ИЗДАТЕЛЬСТВО грамота

Педагогика. Вопросы теории и практики
Pedagogy. Theory & Practice

ISSN 2686-8725 (online)
ISSN 2500-0039 (print)

2026. Том 11. Выпуск 4 | 2026. Volume 11. Issue 4
Материалы журнала доступны на сайте (articles and issues available at): pedagogy-journal.ru

RU

Практика английского языка в виртуальной реальности: анализ результатов и ошибок студентов инженерных специальностей

Корзин А. С., Алексеева Н. А., Шалеева Е. Ф., Дмитриченкова С. В., Круглова Л. В.

Аннотация. Цель настоящего исследования состоит в выявлении факторов, затрудняющих прохождение студентами инженерных специальностей с уровнем владения английским языком A2-B1 учебного диалога на иностранном языке в виртуальной реальности (VR), на основе анализа структуры ошибок, выявленной по статистике выполнения сценария. В статье рассмотрены результаты прохождения сценария “Basic information about the company” в двух режимах (демонстрация и экзамен) и проведено сопоставление распределений пользователей с использованием описательной статистики и графического анализа (гистограммы, оценка плотности распределения (KDE-оценка), диаграмма размаха (box-plot)). В ходе анализа ошибок были выделены и рассмотрены два основных типа: ошибки, связанные с выбором обучающимися неправильного варианта ответа, и ошибки распознавания речи (система не принимает или не распознает корректно произнесенный ответ). Эти ошибки анализировались на каждом этапе диалога. Научная новизна исследования состоит в том, что в нем впервые на материале реальных результатов прохождения VR-диалога студентами инженерных специальностей Российского университета дружбы народов имени Патриса Лумумбы в рамках обучения английскому языку предложен и применен индекс доминирования ошибок распознавания (RDI), позволяющий оценить вклад технического фактора (системы автоматического распознавания речи, ASR) в общую структуру ошибок и отделить его от ошибок, связанных с выбором обучающимися неправильного варианта ответа. В результате установлено, что общая сложность заданий VR-диалогов в режимах «демонстрация» и «экзамен» является сопоставимой (медианные значения около 65-67%), однако результаты по прохождению режима «экзамен» сосредоточены в диапазоне 60-70%, тогда как в «демонстрации» наблюдаются более высокая вариативность результатов и наличие низких показателей успеваемости. Было продемонстрировано, что главной проблемой при прохождении обучения являются ошибки распознавания речи. Они встречаются чаще, чем ошибки, связанные с выбором обучающимися неправильного варианта ответа, – в среднем 41,8% по сравнению с 30,2%. Использование показателя RDI позволяет заключить, что ошибки в основном связаны с распознаванием речи и составляют около 76%. На уровне отдельных реплик максимальные ошибки в распознавании речи встречаются в формулах благодарности, завершающих диалог, и вопросах о специализации компании. Наибольшая доля неверных выборов связана с этапами диалога, где необходимо строго соблюдать речевой этикет и точные формулировки. Результаты показывают, что пользователи испытывают больше трудностей с VR-диалогами не из-за недостатка знаний, а из-за проблем с распознаванием речи и особенностей оформления ответов. Это указывает на необходимость улучшения модуля распознавания речи и дизайна сценариев для оценки коммуникативных навыков.

EN

What hinders practicing English in virtual reality: a data-driven analysis of performance and errors among engineering students

A. S. Korzin, N. A. Alekseeva, E. F. Shaleeva, S. V. Dmitrichenkova, L. V. Kruglova

Abstract. The study aims to identify the factors that hinder engineering students with intermediate English proficiency (A2-B1) from completing a learning dialogue in English as a foreign language within a virtual reality (VR) environment, using an analysis of error patterns based on scenario performance statistics. This study examines user results for the scenario “Basic information about the company” in two modes (demo testing and final exam) and compares performance distributions using descriptive statistics and visual analytics (histograms, kernel density estimation, and box plots). Within the error analysis, two key error types

Научная статья (original research article) | https://doi.org/10.30853/ped20260082

were distinguished and quantitatively described across individual dialogue steps: response selection errors (choosing an incorrect reply option) and speech recognition errors (failure of the system to accept/recognise a correctly spoken answer). The scientific novelty of the study lies in the fact that, using real VR dialogue interaction logs obtained from engineering students at RUDN during an English language training course, it proposes and applies a Recognition Dominance Index (RDI) to estimate the contribution of the technical factor (automatic speech recognition, ASR) to overall failure and to separate it from learning-related difficulties. The results show that the overall task difficulty remains comparable between the demo and exam (median performance around 65-67%); however, exam scores are more concentrated in the 60-75% range, while demo outcomes display higher variability and include low-score outliers. Speech recognition was demonstrated to be the main bottleneck of the scenario: on average, recognition errors occurred more frequently than selection errors (approximately 41.8% versus 30.2%), and the RDI indicated a predominantly recognition-driven nature of failure (approximately 76% on average). At the level of individual utterances, the highest non-recognition rates were observed in the closing expressions of gratitude and in questions about the company's specialisation, whereas the highest incorrect-selection rates were associated with steps that required strict adherence to academic etiquette and precise wording. The findings suggest that successful completion of the VR dialogue is hindered less by users' lack of content knowledge and more by speech recognition limitations and response design features, which highlights the need to improve the ASR component and scenario annotation when assessing communicative skills.

Introduction

The relevance of this study is driven by current trends in immersive and digital education, where increasing attention is paid to how learners acquire communicative skills in technology-mediated environments and how these skills can be assessed reliably. Virtual reality (VR) dialogue training is widely viewed as a promising format for situational language practice because it enables role-based interaction, contextualised prompts, and repeated rehearsal of speech behaviour in near-authentic settings. At the same time, the effectiveness of VR dialogue practice depends not only on learners' linguistic competence but also on the stability of speech interaction, especially automatic speech recognition (ASR), which mediates whether an utterance is accepted by the system. The issue arises when technical constraints affect task completion, as the training environment may produce errors unrelated to pedagogical objectives, thereby distorting performance assessment and diminishing learner engagement. This makes it necessary to analyse VR dialogue outcomes in a way that separates learning-related mistakes from system-driven failures.

This study aimed to identify the factors that hinder engineering students' completion of a structured VR dialogue scenario in English and to describe the error structure of scenario performance using empirical interaction data. To achieve this purpose, the study addresses the following objectives:

1) to identify the pedagogical and technological conditions that determine the effectiveness and validity of VR-based English dialogue training and assessment, with particular attention to ASR-related constraints and the distinction between learner errors and system-related failures;

2) to analyse and compare performance distributions in two assessment conditions (demo testing and the final exam) for engineering students with intermediate English proficiency;

3) to classify and quantify the main types of errors observed during scenario completion within the context of English language learning in VR;

4) to determine the relative contribution of response selection errors and speech recognition errors across individual dialogue steps;

5) to propose and apply an indicator that estimates the dominance of recognition-related failures within the overall error structure;

6) to identify scenario stages and utterances associated with the highest rates of non-recognition and incorrect selection, and outline implications for scenario and assessment design.

The empirical material of the study consists of aggregated statistics and log-derived metrics collected during 172 users' completion of the VR scenario "Basic information about the company", which simulates an academic interaction with a professor in preparation for an industrial internship. The dataset covers two modes of scenario use (demo and exam) and includes step-level counts of attempts and errors, enabling both overall performance comparisons and a fine-grained analysis of problematic dialogue turns.

The theoretical foundations of this research integrate publications on VR-based, scenario-driven language learning and technology-mediated speaking assessment, where immersive tasks are used to practice situated communication and can support language development (Chen, Yuan, 2023; Palmas, Cichor, Plecher et al., 2019; Sadigzade, 2025). Distributional analysis follows the established principles of density estimation and exploratory data analysis (Silverman, 1986; Scott, 2015; Tukey, 1977), while the interpretation of interaction logs and aggregated indicators is grounded in learning analytics and educational data mining (Baker, Inventado, 2014; De Araujo, Papadopoulos, Mckenney et al., 2024) and standard statistical practice (Field, 2018). The operationalisation of error categories is informed by error analysis in language learning (Broussard, 1999) and computer-assisted language learning (CALL) research that distinguishes learner errors from system-driven constraints (Heift, Schulze, 2007). Finally, the study draws on work on ASR in CALL and automated speaking assessment, which shows that recognition performance, especially for non-native speech, can substantially influence observed outcomes and assessment

fairness, motivating the explicit modelling of recognition failures alongside content errors (De Vries, Cucchiarini, Bodnar et al., 2014; Knill, Gales, Kyriakopoulos et al., 2018; Thi-Nhu Ngo, Hao-Jan Chen, Kuo-Wei Lai, 2023; Kang, Jeon, Lee, 2024; Jurafsky, Martin, 2020).

The methodological framework combined descriptive statistics and visual analytics to compare user outcomes across conditions using histograms with kernel density estimation (KDE) and box plots. KDE provides a smooth, non-parametric estimate of score distributions with bandwidth selected via Scott's rule, whereas box plots follow Tukey's five-number summary and identify outliers using the 1.5×IQR criterion. In parallel, the study conducted a dialogue-step error analysis using two operationally defined error rates: (1) the recognition error rate (UR), which is the proportion of spoken responses not accepted by the speech recogniser, and (2) the content error rate (WR), which is the proportion of responses that were selected incorrectly. The study also quantified the contribution of technical factors via the Recognition Dominance Index (RDI), which is defined as the proportion of errors that are recognition errors among all errors recorded at a step. Module-level estimates are additionally reported using attempt-weighted averages with dispersion described by the standard deviation and coefficient of variation.

The practical significance of this research lies in its potential to inform the design of VR dialogue scenarios, assessment criteria, and feedback mechanisms in language training. By distinguishing pedagogical difficulties from ASR-related failures, the proposed approach can support a more valid performance interpretation, improve the learner experience, and guide targeted improvements in speech recognition settings, prompt formulation, and response-option design in immersive educational applications.

Discussion and results

Literature review

Virtual reality is used for dialogue-based language training because it provides immersive and interactive environments that simulate real-life conversational contexts, allowing learners to practice language skills in a controlled yet realistic setting. The immersive nature of VR enables users to engage deeply with virtual interlocutors, fostering authentic dialogue experiences that are often difficult to replicate in traditional classroom settings (Sadigzade, 2025, p. 16).

VR virtual human training programs simulate nuanced interpersonal situations, enabling learners to practice dialogue and communication skills without real-world risks or supervision (Palmas, Cichor, Plecher et al., 2019, p. 463). Moreover, VR environments offer personalised and adaptive learning opportunities, allowing for scaffolding tailored to the individual learner's level and needs, which supports learner autonomy and self-directed improvement (Sadigzade, 2025, p. 19).

The most consistently reported learning benefits of dialogue-based language training using VR include enhanced cognitive engagement and motivation, improved fluency and vocabulary retention, and decreased anxiety during language use. Immersive, authentic contexts stimulate active participation and provide immediate multimodal feedback, which strengthens memory and comprehension through experiential and repetition-enabled learning (Chen, Yuan, 2023, p. 397; Sadigzade, 2025, p. 18). Learners also benefit from the increased frequency and quality of interactions in VR, which are crucial for language development, allowing for practice in pronunciation, syntax, and pragmatic aspects of communication in a naturalistic manner (Sadigzade, 2025, p. 19).

Validity issues in assessing speaking/dialogue skills in VR stem from challenges such as ensuring that virtual scenarios authentically replicate real-life communication contexts and that the measured constructs actually represent speaking or dialogue competence (Lee, 2025, p. 58-59).

Automatic speech recognition systems show notably decreased reliability when processing non-native speech in educational settings, primarily because of accented and non-native pronunciation patterns that diverge from the native speech on which these systems are typically trained. The main factors contributing to non-recognition or misrecognition include specific pronunciation deviations influenced by the speaker's mother tongue and the limited availability of non-native speech data for training ASR models (Radzikowski, Wang, Yoshie et al., 2021, p. 1).

The accuracy of ASR for non-native speakers can be enhanced using methods such as acoustic model adaptation, bilingual models, speaker adaptation, and systems trained on diverse non-native speech data. These methods aim to tailor the acoustic models to better fit non-native speech characteristics and improve performance despite the scarcity of non-native data (Wang, Schultz, Waibel, 2003, p. 540). Advanced approaches that combine semi-supervised learning and transfer learning have been successfully applied in language tutoring systems for non-native learners, enabling improved spontaneous speech recognition accuracy and reliable proficiency evaluation (Kang, Jeon, Lee, 2024, p. 56).

The factors that most often cause non-recognition or misrecognition involve linguistic variability inherent to non-native speech, such as accented pronunciation, phonetic deviations, and disfluencies. Furthermore, listeners themselves tend to process non-native speech less thoroughly, expecting lower reliability and therefore extracting less detailed linguistic information, which parallels the challenges recognised by ASR systems (Lev-Ari, 2015, p. 10-11; Lev-Ari, Keysar, 2012, p. 536-537).

Automatic speech recognition errors can significantly influence learner performance, motivation, and fairness in automated assessment systems. Regarding learner performance, ASR systems used in free-speaking language assessments and CALL must contend with transcription inaccuracies owing to the spontaneous and unscripted nature of learner responses. ASR errors affect the system's ability to accurately detect errors related to pronunciation,

grammar, and relevance to prompts, potentially leading to confusing or misleading feedback for learners. The inability to provide precise corrective feedback due to ASR errors can impair the effectiveness of language learning and progression (De Vries, Cucchiarini, Bodnar et al., 2014, p. 2; Knill, Gales, Kyriakopoulos et al., 2018, p. 1643-1644).

From a motivational perspective, learners who receive immediate corrective feedback based on ASR are more likely to positively evaluate the learning system and remain engaged, even though actual proficiency gains may not differ significantly from those of learners who do not receive feedback. The presence of ASR-based feedback encourages learner appreciation and possibly motivation, but this benefit hinges on the accuracy and clarity of such feedback being maintained to avoid confusion due to errors (De Vries, Cucchiarini, Bodnar et al., 2014, p. 18).

Fairness in automated assessments is notably challenged by ASR errors, which disproportionately affect certain user groups, particularly speakers with dialectal variations or speech disorders. ASR errors also affect the fairness and validity of automated proficiency evaluation. Non-native and speech-disordered speakers may experience higher ASR error rates, leading to inaccurate assessments. Personalised ASR models tailored to speech specifics, including conversational speech characteristics, have improved recognition accuracy but highlight the current limitations of ASR in uniformly and fairly evaluating diverse learner populations. The performance gap caused by linguistic complexity and speech peculiarities can undermine the fairness of automated assessments if unmitigated (Kang, Jeon, Lee, 2024, p. 49; Tobin, Nelson, Macdonald et al., 2024, p. 4184).

From a learning outcome standpoint, ASR has shown medium to large effects in improving pronunciation performance, particularly when combined with explicit corrective feedback, extended treatment duration, and interactive peer practice. However, short-term or indirect feedback and a lack of tailored adaptation reduce effectiveness. This sensitivity to error and individual learner contexts underscores the importance of robust ASR systems and intelligent mitigation strategies to sustain motivation and fairness (Thi-Nhu Ngo, Hao-Jan Chen, Kuo-Wei Lai, 2023, p. 5, 16-17).

In dialogue training systems, distinguishing learner mistakes (such as incorrect choices) from system failures (such as ASR rejection) relies on well-defined error taxonomies and learning-analytics approaches that analyse interaction patterns and system performance.

Errors can be categorised as learner errors, such as interlingual errors influenced by the first language, intralingual errors arising from the process of learning the target language, or communication-strategy-based errors (James, 2013, p. 173-189), or system-related failures. The latter include ASR-driven non-understandings (e.g., no-input or rejection cases) and downstream interpretation problems in spoken dialogue systems (Bohus, Rudnicky, 2005, p. 142; Bohus, 2007, p. 37-41). The classification involves differentiating errors from mistakes and lapses, focusing on recurring learner performance issues rather than one-off mishaps, and offering the granularity needed to separate learner faults from system faults.

In dialogue systems, especially task-oriented and conversational recommender systems, joint learning models combine natural language understanding (NLU) with dialogue management to mitigate noisy outputs and error propagation in the pipelines. These end-to-end approaches can more effectively identify whether errors stem from learner input (e.g., wrong choice) or system processing, such as ASR failures, by backpropagating error signals from system actions to refine the NLU module and dialogue state tracking components (Li, Chang, Wu, 2020, p. 14; Yang, Chen, Hakkani-Tur et al., 2017, p. 5693). This joint learning helps to isolate and classify errors arising from different sources.

Learning analytics methods leverage data-driven techniques to analyse learner interactions and system logs to detect differences between learner-induced errors and system failures. For example, in intelligent tutoring systems with conversational dialogue, the system logs multi-turn interactions and learner responses, allowing for the identification of patterns indicating learner misunderstanding versus system misunderstandings (Graesser, Jordan, Vanlehn et al., 2001, p. 50).

Similarly, sequential pattern mining and dialogue analytics can reveal productive or unproductive learner behaviours and system responses, highlighting where system errors, such as ASR rejections, impede communication flow (De Araujo, Papadopoulos, Mckenney et al., 2024, p. 2701-2702).

In summary, the distinction between learner mistakes and system failures in dialogue training is supported by detailed error taxonomies that classify language learning errors versus technical faults, combined with learning-analytics methods, including joint end-to-end learning models, multi-turn conversational logs, sequential pattern mining, and reinforcement-learning-based dialogue management. Together, these approaches enable systems to identify and appropriately respond to the source of errors, thereby improving feedback accuracy and the learner experience.

Materials and data processing

The study was conducted at the Academy of Engineering of RUDN University between September and December 2025 and was based on interaction logs collected from the VR Supersonic English language learning system during the completion of Module 1.1, “Basic information about the company”. Two types of learner interaction with the VR system within the same module were analysed: (i) demo mode, which involved practice attempts (Figure 1), and (ii) exam mode, which involved final assessment attempts (Figure 2). The sample comprised engineering students (189 participants) with intermediate English proficiency (A2-B1). The system records the attempt outcomes and step-level interaction counts, including incorrect selections and responses that are not recognised by the speech recognition component.

The analysis uses two granularities of data: (1) attempt-level records summarising the overall completion outcome for a simulation run and (2) dialogue-step aggregates describing student interactions within each dialogue turn. All the variables used in the analysis are summarised in Table 1.

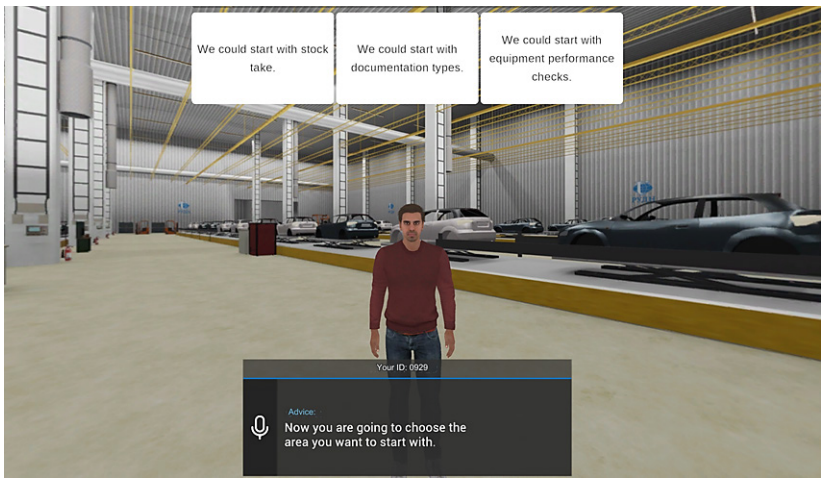

Figure 1. *Dialogue demonstration mode*

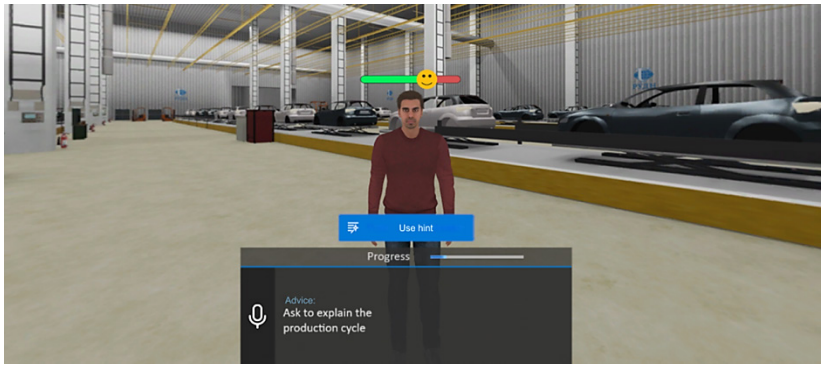

Figure 2. *Exam mode (with hidden answer options)*

Table 1. *Key variables of the dataset under study*

Level	Variable	Description	Data type
Attempt	Status	Attempt completion status	Categorical
Attempt	Overall result, %	Final completion percentage	Numerical (0-100%)
Attempt	User	Student ID	Categorical
Attempt	Simulation type	Test mode	Categorical
Dialogue-step	Module	Module identifier	Categorical (1.1)
Dialogue-step	Dialogue	Dialogue number in the module	Integer (1-12)
Dialogue-step	Wrong_answer	Number of content errors made by students	Absolute frequency
Dialogue-step	Attempts_w	Total number of attempts to complete the dialogue	Absolute frequency
Dialogue-step	Unrecognised	Number of responses not recognised by the system	Absolute frequency
Dialogue-step	Attempts_un	Total number of responses given by the user during the dialogue	Absolute frequency

To improve the reliability of the results, preprocessing was performed prior to the analysis. At the attempt level, incomplete attempts (Status = “Not completed”) were excluded from comparisons between demo and exam performances. The percentage values were validated to ensure that they fell within the 0-100 range.

For the dialogue-step dataset, the following checks were performed: (i) dialogues with missing values in key variables were removed (two dialogues); (ii) Dialogue 10 was excluded because of zero values in Attempts_w, which would make percentage metrics undefined; and (iii) all count variables were validated to ensure non-negative integer values and logical consistency (e.g., Wrong_answer ⩽ Attempts_w). After preprocessing, the analytical sample comprised nine dialogues (Dialogues 1-9), accounting for 94.3% of all recorded interactions in the first module.

To describe user performance and the error structure of the VR module scenario, we applied a set of visualisation techniques and quantitative metrics that are commonly used in educational data analysis and speech-enabled systems. This section defines the measures used in the study and summarises the procedures for computing dialogue- and module-level indicators.

Metrics

Histograms with superimposed probability density curves estimated using KDE were employed to visualise the distribution of the test results. This method allows for a smooth estimate of the distribution density without assuming

a parametric form (Silverman, 1986). The width of the smoothing window was determined using Scott's rule (Scott, 2015), which ensures a balance between the bias and dispersion of the estimate.

Box plots proposed by Tukey (1977) were used to present a five-number summary of distributions. The plots show:

- minimum value (excluding outliers);
- first quartile (Q1, 25th percentile);
- median (50th percentile);
- third quartile (Q3, 75th percentile);
- maximum value (excluding outliers).

Outliers were determined using the 1.5×IQR rule (Frigge, Hoaglin, Iglewicz, 1989), where IQR is the interquartile range (IQR = Q3-Q1).

The recognition error rate was determined using the following formula:

$$UR_i = \frac{U_i}{A_i} \times 100\%,$$

U_i – number of unrecognised responses in the i-th dialogue,

A_i – number of attempts made in the i-th dialogue.

This metric has been adapted from the speech recognition system evaluation methodology (Jurafsky, Martin, 2020) for text-based educational systems.

Error rate (incorrectly selected responses):

$$WR_i = \frac{W_i}{A_i} \times 100\%,$$

W_i – number of incorrect responses in the i-th dialogue,

A_i – number of attempts made in the i-th dialogue.

This metric corresponds to the standard approaches for evaluating the effectiveness of training in computer systems (VanLehn, 2011).

One more metric used is the Recognition Dominance Index (Heift, Schulze, 2007):

$$RDI_i = \frac{U_i}{W_i + U_i} \times 100\%,$$

U_i – number of unrecognised responses,

W_i – number of errors in which the wrong answer was selected.

For example, if RDI > 50%, recognition errors prevail. Accordingly, if RDI < 50%, content errors prevail.

To evaluate the module as a whole, the following were calculated:

- weighted average value: $M = \sum \frac{w_i \times x_i}{w_i}$;
- standard deviation: $SD = \sqrt{\sum}$;
- coefficient of variation: $CV = \frac{S}{M} \times 1000\%$.

All statistical calculations were performed in accordance with the standard methods described by Field (2018) for psychological and educational research. Weighted indicators were used to correct for differences in the volume of the dialogue data (Baker, Inventado, 2014).

Key findings

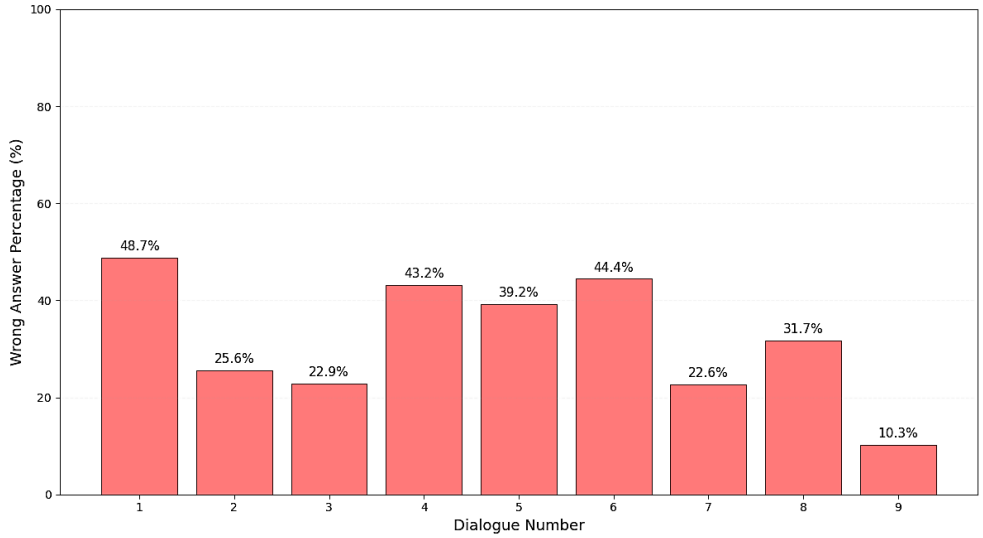

Figure 3. *Wrong answer ratio by dialogue*

Figure 3 summarises the distribution of incorrect response selections across the module dialogues. Overall, incorrect selections constituted 30.2% of the recorded attempts. The highest values were observed in Dialogues 1, 4, and 6, indicating that these turns may be comparatively more demanding for learners and/or more sensitive to the structure of the provided response options.

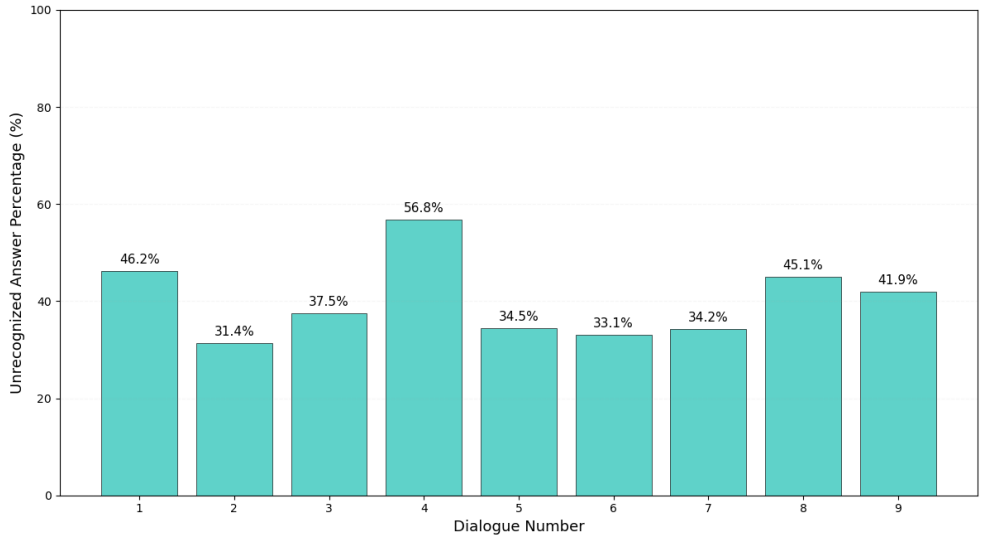

Figure 4. *Unrecognised answer ratio by dialogue*

Figure 4 shows the distribution of unrecognised responses. Across the dataset, 40.08% of the spoken answers were not accepted by the system. The most pronounced case was Dialogue 4, where the unrecognised response rate reached 56.8%, meaning that more than half of the attempts were rejected by speech recognition at this stage.

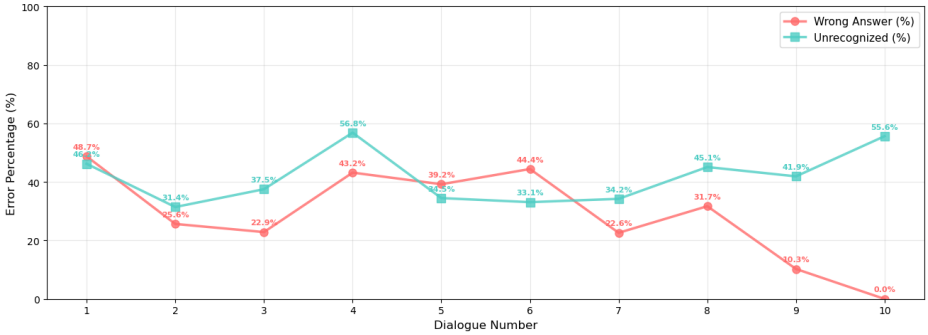

Figure 5. *Dynamics of errors by dialogue*

A direct comparison of the two error categories is shown in Figure 5. Recognition-related failures (40.08%) exceeded incorrect selection errors (32.07%) in six of the nine dialogues, suggesting that the predominance of non-recognition is not confined to a single prompt but represents a recurrent pattern across the scenario.

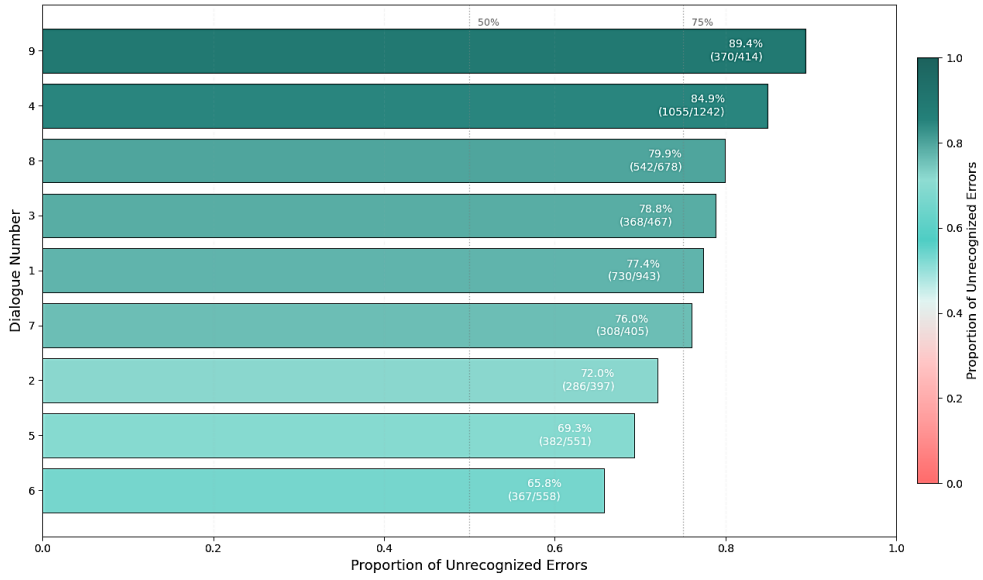

Figure 6. *Heatmap of unrecognised errors proportion by dialogue*

The structural composition of the errors is further visualised in Figure 6 using a heatmap. The scale ranges from red (lower proportion of unrecognised errors; relative predominance of incorrect selections) to dark turquoise (higher proportion of unrecognised errors; predominance of non-recognition). This representation identifies dialogue steps where learner choice errors are comparatively more frequent and steps where the system's recognition performance is the primary constraint.

To quantify the relative contribution of recognition failures, the RDI was calculated as the ratio of unrecognised responses to the total number of errors. The results indicate a strong predominance of recognition-driven failure: on average, unrecognised responses accounted for 76.3% of all errors (SD = 8.8%), implying that most unsuccessful outcomes are attributable to ASR-related rejection rather than to inappropriate response choice.

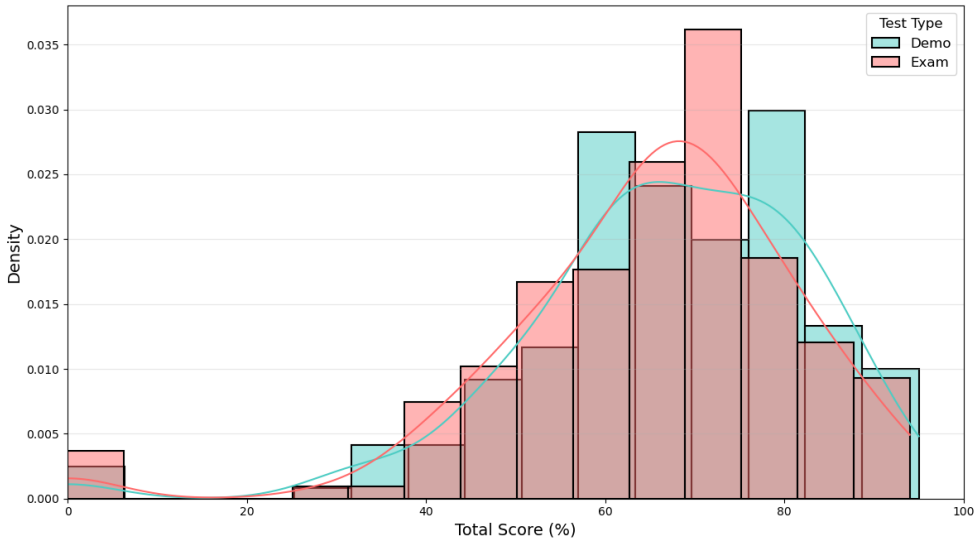

Figure 7. *Distribution comparison: demo vs exam results*

The graph (Figure 7) shows a comparison of the distributions of the demonstration and exam results in the form of combined histograms with superimposed probability density estimation (KDE) curves. The exam distribution is more concentrated in the 60-75% interval, whereas the demo results show greater dispersion across the full score range. The KDE peak was higher and narrower for the exam condition, consistent with reduced variability.

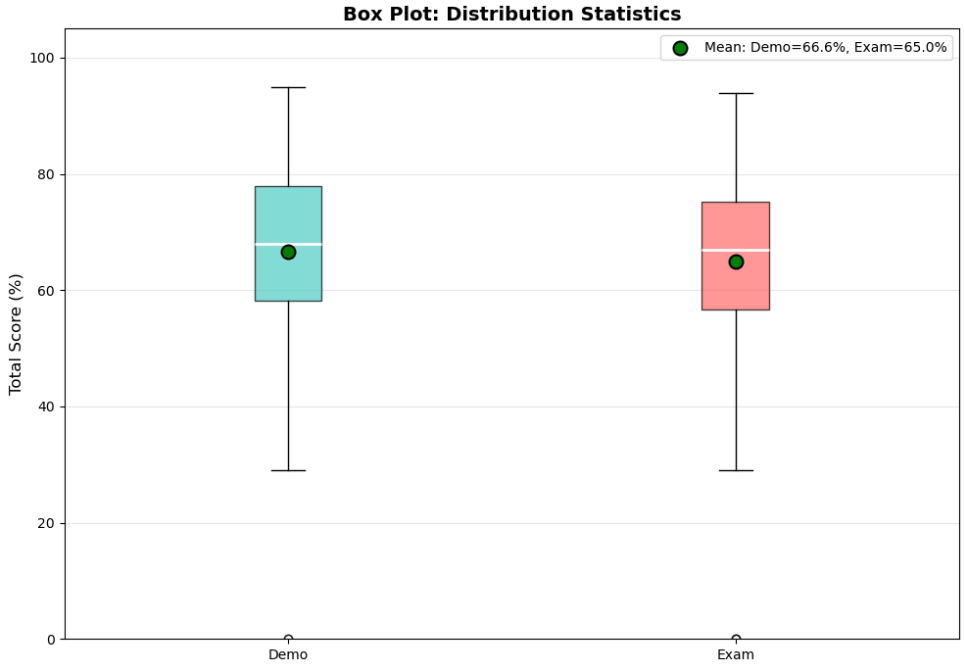

Figure 8. *Box plot: distribution statistics*

This difference is supported by the box plot comparison (Figure 8). The median score in the demo was 66.6%, and the median exam score was 65.0%, yielding a difference of 1.6 percentage points, which suggests a broadly comparable overall task difficulty. However, the exam condition showed a narrower interquartile range (IQR), indicating a tighter clustering of results. Demo scores exhibited a wider IQR and a larger overall range, alongside a higher number of outliers, particularly on the lower end (below 30%).

To assess the impact of the simulator's limited synonym base on performance in the VR simulation, an analysis was conducted of the distribution of the metric "First-attempt success rate, %" across students' final performance categories. The primary visualization tool was the box plot (Figure 9), which enables a clear comparison of central tendencies and dispersion across groups (Tukey, 1977).

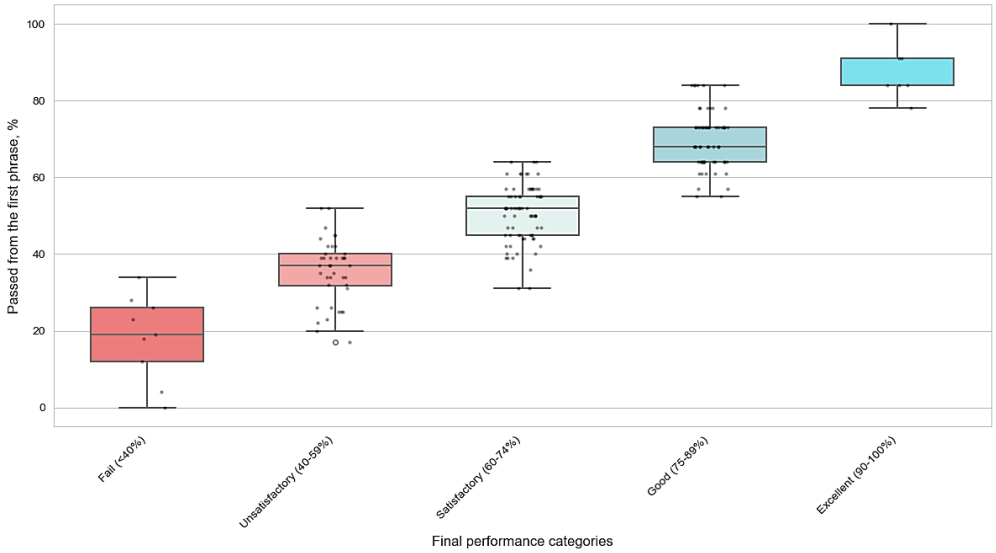

Figure 9. *Distribution of the metric "First-attempt success rate, %" by final performance categories*

Across the entire sample, the mean value of the "First-attempt success rate" metric was 53%, while a standard deviation of 19% indicates substantial variability in results. The median (55%) is close to the mean, suggesting the absence of strong skewness in the overall distribution. However, the interquartile range (39% to 73%) confirms considerable heterogeneity in this indicator.

Figure 9 presents box plots for five final performance categories, ranging from "Fail" to "Excellent". As expected, the median value of the first-phrase metric increases with higher final performance categories – from 18% in the "Fail" group to 84% in the "Excellent" group. However, the principal finding of the study lies not in this upward trend, but in the degree of dispersion within categories, particularly among the higher-performing groups.

The "Good" category (75-89%) is of particular interest. Although its median first-phrase score is 68%, the lower quartile falls to 57%, and the lower whisker extends to 40% and below. This indicates that one quarter of students who ultimately achieved a solid "Good" result began the simulation with a first-phrase score below 57%, with some as low as 40%. Given their eventual success, these low initial scores cannot be attributed to insufficient subject knowledge.

A similar, though less pronounced, pattern is observed in the "Excellent" category (90-100%). Despite a median first-phrase score of 84%, some students recorded initial scores in the 60-70% range – substantially below their final outcomes. Considering that these students ultimately achieved near-perfect performance, their initial difficulties are unlikely to reflect knowledge gaps; rather, they appear to be associated with adaptation to the system's lexical expectations.

The "Satisfactory" category (60-74%) exhibits the greatest dispersion, with first-phrase scores ranging from 20% to 80%. The wide interquartile range and extended whiskers point to marked heterogeneity within this group. Some students began with relatively high scores but did not maintain their performance, while others initially struggled to adapt to the simulation format yet ultimately achieved satisfactory results.

In the lower categories ("Unsatisfactory" and "Fail"), median first-phrase scores are predictably low (38% and 18%, respectively). Nevertheless, even in these groups, outliers with initial scores approaching 50% are observed, suggesting unrealized potential that was not fully expressed during the simulation.

The mistake profile was derived from the dialogue statistics, which reports the most frequent problems by utterance/phrase and separates two error mechanisms: response-selection (wrong option) errors and ASR non-recognition of spoken responses. Overall, the results show that errors are not evenly distributed across the scenario: instead, a small number of dialogue steps account for a disproportionately large share of failures, indicating clear "bottlenecks" in the interaction flow.

For wrong-answer (selection) errors, the highest rates were observed for Phrase #1 ("Thank you, Professor Saleh… Could I get the necessary documents…") with 394/856 incorrect selections (46.03%), Phrase #6 ("Could I learn more about my role…") with 363/846 (42.91%), and Phrase #4 ("That sounds fascinating! Are there any specific areas they specialize in?") with 352/849 (41.46%). These were followed by Phrase #5 (32.22%) and Phrase #8 (28.55%). The concentration of selection errors in these turns suggests that they may be especially sensitive to pragmatic appropriateness (e.g., etiquette, politeness), the wording of response options, or ambiguity in the prompt-option mapping.

The ASR non-recognition analysis shows even stronger peaks in difficulty for particular utterances. The highest non-recognition rate occurred for Phrase #11 ("Thank you so much… I truly appreciate your guidance") at 69/124 (55.65%), closely followed by Phrase #4 at 1740/3273 (53.16%). Elevated non-recognition was also found for Phrase #1 (44.59%)

and Phrase #8 (42.28%), while Phrase #9 (“Thank you… I’m looking forward…”) remained problematic at 37.59%. Notably, Phrases #1, #4, and #8 appear among the most problematic items for both error types, indicating “double-bottlenecks” where learners not only choose incorrect options but also experience frequent recognition rejection; this pattern suggests that successful completion in these steps is jointly constrained by task design (prompt/option clarity) and ASR robustness.

Table 2. *Top 5 phrases that were labeled as ‘Wrong answers’*

	Phrase number	Phrase	Wrong answers	Number of attempts	Percentage of wrong answers
1	#1	Thank you, Professor Saleh. My name is… Could I get the necessary documents for my training?	394	856	46.03
2	#6	Could I learn more about my role during the internship?	363	846	42.91
3	#4	That sounds fascinating! Are there any specific areas they specialize in?	352	849	41.46
4	#5	What company departments can we choose for the training?	271	841	32.22
5	#8	Is there anything specific I should study beforehand to make the most of my internship?	241	844	28.55

Table 3. *Top 5 phrases that were labeled as ‘Unrecognised answers’*

	Phrase number	Phrase	Unrecognised answers	Number of attempts	Percentage of unrecognised answers
1	#11	Thank you so much, Professor Saleh. I truly appreciate your guidance.	69	124	55.65
2	#4	That sounds fascinating! Are there any specific areas they specialize in?	1740	3273	53.16
3	#1	Thank you, Professor Saleh. My name is… Could I get the necessary documents for my training?	1295	2904	44.59
4	#8	Is there anything specific I should study beforehand to make the most of my internship?	912	2157	42.28
5	#9	Thank you, Professor Saleh. I’m looking forward to this internship!	583	1551	37.59

Interpretation of results

The findings strengthen the conclusion that, although overall performance in demo and exam conditions is broadly comparable, the dominant constraint in this VR dialogue activity is system-side recognition failure rather than learners’ inability to select an appropriate response. At the dialogue level, unrecognised spoken responses average 40.08%, while incorrect response selection averages 32.07%, and recognition-related failures exceed selection errors in six out of nine dialogue steps. This pattern is not limited to isolated turns; the stepwise dynamics and heatmap indicate that non-recognition is a recurrent property of the interaction rather than an occasional anomaly. As a result, system behaviour becomes a central determinant of whether users can progress through the scenario successfully, which has immediate implications for both learning interpretation and assessment design.

A key implication concerns the construct representation and validity of technology-mediated speaking tasks. When a substantial share of unsuccessful attempts stems from ASR non-recognition, observed performance may partially reflect construct-irrelevant variance, that is, variation driven by technical acceptance constraints rather than communicative competence. In practical terms, the same learner may possess adequate pragmatic and linguistic knowledge but still fail at steps where the system rejects a correct or near-correct spoken response. This makes it difficult to interpret low scores as clear evidence of deficits in etiquette, vocabulary, or discourse appropriateness because the pathway from competence to score is mediated by recognition reliability. The results therefore suggest that, without additional safeguards, a single composite score risks blending two constructs: (1) communicative competence targeted by the scenario and (2) compatibility of the learner’s pronunciation/lexical choice with the recogniser’s expectations.

The comparison of demo and exam distributions provides complementary evidence of stability and user behaviour. While median outcomes remained close (66.6% in demo vs. 65.0% in exam; $\Delta = 1.6$ pp), the exam distribution was more tightly concentrated in the 60-75% range, whereas demo outcomes showed greater dispersion and a larger number of low-score outliers (below 30%). This pattern is consistent with familiarisation and standardisation effects: in the exam, participation conditions are likely to be more uniform, and learners may be more attentive and motivated, producing a narrower spread of results. Simultaneously, the persistence of non-recognition across multiple dialogue steps implies that familiarisation alone cannot resolve the principal bottleneck. Even if users understand the scenario and expected pragmatic intent, their scores can still be constrained by recognition rejection at critical points in the dialogue.

The Recognition Dominance Index makes this issue explicit by quantifying the extent to which recognition failures contribute to the overall error structure. With a mean RDI of 76.3% (SD = 8.8%), roughly three quarters of recorded errors were recognition-driven rather than caused by incorrect response selection. This has an important interpretive consequence: many failures in the logs are better understood as system-limited interaction events rather than learning errors. In an assessment setting, treating these failures as equivalent to content mistakes would risk penalising learners for factors outside the intended construct of the assessment. Conversely, in a training setting,

repeated technical rejection can create misleading feedback loops (e.g., learners may infer that their pragmatic choice is wrong when the system is rejecting the audio), which can reduce confidence and increase frustration, even when users are conceptually "right".

Dialogue- and utterance-level results point to specific bottlenecks that should be prioritised for redesign. The most severe non-recognition peak occurs in Dialogue 4, where the unrecognised response rate reaches 56.8%, indicating that more than half of the spoken attempts at this stage are rejected. Phrase-level statistics further show that particular utterances experience pronounced difficulty: for non-recognition, the largest rates are observed for Phrase #11 (final gratitude) at 55.65% (69/124) and Phrase #4 at 53.16% (1740/3273); elevated rejection is also reported for Phrase #1 (44.59%), Phrase #8 (42.28%), and Phrase #9 (37.59%). Importantly, these values suggest that recognition failures concentrate in parts of the dialogue that are linguistically formulaic (e.g., gratitude) as well as in informational questions (e.g., specialisation), implying that both "routine" and "content" turns may require improved acceptance strategies rather than only complex, multi-clause utterances.

Selection errors exhibited a different but partially overlapping pattern. The highest incorrect-selection rates occurred for Phrase #1 (46.03%), Phrase #6 (42.91%), and Phrase #4 (41.46%), followed by Phrase #5 (32.22%) and Phrase #8 (28.55%). These findings suggest that certain steps are difficult not because learners "cannot speak", but because they must map the prompt to the correct pragmatic choice among options, often requiring sensitivity to politeness, register, and appropriateness. Moreover, the overlap between the error lists is particularly informative: Phrases #1, #4, and #8 appear among the most problematic items for both error types, indicating "double-bottlenecks" where users are simultaneously vulnerable to option-selection pitfalls and recognition rejection. This interaction can amplify failure: learners may first select an incorrect option, and then, even after choosing correctly, encounter rejection when speaking the corrected response. Consequently, a single dialogue step can generate multiple failed attempts for different reasons, complicating both the learner's experience and performance interpretation.

The extended dataset also introduced the first-attempt success rate as an informative process metric. Across the sample, first-attempt success averaged 53% (SD = 19%) with a median of 55%, indicating substantial variability in whether users could pass steps on their first try. Importantly, distributional patterns across performance categories suggest that low first-attempt success is not always a marker of low competence: within the "Good" category, one quarter of learners fall below 57% first-attempt success (with some near 40%), and even among "Excellent" learners, first-attempt success may remain in the 60-70% range despite near-ceiling final results. This is consistent with the adaptation-to-the-system effect, in which learners converge over repeated attempts on the system's expected lexical form or pronunciation pattern. In such a setting, final success can reflect persistence and iterative guessing as much as competence, whereas first-attempt metrics may better capture how "naturally" the system accepts competent speech. This is important for formative training because repeated attempts can provide learning opportunities. However, for summative assessment, it raises fairness questions if some learners require more attempts simply because their speech is less compatible with the recogniser.

Taken together, these patterns support several design and assessment implications. First, for steps with high non-recognition, especially Dialogue 4 and the phrases with the largest rejection rates, priority should be given to ASR-robust redesign: expanding acceptable synonyms/paraphrases, widening lexical coverage for expected intents, and introducing tolerant acceptance logic (e.g., partial matching, confidence-aware retries, or structured clarification prompts) to prevent repeated technical failure from obscuring learning progress. Second, for steps with elevated selection errors (for example, Dialogues 1, 4, and 6; Phrases such as #1 and #6), improvements should focus on prompt and option engineering by reducing ambiguity between choices, sharpening etiquette cues, and ensuring that distractor options test the intended pragmatic distinction rather than superficial wording differences. Third, reporting practices should separate system-related errors from content errors (UR/WR and RDI alongside the final score), because the RDI values indicate that many "errors" are technical and should not be treated as learner deficits in feedback or grading.

However, the results should be interpreted in light of the methodological constraints. The analysis relies primarily on aggregated step-level statistics and does not directly examine acoustic inputs, pronunciation features, or ASR confidence outputs; therefore, it cannot definitively attribute non-recognition to specific phonetic segments, accent patterns, or system parameter settings. In addition, while the binary separation between selection errors and recognition failures is analytically useful, it does not capture finer-grained pragmatic or linguistic subtypes (e.g., whether incorrect selection reflects politeness violations, attentional lapses, or misunderstanding of task intent). Future work that adds utterance-level evidence, such as ASR confidence scores and, where permissible, transcripts, would allow a more precise diagnosis of why certain phrases act as bottlenecks and would enable the targeted evaluation of redesign efforts.

Overall, the expanded results indicate that VR dialogue training in its current configuration is hindered primarily by ASR reliability and acceptance coverage, and that recognition-driven failures meaningfully shape score distributions, first-attempt success, and users' experience of task difficulty. Addressing these recognition bottlenecks, particularly in the identified "double-bottleneck" steps, appears essential not only to improve usability and reduce frustration, but also to strengthen the interpretability and fairness of performance-based assessment in speech-enabled VR learning environments.

Conclusion

This study examined user performance and error patterns in a VR-based dialogue scenario designed to practice academic etiquette and internship-related communication. By combining distributional analysis of outcomes with step-level error analytics, this study provides an evidence-based account of where learners struggle and to what extent technical constraints, specifically ASR, shape observed performance.

Importantly, the findings suggest that observed underperformance in VR dialogue tasks cannot be attributed solely to learner competence: a substantial proportion of failures arise from recognition-related constraints that can mask true communicative ability and distort feedback. This distinction is critical when VR scenarios are used in assessment contexts, because recognition-driven errors introduce construct-irrelevant variance and may disadvantage learners whose speech patterns fall outside the recogniser's most robust range. From a design perspective, identifying where recognition errors cluster provides actionable guidance for improving both the learning experience and the fairness of automated evaluation.

The objectives of this study were largely achieved.

1. Literature review: the effectiveness of dialogue practice in VR is determined not only by pedagogical factors, but also by technological ones, particularly the quality of automatic speech recognition, which can significantly influence learner performance, the reliability of feedback, and the validity of assessment.

2. Performance distributions (demo vs. exam): The comparison showed similar overall difficulty (median 66.6% in demo vs. 65.0% in exam) but a more concentrated exam distribution in the 60-75% range and greater dispersion with more low-score outliers in the demo.

3. Error classification and quantification: Two main error types were identified and quantified: incorrect response selection (32.7%) and unrecognised spoken responses (40.08%).

4. Relative contribution across dialogue steps: Recognition-related failures were more prevalent than selection errors in six of the nine dialogue steps, indicating a recurrent dominance of non-recognition across the scenario.

5. Indicator of recognition dominance: The Recognition Dominance Index was proposed and applied, yielding a mean of 76.3% (SD = 8.8%) and demonstrating that most errors were recognition-driven.

6. Problematic stages and implications: The most error-prone stages were Dialogues 1, 4, and 6 for incorrect selection and Dialogue 4 for non-recognition (56.8%), implying the need to prioritise ASR-robust redesign of critical prompts and refinement of response options to improve the assessment validity and user experience.

Taken together, the results indicate that the overall difficulty of the scenario is broadly stable across assessment conditions, while the predominant source of failure is ASR non-recognition rather than inappropriate response choice. In practical terms, this means that improving learning outcomes in this module is likely to depend less on changing the pedagogical target and more on reducing technical rejection at the most sensitive dialogue steps; otherwise, learners may be repeatedly penalised for system limitations rather than for communicative errors, and instructors may misinterpret error patterns as evidence of linguistic weakness.

These findings support two practical recommendations: (i) prioritise ASR-robust redesign of the most problematic dialogue turns (e.g., by expanding acceptable variants, adjusting recognition thresholds, and adding fallback interaction paths), and (ii) refine response options and prompts in steps with elevated selection errors to reduce ambiguity and better target the intended pragmatic learning outcomes. In addition, separating system-related failures from learner mistakes in reporting (e.g., presenting recognition failures as a distinct category in dashboards or teacher reports) would improve transparency and help stakeholders make better-informed decisions about learner progress and scenario quality.

Future work will refine the same dialogue by introducing carefully selected synonyms and paraphrase variants to increase ASR acceptance, and will then reassess performance using utterance-level signals (e.g., ASR confidence scores and, where permissible, audio or transcripts) to determine whether these revisions reduce recognition failures. If recognition rates improve without inflating incorrect-selection errors, this would provide empirical support for iterative scenario tuning as a practical pathway to increasing both usability and assessment validity in VR-based dialogue training.

Источники | References

1. Baker R. S., Inventado P. S. Educational Data Mining and Learning Analytics // Learning Analytics: From Research to Practice / ed. by J. A. Larusson, B. White. Springer, 2014. https://doi.org/10.1007/978-1-4614-3305-7_4
2. Bohus D. Error awareness and recovery in conversational spoken language interfaces: Doctoral dissertation. Pittsburgh: Carnegie Mellon University, 2007.
3. Bohus D., Rudnicky A. Sorry and I didn't catch that! – an investigation of non-understanding errors and recovery strategies // Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. Lisbon, 2005.
4. Broussard K. M. Errors in Language Learning and Use: Exploring Error Analysis by Carl James // TESOL Quarterly. 1999. Vol. 33. No. 1. https://doi.org/10.2307/3588202
5. Chen C., Yuan Y. Effectiveness of Virtual Reality on Chinese as a second language vocabulary learning: perceptions from international students // Computer Assisted Language Learning. 2023. Vol. 38 (3). https://doi.org/10.1080/09588221.2023.2192770
6. De Araujo A., Papadopoulos P. M., Mckenney S., De Jong T. A learning analytics-based collaborative conversational agent to foster productive dialogue in inquiry learning // Journal of Computer Assisted Learning. 2024. Vol. 40 (6). https://doi.org/10.1111/jcal.13007
7. De Vries B. P., Cucchiarini C., Bodnar S., Strik H., Van Hout R. Spoken grammar practice and feedback in an ASR-based CALL system // Computer Assisted Language Learning. 2014. Vol. 28 (6). https://doi.org/10.1080/09588221.2014.889713
8. Field A. Discovering Statistics Using IBM SPSS Statistics. 5th ed. Sage Publications, 2018.
9. Frigge M., Hoaglin D. C., Iglewicz B. Some implementations of the Boxplot // The American Statistician. 1989. Vol. 43 (1).
10. Graesser A., Jordan P., Vanlehn K., Rosé C., Harter D.Intelligent tutoring systems with conversational dialogue // AI Magazine. 2001. Vol. 22 (4). https://doi.org/10.1609/aimag.v22i4.1591
11. Heift T., Schulze M. Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues. Routledge, 2007.

12. James C. Errors in language learning and use: Exploring error analysis. Routledge, 2013.
13. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. Pearson, 2020.
14. Kang B. O., Jeon H., Lee Y. K. AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation // ETRI Journal. 2024. Vol. 46 (1). https://doi.org/10.4218/etrij.2023-0322
15. Knill K., Gales M., Kyriakopoulos K., Malinin A., Ragni A., Wang Y., Caines A. Impact of ASR Performance on Free Speaking Language Assessment // Interspeech 2018 (Hyderabad, India, 2-6 September 2018). 2018. https://doi.org/10.21437/interspeech.2018-1312
16. Lee A. Assessing Speaking Skills in Virtual Reality: Impacts and Implications // English Teaching. 2025. Vol. 80 (2).
17. Lev-Ari S.Comprehending non-native speakers: theory and evidence for adjustment in manner of processing // Frontiers in Psychology. 2015. Vol. 5. https://doi.org/10.3389/fpsyg.2014.01546
18. Lev-Ari S., Keysar B. Less-Detailed Representation of Non-Native Language: Why Non-Native Speakers' Stories Seem More Vague // Discourse Processes. 2012. Vol. 49 (7). https://doi.org/10.1080/0163853x.2012.698493
19. Li K.-C., Chang M., Wu K.-H. Developing a Task-Based Dialogue System for English Language Learning // Education Sciences. 2020. Vol. 10 (11). https://doi.org/10.3390/educsci10110306
20. Palmas F., Cichor J., Plecher D. A., Klinker G. Acceptance and Effectiveness of a Virtual Reality Public Speaking Training // IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (Beijing, China). 2019. https://doi.org/10.1109/ismar.2019.00034
21. Radzikowski K., Wang L., Yoshie O., Nowak R. Accent modification for speech recognition of non-native speakers using neural style transfer // EURASIP Journal on Audio, Speech, and Music Processing. 2021. Vol. 2021 (1). https://doi.org/10.1186/s13636-021-00199-3
22. Sadigzade Z. Immersive and Gamified Approaches: VR/AR in Language Learning // Porta Universorum. 2025. Vol. 1 (6). https://doi.org/10.69760/portuni.0106002
23. Scott D. W. Multivariate density estimation: Theory, practice, and visualization. Wiley, 2015.
24. Silverman B. W. Density estimation for statistics and data analysis. Chapman and Hall, 1986.
25. Thi-Nhu Ngo T., Hao-Jan Chen H., Kuo-Wei Lai K. The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis // ReCALL. 2023. Vol. 36 (1). https://doi.org/10.1017/s0958344023000113
26. Tobin J., Nelson P., Macdonald B., Heywood R., Cave R., Seaver K., Desjardins A., Jiang P.-P., Green J. R. Automatic Speech Recognition of Conversational Speech in Individuals with Disordered Speech // Journal of Speech, Language, and Hearing Research: JSLHR. 2024. Vol. 67 (11). https://doi.org/10.1044/2024_jslhr-24-00045
27. Tukey J. W. Exploratory data analysis. Addison-Wesley, 1977.
28. VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems // Educational Psychologist. 2011. Vol. 46 (4).
29. Wang Z., Schultz T., Waibel A.Comparison of acoustic model adaptation techniques on non-native speech // 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (Hong Kong, China): Proceedings. 2003. https://doi.org/10.1109/icassp.2003.1198837
30. Yang X., Chen Y.-N., Hakkani-Tur D., Crook P., Li X., Gao J., Deng L. End-to-end joint learning of natural language understanding and dialogue manager // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (New Orleans, LA, USA): Proceedings. 2017. https://doi.org/10.1109/icassp.2017.7953246

Информация об авторах | Author information

RU

Корзин Андрей Сергеевич[1]
Алексеева Наталия Александровна[2]
Шалеева Елена Федоровна[3]
Дмитриченкова Светлана Владимировна[4], к. пед. н., доц.
Круглова Лариса Владимировна[5], к. техн. н.
[1, 2, 3, 4, 5] Российский университет дружбы народов имени Патриса Лумумбы, г. Москва

EN

Andrey Sergeevich Korzin[1]
Nataliia Alexandrovna Alekseeva[2]
Elena Fedorovna Shaleeva[3]
Svetlana Vladimirovna Dmitrichenkova[4], PhD
Larisa Vladimirovna Kruglova[5], PhD
[1, 2, 3, 4, 5] Peoples' Friendship University of Russia, Moscow

[1] *korzin-as@rudn.ru,* [2] *alexnata02alexeeva@yandex.ru,* [3] *shaleeva-ef@rudn.ru,*
[4] *dmitrichenkova-sv@rudn.ru,* [5] *lar.kruglova@gmail.com*

Информация о статье | About this article

Дата поступления рукописи (received): 01.03.2026; опубликовано online (published online): 08.04.2026.

Ключевые слова (keywords): изучение языков с помощью виртуальной реальности (VR); обучение на основе диалогов; аналитика обучения; ошибки распознавания речи (ASR); оценка эффективности; virtual reality (VR) language learning; dialogue-based training; learning analytics; speech recognition (ASR) errors; performance assessment.