/ ISSN: 1994 – 4233 / E-ISSN: 2181–1237
Uzbekistan State World Languages University
2025 Vol. 1 No. M 3-4
LINGUISTICS
DOI: 10.36078/987656149

Оминахон Муроджонова

Научный руководитель: Иботова З.А.
Узбекский государственный университет мировых языков
E-mail: ominakhanmurodjonova@gmail.com
ORCID: https://orcid.org/0009-0004-9404-0317

ВЛИЯНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ТЕСТИРОВАНИЕ АНГЛИЙСКОГО ЯЗЫКА: ОПЫТ ПРЕПОДАВАТЕЛЕЙ УЗБЕКИСТАНА

Аннотация. Интеграция искусственного интеллекта в языковое тестирование трансформирует образовательную практику, однако восприятие этих технологий преподавателями английского языка в Центральной Азии остаётся малоизученным. Целью исследования является анализ опыта использования AI-инструментов тестирования английского языка преподавателями вузов Узбекистана и выявление основных преимуществ и вызовов данной технологии. Исследование проводилось методом смешанного дизайна с участием 94 преподавателей английского языка из 12 университетов Узбекистана в период с сентября по ноябрь 2024 года. Методология включала онлайн-анкетирование с использованием шкалы Лайкерта и полуструктурированные интервью с 15 респондентами. Анализ показал, что 78.7% преподавателей используют AI-платформы для проверки письменных работ, преимущественно Grammarly (89.4% пользователей) и Turnitin (45.7%). Средняя оценка эффективности AI-тестирования составила 4.1 из 5. Основные преимущества: экономия времени на 35-40%, повышение объективности оценивания (отметили 68.1% респондентов), мгновенная обратная связь. Однако выявлены серьёзные вызовы: ограниченная оценка творческого мышления (72.3% респондентов), технические барьеры доступа к платным версиям (83.0%), этические опасения относительно конфиденциальности данных (54.3%). Исследование демонстрирует необходимость гибридных моделей оценивания, сочетающих AI с экспертной оценкой человека. Результаты могут быть использованы для разработки политики внедрения AI-технологий в систему высшего образования Узбекистана и профессионального развития преподавателей.

Ключевые слова: искусственный интеллект, тестирование английского языка, автоматизированное оценивание, Grammarly, преподаватели вузов, Узбекистан, гибридная модель оценивания.

Ominakhan Murodjonova

Supervisor: Ibotova Z.A.
Uzbekistan State World Languages University
E-mail: ominakhanmurodjonova@gmail.com
ORCID: https://orcid.org/0009-0004-9404-0317

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON ENGLISH LANGUAGE TESTING: UZBEKISTAN TEACHERS' EXPERIENCE

Abstract. The integration of artificial intelligence into language assessment is transforming educational practices, yet teachers' perceptions of these technologies in Central Asia remain understudied. This research analyzes the experience of AI-powered English testing tools among

university teachers in Uzbekistan and identifies key benefits and challenges of this technology. The study employed a mixed-method design with 94 English teachers from 12 universities in Uzbekistan between September and November 2024. Methodology included online survey using Likert scale and semi-structured interviews with 15 respondents. Analysis revealed that 78.7% of teachers use AI platforms for assessing written work, predominantly Grammarly (89.4% of users) and Turnitin (45.7%). Mean effectiveness rating of AI-based testing was 4.1 out of 5. Key advantages included: time savings of 35-40%, enhanced objectivity in grading (noted by 68.1% of respondents), and immediate feedback. However, significant challenges emerged: limited assessment of creative thinking (72.3% of respondents), technical barriers to accessing paid versions (83.0%), and ethical concerns regarding data privacy (54.3%). The study demonstrates the necessity for hybrid assessment models combining AI with human expert evaluation. Findings can inform policy development for AI technology integration into Uzbekistan's higher education system and teacher professional development.

Keywords: artificial intelligence, English language testing, automated assessment, Grammarly, university teachers, Uzbekistan, hybrid assessment model

Ominaxon Murodjonova

Ilmiy rahbar Ibotova Z.A.
O‘zbekiston davlat jahon tillari universiteti
E-mail: ominakhanmurodjonova@gmail.com
ORCID: https://orcid.org/0009-0004-9404-0317

SUN’IY INTELLEKTNING INGLIZ TILI TESTLASH JARAYONIGA TA’SIRI: O‘ZBEKISTON O‘QITUVCHILARINING TAJRIBASI

Annotatsiya. Sun’iy intellektning tilni baholash jarayoniga integratsiyalashuvi ta’lim amaliyotini sezilarli darajada o‘zgartirmoqda, biroq Markaziy Osiyo mintaqasida o‘qituvchilarning mazkur texnologiyalarga nisbatan munosabati va amaliy tajribasi yetarli darajada o‘rganilmagan. Ushbu tadqiqotning maqsadi O‘zbekiston oliy ta’lim muassasalarida faoliyat yuritayotgan ingliz tili o‘qituvchilarining AI-vositalardan foydalangan holda testlash tajribasini tahlil qilish hamda ushbu texnologiyaning asosiy afzalliklari va muammolarini aniqlashdan iborat. Tadqiqot 2024-yil sentyabr–noyabr oylarida O‘zbekistonning 12 ta universitetidan 94 nafar ingliz tili o‘qituvchisi ishtirokida aralash dizayn (mixed-methods) asosida olib borildi. Metodologiya Likert shkalasiga asoslangan onlayn so‘rovnoma hamda 15 nafar respondent bilan o‘tkazilgan yarim tuzilgan intervyularni o‘z ichiga oldi. Tahlil natijalari shuni ko‘rsatdiki, o‘qituvchilarning 78,7 foizi yozma ishlarni tekshirish jarayonida AI-platformalardan foydalanadi, bunda eng ko‘p qo‘llaniladigan vositalar Grammarly (respondentlarning 89,4 foizi) va Turnitin (45,7 foiz) hisoblanadi. AI-asosidagi testlash samaradorligining o‘rtacha bahosi 5 ballik tizimda 4,1 ballni tashkil etdi. Asosiy afzalliklar sifatida vaqtni 35–40 foizga tejash, baholashda ob’yektivlikning oshishi (respondentlarning 68,1 foizi qayd etgan) hamda darhol qaytarilma aloqaning ta’minlanishi ko‘rsatildi. Shu bilan birga, bir qator muhim muammolar ham aniqlangan: ijodiy fikrlashni baholash imkoniyatlarining cheklanganligi (72,3 foiz), pulli versiyalarga kirishdagi texnik to‘siqlar (83,0 foiz) hamda ma’lumotlar maxfiyligi bilan bog‘liq axloqiy xavotirlar (54,3 foiz). Tadqiqot natijalari AI texnologiyalari va inson ekspert baholashini uyg‘unlashtiruvchi gibrid baholash modellarini joriy etish zarurligini ko‘rsatadi. Olingan xulosalar O‘zbekiston oliy ta’lim tizimida AI-texnologiyalarni tatbiq etish siyosatini ishlab chiqish hamda o‘qituvchilarning kasbiy rivojlanishini qo‘llab-quvvatlashda amaliy ahamiyatga ega bo‘lishi mumkin.

Kalit so‘zlar: sun’iy intellekt, ingliz tili testi, avtomatlashtirilgan baholash, Grammarly, universitet o‘qituvchilari, O‘zbekiston, gibrid baholash modeli.

INTRODUCTION

The rapid advancement of artificial intelligence technologies has fundamentally reshaped language education and assessment practices worldwide. Within the past decade, AI-powered tools have transitioned from experimental applications to mainstream educational resources, offering unprecedented capabilities for automated scoring, instant feedback generation, and adaptive testing. Natural language processing algorithms now enable machines to evaluate linguistic competence with increasing sophistication, analyzing not only grammatical accuracy but also lexical diversity, coherence, and pragmatic appropriateness [Burstein et al., 2024]. This technological evolution presents both transformative opportunities and complex challenges for language educators, particularly in contexts where resources and infrastructure may be limited.

English language testing occupies a critical position in global education, serving as a gatekeeper for academic advancement, professional mobility, and international communication. Traditional assessment methods, while theoretically sound, often suffer from practical limitations including scorer inconsistency, time-intensive evaluation processes, and restricted scalability. The emergence of AI-based testing platforms promises to address these constraints through automated scoring systems that can process thousands of responses simultaneously while maintaining standardized evaluation criteria [Sharadgah & Sa'di, 2022]. Platforms such as Grammarly, Turnitin, Duolingo English Test, and ELSA Speak have gained widespread adoption, fundamentally altering how language proficiency is measured and developed.

Recent research demonstrates that AI technologies in language assessment encompass multiple dimensions: automated writing evaluation systems that analyze essay quality across various linguistic features, speech recognition software that assesses pronunciation and fluency, adaptive testing algorithms that adjust question difficulty based on learner performance, and feedback generation mechanisms that provide detailed error analysis [Huang et al., 2023]. These technologies leverage machine learning models trained on extensive corpora of human-evaluated language samples, enabling them to replicate expert judgment patterns with increasing accuracy. Studies report correlation coefficients between AI scores and human raters ranging from 0.75 to 0.92 for writing assessment, suggesting substantial agreement in many contexts [Isaacs et al., 2023].

However, the integration of AI into language testing is not without controversy. Scholars have raised substantive concerns regarding construct validity, questioning whether automated systems truly measure communicative competence or merely surface-level linguistic features [Chapelle, 2025]. Critics argue that current AI technologies struggle to evaluate higher-order skills such as rhetorical effectiveness, cultural sensitivity, creativity, and strategic communication. Furthermore, issues of algorithmic bias, data privacy, test security, and digital inequality have emerged as critical ethical considerations that demand careful attention from educational institutions and policymakers [Burstein, 2023].

In the context of Uzbekistan, the adoption of AI-based English testing tools represents a particularly significant development. As the nation pursues ambitious educational reforms and seeks to enhance English proficiency among its population, university teachers face the challenge of integrating emerging technologies into their assessment practices. Yet empirical research examining how Uzbekistani educators perceive, utilize, and evaluate AI-powered testing platforms remains notably scarce. Existing studies predominantly focus on Western educational contexts, leaving a substantial gap in understanding how these technologies function in Central Asian settings characterized by distinct linguistic, cultural, and infrastructural conditions.

This study addresses this gap by investigating the experiences of English language teachers in Uzbekistan's higher education institutions with AI-based testing tools. The research examines which technologies are currently being employed, how educators assess their effectiveness, what advantages they recognize, and what obstacles they encounter. By centering teacher perspectives, this study contributes valuable insights into the practical realities of AI integration in language assessment, moving beyond theoretical discussions to document actual usage patterns and perceptions.

The specific research questions guiding this investigation are: (1) What AI-powered English testing tools are currently utilized by university teachers in Uzbekistan, and for what purposes? (2) How do teachers perceive the effectiveness of these tools in assessing different language skills? (3) What advantages and challenges do teachers identify in implementing AI-based assessment? (4) What implications emerge for policy and professional development in integrating AI technologies into language testing practices?

METHODS

Research Design. This study employed a convergent mixed-methods design, integrating quantitative survey data with qualitative interview findings to provide a comprehensive understanding of teachers' experiences with AI-based English testing tools. The mixed-methods approach was selected to capitalize on the strengths of both paradigms: quantitative data enabled identification of usage patterns and statistical trends across a broader sample, while qualitative data provided nuanced insights into teachers' reasoning, contextual factors, and lived experiences [Creswell & Clark, 2018]. Data collection occurred simultaneously, with both components given equal priority in the analytical process.

Participants. The study involved 94 English language teachers from 12 universities across Uzbekistan, including institutions in Tashkent, Samarkand, Bukhara, and Fergana. Participants were recruited through professional networks and university English departments using purposive sampling with snowball extension. Inclusion criteria required participants to have at least two years of teaching experience at the university level and to have used at least one AI-based tool for language assessment purposes within the past academic year. The sample comprised 67 female and 27 male teachers, with teaching experience ranging from 2 to 28 years (mean = 8.4 years). Educational backgrounds included 31 participants with doctoral degrees, 47 with master's degrees, and 16 with bachelor's degrees. All participants provided informed consent, and the research protocol received ethical approval from the Uzbekistan State University of World Languages Ethics Committee.

Data Collection Instruments. The quantitative component utilized a structured online questionnaire developed specifically for this study and pilot-tested with 10 teachers not included in the final sample. The questionnaire comprised four sections: (1) demographic information and teaching background; (2) current usage of AI testing tools with multiple-choice and checklist items; (3) perceived effectiveness ratings using 5-point Likert scales (1=very ineffective to 5=very effective) across different language skills and assessment purposes; (4) advantages and challenges experienced, measured through both rating scales and ranking exercises. The instrument demonstrated acceptable internal consistency (Cronbach's alpha = 0.82 for effectiveness items).

For the qualitative component, semi-structured interviews were conducted with 15 participants selected through maximum variation sampling to ensure diversity in institutional affiliation, teaching experience, and reported AI usage patterns. Interview protocols explored participants' initial encounters with AI testing tools, decision-making processes regarding tool selection and implementation, specific examples of successes and failures, comparisons between AI and traditional assessment methods, and recommendations for future development. Interviews averaged 45 minutes in duration and were conducted in English or Russian based on participant preference, audio-recorded with permission, and transcribed verbatim.

Procedure. Data collection occurred between September and November 2024. The online questionnaire was distributed through Google Forms with a cover letter explaining the research purpose, participant rights, and confidentiality measures. Three reminder emails were sent at two-week intervals to enhance response rates. The final response rate was 68%, with 94 completed questionnaires from 138 initial invitations. Interviews were scheduled according to participant availability and conducted via Zoom or in-person at participants' institutions. Participants received a brief summary of interview topics one week in advance to facilitate reflection.

Data Analysis. Quantitative data were analyzed using descriptive statistics (frequencies, percentages, means, standard deviations) to char-

acterize tool usage patterns and effectiveness perceptions. Cross-tabulation analyses examined relationships between demographic variables and AI tool adoption. All quantitative analyses were performed using SPSS version 26. Qualitative interview data underwent thematic analysis following Braun and Clarke's six-phase framework. This involved familiarization through repeated reading of transcripts, generation of initial codes, searching for themes, reviewing and refining themes, defining and naming themes, and producing the analytical narrative. Coding was conducted independently by two researchers, with discrepancies resolved through discussion until consensus was achieved. Integration of quantitative and qualitative findings occurred during the interpretation phase, identifying areas of convergence, divergence, and complementarity between the two data sources.

RESULTS

Current Usage of AI Testing Tools. Survey data revealed that 78.7% (n=74) of participants actively incorporated AI-powered platforms into their English assessment practices. Among users, Grammarly emerged as the predominant tool, utilized by 89.4% of those employing AI technologies, followed by Turnitin at 45.7%, ChatGPT for test item generation at 28.7%, and specialized platforms such as ELSA Speak (12.8%) and Duolingo for supplementary practice (9.6%). The majority of teachers (83.0%) reported using free versions of these platforms rather than institutional subscriptions, reflecting budgetary constraints within many universities. Table 1 presents a comprehensive overview of the AI testing platforms identified in this study, their primary functions, and the percentage of respondents reporting their use.

Table 1.

AI-Powered English testing platforms used by Uzbekistan university teachers

Platform	Primary Function	Skills Assessed	% Users (n=74)
Grammarly	Grammar & spelling check	Writing	89.4%
Turnitin	Plagiarism & AI detection	Writing	45.7%
ChatGPT	Test item generation	All skills	28.7%
ELSA Speak	Pronunciation assessment	Speaking	12.8%
Duolingo	Practice & supplementary testing	All skills	9.6%
Google Gemini	Content analysis & feedback	Writing, Reading	7.4%

Regarding assessment purposes, participants primarily employed AI tools for evaluating written assignments (92.6% of users), followed by providing formative feedback to students (68.9%), detecting plagiarism and AI-generated content (51.4%), generating test items and rubrics (35.1%), and assessing pronunciation (17.6%). The integration of AI technologies was most common in writing courses and academic English programs, with less frequent application in speaking or literature classes.

Perceived Effectiveness. Teachers' effectiveness ratings indicated generally positive perceptions of AI testing tools, with an overall mean score of 4.1 out of 5 (SD = 0.7). As shown in Figure 1, AI tools were rated highest for assessing surface-level linguistic features such as grammar and mechanics, spelling, and plagiarism detection. In contrast, substantially lower ratings were assigned to higher-order skills, including content quality, argumentative coherence, cultural appropriateness, and creativity, highlighting a clear asymmetry in AI assessment capabilities.

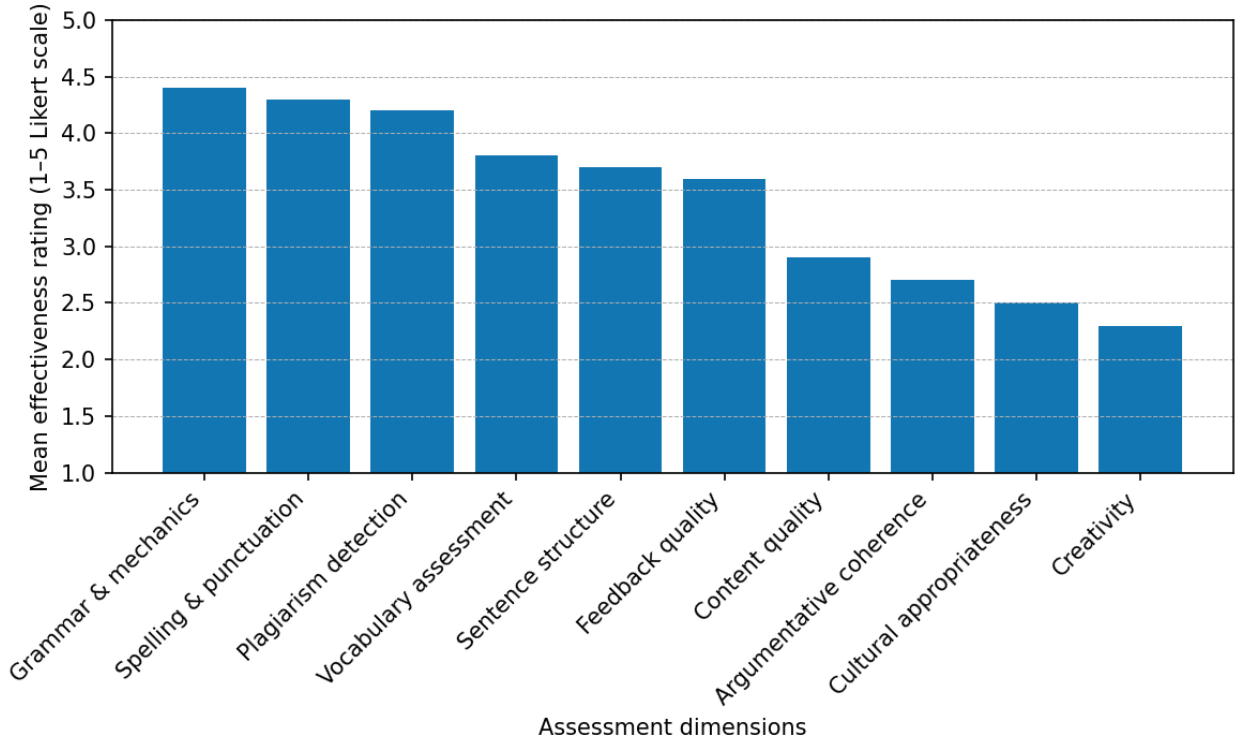


Fig. 1. Teachers' perceived effectiveness of AI-based English testing tools across assessment dimensions.

As illustrated in Figure 1, AI tools were rated as highly effective for surface-level linguistic features, while considerably lower ratings were assigned to higher-order skills such as creativity, cultural appropriateness, and argumentative coherence.

Interview participants elaborated on these quantitative patterns, explaining that AI tools excelled at identifying surface-level errors but struggled with deeper evaluative tasks. One participant noted, "Grammarly catches every misplaced comma and subject-verb disagreement, which saves enormous time. But when I need to assess whether a student's argument is well-developed or culturally sensitive, the AI feedback is generic at best, misleading at worst." Another teacher emphasized the mismatch between AI capabilities and communicative competence: "The tools measure linguistic correctness more than communicative effectiveness. A perfectly grammatical essay can still fail to achieve its rhetorical purpose."

Reported Advantages. Participants identified multiple substantial advantages of AI-based testing tools. The most frequently cited benefit was time savings, with 82.4% of respondents reporting reduced grading time. Teachers estimated that AI tools decreased their assessment workload by 35-40% for written assignments, allowing them to grade a typical class of 30 essays in 4-5 hours rather than 7-8 hours through traditional methods. This efficiency gain enabled teachers to provide more frequent formative feedback and dedicate saved time to lesson planning and individualized student support.

Enhanced objectivity represented another widely recognized advantage, noted by 68.1% of respondents. Participants appreciated that AI scoring eliminated potential biases related to handwriting legibility, student-teacher relationships, or scorer fatigue. As one interviewee explained, "Human graders have good days and bad days. We might be more lenient after coffee or harsher when tired. AI provides consistent standards regardless of when the assessment occurs." However, some teachers also acknowledged that this objectivity was limited to measurable fea-

tures and might not reflect genuine assessment validity.

Immediate feedback availability was valued by 71.3% of participants, particularly for supporting student learning outside classroom hours. Students could submit draft essays to Grammarly or Turnitin and receive instant diagnostic information, enabling iterative revision before final submission. Teachers observed that this immediate feedback loop promoted student autonomy and metacognitive awareness of their own error patterns. Additionally, 59.6% of respondents noted that detailed error reports generated by AI tools provided valuable data for identifying common weaknesses across student cohorts, informing instructional adjustments.

Identified Challenges. Despite recognized advantages, participants articulated substantial challenges in implementing AI-based assessment. The most pressing concern, expressed by 83.0% of respondents, involved financial and technical barriers. Most AI platforms require paid subscriptions for full functionality, creating an insurmountable obstacle for many Uzbekistani universities operating with constrained budgets. Free versions offered limited features, restrictive usage caps, or intrusive advertising that disrupted workflow. Furthermore, inconsistent internet connectivity in some regions posed practical difficulties for cloud-based assessment tools.

Pedagogical limitations were identified by 72.3% of participants, who emphasized that AI tools inadequately assessed higher-order thinking skills central to language competence. Teachers expressed frustration that automated systems could not meaningfully evaluate argumentative quality, rhetorical strategies, audience awareness, or cultural sensitivity. Interview participants shared examples of essays receiving high AI scores despite weak argumentation or cultural inappropriateness. One teacher recounted, "A student wrote an impeccably grammatical essay arguing an ethically problematic position. Grammarly gave it a high score because the grammar was perfect, but the content was deeply flawed."

Ethical concerns regarding data privacy and academic integrity were raised by 54.3% of respondents. Teachers questioned whether student writing submitted to commercial platforms remained confidential and worried about potential misuse of personal data. Some participants expressed discomfort with platforms that retained student essays in databases for training purposes or comparison against future submissions. Additionally, the proliferation of AI writing tools accessible to students created detection challenges; teachers reported increasing difficulty distinguishing between student-written work, AI-generated content, and AI-assisted writing.

Finally, 47.9% of participants noted professional development gaps, indicating they had received no formal training in AI tool selection, implementation, or critical evaluation. Teachers felt uncertain about interpreting AI-generated scores, explaining algorithmic decisions to students, or determining when to override automated assessments with human judgment. This lack of preparation undermined confidence and potentially limited effective tool utilization.

DISCUSSION

Interpretation of Findings. The findings of this study illuminate both the promise and complexity of AI integration in English language testing within Uzbekistan's higher education context. The high adoption rate of 78.7% demonstrates that AI-based tools have achieved substantial penetration among university teachers, reflecting global trends toward technological assessment solutions. However, the predominant reliance on free versions and limited feature sets suggests that full realization of AI's potential remains constrained by institutional resources rather than professional willingness or technological capability.

The effectiveness perception patterns align with broader research literature demonstrating that AI excels at evaluating surface linguistic features while struggling with discourse-level and pragmatic dimensions. Studies by Huang et al. (2023) and Chapelle (2025) similarly report that automated writing evaluation achieves strong correlations with human raters for grammar and mechanics but demonstrates weaker validity for rhetorical quality and argumentation. This capability-limitation profile creates a fundamental tension: the features most amenable to automated assessment may not represent the most critical

components of communicative competence that educators aim to develop.

The reported time savings of 35-40% constitute a substantial practical advantage that should not be dismissed. For teachers managing heavy course loads—common in many Uzbekistani universities—this efficiency gain represents meaningful relief from assessment burden. However, the quality of time saved matters critically. If reduced grading time enables teachers to provide more comprehensive feedback on higher-order concerns or to engage in professional development, the benefit extends beyond mere efficiency. Conversely, if time savings simply shift work elsewhere without improving educational quality, the value becomes questionable.

Implications for Validity and Fairness. The pedagogical limitations identified by teachers raise serious questions about construct validity in AI-based assessment. If automated tools primarily measure linguistic accuracy rather than communicative effectiveness, their increasing adoption risks narrowing the construct being assessed and potentially distorting curriculum and instruction. This phenomenon, often termed "teaching to the algorithm," has been documented in other educational contexts where high-stakes testing drives pedagogical decisions. Teachers may unconsciously prioritize aspects of writing that AI evaluates well—grammatical correctness, vocabulary sophistication—while de-emphasizing dimensions that AI handles poorly, such as rhetorical effectiveness or critical thinking.

The fairness concerns expressed by teachers deserve serious consideration. Research by Burstein (2023) demonstrates that AI scoring models can perpetuate biases present in training data, potentially disadvantaging learners from particular linguistic, cultural, or educational backgrounds. For instance, if training corpora predominantly represent Western rhetorical conventions, non-Western writing patterns might be systematically undervalued. In the Uzbekistan context, where Russian and Uzbek linguistic influences shape learners' English interlanguage, AI systems trained primarily on native English corpora may fail to recognize legitimate variation and creativity.

The Digital Divide and Access Inequality. The financial and technical barriers identified in this study highlight how AI technologies, despite promises of democratizing access, may paradoxically exacerbate educational inequalities. Universities with greater resources can provide institutional subscriptions to premium AI platforms, offering students access to sophisticated assessment and feedback tools. Institutions with limited budgets must rely on free versions with restricted capabilities, creating a tiered system of educational opportunity. This pattern reproduces broader concerns about digital inequality in education, where technological advancement benefits those already privileged while leaving others further behind.

Infrastructure challenges compound access inequality. In regions with unreliable internet connectivity, cloud-based assessment platforms become impractical regardless of cost. Teachers and students cannot depend on technologies that function inconsistently, leading to frustration and abandonment. This technological precarity distinguishes the Uzbekistan context from wealthier nations where high-speed internet access is ubiquitous, reminding us that AI integration strategies must account for local infrastructural realities rather than assuming universal technological capacity.

Toward Hybrid Assessment Models. The findings strongly support hybrid assessment models that strategically combine AI automation with human professional judgment. Such models would leverage AI's strengths—rapid processing of surface features, immediate preliminary feedback, consistent scoring of routine elements—while preserving human capacity for evaluating complex dimensions that require contextual understanding, cultural sensitivity, and pedagogical expertise. Research by Isaacs et al. (2023) similarly advocates for complementary rather than replacement approaches, arguing that optimal assessment systems utilize both automated and human components aligned with their respective capabilities.

Implementing effective hybrid models requires careful design. AI tools might provide initial screening of grammar and mechanics, flagging major errors for student revision before human evaluation of content and argumentation. Alternatively, AI could handle formative feedback during drafting stages while teachers con-

duct summative assessment of final products. The key principle involves matching task complexity to assessor capability—automated systems for straightforward judgments, human expertise for nuanced evaluation. However, successful implementation demands that teachers understand AI limitations and maintain critical oversight rather than uncritically accepting automated scores.

Professional Development Imperatives. The substantial professional development gap identified in this study represents a critical obstacle to effective AI integration. Teachers cannot make informed decisions about tool selection, interpret algorithmic outputs appropriately, or explain automated assessments to students without adequate training. Professional development programs must address multiple dimensions: technical literacy in AI functionality, critical evaluation of algorithmic validity and bias, pedagogical strategies for integrating automated assessment, and ethical considerations regarding data privacy and academic integrity.

Such professional development cannot consist merely of tool training—how to use Grammarly or Turnitin—but must cultivate deeper understanding of AI capabilities and limitations, enabling teachers to serve as informed mediators between technology and pedagogy. Teachers need conceptual frameworks for evaluating whether particular assessment purposes align with AI capabilities, identifying situations where human judgment remains irreplaceable, and recognizing potential algorithmic biases that might disadvantage certain learner populations.

Limitations and Future Research. This study's findings should be interpreted considering several limitations. The sample, while diverse across institutions, represents a convenience sample that may not fully reflect the broader population of English teachers in Uzbekistan. Teachers who volunteered for participation may hold more favorable attitudes toward technology than non-participants, potentially inflating adoption rates and effectiveness perceptions. Additionally, self-reported survey data may be subject to social desirability bias, with participants overstating their AI usage or competence.

Future research should examine student perspectives on AI-based assessment, investigating how learners perceive automated feedback, whether they find it helpful for improving their writing, and how AI assessment affects their motivation and learning strategies. Longitudinal studies tracking changes in student writing quality over time in AI-supported versus traditional assessment contexts would provide valuable evidence regarding learning outcomes. Comparative analysis examining AI assessment implementation across different institutional contexts—well-resourced versus under-resourced universities—could illuminate how institutional capacity shapes technological integration and educational equity.

CONCLUSION

This study investigated the experiences of university English teachers in Uzbekistan with AI-powered testing tools, documenting current usage patterns, perceived effectiveness, identified advantages, and encountered challenges. The findings reveal widespread adoption of AI technologies, particularly Grammarly and Turnitin, for assessing written work and providing formative feedback. Teachers recognize substantial benefits including significant time savings, enhanced objectivity, and immediate feedback availability. However, they simultaneously identify serious limitations regarding construct validity, higher-order skill assessment, technical access barriers, and professional preparation gaps.

These findings underscore that AI technologies represent powerful but fundamentally limited tools for language assessment. While automated systems excel at evaluating surface linguistic features with speed and consistency, they struggle with the complex, contextualized judgments that characterize expert human assessment of communicative competence. The path forward lies not in wholesale adoption or rejection of AI, but in thoughtful integration that leverages technological capabilities while preserving irreplaceable human expertise. Hybrid assessment models that strategically combine automated and human evaluation offer the most promising approach for balancing efficiency with validity.

For policymakers in Uzbekistan's education system, these findings highlight the necessity of addressing financial barriers through institutional AI subscriptions, infrastructure invest-

ment to ensure reliable internet access, and professional development programs that prepare teachers to use AI tools critically and effectively. Educational institutions should develop clear guidelines regarding acceptable AI use in assessment, data privacy protection, and academic integrity standards. Professional development initiatives should equip teachers with both technical skills and conceptual understanding to make informed decisions about when AI assessment is appropriate and when human judgment remains essential.

For teachers, the imperative is to maintain critical engagement with AI technologies rather than uncritical adoption. This involves understanding algorithmic capabilities and limitations, questioning automated scores when they conflict with professional judgment, and prioritizing pedagogical goals over technological convenience. Teachers must advocate for resources and support that enable responsible AI integration while resisting pressures to replace human expertise with automated systems in contexts where such substitution undermines educational quality.

Looking forward, the relationship between AI and language assessment will continue evolving as technologies advance and educational contexts adapt. The challenge lies in ensuring that this evolution serves educational purposes—supporting student learning, providing valid assessment, promoting equitable access—rather than being driven primarily by technological possibility or commercial interest. Achieving this requires ongoing dialogue among teachers, researchers, policymakers, and technology developers, grounded in evidence about what works, for whom, and under what conditions. This study contributes to that dialogue by documenting the lived realities of AI assessment implementation in a context that has received insufficient research attention, offering insights that may inform more equitable and effective integration of these powerful technologies.

REFERENCES

1. Alshehri, A. (2025). AI's effectiveness in language testing and feedback provision. Social Sciences & Humanities Open, 12, Article 101892. https://doi.org/10.1016/j.ssaho.2025.101892
2. Burstein, J., von Davier, A., LaFlair, G., & Yancey, K. (2024). Responsible AI for test equity and quality: The Duolingo English Test as a case study. arXiv preprint arXiv:2409.07476. https://arxiv.org/abs/2409.07476
3. Chapelle, C. A. (2025). Generative AI as game changer: Implications for language education. System, 126, Article 103288. https://doi.org/10.1016/j.system.2025.103672
4. González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. Applied Sciences, 11(12), 5467. https://doi.org/10.3390/app11125467
5. Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Exploring the impact of artificial intelligence on English language teaching: A meta-analysis. Computers & Education: Artificial Intelligence, 5, Article 100163. https://doi.org/10.1016/j.caeai.2023.100163
6. Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. Language Testing, 40(3), 748-770. https://doi.org/10.1177/02655322231158550
7. Latif, A., & Qureshi, M. S. (2024). The intersection of AI and language assessment: A study on the reliability of ChatGPT in grading IELTS writing task 2. Language Teaching Research Quarterly, 43, 22-42. https://doi.org/10.32038/ltrq.2024.43.02
8. Mahfoodh, O. H. A. (2024). Which one? AI-assisted language assessment or paper format: An exploration of the impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. Language Testing in Asia, 14, Article 32. https://doi.org/10.1186/s40468-024-00322-z

9. Mirzaei, M. S., Meshkat, M., & Namaziandost, E. (2024). Application of artificial intelligence in language skills testing. ANGLISTICUM Journal of the Association-Institute for English Language and American Studies, 13(1), 22-34. https://doi.org/10.5281/zenodo.10729638
10. Peñarroja, F. M., Luna, M. E. T., & De Torres, A. J. P. (2024). Learning English as a second language with artificial intelligence for prospective teachers: A systematic review. Frontiers in Education, 9, Article 1490067. https://doi.org/10.3389/feduc.2024.1490067
11. Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelligence in English language teaching (2015-2021). Journal of Information Technology Education: Research, 21, 337-377. https://doi.org/10.28945/4999
12. Tran, H. T. M., & Nguyen, T. L. (2025). Artificial intelligence-based assessment in ELT exam creation: A case study of Van Lang University lecturers. Saudi Journal of Language Studies, 5(1), 45-62. https://doi.org/10.1108/SJLS-06-2024-0030
13. Zhai, C., & Wibowo, S. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. System, 129, Article 103429. https://doi.org/10.1016/j.system.2025.103429