IJBMC
International Journal of Body, Mind and Culture

Volume 13, Issue 11, Article e2026-1783
DOI: 10.61838/ijbmc.v13i11.1783

Article type:
Original Research

1 Satpayev Kazakh National Technical University, Almaty, Kazakhstan.
2 Abai Kazakh National Pedagogical University, Almaty, Kazakhstan.
3 Almaty University of Power Engineering and Telecommunication after G. Daukeev, Almaty, Kazakhstan.

Corresponding author email address:
karlygash.moldabayeva@proton.me

English Discourse Performance Before and After Loora AI Practice in A2 University Students

Sagira. Odanova[1], Akbota. Zhussupova[2], Akzharkyn. Abuova[2], Kuralay. Telgozhayeva[3], Karlygash. Moldabayeva[2*]



Article history:

Received 21 Aug 2025
Revised 08 Oct 2025
Accepted 16 Feb 2026
Published online 01 Nov 2026

How to cite this article:
Odanova, S., Zhussupova, A., Abuova, A., Telgozhayeva, K., & Moldabayeva, K. (2026). English Discourse Performance Before and After Loora AI Practice in A2 University Students. International Journal of Body, Mind and Culture, 13(11), Article e2026-1783. https://doi.org/10.61838/ijbmc.v13i11.1783

ABSTRACT

Objective: To describe English discourse scores reported before and after Loora AI conversation practice among A2-level university students and establish the limits of their interpretation.

Methods and Materials: A single-group pretest–posttest study involved 60 students aged 18–20 at one university in Kazakhstan. The six-week programme scheduled three approximately 20-minute individual sessions per week. Two raters assessed five discourse dimensions using a study-specific rubric, described as 0–20 points per dimension and a summed total of 0–100.

Findings: The reported overall mean increased from 61.5 to 74.4. However, dimension means of 58.7–77.1 exceed the stated 0–20 scale. Dividing the five dimension means by 5 and summing yields 61.52 and 74.44, consistent with a possible but undocumented normalization. Reported $p < .001$ values cannot be verified without variability estimates and test output. ICC point estimates were .86–.91, but model specifications and confidence intervals were absent. Dialogue mechanisms lacked auditable coding or excerpts.

Conclusion: The aggregate pattern is provisional. Scoring confirmation, instrument validation, and complete statistical reporting are required. The uncontrolled design cannot establish effectiveness. Psychological outcomes, cultural identity, Kazakh-language performance, and cross-language transfer were not measured.

Keywords: Artificial intelligence, English language, Discourse performance, Loora, Assessment validity.

Introduction

Conversational artificial intelligence offers additional opportunities for second-language practice, but access to dialogue and demonstrable learning are separate questions. Computer-assisted language learning has developed from restricted exercises toward interactive environments, and conversational agents have long been explored as language-learning tools (Fryer & Carpenter, 2006; Warschauer & Healey, 1998). More recent work considers adaptive conversation, feedback, and the distribution of activity between learners, teachers, and digital systems (Godwin-Jones, 2023; Godwin-Jones, 2024). The educational value of a particular application nevertheless depends on the tasks, learners, assessment, and comparison conditions used to evaluate it.

Evidence concerning chatbots is heterogeneous across designs and outcomes (Huang et al., 2022). Research has examined intelligent personal assistants, speaking practice, task performance, and pragmatic learning rather than a single interchangeable language-learning effect (Dizon & Tang, 2020; Sim et al., 2025; Sok & Shin, 2025; Tai & Chen, 2024). A study of conversational practice should therefore identify what performance is assessed and whether its design can distinguish an application-specific contribution from practice more generally.

Psychological outcomes also require their own observations. Enjoyment, anxiety, willingness to communicate, speaking-related self-efficacy, and teachers' professional experiences have been investigated through designs that explicitly address those constructs (Aditya et al., 2026; Hashemi et al., 2026; Zhang et al., 2024). They cannot be inferred from a discourse score alone. Similarly, studying English performance in Kazakhstan does not by itself measure Kazakh linguistic consciousness, cultural identity, or intercultural awareness. The present study is concerned with English task performance and does not treat these psychocultural constructs as measured outcomes.

The study aimed to describe the discourse scores reported before and after a six-week programme of individual English conversation with Loora AI among A2-level university students. The intended dimensions were discourse organization, topic maintenance, interactional responsiveness, meaning clarification, and pragmatic appropriateness. A secondary interest concerned dialogue processes that might support learning. Because the score scale is inconsistently documented and no comparison group was included, the available results are interpreted as provisional descriptive observations rather than evidence of intervention effectiveness.

Theoretical Background

Communicative and interactional performance

Communicative competence connects grammatical knowledge with sociolinguistic and strategic resources for using language (Canale & Swain, 1980). The CEFR Companion Volume further describes production, interaction, mediation, and plurilingual activity (Council of Europe, 2020). These frameworks support attention to contextual language use, but they do not validate a newly assembled scoring instrument or establish that a learner has progressed between CEFR levels.

Interactional approaches emphasize the coordination of turns, responses to preceding contributions, and work to resolve difficulties. Negotiation for meaning and modified output may create opportunities for attention and learning (Long, 1996; Loewen & Sato, 2018). Empirical interaction research provides a rationale for examining these processes directly rather than assuming that opportunities necessarily produce development (Mackey, 1999). Feedback and subsequent learner uptake are likewise distinct events that require observation (Lyster & Ranta, 1997).

Pragmatic appropriateness concerns choices made in a communicative context, including relevant social relationships and purposes (Culpeper et al., 2018). Interactional organization is sequential, so an isolated response or overall score may not show how a misunderstanding was resolved (Schegloff, 2007). Comparisons of AI and human interlocutors also indicate that the interlocutor condition can shape the interactional

ijbmc.org
E-ISSN: 2345-5802

evidence elicited by a task (Su & Chen, 2026; Timpe-Laughlin et al., 2025). Performance with an AI partner should therefore not be equated automatically with performance in human conversation.

Learning context and human support

Individual AI practice is situated within a wider educational environment. Participation and a sense of community involve social and institutional conditions that extend beyond an individual task score (Yang & Seyed Alitabar, 2024). Research on digital mentorship also emphasizes communication practices and personal connection in virtual learning relationships (Seyed Alitabar & Zadhasn, 2023). These considerations support retaining human learning contexts and teacher support when evaluating conversational technology; neither source establishes a Loora-specific effect.

Materials and Methods

Design and participants

The study used an uncontrolled single-group pretest–posttest design. Sixty students aged 18–20 years were recruited by convenience sampling at a higher education institution in Kazakhstan. Eligibility comprised A2 English proficiency, the specified age range, and availability for all study stages. All 60 students were reported to have participated in the pretest, practice programme, and posttest. The institution, collection dates, gender distribution, language backgrounds, and method used to establish A2 proficiency were not specified. Outcome-specific paired-sample denominators and a response-disposition log were not provided.

There was no random assignment, comparison group, or delayed follow-up. The design can describe a temporal pattern in the same group but cannot separate a specific application effect from repeated testing, ordinary coursework, additional practice, maturation, expectancy, or other concurrent experiences. These are established concerns in single-group before-and-after research (Marsden & Torgerson, 2012).

Practice programme and assessment procedure

The programme was scheduled for six weeks, with three approximately 20-minute sessions each week: 18 sessions and approximately six hours of individual practice. Students interacted one-to-one with Loora AI rather than with peers during these sessions. Eighteen conversational tasks involved answering questions, sustaining a topic, expressing opinions, asking for additional information, clarifying misunderstandings, and responding to a communicative situation. Attendance and application logs were not supplied to verify individual exposure or completion of each scheduled task.

The reported topics included daily life, university study, leisure, friendship, food, travel, technology, future plans, health, work, the environment, shopping, and service encounters. The task sequence began with a question or situation, followed by learner and AI turns; clarification or reformulation was encouraged when needed. The topics were described as selected for A2 learners, but the complete prompts, difficulty-calibration procedure, and instructor involvement were not documented. The developer describes Loora as an English conversation application (Loora, n.d.); the study-specific application version, account configuration, and update history were not reported.

The pretest was completed without Loora. Posttest tasks were described as equivalent in content and linguistic complexity, but the full task sets, interlocutor arrangements, recording conditions, and evidence of equivalence were not supplied. These details matter because familiarity with task content or a change in interlocutor can affect observed performance. No separate human-interlocutor transfer test or assessment of Kazakh-language performance was reported.

Assessment dimensions and rubric

The study used five discourse dimensions (Table 1) and a study-specific analytic rubric with five performance bands (Table 2). Two raters were reported. The same criteria were described as applying at both assessment points. Rater qualifications, training, calibration examples, independence,

ijbmc.org
E-ISSN: 2345-5802

blinding to assessment occasion, and resolution of discrepant ratings were not specified.

Table 1

Discourse dimensions and their intended observational focus

Dimension	Intended focus
Discourse organization	Logical connections between ideas and organization of an extended response
Topic maintenance	Sustaining and developing a topic while avoiding unsupported digressions
Interactional responsiveness	Responding meaningfully to the preceding question or turn
Meaning clarification	Requesting or providing clarification, reformulating, and addressing misunderstandings
Pragmatic appropriateness	Selecting language and response strategies appropriate to the communicative situation

Note. These are the intended assessment targets. They are not separately validated constructs, and their conceptual relationship to communicative competence does not establish the validity of the study-specific scores.

Table 2

Reported analytic rubric and its original score bands

Dimension	0–4 Very limited	5–8 Limited	9–12 Developing	13–16 Good	17–20 Highly effective
Discourse organization	Ideas are fragmented	Limited organization	Partly coherent	Generally coherent	Highly coherent and well organized
Topic maintenance	Topic frequently lost	Limited topic development	Topic partly maintained	Topic generally maintained	Topic consistently maintained and developed
Interactional responsiveness	Responses often inappropriate	Limited response relevance	Partly appropriate responses	Appropriate responses	Consistently relevant and responsive
Meaning clarification	Rarely clarifies meaning	Limited clarification	Uses basic clarification	Effectively clarifies meaning	Consistently negotiates and clarifies meaning
Pragmatic appropriateness	Inappropriate responses	Limited contextual awareness	Partly appropriate	Generally appropriate	Highly appropriate to context

Note. Each dimension is described as scored from 0 to 20, and summing five raw dimensions would yield a total from 0 to 100. The within-band allocation of points, A2-level calibration, rater training, and independent validity evidence were not documented. The dimension means originally reported above 20 cannot be interpreted as raw scores on this rubric.

The rubric's conceptual connection to communicative frameworks is a rationale for its content, not independent evidence of validity. Expert content review, learner piloting, construct structure, relations with an established external assessment, and sensitivity to meaningful change were not documented. The allocation of individual points within each broad band and the interpretation of high-band performance specifically for A2 learners also remain unclear. Reliability and validity address different questions; consistency of ratings cannot by itself establish the interpretation or intended use of a score (American Educational Research Association et al., 2014).

Scoring and arithmetic reconciliation

The stated procedure assigned 0–20 points to each dimension and summed the five raw dimension scores, giving a total from 0 to 100. However, the dimension means in the reported results range from 58.7 to 77.1 (Table 3). They therefore cannot be raw means on the stated dimension scale. The original records do not identify an alternative scale, a normalization step, or the calculation syntax.

A linear reporting convention could explain the arithmetic: multiplying each raw 0–20 dimension by 5 would produce a 0–100 display score for that dimension. Under equal weighting and the same participants across domains, the overall raw total would then equal the average of the five display values. Conversely, dividing each displayed dimension mean by 5 and adding the results yields 61.52 at pretest and 74.44 at posttest, rounding to the reported totals of 61.5 and 74.4. Table 4 makes this possible relationship explicit. It is a conditional reconciliation, not confirmation that normalization was actually performed or that the participant-level scores are correct.

ijbmc.org
E-ISSN: 2345-5802

Statistical analysis and interrater estimates

The reported analysis specified paired-samples t tests, a significance threshold of .05, and intraclass correlation coefficients for the two raters. Means and arithmetic differences can be checked from the aggregate entries. The reported p values are retained as unverified claims. Standard deviations at each occasion, the standard deviation or standard error of paired changes, t statistics, actual analysis degrees of freedom, confidence intervals, and effect sizes were not supplied. These quantities cannot be reconstructed from marginal means and a p-value threshold.

Verification requires the paired raw ratings, the rule for combining raters, the transformation and total-score formulas, missing-data handling, and the original statistical output. A complete analysis should report the paired mean difference and its confidence interval, t with degrees of freedom, and an explicitly defined paired effect size; comparisons across five domains and a total also require a stated multiplicity strategy. No new inferential estimates are presented here. For the ICCs, model, definition, measurement unit, and uncertainty must be specified before interpreting consistency or absolute agreement (Koo & Li, 2016).

Status of the dialogue analysis

The study description identifies question–answer organization, negotiation, clarification, repair, feedback, and pragmatic adaptation as relevant dialogue processes. It does not provide a corpus size, sampling strategy, transcription conventions, coding unit, independently coded material, disagreement procedure, frequency counts, or contextualized extracts. The two assessment raters cannot be assumed to have served as independent qualitative coders. Consequently, the process descriptions are treated as analytic propositions, not verified qualitative findings or demonstrated mechanisms. Table 6 specifies the observations that would be required for a systematic analysis.

Finding and Results

Reported score pattern

All five reported posttest means exceed their corresponding pretest means (Table 3). On the original display scale, the differences are +12.4 for discourse organization, +11.4 for topic maintenance, +14.3 for interactional responsiveness, +15.2 for meaning clarification, and +11.3 for pragmatic appropriateness. The reported overall values are 61.5 and 74.4, a difference of +12.9. These reproduce the aggregate entries; they do not establish a validated raw-score improvement or a causal effect.

Table 3

Reported pretest and posttest means on an unconfirmed display scale

Outcome	Reported pretest M	Reported posttest M	Reported change	Reported p
Discourse organization	61.4	73.8	+12.4	< .001
Topic maintenance	64.2	75.6	+11.4	< .001
Interactional responsiveness	62.8	77.1	+14.3	< .001
Meaning clarification	58.7	73.9	+15.2	< .001
Pragmatic appropriateness	60.5	71.8	+11.3	< .001
Overall discourse interaction score	61.5	74.4	+12.9	< .001

Note. M = mean. All numerical entries reproduce the reported aggregate results; the p column contains reported, unverified paired-test claims. Dimension means exceed the stated raw maximum of 20. No standard deviations, standard errors of change, t statistics, outcome-specific degrees of freedom, confidence intervals, or effect sizes were supplied. Consequently, neither the measurement scale nor statistical significance is independently established by this table. Table 4 gives a conditional arithmetic reconciliation, not validated replacement observations.

Scale audit

Summing the five originally displayed dimension means gives 307.6 at pretest and 372.2 at posttest, exceeding the stated maximum total of 100. Their averages are 61.52 and 74.44, which explain the reported overall values after rounding. Under the conditional conversion in Table 4, the dimension means would instead range from 11.74 to 15.42 on the 0–20 scale. This demonstrates a coherent possible reporting convention, but the original scoring sheets are necessary to distinguish undocumented normalization from an erroneous aggregation or transcription.

Table 4

Conditional reconciliation with the stated zero to twenty rubric

Dimension or summary	Pretest equivalent	Posttest equivalent	Equivalent change
Discourse organization	12.28	14.76	+2.48
Topic maintenance	12.84	15.12	+2.28
Interactional responsiveness	12.56	15.42	+2.86
Meaning clarification	11.74	14.78	+3.04
Pragmatic appropriateness	12.10	14.36	+2.26
Sum of conditional raw dimension means	61.52	74.44	+12.92

Note. Each dimension entry is calculated by dividing its reported display value in Table 3 by 5. This calculation is valid as a raw-score interpretation only if the original domain values were converted linearly from 0–20 to 0–100 before reporting. Such a conversion was not documented. Under equal weighting of five domains for the same participants, the sum of these conditional raw means equals the mean of the five display means. The resulting totals round to the reported overall means. These are arithmetic equivalents, not rescored participant data; the scoring sheets and calculation syntax are needed to confirm the interpretation.

The largest displayed change is in meaning clarification, followed by responsiveness. These are descriptive differences between reported means. Without paired uncertainty estimates, instrument calibration, and a confirmed common metric, the differences between domain changes cannot be interpreted as statistically distinct gains or evidence that one interactional mechanism was more effective than another. The smaller displayed change in pragmatic appropriateness likewise does not establish slower pragmatic development.

Reported interrater coefficients

The reported ICC estimates range from .86 to .91, with .90 for the overall score (Table 5). Interpretation is limited by the missing model and confidence intervals. These coefficients do not resolve the scale inconsistency, validate the rubric, establish absolute agreement, or rule out shared rater expectations.

Table 5

Reported interrater intraclass correlation coefficients

Outcome	Reported ICC
Discourse organization	0.89
Topic maintenance	0.87
Interactional responsiveness	0.91
Meaning clarification	0.86
Pragmatic appropriateness	0.88
Overall discourse interaction score	0.90

Note. ICC = intraclass correlation coefficient. Two raters were reported. The ICC model, single- versus average-measure form, consistency versus absolute-agreement definition, confidence intervals, number of rated observations, and handling of pretest and posttest samples were not specified. Point estimates alone cannot establish absolute agreement, instrument validity, or freedom from rater expectancy effects.

Interactional processes

No auditable episode counts or dialogue excerpts are available to establish how often clarification, repair, feedback uptake, or pragmatic adaptation occurred. The process guide in Table 6 therefore distinguishes plausible opportunities within conversation from observed learner behaviour. No new learner quotations, coding frequencies, or qualitative reliability values are introduced. In particular, the mean score labelled meaning clarification does not independently

ijbmc.org
E-ISSN: 2345-5802

demonstrate that learners performed more successful clarification sequences.

Table 6

Operational guide for future analysis of dialogue processes

Process	Possible unit and coding criterion	Evidence needed before making a finding
Question and answer	An adjacency pair in which a response addresses the prior question	Aligned turns, speaker labels, and a record of relevant and nonresponsive answers
Meaning negotiation	An episode initiated by a signal of nonunderstanding and followed by an attempt to resolve it	The trouble source, repair sequence, and evidence of resolution or unresolved difficulty
Clarification	A request for, or provision of, additional information addressing an identifiable ambiguity	The preceding context and the clarification turn; distinguish learner- from AI-initiated action
Repair	Replacement, repetition, or reformulation directed at a preceding difficulty	Original and revised turns, initiation source, and evidence of uptake
Feedback	An evaluative or corrective turn followed by a learner response	Feedback type, target, subsequent learner turn, and whether uptake occurs
Pragmatic adaptation	A change in linguistic choice in relation to a specified communicative context	Task context, relevant social roles, the language choice, and a justified interpretation

Note. This table clarifies analytic constructs; it does not report completed coding, frequencies, or observed participant excerpts. Negotiation episodes can include clarification, feedback, and repair, so these categories must not be added as independent events without an explicit rule. Corpus sampling, transcription conventions, coder roles, disagreement procedures, and counterexamples must be documented before drawing empirical mechanism claims.

Discussion and Conclusion

Interpretation of the descriptive pattern

The reported aggregate pattern is compatible with higher assessed English discourse performance after a period of regular practice. Its interpretation is constrained first by unresolved scoring and incomplete statistical output, and then by the uncontrolled design. Even a fully verified paired change would not identify the portion attributable specifically to Loora. The data therefore warrant further investigation of structured conversation practice rather than a conclusion that this application is effective.

Interaction theory offers plausible explanations for what a future analysis could examine. Asking for clarification, revising an answer, and responding to feedback may provide opportunities to connect form and meaning (Long, 1996; Loewen & Sato, 2018; Lyster & Ranta, 1997). The current materials do not demonstrate mediation by those processes. Aggregate score changes and theoretical plausibility cannot substitute for coded sequences showing a trouble source, a response, and subsequent uptake or resolution.

The broader literature supports separating task design, interlocutor characteristics, and outcome definition. ChatGPT-based and peer interactions have been compared for specific speech acts, with differing patterns across tasks and retention periods (Sim et al., 2025). Controlled work on speaking and summarizing also distinguishes task conditions rather than assuming a universal chatbot effect (Sok & Shin, 2025). Other studies address speech-act performance and AI or human interactional assessment (Rahman et al., 2025; Su & Chen, 2026). These findings motivate stronger evaluation; they do not validate the present rubric, repair its scores, or establish that results transfer between platforms.

Psychological and cultural limits

The participants studied in Kazakhstan and practised English. Their national or educational location is contextual information, not a measure of cultural cognition. No Kazakh proficiency, linguistic identity, cultural awareness, cultural adaptation, or transfer between English and Kazakh was assessed. The study therefore cannot demonstrate a change in Kazakh linguistic consciousness or psychocultural functioning. Its empirical contribution remains within language-education assessment, and a stronger psychological or cultural contribution would require directly measured constructs and a design capable of addressing them.

Speaking-related beliefs and emotional responses are relevant questions for future research, but they should be assessed separately from observed language performance. Research on EFL self-efficacy and on teachers' responses to AI shows the importance of attending to participants' own accounts and the population being studied (Aditya et al., 2026; Hashemi et al., 2026). Teachers' identity concerns are not interchangeable with students' confidence, and neither is measured by the five discourse dimensions. Likewise, an English practice study provides no evidence that an AI system is linguistically or culturally suitable for Kazakh learners without testing that suitability.

Pedagogical implications

The programme illustrates a schedule for supplementary individual conversation practice, not a validated substitute for teaching. Reviews emphasize the continuing pedagogical role of teachers in AI-supported language learning (Li et al., 2025). Human guidance and relationships also matter in digital mentorship (Seyed Alitabar & Zadhasn, 2023). A useful comparison would hold practice time and task content constant across AI, peer, and teacher-supported conditions, then assess performance with an independent human interlocutor and after a delay.

Classroom participation and sense of community should remain distinct from success within an individual AI exchange (Yang & Seyed Alitabar, 2024). Future evaluation could examine whether additional practice supports participation outside the application, while documenting adverse or unhelpful interactions and unequal access. Such outcomes would connect the intervention to its educational setting without assuming that more app engagement implies broader psychological benefit.

Limitations and research priorities

The central limitations are the unresolved distinction between raw and displayed scores, the undocumented overall-score calculation, and incomplete inferential reporting. The linear reconciliation is mathematically compatible with the aggregate values but requires confirmation. The rubric lacks reported content review, A2-specific calibration, construct evidence, and external validation. ICC point estimates supply only limited information about rating reproducibility and do not remedy these measurement limitations.

Convenience recruitment from one institution, a narrow age and proficiency range, unspecified collection dates and language backgrounds, and uncertain task equivalence limit interpretation and generalizability. There was no comparison group, follow-up, or verified human-conversation transfer outcome. Concurrent coursework, practice outside the programme, adherence to the scheduled dose, and rater expectancy were not adequately controlled. The qualitative component lacks the documentary basis needed for independent assessment.

The next steps are to audit the original rating files and calculation syntax, verify the scale and paired analyses, and document the instrument's intended interpretation and calibration. Subsequent controlled studies should use prespecified outcomes, independently rated equivalent tasks, a practice-matched comparator, delayed assessment, and transparent dialogue coding. Psychological outcomes, cultural interpretation, and Kazakh-language transfer should be added only when directly assessed with appropriate evidence for their measures. Ethics approval or exemption and platform-specific data arrangements also require documentation.

The reported means are higher after six weeks of Loora-based English conversation practice, but scoring inconsistencies and incomplete statistical output prevent firm conclusions about the magnitude and statistical reliability of change. An undocumented linear normalization could explain the aggregate arithmetic, yet it cannot be confirmed without the original ratings. The uncontrolled design cannot establish intervention effectiveness, and psychological, cultural, and Kazakh-language outcomes were not measured. The study provides provisional descriptive observations that require measurement verification and controlled replication.

Ijbmc.org
E-ISSN: 2345-5802

Ethics and Data Protection

Participants were reported to have received information about the purpose and procedures, provided informed consent, participated voluntarily, and retained the right to withdraw without adverse consequences. No approving research ethics committee, approval number or date, or documented exemption was identified. Thus, consent is reported, while institutional ethical oversight cannot be confirmed from the study documentation. The arrangements separating voluntary research participation from course assessment were not specified.

The report states that personally identifying information was omitted from the research dataset. This does not establish that voice recordings, transcripts, or commercial-platform accounts were anonymous. Storage locations, access permissions, encryption, retention and deletion periods, transfer between the platform and researchers, and safeguards against identification from speech were not described. Nor was it documented whether participants were informed about third-party processing, secondary uses, or any use of their interactions for model development. Current vendor materials cannot retrospectively establish the settings and terms that applied during this study. Human-centred AI research requires explicit attention to these data-governance arrangements (Miao & Holmes, 2023).

Funding and Acknowledgements

The authors reported financial support from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant IRN AP26100757, for a project concerning IT tools and programmes for level-based Kazakh-language teaching and development of learners' soft skills using generative artificial intelligence. The project's target language does not change the English-language outcomes actually assessed in this study. The funder's role in design, analysis, interpretation, and publication decisions was not specified.

References

Aditya, M. Y., Nugroho, H. A., & Mustofa, A. (2026). EFL speaking self-efficacy among English education students in Bangkalan Madura. *International Journal of Body, Mind and Culture*, *13*(11), Article e2026–1379. https://ijbmc.org/index.php/ijbmc/article/view/1379

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. https://rm.coe.int/16809ea0d4

Culpeper, J., Mackey, A., & Taguchi, N. (2018). *Second language pragmatics: From theory to research*. Routledge. https://doi.org/10.4324/9781315692388

Dizon, G., & Tang, D. (2020). Intelligent personal assistants for autonomous second language learning: An investigation of Alexa. *The JALT CALL Journal*, *16*(2), 107–120. https://doi.org/10.29140/jaltcall.v16n2.273

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, *10*(3), 8–14. https://doi.org/10.64152/10125/44068

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. *Language Learning & Technology*, *27*(2), 6–27. https://doi.org/10.64152/10125/73501

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Language Learning & Technology*, *28*(2), 5–30. https://doi.org/10.64152/10125/73570

Hashemi, M., Ebrahimi, F., & Hashemi, F. (2026). English language teachers' psychological responses to AI integration: A grounded theory approach. *International Journal of Body, Mind and Culture*, *13*(1), 46–54. https://doi.org/10.61838/ijbmc.v13i1.1149

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Li, Y., Zhou, X., & Chiu, T. K. F. (2025). Systematics review on artificial intelligence chatbots and ChatGPT for language learning and research from self-determination theory (SDT): What are the roles of teachers? *Interactive Learning Environments*, *33*(3), 1850–1864. https://doi.org/10.1080/10494820.2024.2400090

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching*, *51*(3), 285–329. https://doi.org/10.1017/S0261444818000125

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. https://doi.org/10.1016/B978-012589042-7/50015-3

Loora. (n.d.). *Loora*. https://www.loora.com/

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*(1), 37–66. https://doi.org/10.1017/S0272263197001034

Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*(4), 557–587. https://doi.org/10.1017/S0272263199004027

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583–616. https://doi.org/10.1080/03054985.2012.731208

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Rahman, G., Mudhsh, B. A., Almutairi, M., & Kouki, M. (2025). Optimizing ESL learners' speech act performance: The role of AI-powered chatbots in pragmatic competence development. *Theory and Practice in Language Studies*, *15*(10), 3187–3198. https://doi.org/10.17507/tpls.1510.05

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge University Press. https://doi.org/10.1017/CBO9780511791208

Seyed Alitabar, S. H., & Zadhasn, Z. (2023). Digital mentorship: Mentor–mentee relationships in virtual spaces. *AI and Tech in Behavioral and Social Sciences*, *1*(2), 12–18. https://doi.org/10.61838/kman.aitech.1.2.3

Sim, S., Kim, Y., & Ku, K. (2025). Exploring generative AI as a roleplay interlocutor in L2 task-based pragmatics learning: Comparing ChatGPT-learner and technology-mediated peer interactions. *TESOL Quarterly*, *59*(S1), S86–S116. https://doi.org/10.1002/tesq.70010

Sok, S., & Shin, H. W. (2025). Do interactions with ChatGPT influence L2 learners' oral speaking ability, summarization ability, and perceptions of generative AI tasks? *TESOL Quarterly*, *59*(S1), S19–S51. https://doi.org/10.1002/tesq.70001

Su, Y., & Chen, X. (2026). Second language interactional competence in roleplay: Comparing generative artificial intelligence and human interlocutors. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.70040

Tai, T.-Y., & Chen, H. H.-J. (2024). Improving elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions. *Computers & Education*, *220*, Article 105112. https://doi.org/10.1016/j.compedu.2024.105112

Timpe-Laughlin, V., Divekar, R., Sydorenko, T., Dombi, J., & Oh, S. (2025). Intent-based versus GPT-based conversational agents: Benefits and challenges for practicing and assessing oral interaction. *TESOL Quarterly*, *59*(S1), S117–S149. https://doi.org/10.1002/tesq.3409

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, *31*(2), 57–71. https://doi.org/10.1017/S0261444800012970

Yang, J., & Seyed Alitabar, S. H. (2024). The effects of school size on student participation and sense of community. *Iranian Journal of Educational Sociology*, *7*(1), 205–211. https://doi.org/10.61838/kman.ijes.7.1.20

Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. *System*, *121*, Article 103259. https://doi.org/10.1016/j.system.2024.103259