Balancing interactional authenticity and variability in the assessment of interactional competence: A comparative study of human interlocutors and conversational virtual agent

Ryo Takagi, Shungo Suzuki, Mao Saeki, and Yoichi Matsuyama
Waseda University

Considering the authenticity of speaking tests, scholars have explored the test design that can capture test takers' interactional competence, as opposed to what can be assessed in monologic tasks (for an overview, see Galaczi & Taylor, 2018). Roever and Ikeda's (2021) study supports such potential gap between test scores in monologic and dialogic tasks, reporting only the 56% of shared variance between them. One important issue in designing the task for interactional competence is the interactional authenticity, namely, to what extent the task can elicit symmetrical interaction between test-takers and examiners, such as turn-taking management and topic negotiation (e.g., Galaczi & Taylor, 2018). Meanwhile, dialogic tasks entail variability in examiners' behaviours due to the co-constructed nature of interaction (i.e., interactional variability; Galaczi & Taylor, 2018), which subsequently affect the reliability of the tests.

One possible solution to this trade-off between interactional authenticity and variability is the use of conversational artificial intelligent (AI) virtual agents (hereafter, AI agent), which systematically and consistently respond to test-takers' reponses with assistance of natural language processing and multimodal machine learning techniques, as an interlocutor in interactive speaking tasks. Thanks to the systematicity of AI agent's behaviour, the challenges in reliability in interactive tasks could be reduced. However, another issue that needs to be addressed includes how similar test-takers' behaviours are between human and AI agent examiners and to what extent an AI agent can elicit ratable speech samples for the given assessment purpose (e.g., placement, diagnosis). In order to address these issues, we plan to conduct a comparative study on the ratability of interactional speech samples between human and AI agent interlocutors.

The current project aims to develop an English speaking placement test at a private university in Japan. The majority of the target population of the test is Japanese learners of English whose proficiency level ranges from A1 to C1 level on the CEFR scale. The current study thus plans to recruit 10 students from each proficiency level (N = 50 in total). In-service teachers in the curriculum will serve as human interviewers. Students will perform two speaking tasks—interview and role play—with two conditions of interlocutors (human vs. AI agent; a total of 200 interactional samples). The interview task provides an asymmetrical interaction where the interviewer constantly asks questions to students on a range of topics including travel and social media, while the role play task offers a symmetrical interaction where a student takes turns with an interlocutor as a tour guide to book a tour plan.

The interactional data will be analyzed in terms of the proximity of AI agents to human interlocutors using interactive evaluation protocols, in which experts judge the quality of various dimensions of interaction with scaler ratings (Finch et al., 2020). The ratability of interactional data will also be evaluated in terms of the amount of interactional features across interactional

conditions, following previous studies (e.g., Galaczi, 2014; Galaczi & Taylor, 2018; May, 2020). In the presentation, the results of a small pilot will be discussed with the audience for exploring more valid test design and appropriate assessment criteria.

(250-word summary)

One of the issues in the assessment of interactional competence is the balance between interactional authenticity (i.e., symmetricity of interaction between test-takers and examiners) and interactional variability (i.e., variability in examiners' behaviours; see Galaczi & Taylor, 2018). One possible solution to this trade-off between interactional authenticity and variability is the use of conversational artificial intelligent (AI) virtual agents (hereafter, AI agent), which systematically and consistently respond to test-takers' responses as an interlocutor in interactional speaking tasks. However, another issue that needs to be addressed includes how similar test-takers' behaviours are between human and AI agent examiners and to what extent an AI agent can elicit ratable speech samples for the given assessment purpose (e.g., placement, diagnosis). In order to address these issues, we plan to conduct a comparative study on the ratability of interactional speech samples between human and AI agent interlocutors. The current study thus plans to recruit 50 students with varying proficiency levels. Students will perform two speaking tasks—interview and role play—with two conditions of interlocutors (human vs. AI agent). The interactional data will be analyzed in terms of the proximity of AI agents to human interlocutors using interactive evaluation protocols (Finch et al., 2020). The ratability of interactional data will also be evaluated in terms of the amount of interactional features across interactional conditions following previous studies (e.g., Galaczi, 2014; May, 2020). In the presentation, the results of a small pilot will be discussed with the audience for exploring more valid test design and appropriate assessment criteria.