Generative AI as Writing or Speaking Partners in L2 Learning: Implications for Learning-Oriented Assessments

Soo Hyoung Joo
Teachers College, Columbia University

The advent of generative AI (GenAI) technology has impacted second language (L2) learning and assessment, offering new opportunities for learners to practice and improve their skills. One approach gaining interest is employing GenAI tools as writing or speaking partners to provide personalized, real-time feedback and assistance to learners. These interactions allow learners to practice their writing and speaking skills while receiving assessment information on various aspects of language, including grammar, vocabulary, and pronunciation. Considering the potential of GenAI tools to enrich assessment and learning experiences, it is worth examining recent research on the use of this technology for this purpose.

This paper reviews the literature on the use of GenAI as writing and speaking partners through the lens of the Learning-Oriented Assessment (LOA) framework (Purpura, 2024; Turner & Purpura, 2016) to explore how assessment data from GenAI tools could be leveraged to further learning. The LOA framework provides a structured approach to examine the educational value of engaging with assessment using GenAI tools across eight dimensions. These include the *contextual* dimension, which situates the performance within specific real-life language use domains; the *proficiency* dimension, which outlines the linguistic or topical resources assessed; and the *elicitation* dimension, which concerns how these resources are elicited through tasks. Five other dimensions relate to moderators of performance: the *socio-cognitive* dimension, which involves the cognitive processes and strategies learners employ during the assessment; the *instructional* dimension, which considers how the assessment aligns with the curriculum, teaching methods, and learning objectives; the *affective* dimension, which takes into account learners' emotions, attitudes, and motivation; the *social-interactional* dimension, which concerns the interactional practices required in the sociocultural context; and the *technological* dimension, concerning learners' technological literacy in relation to the capabilities and constraints of technology.

While research on GenAI tools from an LOA perspective remains limited, this paper seeks to classify the key aspects identified in the existing literature according to the eight LOA dimensions and discuss their potential implications. As the review progresses, relevant LOA dimensions identified by the author will be noted in parentheses. Table 1 provides a summary of the studies reviewed, along with the associated LOA dimensions.

TABLE 1
Studies Reviewed in the Paper Categorized by LOA Dimensions

Skill	Study	Method	Research Topic	LOA Dimensions
Speaking	Godwin-Jones (2022)	Literature review	AI tools for writing assistance in L2 learning contexts	technology, instructional socio-cognitive, contextual
	Escalante et al. (2023)	Quasi-experimental	The efficacy of GenAI feedback on writing performance and perceptions among L2 learners	socio-cognitive, proficiency, social-interactional, affective, technology
	Cheng et al. (2024)	Experimental	A broadening of the construct of writing ability to account for human-AI collaborative writing	elicitation, socio-cognitive, proficiency, technology
Writing	Youn (2023)	Literature review	The potential of spoken dialogue systems and intelligent personal assistants in assessing L2 speaking	elicitation, proficiency, social-interactional, technology
	Fathi et al. (2024)	Mixed-methods	The impact of AI-mediated interactive speaking activities on L2 learners' speaking skills and WTC	socio-cognitive, proficiency, social-interactional, affective, instructional, technology
	Wan and Moorhouse (2024)	Technology review	The functionality and potential of "Call Annie" in facilitating L2 development	technology, socio-cognitive, affective, instructional

First, this review considers the research on AI tools for writing. As outlined above, three recent studies have explored the role of GenAI as writing partners, each examining different aspects of AI-assisted writing. To begin with, Godwin-Jones (2022) provided a comprehensive review of AI tools for writing assistance in L2 learning contexts, including Gen-AI tools, discussing the technical affordances and challenges (technology dimension), instructional considerations (instructional dimension), and the efficacy in enhancing writing performances (socio-cognitive dimension). The study highlighted the challenges of using AI tools as writing partners, including concerns about the authenticity and creativity of the written product, as well as the potential for biased views and hateful language in generated texts. For instance, AI-human collaborative writing typically involves an iterative process where humans consult the AI to refine the content and language. As the contributions often blend together, the distinction between human and AI-generated content can be blurred. Moreover, the study noted that gen-AI models might generate texts that read well but lack substantive content, which poses challenges for using them to provide content feedback (Zhang & Li, 2021).

To address these challenges, Godwin-Jones (2022) emphasized the importance of developing learners' AI literacy, including understanding how AI-based systems work and how to use them in writing effectively. The study also highlighted the need to integrate AI tools thoughtfully into writing instruction, ensuring that they take a supporting role rather than lead the learning process. As Link et al. (2020) suggested, AI tools can provide sentence-level feedback while teachers focus on higher-level writing issues. Furthermore, Godwin-Jones (2022) underscored the implications for research, emphasizing the need for more systematic and critical empirical studies to identify the individual and contextual factors that influence the effectiveness of AI-generated feedback (contextual dimension). This includes validating the claims made by software companies and holding them accountable (Chapelle et al., 2015; Ranalli, 2021).

As a response to this call, Escalante et al.'s (2023) quasi-experimental study investigated the effectiveness of GenAI feedback in improving writing performance (socio-cognitive,

instructional dimensions) and perceptions (social-interactional, affective dimensions) among L2 learners. The study compared the writing development of students who received feedback from either ChatGPT, powered by GPT-4, or a human tutor over a six-week period. Additionally, it examined L2 learners' preferences and perceptions of AI-generated versus human tutor feedback. The results showed no significant differences in writing scores between AI-generated and human tutor feedback conditions, suggesting that AI-generated feedback was as effective as human feedback in facilitating writing development. Also, in terms of the perceptions, students were nearly equally divided in their preference between AI-generated feedback and human feedback. Students who preferred human feedback valued the engagement and interaction afforded by face-to-face tutoring, while those who preferred AI feedback appreciated its clarity, specificity, and 24/7 availability. Based on this mixed result, the authors propose a blended approach leveraging the strengths of both AI and human feedback to optimize the writing feedback process.

Taking a step beyond investigating the role of AI-generated feedback in enhancing writing performances, Cheng et al. (2024) highlighted the importance of broadening the construct of writing ability to account for human-AI collaborative writing (proficiency dimension). Cheng et al. (2024) adopted the evidence-centered design framework (e.g., Mislevy et al., 2003) to develop a writing assessment tool, "CoAuthor," that accounts for human-AI collaborative writing behaviors (elicitation dimension) and analyzed data from 1,445 writing sessions. The study examined two key conditions in the CoAuthor dataset: 1) ownership, as defined by the percentage of the text written by the human versus AI, and 2) the type of writing prompt, such as a creative essay or argumentative essay prompt. The results showed that writing processes differed across these conditions. When the human ownership was higher, the text composition and revision were more extensive. On the other hand, when the human ownership was lower, the AI suggestions were used more verbatim. The study also found that creative writing prompts led to more exploration of AI suggestions, while argumentative writing prompts led to more integration through revision. The study's findings on ownership may be flawed due to its circular definition of the variable, which defines ownership as the low frequency of adapting AI suggestions, thereby predetermining the outcome that those with higher ownership tend to produce more original texts and revise AI suggestions more frequently. This limitation notwithstanding, the study still offers a valuable framework for assessing human-AI collaborative writing and underscores the importance of considering factors such as ownership and writing prompt type in understanding the socio-cognitive dimension of co-authoring.

In addition to being used for writing, GenAI has also opened up new possibilities for interactive speaking practices, as Youn (2023), Fathi et al. (2024), and Wan and Moorhouse (2024) illustrate. Youn (2023) investigated the potential of spoken dialogue systems (SDS) and intelligent personal assistants (IPAs) in assessing L2 interactive speaking (social-interactional). Youn (2023) highlighted that SDS and IPAs can provide L2 learners with opportunities for authentic, interactive speaking practice, even in the absence of human interlocutors. For example, Ockey and Chukharev-Hudilainen (2021) designed an SDS for a paired speaking task (elicitation), demonstrating the potential of AI to elicit evidence of interactional competence (proficiency dimension). However, Youn (2023) also raised concerns about the authenticity of the interaction and the ability of AI-powered voice recognition technology to accommodate diverse English varieties and proficiency levels. Despite these limitations, Youn (2023) argued that the judicious incorporation of emerging AI technologies, like SDS and IPAs, into teaching

and assessment practices could expand opportunities for learning and assessing L2 interactive speaking.

While AI technologies hold promise for interactive speaking practices, further investigation into their effectiveness as well as their impact on learners' psychological and emotional dispositions, is critical. To bridge this gap, Fathi et al. (2024) investigated the impact of AI-mediated interactive speaking activities on L2 learners' speaking skills (proficiency, socio-cognitive, instructional) and willingness to communicate (WTC) in a mixed-methods study incorporating a semi-structure interview probing the perceptions towards the AI speaking partner (affective, socio-interactional). The results revealed that AI-mediated activities were more effective in enhancing speaking skills and WTC compared to face-to-face instruction. Furthermore, learners held positive attitudes and perceptions towards AI-mediated speaking instruction, citing benefits such as personalized feedback and a low-pressure environment for practice.

Building on the development of voice communication partners, recent advancements in GenAI technology have enabled the integration of video capabilities, allowing for the incorporation of nonverbal cues, such as facial expressions and body language, in these speaking partners. Wan and Moorhouse (2024) provide a review of "Call Annie," a GenAI video chatbot designed to support L2 learners in developing their speaking skills. Wan and Moorhouse (2024) mainly focus on the technological affordances (technology dimension) but also highlight the chatbot's potential in facilitating L2 speaking development (socio-cognitive dimension) by providing opportunities for meaningful interaction, personalized feedback, and anxiety reduction (affective dimension). They also suggest possible applications in classroom settings (instructional dimension), such as integrating the chatbot into self-directed learning, in-class activities, and homework assignments. However, they call for empirical research to investigate the tool's long-term effectiveness and address potential limitations.

While the research on GenAI's role as a writing or speaking partner is still in its preliminary stages, it has significant implications for LOAs in that the assessment information from GenAI can play a critical role in promoting learning, and the assessment process itself carries inherent learning values. GenAI can be incorporated into LOAs in two ways: 1) as partners in interaction and collaboration and 2) as partners for feedback and instruction. When considering the role of GenAI as writing collaborators or speaking interlocutors, a theme that emerged in the studies was the need to redefine the construct of L2 proficiency (proficiency dimension; Cheng et al., 2024; Youn, 2023) to mirror the changing landscape of communication (contextual dimension), where the interaction with GenAI tools becomes increasingly common. There has been a growing effort to create assessments (elicitation dimension) that mirror this change by measuring interactional or collaborative competence (Cheng et al., 2024; Ockey & Chukharev-Hudilainen, 2021). Although these studies have attempted to account for the interactional aspect of human-AI collaboration (social-interactional dimension) and the cognitive processes involved in these collaborations (socio-cognitive dimension), further research is necessary on how these performance moderators can be engineered into the assessment design.

Studies have also claimed that AI-generated feedback can be perceived as clear and specific (Escalante et al., 2023; Fathi et al., 2024) and can promote the development of language performances (socio-cognitive dimension). However, given the concerns regarding the overfocus on surface-level features and the role of learner "ownership" of the performance indicators, as outlined in Cheng et al. (2024), it is recommended to conduct explicit training on AI literacy (technology dimension) and take a blended approach where AI takes a supporting role in teacher

instruction (instructional dimension). Also, further research on contextual factors, such as administrative requirements, technology infrastructure, and trust in technology and its impact on learning, should also be considered (Godwin-Jones, 2022; contextual dimension).

As the use of GenAI as writing and speaking partners becomes more prevalent, it is crucial to adopt a systematic approach to define the skills and competencies involved in AI-human language interactions. One way to do this is by adopting a framework like LOA, which can help us thoughtfully define the construct of AI-human language competencies and integrate GenAI tools into language assessment to promote learning. Given the capacity of GenAI to elicit evidence of complex competencies and provide immediate, personalized feedback, GenAI speaking and writing partners can support self-assessment and reflection, helping learners track progress over time. Moreover, these tools can offer educators valuable insights into individual learning trajectories, allowing them to design more effective instructional strategies.

REFERENCES

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Cheng, Y., Lyons, K., Chen, G., Gašević, D., & Swiecki, Z. (2024). Evidence-centered assessment for writing with generative AI. In *The 14th Learning Analytics and Knowledge Conference (LAK '24)*, March 18–22, 2024, Kyoto, Japan. ACM. https://doi.org/10.1145/3636555.3636866

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(57). https://doi.org/10.1186/s41239-023-00425-2

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, *121 (*103254). https://doi.org/10.1016/j.system.2024.103254

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. http://doi.org/10125/73474

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2020.1743323

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Ockey, G., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*, *42*(5), 924–944. https://doi.org/10.1093/applin/amaa067

Purpura, J. E. (2024). Learning-oriented language assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*. Wiley.

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*(100816). https://doi.org/10.1016/j.jslw.2021.100816

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–273). De Gruyter.

Wan, Y., & Moorhouse, B. L. (2024). Using Call Annie as a generative artificial intelligence speaking partner for language learners. *RELC Journal*. Advance online publication. https://doi.org/10.1177/00336882231224813

Youn, S. J. (2023). Test design and validity evidence of interactive speaking assessment in the era of emerging technologies. *Language Testing, 40*(1), 54–60. https://doi.org/10.1177/02655322221126606

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research, 1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Soo Hyoung Joo is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her research interests include assessing 21C competencies through scenario-based language assessment and the role of domain knowledge in specific-purpose language assessment. Correspondence should be sent to E-mail: shj2118@tc.columbia.edu.