Assessing Language Proficiency through AI Chatbot-Based Evaluations

Fithriyah Inda Nur Abida[1✉], Rahayu Kuswardani[2], Oikurema Purwati[3], Abdul Rosyid[4], Etik Minarti[5]
Universitas Negeri Surabaya, Indonesia[1,2,3]
Universitas Trunojoyo Madura, Indonesia[4]
SMP 28 Surabaya, Indonesia[5]
✉fithriyahabida@unesa.ac.id

Abstract:

This study aims to investigate the effectiveness of using AI chatbots for assessing language proficiency. The study employs a mixed-methods approach, combining quantitative analysis of chatbot evaluation scores with qualitative analysis of participant feedback. The AI chatbot-based evaluations were conducted with a sample of language learners from the State University of Surabaya. The evaluations focused on various linguistic aspects, including grammar, vocabulary, pronunciation, and fluency. The results demonstrate that the AI chatbot-based evaluations provide accurate and consistent language proficiency assessments. The correlation between the chatbot evaluation scores and standardized language proficiency scores is found to be strong, indicating the reliability of the chatbot assessments. Furthermore, participants reported positive experiences with the chatbot assessments, highlighting the usefulness of the instant feedback provided. The chatbot was able to identify language errors and offer targeted suggestions for improvement, enhancing the language learning process. These findings support the effectiveness and potential of AI chatbot-based evaluations for assessing language proficiency. The integration of chatbots in language education can provide valuable tools for both learners and educators, offering objective assessments and personalized feedback.

Keywords: assessment; language proficiency; AI chatbot; evaluations

INTRODUCTION

Assessing language proficiency plays a crucial role in language learning as it helps measure an individual's ability to understand, use, and communicate in the target language (Richards, 2015; Canh & Willy, 2017). Accurate and objective assessments are necessary to identify learners' strengths and weaknesses, determine their level of progress, and design appropriate teaching strategies. However, the traditional evaluation methods commonly used today often involve human assessments with several limitations. Firstly, the human assessment process is time-consuming and costly. For example, in oral assessments, an assessor needs to manually listen to learners' conversations or presentations, analyze errors, and provide relevant feedback. This process can be time-consuming, especially in large classes. Additionally, the subjectivity of human assessment can influence evaluation outcomes (Norbert et al., 2011; Tran and Jarvinen, 2022). Assessments can be influenced by assessors' preferences, perceptions, or personal judgments, leading to an imbalance in the assessment's objectivity and consistency.

When language assessments are conducted by different individuals, the results can vary due to differences in subjectivity, preferences, and perceptions of individuals (Burke, 2009). This can result in inconsistent and unreliable evaluations. For instance, two different assessors may provide different assessments for the same quality or level of errors in a learner's language response. The use of chatbots as language evaluation tools offers the potential to overcome these limitations. Chatbots can be programmed with consistent and objective natural language processing algorithms and models (Pranav, 2022). This means that chatbots will evaluate learners' language responses based on predefined rules and criteria without any individual preferences or perceptions that can influence the assessment. By using chatbots, each language response will be assessed based on the same parameters, providing more consistent and reliable evaluations. If learners' language responses are analyzed by multiple identical chatbots, the evaluation results given will be similar for the same response, without variations influenced by subjective factors.

By leveraging artificial intelligence (AI) and natural language processing (NLP) technology, chatbots can provide objective and consistent assessments (Kooli, 2023). Chatbots can automatically analyze learners' language responses. Through natural language processing, chatbots can understand the text or conversation learners provide, analyze its structure and content, and identify errors made (Jianfeng et al., 2019), (Chang Lin, 2023). This enables chatbots to provide specific and relevant feedback to learners, guiding them to correct their mistakes and understand areas that need improvement in their language proficiency. Additionally, chatbots can offer instant feedback. With the speed of AI processing, chatbots can provide feedback immediately after receiving a response from learners. This reduces the time required for feedback, allowing learners to promptly receive information about their strengths and weaknesses. This instant feedback can assist learners in adaptive learning and quickly improve their language skills (Lin Mubarok, 2021). By combining artificial intelligence with language assessment, the use of chatbots as evaluation tools has the potential to enhance efficiency, accuracy, and consistency in the language proficiency assessment process.

METHODOLOGY

This research conducted a mixed-methods experimental design. This approach combines quantitative analysis, focusing on numerical data derived from chatbot evaluation scores, with qualitative analysis, which involves examining participant feedback in more detail (Cassel Symon, 2014; Sugiyono, 2015; Cresswell, 2018). Quantitative analysis involves collecting and analyzing numerical data obtained from the chatbot evaluations. This could include variables such as overall proficiency scores, scores for specific language skills or competencies, or comparisons between chatbot evaluation scores and standardized language proficiency scores. By quantifying the data, researchers can identify patterns, trends, and correlations that provide insights into the effectiveness of the chatbot-based evaluations. Qualitative analysis involves examining the participant feedback obtained during the evaluation process. This feedback may include comments, opinions, suggestions, or explanations provided by the participants in response to the chatbot's assessment.

The researchers conducted an experimental study that included two distinct groups: a control group and an experimental group. The participants were recruited from undergraduate students from the State University of Surabaya on a voluntary basis. There are two groups of students: experimental and control groups (Campbell and Stanley,1963). The initial participant number of each group is the same, with 30 students. The control group underwent the language proficiency evaluation using the traditional method, where their responses were assessed by experienced human raters.

On the other hand, the experimental group completed the same evaluation tasks but with the assistance of an AI chatbot specifically developed for this study. The evaluation tasks were carefully designed to cover various language aspects and difficulty levels. They were adaptable to be evaluated by both human raters and the AI chatbot. The participants in both groups were provided with the same set of tasks and instructions to ensure fairness and consistency.

During the evaluation process, the researchers collected data, including the participants' responses to the evaluation tasks. This data also encompassed the feedback human raters and the chatbot provided. The researchers ensured the privacy and confidentiality of the participants' data and adhered to ethical guidelines throughout the study (Cresswell, 2012). Following data collection, the researchers conducted a comparative analysis between the control group and the experimental group. They employed statistical analysis techniques to assess the alignment of evaluation results between human raters and the chatbot. This analysis aimed to determine the reliability and validity of the chatbot-based evaluation method compared to the traditional human assessment.

RESULT AND DISCUSSION

The research findings indicate that chatbot assessments have a high level of consistency with human assessments. Chatbots are able to generate evaluations similar to those conducted by humans and provide accurate and consistent assessments of participants' language proficiency. The comparison between chatbot evaluation scores and participant language proficiency scores measured using standardized assessment tools shows a strong correlation. In this study, chatbot evaluation scores were compared to participant language proficiency scores measured using standardized assessment tools. The results of this comparison demonstrate a strong correlation between chatbot evaluation scores and participant language proficiency scores. The strong correlation indicates that the evaluations conducted by the chatbot align with the assessments conducted using standardized evaluation tools (Broom, 2012). This means that chatbots are capable of providing assessments comparable to human assessments, validating the chatbot's ability to measure participants' language proficiency. When chatbot evaluation scores and participant language proficiency scores have a strong correlation, it indicates that the chatbot can be trusted in conducting language assessments. This has important implications in the context of language teaching and learning as it can provide an effective and efficient alternative for evaluating participants' language proficiency.

The research also demonstrates that chatbots are capable of providing specific feedback to participants. Chatbots have the ability to identify language errors made by participants and offer precise suggestions for improvement tailored to the relevant context. This adds value to the language learning process as learners can receive focused feedback that aids in enhancing their language proficiency. In the context of language education, specific feedback is crucial in helping learners understand their mistakes and guiding them toward betterment (Brown, 2010). By employing artificial intelligence (AI) and natural language processing (NLP) technologies, chatbots can analyze participants' language responses in detail and recognize common error patterns.

Consequently, chatbots can provide appropriate suggestions for improvement based on the errors made by the participants. The advantage of specific feedback from chatbots lies in their ability to offer targeted and relevant information and provide examples or additional explanations that help learners comprehend their errors. This enables learners to rectify their mistakes more effectively and enhance their language skills in a more focused manner. With the capability of chatbots to provide specific feedback, learners can feel supported and guided throughout the language learning process. They can swiftly identify areas that require improvement and concentrate on specific aspects that demand further attention. In the long run, this can assist learners in developing their language proficiency more effectively and efficiently.

The use of chatbots in assessing language proficiency also offers advantages in terms of efficiency (Russel, 2016), (Pérez, 2020). Chatbots can provide instant feedback, reducing the time required for manual feedback by human evaluators. When participants answer questions or complete language tasks evaluated by chatbots, they promptly receive feedback regarding errors or areas that need improvement. In traditional assessments involving human evaluators, feedback often takes considerable time to be provided. Evaluators have to read, analyze, and provide feedback to each participant manually. This process can be time-consuming, especially when dealing with a large number of participants. However, with the use of chatbots, feedback can be delivered instantly. Chatbots have the ability to automatically analyze participants' language responses and provide pre-programmed feedback (Gao Li, 2020). This reduces the time needed to provide feedback to each individual participant. This advantage brings efficiency to the assessment process. Learners can immediately see their evaluation results and receive relevant feedback quickly. This allows learners to promptly address their mistakes and develop their language skills without having to wait long for feedback.

Furthermore, the use of chatbots in language assessment also reduces the workload of human evaluators. With chatbots taking on the role of providing instant feedback, human evaluators can focus on tasks that require more in-depth assessment or aspects that cannot be handled by chatbots, such as content evaluation or authenticity. Thus, using chatbots to assess language proficiency provides a time efficiency advantage. Learners can receive feedback quickly, while human evaluators can allocate their time and resources to more complex assessment aspects. This enhances overall efficiency in the assessment process and accelerates learners' progress in developing their language skills.

Although chatbots have proven to be effective in evaluating language proficiency, research also identifies several challenges and opportunities for improvement. One of the challenges faced in using chatbots for language assessment is their ability to understand complex language contexts and nuances. While chatbots can effectively identify grammatical errors and clear pronunciation, they may not yet fully capture differences in the use of more subtle language elements, such as idioms, slang, or cultural references. Therefore, developing more sophisticated chatbots sensitive to language context and nuances presents a potential opportunity for improving evaluation accuracy. Additionally, chatbots in language assessment may be limited in providing deeper and contextual feedback. The feedback provided by chatbots tends to be generic and cannot take into account individual aspects of learners. Furthermore, chatbots cannot fully replace the role of human assessors in language evaluation (Kooli, 2023). The expertise and understanding of humans in comprehending language complexities and providing more holistic assessments cannot be easily replaced by chatbots. Therefore, there needs to be a good balance between using chatbots and human assessment in the evaluation process, where chatbots can be used as a supporting tool to provide instant feedback. In contrast, human assessors continue to play a crucial role in deeper and contextual assessment.

CONCLUSION

The research findings indicate that chatbots possess the capacity to deliver accurate, reliable, and impartial assessments of participants' language skills. The strong correlation observed between chatbot assessment scores and standardized language proficiency scores further supports this claim. Additionally, chatbots can provide participants with tailored feedback and offer efficiency advantages. The demonstrated precision, consistency, and efficiency of chatbots make their incorporation into language education highly advantageous. These findings hold the potential to influence instructional design, curriculum development, and evaluation practices in the realm of language education. However, there are two specific aspects that chatbots still cannot accomplish in language assessment: subjective assessment and contextual depth. The feedback provided by chatbots tends to be generic and cannot take into account individual aspects of learners. Furthermore, chatbots cannot fully replace the role of human assessors in language assessment. The expertise and understanding of humans in comprehending the complexity of language cannot be replicated by chatbots, thus indicating that more holistic assessments still require human assistance. Therefore, achieving a proper balance between the utilization of chatbots and human assessors in the evaluation process becomes essential. Chatbots can serve as valuable supplementary tools for providing immediate feedback, while human assessors remain vital in conducting more in-depth and contextually nuanced assessments.

REFERENCES

Broom, C. (2012). Assessment and Evaluation: Exploring their Principles and Purposes in Relation to Neoliberalism through a Social Studies Case Study. *Canadian Social Studies, 45*(2), 17–36.

Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices* (2nd edition). White Plains, NY: Pearson Education.

Burke, K. (2009). *How to Assess Authentic Learning* (5th edition). Thousand Oaks, CA: Corwin.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Design for Research*. Chicago, IL: Rand Mc Nally College Publishing Company.

Canh, L. V., & Renandya, W. A. (2017). Teachers' English Proficiency and Classroom Language Use: A Conversation Analysis Study. *RELC Journal, 48*(1–2), 1–15. https://doi.org/10.1177/0033688217690935

Cassel, C., & Symon, G. (2014). *Essential Guide to Qualitative Methods in Organizational Research*. Los Angeles, CA: Sage.

Chang Lin, C., Huang, A., & Yang, S. J. H. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability, 15*(5), 4012. https://doi.org/10.3390/su15054012

Creswell, J. W. (2012). *Educational Research*. Pearson.

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods*. California: SAGE Publications.

Gao, X., Zhang, Y., & Li, Q. (2020). An AI chatbot-assisted language learning system for EFL students. *IEEE Transactions on Learning Technologies, 14*(4), 520–531.

Jianfeng, G., Michel, G., & Lihong, L. (2019). Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval, 13*(2–3), 127–298.

Kooli, C. (2023). Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions. *Sustainability, 15*(7), 5614. https://doi.org/10.3390/su15075614

Lin, C. J., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society, 24*(3), 16–35.

Norbert, M., Schwan, W., & Bless, H. (2011). Subjective Assessments and Evaluations of Change: Some Lessons from Social Cognition. *European Review of Social Psychology, 22*(1), 181–210. https://doi.org/10.1080/10463283.2011.567489

Pérez, J. Q., Daradoumis, T., & Marquès Puig, J. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education, 28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

Pranav, D. S., Mutreja, M., Punj, D., & Chawla, P. (2022). Natural Language Processing in Chatbots. In Emerging Technologies in Data Mining and Information Security (pp. 87–98). *Proceedings of IEMIS 2022*, *Volume 3*. Conference proceedings. Institute of Engineering & Management, Kolkata, India. https://doi.org/10.1007/978-981-19-4193-1_9

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Sugiyono. (2015). Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D). Bandung: Alfabeta.

Tran, T.-V., & Järvinen, J. (2022). Understanding the concept of subjectivity in performance evaluation and its effects on perceived procedural justice across contexts. *Accounting and Finance, 62*(3), 1025–1056. https://doi.org/10.1111/acfi.12916