Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews

José Antonio Parejo[1], Amador Durán Toro[1], and Pablo Fernández Montes

SCORE Lab, University of Sevilla, Sevilla, Spain
japarejo@us.es,amador@us.es,pablofm@us.es

Abstract. Requirements elicitation interviews constitute a core yet challenging competence in Requirements Engineering (RE) education. Prior empirical research has identified 34 recurrent mistakes commonly committed by novice analysts during elicitation interviews, grouped into seven high-level categories. Recent work on AI-based interactive training suggests that LLM-backed clients can provide scalable and realistic practice environments for developing interview skills.

This paper presents the design of a confirmatory empirical study evaluating whether AI-supported training improves students' performance in subsequent peer-to-peer elicitation interviews. Students first interact with an AI-based client simulation and later conduct a live interview with a peer role-playing as the client. The study evaluates (i) the frequency of empirically validated interview mistakes and (ii) the structural quality of the resulting UML class diagrams (Mermaid format).

This submission follows the early review modality to obtain methodological feedback prior to execution.

Keywords: Requirements Engineering · Artificial Intelligence · Empirical Software Engineering · Computing Education

1 Introduction and Motivation

Requirements elicitation is a fundamental activity in software engineering, requiring both analytical and interpersonal skills. Interview-based elicitation is particularly demanding because the quality of the gathered information strongly depends on the interviewer's ability to formulate questions, manage the interaction, and progressively construct a shared understanding of the domain.

Empirical evidence shows that novices frequently commit systematic mistakes during elicitation interviews. A controlled study involving 110 students identified 34 unique mistakes grouped into seven categories: Question Formulation, Question Omission, Order of Interview, Communication Skills, Analyst Behaviour, Customer Interaction, and Teamwork and Planning [1]. These mistakes directly affect the quality of elicited requirements and subsequent modeling tasks. To address these challenges, peer-to-peer interviews involving role-playing and active reflection on the mistakes committed have been identified as valuable mechanisms to support training in elicitation skills [2,3].

At the same time, conducting realistic interview training in educational contexts is logistically expensive. Traditional approaches rely on static transcripts of

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

interviews, which train information extraction skills but not the dynamic ability to conduct an interview itself. Recent work proposes the use of generative AI to simulate interactive clients, enabling scalable practice of interpersonal interview skills [4]. Results suggest that AI-based interviews are perceived as more realistic and engaging than transcript-based exercises.

Building on this evidence, the present study evaluates whether AI-supported training effectively transfers to human interactions, improving students' performance in subsequent live elicitation interviews conducted between peers.

2 Educational Activity and Experimental Design

This study follows a between-subjects experimental design embedded within an undergraduate Requirements Engineering course. The activity spans two main sessions separated by a one-week interval.

2.1 Phase 1: Baseline Training (Session 1)

All participants receive a short lecture (20 minutes) on common mistakes in requirement elicitation interviews. To ensure a uniform baseline, all students perform an initial identical conceptual modelling exercise based on an interview transcript.

2.2 Phase 2: Experimental Treatment (Homework)

Following Session 1, students are randomly assigned to either the control or the experimental group. Both groups must complete two homework exercises regarding requirements elicitation:

- **Control Group (Transcript-based):** Students receive a static transcription of an interview for each exercise. They must critically assess it using an evaluation form structured around the taxonomy of common mistakes, which includes examples to guide the evaluation.
- **Experimental Group (AI-supported):** Students conduct a live interview with a simulated non-technical client powered by a Large Language Model (LLM), following the approach described in [5]. The AI client is prompted to avoid technical terminology, provide incremental information, require precise questions, and simulate realistic conversational behavior. After the interview, students evaluate the auto-generated transcript of their own session using the same evaluation form as the control group.

2.3 Phase 3: Peer-to-Peer Evaluation (Session 2)

One week later, all participants conduct a live elicitation interview in pairs using MS Teams.

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

- **The Client Role:** One student acts as the client and receives a short contextual document detailing the organization, their role, business objectives, and the "ground truth" solution (a UML class diagram containing a conceptual model of the actual information requirements).
- **The Requirements Engineer/Analyst Role:** The other student conducts a semi-structured interview, identifies the information requirements, and produces a UML class diagram in Mermaid format.

Dyadic Configurations: Since students from both groups are randomly paired, this naturally creates a 2×2 factorial design based on the training background of both actors in the interview. The four possible dyads [Client - Engineer] are:

1. **[Control - Control]:** Both trained with transcripts.
2. **[Experimental - Experimental]:** Both trained with AI.
3. **[Control - Experimental]:** Client trained with transcripts, Analyst trained with AI.
4. **[Experimental - Control]:** Client trained with AI, Analyst trained with transcripts.

This structure allows us to evaluate the primary effect of the analyst's training while explicitly controlling for any potential confounding influence or interaction effect originating from the client's training background.

After completing the first interview, students swap roles and are provided with a new, structurally equivalent context problem. To ensure comparable difficulty, all exercises involve domains with similar numbers of classes, attributes, and associations (e.g., university enrollment system, pizza shop management, theater ticketing). The domains use common knowledge to prevent domain-expertise bias.

Assessment: After each interview, the student playing the client evaluates the engineer's performance using the standard mistake-taxonomy form, providing a grade from 0 to 10 for each error area along with textual justifications. Furthermore, the client evaluates the resulting UML diagram against the ground-truth solution. MS Teams transcripts are saved for data validation.

3 Research Questions and Hypotheses

This confirmatory study aims to answer the following Research Questions (RQs): **RQ1:** Does prior AI-supported interview training reduce the number of empirically validated elicitation mistakes committed by the engineer during peer-to-peer interviews?

- $H0_1$: There is no difference in the grading of the interview mistakes between AI-trained and transcript-trained engineers.
- $H1_1$: Students receiving AI-supported training commit obtain a higher grade for the interview mistakes than those in the control group.

Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

RQ2: Does AI-supported training improve the structural quality of the resulting UML class diagrams?

- $H0_2$: There is no difference in UML class diagram quality between the groups.
- $H1_2$: AI-trained engineers produce higher-quality diagrams.

RQ3: Is there a correlation between the frequency of interview mistakes and the resulting model's quality?

- $H0_3$: There is no correlation between mistake frequency and diagram quality.
- $H1_3$: A negative correlation exists between mistake frequency and diagram quality.

The significance level is set to $\alpha = 0.05$.

4 Operationalization

4.1 Measurement of Interview Mistakes

The evaluation form groups the 34 common errors identified by [1] into seven specific error areas.

- **Area Grade:** Integer between 0 and 10 representing performance in a specific mistake category, where 0 means no mistakes on this area are present the interview, and 10 means that almost all the interactions/questions in the interview present a common mistake instance on this area.
- **Global Interview Score:** A continuous variable between 0 and 10, computed as the average of the grades across all seven areas.

4.2 Measurement of UML Model Quality

Diagrams are evaluated by the peer client using a structured rubric comparing the output against the ground-truth model. Measures include:

- **Domain Classes:** Correct identification (0-10).
- **Attributes:** Completeness and correctness of attribute allocation (0-10).
- **Associations:** Correct relationships and multiplicities (0-10).
- **Global Model Score:** Average score of the modeling dimensions (0-10).

5 Data Analysis Plan

5.1 Descriptive Analysis

We will report the mean, median, and standard deviation of interview gradings and model quality scores for all four dyadic configurations. Boxplots will be used to visualize the distribution of grades per mistake category and model dimension across the groups.

Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

5.2 Inferential Analysis

- **Assumptions check:** Normality will be assessed using the Shapiro–Wilk test, and homogeneity of variances via Levene's test.
- **Hypothesis Testing (RQ1 & RQ2):** A **Two-Way ANOVA** will be conducted to evaluate the main effect of the Engineer's training (primary independent variable), the main effect of the Client's training, and their interaction effect on interview grades and diagram quality. If assumptions of normality are violated, the Aligned Rank Transform (ART) for non-parametric factorial ANOVA will be used.
- **Effect Size:** Partial Eta Squared (η_p^2) will be calculated to measure the magnitude of the main and interaction effects.
- **Correlation (RQ3):** Pearson or Spearman correlation will evaluate the relationship between interview grades and model quality.
- **Regression:** A linear regression model will be fitted to predict diagram quality based on specific sub-categories of interview mistakes.

6 Importance and Expected Contribution

Interview-based elicitation remains difficult to teach due to its experiential nature. While traditional transcript-based approaches fail to develop interactive skills, AI-based clients provide scalable, low-risk practice environments.

This study will contribute to the RE education community by:

- Providing robust empirical evidence on whether AI-supported practice transfers to actual human-to-human interview settings, while accounting for different dyadic combinations.
- Validating the use of the 34-mistake taxonomy [1] as an active peer-assessment tool.
- Establishing quantitative links between interpersonal interview quality and conceptual modeling accuracy.

7 Threats to Validity

7.1 Internal Validity:

Differences in students' prior modeling experience could skew results. This is mitigated by assessing the baseline models created by students in Phase 1 and using these baseline scores as a covariate/moderator in our final analysis. Additionally, the behavior of the peer acting as the client might influence the engineer's performance depending on the client's own training background. This interaction threat is mitigated by the 2×2 factorial design, which explicitly incorporates the client's background into the statistical model.

Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

Parejo, J. A., Durán Toro, A., Fernandez, P.: Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

7.2 Construct Validity:

The primary threat is the subjectivity and reliability of peer-assessment (students grading students). To mitigate this, we provide a standardized rubric with explicit examples of mistakes and non-mistakes (e.g., valid use of synonyms). Additionally, the research team will cross-evaluate a random sample of the recorded MS Teams transcripts to calculate inter-rater reliability (e.g., Cohen's Kappa) and validate the peer scores.

7.3 External Validity:

The study is conducted in a specific academic context with undergraduate students, which may not perfectly generalize to experienced practitioners in industry. However, the focus of the study is specifically on novice skill acquisition, making the population appropriate.

8 Reproducibility Plan

To ensure open science principles, a replication package will be made available via Zenodo/OSF, including:

- Teaching materials on common mistakes on requirements elicitation interviews.
- UML diagrams provided as gloden solution and exercise context descriptions.
- Anonymised interview transcripts.
- UML diagrams provided as solution for each interview.
- Evaluation forms.

Acknowledgments. This work was partially supported by grants PID2021-126227NB-C21 and PID2021-126227NB-C22 funded by MCIN / AEI / 10.13039 / 501100011033 / FEDER, UE "ERDF"; PID2024-155693NB-C44 funded by MICIU / AEI / 10.13039/ 501100011033 / FEDER, UE; and by "Smart Computer systems Research and Engineering (SCORE)", funded by the Consejería de Transformación Económica, Industria, Conocimiento y Universidades of Junta de Andalucía under the "Programa de Fortalecimiento de Institutos Universitarios de Investigación de la Universidades Andaluzas, Centros e Infraestructuras para la adquisición del sello 'Severo Ochoa' o 'María de Maeztu'".

AI Use Disclosure Artificial intelligence (AI)–based language tools were used solely to assist with language editing and improving clarity in the preparation of this manuscript. All generated content was carefully reviewed and revised by the authors, who assume full responsibility for the accuracy, integrity, and originality of the work. The AI tools did not contribute to the design, analysis, results interpretation, or conclusions.

This work is licensed under a Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International" license.

References

1. Bano, M., Zowghi, D., Ferrari, A., Spoletini, P., Donati, B.: Learning from mistakes: An empirical study of elicitation interviews performed by novices. In: 2018 ieee 26th international requirements engineering conference (re). pp. 182–193. IEEE (2018)
2. Ferrari, A., Spoletini, P., Bano, M., Zowghi, D.: Learning requirements elicitation interviews with role-playing, self-assessment and peer-review. In: 2019 IEEE 27th international requirements engineering conference (RE). pp. 28–39. IEEE (2019)
3. Ferrari, A., Spoletini, P., Bano, M., Zowghi, D.: Sapeer and reversesapeer: teaching requirements elicitation interviews with role-playing and role reversal. Requirements Engineering **25**(4), 417–438 (2020)
4. Lojo, N., González, R., Philip, R., Parejo, J.A., Toro, A.D., Fox, A., Fernández, P.: C-leia: An interactive llm-based simulator for requirements interviews. In: 2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW). pp. 59–61. IEEE (2025)
5. Lojo, N., González, R., Philip, R., Parejo, J.A., Toro, A.D., Fox, A., Fernández, P.: Using large language models to develop requirements elicitation skills (2025), `https://arxiv.org/abs/2503.07800`

Parejo, J. A., Durán Toro, A., Fernandez, P.; Evaluating the Impact of AI-Simulated Clients on Novice Performance in Elicitation Interviews. In: Cetina, C. (ed.) Actas de las XXX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2026). Sistedes (2026). https://hdl.handle.net/11705/JISBD/2026/73

This work is licensed under a Creative Commons “Attribution-NonCommercial-NoDerivatives 4.0 International” license.