




REVIEW

Exploring a Decade of Research: A Systematic Review of Computer-Based English Speaking Tests

Hengzhi Hu¹ , Qiuyu Gong¹ , Nur-Ehsan Mohd-Said^{1,2*} 

¹ Faculty of Education, Universiti Kebangsaan Malaysia, Bangi Selangor 43600, Malaysia

² Centre for Shaping Advanced & Professional Education, Universiti Kebangsaan Malaysia, Selangor 43600, Malaysia

ABSTRACT

The rapid integration of technology into educational assessment has revolutionized the evaluation of English speaking proficiency. Computer-based English speaking tests (CBESTs) have emerged as scalable and efficient solutions, which offer enhanced consistency and accessibility in high-stakes and large-scale testing contexts. However, existing studies on CBESTs have primarily focused on specific aspects of their design, implementation, and impact, leaving a fragmented understanding of their broader implications. As such, this systematic review synthesizes empirical research on CBESTs published between 2014 and 2024 to identify overarching trends, challenges, and opportunities. Employing the PRISMA methodology, the review analyzed 36 studies identified from three databases: Web of Science, Scopus, and Google Scholar. The findings highlight diverse research foci, including advancements in automated scoring, test validity, and the influence of cognitive and affective factors on performance. Studies also explored test-taker perceptions and experiences, which revealed mixed attitudes toward fairness and authenticity. Research methodologies ranged from quantitative correlational studies and qualitative case studies to mixed-methods designs, reflecting a diverse yet fragmented body of work. The review highlights the need for continued innovation in CBEST design and emphasizes the importance of hybrid models that integrate automation with human judgment. For test developers and policymakers, the findings underscore the importance of equitable implementation, technical refinement, and alignment with pedagogical goals. Future research should explore underrepresented areas such as long-term learning impacts and broader inclusivity to enhance the utility

*CORRESPONDING AUTHOR:

Nur-Ehsan Mohd-Said, Faculty of Education, Universiti Kebangsaan Malaysia, Bangi Selangor 43600, Malaysia; Centre for Shaping Advanced & Professional Education, Universiti Kebangsaan Malaysia, Selangor 43600, Malaysia; Email: nurehsan@ukm.edu.my

ARTICLE INFO

Received: 5 March 2025 | Revised: 4 April 2025 | Accepted: 11 April 2025 | Published Online: 14 April 2025

DOI: <https://doi.org/10.30564/fls.v7i4.8978>

CITATION

Hu, H., Gong, Q., Mohd-Said, N.-E., 2025. Exploring a Decade of Research: A Systematic Review of Computer-Based English Speaking Tests. *Forum for Linguistic Studies*. 7(4): 788–803. DOI: <https://doi.org/10.30564/fls.v7i4.8978>

COPYRIGHT

Copyright © 2025 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

and fairness of CBESTs.

Keywords: Computer-Based Tests; Assessment; English Speaking; Systematic Review

1. Introduction

The rapid advancement of technology has revolutionized educational assessment and offered unprecedented opportunities to enhance fairness, accessibility, and efficiency. Nowhere is this transformation more evident than in the domain of language assessment, particularly in evaluating English as a lingua franca. English serves as a global means of communication, connecting individuals in academic, professional, and social contexts^[1]. As its role continues to expand, the demand for innovative, scalable, and reliable assessment methods has grown exponentially^[2]. Technology, acting as a pivotal force, has fundamentally reshaped how English proficiency is measured and validated and challenged traditional methods of language evaluation.

Among the technological advancements, computer-based testing (CBT) has emerged as a cornerstone of modern language assessment practices. Promising enhanced consistency, scalability, and accessibility^[3, 4], CBT has been widely adopted across diverse contexts. In large-scale standardized tests such as the Test of English as a Foreign Language (TOEFL), Pearson Test of English Academic (PTE Academic), and Duolingo English Test, CBT leverages automated scoring and adaptive testing mechanisms^[5]. Beyond standardized examinations, CBT is gaining traction in institutional and classroom-based assessments, which supports individualized learning pathways and formative evaluations^[6, 7]. These developments reflect the versatility of CBT in meeting the diverse needs of educators and learners to foster equitable access to reliable language assessment tools.

Speaking, one of the most complex components of language use, presents unique challenges for assessment, as it is a real-time skill that involves immediate processing and production of language^[8, 9], the complexities of which are further magnified when assessments aim to capture authentic communicative abilities that often involve dynamic, unpredictable, and context-dependent language use^[10]. Computer-based English speaking tests (CBESTs) have emerged as a solution, which incorporate various approaches to evaluate speaking proficiency. CBESTs can be broadly categorized

into various types based on their design, purpose, and implementation. These include automated speaking tests, semi-automated tests, and human-rated computer-assisted tests. Each type leverages technology to address specific aspects of speaking assessment, such as accuracy, fluency, coherence, and interactional competence^[11, 12]. While automated tests often rely on speech recognition and natural language processing technologies for scoring^[13, 14], semi-automated and human-rated tests integrate human judgment to enhance reliability and validity^[15].

However, challenges persist in ensuring the validity, reliability, and fairness of CBESTs. Technological limitations, such as inaccuracies in speech recognition systems when handling diverse accents, speech patterns, and linguistic variations, can undermine the fairness of these assessments, particularly for non-native speakers from underrepresented backgrounds^[6]. Additionally, while automated scoring offers consistency, it may fail to account for nuanced aspects of communication, such as pragmatics, intonation, and cultural context, which are critical to authentic speaking proficiency^[15, 16]. The cost of developing and maintaining advanced CBEST platforms, coupled with disparities in access to technology, poses barriers to equitable implementation, especially in resource-limited settings^[3]. Furthermore, test-takers' unfamiliarity with computer-based interfaces or anxiety associated with digital assessments may affect their performance, introducing variability unrelated to their actual language ability^[17]. These challenges underscore the need for continued innovation and rigorous research to refine CBESTs to ensure they deliver accurate, fair, and contextually relevant evaluations of speaking proficiency.

In light of these complexities, research into CBESTs has expanded over the past years and delved into various aspects such as the development of automated scoring technologies^[18], the design of adaptive tasks to assess speaking proficiency^[3], the validity and reliability of the tests^[6], and stakeholder perceptions of the practical and ethical dimensions of CBEST implementation^[10, 17, 19]. However, while these investigations have provided valuable findings, they remain fragmented, and a notable gap in the literature is

the lack of a comprehensive review consolidating research on CBESTs conducted to identify overarching trends, challenges, and implications. To address this gap, this systematic review aims to synthesize empirical research on CBESTs published in the last decade, from 2014 to 2024. The review will explore the focus areas of previous studies, their research designs, as well as the effectiveness and challenges associated with CBESTs. By synthesizing a decade of research, this review seeks to provide a comprehensive understanding of the CBESTs, highlighting their potential to reshape English speaking assessment in diverse educational contexts.

2. Methodology

The systematic review adopted the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) approach^[20], which ensures transparency and comprehensiveness in the review process. This approach involves a structured methodology for identifying, screening, and selecting relevant studies to address the research objectives, as demonstrated in **Figure 1**.

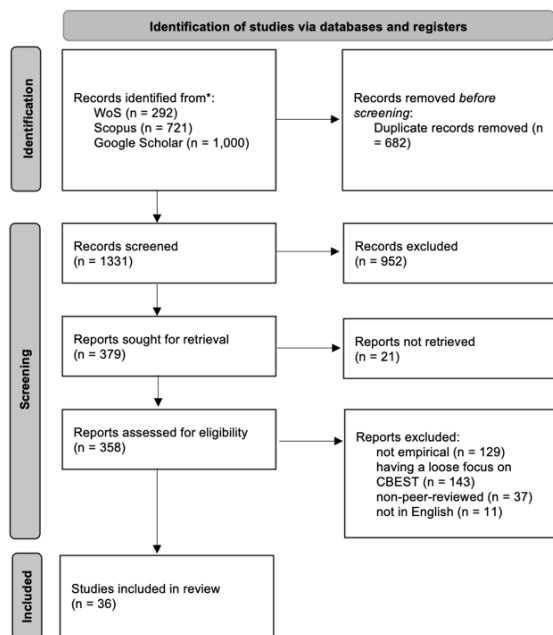


Figure 1. The Flowchart of PRISMA.

2.1. Identification

Three databases were included in the review: Web of Science (WoS), Scopus, and Google Scholar. WoS and Scopus were chosen because of their comprehensive coverage

of high-quality, peer-reviewed literature across multiple disciplines^[21], including educational technology and language assessment. These databases are particularly well-suited for identifying studies in well-established journals and conference proceedings. Google Scholar was used as a supplementary source to capture conference papers and other potentially relevant studies that might not be indexed in WoS or Scopus, ensuring a broader coverage of the topic^[21].

The search strings used for the databases were as follows:

- Web of Science (WoS): *AB = ("speaking" AND ("computer-based" OR "computer-assisted" OR "automated")) AND ("assessment" OR "evaluation" OR "examination" OR "test")) AND PY = (2014-2024)*
- Scopus: *TITLE-ABS-KEY ("speaking" AND ("computer-based" OR "computer-assisted" OR "automated")) AND ("assessment" OR "evaluation" OR "examination" OR "test")) AND PUBYEAR > 2013 AND PUBYEAR < 2025*
- Google Scholar: *"speaking" AND ("computer-based" OR "computer-assisted" OR "automated") AND ("assessment" OR "evaluation" OR "examination" OR "test") AND after:2013 AND before:2025*

Each search string was tailored to the specific syntax requirements of the respective database to ensure comprehensive and relevant results within the publication timeframe of 2014–2024. As a result, 292,721, and 1,000 records were identified from WoS, Scopus, and Google Scholar, respectively (*Note: For Google Scholar, due to its limitations in advanced search functionalities and precise filtering, a large number of results were generated. However, only the first 1,000 records are displayed, as this is the maximum result limit for Google Scholar*)^[21]. With duplicates (n = 682) removed from the initial search using the software Covidence, the remaining records were screened.

2.2. Screening

The remaining records (n = 1,331) were initially screened based on the title, abstract, and keywords. A total of 952 records were excluded during this stage as preliminary examination revealed they were not related to CBEST. The remaining records (n = 379) were then retrieved in full text. Of these, 21 could not be accessed, leaving 358 records for detailed review by the research team using a predefined set

of criteria. Following this review, 129 records were excluded as they did not present empirical evidence and were instead reviews or opinion articles; 143 records were excluded due to a loose focus on CBEST; 37 records were excluded because they were non-peer-reviewed literature, such as theses or reports; and 11 records were excluded as they were presented in a language other than English.

2.3. Included

Consequently, 36 records were included in the review. Before conducting thematic analysis, the records were organized using EndNote, and a panel of experts appraised their quality. The evaluation focused on whether the research presented demonstrated methodological robustness, with research designs and findings effectively addressing the stated research purpose. Studies that met these criteria were prioritized for deeper analysis to ensure the review's validity and reliability. As a result, all the records were deemed to be robust for the review. They are presented in **Appendix A Table A1**, following the matrix of research purpose, design, and findings^[22–58].

3. Findings

3.1. Areas of Focus in Previous Research

The body of research on the CBEST has explored a range of areas, from technical advancements and test validity to the cognitive and emotional factors influencing test-taker performance. These studies provide a rich foundation for understanding CBEST but also highlight areas where further exploration is warranted. One key focus has been the technical aspects of CBEST, including their validity and reliability. Researchers have extensively investigated the comparability of automated scoring systems with human raters. Findings suggest that while automated systems generally exhibit high inter-rater reliability, they may be slightly more lenient, particularly for low-proficiency speakers, compared to human examiners^[23]. Additionally, low bias rates observed in automated scoring indicate better internal consistency compared to human raters^[24]. Studies evaluating the validity of CBEST frameworks have demonstrated alignment with established communicative language testing criteria, supporting their use in tertiary education settings^[25]. Furthermore,

innovative machine-learning algorithms for fluency and pronunciation assessment have shown promising results, with fluency predictions achieving 94% accuracy and pronunciation predictions reaching 99.9%, setting new benchmarks in the literature^[26]. Such advancements underline the potential of automated scoring systems for reliable and efficient assessment in diverse educational contexts.

Another area of focus involves test-taker perceptions and experiences with CBEST. Studies have revealed mixed attitudes toward CBEST, with many test-takers acknowledging the format as valid and motivating. However, concerns about fairness and the lack of authentic communication persist^[27]. For example, in Thai university contexts, students demonstrated confidence in operating CBEST equipment but expressed reservations about its ability to replace traditional methods^[28]. Additionally, test-takers' perceptions of CBEST validity often influence their test-taking effort indirectly, mediated by their perception of the test's importance^[29]. Psychological factors, particularly speaking-related anxiety, have also been widely examined. High levels of anxiety, including communication apprehension and test anxiety, have been found to negatively impact test performance^[30]. Notably, fear of negative classroom feedback has emerged as a significant predictor of speaking-related anxiety in CBEST^[31].

Cognitive factors, such as strategic competence and response processes, have also been extensively studied. For instance, planning, problem-solving, monitoring, and evaluating have been identified as essential metacognitive strategies for CBEST test-takers, highlighting the importance of fostering these skills in test preparation^[32]. Eye-tracking studies have further revealed differences in how native and non-native speakers process test materials, with non-native speakers focusing more on countdown timers and native speakers prioritizing content-related features such as on-screen pictures^[33]. These findings emphasize the role of cognitive strategies in navigating CBEST tasks effectively.

Comparative studies examining different testing modes—such as face-to-face, computer-delivered, and emerging platforms such as the metaverse—have shed light on how mode effects influence performance. For example, metaverse-based testing has demonstrated significant potential, with participants performing better in this mode compared to traditional and video-conferencing formats^[34]. Test-

takers have also expressed a preference for the metaverse platform, citing its engaging and supportive features. However, other studies have found no significant performance differences between computer-delivered and face-to-face tests, though test-takers often preferred traditional formats for their perceived interactivity^[35]. In Vietnamese contexts, specific mode effects were observed in criteria such as content development and pronunciation, highlighting nuanced differences between test formats^[36].

Finally, linguistic and discourse features influencing CBEST performance have been explored. Studies have identified significant correlations between linguistic features, such as vocabulary indices, and speaking proficiency. Vocabulary emerged as the strongest predictor of proficiency, surpassing utterance and syntactic features^[37]. Discourse analysis has also revealed patterns in linguistic errors, with common issues including morphemes and subject-verb agreement^[38]. Additionally, research on the effects of pre-task and online planning has revealed nuanced relationships between complexity, fluency, and accuracy measures, indicating that planning time can enhance fluency and accuracy without significantly affecting test scores^[39].

3.2. Designs of Previous Research

The research methodologies employed in studies on the CBEST span quantitative, qualitative, and mixed-methods approaches, each with specific designs tailored to investigate various aspects of CBEST performance and applications. Quantitative research has been the most widely used approach, offering structured and statistical analyses to explore the factors influencing CBEST. Within this framework, correlational designs have been utilized to examine relationships between variables related to CBEST performance. For instance, researchers have investigated the relationship between speaking patterns, emotional states, and test grades using advanced models such as HuBERT and neural networks^[28], while others have focused on correlations between linguistic features such as vocabulary and syntax and speaking proficiency^[37]. Survey-based designs have also been prominent, with structured questionnaires used to assess test-taker perceptions, anxiety, and attitudes toward CBEST^[28, 30, 31]. These studies have measured speaking-related anxiety and its impact on performance, as well as test-takers' perceptions of validity and fairness^[27, 29]. Vali-

dation and reliability designs have further examined the credibility of CBEST, employing methods such as many-facet Rasch analysis to compare automated scoring with human ratings and confirm alignment with established communicative language testing frameworks^[24, 25]. Experimental and quasi-experimental designs have manipulated variables such as task design and planning conditions to evaluate their effects on performance, including studies on planning time and comparative analyses of different test modes, such as face-to-face, computer-delivered, and metaverse-based tests^[34, 36, 39].

Qualitative research approaches have contributed in-depth insights into test-taker experiences and perceptions. Case study designs have explored specific contexts, such as Chinese high school students' views on CBEST validity and suggestions for improvement or Thai university students' perceptions of CBEST^[17, 40]. Additionally, cognitive process studies have examined the behaviors and thought processes of test-takers during CBEST tasks. These studies have used methods such as eye-tracking combined with interviews to understand how native and non-native speakers process test tasks^[33], as well as the development of metacognitive strategy inventories to investigate strategic competence^[32].

Mixed-methods research integrates quantitative and qualitative approaches, offering a comprehensive understanding of CBEST. Concurrent mixed-methods designs have been used to triangulate findings by collecting qualitative and quantitative data simultaneously, as in studies exploring test-taker perceptions and performance differences across testing modes or validating CBEST systems through performance analysis and expert evaluations^[25, 36, 41]. Sequential mixed-methods designs, which collect quantitative data followed by qualitative insights, have been employed to investigate relationships between test-taker perception, motivation, and performance^[29], or to validate CBEST systems through expert interviews after analyzing performance data^[42]. Multiphase mixed-methods designs, involving multiple stages of data collection and analysis, have been adopted for large-scale test development and validation projects, such as the creation of speaking tests for Uruguayan students transitioning from primary to secondary education^[43].

3.3. Effectiveness of the CBEST

The effectiveness of the CBEST has been explored extensively, with studies investigating its reliability, validity,

practicality, and perceptions. These investigations have provided valuable insights into the strengths in the CBEST, supported by diverse methodological approaches. Specifically, the reliability of the CBEST has been a primary focus, particularly in terms of scoring consistency between automated systems and human raters. Studies employing quantitative correlational designs have demonstrated that automated systems often exhibit high inter-rater reliability, although they tend to be slightly more lenient, particularly with low-proficiency speakers^[24, 44]. The integration of advanced algorithms, such as those utilizing support vector regression, has further enhanced the accuracy of automated fluency and pronunciation evaluations, achieving up to 99.9% accuracy^[23]. Similarly, the development of multi-turn oral discourse tasks has shown promise, with a 72% exact agreement between human and machine scores^[45].

The validity of CBEST has also been substantiated, with multiple studies confirming alignment with established communicative language testing frameworks. For instance, research on the Vietnamese Standardized Test of English Proficiency revealed comparable scores between computer-delivered and face-to-face formats, highlighting the CBEST's applicability across diverse contexts^[46]. Moreover, experimental designs evaluating platforms such as the metaverse and Skype have demonstrated their potential in fostering valid and interactive assessments^[27, 37].

Practicality and scalability are additional strengths of CBEST. Studies have illustrated how innovative platforms, such as WhatsApp and customized Skype systems, facilitate widespread implementation while maintaining data security and operational feasibility^[27, 47]. These systems enable cost-effective and scalable solutions, particularly in high-stakes and resource-limited settings. However, perceptions of CBEST have been mixed, with studies revealing varied attitudes across different demographic groups. For instance, while many test-takers appreciate the efficiency and accessibility of CBEST, concerns persist regarding fairness and authenticity^[17, 48]. This dichotomy underscores the need for continuous improvement in addressing user preferences and expectations.

3.4. Challenges in the CBEST

The implementation of CBEST faces several challenges that significantly impact its reliability, validity, and user ac-

ceptance. Regarding technical and design limitations, one of the core technical challenges in CBEST lies in ensuring the reliability and fairness of automated scoring systems. Studies have highlighted discrepancies in scoring between automated systems and human raters. While automated systems generally demonstrate high inter-rater reliability, they are occasionally more lenient, especially with low-proficiency speakers^[24]. Furthermore, systems exhibit lower intra-rater reliability compared to human raters, raising concerns about consistency under strict conditions^[44]. Another technical hurdle involves the development and validation of machine-learning-based algorithms. Despite notable successes in achieving high accuracy in fluency and pronunciation predictions^[26], such systems still face challenges in matching human evaluators' nuanced judgment. Additionally, the integration of novel platforms such as the metaverse for test delivery has shown potential, but such advancements introduce issues related to accessibility, familiarity with the platform, and equitable implementation^[34].

Regarding cognitive challenges, factors such as strategic competence and task familiarity also present significant challenges. Test-takers' ability to effectively plan, monitor, and evaluate their responses varies widely, impacting their performance^[32]. Research involving response processes and strategic behavior reveals differences between native and non-native speakers in navigating test interfaces, with non-native speakers showing higher cognitive loads^[33]. Furthermore, the absence of an interlocutor in the CBEST removes critical interactional elements essential to real-life communication. This gap in assessing interactional competence, often highlighted in face-to-face tests, challenges the validity of the CBEST^[46].

Affective factors, particularly test-related anxiety, also pose substantial barriers to the effectiveness of the CBEST. High levels of speaking anxiety, including communication apprehension and test anxiety, are negatively correlated with performance, as highlighted in some studies^[30, 31]. The lack of authentic interaction and immediate feedback exacerbates these anxieties, especially in culturally diverse contexts where verbal communication norms vary. Additionally, mixed attitudes toward the fairness and validity of the CBEST contribute to its affective challenges. For instance, while some test-takers perceive the format as motivating and efficient, others express concerns about its fair-

ness and authenticity^[27]. Such perceptions directly influence test-takers' effort and engagement, ultimately affecting their performance^[29].

Lastly, the practical deployment of CBEST involves overcoming barriers related to accessibility, resource allocation, and user training. For example, the successful development of a CBEST embedded in platforms such as Skype or WhatsApp underscores the need for robust infrastructure and technical support^[27, 47]. However, these innovations often fail to address broader inclusivity concerns, such as accessibility for students in rural or underserved areas. Moreover, the lack of standardization in the CBEST frameworks across different educational settings leads to variability in their application and acceptance. While some systems align with established language testing frameworks^[25], others struggle to achieve comparability across different testing modes, such as computer-delivered versus face-to-face formats^[3, 46].

4. Discussion

The systematic review underscores both the potential and challenges of the CBEST as a transformative tool in language assessment. While they offer scalability, consistency, and adaptability, their implementation and outcomes pose several critical issues that merit deeper reflection, particularly in terms of reliability, validity, equity, and the role of cognitive and affective factors. A central theme in CBEST research is the tension between reliability and validity. Automated scoring systems, while consistent and scalable, often struggle to capture the nuanced elements of speaking proficiency (e.g., pragmatics, intonation, and cultural context) that are integral to real-world communication^[24, 26, 44]. Although many studies confirm the alignment of CBEST frameworks with communicative language testing principles^[25, 36], the absence of interactional components—such as real-time dialogue with interlocutors—limits the assessment of dynamic and context-sensitive language use. This gap is particularly concerning given the increasing emphasis on assessing communicative competence in diverse global contexts. Hybrid models that combine automated scoring with human judgment may offer a solution, balancing scalability with the depth of evaluation^[41].

The cognitive and affective dimensions of CBESTs further highlight important challenges. Test-related anxiety,

particularly speaking anxiety and communication apprehension, consistently emerges as a significant barrier to performance^[30, 31]. Unlike face-to-face tests, which often involve personal interaction, the perceived impersonality of CBESTs can exacerbate test-takers' stress levels. This effect is especially pronounced among non-native speakers who may struggle with unfamiliar digital interfaces or fear judgment in automated environments^[28, 33]. Cognitive factors, such as strategic competence and task familiarity, further influence performance. Studies demonstrate that test-takers with stronger metacognitive strategies—such as planning, monitoring, and problem-solving—navigate CBEST tasks more effectively than those without such skills^[32]. To mitigate these challenges, preparatory interventions that focus on familiarizing test-takers with CBEST platforms and building metacognitive awareness are essential.

The absence of interactional competence assessment in CBESTs also warrants critical attention. While traditional face-to-face assessments allow for dynamic interaction between participants and assessors, CBESTs often rely on static prompts or pre-programmed tasks. This limitation undermines their ability to measure one of the most critical components of speaking proficiency—interactional competence^[46]. Recent innovations, such as video-mediated tests, show promise in addressing this gap by enabling real-time interaction while maintaining the scalability of CBESTs^[34, 41]. However, these advancements also introduce new challenges, such as ensuring equitable access to technology and addressing potential biases in virtual communication environments.

Equity and accessibility remain pressing concerns in the implementation of CBESTs. While these tests are often lauded for their scalability and potential to democratize language assessment, disparities in technological access and infrastructure create significant barriers for learners in under-resourced settings. For example, students in rural or low-income areas may lack stable internet connections or familiarity with digital tools, limiting their ability to perform well in CBEST environments^[6]. Such inequities not only undermine the fairness of CBESTs but also risk exacerbating existing educational divides. The practical deployment of CBESTs also raises questions about standardization and comparability. The variability in how CBEST frameworks are designed and implemented across different educational and cultural contexts presents a significant challenge. For

instance, differences in task design, scoring algorithms, and test administration modes can lead to inconsistent evaluations of speaking proficiency^[25, 36]. Developing robust, universally applicable frameworks that ensure comparability across settings is crucial to addressing this issue. Moreover, integrating ethical considerations into CBEST design, such as transparency in scoring algorithms and sensitivity to cultural variations in communication norms, will further enhance their credibility and acceptance^[10, 24].

Despite these challenges, the potential of CBESTs remains significant. They provide scalable and efficient solutions for assessing speaking proficiency, particularly in high-stakes and large-scale testing scenarios. However, realizing this potential requires continuous innovation and collaborative efforts among researchers. A review of the research designs employed in CBEST studies reveals that quantitative approaches dominate the field, with correlational studies examining variables such as speaking patterns, emotional states, and linguistic features^[24, 28, 37]. Survey-based designs further contribute by capturing test-taker perceptions and anxiety, highlighting the psychological dimensions of CBEST experiences^[28, 30, 31]. While these methods offer valuable statistical insights into test reliability and validity, they often neglect the nuanced, context-specific factors that qualitative studies uncover. For instance, qualitative case studies have illuminated the cognitive and emotional challenges faced by non-native speakers and test-taker responses to varying test interfaces^[33, 40]. Mixed-methods approaches, though less prevalent, have demonstrated their utility in triangulating quantitative and qualitative findings to validate CBEST systems and assess their impact across diverse contexts^[25, 29, 41]. These methodological trends underscore the need for more interdisciplinary research that integrates technological, linguistic, and psychological perspectives. However, future studies should prioritize longitudinal designs to investigate the long-term impacts of CBESTs on learning outcomes and experimental designs to test their adaptability in assessing interactional competence and pragmatic skills.

Future research should also prioritize refining the technical aspects of CBESTs, exploring hybrid models that integrate human judgment with automated systems to balance consistency with the nuanced understanding of communication. For instance, hybrid scoring models that leverage speech recognition technology while incorporating human

evaluators for aspects such as intonation and pragmatic competence could address limitations in fully automated systems^[13, 15]. Such advancements require collaborative efforts between researchers and test developers to ensure alignment with established communicative language testing frameworks and adaptability across diverse educational contexts. Test developers, along with policymakers, should also address the broader implications of CBEST deployment in educational systems. This includes aligning CBEST objectives with curriculum goals and teacher training programs to ensure that assessments not only measure proficiency but also support pedagogical outcomes. Policymakers should consider integrating CBESTs into national assessment frameworks, while offering professional development opportunities for educators to understand and implement these tools effectively. Emphasizing collaboration among these stakeholders will be critical in scaling CBESTs as reliable, fair, and effective instruments for assessing speaking proficiency.

5. Conclusions

This systematic review highlights the significant advancements in the CBEST over the past decade, focusing on areas such as technological innovations, test validity, cognitive and affective factors influencing performance, and diverse methodological approaches. The review underscores the growing reliance on automated systems for scoring and the promising potential of hybrid models that combine automation with human judgment to address nuanced aspects of communicative competence. Test-taker perceptions, strategic competence, and anxiety-related challenges emerged as critical areas of focus, with studies revealing the interplay between these factors and performance outcomes. The methodologies employed reflect a rich diversity, with quantitative, qualitative, and mixed-methods designs contributing to a comprehensive understanding of CBEST effectiveness and challenges. These findings suggest that while CBESTs offer scalable and efficient solutions, their optimal implementation requires sustained innovation, equitable access, and alignment with pedagogical objectives.

However, this review is not without limitations. First, while the inclusion of multiple databases ensured broad coverage, the reliance on English-language publications may

have excluded valuable insights from non-English studies, potentially limiting the global scope of the findings. Additionally, the review did not analyse the pedagogical applications of CBESTs in depth, such as their integration into classroom instruction or their impact on long-term language learning outcomes—areas that are critical to understanding the full educational implications of such assessments. The emphasis on studies published between 2014 and 2024 may also have excluded very recent developments or emerging innovations still in progress. Moreover, the review is subject to the inherent constraints of systematic review methodology, which, while offering a structured and replicable approach to evidence synthesis, may unintentionally prioritize studies that are more easily discoverable, more frequently cited, or published in high-impact journals—thus reinforcing existing research trends and overlooking less prominent but equally valuable contributions. Additionally, the methodological heterogeneity of the included studies, while reflecting the interdisciplinary nature of CBEST research, posed challenges for synthesising findings across diverse theoretical frameworks, research designs, and assessment contexts. Future reviews should aim to address these limitations by incorporating a broader range of sources—including grey literature, institutional reports, and multilingual databases—and by adopting more inclusive review strategies such as narrative synthesis or meta-analysis that can account for such methodological variation and emerging developments in the field.

For future research, several promising avenues should be explored to deepen our understanding of CBESTs and enhance their application in language assessment. One key area involves investigating the long-term impact of CBESTs on language development. While existing studies have primarily focused on immediate test performance and perceptions, longitudinal research could examine how repeated exposure to CBESTs influences learners' language proficiency, strategic competence, and test-taking behaviours over time. Such studies could adopt mixed-methods or longitudinal cohort designs, integrating performance data with interviews and learning analytics to track progress across semesters or academic years. Another crucial direction involves examining the effectiveness of feedback mechanisms in automated scoring systems. As automated feedback becomes increasingly prevalent, research is needed to assess

how different types of feedback—such as immediate vs. delayed, formative vs. summative, or linguistic vs. strategic—affect learner uptake and improvement. Experimental and quasi-experimental designs could be employed to compare learner outcomes across feedback conditions, providing empirical evidence to inform feedback integration in CBEST platforms. Further research should also address the pedagogical integration of CBESTs in classroom settings. Questions remain about how these assessments can be used not only for summative purposes but also for formative and diagnostic goals. Action research or classroom-based intervention studies could examine how teachers incorporate CBEST results into their instruction, how students respond to CBEST-informed teaching, and what institutional support is needed to facilitate this integration. Such work will be essential in ensuring that CBESTs are not only technologically advanced but also pedagogically meaningful, ethically implemented, and learner-centred.

Author Contributions

Conceptualization, H.H.; methodology, H.H. and Q.G.; software, H.H. and Q.G.; validation, N.-E.M.-S.; formal analysis, H.H. and Q.G.; investigation, H.H. and Q.G.; resources, H.H.; data curation, H.H. and Q.G.; writing—original draft preparation, H.H. and Q.G.; writing—review and editing, N.-E.M.-S.; visualization, H.H.; supervision, N.-E.M.-S.; project administration, N.-E.M.-S.; funding acquisition, N.-E.M.-S. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by Universiti Kebangsaan Malaysia (TAP-K017971), the Faculty of Education.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were generated or analyzed in this study. All data supporting the findings of this systematic review are derived from previously published studies, which are fully

cited in the references.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix A

Table A1. Summary of Reviewed Research.

Source	Indexing	Research Purpose	Research Design	Main Findings
[23]	WoS, Scopus, Google Scholar	to develop and evaluate a system for automatically assessing learner speech from a multi-turn oral discourse completion task	quantitative design with a tripartite structure (comprising: automatic speech recognizer, modules computing speech features, and scoring model)	The test achieved a 72% exact agreement between human and machine scores, comparable to results reported in existing literature, demonstrating the potential for use in low-stakes practice environments.
[24]	WoS, Scopus, Google Scholar	to examine the influence of the absence of an interlocutor on speaking test performance	quantitative design (with data collected from Japanese university students who attended a CBEST and a face-to-face speaking test)	Students who took different test modes did not show any statistical difference in the scores obtained, with further analysis showing no differences in factor structures of the two tests.
[25]	Google Scholar	to evaluate the effectiveness of a CBEST supported by the PSO algorithm and the artificial neural network	quantitative correlational design (involving Chinese university students taking the test)	The developed CBEST showed a high level of validity when comparing subjective human ratings with the scores generated by the testing system.
[26]	WoS, Scopus, Google Scholar	to investigate the role of monitoring in self-regulated learning during the CBEST and its effects on the performance of Chinese learners	quantitative correlational design (with data collected from students' test performance and questionnaires)	Monitoring was reported as being used frequently during the speaking tests; it showed no substantial impact on learners' performance in the speaking tests.
[27]	Google Scholar	to evaluate the effectiveness of a CBEST embedded in Skype	quantitative survey design (with questionnaire data collected from Japanese secondary students)	Students had a generally positive attitude to the test, believing it was motivating and efficient and had little washback effect.
[28]	Scopus, Google Scholar	to explore whether speaking patterns and emotional states could predict students' grades in a CBEST for Brazilian university students	quantitative correlational design (utilized a pretrained HuBERT model and an InceptionTime network for speech emotion recognition)	Higher speaking ratios were positively correlated with better grades. The majority of students exhibited a neutral emotional state during the CBEST, with sadness observed during hesitation moments. However, emotional states were not significantly correlated with students' grades.
[29]	Google Scholar	to examine the validity and reliability of a CBEST at China's tertiary settings based on the communicative language testing framework	mixed-methods concurrent design (with data collected from both test scores and test transcripts)	The developed test, within the communicative language testing framework, had satisfying validity and reliability.
[30]	WoS, Scopus, Google Scholar	to explore Chinese high school students' views of the CBEST	qualitative case study design (with students participating in semi-structured interviews)	Test takers had varying levels of confidence in the validity of the CBEST, and they supposed optimizing the test design and investing in examination practice and preparation for better performance.
[31]	Google Scholar	to develop and validate an online English speaking test for Uruguayan students transitioning from primary to secondary school	mixed-methods design (involving three phases: small-scale and large-scale trials with learners and examiners; CEFR-linking exercise with expert panelists; combined evidence sources to build a validity argument for the test)	The test demonstrated alignment with CEFR levels pre-A1 to A2, with results supporting the use of technology and tailored design to assess and promote speaking proficiency effectively.

Table A1. *Cont.*

Source	Indexing	Research Purpose	Research Design	Main Findings
[32]	WoS, Scopus, Google Scholar	to examine Chinese college students' anxiety levels when taking the CBEST	quantitative survey design (with questionnaires focusing on micro and macro dimensions)	Students demonstrated high levels of speaking anxiety, especially test anxiety, communication apprehension, and anxiety from the computer-based assessment process, when taking the CBEST, which was negatively correlated with test performance.
[33]	WoS, Scopus, Google Scholar	to investigate learners' strategic competence in the CBEST through the development and validation of a strategic competence inventory	quantitative survey design (with questionnaires focusing on planning, problem-solving, monitoring, and evaluating as strategic competences)	The study confirmed that planning, problem-solving, monitoring, and evaluating were essential metacognitive strategies for test-takers during the CBEST.
[34]	Scopus, Google Scholar	to investigate the effects of pre-task and online planning on discourse features (complexity, fluency, and accuracy) and test scores in a CBEST	quantitative quasi-experimental design (with students under three planning time conditions, namely pre-task, on-line, and no planning time, and data collected from human ratings and analytic complexity, fluency, and accuracy indices.	There were no significant differences in discourse measures for the CBEST performance across planning conditions. Test-takers produced more fluent and accurate language with planning time compared to no planning time.
[35]	Scopus, Google Scholar	to examine the impact of formulaic language sequences on fluency in young learners' CBEST tasks in America	quantitative design (with formulaic language sequences coded for discourse function)	The use of formulaic sequences was a significant predictor of fluency in the CBEST, and the most frequently used sequences were for clarification and for comparing.
[36]	WoS, Scopus, Google Scholar	to explore the relationships between test-taker perception, test-taking motivation, and test performance in a CBEST (i.e., TOEIC)	mixed-methods sequential design (with Japanese university students completing questionnaires and attending interviews)	Students believed the CBEST had significant validity and were greatly motivated, though they had reservations about the test delivery mode. Test-taker perception and motivation appeared to have a minimal impact on test performance. Participants' views on computer-delivered testing were directly linked to their test-taking effort. However, their perception of test validity seemed to influence test-taking effort indirectly, mediated by their perception of the test's importance.
[37]	WoS, Scopus, Google Scholar	to explore the evolution of English speaking proficiency assessment methods, focusing on the potential of the metaverse	quantitative design (with students taking different forms of tests and completing questionnaires)	Test-takers performed significantly better in the metaverse test mode compared to the face-to-face and Zoom modes, and questionnaire responses indicated that the metaverse platform was the most helpful and preferred mode among participants.
[38]	Google Scholar	to identify the linguistic features (utterance, vocabulary, and syntactic) of English recognized as salient by native speakers within a CBEST	quantitative design (with discourse analysis made of Korean university students' test responses)	Significant differences in utterance, vocabulary, and syntactic indices were observed between upper and lower proficiency groups, and most indices showed significant correlations with speaking proficiency, with vocabulary indices exhibiting the strongest correlations. A very high correlation was found between holistic and analytic scoring methods.
[39]	Google Scholar	to propose and evaluate an automatic fluency evaluation algorithm for English speaking tests, utilizing acoustic features and support vector regression (SVR) to estimate fluency scores.	quantitative design (acoustic features, including speech rate, articulation rate, and mean length of runs, were extracted from spoken utterances, and SVR was used to compute fluency scores based on human-rated training data)	Speech rate, articulation rate, and mean length of runs were the most effective features for fluency evaluation. The algorithm achieved a high correlation with human scores across the CBEST, demonstrating its potential as a reliable secondary fluency evaluation tool.

Table A1. *Cont.*

Source	Indexing	Research Purpose	Research Design	Main Findings
[40]	Scopus, Google Scholar	to explore Thai university students' perceptions of CBT, including the CBEST	quantitative survey design (with questionnaires focusing on students' perceptions)	Students had great confidence in operating the equipment for the CBEST, though their English proficiency was moderate.
[41]	Google Scholar	to examine the correlations between students' scores of a CBEST (i.e., TOEIC) with their listening and reading scores and to investigate test-takers' perceptions	quantitative design (with statistics obtained from Japanese university students' test performances and questionnaires)	The speaking scores had moderate correlation with the listening and reading scores, and test-takers had a positive perception of the CBEST.
[42]	Google Scholar	to examine whether different testing mode would affect students' speaking performance	quantitative comparative design (with Japanese university students taking either a CBEST or a face-to-face test and completing questionnaires)	Different testing modes did not influence students' speaking performance, though students preferred the traditional face-to-face speaking test mode rather than the CBEST.
[43]	Scopus, Google Scholar	to compare students' performances and perceptions of two modes of English speaking test (i.e., CBEST and face-to-face test) in a Thai university	mixed-methods concurrent design (with data collected from students' test performance, interviews, and questionnaires)	Students performed better in the CBEST than in the face-to-face test, though they considered both to be valid. While the face-to-face test was praised to be interactive, students preferred taking the CBEST for the effect of reduced anxiety.
[44]	Scopus, Google Scholar	to explore the differences in rater severity and consistency between automatic scoring and human rating	quantitative design using a many-facet Rasch model measurement computer program (with data collected from Chinese secondary students taking a CBEST)	Differences in rater severity between computer automated scoring and expert human raters do not significantly impact the distribution of examinees' scores. The low bias rate of computer automated scoring suggests it outperforms human raters in terms of internal consistency.
[45]	WoS, Scopus, Google Scholar	to develop a supervised machine-learning method for automatically evaluating fluency and pronunciation levels of language learners and detecting specific pronunciation errors	quantitative design (with data collected from audio samples from English-learning students and from test-takers' test performances)	The developed test achieved 94% accuracy for fluency predictions and 99.9% accuracy for pronunciation predictions, the highest reported in the literature for such evaluations.
[46]	WoS, Scopus, Google Scholar	to investigated the comparability of performance scores between computer-delivered and face-to-face formats for two speaking tests in the Vietnamese Standardized Test of English Proficiency (VSTEP)	mixed-methods concurrent design (test scores obtained from university students and interviews with selected cases)	The results showed comparable scores for the VSTEP.2 test but higher scores for the face-to-face mode in the VSTEP.3–5 test, with mode effects observed only in specific criteria (content development and pronunciation). Interviews revealed test-taker preferences shaped by the affordances and constraints of each mode.
[47]	WoS, Scopus, Google Scholar	To investigate the feasibility of Mobile-Assisted Language Assessment and learners' attitudes towards its use, specifically through electronic portfolios and the WhatsApp platform	mixed-methods concurrent design (with detailed description of the use of WhatsApp in the CBEST and interview data on students' perceptions)	The CBEST was successfully implemented using WhatsApp as a platform for speaking assessments. However, mixed attitudes were observed, with concerns centering around fairness and the lack of authentic communication.
[48]	Google Scholar	to evaluate a self-developed application for English speaking assessment in an Indonesian university	quantitative survey design (with English teachers completing questionnaires)	Participants supposed the application was easy to use and practical, which could efficiently measure students' English speaking proficiency.
[49]	WoS, Scopus, Google Scholar	to investigate young learners' response processes when taking the CBEST	mixed-methods concurrent design (with eye-movement indices captured and connected with speaking performance, supplemented by qualitative analysis of interview and drawing data)	Non-native English-speaking children scored significantly lower in speech production compared to native English-speaking children, with the former exhibiting longer fixation durations and more frequent gazes at the countdown timer, and the latter focusing more on content-related features (e.g., on-screen pictures).

Table A1. *Cont.*

Source	Indexing	Research Purpose	Research Design	Main Findings
[50]	Scopus, Google Scholar	to investigate test-takers' views on a CBEST (i.e., Aptis) conducted in university settings	mixed-methods concurrent design (with data collected from both questionnaires and interviews)	Test-takers had a generally positive perception of the CBEST, believing it could validly measure one's English speaking proficiency. However, test-takers' personal characteristics might determine whether the CBT mode was truly suitable or not.
[51]	Google Scholar	to explore the comparability of the automated scores of a CBEST with human ratings	quantitative design with many-facet Rasch analysis (with data obtaining from Chinese secondary students)	Both automated and teacher raters exhibit strong inter-rater reliability. However, the automated rater displays lower intra-rater reliability compared to college and high school teacher raters under strict infit criteria. Neither the automated nor human raters exhibit central tendency bias or random effects.
[52]	Scopus, Google Scholar	to develop and implement the Kyoto Institute of Technology Speaking Test, to assess the English speaking ability of Japanese undergraduate students	quantitative design (with data collected from two large-scale feasibility tests)	The study successfully demonstrated the feasibility of using secure data-sharing systems and Windows custom imaging tools in high-stakes computer-based speaking tests. These systems effectively preserved data security and facilitated external grading processes while accommodating the operational demands of PC classrooms.
[53]	Google Scholar	to identify the most frequently occurring linguistic and surface structure errors made by Korean university students in a CBEST	quantitative design (using linguistic and surface structure error taxonomies for analyzing students' test responses)	Errors with morphemes (-ing/-ed) and subject-verb agreement frequently occurred in the CBEST, as well as addition errors.
[54]	Google Scholar	to investigate the speaking-related anxiety of Chinese university students during the CBEST	quantitative survey design (with questionnaires completed by students)	Fear of negative classroom feedback was the strongest predictor of speaking-related anxiety in the CBEST, and classroom communication apprehension also positively predicted speaking-related anxiety during the test.
[55]	Scopus, Google Scholar	to explore the feasibility of using computer-mediated video technology, specifically Skype, to deliver the CBEST that measures a broad construct of oral ability, including interactional competence	mixed-methods concurrent design (with data collected from questionnaires, focus group discussions, and test data)	Tasks were viewed as representative of interactive speaking activities encountered in real-life language use, which provided opportunities for participants to demonstrate their oral abilities. When technology functioned effectively, the tasks were successfully completed in a computer video-mediated environment.
[56]	Google Scholar	to examine the relationships between working memory capacity, L2 motivation, and Japanese EFL learners' speaking skills, and to determine the respective contributions of these factors to overall L2 speaking skills	quantitative correlational design (with data collected from a CBEST and questionnaire)	Significant correlations were found between L2 speaking skills and both working memory capacity and L2 motivation within the CBEST. Both working memory capacity and motivation significantly explained variance in overall L2 speaking skills. Motivation had a stronger influence on speaking subcomponents, except for speaking grammar, where working memory had a greater impact.
[57]	WoS, Scopus, Google Scholar	to investigate the reliability of an automarker for scoring candidate responses in an online oral English test and evaluate its alignment with examiner scores	quantitative design (with data collected from the test and analyzed by agreement analysis)	The automarker demonstrated excellent internal consistency. The automarker was slightly more lenient than examiner fair average scores, especially for low-proficiency speakers. The Language Quality measure effectively predicted automarker reliability and identified abnormal speech.

Table A1. *Cont.*

Source	Indexing	Research Purpose	Research Design	Main Findings
[58]	Google Scholar	to develop and evaluate a computerized system for grading spontaneous spoken language of ESL learners	mixed-methods design (with data collected from speech corpus of English learners in Taiwan)	The automated system demonstrated high potential for use in speaking assessment. Predictive results from the system were more reliable than those of human experts.

References

- [1] Hu, H., Zhou, Q., 2024. The subterranean English training market: Examining grassroots resistance amidst China's double-reduction policies. In: Alam, M.B. (ed.). *Shadow Education in Asia: Policies and Practices*. IGI Global: Hershey, PA, USA. pp. 160–180.
- [2] Liu, Y., 2023. The Application of Portfolio Assessment in English Continuation Writing for Senior High Schools. *Journal of Advanced Research in Education*. 2(4), 41–46. DOI: <https://doi.org/10.56397/JARE.2023.07.07>
- [3] Brahim, Y., 2023. Computer-Based Vs. Face-to-Face Speaking Assessment: Fitness for Purpose from a Communicative Language Testing View. *International Journal of Social Science and Human Research*. 6(1), 22–30. DOI: <https://doi.org/10.47191/ijsshr/v6-i1-04>
- [4] Tan, J., 2021. Research on Computer-Aided English Language Evaluation System. *Journal of Physics Conference Series*. 1992(3), 032103. DOI: <https://doi.org/10.1080/1742-6596/1992/3/032103>
- [5] Kunnan, A.J. (ed.), 2024. *The Concise Companion to Language Assessment*. Wiley: Hoboken, NJ, USA. pp. 1–720.
- [6] Li, W., 2023. A Critique of the Computer-Based English Speaking Test in Fujian (CEST-FJ). *English Language Teaching and Linguistics Studies*. 5(3), 123–141. DOI: <https://doi.org/10.22158/eltls.v5n3p123>
- [7] Shoja, L., Maadikhah, M.M., 2024. From CALT to AI: Reviewing the Evolution of Technology-based Language Testing and Assessment [Presentation]. Ilam University: Ilam, Iran. DOI: <https://doi.org/10.13140/RG.2.2.13997.81125>
- [8] Hu, H., Said, N.E.M., Hashim, H., et al., 2022. Killing Two Birds with One Stone? A Study on Achievement Levels and Affective Factors in Content and Language Integrated Learning (CLIL). *International Journal of Learning, Teaching and Educational Research*. 21(4), 150–167. DOI: <https://doi.org/10.26803/ijlter.21.4.9>
- [9] Zheng, L., Ismail, H.H., Hashim, H., et al., 2025. Storytelling in English Language Education in China: A Systematic Review of Empirical Research from the Past Decade (2014–2024). *Forum for Linguistic Studies*. 7(2), 280–295. DOI: <https://doi.org/10.30564/fls.v7i2.8314>
- [10] Aydoğdu, Ç., Kaplan, Y.Ü., 2024. Voices Regarding Online Assessment: Students' Perceptions, Challenges and Proposed Solutions. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*. 12(3), 844–865. DOI: <https://doi.org/10.18506/anemon.1446039>
- [11] Sun, H., 2022. How to Teach Spoken English in Junior High Schools to Cope with the “Human-Computer Dialogue” Test. *Learning Week*. 10(10), 77–79. DOI: <https://doi.org/10.16657/j.cnki.issn1673-9132.2022.10.026>
- [12] Giraldo, F., 2023. *Fostering Pre-Service Teachers' Language Assessment Literacy*. Sello Editorial Universidad de Caldas: Caldas, Colombia. pp. 1–256. DOI: <https://doi.org/10.2307/jj.8973304>
- [13] Feng, S., 2020. A Preliminary Study on English APP-Assisted Oral English Personalized Learning. *Education Research*. 3(4), 17–18. DOI: <https://doi.org/10.32629/er.v3i4.2655>
- [14] Zhu, B., Zhong, Z., 2024. Application Research of Computer-Assisted Technologies in EAP Module Learning. *Advances in Educational Technology and Psychology*. 8(2), 123–128. DOI: <https://doi.org/10.23977/aetp.2024.080218>
- [15] Zhang, X., 2020. Improve English Speaking Skills in the “Human-Computer Dialogue” Examination. *Asia Pacific Education*. 22, 191–192. DOI: <https://doi.org/10.12240/j.issn.2095-9214.2020.22.091>
- [16] Hu, H., Du, K., 2022. TikTok in Mobile-Assisted English Language Learning: An Exploratory Study. *International Journal of Information and Education Technology*. 12(12), 1311–1320. DOI: <https://doi.org/10.18178/ijiet.2022.12.12.1755>
- [17] Hu, H., 2022. Computer-Delivered English Listening and Speaking Test in Zhongkao: Test-Taker Perception, Motivation and Performance. *Proceedings of SOCIOINT 2022-9th International Conference on Education and Education of Social Sciences*; 13–14 June 2022; International Organization Center of Academic Research (OCERINT), Istanbul, Turkey (Virtual-Online). pp. 59–75.
- [18] Klebanov, B.B., Madnani, N., 2021. *Automated Essay Scoring*. Morgan & Claypool Publishers: Vermont, USA.
- [19] Ghumssani, B.H.H., 2024. Saudi EFL Universities Students' Perceptions of Taking Computer-Based Tests and Paper-Based Tests. *Arab World English Journal*. 1–23. DOI: <https://doi.org/10.24093/awej/th.314>
- [20] PRISMA, 2024. PRISMA Flow Diagram. Available from: <https://www.prisma-statement.org/prisma-a-2020-flow-diagram> (cited 11 November 2024).

- [21] Brown, C.C., 2021. Librarian's Guide to Online Searching: Cultivating Database Skills for Research and Instruction. Libraries Unlimited: Santa Barbara, CA, USA.
- [22] Hu, H., Said, N.E.M., Hashim, H., et al., 2023. Sustaining Content and Language Integrated Learning in China: A Systematic Review. *Sustainability*. 15(5), 3894. DOI: <https://doi.org/10.3390/su15053894>
- [23] Hayashi, Y., Kondo, Y., 2024. Automated Speech Scoring of Dialogue Response by Japanese Learners of English as a Foreign Language. *Innovation in Language Learning and Teaching*. 18(1), 32–46. DOI: <https://doi.org/10.1080/17501229.2023.2217181>
- [24] Zhou, Y., 2015. Computer-Delivered or Face-to-Face: Effects of Delivery Mode on the Testing of Second Language Speaking. *Language Testing in Asia*. 5(1), 2. DOI: <https://doi.org/10.1186/s40468-014-0012-y>
- [25] Min, Y., Li, C., Wang, X., et al., 2020. Computer Based English Speaking Test Based on Artificial Neural Network. *Computer Science & IT Research Journal*. 1(1), 29–36. DOI: <https://doi.org/10.51594/csitrj.v1i1.132>
- [26] Zhang, W., Wilson, A., 2023. From Self-Regulated Learning to Computer-Delivered Integrated Speaking Testing: Does Monitoring Always Monitor? *Frontiers in Psychology*. 14, 1028754. DOI: <https://doi.org/10.3389/fpsyg.2023.1028754>
- [27] Kanzawa, K., Mitsunaga, H., Edmonds, G., et al., 2022. Development and Administration of a Skype-Based English Speaking Test in a Japanese High School. *Bulletin of Kyoto Institute of Technology*. 14, 27–47.
- [28] Beccaro, W., Arjona Ramirez, M., Liaw, W., et al., 2024. Analysis of Oral Exams with Speaker Diarization and Speech Emotion Recognition: A Case Study. *IEEE Transactions on Education*. 67(1), 74–86. DOI: <https://doi.org/10.1109/TE.2023.3321155>
- [29] Lu, Z., Li, Z., Hou, L., 2016. On the Validity and Reliability of a Computer-Assisted English Speaking Test. *Proceedings of the 2016 International Conference on Intelligent Control and Computer Application*; 16–17 January 2016; Zhengzhou, China. Atlantis Press: Zhengzhou, China. pp. 187–193.
- [30] Zhan, Y., Wan, Z.H., 2016. Test Takers' Beliefs and Experiences of a High-Stakes Computer-Based English Listening and Speaking Test. *RELC Journal*. 47(3), 363–376. DOI: <https://doi.org/10.1177/0033688216631174>
- [31] Khabbazzashi, N., Nakatsuhara, F., Inoue, C., et al., 2022. The Design and Validation of an Online Speaking Test for Young Learners in Uruguay: Challenges and Innovations. *International Journal of TESOL Studies*. 4(1), 141–168. DOI: <https://doi.org/10.46451/ijts.2022.01.10>
- [32] Yang, Y., 2017. Test Anxiety Analysis of Chinese College Students in Computer-Based Spoken English Test. *Educational Technology & Society*. 20(2), 1–12.
- [33] Zhang, W., Zhang, L.J., Wilson, A.J., et al., 2021. Supporting Learner Success: Revisiting Strategic Competence Through Developing an Inventory for Computer-Assisted Speaking Assessment. *Frontiers in Psychology*. 12, 689581. DOI: <https://doi.org/10.3389/fpsyg.2021.689581>
- [34] Joo, M., 2022. Effects of Pre-Task and On-Line Planning on Complexity, Fluency, and Accuracy in Computer-Based English Speaking and Writing Tests. *Korean Journal of English Language and Linguistics*. 22, 938–956. DOI: <https://doi.org/10.15738/kjell.22.202210.938>
- [35] François, J., Albakry, M., 2021. Effect of Formulaic Sequences on Fluency of English Learners in Standardized Speaking Tests. *Language Learning & Technology*. 25(2), 26–41. Available from: <http://hdl.handle.net/10125/73429>
- [36] Zhou, Y., Yoshitomi, A., 2019. Test-Taker Perception of and Test Performance on Computer-Delivered Speaking Tests: The Mediation Role of Test-Taking Motivation. *Language Testing in Asia*. 9(1), 10. DOI: <https://doi.org/10.1186/s40468-019-0086-7>
- [37] Shin, D., Kwon, S.K., Noh, W.I., et al., 2024. Exploring the Role of the Metaverse in English Speaking Proficiency Tests. *Journal of Computer Assisted Learning*. DOI: <https://doi.org/10.1111/jcal.13108>
- [38] Moon, Y.-S., Choi, I.-C., 2019. Salient Linguistic Features of EFL Learner Spoken Corpus Elicited by a Computerized Speaking Test. *Multimedia-Assisted Language Learning*. 22(3), 54–83. DOI: <https://doi.org/10.15702/mall.2019.22.3.54>
- [39] Jang, B.Y., Kwon, O.W., 2016. Computer-Based Fluency Evaluation of English Speaking Tests for Koreans. *Phonetics and Speech Sciences*. 6(2), 9–20. DOI: <https://doi.org/10.13064/KSSS.2014.6.2.009>
- [40] Chaisuriya, A., 2023. Readiness for Computer-Based English Tests Among College Students in Regional Thailand. *Theory and Practice in Language Studies*. 13(2), 370–375. DOI: <https://doi.org/10.17507/tpls.1302.11>
- [41] Kanzaki, M., 2017. TOEIC Speaking Test: A Correlational Study and Test Takers' Reactions. In: Clements, P., Krause, A., Brown, H. (eds.). *Transformation in Language Education*. JALT: Tokyo, Japan. pp. 441–448.
- [42] Yonezaki, M., 2016. A Comparative Analysis of Semi-Direct Speaking Testing and Direct Speaking Testing for Japanese EFL Learners. *International Journal of Curriculum Development and Practice*. 18(1), 27–38. DOI: https://doi.org/10.18993/jcrdaen.18.1_27
- [43] Sangsuwan, W., Rukthong, A., 2023. Test-Takers' Performances on and Perceptions of Two Different Modes of Online Speaking Tests. *LEARN Journal: Language Education and Acquisition Research Network*. 16(2), 168–183.
- [44] Liu, J., Zhang, B., 2020. Multi-Level Rasch Model

- Analysis of Computer-Assisted Automated Scoring of English Listening and Speaking Tests. *Proceedings of 2020 International Conference on Computer Engineering and Application*; 18–20 March 2020; Guangzhou, China. pp. 632–636. DOI: <https://doi.org/10.1109/ICCEA50009.2020.00138>
- [45] Brena, R.F., Zuvirie, E., Preciado, A., et al., 2021. Automated Evaluation of Foreign Language Speaking Performance with Machine Learning. *International Journal on Interactive Design and Manufacturing*. 15, 317–331. DOI: <https://doi.org/10.1007/s12008-021-00759-z>
- [46] Nguyen, T.H.H., Nguyen, B.T.T., Hoang, G.T.L., et al., 2024. Computer-Delivered vs. Face-to-Face Score Comparability and Test Takers' Perceptions: The Case of the Two English Speaking Proficiency Tests for Vietnamese EFL Learners. *Language Testing in Asia*. 14(1), 6. DOI: <https://doi.org/10.1186/s40468-024-00277-1>
- [47] Tarighat, S., Khodabakhsh, S., 2016. Mobile-Assisted Language Assessment: Assessing Speaking. *Computers in Human Behavior*. 64, 409–413. DOI: <https://doi.org/10.1016/j.chb.2016.07.014>
- [48] Wiannastiti, M., 2016. Assessing Speaking for a Large Number of Students by Using Bingar Application. *Proceedings of the 5th ELTLT International Conference Proceedings*; 8–9 October 2016; State University of Semarang: Semarang, Indonesia. pp. 426–430.
- [49] Lee, S., Winke, P., 2017. Young Learners' Response Processes When Taking Computerized Tasks for Speaking Assessment. *Language Testing*. 35(2), 239–269. DOI: <https://doi.org/10.1177/0265532217704009>
- [50] Amengual-Pizarro, M., García-Laborda, J., 2017. Analysing Test-Takers' Views on a Computer-Based Speaking Test. Profile: Issues in Teachers' Professional Development. 19(S1), 23–38. DOI: https://doi.org/10.15446/profile.v19n_sup1.68447
- [51] Cao, L., 2020. Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test. *English Language Teaching*. 13(1), 18–30. DOI: <https://doi.org/10.5539/elt.v13n1p18>
- [52] Masuda, H., Mori, M., Kanzawa, K., et al., 2016. Secure Data Management in an English Speaking Test Implemented in General-Purpose PC Classrooms. *Proceedings of the 2016 ACM SIGUCCS Annual Conference*; 6–9 November 2016; Denver Colorado USA. Association for Computing Machinery: New York, NY, USA. pp. 135–138.
- [53] Munkh-ochir, G., Lee, C.-i., 2015. A Study on Error Types Represented in Computer-Based Speaking and Computer-Based Writing: Centered Around Korean Learners. *Multimedia-Assisted Language Learning*. 18(3), 112–141. DOI: <https://doi.org/10.15702/mall.2015.18.3.122>
- [54] Xu, L., Zhao, X., Zheng, C., et al., 2017. Speaking-Related Anxiety in Computer-Assisted Language Testing Settings. In: Hayashi, Y. (ed.). *Workshop Proceedings of the 25th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education: Christchurch, New Zealand.
- [55] Ockey, G.J., Timpe-Laughlin, V., Davis, L., et al., 2019. Exploring the Potential of a Video-Mediated Interactive Speaking Assessment. *ETS Research Report Series*. 2019(1), 1–29. DOI: <https://doi.org/10.1002/ets2.12240>
- [56] Kondo, A., 2021. The Effects of Individual Differences in Speaking Skills of Japanese EFL Learners: Aptitude (Phonological Working Memory) and Attitude (L2 Motivation). *Studies in Language Sciences*. 19, 1–18. DOI: https://doi.org/10.34609/sls.19.0_1
- [57] Xu, J., Jones, E., Laxton, V., et al., 2021. Assessing L2 English Speaking Using Automated Scoring Technology: Examining Automarker Reliability. *Assessment in Education: Principles, Policy & Practice*. 28(4), 411–436. DOI: <https://doi.org/10.1080/0969594X.2021.1979467>
- [58] Wu, T.-I., Lo, T.-H., Chao, F.-A., et al., 2022. A Preliminary Study on Automated Speaking Assessment of English as a Second Language (ESL) Students. *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing*; 21–22 November 2022; Taipei, Taiwan. pp.174–183.