many teachers expressed that they did not feel confident in their AI-related technical skills and would like to receive formal training on how to use these tools effectively. This sends a clear signal that educational institutions should provide formal training programs and seminars that focus on enhancing ESL teachers’ AI competencies. Moreover, the participants clarified that clear institutional guidelines and ethical frameworks need to be developed before they would be willing to use AI tools more regularly in the classroom. Policymakers could use these insights to support the creation of new policies that promote AI integration as a supplement to human assessment rather than as a replacement, thus ensuring the quality of education and maintaining the critical pedagogical role of teachers.

This research was ultimately limited by the number of participants (n=53), and further research of this kind would benefit from having a higher number of participants as well as moving beyond self-reports and experimentally comparing ChatGPT’s performance with human ratings across closed and open-ended tasks, as well as different skills to pinpoint whether it can be utilized as a reliable supplementary assessor. Additionally,

future studies should test whether targeted training in prompt design, rubric alignment, and verification strategies leads to more accurate and pedagogically consistent use. Because teachers highlighted institutional ambiguity and ethical concerns, comparative research is also needed to examine how clear policies and support structures shape actual classroom practices and trust in AI-assisted assessment.

Works Cited

- ALBUQUERQUE DA SILVA, Daisy, et al. "Analysis of the effectiveness of large language models in assessing argumentative writing and generating feedback." *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, 2024, pp. 573–582, <https://doi.org/10.5220/0012466600003636>.
- BALASKAS, Stefanos, et al. "Determinants of CHATGPT adoption intention in higher education: Expanding on Tam with the mediating roles of trust and risk." *Information*, vol. 16, no. 2, 22 Jan. 2025, p. 82, <https://doi.org/10.3390/info16020082>.
- BUCOL, Junifer Leal, and Napattanissa SANGKAWONG. "Exploring chatgpt as a writing assessment tool." *Innovations in Education and Teaching International*, vol. 62, no. 3, 7 June 2024, pp. 867–882, <https://doi.org/10.1080/14703297.2024.2363901>.
- CONDON, William. "Large-scale assessment, locally-developed measures, and Automated Scoring of essays: Fishing for red herrings?" *Assessing Writing*, vol. 18, no. 1, Jan. 2013, pp. 100–108, <https://doi.org/10.1016/j.asw.2012.11.001>.
- DAHRI, Nisar Ahmed, et al. "Extended Tam based acceptance of ai-powered chatgpt for supporting metacognitive self-regulated learning in education: A mixed-methods study." *Heliyon*, vol. 10, no. 8, Apr. 2024, <https://doi.org/10.1016/j.heliyon.2024.e29317>.
- DAVIS, Fred D. "Perceived usefulness, perceived ease of use, and user acceptance of Information Technology." *MIS Quarterly*, vol. 13, no. 3, Sept. 1989, p. 319, <https://doi.org/10.2307/249008>.
- DIKLI, Semire. "An Overview of Automated Scoring of Essays." *The Journal of Technology, Learning, and Assessment*, vol. 5, no. 1, 16 Aug. 2006, <https://ejournals.bc.edu/index.php/jtla/article/view/1640>
- HOSSAIN, Kamal, and Abdullah AL YOUNUS. "Teachers' perspectives on integrating CHATGPT into EFL writing instruction." *TESOL Communications*, 6 Jan. 2025, <https://doi.org/10.58304/tc.20250103>.
- HUSSEIN, Mohamed Abdellatif, et al. "Automated language essay scoring systems: A literature review." *PeerJ Computer Science*, vol. 5, 12 Aug. 2019, <https://doi.org/10.7717/peerj-cs.208>.
- JEON, Jaeho, et al. "A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models." *Interactive Learning Environments*, vol. 32, no. 8, 5 May 2023, pp. 4613–4631, <https://doi.org/10.1080/10494820.2023.2204343>.
- KORAISHI, Osama. "Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment." *Language Education and Technology (LET Journal)*, vol. 3, no. 1, 12 May 2023, pp. 55–72, <https://langedutech.com/letjournal/index.php/let/article/view/48>
- KOLTOVSKAIA, Svetlana. "Student engagement with automated written corrective feedback (AWCF) provided by grammarly: A multiple case study." *Assessing Writing*, vol. 44, Apr. 2020, p. 100450, <https://doi.org/10.1016/j.asw.2020.100450>.
- LAGAKIS, Paraskevas, and Stavros DEMETRIADIS. "Automated Essay Scoring: A review of the field." 2021 *International Conference on Computer, Information and Telecommunication*

- Systems (CITS)*, 11 Nov. 2021, pp. 1–6, <https://doi.org/10.1109/cits52676.2021.9618476>.
- LANTOLF, James P., and Tracy G. BECKETT. "Sociocultural theory and Second language acquisition." *Language Teaching*, vol. 42, no. 4, Oct. 2009, pp. 459–475, <https://doi.org/10.1017/s0261444809990048>.
- LO, Chung Kwan, et al. "Exploring the application of Chatgpt in ESL/EFL Education and Related Research Issues: A systematic review of empirical studies." *Smart Learning Environments*, vol. 11, no. 1, 7 Nov. 2024, <https://doi.org/10.1186/s40561-024-00342-5>.
- MAHAPATRA, Santosh. "Impact of CHATGPT on ESL students' academic writing skills: A mixed methods intervention study." *Smart Learning Environments*, vol. 11, no. 1, 13 Feb. 2024, <https://doi.org/10.1186/s40561-024-00295-9>.
- MARSH RUNYON, K. Rebecca, et al. "ETS® AI Labs™ Ways of Working Tutorial: How to build evidence-based, user-obsessed, AI-Enabled Learning Solutions in an agile framework." *Lecture Notes in Computer Science*, 2022, pp. 119–122, https://doi.org/10.1007/978-3-031-11647-6_21.
- MIZUMOTO, Atsushi, and Masaki EGUCHI. *Exploring the Potential of Using an AI Language Model for Automated Essay Scoring*, 2023, <https://doi.org/10.2139/ssrn.4373111>.
- MUNIANDY, Jayanthi, and Mahiswaran SELVANATHAN. "CHATGPT, a partnering tool to improve ESL learners' speaking skills: Case study in a public university, Malaysia." *Teaching Public Administration*, vol. 43, no. 1, 29 Jan. 2024, pp. 4–20, <https://doi.org/10.1177/01447394241230152>.
- RADFORD, Alec, et al. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, vol. 1, no. 8, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- REZA ABEDI, Seyed, et al. "Harnessing Artificial Intelligence for ESL assessments: Efficiency, challenges, and future directions." *Language, Technology, and Social Media*, vol. 3, no. 1, 3 Feb. 2025, pp. 119–130, <https://doi.org/10.70211/ltsm.v3i1.83>.
- VENKATESH, Viswanath, et al. "User acceptance of information technology: Toward a unified view." *MIS Quarterly*, vol. 27, no. 3, 2003, p. 425, <https://doi.org/10.2307/30036540>.
- WANG, Zhiyi, et al. "Assessing phrase break of ESL speech with pre-trained language models and large language models." *INTERSPEECH 2023*, 20 Aug. 2023, pp. 4194–4198, <https://doi.org/10.21437/interspeech.2023-910>.
- YOON, Su-Youn, et al. "Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion." *arXiv*, 10 Oct. 2023. <https://doi.org/10.48550/arXiv.2310.06505>
- ZAIARNA, Inna, et al. "Chatgpt in Foreign Language Teaching and assessment: Exploring EFL instructors' experience." *Information Technologies and Learning Tools*, vol. 102, no. 4, 30 Sept. 2024, pp. 176–191, <https://doi.org/10.33407/itlt.v102i4.5716>.

Алекса Ђ. Стошић

Марија В. Милојковић

ОЦЕЊИВАЊЕ У НАСТАВИ ЕНГЛЕСКОГ ЈЕЗИКА КАО СТРАНОГ: СТАВОВИ СРПСКИХ НАСТАВНИКА ПРЕМА УПОТРЕБИ ЧЕТ ЏИ-ПИ-ТИЈА

Резиме

Вештачка интелигенција и језички модели почели су све чешће да се користе у настави енглеског језика као страног од како су постали доступни јавности. Скорашња истраживања указују да ови алати утичу и на начин оцењивања знања тако што омогућују прецизније, ефикасније и персонализоване процене. Ипак, и даље се испитује како и у којој мери наставници користе ове алате, и да ли разумеју како да их употребе на ефективан начин. Овим истраживањем испитујемо како наставници енглеског у Србији дожи-

вљавају и користе Чет Џи-Пи-Ти као алат за оцењивање ученичких радова. Истраживање примењује комбиновани метод, анкету са 53 учесника и 10 пост-анкетних интервјуа, како бисмо испитали склоности, навике и ставове наставника у вези коришћења Чет Џи-Пи-Тија током оцењивања. Резултати указују да наставници најчешће примењују вештачку интелигенцију при оцењивању граматике и вокабулару, док се вештине попут говора, слушања и критичког мишљења ретко оцењују уз помоћ вештачке интелигенције. Наставници указују умерено поверење у способност вештачке интелигенције да уочи граматичке и синтаксичке грешке, али такође указују на одређена ограничења. Ограничења обухватају способност вештачке интелигенције да пружи тачну процену, њену склоност ка пружању сувише генералних повратних информација и недостатак разумевања контекста. Иако многи схватају да ови алати могу да уштеде време при оцењивању, истовремено наглашавају сопствено ограничено искуство са употребом алата вештачке интелигенције и истицали потребу за људским надзором, формалном обуком и механизмима који би обезбедили њихову етичку употребу. Резултати указују да Чет Џи-Пи-Ти још нема широку примену у српским учионицама. Наставници сматрају да га углавном треба користити као допунски алат за пружање подршке наставницима, али и да је неопходно обезбедити професионалну обуку за његову употребу пре него што би га спремно увели у наставу.

Кључне речи: Чет Џи-Пи-Ти, Вештачка интелигенција, Настава енглеског језика, оцењивање, ставови наставника