Junhong Xiao* and Yiwei Peng

Critiquing research published in SSCI and ESCI CALL journals between 2023 and 2024: a research design perspective

https://doi.org/10.1515/jccall-2025-0015
Received April 24, 2025; accepted June 7, 2025; published online July 7, 2025

Abstract: CALL publications are not short of systematic literature reviews. The majority of these reviews aim to identify publication trends, research themes, types of technology used, factors influencing technology adoption, task/learning activity design, effectiveness of CALL, theoretical underpinning and so on. No previous publication has focused exclusively on CALL research from the perspective of research design although appropriate design is the first step in ensuring the quality of research outcomes. This study reports on the findings of a research design-focused review of 338 full articles published in five SSCI and ESCI CALL journals between 2023 and 2024. Strengths identified include clear statements of context of study and research purpose, question and/or hypothesis; increasing adoption of mixed-method approach; popularity of experimental research; sound theoretical foundation; multiple data sources/triangulation; specific accounts of data collection and analysis; and acknowledgement of limitations. Aspects which are found to be weak include predominance of micro-level research; scarcity of replication studies; insensitivity to researcher biases; overwhelming dominance of non-probabilistic, convenience sampling; shortage of longitudinal research; undue reliance on self-reported data and unverified and/or unavailable instruments; insufficient attention to ethical issues; over-favoring case study and survey research; and use of ambiguous units of measurement. Implications from the findings are discussed. Broader issues beyond the findings are raised which warrant more attention. By highlighting the strengths and weaknesses of CALL journal publications in terms of research design rather than appraising individual studies, it is hoped that researchers and gatekeepers (editors/reviewers) alike will pay more attention to the design of a study.

Keywords: research design; empirical study; computer assisted language learning (CALL); systematic literature review; research robustness

***Corresponding author: Junhong Xiao**, The Open University of Shantou, 8 Leshan Road, Shantou, Guangdong, 515000, China, E-mail: frankxjh@outlook.com. https://orcid.org/0000-0002-5316-2957
Yiwei Peng, The Open University of Shantou, 8 Leshan Road, Shantou, Guangdong, 515000, China, E-mail: diana66912@yahoo.com. https://orcid.org/0000-0001-5341-2568

1 Introduction

In the broader field of applied linguistics, computer assisted language learning (CALL) is one of the most vibrant subfields, albeit not the most long-established, with an abundance of research publications. Nevertheless, as aptly observed by the guest editors of the *Applied Linguistics Review* (2024, Volume 15 Issue 4) Special Issue: *Research Synthesis in Language Learning and Teaching*, there exists "the research-practice chasm" which refers to the phenomenon that "educational research evidence is rarely used by teachers to inform their practice" (Chong et al. 2024, p. 1564). These editors also cited reasons given by other researchers, including mutual distrust between researchers and teachers, failure to engage teachers in setting research agendas, inaccessibility of research publications to teachers, and teachers' shortage of time to read CALL publications, among other things. Research quality is also a concern. "Practitioners or policy makers should not be blamed for not applying research results" according to Xiao (2023a) who maintains that "if research results are conducive to practice, the temptation to draw on research will overcome resistance to changes" (p. 215). A major threat to research quality is lack of rigor in terms of research design (Peng and Xiao 2022; van Drimmelen et al. 2024). Unfortunately, it happens that important research design details may be missing as reported in the systematic review of meta-analyses in second language research by Vuogan and Li (2024). Therefore, a comprehensive review of the design of CALL research may contribute to the sustainable development of the field.

CALL publications are not short of systematic literature reviews, including meta-reviews/research syntheses. As can be seen in Section 2.1 below, literature reviews constitute over 10 % of the publications in five CALL journals listed in Social Science Citation Index (SSCI) or Emerging Sources Citation Index (ESCI) (see Table 2) between 2023 and 2024. The foci of published literature reviews are many and varied. For example, Xiao et al. (2024) reviews the main ways in which conversational agents are used in second language learning while Zhai and Wibowo (2023) set out to identify factors that influence the use of artificial intelligence (AI) dialogue systems for learning English as a foreign language (EFL) and existing gaps as reported in the reviewed studies (also see DeMolder et al. 2023). Kim and Namkung (2024) focus on methodological features, research themes, and types of technology used, and on top of these Zheng et al. (2022) attempts to reveal general publication trends (also see DeMolder et al. 2023) as well as benefits and challenges brought about by virtual reality to language education. Akayoğlu (2019) seeks to identify prominent theoretical frameworks underpinning CALL publications in *Computer Assisted Language Learning*, *System*, *British Journal of Educational Technology*, and *Language Learning and Technology* between 1997 and 2018 while Li et al. (2024) "aims to (i) suggest the

roles that teachers play in the AI chatbots (including ChatGPT)-assisted language learning process, (ii) discuss how those roles satisfy the SDT needs of the students, and (iii) discuss the challenges faced by teachers in this learning" (pp. 2–3). In contrast, Shortt et al. (2023) centers around publication output by year, authorship nationalities, target languages, underlying theoretical frameworks, research approaches, types of research questions, sampling and participants as well as data collection and analysis (also see Zhang et al. 2023). Finally, Mohsen et al. (2024) presents a comprehensive overview of the CALL research landscape in their scientometric study of hotspot research and trending issues in the past 42 years' CALL research and findings from previous CALL syntheses.

Despite the significant presence of literature reviews in the CALL research landscape, it seems that no previous publication has focused exclusively on the robustness of CALL studies from the perspective of research design. There are reviews which include certain aspects of the design of a CALL study (for example, Burston et al. 2024; Kim and Namkung 2024; Shortt et al. 2023; Xiao et al. 2024; Zheng et al. 2022, 2023). Nevertheless, no study is totally dedicated to this theme, i.e. covering all aspects – from the context in which a study is situated and the selection of research area which can be categorized into issues at macro, meso, and micro levels to how the researcher manages his/her potential bias, handles ethical issues, and reflects on the limitation(s) of the study. In other words, in this article, by research design, we mean things that should or can be planned and deliberated before a study is conducted.

Without doubt, it is important to identify research themes and trends, types of technology used and main uses of a technology, factors influencing the application of a technology and the extent to which it is aligned with the underlying principles of a (language) learning theory, task designs and learning activities, student engagement and learning outcomes as well as theoretical frameworks adopted. Equally important are findings from bibliometric analyses. All these efforts can move the field forward. Nonetheless, it is no less relevant to evaluate the rigorousness and trustworthiness of the research design followed by CALL studies.

Appropriate design is the first step in ensuring research quality. The validity, reliability, and even generalizability of the findings of a study depends, to a considerable extent, on rigorous design and appropriate administration. In reality, this appropriateness and rigorousness is often taken for granted. For instance, Anderson's (2003) equivalency theorem is intended to help design interaction activities in open, distance, and digital education (ODDE). According to this theorem, of the three types of interaction, namely teacher–student, student–student, and student–content interaction, a high level of one type will lead to good learning outcomes

even if the other two are sparingly or not at all offered. Studies claiming to prove the robustness of the equivalency theorem are too many to list. A typical case in point is the meta-analysis of Bernard et al. (2009). Nonetheless, Xiao's (2017) appraisal of the studies included in this meta-analysis shows that none of them involved only one type of interaction because they "were categorized by the most prevalent interaction type contained in the independent variable" and did not take into account the other type(s) of interaction actually happening (Bernard et al. 2009, p. 1253). Therefore, none of the studies can prove that the effectiveness of learning can be enhanced even if there is only one type of interaction.

It would be no exaggeration to say that defects in the design of a study may misdirect and misinform future research and practice, likely to do more harm than good to learning and teaching eventually. Given that no study has been conducted with an exclusive focus on reviewing CALL publications from the perspective of research design, this study aims to fill the gap by answering the following question: What are the strengths and weaknesses of CALL research in terms of research design, as evidenced by publications in CALL-core journals?

2 Methodology

2.1 Sampling

2.1.1 Selected journals

Selecting journals was the first step. To this end, the inclusion and exclusion criteria were established and applied (see Table 1).

Only journals that meet all the above criteria were selected. The search resulted in five journals (see Table 2). All of them have CALL in their titles except *Language Learning & Technology* (*LL&T*) which is counted in as a CALL journal in Choubsaz et al. (2024) and Mohsen et al. (2024).

Table 1: Inclusion and exclusion criteria for journal selection.

Inclusion	Exclusion
– Listed in SSCI or ESCI;	– Neither listed in SSCI nor ESCI;
– Published in English;	– Published in languages other than English;
– Had an exclusive or major focus on CALL;	– Did not have an exclusive or major focus on CALL;
– Peer-reviewed.	– Not peer-reviewed.

Table 2: Selected journals.

Journal titles	SSCI/ESCI
Computer Assisted Language Learning	SSCI
ReCALL	SSCI
Language Learning & Technology (*LL&T*)	SSCI
CALICO Journal	ESCI
International Journal of Computer-Assisted Language Learning and Teaching (*IJCALLT*)	ESCI

2.1.2 Selected articles

After journal selection, the next step was to select articles published in the sample journals. These articles had to be analyzed manually piece by piece because no software is available to process data for the purpose of this study. To avoid an excessive workload, the scope of selection was limited within a two-year timeframe between 2023 and 2024. The inclusion and exclusion criteria were then established and applied (see Table 3).

The final sample consists of 338 articles (see Table 4). The articles included were then categorized using the three-level research framework widely adopted in the field of ODDE (Zawacki-Richter 2009; Zawacki-Richter and Anderson 2014). According to this framework, macro-level research covers aspects of education systems and theories, including access, equity and ethics; globalization; teaching/delivery systems and institutional partnership in developing cross-border programs; theories/theoretical research; research and knowledge transfer (including literature reviews). Meso-level research mainly refers to institution-level management, organization, and technology use/adoption (hence including technology/media review). Micro-level research focuses on learning and teaching practices, mainly with students and teachers as subjects of investigation. Table 4 displays the distribution of these three levels of research.

Included in the 338 articles selected are 282 articles reporting on empirical studies and 56 non-empirical research articles, namely theoretical/conceptual

Table 3: Inclusion and exclusion criteria for article selection.

Inclusion	Exclusion
– Research articles including systematic literature review and technology/media review;	– Editorial and book review;
– Published in 2023 or 2024;	– Not published in 2023 or 2024;
– Full-text available.	– Full-text unavailable.

Table 4: Publications in the selected journals.

Journals selected		*Computer Assisted Language Learning*	*ReCALL*	*LL&T*	*CALICO Journal*	*IJCALLT*	Total number
All publications		143	47	99	55	32	376
Full-text unavailable					3		3
Articles excluded			5	9	18	3	35
Articles included (n = 338)	Micro-level: teaching and learning	117	26	72	17	24	256
	Meso-level: management, organization and technology	4	3	10	13	2	32
	Macro-level: education systems and theories	22	13	8	4	3	50
	Number of included articles	143	42	90	34	29	338

articles (macro-level category), literature reviews (macro-level category) and technology/media reviews (meso-level category).

2.2 Assessment scheme

The current study is an approximate replication of an earlier review with a dedicated focus on the design of ODDE research (Peng and Xiao 2022). Informed by relevant research design literature, Peng and Xiao (2022) designed and developed an assessment scheme specifically for the purpose. The original scheme, although evaluated by experienced researchers and gatekeepers in the field, was further adapted and revised to better serve the purpose of the current study. For example, author affiliation, research area, and theoretical foundation were added as new dimensions and some hints were tweaked with the benefit of hindsight from the previous study (see Table 5). Theoretical foundation, an element that can and should be deliberated before a study is conducted, was not included in the original scheme. It is the same case with research area. Therefore, both theoretical foundation and research area were incorporated into the assessment scheme. And according to a systematic review, research on AI applications in higher education tends to lack theoretical underpinnings, which may be because the majority of researchers are not affiliated to teaching and learning departments of an institution (Zawacki-Richter et al. 2019). We added author affiliation with the purpose of finding whether this is the case with our study.

Table 5: Assessment scheme (adapted from Peng and Xiao 2022).

Aspects to be assessed	Hints	
Author affiliation	Was the author affiliated to a teaching and learning department or a technology department in an organization? Or was his/her disciplinary background education-oriented or technology-oriented if no information was available about his/her specific affiliation?	
Research area	Did the study aim at a macro, meso, or micro-level issue?	
Context of study	Was the context of study clearly stated?	
Hypothesis, research question or purpose	Was the hypothesis, research question or purpose of the study clearly and adequately stated?	
Theoretical foundation	Was the study informed by any (language) learning theory, hypothesis, model, or framework?	
Approach	Did the study adopt a quantitative, qualitative or mixed-method approach?	
Design	Experimental	If the study was designed as an experiment, was it a true experiment or quasi-experiment?
	Non-experimental	If it was not an experimental study, what was it? Did the author name the research design used? If not, look for evidence and classify it into a proper design category.
	Replication	Was it a replication study?
Duration	If the study involved an intervention/treatment or was length-sensitive, was the duration specified?	
Sampling strategy	Did the author name the sampling strategy used? If not, look for evidence and classify it into a proper category.	
Data collection	Source of data	What kind of data was collected to answer the research question(s)? Was data from more than one source collected for the purpose of triangulation?
	Instrument (questionnaire, scale, rubric, interview protocol, and test)	Was any instrument such as questionnaire, scale, rubric, interview protocol, or test used to collect data? If yes, was it developed exclusively for the study or adapted from an existing one by the researcher? Or was it an existing instrument developed elsewhere? Was the instrument tested through statistical methods or reviewed and piloted, before it was used to collect data for this study? Or was it only reviewed, or only piloted, or neither reviewed nor piloted? Was the content (for example, questionnaire items, interview protocol, or test items) available in its entirety?
	Administration	Was data collection, including the how, who and when, clearly stated?
	Cross-sectional or longitudinal	Was the data collected only at one point in time or over a period?

Table 5: (continued)

Aspects to be assessed	Hints
Data analysis	Was the method of analysis clearly stated?
Researcher bias	Did the author explain how his/her own potential bias and influence was managed throughout the study, including sampling, data collection, and/or intervention administration?
Ethical issue	Were ethical issues, if applicable, adequately addressed?
Limitation	Was there any limitation acknowledged by the author?

It is worth mentioning that this assessment scheme is applicable to articles reporting on empirical studies (n = 282). As for the 56 non-empirical research articles, only author affiliations and categorization of research areas were noted down; the other dimensions of the scheme were not applicable.

2.3 Sample analysis

In view of the purpose of the study, the method of directed qualitative content analysis was employed to analyze the articles (Hsieh and Shannon 2005). Prior to the formal analysis, the two authors used the adapted assessment scheme to review 10 randomly selected articles independently. The results were compared and consensus achieved about discrepancies through re-reading the articles concerned and scrutinizing controversial contents together. To further reduce differences, another 10 articles were randomly selected for a second round of tentative analysis. This time, only minor differences existed in the results.

The two authors then conducted a formal analysis of the remaining 318 articles independently, following the same procedure used in the two rounds of trial analysis to finalize the results. To further reduce potential researcher bias and ensure coding reliability and validity, an independent researcher was invited to use the framework to analyze 40 articles of his random choice. Only a few minor discrepancies exist between his analysis results and the authors', which shows that the authors' coding was reliable.

2.4 Limitations of this study

A key limitation is that this review is intended to provide an overarching perspective on the design of CALL research rather than a fine-grained analysis of individual articles. In a sense, its method of analysis is more quantifiable than qualitative. For

example, it assessed whether the data analysis process was clearly explained, not whether it was appropriate for the specific study. Similarly, it assessed whether a study adopted a quantitative, qualitative, or mixed-method approach and identified the type of study design followed, not whether the research approach or study design adopted was appropriate for the purpose of the study. In addition, it neither evaluated how well the findings were interpreted and discussed nor judged whether the conclusion was adequately supported by the findings. A fine-grained approach, namely taking each individual article as a unit of analysis and analyzing how well each component of its research design aligns with each other as a whole, would yield more insights into the rigorousness and trustworthiness of CALL research design.

Furthermore, as readers may have already noticed, our selection of articles was limited to those published in English in five CALL-focused journals within a relatively short timeframe of two years. The findings may not be able to capture the full CALL research landscape despite the fact that these journals are internationally reputed. Future efforts to overcome these limitations are worthwhile.

3 Findings

3.1 Author affiliation and research area

Statistics show that 86.4 % of the articles (n = 292 out of 338) were (co-)authored by researchers affiliated to institutional departments for learning and teaching as well as educational research or with an education-oriented background and 13.6 % (n = 46) were co-authored by researchers from both education-oriented and technology-oriented departments or with both backgrounds. In other words, none of the studies was conducted by researchers only from departments for technology research and development (R&D) or only with a technology R&D background.

As for categorization of research areas, three quarters of the articles (75.7 %, n = 256 out of 338) fall within the category of micro-level research, mainly about "This is what I did to my course or my students with a technology and it worked really well" or the relationship between student characteristics and CALL. Meso and macro-level research only accounts 9.5 % (n = 32) and 14.8 % (n = 50) respectively. All the meso-level studies are concerning new trends/affordances of educational technology for language learning and the design, development, testing, and/or validation of a technology/tool/program/software/system, with only one exception which aims to adapt English Language Proficiency Admissions Assessments for postsecondary enrollments to address technology-mediated language skills in the digital age (Cardwell et al. 2024). Of the 50 macro-level studies, 38 are literature reviews. As for the remaining 12, three are conceptual, one aims to identify CALL researchers'

genuine research strategies (Meihami 2024) and another attempts "to construct a working theory of IVR as a catalyst for understanding and creating multiple forms of language in use" (Karimi et al. 2023, p. 24) while the others intend to develop, test and/or validate a model, taxonomy, scale or pedagogy.

3.2 Context and purpose of study

All the empirical research articles (n = 282) provide background information concerning what was investigated to enable readers to situate the study in the broader context when interpreting its findings and evaluating (possible) relevance to other situations. Only less than two percent (1.8 %, n = 5) of these articles do not have a clear or correct account of their context. For example, Lee and Lu (2023), situated in a public secondary school in Southeastern China, claim that "70 % of the participants (n = 292) reported having spent their daily time on extramural English activities on the Internet (e.g., watching YouTube clips in English or chatting with others in English via social media)…" (p. 131). Nevertheless, it is open to doubt that these secondary school students were able to access YouTube and the like to practice using English. Another case in point is Jakonen and Jauni (2024) conducted in the Finnish higher education context. "The courses … were part of the university's regular curriculum" (Jakonen and Jauni 2024, p. 878). However, no information about the Finnish higher education context and "the university" was available. Further, readers have no idea who the students were except that they were learners of different second languages.

Integrally related to the research context is the statement of the purpose of study or the framing of the research questions or hypotheses. All the empirical research articles include the purpose, research questions, and/or hypotheses of the study. However, in 22 % of the cases (n = 62), relevant accounts are inadequate and likely misleading in that they imply the researchers' preconception about the research question. Take Feng and Ng (2024). The research question – "What are the spatial-related factors in virtual environment that affect learners' vocabulary learning? " – suggests the researchers' belief in the influence of spatial-related factors on learners' vocabulary learning even before the study was conducted (Feng and Ng 2024, p. 2). Typical research questions which implicate the researchers' preconception or assumption include "What are the differences between online Chinese language learners' LLS use in the asynchronous and synchronous environments, and what are the influential factors?" (Chen and Rodway 2023, p. 202), "How does ELLs' initial level of GRV knowledge impact the effect of a texting-based intervention on their target vocabulary acquisition…?" (Li et al. 2023, p. 56), "What professional

benefits did the teacher educators identify from facilitating the multinational telecollaborations?" (Wach et al. 2024, p. 2), and "How much does the use of Google Assistant significantly improve the adolescent EFL learners' oral proficiency?" (Tai 2024, p. 1286).

3.3 Research approach, design and theoretical framework

In terms of research approach, nearly half of the empirical studies (46.4 %, *n* = 131 out of 282) adopt a mixed-method approach while quantitative and qualitative studies represent 30.9 % (*n* = 87) and 22.7 % (*n* = 64) respectively.

As for types of design, 48.6 % (*n* = 137) are experimental and 51.4 % (*n* = 145) non-experimental. Of the experimental studies, over one third (36.5 %, *n* = 50 out of 137) are true experiments and 63.5 % (*n* = 87 out of 137) quasi-experiments.

Nearly half of the non-experimental studies adopt the design of case study (46.9 %, *n* = 68 out of 145) and 29.7 % (*n* = 43) belong to the category of survey research. The number of studies following other designs is apparently insignificant (see Table 6). On the other hand, only 17 out of 282 (6 %) are replication studies, five of which are published in a special issue on replication of *ReCALL* (volume 35, issue 2).

When it comes to theoretical underpinnings, two-thirds of the empirical studies (67.4 %, *n* = 190) are informed by relevant theories, models, hypotheses, or frameworks, with 29.4 % of them including a (sub)section specifically to explain the theoretical underpinnings of their studies. However, it is worth noting that 32.6 % (*n* = 92) are atheoretical in that no specific theory is mentioned as the foundation for designing the study, analyzing the data and interpreting the findings.

Table 6: Non-experimental designs (*n* = 145).

Type of non-experimental design	*n*	Percentage
Case study (ethnography)	68	46.9 %
Survey research	43	29.7 %
Scale/model/tool development and/or validation	11	7.6 %
Correlational research	8	5.5 %
Action research	3	2 %
Design-based research	3	2 %
Grounded theory research	2	1.4 %
Narrative research	2	1.4 %
Observational research	2	1.4 %
Phenomenological research	2	1.4 %
Self-study research	1	0.7 %

3.4 Sampling and duration

Only slightly over one-fifth (21.3 %, n = 60 out of 282) of the empirical research articles name their sampling strategies. Of the named strategies, eight (13.3 %) were probabilistic, mostly random sampling and 86.7 % (n = 52 out of 60) non-probabilistic, mostly convenience or purposive sampling (see Table 7). Of the 222 studies (78.7 %) which do not specify the sampling strategies used, evidence collected to identify their sampling strategies shows that 80.2 % (n = 178 out of 222) use convenience sampling and 19.8 % (n = 44 out of 222) adopt purposive sampling.

In other words, of the 282 empirical studies under review, probabilistic and non-probabilistic sampling strategies are used by 2.8 % (n = 8) and 97.2 % (n = 274) of the studies respectively. In the non-probabilistic sampling studies (n = 274), 76.3 % (n = 209 out of 274) follow convenience sampling and 21.5 % (n = 59 out of 274) adopt purposive sampling, with the remaining 2.2 % (n = 6 out of 274) using a mixture of convenience and snowballing, purposive and snowballing, or convenience and purposive sampling (see Table 7).

Of the 229 studies involving intervention or treatment, only twenty studies (8.7 %) do not specify the duration. It is unrealistic to classify duration into different types of length because different units of measurement are applied in the studies, including minute, hour, day, week, month, semester/term, (school) year, course, module, class, session, meeting, and episode. Only some of the articles further specify the details of units of measurement. For example, "participants completed elicitation and treatment tasks via computer in a single 90-min session" (Richards 2024, p. 5). "This study lasted for one semester, a total of 17 weeks…the three groups were required to interact with either Google Assistant, the L1 English speakers, or the L2 English speakers in their free time in 10-min sessions twice a week" (Tai 2024, p. 1290). However, many of these units of measurement may mean different lengths of time in different contexts or are ambiguous even in the same context but are not specified, hence likely to cause confusion. A typical case in point is the use of "episode" in this

Table 7: Named sampling strategies (n = 60).

Probabilistic sampling		**Non-probabilistic**	
Random sampling	5	Convenience sampling	31
Stratified random sampling	1	Purposive sampling	15
Cluster random sampling	1	Convenience & snowballing sampling	4
Hierarchical cluster sampling	1	Purposive & snowballing sampling	1
		Convenience and purposive sampling	1
Total	8		52

statement "we present a single case analysis based on one episode which includes all identified practices in the dataset" (Badem-Korkmaz and Balaman 2024, p. 1889). There is no knowing how long one episode lasts.

The shortest duration is about 4 mins in a study investigating L2 learner output in face-to-face versus fully automated role-plays which "used a multi-turn conversation task that was developed to elicit approximately 2 mins of oral interaction" with participants completing the task in two different formats (Timpe-Laughlin et al. 2024, p. 158). The longest duration is 1–9 years in an investigation of "how learners engage in self-initiated and self-directed feedback practices beyond the classroom in online spaces" (Lyu and Lai 2024, p. 114).

3.5 Data collection

Sixty-seven percent (n = 189 out of 282) of the empirical studies triangulate data from more than one source to enhance the credibility and validity of their findings. Hence, the identified sources of data outnumber the reviewed studies (575 vs. 282). Questionnaire/scale/rubric, test, and interview top the list of sources, trailed distantly by other sources (see Table 8). Slightly over half of the studies (51.1 %, n = 144 out of 282) collected their data from questionnaire, scale or rubric, with 61.1 % (n = 88 out of 144) closed-ended in nature. Test (35.8 %, n = 101 out of 282) is the second most popular source of data while interview (34 %, n = 96 out of 282) is the third most frequent source, mostly open-ended (94.8 %, n = 91 out of 96). However, it should be pointed out that in nine cases, no information is available about the questionnaire, scale or rubric and the interview protocol used. The remaining sources of data are far less significant. Ranking the 4th and 5th are participant performance (writing/oral presentation) and online (text/video/human-chatbot) interaction (including social media posts), which are used in 13.1 % and 12.4 % studies respectively (see Table 8 for other sources of data).

In addition to collecting data from different sources, it is not unusual that studies employ more than one instrument to collect data from one source. Therefore, the instruments used outnumber the studies reviewed (360 vs. 282). For example, Gok et al. (2023) used five data collection instruments – "University Placement Test scores, TOEFL reading test, Demographic Questionnaire, FLCAS and FLRAS" (p. 848), that is, two tests, one questionnaire, and two scales. As can be seen in Table 5, by instrument, we refer only to the three most frequent ones – questionnaire/scale/rubric, interview, and test. Therefore, the actual instruments used are more than 360.

Of the 360 instruments used, 70 % (n = 252) are developed by the authors specifically for the purpose of their study, 13.9 % (n = 50) are adapted from previous

Table 8: Sources of data (n = 575).

Sources of data	Number
Questionnaire/scale/rubric	144
	Closed-ended (n = 88)
	Open-ended (n = 21)
	Closed and open-ended (n = 28)
	No information (n = 7)
Test	101
Interview	96
	Open-ended (n = 91)
	Closed and open-ended (n = 3)
	No information (n = 2)
Participant performance (writing/oral presentation)	37
Online (text/video/human-chatbot) interaction (including social media posts)	35
Video recording/screen capture video	27
Written reflection	25
Platform/software data (e.g. eye-tracker data)	24
Observation/field note	21
Course artefact (e.g. multimodal poster)	20
Course activity	11
Stimulated retrospective recall	8
Learning journal	8
Focus group discussion	8
Feedback comment	6
Thinking aloud	1
Visualized vocabulary knowledge mind mapping	1
Course syllabus and description	1
E-portfolios	1

studies, and 16.1 % (n = 58) are originally developed in previous studies but used in the current studies without making any change (see Table 9).

The majority of these instruments (77.2 %, n = 278 out of 360) are neither reviewed nor piloted or statistically tested before being used to collect data or in the case of existing instruments, not given reasons why they are fit for the current study. Put specifically, this is the case with 77.4 % of the newly developed instruments (n = 195 out of 252), 78 % of the adapted instruments (n = 39 out of 50), and 75.9 % of the existing instruments (n = 44 out of 58). In contrast, only 10.8 % (n = 39 out of 360) are both reviewed and piloted or statistically tested and 12 % (n = 43 out of 360) either reviewed or piloted, before being put to use.

Table 9: Instruments (questionnaire/scale/rubric/interview protocol/test) (n = 360).

	Instrument	Number
Newly-developed (n = 252)	Reviewed and piloted (or statistically tested)	24
	Reviewed	11
	Piloted	22
	Neither reviewed nor piloted	195
	Content availability in entirety	142
Adapted (n = 50)	Reviewed and piloted (or statistically tested)	4
	Reviewed	1
	Piloted	6
	Neither reviewed nor piloted	39
	Content availability in entirety	27
Existing (n = 58)	Reviewed and piloted or explanation given	11
	Reviewed	2
	Piloted	1
	Neither reviewed nor piloted or no explanation given	44
	Content availability in entirety	21

As for content availability, slightly over half (52.8 %, n = 190) are available in their entirety, either included in the main text, as appendixes or with a link to where they are digitally stored. Put specifically, this availability applies to 56.3 % (n = 142 out of 252) of the new category, 54 % (n = 27 out of 50) of the adapted category, and 36.2 % (n = 21 out of 58) of the existing category.

An overwhelming majority of the empirical research articles (94 %, n = 265 out of 282) explain in clear terms how the data are collected, with 3.9 % (n = 11) not clear enough and 2.1 % (n = 6) failing to clarify this procedure. For instance, Han et al. (2023) "collected two sources of data. First, it collected the participants' CFL teaching practices. Second, it collected their reflections as recorded in the evidentiary chapters of their thesis and a focus group discussion (conducted for 90 min and audio taped)" (p. 6). No further information is provided concerning how the participants' teaching practices were collected and how the focus group discussion was organized and conducted. It is the same case with Alharbi (2024) with its Data Collection section detailing how the content of the instruments address "specific aspects of the research questions" rather than the how, who, and when of data collection (p. 7).

Last but not least, it is important to note that 80.1 % (n = 226 out of 282) of the studies collected data at one point in time. In other words, only one-fifth (19.9 %, n = 56) are longitudinal in nature in that data collection takes place more than once over a period.

3.6 Data analysis

Most of the empirical research articles (81.2 %, *n* = 229 out of 282) state the process of data analysis in clear terms. However, nearly 20 % are either less informative (16.3 %, *n* = 46) or do not explain how their data is analyzed (2.5 %, *n* = 7). Han et al. (2023) cited above is a case in point that does not explain how the participants' teaching practices, their reflections, and focus group discussion were analyzed. It is the same case with Smith et al. (2023). Another example is Wen et al. (2023) that does not mention how the qualitative data from the open-ended part of the questionnaire was analyzed.

3.7 Researcher bias, ethical concern and limitation

Less than 10 percent of the empirical research articles explain how the researchers' biases are addressed (8.5 %, *n* = 24 out of 282). Good examples include Ekmekçi (2023) that "deliberately conducted the interview just after the final performances of the students had been graded in order to avoid any possible concern of the students about the objectivity of grades" (p. 1016) and Zhang et al. (2024), whose interview questions "were not directly related to the usage of mobile phones or other types of electronic devices for learning medical vocabulary" to "avoid biased responses from the interviewees" (p. 2013). Only one article (Kessler 2023) has a section exclusively for the purpose, with the heading "Researcher Positioning", explaining how the researcher managed to avoid his influence on the students' reflections on the learning experience.

In contrast, as high as 91.5 % (*n* = 258) do not clarify the ways researcher biases are reduced or eliminated throughout the research process so that the validity and reliability of the findings are not undermined. In fact, this situation also begs the question whether the researchers/authors realize or acknowledge the existence of their biases. They may even take possible biases as advantages. A case in point is Jiang et al. (2024) which sets out to prove the effectiveness of the Duolingo English course. No researcher bias is acknowledged although three of the four authors work for Duolingo. Similarly, Mendes de Oliveira et al. (2023) "was conducted in collaboration with the teams responsible for enterprise sales and academic research at the language learning company Babbel" to investigate users' experience and perceptions of Babbel's virtual classroom solution without explaining how to reduce or avoid possible impact of this collaboration on the research findings (p. 1509).

Ethical concerns are issues that should be properly addressed in studies involving human subjects. Nearly 60 % of the empirical studies (59.6 %, *n* = 168 out of

282) explain the ways ethical issues are handled. However, the measures taken may not be sufficient to avoid possible negative ethical consequences. Informed consent is a cliché most often used although it happens that no information is given regarding what the participants consented to or whether they really knew what they consented to. Take García-Pastor and Calatayud (2023) whose participants were aged 16. A brief statement that they all "provided informed consent for the research" cannot dismiss ethical concerns related to this study (García-Pastor and Calatayud 2023, p. 319). It is the same case with the statement that "informed consent was obtained from the participants for experimentation" in a study involving 9th-grade students as participants (Chen and Lee 2023, p. 1092). Similarly, simply anonymizing or giving monetary reward is far from appropriate.

Of the remaining articles that make no statement of how ethical concerns are addressed (40.4 %, n = 114 out of 282), this requirement is not applicable to eight of them either because it is a self-study (e.g. Wach et al. 2024), "is based on publicly available anonymized corpus data" (Blázquez-Carretero 2023, p. 336), or does not involve human subjects (e.g. Díez-Arcón and Agonács 2024). Therefore, there are 106 articles (37.6 %) that do not take into account ethical issues related to their research.

All studies have limitations. Over ninety percent of our sample articles reflect on the limitations pertaining to their studies (92.2 %, n = 260 out of 282). However, as is the case with ethical concerns, some articles tend to understate their limitations by only mentioning those that readers can easily identify such as small sample size, no triangulation, or lack of generalizability.

4 Discussion

Rigorous design is the first step in ensuring the quality of scientific inquiry (Simonson et al. 2011). Therefore, attention to the design of a study cannot be overemphasized. Whether a study is properly designed may influence the reliability, validity, generalizability, and/or replicability of its findings and, needless to say, the robustness of its conclusions. Critiquing CALL publications with an exclusive focus on research design may shed light on the trustworthiness of CALL research and carry implications for future research.

This section will highlight both strengths and weaknesses in the landscape of CALL research publications. It should be borne in mind that even in the strength areas there is still room for improvement as can be seen in the discussion below.

4.1 Strengths

4.1.1 Clear statements of context and purpose of study

Education is contextualized and influenced by contextual factors. Therefore, educational research should also be situated. Previous reviews tend to focus on the types of context in which CALL studies are situated (e.g., DeMolder et al. 2023; Klímová and Seraj 2023; Zhang et al. 2023). The current study differs from previous ones in that it centers on whether sufficient contextual information is provided which is key to the interpretation of the findings and to the replication of a study (Shortt et al. 2023). Researchers should have a clear idea of the context of study at the very beginning and state relevant details clearly in their publication. Clear description of the context of study is a strength in the empirical studies reviewed, with only 1.8 % less clear or descriptive.

Unlike Shortt et al. (2023) that identifies types of research questions (i.e. performance-, attitude and motivation-, or design-oriented) used in their samples, our study focuses on the framing to see whether they embed researcher bias, preconception or assumption which may mislead the research process, including data collection, analysis and interpretation. The purpose, research questions or hypotheses "provide critical information to readers about the direction of a research study…also raise questions that the research will answer through the data collection process" (Creswell 2012, p. 109), consequently defining "the most appropriate participants, source of data, and method of data analysis for the study" (Peng and Xiao 2022, p. 10). For example, Luo and Watts (2024) aim "to explore the nature of smartphone-assisted ELL…" (p. 614), hence requiring that the participants had smartphones and used them to learn English. However, the article does not provide this information. All the empirical research articles in our review state their purpose of study or research questions/hypotheses although 22 % of them are inadequate in that they are somewhat biased. For example, "How much does the use of Google Assistant significantly improve the adolescent EFL learners' oral proficiency?" (Tai 2024, p. 1286) – this research question suggests that the use of Google Assistant has a positive effect on adolescent EFL learners' oral proficiency. Nevertheless, this does not make sense in that the purpose of the study was to find out whether Google Assistant was conducive to adolescent learners' EFL. With this preconception or assumption in mind, the researcher may not have been able to maintain neutrality in collecting data and interpreting the findings.

4.1.2 Increasing adoption of mixed-method approach

Another strength is the popularity of mixed-method approach studies representing nearly half of the empirical research samples, in comparison with purely quantitative or qualitative ones. Hubbard (2009) aptly observes that "although quantitative studies probably dominated in the early literature, qualitative and mixed-method studies are now common, especially in the area of computer mediated communication (CMC)" (p. 5). Our finding reinforces findings from other reviews (e.g., Lee 2023; Peterson 2023; Shadiev and Yu 2024; Shi and Aryadoust 2024). The proportions of different research approaches may vary a little bit in some reviews (e.g., Kim and Namkung 2024; Zhang and Sun 2023; Zheng et al. 2022), which may be due to the difference in the activities/tasks/interventions involved, the technologies used, or the specific topic reviewed, in other words, the subfields that they focus on. Our sample empirical studies are not focused on any particular subfield, hence more representative.

4.1.3 Substantial proportion of experimental studies

Study design is also a strength, with nearly half of the empirical studies (48.6 %) experimental in nature, far more than those in Loncar et al. (2023) (21 %), a review of technology-mediated feedback for L2 English writing literature and Shi and Aryadoust (2024) (17.6 %), a review of AI-based automated written feedback research. Increase in experimental studies may be an emerging feature of the CALL research landscape because, as pointed out above, our review is more comprehensive instead of focusing on a particular subfield. Of the experimental studies in our samples, nearly two-thirds (63.5 %) are quasi-experiments. Given the complexity of educational research, true experiments may not always be feasible, especially when taking logistical challenges and ethical consequences into consideration. Researchers are therefore justified in exercising certain degree of freedom or discretion in deciding how a project is conducted (van Drimmelen et al. 2024). In this sense, it is understandable that there are more quasi-experiments than true experiments.

4.1.4 Significant proportion of theoretically-informed studies

Unlike Shadiev and Yu (2024) that investigate the theory, hypothesis, model, or framework upon which their sample studies are based, our review focuses on whether a study is underpinned by any theory, hypothesis, model, or framework. Lack of theoretical underpinnings is a common feature of educational technology research in general (Bond et al. 2019; Bulfin et al. 2014; Prinsloo 2018). A review of three top educational technology journals *Computers & Education*; *Learning, Media*

and Technology, and *British Journal of Educational Technology* shows that 40 % of their publications are atheoretical (Hew et al. 2019), consistent with the findings of a review of literature on personalization in educational technology between 1960 and 2015 (Bartolomé et al. 2018). A review of research on AI applications in higher education also reveals a weak connection of their sample studies to theoretical pedagogical perspectives possibly due to "the low presence of authors affiliated with Education Departments" (Zawacki-Richter et al. 2019, p. 22). In comparison, two-thirds of the empirical studies in our review (67.4 %) are underpinned by relevant theories possibly due to the higher percentage of authors affiliated to education-related departments (86.4 %). This is no doubt a positive trend. Nevertheless, more needs to be done given that one-third are still atheoretical, as in the case of Shadiev and Yu (2024), and especially in view of a longitudinal analysis of highly cited articles published between 1983 and 2019 in four CALL journals, according to which 18 % of the seminal studies "did not explicitly adopt any theory to frame their research" (Choubsaz et al. 2024, p. 49).

4.1.5 Frequent data triangulation and clear statements of data collection and analysis

The sources of data are twice the sample studies (575 vs. 282) because two-thirds of the empirical studies employ triangulation, which can definitely contribute to validity and reliability. In addition, only 6 % of the empirical research articles do not explain their data collection process clearly to enable readers to judge the rigorousness of data collection and facilitate replicability while about 20 % are not articulate enough about data analysis. Overall, data triangulation, collection and analysis can count as a strength. Having said that, like data collection, descriptions of data analysis should be as clear and specific as possible to enable readers to assess the trustworthiness of the findings and replicate the study if they so wish.

4.1.6 Acknowledgement of limitations

Finally, over ninety percent of the empirical research articles acknowledge the limitations of their studies. No doubt, this is a strength. No study is immune from limitations (Creswell 2012). Reflecting on the limitations of a study can help contextualize the interpretation of its findings and serve as directions of future research not only for the author(s) but also for other researchers. This process should not be taken as a procedural requirement of academic publication with perfunctory acknowledgement. A good case in point is Roy (2024) which, together with the easily identified limitations, points out that other possible and yet less easily spotted weaknesses "such as teacher quality, student motivation, or classroom environment

may have influenced the observed differences between groups" and that "technical issues, such as occasional problems with the DST method, may impact the overall experience and results" (p. 15) (also see Kourtali and Borges 2024). It should be noted, though, that the limitations stated in some articles seem to be superficial, only mentioning those that readers can easily identify such as small sample size, short duration, or no triangulation.

4.2 Weaknesses

4.2.1 Predominance of micro-level research

The current landscape is dominated by micro-level research typically exemplified by "This is what I did to my course or to my students". Micro-level research constitutes the foundation of the field. However, learning and teaching is affected by factors not only at the micro level but also at the meso and macro levels, for example, socio-economic development, socio-cultural context, institutional infrastructure, cost and finance, quality evaluation and assurance policy and mechanism, educational resources as well as support for both faculty and students (Xiao 2023a). Research into these macro- and meso-level issues is essential to the dissemination or generalization of findings from micro-level research, hence the sustainable development of CALL. Meso and macro-level research represents only about one quarter of all the articles in our review, with the former focusing on technology trends and design, development, testing or validation and the latter falling within the areas of theoretical research, research and knowledge transfer (literature review). This lack of interest in meso and macro-level issues may be due to the high percentage (86.4 %) of authors with learning and teaching or education-related backgrounds, namely CALL practitioners. Researchers need to explore and understand the big picture of CALL rather than confine research to issues directly related to micro-level practice.

4.2.2 Scarcity of replication studies

Replication studies are "critical to the growth and credibility of our discipline" (McManus 2024, p. 1), refining findings, contributing to generalizability, and strengthening conclusions (Gass et al. 2021). It is "a sign of a field's maturity" (Smith and Schulze 2013, p. i). However, only 6 % of our empirical samples replicate previous studies, a situation that warrants attention and needs redressing. This is an obvious weakness in that there is no knowing "if the results hold for a different population, in a different setting, or for a different modality" (Polio and Gass 1997, p. 502). A decade ago, *CALICO Journal* published a special issue on replication and evaluation in CALL,

hoping that this would be "the beginning of many replication studies" to be published in this journal (Bikowski and Schulze 2015). Despite the editors' promise that they will be "providing a venue for CALL scholars to disseminate their replication research" (Smith and Schulze 2013, p. ii) and the consensus on the imperative of replication research in CALL, this kind of research remains scarce (Tschichold 2023). Difficulties in conducting replication studies, in particular exact replication (Chun 2012; Foung and Kohnke 2023; Tschichold 2023) cannot justify the scarcity. More needs to be done.

4.2.3 Insensitivity to researcher biases

Another apparent weakness is that over 90 % of the empirical research articles do not acknowledge potential researcher biases and explain what measures the researchers have taken to prevent possible interference of these biases in the research process. The researcher's neutrality in the research process, researcher-participant relationship, and interpretation of data and findings may suffer consequences because of these biases (Werth and Williams 2021). Researchers' awareness of and actions to minimize own subjectivity cannot be overemphasized (Jung 2025). Negligence in this regard is likely to affect various stages of the research process, hence undermining the trustworthiness of the research outcomes. This seems to be a weakness of the research landscape of technology-enhanced education in general (Peng and Xiao 2022; Zhang et al. 2023).

4.2.4 Overuse of non-probabilistic, convenience sampling

Sampling is also found to be a weakness of the empirical studies reviewed in our study, echoing Ballance's (2024) conclusion that "little attention is given to sampling" (p. 58). Only about one-fifth name their sampling strategies. Furthermore, of all the 282 empirical studies reviewed, only 2.8 % (n = 8) employ probabilistic sampling strategies, mostly random sampling while the overwhelming majority apply non-probabilistic sampling strategies, with three quarters of them convenience sampling. The dominance of non-probabilistic strategies, especially convenience sampling may undermine the reliability and validity of the findings (Ballance 2024; Vehovar et al. 2016). Regrettably, it is a phenomenon commonly found in other reviews (e.g. Shortt et al. 2023). While acknowledging the difficulties that may arise from other sampling strategies, in particular probabilistic sampling, we should not allow any particular sampling technique, especially non-probabilistic technique to prevail. Sampling matters to empirical studies because it affects whether "the effect of unknown sources of variance" can be neutralized (Ballance 2024, p. 64).

4.2.5 Shortage of longitudinal research

The shortage of longitudinal studies is a weakness, with only 20 % collecting their data more than one time. This also features in other reviews (e.g., Soyoof et al. 2023). As argued by Kim and Namkung (2024), "more longitudinal projects are warranted to understand learner perception dynamics … over time"(p. 17) in order to "provide more reliable and/or generalizable results" (Ballance 2024, p. 60), an argument supported by Xiao et al. (2024). "The effectiveness of an educational intervention needs to be tested over time" to overcome novelty effect (Xiao 2023a, p. 215). Reeves and Lin (2020) suggest a new direction of research which may be equally applicable to the field of CALL "whereby we develop robust, multi-year research agendas focused on important problems and innovative solutions, judge our worthiness for promotion and tenure on evidence of impact rather than simple article counts, closely collaborate with practitioners, and establish our field as preeminent in meeting global problems related to education" (p. 1999). As is the case with other weaknesses, while there are challenges with longitudinal empirical research, for example, time commitments, funding, and sample stability, among other things (Barkhuizen 2009; Jenkins et al. 2011), more longitudinal studies are needed to consolidate the knowledgebase of CALL.

4.2.6 Undue reliance on self-reported data and unverified and/or unavailable instruments

Another weakness relates to the undue reliance on self-reported data and unverified and/or unavailable instruments. Of the 19 sources identified, survey (questionnaire/scale/rubric and interview) and test are the most frequent sources, far outnumbering the other sources, similar to Kim and Namkung (2024) and Zheng et al. (2022). Questionnaire and interview are also found to be the most popular data sources in Shadiev and Yu (2024). On the other hand, although our sample empirical studies often use more than one instrument to collect data from survey, survey data is self-reported in nature and may not reflect the participants' true feelings, ideas, or conditions, due to various factors such as social desirability bias. This is especially the case when only about 10 % of the instruments are both reviewed and piloted or statistically tested before they are applied in the current studies and 47.2 % are unavailable in their entirety. This situation may jeopardize the quality of the data obtained as well as research replicability. Instrument availability in its entirety can also enable readers to assess whether the instrument is fit for purpose or whether it is biased. For example, Yang et al. (2023) claims to use the Chinese version of the Foreign Language Classroom Anxiety Scale (FLCAS) by Horwitz et al. (1986) in their study but goes on to say "the revised FLCAS consisted of 28 items" (p. 1594). No

Chinese version is available; no explanation is given as to what revisions had been made and why. Therefore, readers cannot possibly assess the suitability of this instrument. In contrast, some interview questions used in Hwang et al. (2024) may be misleading such as "Do you like to use Smart UEnglish to practice English? Why?" and "Which function of Smart UEnglish/UEnglish do you like most? Why?" (p. 1644). These two questions imply that the interviewees liked to use Smart UEnglish/UEnglish. The first question is not suitable especially for the interviewees from the control group in that the UEnglish they used was without smart mechanisms. Equally inappropriate for the control group interviewees are two other questions – "Do you want to keep using Smart UEnglish to practice English? Why?" and "Do you think using Smart UEnglish is helpful to your English? Why?" (Hwang et al. 2024, p. 1644). The instruments applied in both studies were used without passing through the stage of review and piloting.

Research instruments need to be carefully designed and rigorously reviewed and piloted before being used and readers should be able to access full contents conveniently. On the other hand, many of the less frequent sources may provide valuable data about pedagogical and psychological change in language learning, in particular those that reflect the participants' actual performance or mental activities. In other words, more studies using an interpretative and naturalist approach to data should be encouraged (Zhang et al. 2023).

4.2.7 Insufficient attention to ethical issues

Given that 40 % of the empirical research articles do not include ethical statements, this also constitutes a weakness. Ethical issues arise "in all research designs involving human respondents owing to an intrinsic tension between the needs of the researcher to collect personal data on which to base generalizations and the rights of the participants to maintain their dignity and privacy", in particular when there is a fiduciary relationship between the researcher(s) and the participants, for example, teacher-student relationship (Ferguson et al. 2004, p. 57). Therefore, ethical concerns, "unless properly addressed, may lead to resistance from participants and consequently data inaccuracy" (Peng and Xiao 2022, p. 12). On the other hand, although 60 % of our samples explain how ethical issues are addressed, the measures taken may not be adequate in all cases. For example, simply obtaining informed consent from participants "who are in dependent or restricted relationships with the researcher" is far from acceptable because there may be "a coercive element" in it (Ferguson et al. 2004, p. 58). This is especially the case with those participants who are minors and may be more vulnerable to pressure from the researchers or do not

really know what they consent to. In the so-called intelligent age, informed consent often becomes "the pretext to justify the misuse of educational data" in particular when "people are forced to give their consent if they want to use" a smart app, software or device (Xiao et al. 2025, p. 6).

Equally insufficient are measures such as material or monetary rewards or giving extra score as a bonus. For example, in a study investigating the effects of individual versus collaborative processing of ChatGPT-generated assessment feedback, "to maintain students' engagement and motivation in the project, the rated writing products were used as assignments for the grading of the course *Intermediate English Writing*, a compulsory course for all sophomore EFL students at the university. Outstanding performance in seeking, processing, and using ChatGPT-generated AF was awarded bonus marks" (Yan 2024, p. 5). This not only reflects the researcher's stance/bias on the assumed positive impact of ChatGPT on student learning, exacerbating social desirability bias but is also ethically controversial because the non-participants were not fairly treated. An excellent example of how ethical concerns are addressed is Jensen (2024), a case study of an 11-year-old girl. In addition to registering the study with the relevant department of the home university and following the university's ethical guidelines, the researcher sought informed consent from the girl's parents and informed assent from the participant herself. Further, the participant and other people involved were anonymized. Another good example is Roy (2024) whose participants were 8th-grade students. In other words, measures to handle ethical issues should be specific and relevant to the participants of a study.

4.2.8 Over-favoring case study and survey research in non-experimental research

Case study (46.9 %) and survey research (29.7 %) account for nearly 80 % of non-experimental design studies, with the remaining nine designs combined representing less than a quarter (see Table 6). This is also a weakness. As argued in Peng and Xiao (2022), educational research may benefit from "designs such as action research, design-based research, narrative research, grounded theory research, phenomenological research" (p. 11). Given the longitudinal and situated nature of language learning, innovations are needed in CALL research design. A mixture of designs in a study may be a worthwhile direction for CALL research. For example, both narrative and observational research designs may be adopted in a study and so are case study and phenomenological research designs; experimental designs may be embedded into a design-based research study.

4.2.9 Use of ambiguous units of measurement

Finally, when it comes to duration of intervention/treatment, the greatest problem is not that a significant percentage of empirical studies do not indicate time span as in the case of Shi and Aryadoust (2024) which finds that 42.9 % of the studies on AI-based automated written feedback research fail to specify the duration of their intervention/treatment. There are only 8.7 % of studies in our review which do not detail how long the intervention/treatment lasted. The biggest problem with our samples is the use of units of measurement which may be context-sensitive but not further specified. For example, how long is a semester, course, module, class, session, and meeting? As pointed out by Peng and Xiao (2022), the lengths of these units may vary "a great deal across different countries or educational institutions, and even within an institution" (p. 7). Specification of the duration is essential to enhancing replicability of a study.

5 Concluding remarks: implications from the findings and beyond

Overall, CALL research has its strengths and weaknesses in terms of research design. Strengths identified are clear statements of context of study and research purpose, question and/or hypothesis; increasing adoption of mixed-method approach; popularity of experimental research; sound theoretical foundation; multiple data sources (triangulation); specific accounts of data collection and analysis; and acknowledgement of limitations. Aspects which are found to be weak or far from rigorous include predominance of micro-level research; scarcity of replication studies; insensitivity to researcher biases; overwhelming dominance of non-probabilistic, convenience sampling; shortage of longitudinal research; undue reliance on self-reported data and unverified and/or unavailable instruments; insufficient attention to ethical issues; over-favoring case study and survey research in non-experimental research; and the use of ambiguous units of measurement. Both strengths and weaknesses are discussed in depth in Section 4 above, including room for further improvement concerning each strength. The discussion may be taken as caveat emptor for readers and guidelines/checklist for researchers. This concluding section will reiterate some fundamental issues pertaining to the findings as well as related broader issues, issues that can serve as future research directions. The first three lines of research center on macro and meso-level issues and the last three lines on micro-level issues.

The first line of research is theory building. Compared with research on technology-enhanced education in general, the subfield of CALL research is much

better informed theoretically, mainly by second language acquisition theories, linguistic theories, and pedagogical, learning or educational theories (Akayoğlu 2019; Hubbard 2008; Mohsen et al. 2024; Shadiev and Yu 2024). Nonetheless, there is no dedicated CALL theory (Hubbard 2020) despite Oller's (2013) attempt to build such a one. Given the contribution of theory building to a field of inquiry, building a grand CALL theory should be put on the research agenda.

The second line of research is undertaking cost-effectiveness and affordability research. Low cost, high quality and wider access are what drive the use of technology in education (Xiao 2023b). A review of over 3,000 studies on technology application in learning and teaching shows that only about 9 % took into account cost-effectiveness and/or accessibility in their research designs (Xiao 2023b). Furthermore, none of these studies examined the costs that students had to bear, that is, whether the technology used was affordable to students in actual life. It is the same case with our current review. None of the studies considered the variable of cost in their research designs. Unless CALL is cost effective and affordable to educational institutions and individual students alike, it is only feasible in the idealized, controlled experimental conditions, hence no accessibility and equity to speak of eventually. Research of the like is of limited practical relevance even if it can enhance learning effectiveness.

The third line of research is conducting research into institution-wide issues. For example, what are the implications of the institution-wide implementation of CALL for an institution's management and administration? How does the institution-wide implementation of CALL impact on the financial management and eventually business model of an institution? What policy or mechanism should be in place to evaluate and assure the quality of CALL? What professional development opportunities should be provided for faculty and staff in relation to CALL? This kind of research is needed to ensure that all stakeholders in an institution will work in synergy with each other and give full play to CALL.

The fourth line of research is rethinking the relationship between technology and language learning. One of the reasons for the scarcity of replication studies is said to be rapid developments in technology because "often, older technologies are completely replaced by newer versions and cannot even be accessed or used any longer" (Chun 2012, p. 596). So, even though "the older technology is superior to the newer technology" (Chun 2012, p. 595) and we have yet to fully understand "how the older technology could be used most effectively", "one is compelled to use a newer technology" (p. 596). Is it true that we have no choice but to use a newer technology? The history of technology-enhanced education tells a different story. "Educational reform has become a race for new technologies…what people are doing is to reform education to ensure a particular new technology will be used" rather than thinking about which educational problem needs to be fixed by which technology, be it old or

new (Xiao and Bozkurt 2025, p. 35). When it comes to the use of technology in education, the newer is not necessarily the better. A technology which is fit for a particular purpose is the best for that purpose. Strictly speaking, there is no technology which is outmoded.

The fifth line of research is conducting replicative, multi-case longitudinal research. A single case/context study needs to be replicated in other cases/contexts while being longitudinal also means being iterative by "involving different learners and instructors in different learning environments and with different learning objectives and domains of knowledge" (Xiao 2023a, p. 215) instead of being limited to collecting data over a period of time such as post-test. This is an effective way to scale up a study and generalize the findings. The dominance of short-term intervention, single-case, cross-sectional studies in technology-enhanced education (Reeves and Lin 2020; Scully et al. 2018; Song and Xiao 2017) needs to be challenged; CALL research is no exception.

The sixth line of research is maintaining a positive stance on failures. Educational technology research "is as much about investigating the imperfect 'state of the actual' as it is about exploring the perfected 'state-of-the-art'" (Selwyn 2012, p. 216). Nevertheless, it seems that the purpose of research is to find out what works, not what does not work and so only successes are worth publishing (Bulfin et al. 2014; Reeves and Lin 2020). This has become an unwritten rule in academia. Researchers even wonder whether we can talk about failures (Prinsloo 2018). Generally speaking, scientific research is a designed endeavor, which means that the chances of success are much higher than those of failure. However, failures are unavoidable; there are times when things do not go as planned. Common sense tells us that lessons from failures are as valuable as best practices. Only reporting successes and ignoring or even choosing to ignore failures will definitely distort the CALL research landscape, doing more harm than good. No failure in CALL research is reported in our samples as well as in other reviews.

In summary, we hope that future CALL research will center on broader meso and macro-level issues while continuing to investigate micro-level learning and teaching practice. This will bring CALL research to a higher level, accelerating the full maturity of the field.

Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Research funding: None declared.
Ethical approval: Not applicable.

References

Akayoğlu, Sedat. 2019. Theoretical frameworks used in CALL studies: A systematic review. *Teaching English with Technology* 19(4). 104–118.

Alharbi, Wael. 2024. Mind the gap, please!: Addressing the mismatch between teacher awareness and student AI adoption in higher education. *International Journal of Computer-Assisted Language Learning and Teaching* 14(1). 1–28.

Anderson, Terry. 2003. Getting the mix right again: An updated and theoretical rationale for interaction. *The International Review of Research in Open and Distance Learning* 4(2). https://doi.org/10.19173/irrodl.v4i2.149.

Badem-Korkmaz, Fatma & Ufuk Balaman. 2024. Eliciting student participation in video-mediated EFL classroom interactions: Focus on teacher response-pursuit practices. *Computer Assisted Language Learning* 37(7). 1881–1903.

Ballance, Oliver J. 2024. Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation. *ReCALL* 36(1). 58–71.

Barkhuizen, Gary P. 2009. Topics, aims, and constraints in English teacher research: A Chinese case study. *TESOL Quarterly* 43(1). 113–125.

Bartolomé, Antonio, Linda Castañeda & Jordi Adell. 2018. Personalisation in educational technology: The absence of underlying pedagogies. *International Journal of Educational Technology in Higher Education* 15(14). https://doi.org/10.1186/s41239-018-0095-0.

Bernard, Robert M., Philip C. Abrami, Eugene F. Borokhovski, C. Anne Wade, Rana M. Tamim, Michael A. Surkes & Edward C. Bethel. 2009. A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research* 79. 1243–1289.

Bikowski, Dawn & Mathias Schulze. 2015. Replication and evaluation in CALL. *CALICO Journal* 32(2). i–v.

Blázquez-Carretero, Miguel. 2023. Building a pedagogic spellchecker for L2 learners of Spanish. *ReCALL* 35(3). 321–338.

Bond, Melissa, Olaf Zawacki-Richter & Mark Nichols. 2019. Revisiting five decades of educational technology research: A content and authorship analysis of the British Journal of Educational Technology. *British Journal of Educational Technology* 50(1). 12–63.

Bulfin, Scott, Michael Henderson, Nicola F. Johnson & Neil Selwyn. 2014. Methodological capacity within the field of "educational technology" research: An initial investigation. *British Journal of Educational Technology* 45(3). 403–414.

Burston, Jack, Androulla Athanasiou & Konstantinos Giannakou. 2024. Quantitative experimental L2 acquisition MALL studies: A critical evaluation of research quality. *ReCALL* 36(1). 22–39.

Cardwell, Ramsey, Ben Naismith, Jill Burstein, Steven Nydick, Sarah Goodwin & Anthony Verardi. 2024. From pen to pixel. *CALICO Journal* 41(2). 209–2034.

Chen, Hsieh & Ju S. Lee. 2023. Digital storytelling outcomes, emotions, grit, and perceptions among EFL middle school learners: Robot-assisted versus PowerPoint-assisted presentations. *Computer Assisted Language Learning* 36(5–6). 1088–1115.

Chen, Lijuan & Claire Rodway. 2023. Distance students' language learning strategies in asynchronous and synchronous environments. *CALICO Journal* 40(2). 198–217.

Chong, Sin W., Melissa Bond & Hamish Chalmers. 2024. Opening the methodological black box of research synthesis in language education: Where are we now and where are we heading? *Applied Linguistics Review* 15(4). 1557–1568.

Choubsaz, Yazdan, Alireza Jalilifar & Alex Boulton. 2024. A longitudinal analysis of highly cited papers in four CALL journals. *ReCALL* 36(1). 40–57.

Chun, Dorothy M. 2012. Replication studies in CALL research. *CALICO Journal* 29(4). 591–600.
Creswell, John W. 2012. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*, 4th ed. Boston, MA: Pearson.
DeMolder, Jessica, David Wiseman, Charles Graham & Camellia Hill. 2023. Toward blended language learning frameworks: A systematic review. *CALICO Journal* 40(2). 218–237.
Díez-Arcón, Paz & Nikoletta Agonács. 2024. Conceptualising language MOOC diversity: The creation of a defined taxonomy through the analysis of course indicators. *ReCALL* 36(3). 324–342.
Ekmekçi, Emrah. 2023. Pursuing a standardized content of a CALL course for pre-service EFL teachers: The procedure, impacts, and reflections. *Computer Assisted Language Learning* 36(5–6). 1005–1039.
Feng, Baoxin & Lee-Luan Ng. 2024. The spatial influence on vocabulary acquisition in an immersive virtual reality-mediated learning environment. *International Journal of Computer-Assisted Language Learning and Teaching* 14(1). 1–17.
Ferguson, Linda M., Olive Yonge & Florence Myrick. 2004. Students' involvement in faculty research: Ethical and methodological issues. *International Journal of Qualitative Methods* 3(4). 56–68.
Foung, Dennis & Lucas Kohnke. 2023. Beyond replication: An exact replication study of Łodzikowski (2021). *ReCALL* 35(2). 225–238.
García-Pastor, María D. & Jorge Piqueres Calatayud. 2023. Crafting L2 multimodal composing identities: A study with secondary EFL learners. *CALICO Journal* 40(3). 313–334.
Gass, Susan, Shawn Loewen & Luke Plonsky. 2021. Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching* 54(2). 245–258.
Gok, Duygu, Hilal Bozoglan & Bahadir Bozoglan. 2023. Effects of online flipped classroom on foreign language classroom anxiety and reading anxiety. *Computer Assisted Language Learning* 36(4). 840–860.
Han, Jinghe, Qiaoyun Liu & Ruiyan Sun. 2023. A multimodal approach to teaching Chinese as a Foreign Language (CFL) in the digital world. *International Journal of Computer-Assisted Language Learning and Teaching* 13(1). 1–16.
Hew, Khe F., Min Lan, Ying Tang, Chengyuan Jia & Chung K. Lo. 2019. Where is the "theory" within the field of educational technology research? *British Journal of Educational Technology* 50(3). 956–971.
Horwitz, Elaine K., Michael B. Horwitz & Joann Cope. 1986. Foreign language classroom anxiety. *The Modern Language Journal* 70(2). 125–132.
Hsieh, Hsiu-Fang & Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15(9). 1277–1288. https://10.1177/1049732305276687.
Hubbard, Philip. 2008. Twenty-five years of theory in the CALICO Journal. *CALICO Journal* 25(3). 387–399.
Hubbard, Philip. 2009. A general introduction to computer assisted language learning. In Philip Hubbard (ed.), *Computer assisted language learning*, 1–20. London: Routledge.
Hubbard, Philip. 2020. An invitation to CALL: Foundations of computer-assisted language learning (Unit 6: CALL theory and research). Available at: https://web.stanford.edu/~efs/callcourse2/Invitation-to-CALL-Unit6.pdf.
Hwang, Wu-Yuin, Bo-Chen Guo, Anh Hoang, Ching-Chun Chang & Nien-Tsu Wu. 2024. 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.
Jakonen, Teppo & Heidi Jauni. 2024. Managing activity transitions in robot-mediated hybrid language classrooms. *Computer Assisted Language Learning* 37(4). 872–895.
Jenkins, Andrew, Rodie Akerman, Lara Frumkin, Emma Salter & John Vorhaus. 2011. *Literacy, numeracy and disadvantage among older adults in England*. Institute of Education, University of London. Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2019/11/Older20adults20and20writing.pdf.

Jensen, Signe Hannibal. 2024. Doing being on social media: “pls like, comment, and subscribe!”. *CALICO Journal* 41(1). 25–47.

Jiang, Xiangying, Ryan Peters, Luke Plonsky & Bozena Pajak. 2024. The effectiveness of Duolingo English courses in developing reading and listening proficiency. *CALICO Journal* 41(3). 249–272.

Jung, Insung. 2025. *Pathways to international publication in the social sciences: A guide for early career and non-native English researchers*. Singapore: Springer.

Karimi, Honeiah, David Joshua Sañosa, Kevin Hernandez Rios, Phoebe Tran, Dorothy M. Chun, Richert Wang & Diana J. Arya. 2023. Building a city in the sky: Multiliteracies in immersive virtual reality. *CALICO Journal* 40(1). 24–44.

Kessler, Matt. 2023. Supplementing mobile-assisted language learning with reflective journal writing: A case study of Duolingo users’ metacognitive awareness. *Computer Assisted Language Learning* 36(5–6). 1040–1063.

Kim, YouJin & Yoon Namkung. 2024. Methodological characteristics in technology-mediated task-based language teaching research: Current practices and future directions. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190524000096 (Epub ahead of print).

Klímová, Blanka & Prodhan Mahbub Ibna Seraj. 2023. The use of chatbots in university EFL settings: Research trends and pedagogical implications. *Frontiers in Psychology* 14. 1131506.

Kourtali, Nektaria-Efstathia & Lais Borges. 2024. The effects of feedback timing on L2 development in written SCMC. *Computer Assisted Language Learning* 37(8). 2291–2319.

Lee, Sangmin-Michelle. 2023. The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning* 36(1–2). 103–125.

Lee, Ju Seong & Ying Lu. 2023. L2 motivational self-system and willingness to communicate in the classroom and extramural digital contexts. *Computer Assisted Language Learning* 36(1–2). 126–148.

Li, Jia, Linying Ji & Qizhen Deng. 2023. The heterogeneous and transfer effects of a texting-based intervention on enhancing university English learners’ vocabulary knowledge. *Computer Assisted Language Learning* 36(1–2). 52–80.

Li, Yan, Xinyan Zhou & Thomas K. F. Chiu. 2024. Systematics review on artificial intelligence chatbots and ChatGPT for language learning and research from self-determination theory (SDT): What are the roles of teachers? *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2024.2400090 (Epub ahead of print).

Loncar, Michael, Wayne Schams & Jong-Shing Liang. 2023. Multiple technologies, multiple sources: Trends and analyses of the literature on technology-mediated feedback for L2 English writing published from 2015–2019. *Computer Assisted Language Learning* 36(4). 722–784.

Luo, Yujuan & Mike Watts. 2024. Exploration of university students’ lived experiences of using smartphones for English language learning. *Computer Assisted Language Learning* 37(4). 608–633.

Lyu, Boning & Chun Lai. 2024. Analyzing learner engagement with native speaker feedback on an educational social networking site: An ecological perspective. *Computer Assisted Language Learning* 37(1–2). 114–148.

McManus, Kevin. 2024. The future of replication in applied linguistics: Toward a standard for replication studies. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190524000011 (Epub ahead of print).

Meihami, Hussein. 2024. Investigating CALL researchers’ strategies to conduct genuine CALL research: A community of practice perspective. *Computer Assisted Language Learning* 37(3). 307–332.

Mendes de Oliveira, Milene, Zachary Sporn, Lea Kliemann, Alexandra Borschke & Meike Meyering. 2023. Online language learning and workplace communication: A study on Babbel’s virtual-classroom solution. *Computer Assisted Language Learning* 36(8). 1501–1527.

Mohsen, Mohammed Ali, Sultan Althebi, Rawan Alsagour, Albatool Alsalem, Amjad Almudawi & Abdulaziz Alshahrani. 2024. Forty-two years of computer-assisted language learning research: A scientometric study of hotspot research and trending issues. *ReCALL* 36(2). 230–249.

Oller, John W. 2013. Toward a theory of technologically assisted language learning/instruction. *CALICO Journal* 13(4). 19–43.

Peng, Yiwei & Junhong Xiao. 2022. Is the empirical research we have the research we can trust? A review of distance education journal publications in 2021. *Asian Journal of Distance Education* 17(2). 1–18. http://asianjde.com/ojs/index.php/AsianJDE/article/view/659.

Peterson, Mark. 2023. Digital simulation games in CALL: A research review. *Computer Assisted Language Learning* 36(5–6). 943–967.

Polio, Charlene & Susan Gass. 1997. Replication and reporting: A commentary. *Studies in Second Language Acquisition* 19(4). 499–508.

Prinsloo, Paul. 2018. 反思 2017 年多伦多在线学习世界大会: 我听到的和没有听到的 [What I heard and what I did not hear: Reflections on the world conference on online learning, Toronto, 2017]. *Distance Education in China* 12. 5–11.

Reeves, Thomas C. & Lin Lin. 2020. The research we have is not the research we need. *Educational Technology Research and Development* 68. 1991–2001.

Richards, Paul. 2024. Pragmatic feedback on refusals in a computer-simulated advising session. *Language Learning & Technology* 28(1). 1–26. https://hdl.handle.net/10125/73549.

Roy, Abhipriya. 2024. Impact of digital storytelling on motivation in middle school English classrooms. *International Journal of Computer-Assisted Language Learning and Teaching* 14(1). 1–20.

Scully, Darina, Michael O'Leary & Mark Brown. 2018. *The learning portfolio in higher education: A game of snakes and ladders*. Dublin City University, Centre for Assessment Research Policy & Practice in Education (CARPE), and National Institute for Digital Learning (NIDL). Available at: https://www.dcu.ie/sites/default/files/inline-files/Learning%20Portfolios%20in%20Higher%20Education%202018.pdf.

Selwyn, Neil. 2012. Ten suggestions for improving academic research in education and technology. *Learning, Media and Technology* 37(3). 213–219.

Shadiev, Rustam & Jiatian Yu. 2024. Review of research on computer-assisted language learning with a focus on intercultural education. *Computer Assisted Language Learning* 37(4). 841–871.

Shi, Huawei & Vahid Aryadoust. 2024. A systematic review of AI-based automated written feedback research. *ReCALL* 36(2). 187–209.

Shortt, Mitchell, Shantanu Tilak, Irina Kuznetcova, Bethany Martens & Babatunde Akinkuolie. 2023. Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning* 36(3). 517–554.

Simonson, Michael, Charles Schlosser & Anymir Orellana. 2011. Distance education research: A review of the literature. *Journal of Computing in Higher Education* 23. 124–142.

Smith, Bryan & Mathias Schulze. 2013. Thirty years of the CALICO Journal – Replicate, replicate, replicate. *CALICO Journal* 30(1). i–iv.

Smith, Sara A., María Soledad Carlo, Sanghoon Park & Howard Kaplan. 2023. Exploring the promise of augmented reality for dual language vocabulary learning among bilingual children: A case study. *CALICO Journal* 40(1). 91–112.

Song, Yilin & Junhong Xiao. 2017. 再谈移动学习—访英国移动学习教授约翰·特拉克斯勒 [Mobile learning revisited—An interview with Professor John Traxler]. *Distance Education in China* 11. 43–46.

Soyoof, Ali, Barry Lee Reynolds, Boris Vazquez-Calvo & Katherine McLay. 2023. Informal digital learning of English (IDLE): A scoping review of what has been done and a look towards what is to come. *Computer Assisted Language Learning* 36(4). 608–640.

Tai, Tzu-Yu. 2024. Effects of intelligent personal assistants on EFL learners' oral proficiency outside the classroom. *Computer Assisted Language Learning* 37(5–6). 1281–1310.

Timpe-Laughlin, Veronika, Tetyana Sydorenko & Judit Dombi. 2024. Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning* 37(1–2). 149–178.

Tschichold, Cornelia. 2023. Replication in CALL. *ReCALL* 35(2). 139–142.

van Drimmelen, Tom, M. Nienke Slagboom, Ria Reis, Lex M. Bouter & Jenny T. van der Steen. 2024. Decisions, decisions, decisions: An ethnographic study of researcher discretion in practice. *Science and Engineering Ethics* 30. 59.

Vehovar, Vasja, Vera Toepoel & Stephanie Steinmetz. 2016. Non-probability sampling. In Christof Wolf, Dominique Joye, Tom W Smith & Yang-chih Fu (eds.), *The Sage handbook of survey methods*, 329–345. Thousand Oaks, CA: Sage Publications.

Vuogan, Alyssa & Shaofeng Li. 2024. A systematic review of meta-analyses in second language research: Current practices, issues, and recommendations. *Applied Linguistics Review* 15(4). 1621–1644.

Wach, Aleksandra, Shannon Tanghe & De Zhang. 2024. Multinational telecollaboration in language teacher education: Teacher educators' perspectives. *Language Learning & Technology* 28(1). 1–13. https://hdl.handle.net/10125/73590.

Wen, Yiran, Jian Li, Hongkang Xu & Hanwen Hu. 2023. Restructuring multimodal corrective feedback through Augmented Reality (AR)-enabled videoconferencing in L2 pronunciation teaching. *Language Learning & Technology* 27(3). 83–107. https://hdl.handle.net/10125/73533.

Werth, Eric & Katherine Williams. 2021. What motivates students about open pedagogy? Motivational regulation through the lens of self-determination theory. *The International Review of Research in Open and Distributed Learning* 22(3). 34–54.

Xiao, Feiwen, Priscilla Zhao, Hanyue Sha, Dandan Yang & Warschauer Mark. 2024. Conversational agents in language learning. *Journal of China Computer-Assisted Language Learning* 4(2). 300–325.

Xiao, Junhong. 2017. Learner-content interaction in distance education: The weakest link in interaction research. *Distance Education* 38(1). 123–135.

Xiao, Junhong. 2023a. Critical issues in open and distance education research. *The International Review of Research in Open and Distributed Learning* 24(2). 213–228.

Xiao, Junhong. 2023b. Critiquing sustainable openness in technology-based education from the perspective of cost-effectiveness and accessibility. *Open Praxis* 15(3). 244–254.

Xiao, Junhong & Aras Bozkurt. 2025. Prophets of progress: How do leading global agencies naturalize enchanted determinism surrounding artificial intelligence for education? *Journal of Applied Learning & Teaching* 8(1). 28–40.

Xiao, Junhong, Aras Bozkurt, Mark Nichols, Angelica Pazurek, Christian M. Stracke, John Y. H. Bai, Robert Farrow, Dónal Mulligan, Chrissi Nerantzi, Ramesh Chander Sharma, Lenandlar Singh, Isak Frumin, Andrew Swindell, Sarah Honeychurch, Melissa Bond, Jon Dron, Stephanie Moore, Jing Leng, Patricia J. Slagter van Tryon, Manuel Garcia, Evgeniy Terentev, Tlili Ahmed, Thomas K. F. Chiu, Charles B. Hodges, Petar Jandrić, Alexander Sidorkin, Helen Crompton, Stefan Hrastinski, Apostolos Koutropoulos, Mutlu Cukurova, Peter Shea, Steven Watson, Kai Zhang, Kyungmee Lee, Eamon Costello, Mike Sharples, Anton Vorochkov, Bryan Alexander, Maha Bali, Robert L. Moore, Olaf Zawacki-Richter, Tutaleni Iita Asino, Henk Huijser, Chanjin Zheng, Sunagül Sani-Bozkurt, Josep M. Duart & Chryssa Themeli. 2025. Venturing into the unknown: Critical insights into grey areas and pioneering future directions in educational generative AI research. *TechTrends* 69. 582–597.

Yan, Da Alex. 2024. Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology* 28(1). 1–19. https://hdl.handle.net/10125/73597.

Yang, Yu-Fen, Alexis P. I. Goh, Yi-Chun Hong & Nian-Shing Chen. 2023. Primary school students' foreign language anxiety in collaborative and individual digital game-based learning. *Computer Assisted Language Learning* 36(8). 1587–1607.
Zawacki-Richter, Olaf. 2009. Research areas in distance education: A Delphi study. *The International Review of Research in Open and Distributed Learning* 10(3). https://doi.org/10.19173/irrodl.v10i3.674.
Zawacki-Richter, Olaf & Terry Anderson (eds.). 2014. *Online distance education – Towards a research agenda*. Athabasca University Press. Available at: http://www.aupress.ca/index.php/books/120233.
Zawacki-Richter, Olaf, Victoria I. Marín, Melissa Bond & Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education* 16. 39.
Zhai, Chunpeng & Santoso Wibowo. 2023. A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence* 4. 100134.
Zhang, Yining & Ruoxi Sun. 2023. LMOOC research 2014 to 2021: What have we done and where are we going next? *ReCALL* 35(3). 356–371.
Zhang, Meixiu, Miriam Akoto & Mimi Li. 2023. Digital multimodal composing in post-secondary L2 settings: A review of the empirical landscape. *Computer Assisted Language Learning* 36(4). 694–721.
Zhang, Huiwan, Wei Wei & Yiqian Cao. 2024. Using computer-assisted language teaching technologies to develop multiple-levels of medical vocabulary knowledge for second language medical students. *Computer Assisted Language Learning* 37(7). 2007–2027.
Zheng, Chunping, Miao Yu, Zhiyan Guo, Hanyong Liu, Mengya Gao & Ching Sing Chai. 2022. Review of the application of virtual reality in language education from 2010 to 2020. *Journal of China Computer-Assisted Language Learning* 2(2). 299–335.

Bionotes

Junhong Xiao
The Open University of Shantou, 8 Leshan Road, Shantou, Guangdong, 515000, China
frankxjh@outlook.com
https://orcid.org/0000-0002-5316-2957

Junhong Xiao is Emeritus Professor at the Open University of Shantou, China. His research interests include open and digital (language) learning, digital (language) learning, and open, distance, and digital education theory.

Yiwei Peng
The Open University of Shantou, 8 Leshan Road, Shantou, Guangdong, 515000, China
diana66912@yahoo.com
https://orcid.org/0000-0001-5341-2568

Yiwei Peng is associate professor at the Open University of Shantou, China. Her research interests include distance language learning and teaching and learner support.