ReCALL (2025), 1–17
doi:10.1017/S0958344025100268

RESEARCH ARTICLE

Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis‡

Zhuohan Hou
Zhejiang University, China (h_zh@zju.edu.cn)

Shangchao Min
Zhejiang University, China (msc@zju.edu.cn)

Abstract

Speaking is often challenging for language learners to develop due to factors such as anxiety and limited practice opportunities. Dialogue-based computer-assisted language learning (CALL) systems have the potential to address these challenges. While there is evidence of their usefulness in second language (L2) learning, the effectiveness of these systems on speaking development remains unclear. The present meta-analysis attempts to provide a comprehensive overview of the effect of dialogue-based CALL in facilitating L2 speaking development. After an extensive literature search, we identified 16 studies encompassing 89 effect sizes. Through a three-level meta-analysis, we calculated the overall effect size and investigated the potential moderating effect of 13 variables spanning study context, study design and treatment, and measures. Results indicated a moderate overall effect size ($g = .61$) of dialogue systems on L2 learners' speaking development. Notably, three moderators were found to have significant effects: *type of system*, *system meaning constraint*, and *system modality*. No significant moderating effect was identified for *education stage*, *L2 proficiency*, *learning location*, *corrective feedback*, *length of intervention*, *type of interaction*, *measure*, and *key assessment component*. These findings suggest directions for future research, including the role of corrective feedback in dialogue-based CALL, the effectiveness of such systems across proficiency levels, and their potential in diverse learning contexts with the integration of generative artificial intelligence.

Keywords: dialogue systems; chatbots; L2 speaking; meta-analysis

1. Introduction

Dialogue systems are intelligent computer programs that can engage humans in natural conversational interactions. Usually empowered by automatic speech recognition (ASR) and natural language processing technologies, along with machine learning techniques, these systems provide learners with interaction opportunities varying from the most constrained tasks, such as read-aloud and elicited imitation, to more spontaneous and contextualized interactions, such as

Corresponding author: Shangchao Min; Email: msc@zju.edu.cn

‡This article was originally published without the corresponding author being identified. All versions of the article have since been updated and an correction notice published.

Cite this article: Hou, Z. & Min, S. (2025). Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis. *ReCALL* FirstView, 1–17. https://doi.org/10.1017/S0958344025100268

role-plays and paired discussions (e.g. Gokturk & Chukharev-Hudilainen, 2023). The advent of generative artificial intelligence (GenAI) starts a new era of dialogue systems that facilitate autonomous, multi-turn, and coherent conversations. Debates about GenAI applications in language learning, teaching, assessment, and policymaking have been raised (e.g. Voss *et al.*, 2023). We conducted the present study to offer a holistic portrayal of dialogue system applications[1] for advancing L2 speaking skills, using a multilevel quantitative meta-analysis. Our study builds upon the existing body of knowledge that can help the field prepare for the incoming substantial challenges in GenAI applications in educational contexts, especially for L2 speaking, an area that remains under-researched.

Using these intelligent systems for speaking practice offers noticeable advantages in both affective and cognitive domains. The low-stakes and non-judgmental practice environment with a virtual interlocutor helps alleviate L2 learners' speaking anxiety and enhances their willingness to communicate (Jeon, 2024; Kohnke, 2023; Shafiee Rad, 2024; Tai & Chen, 2023). Beyond the observed affective advantages, dialogue systems can also function as effective self-learning tools, providing real-time multimodal corrective feedback (Petersen, 2010; Tai, 2022), an authentic speaking environment (Hwang *et al.*, 2022), and adequate interactive exercises (Hsu, Chen & Yu, 2023). A well-trained dialogue system can also enable learners with increased exposure to high-quality oral input (see Gokturk & Chukharev-Hudilainen, 2023, for a typical dialogue system architecture).

Dialogue systems can impact the development of L2 speaking abilities. However, their effectiveness varies across different areas like system delivery mode, learning environments, and corrective feedback provision, indicating inconclusive research findings. For instance, while the use of a single modal voice-based or text-based chatbot for speaking showed advantages (e.g. Hsu, Chen & Yu, 2023; Kim, Kim & Cha, 2021; Tai, 2022), a superior multimodal presentation of the feedback to promote speaking proficiency was also found (e.g. Hwang *et al.*, 2022; Liu, Hwang & Su, 2024). Another example could be the findings on the use of dialogue systems under informal and formal learning contexts. While many studies suggested the effectiveness of using dialogue systems in an in-class instructional context (Dizon, 2020; Kim, Kim & Cha, 2021), research also found the special advantages of using dialogue systems as a valuable extension of in-class teaching for L2 speaking development (Liu, Hwang & Su, 2024; Tai, 2022). Additionally, as dialogue systems leverage different technologies that influence user speech production, the interaction tasks for speaking practice vary. It remains unclear which types of dialogue systems or interaction tasks are more conducive to achieving effective L2 speaking proficiency. These findings underscore the importance of conducting a comprehensive review to synthesize the existing literature and identify gaps in our understanding, enabling us to develop more effective strategies for using dialogue systems to enhance speaking proficiency.

Many narrative reviews have explored the application of dialogue systems in language learning. Ji, Han and Ko (2023) examined the collaboration between conversational AIs and teachers. Huang, Hew and Fryer (2022) synthesized chatbot affordances from technological, pedagogical, and social perspectives. Bibauw, François and Desmet (2019) discussed definitions and research trends of dialogue-based computer-assisted language learning (CALL) systems, while Litman, Strik and Lim (2018) reviewed speech technologies used for language assessment. Quantitative meta-analyses investigating the effectiveness of chatbots (Lee & Hwang, 2022; Zhang *et al.*, 2023), social robots (Lee & Lee, 2022), and dialogue systems (Bibauw, Van den Noortgate *et al.*, 2022) for general language learning and specifically ASR for L2 pronunciation (Ngo, Chen & Lai, 2024) can also be found. To the best of our knowledge, no meta-analysis focused on L2 speaking development. This paucity in the L2 speaking domain can be partially explained by the complexity

[1]The dialogue systems employed in the study pool of the analysis are mostly rule-based systems that are not powered and trained by GenAI.

of coordinating experiments using dialogue systems in learning contexts (Bibauw, François & Desmet, 2019). The diverse terminology and technology used for dialogue systems also increase the data collection and coding difficulty from primary studies. Regarding this issue, we adopted the umbrella term of dialogue-based CALL[2] as proposed by Bibauw, François and Desmet (2019), which refers to the learners' activities of using any system or application to engage in a dialogue with an automated interlocutor in an L2. This adoption encompasses all forms of dialogue systems that emphasize the interactive process with virtual agents, as the specific typology of dialogue systems is not the focal point of the present review.

A meta-analysis focused on L2 speaking in dialogue-based CALL is therefore essential. Unlike other language skills, speaking requires interactive engagement and immediate feedback, both of which are central to dialogue-based learning environments. By concentrating on L2 speaking within these systems, this study offers a comprehensive review to evaluate their effectiveness, identify trends, and highlight gaps, all of which can guide future research and pedagogical practices. The present study purports to depict a general research picture of dialogue-based CALL on L2 speaking development by synthesizing the overall effect size and related moderators affecting their effectiveness. To this end, we conducted a three-level meta-analysis addressing the following research questions (RQs):

RQ1. To what extent do dialogue-based CALL systems promote L2 speaking development?
RQ2. What are the significant moderator variables in using dialogue-based CALL for L2 speaking development?
RQ3. To what extent do these moderators affect the L2 speaking development?

2. Methodology

2.1 Literature search

To have a broad inclusion of eligible research, we conducted an extensive literature search for studies indexed in well-known online databases, including Scopus, Taylor & Francis Online, and Web of Science. Additionally, dissertations were searched in ProQuest and CNKI China. Due to our language repertoire, only research published in English and Chinese was included. Two search term sets regarding technology and L2 speaking were combined and applied. Informed by Bibauw, François and Desmet (2019) and Huang, Hew and Fryer's (2022) review on dialogue-based CALL, technology-related keywords were entered in the database as "chatbot*," "intelligent personal assistant*," "conversation* agent*," "spoken dialog* system*," "spoken dialog* technolog.*" Search terms for L2 speaking included "speech*, oral, conversation*, interaction*, talk*, speak*, and language*". A secondary search strategy was also applied by checking the research reference lists to include additional eligible research.

For all the search procedures, we filtered the search results in terms of research fields to computer and science, linguistics, education research, arts and humanities, and social science. The search period ended in May 2023.

2.2 Inclusion and exclusion criteria

We defined the inclusion and exclusion criteria from three aspects: technology used, target language proficiency, and research design. Table 1 provides detailed descriptions of the inclusion and exclusion criteria. For L2 speaking proficiency, a broad definition is adopted, including but not limited to conventional speaking proficiency in a psycholinguistic-individualist perspective (e.g. fluency, pronunciation, grammatical accuracy, lexical diversity) and sociolinguistic-

[2]These days, "conversational AI" is also an umbrella term popularized in the context of commercial applications/chatbots and by ChatGPT.

Table 1. Inclusion and exclusion criteria

Inclusion criteria: The study ...	Exclusion criteria: The study ...
1. used at least one dialogue system	1. only used portable devices that did not show interactions with an automated interlocutor
2. investigated L2 speaking proficiency	2. investigated L2 proficiency in non-speaking domains
3. provided required descriptive statistics to estimate the effect size of dialogue systems for L2 speaking development	3. failed to provide descriptive statistics to estimate the effect size of dialogue systems for L2 speaking development
4. employed (quasi-) experimental treatment and control groups or pre-test and post-test group designs	4. fell into case studies, qualitative research, or survey research

interactional perspective (e.g. interactional competence; IC) (Roever & Kasper, 2018). Ultimately, we obtained 16 studies for coding and analysis. Figure 1 illustrates the research search and inclusion in a PRISMA flowchart.

2.3 Coding scheme

We adopted a coding scheme from Plonsky and Oswald (2012) to explore the included research from three general perspectives: study context, design and treatment, and measure. Overall, 13 codes were finally investigated within these three categories. Specifically, for the dialogue systems coding, we adopted Bibauw, François and Desmet's (2019) typologies of systems, interactions, and degree of constraints on meaning. The detailed coding scheme can be found in supplementary material S1.

2.4 Inter-coder reliability

The overall coding process involved several coding cycles by two independent researchers. Initially, the primary inclusion of the effect sizes in the present analysis showed an ideal 94.16% agreement (Mackey & Gass, 2016). Subsequently, the effect size calculation was also checked with an acceptable 86.52% agreement. Any discrepancies arising from effect size inclusion and calculation were resolved through careful discussions between the coders, resulting in a final inclusion of 89 effect sizes for the subsequent coding and analysis. Furthermore, in coding the 13 moderator variables, the average Cohen's kappa coefficient was 0.93. This value falls within the "excellent" range of coding reliability (i.e. from 0.8 to 1), indicating a robust and dependable coding process (Mackey & Gass, 2016: 141).

2.5 Effect size calculation and interpretation

We calculated effect sizes based on the standardized mean difference. To achieve an unbiased estimate of the standardized mean difference in small sample analysis, we used Hedges's *g* (Hedges & Olkin, 2014) to remove the overestimation that tends to be observed in the estimate of Cohen's *d* (Cohen, 2013). To combine the effect sizes across study designs, we adopted Morris and DeShon's (2002) formulas to transform effect sizes into a comparable metric (see supplementary material S2 for details). We followed the field-specific guideline from Plonsky and Oswald (2014), with Cohen's *d* close to 0.40 as a small effect, 0.70 as a medium effect, and 1.0 as a large effect. This benchmark on Cohen's *d* is directly applied to Hedges's *g* since Hedges's *g* is simply an unbiased estimate that corrects the overestimation in small samples.

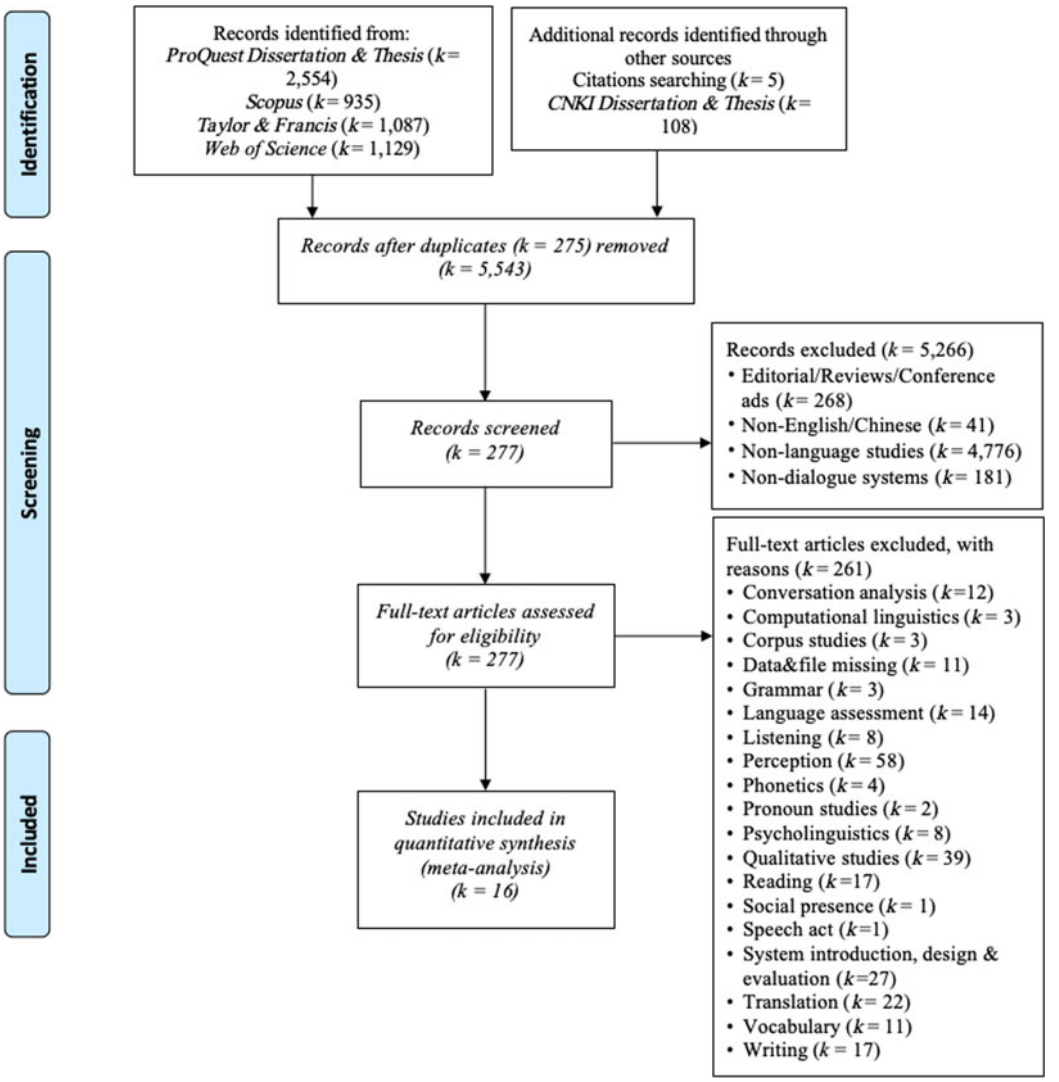

Figure 1. PRISMA flowchart of article search and selection (adapted from Page *et al*., 2021).

2.6 Data analysis using a three-level meta-analytical model

We utilized a multileveled (three-level) meta-analytical model, a robust but not widely applied method in L2 studies, to account for the non-independent effect size inclusion. Research in dialogue-based CALL on L2 speaking development tends to contribute to more than one effect size within the study, leading to a correlated relation among included effect sizes. This correlation (i.e. effect size dependence) threatens the validity of meta-analyses (Matt & Cook, 2009). Using a three-level meta-analytical model would allow effect size to vary among participants (level 1, sample variance), outcomes (level 2, within-study variance), and studies (level 3, between-study variance), suggesting more precise estimates (Assink & Wibbelink, 2016). This approach also allows for more effect size inclusion, which can increase the statistical power of the analysis. Additionally, as effect sizes are extracted from different outcome variables, more effect size inclusion provides opportunities to test more study characteristics (see Cheung, 2019, for further discussion).

Table 2. Overall effect size and results of heterogeneity tests at different levels

Effect size			95% CI		Heterogeneity						
k	*g*	*SE*	Lower	Upper	*Q*	*df*	*p*	τ^2_{level2}	I^2_{level2}	τ^2_{level3}	I^2_{level3}
89	0.61	0.14	0.34	0.89	329.75	88	<.0001	0.18	38.58%	0.20	42.45%

Note. CI = confidence interval.

We used R (R Core Team, 2018) for data analysis. The package metafor (Viechtbauer, 2010) was fitted to our data using its rma.mv function. This function can fit suitable meta-analytic multivariate/multilevel models to account for non-independence in the effects/outcomes. We fit a three-level random-effects model using this function, in which a restricted maximum likelihood estimation method (REML) was used for estimating the parameters in the model. This model accounts for the heterogeneity in effect sizes both within and between studies, accommodating the nested structure of multiple effect sizes per study. Detailed information on the codes, formulas, and a step-by-step guide on performing a three-level meta-analysis in R can be found in Assink and Wibbelink (2016) and Harrer *et al.* (2021). Four outliers in the model were detected and replaced by winsorizing along with a cutoff value at 2 standard deviations from the mean of all effect sizes ($g_{low} = -1.44$, $g_{up} = 2.98$) (Lipsey & Wilson, 2001: 108).

3. Results

3.1 Overall effect of the dialogue-based CALL systems on L2 speaking development

Table 2 provides the overall effect size and results of heterogeneity tests across levels. The estimated average effect was $g = .61$ (95% CI [0.34, 0.89]), indicating a significant medium effect of dialogue systems on L2 speaking development.

3.2 Heterogeneity and publication bias

As shown in Table 2, the result of the *Q* test was significant ($p < .0001$), suggesting significant variations in the outcomes of the primary studies and the need for moderator analyses. The estimates of variance components were $\tau^2_{level2} = 0.18$ and $\tau^2_{level3} = 0.20$, indicating that $I^2_{level2} = 38.58\%$ of the total variation can be attributed to within-study heterogeneity, whereas $I^2_{level3} = 42.45\%$ can be attributed to the between-study heterogeneity. In other words, more variations were observed at the between-study level of the model. Additionally, the three-level model provided a significantly better fit than the two-level model in which level 3 heterogeneity was constrained to zero ($\chi_1{}^2 = 14.54$; $p = .0001$).

For the multilevel meta-analytic model employed, quantifying the relationship between study size and effect size (publication bias) lacks appropriate tests (Assink & Wibbelink, 2016). Consequently, a contoured funnel plot was used to visually assess this association without statistical symmetry evaluation (Figure 2). The asymmetrical plot indicates potential publication bias, with missing studies in the lower left side of the funnel. However, sample sizes did not significantly moderate the effect ($b = 0.00$, 95% CI [−0.01, 0.02], $p = 0.6$), suggesting that larger studies would not produce more negative effect sizes than smaller ones. Therefore, the funnel plot's lack of strong negative effects likely reflects the true effect size distribution rather than publication bias. Tests using a traditional two-level meta-analytic model also indicate publication bias but with minimal impact on the effect (see supplementary material S3 for details).

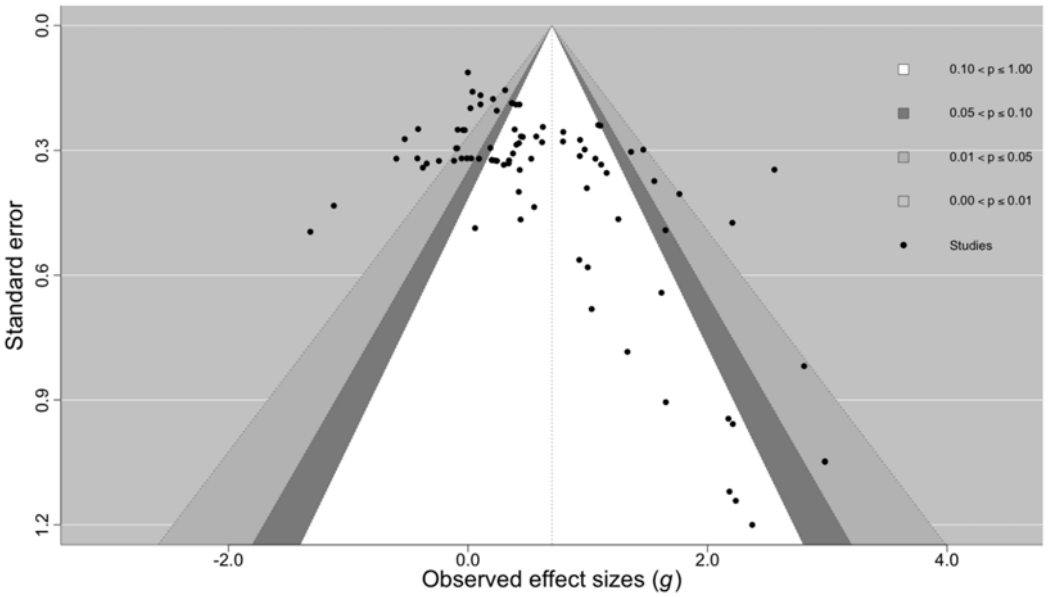

Figure 2. Contoured funnel plot of the standard error of Hedges's *g*.

Table 3. Moderator analyses in data from study context

Moderators	*df*	*Q*	*p*	Categories	*n*	*k*	*g* [95% CI]
Education stage	1	0.00	.98	K-12 education*	27	5	0.62 [0.14, 1.11]
				Higher education**	62	11	0.62 [0.25, 0.98]
L2 proficiency	2	1.07	.59	Beginner (A1/2)***	23	7	0.86 [0.45, 1.26]
				Intermediate (B1/2)**	23	5	0.59 [0.15, 1.02]
				Advanced (C1/2)	5	1	0.61 [0.15, 1.02]

Note. CI = confidence interval; *n* = number of effect sizes; *k* = number of studies; *g* = Hedges's *g*. *$p < .05$. **$p < .01$. ***$p < .001$.

3.3 Moderator analyses

Significant variations across the three levels necessitated moderator analyses for study context, study design and treatment, and measurement variables. A series of omnibus tests were used to address the difference of each variable across different subgroups. For categorical variables, we reported a *Q* test to suggest the possible significant effect of the moderators with their estimated Hedges's *g* value in each level. For continuous variables, specifically the three codes as duration in weeks and sessions and time per session, we reported the regression coefficient *b* to suggest the effect increased by an additional unit. Specifications of the included studies' features and system designs are provided in supplementary material S4.

3.3.1 Study context

Table 3 presents the moderator analyses results for study context variables. The dominant research focuses on targeted higher education ($k = 11$, 69%), with limited investigations in K-12 settings ($k = 5$, 31%). However, the dialogue systems' effect on L2 speaking development does not significantly differ between K-12 and higher education contexts, exhibiting a similar

Table 4. Moderator analyses in data from the design and treatment

Moderators	*df*	*Q*	*p*	Categories	*n*	*k*	*g* [95% CI]
Learning location	1	0.41	.52	In-classroom***	32	8	0.71 [0.31, 1.12]
				Out-of-classroom**	57	8	0.53 [0.14, 0.92]
Corrective feedback	1	0.08	.77	Presence**	49	7	0.57 [0.15, 1.00]
				Absence***	40	9	0.66 [0.28, 1.04]
Duration (weeks)	1	1.20	.27	+1 week (*b*)	89	16	0.03 [−0.02, 0.08]
Duration (sessions)	1	1.60	.21	+1 session (*b*)	89	16	0.02 [−0.00, 0.05]
Time per session	1	2.51	.11	+1 hour (*b*)	80	13	0.49 [−0.12, 1.09]
Types of interaction	2	4.67	.10	Task-oriented***	32	9	0.85 [0.47, 1.23]
				Open-ended*	45	6	0.46 [0.07, 0.84]
				System-guided	12	1	0.08 [−0.79, 0.96]
Types of system*	2	6.82	.03	Form-focused	12	1	0.08 [−0.73, 0.90]
				Goal-oriented***	28	6	1.04 [0.62, 1.46]
				Reactive**	49	9	0.43 [0.11, 0.75]
System modality*	2	8.44	.01	Text-based	15	2	0.07 [−0.46, 0.59]
				Voice-based*	32	7	0.36 [0.01, 0.72]
				Mixed***	42	9	0.93 [0.59, 1.27]

Note. CI = confidence interval; *n* = number of effect sizes; *k* = number of studies; *g* = Hedges's *g*. *$p < .05$. **$p < .01$. ***$p < .001$.

medium effect. Concerning L2 proficiency, dialogue systems demonstrate effects for beginner and intermediate learners, although proficiency level does not moderate the overall effect significantly. Notably, beginner learners appear to benefit more, with a large effect size, compared to a medium effect for intermediate learners. For advanced proficiency learners, the results failed to reach significance, potentially due to the small sample size from a single study ($n = 5$, $k = 1$).

3.3.2 Study design and treatment data

Table 4 reports the moderator analyses results for design and treatment data encompassing seven subgroups. Learning location does not significantly impact the results. Nevertheless, dialogue systems show effectiveness in both out-of-classroom and in-classroom learning contexts, with a noticeably large effect size during informal out-of-classroom learning scenarios. Similarly, no significant difference is observed in oral corrective feedback (CF) presence. The mean effectiveness of the system intervention reaches a medium to large level when CF is absent, while those incorporating CF demonstrate a medium effect on learners' L2 speaking development.

Looking at treatment duration coded as the overall duration in the number of weeks or sessions and task time per session in hours, neither reaches statistical significance. However, they appear to influence the speaking outcome, as their positive regression coefficients suggest. Additionally, time per session in hours seems to present a higher outcome than the other two. The type of interaction is not a differential moderator as well. Task-oriented interaction was predominantly implemented ($k = 9$, 56%), followed by open-ended interaction ($k = 6$, 38%). Both task-oriented and open-ended interactions exhibit a significant difference from the null effect, unlike system-guided interaction, probably due to the small sample size ($n = 1$, $k = 12$).

Table 5. Moderator analyses in data from measures

Moderators	*df*	*Q*	*p*	Categories	*n*	*k*	*g* [95% CI]
Measure	1	0.97	.32	Analytical measure**	35	7	0.52 [0.15, 0.89]
				Holistic measure***	44	15	0.73 [0.39, 1.06]
Speaking rating criteria	4	2.05	.73	Grammatical accuracy	8	5	0.43 [−0.12, 0.97]
				Fluency*	7	5	0.69 [0.12, 1.26]
				Pronunciation*	7	5	0.58 [0.02, 1.14]
				Task completion**	6	4	0.97 [0.30, 1.64]
				Vocabulary**	7	5	0.79 [0.23, 1.35]

Note. CI = confidence interval; *n* = number of effect sizes; *k* = number of studies; *g* = Hedges's *g*. *$p < .05$. **$p < .01$. ***$p < .001$.

Significant differences emerge across system design features. Narrative systems were omitted due to a lack of applications. Most research utilized reactive ($k = 9$, 56%) and goal-oriented systems ($k = 6$, 38%). Goal-oriented systems exhibit a large effect, followed by a medium effect for reactive systems. Form-focused systems provide a non-significant small effect, likely due to the limited number of studies analyzed. Regarding system meaning constraints, results mirrored the system type moderator unsurprisingly, as system coding partially relied on constraints on learners' production meaning and form (see supplementary material S1). To avoid redundancy, the meaning constraint moderator was omitted from the table.

Lastly, system modality significantly impacts overall effectiveness. Most systems employed a mixed multimodal interface ($k = 9$, 56%), followed by voice-based systems ($k = 7$, 44%). Learners benefited most from the mixed mode, exhibiting a large effect size. Voice-based systems, relying solely on sound recognition and production, provide a small effect. Text-based systems demonstrate the lowest effect among the three, although the difference was not statistically significant.

3.3.3 Measures

Table 5 shows the mean effect sizes of two outcome variables, including measure and speaking rating criteria. A dominant report in holistic proficiency is observed ($n = 44$, $k = 15$). No significant difference is found between the effect sizes observed in terms of measure. While speaking performance graded using the analytical scale shows a medium effect, holistic grading presents a large effect, both showing differences from the null effect. Regarding rating criteria, the result shows no significant difference across the five components. Although a large effect size was obtained in task completion, vocabulary, and fluency, dialogue systems show a medium effect on pronunciation ($g = .58$). Their effect on speaking grammatical accuracy remains underdetermined, as this domain failed to reach significance. Since we found only one study measuring IC (Kim, 2017), conducting a moderator analysis would be biased. Therefore, we omitted this variable.

4. Discussion

The present meta-analysis synthesized the results of 16 studies to assess the effectiveness of dialogue-based CALL systems in enhancing L2 speaking skills. Meanwhile, the analysis incorporated a few moderator variables and examined their effect on L2 speaking development. The following section answers the three research questions by discussing the overall effectiveness of dialogue systems for L2 speaking development and the moderators influencing their effects.

4.1 The effectiveness of dialogue-based CALL (RQ1)

In general, dialogue-based CALL exhibited a significantly positive and medium effect on L2 learners' speaking development. This finding is consistent with the previous meta-analyses but with a slightly larger effect (e.g. Bibauw, Van den Noortgate, *et al.*, 2022; Zhang *et al.*, 2023). The speaking gains observed are similar to those reported by Zhang *et al.* (2023) and Lee and Hwang (2022).

This effectiveness could be explained by several advantages of dialogue systems for speaking practice, including but not limited to their ability to (1) create continuing and meaningful interactive opportunities (Han, 2020; Hsu, Chen & Todd, 2023); (2) construct authentic speaking contexts (Hwang *et al.*, 2022); (3) provide multimodal feedback (Tai & Chen, 2022); and (4) engage L2 learners with a stress-free, interactive environment (Hsu, Chen & Yu, 2023; Tai, 2022). As the realm of AI continues to evolve, future research can further explore the ubiquity, interactivity, and authenticity afforded by dialogue systems for enhancing speaking practice.

However, the effectiveness of dialogue systems for L2 speaking development is tempered by several methodological limitations, particularly small sample sizes. Nearly all the studies ($k = 15$, 94%) involved fewer than 60 learners, with the largest sample being 73 (e.g. Kim, Kim, Cha, 2021). Moreover, checking participants' homogeneity before trials was often overlooked. While some studies employed analysis of covariance (ANCOVA) with speaking pre-test score as covariates to address group differences (e.g. Hsu, Chen & Todd, 2023; Hwang *et al.*, 2022; Yang, Lai & Chen, 2022), they did not examine the assumption of homogeneity of regression slope. Assumptions for parametric analysis were rarely reported, with normality being the only assumption addressed in just one study (e.g. Yang, Lai & Chen, 2022). This lack of methodological rigor may lead to inaccurate results. Furthermore, studies using a mixed-design method frequently neglected to incorporate pre-test and post-test within-group repeated measures ($k = 9$, 70%), potentially leading to biased conclusions regarding intervention effectiveness. This also poses challenges for meta-analysis, as calculating sampling variance for effect sizes often requires *t*-values from repeated measures. Future research should overcome these methodological limitations by using larger sample sizes, ensuring participant homogeneity, including pre-test and post-test repeated measures in mixed-design studies, and reporting parametric analysis assumptions.

4.2 Toward a comprehensive understanding of effective dialogue-based CALL for L2 speaking development (RQ2 and RQ3)

Compared to the established effectiveness of dialogue-based CALL for L2 speaking, more crucial and useful questions concern when, where, how, and for whom this effectiveness could be realized. The current study addresses these questions through the comprehensive exploration of multiple moderators in dialogue systems interventions. It is noteworthy that dialogue-based CALL remains a burgeoning area of research, as evidenced by the restricted sample sizes in the present and related review of research (e.g. Bibauw, François & Desmet, 2019; Zhang *et al.*, 2023). Therefore, instead of drawing definitive conclusions, we hope the following findings stimulate greater research attention and provoke further testing and analysis.

4.2.1 Study context

Regarding educational stages, all stages showed a shared medium effect, albeit no significant difference was observed in this moderator. This finding aligns with previous findings (e.g. Bibauw, Van den Noortgate, *et al.*, 2022; Zhang *et al.*, 2023), suggesting that dialogue systems benefit students across levels of education. While prior studies found an advantage for younger learners (e.g. Zhang *et al.*, 2023), the present analysis had a limited representation of K-12 learners ($n = 5$, 31%), indicating insufficient empirical evidence. Therefore, it is still early to conclude that dialogue-based CALL favors L2 learners at specific education stages. Further investigations are

warranted to explore the applications of dialogue systems in relevant teaching contexts and potentially uncover differential effects across educational levels.

L2 proficiency is also a non-significant moderator, while lower (A1/A2) and intermediate (B1/B2) proficiency learners demonstrated certain learning gains. This finding may be ascribed to the social and psychological support within the dialogue-based CALL environment for less proficient L2 learners. Dialogue systems can offer multimodal feedback, enhancing comprehension and consequently production for less proficient learners (Tai, 2022). This advantage for less proficient L2 learners also coincides with Bibauw, Van den Noortgate, *et al.* (2022), wherein a special effect of dialogue-based CALL in the consolidation stages of learning was hypothesized. For advanced L2 proficiency learners, dialogue systems appeared less effective (e.g. Kim, 2016; Tai, 2022), although the limited sample size precluded a significant effect. Tai (2022) posited that ASR technology sacrifices sentence length and complexity to maintain high recognition rates, suggesting less challenging interaction tasks for proficient L2 speakers. In contrast, Hsu, Chen and Todd (2023) observed better interaction experiences for advanced learners due to fluent conversation flow and fewer communication breakdowns resulting from adequate L2 proficiency. While supporting Bibauw, Van den Noortgate, *et al.*'s (2022) hypothesis regarding dialogue-based CALL's effect in the early consolidation stage for less proficient learners, our findings indicate varying experiences when interacting with virtual interlocutors across proficiency levels. We call for research investigating dialogue-based CALL targeting different participant populations, particularly considering potential communication breakdowns across proficiency groups.

4.2.2 Study design and treatment

Dialogue systems have shown effectiveness in both in-classroom and out-of-classroom settings for speaking practice. Integrating them with mobile devices allows dialogue-based CALL to enjoy the mobility, ubiquity, and flexibility of mobile-assisted language learning (MALL), characterized as anytime, anywhere learning (Kukulska-Hulme & Shield, 2008). This facilitates authentically contextual conversation by connecting knowledge with learners' surroundings, using language in meaningful contexts, and stimulating learners' interests (Hsu, Chen & Todd, 2023; Hwang *et al.*, 2022; Tai, 2022). For instance, Tai (2022) encouraged out-of-classroom dialogue-based CALL activities as meaningful extensions of classroom learning. This integration raises the important issue of the teacher's role in dialogue-based CALL. We agree with Ji, Han and Ko (2023) that collaboration between teachers and machines (i.e. dialogue systems) is pivotal for successful AI-integrated language learning. Dialogue systems could help teachers to better allocate teaching resources in combination with classroom-based interactive practices (Tai, 2022). Teachers can also help to maintain learners' learning interests, especially when the novelty effect wears off (El Shazly, 2021). Compared with the rich explorations of dialogue systems' role in helping learners with L2 learning (e.g. Kohnke, 2023; Tai & Chen, 2023), limited research investigates how language teachers can guide students during dialogue-based CALL. Stronger orchestration between teachers and technology would be required in future classrooms (Roschelle, Lester & Fusco, 2020). More studies, especially empirical research, are needed to explore the teachers' participation in dialogue-based CALL, from course design and practical teaching to class management and language assessment.

Second, the presence or absence of CF did not make a significant difference in L2 speaking practice. Notably, the effect sizes in our study indicate much more proximity between the two conditions than what was reported in Bibauw, Van den Noortgate, *et al.* (2022), with more nuanced classifications of CF. This suggests that while we did not find a significant difference, there may still be potential benefits of CF that are not fully captured by our results. Therefore, this finding might not contradict the literature on CF's benefits in L2 learning, particularly in traditional classrooms (e.g. Lyster, Saito & Sato, 2013; Nassaji & Kartchava, 2021) or in technology-enhanced learning environments for pronunciation and speech fluency (Gu *et al.*,

2021; Ngo, Chen & Lai, 2024). The non-significant difference may stem from CF's heterogeneous nature, characterized by disparate techniques, objectives, and instructional contexts across the studies. In dialogue-based CALL, providing CF faces challenges, as it risks disrupting learner interaction and willingness to communicate (Hwang *et al*., 2022). System feedback in the form of silence or erroneous responses can prompt immediate self-correction among learners, particularly in pronunciation. In evaluating feedback of this nature, discerning its corrective intent is challenging, as these potential implicit instances of CF may be present but remain unreported. Under this circumstance, we applied a dichotomous coding to suggest the distinction between studies with and without a clear report of corrective notifications (e.g. incorrect pronunciation notifications in Hsu, Chen & Yu, 2023; Tai, 2022) or CF moves during interaction (e.g. recast in Petersen, 2010). This yes or no coding might have affected the finding. Given that dialogue-based CALL for speaking is still evolving, the specific use of CF across different intelligent dialogue systems warrants further empirical investigations. Discussions should be made upon the context-specific CF to explore its unique contributions to L2 speaking development under dialogue-based CALL, especially employing GenAI-based systems for CF delivery.

Third, although not achieving significance, the intervention duration seems to affect the overall effectiveness of dialogue-based CALL for L2 speaking development, particularly with longer individual sessions. Studies in this domain often omit precise time-on-task data in favor of reporting overall session duration, leading to ambiguity and inconsistency in defining intervention length and frequency. This tendency may potentially contribute to the non-significant findings. Interestingly, while increasing the number of weeks or sessions shows relatively small effects, longer individual sessions seem to have a more pronounced impact. Bibauw, Van den Noortgate, *et al*. (2022) reported higher learning outcomes for dialogue-based CALL studies using a packed practice. Together with our findings, these might imply that both frequency and depth of use matter for effective learning. When learners use the system frequently and each session is long enough for meaningful engagement, the combined effect may produce the best outcomes. Additionally, our findings could also indicate a novelty effect, where learners initially engage more deeply with the system but experience diminishing returns as they grow familiar with it over time. Given that the impact of intervention duration remains unclear, this potential novelty effect, observed in other technology-enhanced learning environments like MALL (e.g. Tseng *et al*., 2022), warrants further investigations. While researchers should strive for greater control and clarity in reporting intervention duration, future studies should examine how both duration and frequency affect dialogue systems for L2 speaking development.

Fourth, our analysis reveals a moderating effect from diverse system designs and meaning constraints of learners' production, while interaction type does not differentially impact L2 speaking development. This established impact suggests that systems possess distinctive instructional and interactional values for L2 speaking development. Goal-oriented systems show advantages for L2 speaking development, backing Bibauw, François and Desmet's (2022) claim that form-focused and goal-oriented systems offer the most promising affordances for language learning, while the impact of form-focused systems remains uncertain due to limited studies involved. Goal-oriented systems emphasize implicit meaning constraints in learner production, prompting learners to engage in dialogic interaction to achieve a specific goal. Unlike open-ended free dialogue, their interactional value emphasizes collaborative activity to accomplish a task, known as task-oriented interactions. Tasks for speaking development vary widely from everyday transactions like travel (Park, 2022) and daily life (Hwang *et al*., 2022) to exam-oriented tasks (Hsu, Chen & Yu, 2023). Learners with these tasks and systems have full interactivity and high user initiative toward predetermined learning goals. Compared to open-ended interactions facilitated by reactive systems, it appears that dialogue tasks for L2 speaking development are better set within a specific context or domain rather than being left entirely to user discretion in unrestricted communications.

The effectiveness of contextualized interaction in goal-oriented systems can also explain the non-significant impact established for interaction type. Some studies in the analysis employed reactive systems for task-oriented interaction activities, predominantly oriented around speech topics (e.g. Dizon, 2020; Tai & Chen, 2022; Yang, Lai & Chen, 2022). These tasks demonstrated more open-ended interactions in nature, especially conducted with intelligent personal assistants (i.e. reactive systems such as Google Assistant). Reactive systems operate solely in response to prompts or questions, providing limited contextual interaction. In contrast, goal-oriented systems incorporate tasks with diverse technological affordances such as CF (Hsu, Chen & Yu, 2023), virtual reality (VR) learning environments (Park, 2022), and a blend of controlled and free-speaking practices (Hwang *et al*., 2022). To effectively enhance L2 speaking, it is advisable to use goal-oriented systems for task-oriented interactions that guide the student through the steps required to accomplish tasks. Additionally, our findings diverge from Bibauw, Van den Noortgate, *et al*.'s (2022) view that learners benefit the most from system-guided interactions and form-focused systems, where systems guide learners through predetermined activities. This contrast also underscores the unique interactional and instructional demands of dialogue systems for L2 speaking practice, which is technologically more challenging to develop. With advanced complex dialogue management techniques, future research can further explore the design or application of different systems varying in user control and interactivity levels.

Lastly, the pivotal role of system modality emerged as a noteworthy moderator. Dialogue systems with mixed system modality integrating a diverse spectrum, encompassing voice, text, and additional facets like VR, yielded a large and significant effect. This finding fits the found modality impact of dialogue systems for L2 speaking in Tai and Chen (2022). While learners prefer voice chatting over text chatting (Kim, Kim & Cha, 2021), a mixed written and spoken interface can increase the intelligibility of the interaction and thus facilitate optimal communication. For less proficient L2 learners, relying solely on auditory feedback from the system may lead to processing and retrieval difficulties, particularly in cases of miscommunication (Tai & Chen, 2022). Visual support in feedback, provided through screen displays or VR equipment, enables learners to pinpoint sources of miscommunication, thereby promoting self-directed learning and correction. Furthermore, the multimodal feedback presentation can motivate learners to explore unknown information and enhance processing and comprehension (Tai & Chen, 2022). Similar significant effects have been reported for mixed interaction modes in studies by Zhang *et al*. (2023) and Lee and Hwang (2022). With more advanced dialogue systems, future research for L2 speaking practice should consider having a multimodal interface plugging various visual and auditory modes.

4.2.3 Measures

Overall, no significant difference is observed in speaking gains measured using holistic or analytical scales in dialogue-based CALL for L2 speaking development, although both indicate some effect. The prevalent use of holistic proficiency scales provides limited insights into specific areas like pronunciation, grammar, and vocabulary. Future studies should explore dialogue systems' effect on specific aspects of speaking, utilizing more informative inquiries, including linguistic features (e.g. speech fluency, lexical diversity, or syntactic complexity) and interactional patterns (e.g. discourse markers, conversational repair strategies, and IC). Concerning the limited report on specific aspects of speaking, dialogue systems seem to improve fluency, pronunciation, task completion, and vocabulary, but not grammatical accuracy. Tai (2022) attributed the limited effectiveness of IPA-mediated interaction in grammar to technological constraints, particularly ASR's struggle to accurately recognize longer sentences. Consequently, participants often use simpler grammatical structures to sustain conversational fluency, in which learners focus on meaning and fluency during free practice with a native-like virtual interlocutor. However, this finding contradicts Hwang *et al.* (2022), where free talk with a chatbot improved grammatical

accuracy and no established relation was found in controlled talks practicing predetermined sentence structures. Given the insufficient empirical evidence, it is premature to draw firm conclusions about dialogue-based CALL's effect on specific speaking areas. Nevertheless, dialogue systems seem to effectively enhance L2 learners' vocabulary, speech fluency, pronunciation, and task completion. Its impact on grammatical accuracy remains to be established.

It is important to address language proficiency coding, particularly with the psycholinguistic-individualist and IC domains. Notably, only Kim (2017) assessing speaking proficiency in the negotiation of meaning is related to IC. However, considering the non-equivalence of L2 speaking proficiency between conventional psycholinguistic-individualist and IC domains (Roever & Ikeda, 2022), it is worth exploring the impact of dialogue-based CALL on learners' IC. Moreover, as highlighted earlier, there is a noticeable research gap concerning the scarcity of studies targeting different proficiency levels. Specifically for studies of advanced learners, utilizing intelligent dialogue systems to develop and assess their IC emerges as a promising avenue.

5. Conclusion

This meta-analysis aims to present a general picture of the effect of dialogue-based CALL on L2 speaking development. After a stringent research search and inclusion process, we identified 16 eligible studies. Results showed a moderate effect ($g = .61$) of dialogue systems for L2 speaking development. Three significant moderators were found: types of systems, the meaning constraint of learner production, and system output modalities that can moderate the effect of dialogue systems for speaking. Learners benefit more when they use goal-oriented systems, stressing the implicit meaning constraints of learner productions. Regarding system modalities, mixed modalities are the most effective, highlighting the need to integrate visual and audio modes.

The present analysis also brings several implications that highlight potential directions for future research. First, given that providing immediate, continuous feedback is one of the key features of present intelligent dialogue systems, it is important to discern the appropriate typology of CF within dialogue systems to ascertain its effectiveness. Second, future research can target the unique affordances of dialogue systems upon learners across proficiency levels. Third, although it did not reach statistical significance, the contrast between in-classroom and out-of-classroom prompts further investigations of dialogue-based CALL within the context of MALL to uncover its adaptivity and mobility. Additionally, understanding the role of teachers in both formal and informal learning contexts is paramount. Collaborative efforts between dialogue systems and language teachers warrant exploration, encompassing aspects like course design, practical teaching, classroom management, and language assessment. Fourth, given the limitation of time control in the field, there is a clear need for further investigation into the effects of intervention duration and frequency of dialogue system applications on L2 speaking development. Lastly, the established effectiveness of goal-oriented systems suggests a future research agenda to develop task-based speaking activities simulating real-life situations. Cross-disciplinary collaborations are encouraged to leverage advanced dialogue manager modules empowered by GenAI for highly contextualized interactions. To conclude, while the current analysis offers insights, the limited number of studies underscores that the use of dialogue systems for L2 speaking development remains a nascent field. With advancements in GenAI-powered dialogue systems, it is imperative to advocate for further research into the potential of dialogue-based CALL for enhancing speaking proficiency.

This study is not without limitations. The analysis falls short of representing a global spectrum of dialogue-based CALL systems for speaking proficiency. The scope of this study was confined by the language proficiency of the researchers, encompassing solely empirical studies published in English and Chinese. Furthermore, the limited number of studies included in the current meta-

analysis also underscores the lack of conclusive evidence and signifies the preliminary phase of this research domain, thus limiting the strength of the effects observed. Due to this limited number of included studies, the potential publication bias that might be indirectly observed in the moderator analysis should also be noted. Additionally, this study only investigated publications from journals, indicating potential incomplete representation of the field. Future research could also consider other publication sources, such as conference proceedings and book chapters. Lastly, the missed search term "robot" also implies incomplete coverage of the field, given that contemporary educational robots often integrate dialogue systems to facilitate human-like interactions.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344025100268

Data availability statement. Data available on request from the authors.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful feedback on earlier drafts. Special thanks goes to Professor Serge Bibauw for sharing the R codes that informed our analysis, and to Jili Shen and Hui Wang for their support with the coding. Any remaining limitations are our own.

Authorship contribution statement. Zhuohan Hou: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. Shanghchao Min: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding disclosure statement. This research was supported by the Zhejiang Provincial Planning Office of Philosophy and Social Science [22ZJQN16YB] and Zhejiang Provincial Graduates' Science and Technology Innovation Program [2023R401174].

Competing interests statement. The authors declare no competing interests.

Ethical statement. Ethical approval was not required.

GenAI use disclosure statement. ChatGPT (Version 4) was used to revise some of the sentences for clarity.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Assink, M. & Wibbelink, C. J. M. (2016) Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, 12(3): 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Bibauw, S., François, T. & Desmet, P. (2019) Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, 32(8): 827–877. https://doi.org/10.1080/09588221.2018.1535508

Bibauw, S., François, T. & Desmet, P. (2022) Dialogue systems for language learning: Chatbots and beyond. In Ziegler, N. & González-Lloret, M. (eds.), *The Routledge handbook of second language acquisition and technology*. New York: Routledge, 121–134. https://doi.org/10.4324/9781351117586-12

Bibauw, S., Van den Noortgate, W., François, T. & Desmet, P. (2022) Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, 26(1): 1–24. https://hdl.handle.net/10125/73488

Cheung, M. W.-L. (2019) A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, 29(4): 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cohen, J. (2013) *Statistical power analysis for the behavioral sciences*. New York: Academic Press. https://doi.org/10.4324/9780203771587

*Dizon, G. (2020) Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, 24(1): 16–26. http://hdl.handle.net/10125/44705

*El Shazly, R. (2021) Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems*, 38(3): Article e12667. https://doi.org/10.1111/exsy.12667

Gokturk, N. & Chukharev-Hudilainen, E. (2023) Strategy use in a spoken dialog system–delivered paired discussion task: A stimulated recall study. *Language Testing*, 40(3): 630–657. https://doi.org/10.1177/02655322231152620

Gu, L., Davis, L., Tao, J. & Zechner, K. (2021) Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study. *Assessment in Education: Principles, Policy & Practice*, 28(1): 58–76. https://doi.org/10.1080/0969594X.2020.1735995

*Han, D.-E. (2020) The effects of voice-based AI chatbots on Korean EFL middle school students' speaking competence and affective domains. *Asia-Pacific Journal of Convergent Research Interchange*, 6(7): 71–80. https://doi.org/10.47116/apjcri.2020.07.07

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). Doing meta-analysis with R: A hands-on guide (1st ed). Chapman & Hall/CRC Press. https://doi.org/10.1201/9781003107347.

Hedges, L. V. & Olkin, I. (2014) *Statistical methods for meta-analysis*. Orlando: Academic Press.

*Hsu, H.-L., Chen, H. H.-J. & Todd, A. G. (2023) Investigating the impact of the Amazon Alexa on the development of L2 listening and speaking skills. *Interactive Learning Environments*, 31(9): 5732–5745. https://doi.org/10.1080/10494820.2021.2016864

*Hsu, M.-H., Chen, P.-S. & Yu, C.-S. (2023) Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments*, 31(7): 4297–4308. https://doi.org/10.1080/10494820.2021.1960864

Huang, W., Hew, K. F. & Fryer, L. K. (2022) Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1): 237–257. https://doi.org/10.1111/jcal.12610

*Hwang, W.-Y., Guo, B.-C., Hoang, A. & Chang, C.-C. (2022) Facilitating authentic contextual EFL speaking and conversation with smart mechanisms and investigating its influence on learning achievements. *Computer Assisted Language Learning*, 37(7): 1632–1658. https://doi.org/10.1080/09588221.2022.2095406

Jeon, J. (2024) Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, 37(1–2): 1–26. https://doi.org/10.1080/09588221.2021.2021241

Ji, H., Han, I. & Ko, Y. (2023) A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1): 48–63. https://doi.org/10.1080/15391523.2022.2142873

*Kim, H.-S., Kim, N. Y. & Cha, Y. (2021) Is it beneficial to use AI chatbots to improve learners' speaking performance? *The Journal of Asia TEFL*, 18(1): 161–178. https://doi.org/10.18823/asiatefl.2021.18.1.10.161

*Kim, N.-Y. (2016) Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning*, 19(4): 63–88. https://doi.org/10.15702/mall.2016.19.4.63

*Kim, N.-Y. (2017) Effects of types of voice-based chat on EFL students' negotiation of meaning according to proficiency levels. *English Teaching*, 72(1): 159–181. https://doi.org/10.15858/engtea.72.1.201703.159

Kohnke, L. (2023) L2 learners' perceptions of a chatbot as a potential independent language learning tool. *International Journal of Mobile Learning and Organisation*, 17(1–2): 214–226. https://doi.org/10.1504/IJMLO.2023.128339

Kukulska-Hulme, A. & Shield, L. (2008) An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, 20(3): 271–289. https://doi.org/10.1017/S0958344008000335

Lee, H. & Lee, J. H. (2022) The effects of robot-assisted language learning: A meta-analysis. *Educational Research Review*, 35: Article 100425. https://doi.org/10.1016/j.edurev.2021.100425

Lee, J.-Y. & Hwang, Y. (2022) A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language & Literature*, 48(1): 213–243. https://doi.org/10.21559/aellk.2022.48.1.011

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

Litman, D., Strik, H. & Lim, G. S. (2018) Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3): 294–309. https://doi.org/10.1080/15434303.2018.1472265

Liu, Y.-F., Hwang, W.-Y. & Su, C.-H. (2024) Investigating the impact of context-awareness smart learning mechanism on EFL conversation learning. *Interactive Learning Environments*, 32(8): 4122–4137. https://doi.org/10.1080/10494820.2023.2194931

Lyster, R., Saito, K. & Sato, M. (2013) Oral corrective feedback in second language classrooms. *Language Teaching*, 46(1): 1–40. https://doi.org/10.1017/S0261444812000365

Mackey, A. & Gass, S. M. (2016) *Second language research: Methodology and design* (2nd ed.). New York: Routledge.

Matt, G. E. & Cook, T. D. (2009) Threats to the validity of generalized inferences. In Cooper, H., Hedges, L. V. & Valentine, J. C. (eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation, 537–560.

Morris, S. B. & DeShon, R. P. (2002) Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1): 105–125. https://doi.org/10.1037/1082-989x.7.1.105

Nassaji, H. & Kartchava, E. (2021) *The Cambridge handbook of corrective feedback in second language learning and teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108589789

Ngo, T. T.-N., Chen, H. H.-J. & Lai, K. K.-W. (2024) The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL*, 36(1): 4–21. https://doi.org/10.1017/S0958344023000113

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … & Moher, D. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372(71): 1–9. https://doi.org/10.1136/bmj.n71

*Park, H. (2022) Effects of virtual reality-based English learning on Korean university students' speaking ability. *Multimedia-Assisted Language Learning*, 25(4): 93–119 https://doi.org/10.15702/mall.2022.25.4.93
*Petersen, K. A. (2010) *Implicit corrective feedback in computer-guided interaction: Does mode matter?* Georgetown University, doctoral dissertation.
Plonsky, L. & Oswald, F. L. (2012) How to do a meta-analysis. In Mackey, A. & Gass, S. M. (eds.), *Research methods in second language acquisition: A practical guide*. Chichester: Blackwell Publishing, 275–295. https://doi.org/10.1002/9781444347340.ch14
Plonsky, L. & Oswald, F. L. (2014) How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4): 878–912. https://doi.org/10.1111/lang.12079
R Core Team (2018) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/
Roever, C. & Ikeda, N. (2022) What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, 39(1): 7–29. https://doi.org/10.1177/02655322211003332
Roever, C. & Kasper, G. (2018) Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, 35(3): 331–355. https://doi.org/10.1177/0265532218758128
Roschelle, J., Lester, J. & Fusco, J. (2020) *AI and the future of learning: Expert panel report*. Washington: Digital Promise. https://doi.org/10.51388/20.500.12265/106
Shafiee Rad, H. (2024) Revolutionizing L2 speaking proficiency, willingness to communicate, and perceptions through artificial intelligence: A case of Speeko application. *Innovation in Language Learning and Teaching*, 18(4): 364–379. https://doi.org/10.1080/17501229.2024.2309539
*Tai, T.-Y. (2022) Effects of intelligent personal assistants on EFL learners' oral proficiency outside the classroom. *Computer Assisted Language Learning*, 37(5–6): 1281–1310. https://doi.org/10.1080/09588221.2022.2075013
*Tai, T.-Y. & Chen, H. H.-J. (2022) The impact of intelligent personal assistants on adolescent EFL learners' speaking proficiency. *Computer Assisted Language Learning*, 37(5–6): 1224–1251. https://doi.org/10.1080/09588221.2022.2070219
Tai, T.-Y. & Chen, H. H.-J. (2023) The impact of Google Assistant on adolescent EFL learners' willingness to communicate. *Interactive Learning Environments*, 31(3): 1485–1502. https://doi.org/10.1080/10494820.2020.1841801
Tseng, W.-T., Chen, S., Wang, S.-P., Cheng, H.-F., Yang, P.-S. & Gao, X. A. (2022) The effects of MALL on L2 pronunciation learning: A meta-analysis. *Journal of Educational Computing Research*, 60(5): 1220–1252. https://doi.org/10.1177/07356331211058662
Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3): 1–48. https://doi.org/10.18637/jss.v036.i03
Voss, E., Cushing, S. T., Ockey, G. J. & Yan, X. (2023) The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly*, 20(4–5): 520–532. https://doi.org/10.1080/15434303.2023.2288256
*Yang, C. T.-Y., Lai, S.-L. & Chen, H. H.-J. (2022) The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking. *Interactive Learning Environments*, 32(5): 2175–2195. https://doi.org/10.1080/10494820.2022.2141266
Zhang, S., Shan, C., Lee, J. S. Y., Che, S. & Kim, J. H. (2023) Effect of chatbot-assisted language learning: A meta-analysis. *Education and Information Technologies*, 28(11): 15223–15243. https://doi.org/10.1007/s10639-023-11805-6

About the authors

Zhuohan Hou is currently a PhD candidate in Applied Linguistics at the School of International Studies, Zhejiang University, Hangzhou, China. Her research interests include speaking and listening assessment, computer-assisted language learning, and quantitative research methods. Her work has been published in peer-reviewed journals such as *System*.

Shangchao Min is a professor of Applied Linguistics at the School of International Studies, Zhejiang University, Hangzhou, China. Her research interests include language testing and assessment, educational measurement, and second language acquisition. She serves on the editorial boards of *Language Testing* and *Language Assessment Quarterly*. She has undertaken several research projects funded by the Ministry of Education and the National Social Science Foundation of China as a principal investigator.