

Mapping Technology-Enhanced Language Assessment in TVET and Occupational Education: Bibliographic Structure and Emerging Directions

Muslihah binti Saman^{1*} and Mohd Fahmi bin Adnan²

^{1,2}Department of Advanced Technical and Vocational Education and Training, Faculty of Educational Sciences and Technology, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.1026EDU0520>

Received: 14 August 2026; Accepted: 20 August 2026; Published: 24 August 2026

ABSTRACT

Technology-enhanced language assessment increasingly combines online testing, automated scoring, artificial-intelligence-supported feedback and multimodal performance capture. In vocational and occupational education, however, this literature is dispersed across language testing, English for specific purposes, workplace communication and professional education. This study maps the bibliographic structure of that intersection and examines what the underlying studies show about occupational authenticity, human-AI judgement, validity evidence and assessment governance. A Scopus search conducted on 5 August 2026 retrieved 253 records. After document-type, language, publication-year and title-abstract screening, 31 journal articles meeting all three substantive eligibility dimensions formed the final bibliometric corpus. Performance analysis, a reproducible Python/NetworkX document-level bibliographic-coupling analysis, author-keyword analysis and structured full-text synthesis were combined. Twenty-two articles (71.0%) were published from 2021 to 2026. At a threshold of at least one shared cited work, the coupling network contained 44 unique document-pair links with a summed edge weight of 65; the largest component included 19 documents and 12 documents were isolated. The same largest-component size and isolate count were retained across exact, prefix and fuzzy canonicalisation variants, although link counts and summed edge weights varied. Author-keyword recurrence was sparse, with only six terms occurring in at least two documents. Full texts were available for 26 articles and supported four cross-cutting findings: human-AI assessment is more defensible as complementarity than substitution; occupational authenticity depends on task-construct alignment; validity evidence remains uneven across assessment uses; and governance and score-use evidence are less developed than delivery, scoring and feedback. The field is therefore recent but structurally fragmented. Future research should prioritise occupationally authentic validation, transparent human-AI decision rules, fairness, consequential score use and traceable assessment governance.

Keywords: bibliometric analysis, occupational communication, technology-enhanced language assessment, TVET, workplace assessment

INTRODUCTION

Digital platforms, automated scoring, generative artificial intelligence (GenAI), multimodal interfaces and data-rich feedback systems are changing how language performance is elicited and judged. These technologies influence task delivery, performance capture, scoring, feedback and the division of judgement between automated systems and human assessors. Recent reviews of AI-enabled language assessment show both substantial potential and persistent concerns involving validity, reliability, fairness and practical implementation (Chen et al., 2025; Chuang & Yan, 2025).

The issue becomes more consequential when language assessment is tied to occupational participation. In technical and vocational education and training (TVET), engineering, healthcare, tourism, business and other

professional settings, successful communication is not simply general language proficiency. It may involve explaining a technical process, interpreting instructions, documenting progress, responding to a client, presenting a product or adapting language to professional norms. Language-for-specific-purposes assessment therefore derives its content and methods from the target language-use situation rather than treating context as a decorative label (Douglas, 2000).

Technology can improve speed and scale without resolving the construct problem. A computer-based test may be efficient yet weakly connected to occupational communication. An automated writing or speaking score may correlate with a human score without demonstrating agreement, appropriate interpretation or safe score use. Likewise, a classroom intervention may improve a locally scored task without establishing transfer to workplace performance. Lockwood (2015) makes a useful distinction between assessment for work, which predicts readiness or entry, and assessment at work, which evaluates communication in authentic performance and quality-assurance contexts.

The relevant evidence is distributed across several research communities. Broad bibliometric studies have mapped English language assessment (Yang & Wang, 2025), computer-assisted language learning (Mohsen et al., 2024), informal digital learning of English (Liu et al., 2025), and vocational training education more generally (Rajamanickam et al., 2025). Recent education bibliometrics have also linked teacher competency, digital technology integration and assessment-related school implementation (Zaki et al., 2026). These mappings establish the growth of technology-related language, education and vocational research, but they do not isolate the intersection where technology materially shapes language or communication assessment for vocational, workplace or discipline-specific professional purposes. Within the present corpus, recent vocationally explicit studies include AI-augmented speaking assessment in vocational EFL classrooms (Dai et al., 2026) and speech-recognition-based pronunciation assessment in higher vocational English (He et al., 2026), illustrating that the intersection is substantive even though it remains small. Recent bibliographic-coupling research demonstrates that specialised intellectual structures can be mapped through shared-reference networks, including within IJRISS (Yusoff et al., 2026), while current Python-based studies show that NetworkX can support transparent network construction and community analysis (Hamaniuk et al., 2026; Mirabile et al., 2026).

This study addresses that intersection through a bibliometric-assisted design that combines performance profiling, document-level bibliographic coupling, author-keyword analysis and structured full-text synthesis. The bibliometric layer is used to describe the structure of the literature, while full-text reading is used to interpret assessment-related claims and their evidentiary scope. This separation is important in a small and heterogeneous corpus because network proximity does not by itself establish thematic equivalence or assessment quality. The study therefore asks:

1. RQ1. How has research on technology-enhanced language assessment in TVET and occupational education developed over time?
2. RQ2. What structural relationships are visible through document-level bibliographic coupling, and how sensitive are those relationships to shared-reference thresholds?
3. RQ3. What conceptual priorities are visible in the author-keyword structure, and where is that structure too sparse for stable co-word interpretation?
4. RQ4. What does full-text synthesis reveal about occupational authenticity, human-AI judgement, validity evidence and assessment governance?

LITERATURE REVIEW

Technology-enhanced language assessment

In this study, technology-enhanced language assessment is defined by function. Digital technology must materially shape assessment delivery, performance capture, scoring, feedback, task generation, interpretation, reporting, rater processes or access. The definition therefore includes computer- and web-based tests, automated writing or speaking assessment, AI-supported feedback, digitally mediated performance tasks and

multimodal systems. It excludes studies in which technology is used primarily for instruction while assessment appears only as a routine pre-test or post-test.

This boundary is important because the presence of a chatbot, learning management system or mobile application does not make a study an assessment study. The assessment object must remain substantive. Recent AI-focused reviews similarly show that design and score interpretation matter as much as automation itself, especially where generated feedback or machine scoring is introduced into consequential decisions (Chen et al., 2025; Chuang & Yan, 2025).

Vocational, occupational and discipline-specific language assessment

Specific-purpose assessment links test content to a target domain. In occupational contexts, language, technical knowledge and action may operate together, which means that task authenticity cannot be inferred from a professional label alone. Douglas (2000) emphasises target language-use analysis, while Lockwood (2015) shows how the purpose, location, task, rater and criteria can distinguish predictive assessment for work from authentic assessment at work.

The operational boundary used here is intentionally broader than the acronym TVET. Engineering, medical, business, tourism and workplace studies were retained when they involved a meaningful professional or vocational context. This avoids missing assessment research that addresses occupational communication without using TVET terminology, while still excluding general language-learning studies with no substantive assessment function.

Positioning against existing mappings

The closest published mappings differ either in substantive scope or in analytical purpose. Table 1 positions the present study against selected recent studies that are relevant to language assessment, vocational education and computational bibliometrics. The comparison is used to define the study boundary rather than to imply that one software environment is inherently superior to another.

Table 1: Positioning of the present study against selected bibliometric mappings

Study	Database/corpus	Primary focus	Relevance to the present study
Yang & Wang (2025)	WoS; 927 publications	English language assessment broadly	Does not isolate technology-mediated vocational or workplace assessment.
Mohsen et al. (2024)	Scopus scientometric study	Computer-assisted language learning	Maps CALL broadly rather than occupational assessment objects and score use.
Zaki et al. (2026)	Scopus; 79 documents	Teacher competency and pedagogical models for digital technology integration in school physical education	Current IJRISS Education benchmark combining Scopus, PRISMA-aligned screening and network-based thematic mapping; vocational language assessment is outside its scope.
Rajamanickam et al. (2025)	Scopus; 366 articles	Vocational training education	Broad TVET benchmark using bibliometric mapping and bibliographic coupling; assessment is not the central unit.
Hamaniuk et al. (2026)	WoSCC; 1,552 publications	Multilingual repertoires and linguistic justice in education	Current education precedent for Python-based network bibliometrics.
Mirabile et al. (2026)	Scopus; 114 publications	Trust in human-AI collaboration in finance	Current precedent for Python/NetworkX bibliographic coupling combined with full-text synthesis; substantive field differs.

Study	Database/corpus	Primary focus	Relevance to the present study
Present study	Scopus; 31 journal articles	Technology-enhanced language assessment in vocational and occupational contexts	Uses reproducible Python/NetworkX coupling and evidence-sensitive full-text synthesis without forcing unstable co-word clusters.

METHODOLOGY

Design and database

The study used a bibliometric-assisted mapping design combining performance profiling, document-level bibliographic coupling, author-keyword analysis and structured full-text synthesis. Scopus was used as the sole bibliographic source because the network depends on internally consistent cited-reference strings, author keywords and document metadata. This was a deliberate consistency choice, not a claim that Scopus is exhaustive: bibliographic databases differ in disciplinary coverage, so database selection should match the retrieval purpose (Gusenbauer, 2022). Studies comparing and combining Scopus with Web of Science show that the indexes overlap only partially, that bibliometric structures depend on the citation index selected, and that merging citation data requires substantial unification and repair before network analysis (Kumpulainen & Seppänen, 2022). For this small specialised corpus, retaining one citation index therefore avoids adding cross-index metadata variation to the document-coupling network. The resulting map is explicitly Scopus-bounded. Bibliometric methods are appropriate for describing publication patterns and relationships among research constituents, while small specialised corpora require caution when network partitions are interpreted (Donthu et al., 2021; Rogers et al., 2020). The computational strategy is consistent with recent Python/NetworkX bibliometric applications (Hamaniuk et al., 2026; Mirabile et al., 2026), but the present study does not assume that algorithmic communities automatically constitute substantive themes.

Search development and study selection

Two Scopus advanced-search pilots were conducted on 5 August 2026. P1 used a strict intersection of language assessment, assessment technology and TVET terminology and returned 33 records. P2 broadened the context block to include vocational, occupational, workplace, English for specific purposes (ESP), English for occupational purposes (EOP) and professional-communication terminology and returned 253 records. All 33 P1 records were contained within P2. P1 captured only 3 of the 31 final eligible articles, whereas P2 captured all 31; P2 was therefore retained as the final retrieval query because it provided substantially better coverage of the defined intersection.

Table 2: Search-pilot development

Pilot	Function	Raw records	Decision
P1	Strict TVET–technology–language–assessment intersection	33	Sensitivity pilot; captured 3/31 final articles.
P2	Expanded vocational, occupational, workplace, ESP/EOP and professional context	253	Final retrieval query; captured 31/31 final articles.

The exact search syntax, raw exports and search evidence were preserved during study development. Search reporting was guided by PRISMA 2020 and PRISMA-S principles for transparent identification and reproducible literature searching (Page et al., 2021; Rethlefsen et al., 2021). The final P2 search string is reproduced in Appendix A.

Eligibility and screening

Administrative eligibility was limited to English-language Articles or Reviews published from 2000 through 2026. Conference papers and conference reviews, books and chapters, errata and retracted documents were

removed before title-abstract screening. Substantive eligibility required three simultaneous elements: (1) a language or communication assessment object; (2) a material technological role in assessment delivery, scoring, feedback, task generation, reporting, rater processes or access; and (3) a vocational, occupational, workplace or discipline-specific professional-education context.

Open-access status was not an eligibility criterion. Records that met the scientific criteria remained in the bibliometric corpus regardless of publisher access model. Full-text availability was considered only for the separate interpretive synthesis so that bibliometric inclusion was not determined by access status.

Table 3: Screening reconciliation

Screening stage	Records
Ineligible document type	105
Non-English	5
Outside 2000–2026	6
Title–abstract records assessed	137
Excluded after title–abstract assessment	106
Final bibliometric corpus	31

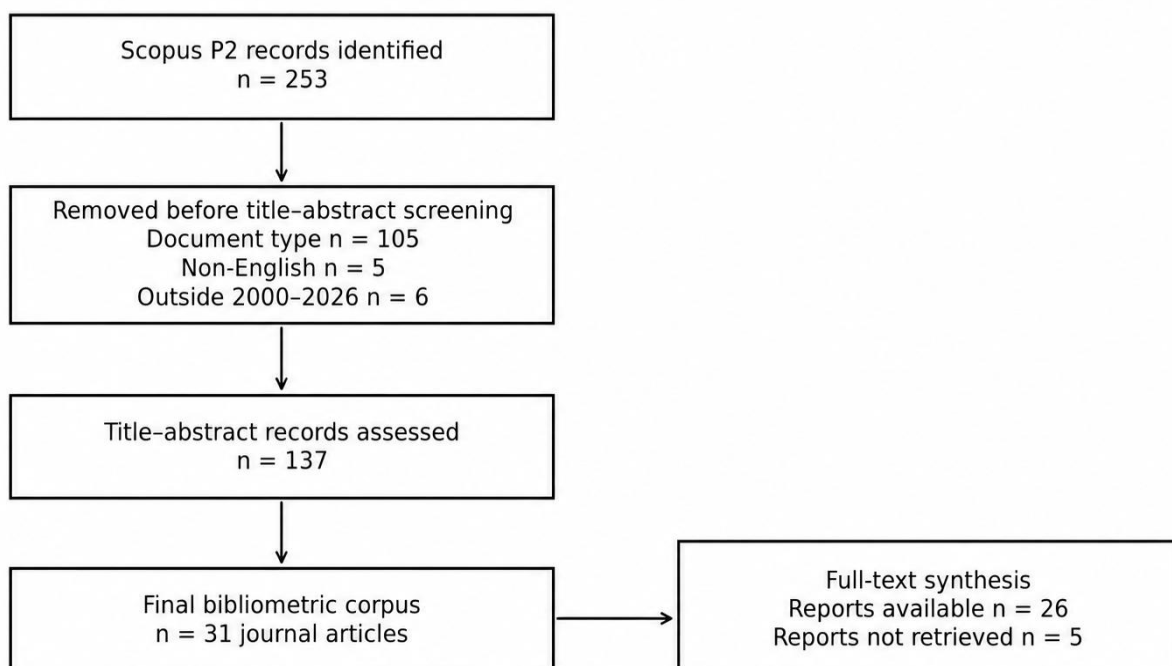


Figure 1: Identification, screening and full-text availability for the final corpus.

Bibliometric and network analysis

Performance analysis described publication year, source dispersion and Scopus citation counts at the export date. Bibliographic coupling was implemented in Python from the final Scopus References field. Because semicolons in the flat export can occur both within author strings and between cited works, cited-reference boundaries were reconstructed by accumulating tokens until a parenthesised publication-year terminator was reached. Across the 31 included records, 3,323 semicolon tokens were reconstructed into 1,518 cited-reference blocks; 1,481 blocks (97.6%) yielded a usable year, first-author surname and normalised cited-title stem. Thirty-seven blocks were unusable for matching and no residual unclosed blocks remained. Exact normalised

title-year matches were unified; within the same first-author/year group, variants were additionally merged when normalised title similarity was at least 0.88 or one title was an exact token-prefix of the other with at least three tokens. Ten cross-record canonicalisation decisions were manually checked before analysis.

Documents were treated as nodes and an edge was created when two documents shared at least one canonical cited work. Edge weight was the raw number of shared canonical cited works. For descriptive comparison only, cosine-normalised overlap was calculated as $c_{ij} / \sqrt{(|R_i| |R_j|)}$, where c_{ij} is the shared-reference count and R_i and R_j are the two usable canonical cited-reference sets. This normalised value did not determine edge inclusion, node size or the network layout. Network construction and analysis used Python 3.13.5 and NetworkX 3.6.1, with Matplotlib 3.10.8 for static figures. The primary network used a threshold of at least one shared cited work, with sensitivity checks at thresholds of two and three. Density, connected components, isolates and clustering measures were computed in NetworkX. Deterministic greedy-modularity communities on the non-isolated subgraph were used only as structural diagnostics. The static map used a fixed spring-layout seed of 42. The workflow was rerun twice in the same software environment and returned identical generated outputs. Canonicalisation sensitivity was examined using exact-only, exact-plus-prefix and fuzzy thresholds of 0.95, 0.90, 0.88 and 0.85.

Author-keyword analysis and full-text synthesis

Author keywords were first assessed for coverage and recurrence. A dense co-word solution was not forced when recurrence was insufficient. Full texts were then reviewed separately for occupational authenticity, human-AI judgement, validity evidence and governance/score use. Peer-reviewed full texts were available for 26 of the 31 articles: 25 publisher or official final versions and one peer-reviewed accepted manuscript. Five final reports were not available for content synthesis and were not used as primary evidence for interpretive claims.

The 26 available studies used heterogeneous designs, including experiments, surveys, computational models, psychometric studies, qualitative studies and conceptual work. Interpretation therefore focused on the type and scope of evidence reported rather than assigning a single quality score across incompatible designs. Correlation was not treated as agreement, nonsignificance was not treated as equivalence, self-reported competence was distinguished from observed performance, and psychometric evidence was interpreted only for the aspects it directly addressed.

RESULTS

Corpus selection, growth and source dispersion

The P2 search yielded 253 records. Administrative screening removed 105 ineligible document types, five non-English records and six records outside 2000-2026, leaving 137 records for title-abstract assessment. A further 106 records failed the substantive assessment, producing a final bibliometric corpus of 31 journal articles. The strict P1 pilot would have captured only 3 of these 31 articles, whereas P2 captured all 31, confirming the need for the expanded occupational and ESP/EOP context block. No record was excluded because it was behind a paywall. The $n=31$ corpus is therefore the complete set of retrieved Scopus records meeting the predefined search-and-screening protocol, rather than a random subsample of the 137 records assessed at title-abstract level.

The included literature was temporally concentrated in the current decade. Twenty-two of the 31 articles (71.0%) were published from 2021 through 2026. Six articles were published in 2025 and five in 2026, while the earliest included records appeared in 2001 and 2002. The 31 records were distributed across 27 source titles. *Frontiers in Psychology*, *Assessing Writing*, *Journal of Writing Analytics* and *Language Testing* each contributed two articles; every other source contributed one.

Scopus citation counts at the 5 August 2026 export totalled 304 and were unevenly distributed. The five most cited included records were Shih (2013; 61 citations), McNamara (2014; 48), Kohnke et al. (2021; 36), Corrigan and Slomp (2021; 33), and Wolf and Wolf (2023; 19). These counts are reported as a performance indicator at the export date and are not treated as evidence of methodological quality.

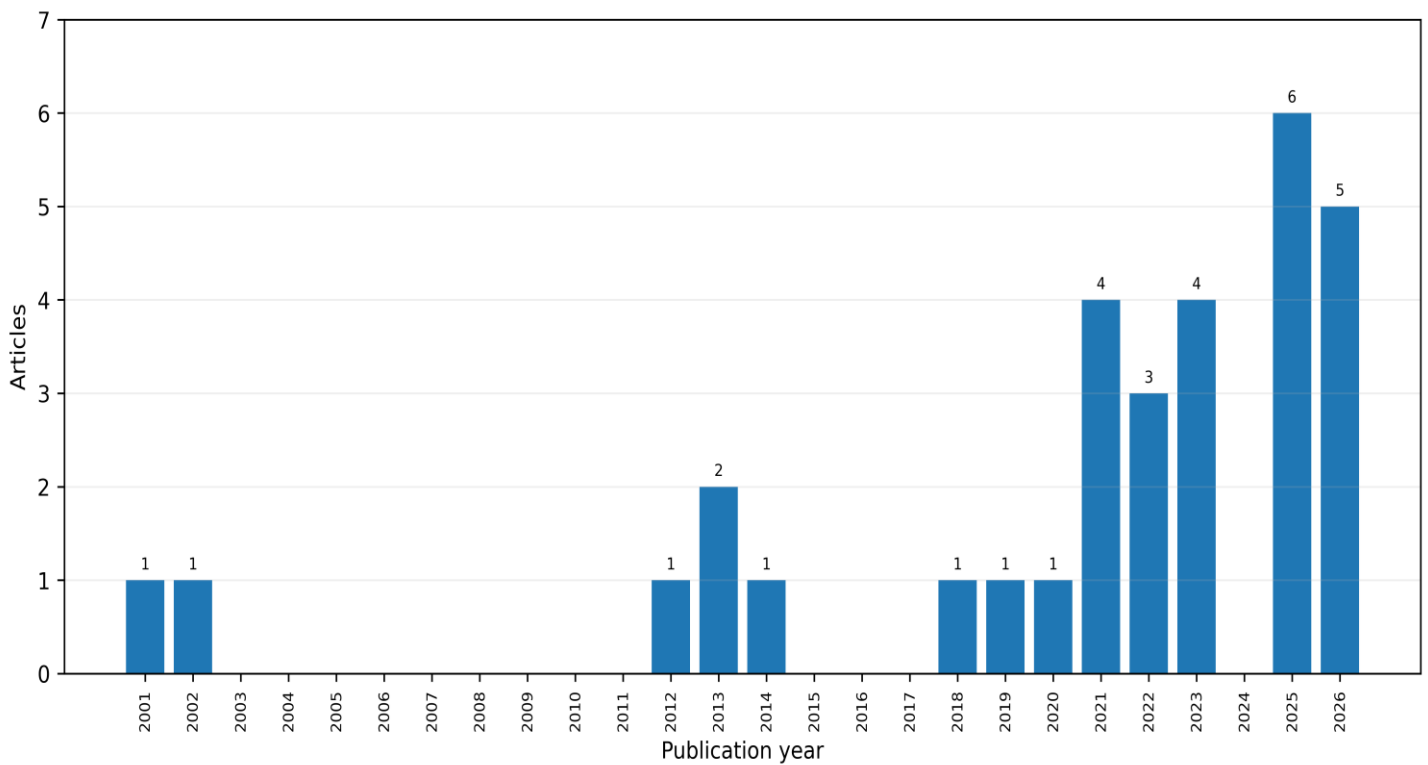


Figure 2: Annual publication output in the final corpus from 2001 to 2026 (n = 31); years with zero included publications are shown explicitly.

Table 4: Corpus profile

Indicator	Result
Database	Scopus
Search date	5 August 2026
Publication years represented	2001–2026
Final bibliometric corpus	31 journal articles
Published 2021–2026	22 (71.0%)
Distinct source titles	27
Scopus citations at export	304
Full texts available for synthesis	26
Final reports not retrieved	5

Bibliographic coupling: a sparse and threshold-sensitive structure

The Python/NetworkX workflow produced 44 unique document-pair links at the primary threshold of at least one shared canonical cited work. The sum of raw shared-reference edge weights across these unique pairs was 65 and network density was 0.0946. The largest connected component contained 19 articles and 12 articles were isolated, producing 13 components overall. The structure was sensitive to stronger coupling thresholds: requiring at least two shared cited works reduced the network to 18 links, a 15-document largest component and 16 isolates; at three shared cited works, only three links remained, the largest component contained three documents and 26 documents were isolated. The non-isolated subgraph produced four greedy-modularity communities (modularity = 0.363), but these partitions were treated as structural diagnostics rather than as self-validating thematic clusters.

Table 5: Bibliographic-coupling threshold sensitivity

Threshold	Links	Largest component	Isolates	Components
≥ 1 shared cited work	44	19	12	13
≥ 2 shared cited works	18	15	16	17
≥ 3 shared cited works	3	3	26	28

Canonicalisation sensitivity showed that the component-level conclusion did not depend on a single fuzzy threshold. Exact title-year matching alone produced 39 links (sum of edge weights = 52), while adding prefix matching produced 44 links (sum of edge weights = 61). Fuzzy thresholds of 0.95, 0.90, 0.88 and 0.85 all produced 44 links, with summed edge weights ranging from 64 to 65. Across all tested variants, the largest component remained 19 documents with 12 isolates and 13 connected components. Community membership changed more readily than component-level structure, supporting a cautious interpretation of algorithmic partitions.

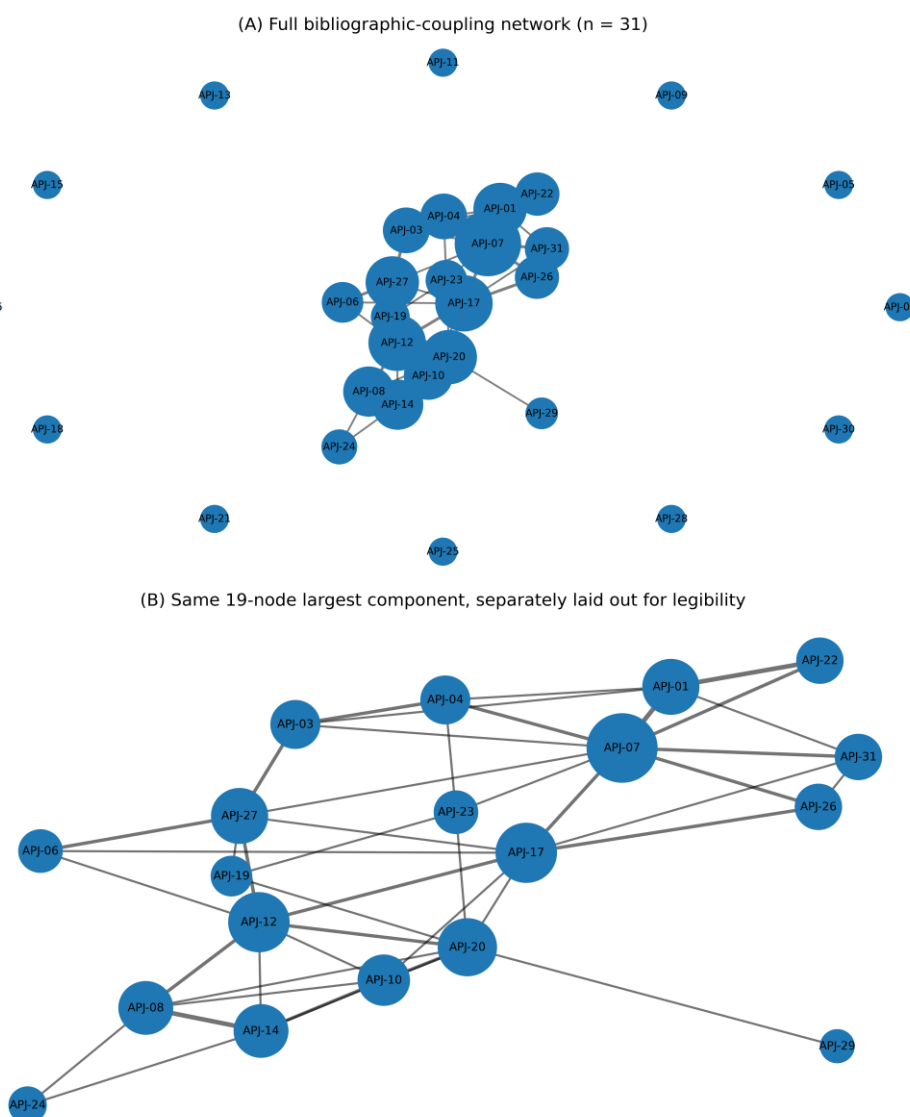


Figure 3: Reproducible document-level bibliographic-coupling network generated in Python 3.13.5/NetworkX 3.6.1 at a threshold of at least one shared cited work.

Panel A shows all 31 records, including 12 isolates; the isolates are placed around the perimeter solely for legibility and their positions do not encode similarity or distance. Panel B shows the same 19-node largest connected component with a separately recomputed display layout; no nodes, edges, edge weights, analytical

thresholds or corpus records are changed. Node size represents weighted degree and edge width represents shared-reference count. Both layouts use seed 42; layout coordinates are display variables, not analytical results.

The sparse network is itself informative. Three document pairs reached the maximum observed coupling strength of three shared cited works, whereas 26 of the 44 links were supported by only one shared cited work. Stronger pairings tended to occur among studies drawing on common assessment-validity, writing-assessment or methodological literatures, while many papers entered through locally specific technical or pedagogical traditions. Panel B is presented only to make the connected structure legible; all 31 records remain in every reported network metric. Because the network fragmented quickly under stricter thresholds, interpretation of the field was not based on network colour or community assignment alone.

Author-keyword structure

Author keywords were available for 28 of 31 records, but recurrence was low. Only six exact terms occurred in at least two documents: writing assessment (3), language for specific purposes (3), ESP (2), formative assessment (2), engagement (2) and writing analytics (2). Three records had no author keywords. This sparsity does not support a stable dense co-word map; descriptive recurrence is therefore reported rather than forcing a thematic clustering solution.

Table 6: Author keywords recurring in at least two documents

Exact author keyword	Documents
writing assessment	3
language for specific purposes	3
ESP	2
formative assessment	2
engagement	2
writing analytics	2

Full-text synthesis

Full texts were available for 26 articles. The synthesis identified four cross-cutting patterns that were sufficiently supported across the available studies: human-AI judgement and feedback, occupational authenticity through task-construct alignment, uneven validity evidence, and governance/score use. Five reports remained unavailable for full-text synthesis: van den Branden et al. (2002), Motteram et al. (2023), McNamara (2014), Lockwood (2022), and Salem (2026). These records remained in the bibliometric network but did not contribute primary evidence to the interpretive findings.

Table 7: Cross-study synthesis from the 26 available peer-reviewed full texts

Cross-cutting pattern	Illustrative studies	Synthesis
Human-AI judgement and feedback	Lan et al. (2025); Pan (2025, 2026); Zhu & Dong (2025); Doykova (2025); Gao (2026)	Automation supports bounded scoring and feedback functions, but contextual, pragmatic and consequential judgement continues to require human oversight.
Occupational authenticity through task-construct alignment	Dursun et al. (2020); Waiprakhon & Jaturapitakkul (2018); Jaturapitakkul et al. (2025); Fortanet-Gómez & Edo-Marzá (2022); Kohnke et al. (2021)	Occupational grounding is strongest when target communication, task design, response processes and intended constructs are explicitly connected.

Cross-cutting pattern	Illustrative studies	Synthesis
Validity evidence across assessment uses	Nazim & Alzubi (2022); Jaturapitakkul et al. (2025); Phaiboonnugulkij & Prapphal (2013); Lan et al. (2025)	Reliability, item functioning, correlations, model performance and response-process evidence answer different questions and should not be treated as interchangeable validity evidence.
Governance, score use and implementation	Zapata-Rivera et al. (2021); Hamilton et al. (2001); Giray & Ramos (2025)	A smaller strand addresses score communication, rater preparation, implementation constraints and assessment security, leaving consequential score use comparatively underdeveloped.

Evidence quality and interpretation

The full-text synthesis also showed why evidence type matters when interpreting technology-assisted assessment. Lan et al. (2025), for example, reported a moderate total-score correlation with experienced teachers ($r = .424$) and weaker analytical-domain correlations ($r = .186-.364$). These results support partial alignment rather than interchangeability between human and machine scoring. Pan (2026) did not detect a between-group writing advantage after baseline control, although qualitative evidence indicated engagement with AI-supported revision. Zhu and Dong (2025) likewise highlighted contextual and pragmatic differences between human and ChatGPT ratings. Together, these studies support a cautious complementarity position rather than a general substitution claim.

Other studies were most informative when their evidence was kept aspect-specific. Dursun et al. (2020) supports workplace-domain analysis and assessment-task design; Jaturapitakkul et al. (2025) provides Rasch evidence on item functioning, targeting, reliability and differential item functioning; De Felice and Deane (2012) demonstrates an early speech-act classification component; and Hamilton et al. (2001) documents implementation constraints in online rater training. These contributions are valuable without extending them beyond the designs and evidence actually reported.

DISCUSSION

Growth without consolidation

The publication profile is strongly concentrated in the current decade, but this recent concentration has not produced a dense intellectual core. Source dispersion is high, author-keyword recurrence is low and the coupling network fragments quickly as the shared-reference threshold increases. This combination suggests an intersection assembled from multiple neighbouring traditions rather than a mature specialty with a stable shared vocabulary. The result also explains why the literature is difficult to retrieve with a narrow TVET search string.

The small corpus reflects the deliberately narrow three-part eligibility intersection and should not be defended by invoking a universal bibliometric minimum. Rogers et al. (2020) examined sample-size stability for institutional normalised-citation-impact estimation rather than proposing a universal threshold for topic mapping. Their analysis nevertheless underlines the instability of quantitative inference in small sets. Accordingly, the present $n=31$ corpus is not used to support institutional-style impact ranking, fine-grained cluster claims or a statistically stable map of the entire field. It is analysed as the complete Scopus-eligible set under the stated protocol, with threshold and canonicalisation sensitivity reported explicitly and full-text synthesis used to bound substantive interpretation.

Occupational authenticity is an assessment-design problem

The full-text evidence suggests that occupational authenticity is strongest when target communication, task design, evidence and intended score use are connected. Dursun et al. (2020) begins from workplace-domain

analysis and evidence-centred design. TETET studies connect computer-based reading tasks to engineering and technology contexts and then examine response processes or item functioning (Waiprakhon & Jaturapitakkul, 2018; Jaturapitakkul et al., 2025). Product-pitch and infographic studies extend this logic to multimodal and discipline-specific performance (Fortanet-Gómez & Edo-Marzá, 2022; Kohnke et al., 2021).

These studies also show why authenticity should not be equated with realism. A task may look occupational while its scoring evidence remains narrow. A digital simulation or professional genre still requires defensible construct representation, scoring, reliability and use. The field would benefit from clearer statements of whether an assessment is intended for learning, selection, certification, diagnosis, coaching or quality assurance, because the validity evidence required for these uses is not interchangeable.

Human-AI judgement: complementarity is more defensible than substitution

Recent AI studies are prominent but heterogeneous in design and evidentiary strength. Lan et al. (2025) examined a customised Claude-3-Opus chatbot for engineering technical progress reports and found moderate overall alignment but weaker domain-level relationships with experienced teachers. Pan (2025) reported gains within a hybrid GenAI-plus-peer-feedback condition, but the design bundled feedback sources and involved cohort differences. Pan (2026) found no detected between-group writing advantage after baseline control while reporting stronger revision engagement with AI feedback. Doykova (2025) compared AI- and instructor-designed tasks in an exploratory design, and Zhu and Dong (2025) showed that human and machine ratings can diverge on pragmatic and contextual features.

Taken together, the evidence supports complementarity more strongly than substitution. Automated systems can assist with repeatable, bounded features such as linguistic feedback, speech-act classification or pattern-based scoring. Human judgement remains important when meaning depends on task fulfilment, audience, tone, empathy, disciplinary context or the consequences of a score. This conclusion is consistent with wider systematic evidence warning against unselective adoption of GenAI in language assessment (Chuang & Yan, 2025).

Validity needs to be decomposed rather than declared

Validity evidence in the corpus is heterogeneous and often aspect-specific. Item-total correlations, internal consistency, Rasch fit, human-machine correlation and model-performance coefficients address different properties; none alone establishes that scores are appropriate for a workplace decision. Nazim and Alzubi (2022), for example, reported item-functioning and internal-consistency evidence. Vernice et al. (2019) is best interpreted as an exploratory screening study rather than evidence of a fully validated occupational diagnostic. The TETET Rasch study contributes evidence on item functioning, targeting, reliability and gender DIF, while external, cross-cultural and consequential validity remain separate questions.

Future studies should therefore separate at least four layers: representation of the occupational construct, scoring quality, comparability across relevant groups or conditions, and consequences of score use. AI systems add another layer: the quality and stability of the machine judgement itself. Repeated-run reliability, participant-disjoint evaluation, human reference-score quality, subgroup error analysis and transparent model/version documentation should become routine where automated scores are proposed.

Governance and score use are underdeveloped

Delivery, scoring and feedback dominate the visible technology story, while governance and score use are less developed. Zapata-Rivera et al. (2021) treats score communication as a design problem, Hamilton et al. (2001) illustrates the organisational realities of online rater training, and Giray and Ramos (2025) raise task-design questions in response to AI-assisted plagiarism. These strands point beyond technical accuracy toward accountability: who reviews automated output, what evidence is retained, how disagreements are resolved, and how learners or employers interpret the result.

For TVET and occupational education, these questions are not secondary. Assessment results may influence progression, certification, employability or workplace coaching. A mature research programme therefore needs

to connect the assessment lifecycle from task specification and evidence capture through scoring, review, reporting, appeals and consequential use.

FUTURE RESEARCH AGENDA

The combined bibliometric and full-text findings point to six priorities for a more mature research programme in technology-enhanced vocational and occupational language assessment.

Table 8: Evidence-based research agenda

Priority	Research question	Evidence needed
Occupational authenticity	How closely do digitally elicited performances represent target workplace communication?	Target-task analysis, subject-matter experts, authentic artefacts and criterion evidence.
Human-AI decision rules	Which components can be automated, which require human judgement, and how should disagreements be resolved?	Blinded human-AI comparison, agreement analysis, repeated-run stability and error typology.
Validity and fairness	Do scores support intended interpretations across groups and contexts?	Construct evidence, external criteria, subgroup/DIF analyses and accessibility testing.
Score use and consequences	How are scores, feedback and reports interpreted and acted upon?	Score-report studies, decision tracking, user interviews and consequential evidence.
Governance and traceability	What evidence is retained for review, appeals and quality assurance?	Version control, traceable decision records, model documentation, rater governance and data-protection procedures.
Transfer to work	Does educational assessment performance predict or support authentic occupational communication?	Longitudinal education-workplace linkage and workplace-expert judgement.

LIMITATIONS

The study is limited to Scopus and English-language records. Scopus-only retrieval was chosen to preserve within-index consistency for cited-reference coupling, but database coverage is discipline-dependent and no single bibliographic source should be assumed to capture the full literature (Gusenbauer, 2022; Kumpulainen & Seppänen, 2022). The expanded P2 query substantially improved relevant-record coverage within Scopus: all 31 final articles were retrieved by P2, whereas the strict P1 query captured only three. Even so, the search may miss records indexed elsewhere or studies using local terminology that does not appear in titles, abstracts or keywords. The final corpus is intentionally narrow and should be interpreted as a Scopus-bounded map of a specialised intersection rather than a complete map of language assessment or TVET research.

The bibliographic-coupling network was derived from flat Scopus reference strings and therefore required deterministic reconstruction and canonicalisation. Sensitivity analysis was stable at the component level: exact-only, prefix and tested fuzzy-matching variants retained a 19-document largest component, 12 isolates and 13 connected components, although link counts, summed edge weights and community assignments varied. These checks reduce but do not eliminate the possibility that unusual citation variants were missed or merged.

Full texts were available for 26 of the 31 articles. The five unavailable reports remained in the bibliometric corpus but were not used as primary evidence for the interpretive synthesis. Author-keyword recurrence was also sparse, which limited the value of conventional co-word clustering. These constraints are reported directly because forcing stronger thematic structure than the data support would reduce, rather than improve, interpretive reliability.

CONCLUSION

Technology-enhanced language assessment in TVET and occupational education is a small but increasingly visible intersection. Its literature spans automated scoring, AI feedback, computer-based testing, multimodal performance, rater support and workplace-oriented task design, yet the bibliographic network and keyword structure remain fragmented.

The strongest direction is not simply more automation. It is better alignment among occupational constructs, tasks, scoring evidence, human judgement and intended score use. Full-text synthesis shows why that distinction matters: technically sophisticated systems can still have limited evidence for particular uses, while focused design or psychometric studies can make valuable contributions when interpretations remain aligned with the evidence reported.

Future research should move from technology presence to evidence quality. Occupational authenticity, transparent human-AI decision rules, fairness, consequential score use and traceable governance provide a more defensible agenda for technology-enhanced language assessment in vocational and professional education.

ETHICS STATEMENT

This study analysed bibliographic records and published literature and did not collect primary data from human participants or animals. Institutional ethics approval and informed consent were therefore not applicable.

ACKNOWLEDGEMENTS

The authors acknowledge Universiti Teknologi Malaysia and the academic supervision received in developing the thesis-based foundation of this paper.

FUNDING

This work received no external funding.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY

The exact Scopus search strategy is provided in Appendix A. Screening decisions, processed bibliographic-coupling outputs and the Python/NetworkX analysis scripts are retained by the authors and can be provided on reasonable request. Raw Scopus exports are subject to database licensing and are therefore not redistributed publicly.

AUTHORS' CONTRIBUTIONS

Muslihah binti Saman: Conceptualisation, methodology, investigation, data curation, formal analysis, and writing – original draft. Mohd Fahmi bin Adnan: Methodology, supervision, and writing – review and editing. Both authors reviewed and approved the submitted manuscript.

DECLARATION OF AI-ASSISTED TOOLS

During manuscript preparation, the authors used OpenAI ChatGPT to support language editing and to assist with development and checking of the reproducible Python workflow and documentation. The authors independently reviewed the search strategy, screening decisions, source evidence, numerical results, interpretations and final manuscript, and take responsibility for the submitted work. All analytical figures were generated from the study data using Python, Matplotlib and NetworkX; no generative image model was used.

APPENDIX A: SCOPUS SEARCH STRATEGY

Final P2 Scopus advanced-search query used on 5 August 2026:

TITLE-ABS-KEY(

(
 "language assessment"
 OR "language testing"
 OR "English assessment"
 OR "English testing"
 OR "English language assessment"
 OR "English language testing"
 OR "second language assessment"
 OR "second language testing"
 OR "foreign language assessment"
 OR "foreign language testing"
 OR "L2 assessment"
 OR "L2 testing"
 OR "language proficiency assessment"
 OR "English proficiency assessment"
 OR "speaking assessment"
 OR "oral assessment"
 OR "oral proficiency assessment"
 OR "writing assessment"
 OR "pronunciation assessment"
 OR "communication skills assessment"
 OR "communicative competence assessment"
 OR (language W/3 assess*)
 OR (language W/3 test*)
 OR (English W/3 assess*)
 OR (English W/3 test*)
)

AND

(
 digital
 OR technolog*
 OR online
 OR computer*
 OR mobile
 OR automated
 OR automatic
 OR "e-assessment"
 OR "electronic assessment"
 OR "computer-based assessment"
 OR "computer-assisted assessment"
 OR "computer-adaptive testing"
 OR "online assessment"
 OR "digital assessment"
 OR "technology-enhanced assessment"
 OR "technology-based assessment"
 OR "technology-mediated assessment"
 OR "automated scoring"
 OR "automatic scoring"
 OR "automated feedback"
 OR "automatic feedback"
)

OR "artificial intelligence"
OR "generative AI"
OR "large language model*"
OR ChatGPT
OR chatbot*
OR "machine learning"
OR "natural language processing"
OR "speech recognition"
OR "learning analytics"
OR "multimodal assessment"
OR "simulation-based assessment"
OR "virtual reality"
OR "augmented reality"
OR "e-portfolio"
OR "electronic portfolio"
OR "digital portfolio"

)

AND

(

TVET
OR vocation*
OR polytechnic*
OR "technical education"
OR "career and technical education"
OR "occupational education"
OR "higher vocational education"
OR "secondary vocational education"
OR "vocational higher education"
OR "vocational college*"
OR "technical college*"
OR "skills training"
OR "work-based learning"
OR apprenticeship*
OR "competency-based education"
OR "competence-based education"
OR "competency-based training"
OR "competence-based training"
OR "vocational English"
OR "workplace English"
OR "professional English"
OR "English for occupational purposes"
OR "English for specific purposes"
OR "language for specific purposes"
OR EOP
OR ESP
OR "workplace communication"
OR "occupational communication"
OR "professional communication"
OR "workplace language"
OR "occupational language"

)

)

REFERENCES

1. Chen, A., Zhang, Y., Jia, J., Liang, M., Cha, Y., & Lim, C. P. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning*, 41(1), e13064. <https://doi.org/10.1111/jcal.13064>
2. Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, 134, 103846. <https://doi.org/10.1016/j.system.2025.103846>
3. Corrigan, J. A., & Slomp, D. H. (2021). Articulating a sociocognitive construct of writing expertise for the digital age. *Journal of Writing Analytics*, 5, 142–195. <https://doi.org/10.37514/JWA-J.2021.5.1.05>
4. Dai, X., Wang, Y., Yu, J., Qiu, B., Wang, R., & Na, M. (2026). Motivation and engagement as pathways: How AI-augmented online assessment shapes English-speaking competency in vocational EFL classrooms. *Frontiers in Psychology*, 16, Article 1730953. <https://doi.org/10.3389/fpsyg.2025.1730953>
5. De Felice, R., & Deane, P. (2012). Identifying speech acts in e-mails: Toward automated scoring of the TOEIC® e-mail task. *ETS Research Report Series*, 2012(2), i–62. <https://doi.org/10.1002/j.2333-8504.2012.tb02298.x>
6. Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
7. Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.
8. Doykova, I. (2025). Comparing AI- and instructor-designed ESP tasks for medical English students. *Analele Universității „Ovidius” Constanța. Seria Filologie*, 36(2), 1128–1141.
9. Dursun, A., Morris, J. K., & Ünalı, A. (2020). Designing proficiency-oriented performance tasks for the 21st-century workplace written communication: An evidence-centered design approach. *Assessing Writing*, 46, Article 100487. <https://doi.org/10.1016/j.asw.2020.100487>
10. Fortanet-Gómez, I., & Edo-Marzá, N. (2022). New genres and new approaches: Teaching and assessing product pitches from a multimodal perspective in the ESP classroom [Nuevos géneros y nuevos enfoques: La enseñanza y el aprendizaje del product pitch desde una perspectiva multimodal en la clase de inglés especializado]. *Porta Linguarum*, 38, 65–81. <https://doi.org/10.30827/portalin.vi38.21621>
11. Gao, F. (2026). MFF-PEA: An automatic assessment model for professional spoken English via multimodal feature fusion. *International Journal of Information and Communication Technology*, 27(5), 1–15. <https://doi.org/10.1504/IJICT.2026.151530>
12. Giray, L., & Ramos, C. C. (2025). Human subjectivity as defense against AI plagiarism in English for Specific Purposes (ESP) writing. *Journal of Teaching English for Specific and Academic Purposes*, 13(3), 391–404. <https://doi.org/10.22190/JTESAP250425030G>
13. Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, 127, 2683–2745. <https://doi.org/10.1007/s11192-022-04289-7>
14. Hamaniuk, V. A., Bilozir, O. S., & Semerikov, S. O. (2026). Mapping the research landscape of multilingual repertoires and linguistic justice in education: A systematic bibliometric analysis. *Educational Dimension*, 15, 239–270. <https://doi.org/10.31812/ed.836>
15. Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System*, 29(4), 505–520. [https://doi.org/10.1016/S0346-251X\(01\)00036-7](https://doi.org/10.1016/S0346-251X(01)00036-7)
16. He, T., Yao, J., Lai, Y., & Zou, Y. (2026). Optimization design of intelligent pronunciation assessment system based on speech recognition technology in higher vocational English classroom teaching under the background of industry-teaching integration. *International Journal of Computer Information Systems and Industrial Management Applications*, 18, 12. <https://doi.org/10.70917/ijcisim-2026-0029>
17. Jaturapitakkul, N., Horness, P., & Pansa, D. (2025). A Rasch-based validation of the reading section of Test of English for Thai Engineers and Technologists (TETET). *rEFLections*, 32(3), 1660–1681. <https://doi.org/10.61508/refl.v32i3.285703>
18. Kohnke, L., Jarvis, A., & Ting, A. (2021). Digital multimodal composing as authentic assessment in discipline-specific English courses: Insights from ESP learners. *TESOL Journal*, 12(3), Article e600. <https://doi.org/10.1002/tesj.600>
19. Kumpulainen, M., & Seppänen, M. (2022). Combining Web of Science and Scopus datasets in citation-

- based literature study. *Scientometrics*, 127, 5613–5631. <https://doi.org/10.1007/s11192-022-04475-7>
20. Lan, G., Li, Y., Yang, J., & He, X. (2025). Investigating a customized generative AI chatbot for automated essay scoring in a disciplinary writing task. *Assessing Writing*, 66, Article 100959. <https://doi.org/10.1016/j.asw.2025.100959>
21. Liu, A., Abdullah, A., & Meng, F. (2025). Mapping informal digital learning of English (IDLE): A bibliometric analysis of citation, co-occurrence, and co-citation (2000–2024). *Asia Pacific Journal of Educators and Education*, 40(3), 355–379. <https://doi.org/10.21315/apjee2025.40.3.14>
22. Lockwood, J. (2022). The design of a webchat assessment framework for contact centres in Asia. *Sociolinguistic Studies*, 16(1), 39–63. <https://doi.org/10.1558/sols.42361>
23. Lockwood, J. E. (2015). Language for Specific Purpose (LSP) performance assessment in Asian call centres: Strong and weak definitions. *Language Testing in Asia*, 5, Article 3. <https://doi.org/10.1186/s40468-014-0009-6>
24. McNamara, T. (2014). 30 years on—Evolution or revolution? *Language Assessment Quarterly*, 11(2), 226–232. <https://doi.org/10.1080/15434303.2014.895830>
25. Mirabile, M., Corazza, G. E., & Alonso-Moral, J. M. (2026). Trust in human–AI collaboration in finance: A bibliometric–systematic literature review. *AI & Society*. Advance online publication. <https://doi.org/10.1007/s00146-026-03049-y>
26. Mohsen, M. A., Althebi, S., Alsagour, R., Alsalem, A., Almudawi, A., & Alshahrani, A. (2024). Forty-two years of computer-assisted language learning research: A scientometric study of hotspot research and trending issues. *ReCALL*, 36(2), 230–249. <https://doi.org/10.1017/S0958344023000253>
27. Motteram, J., Spiby, R., Bellhouse, G., & Sroka, K. (2023). Implementation of an accommodations policy for candidates with diverse needs in a large-scale testing system. *Language Testing*, 40(4), 904–932. <https://doi.org/10.1177/02655322231166587>
28. Nazim, M., & Alzubi, A. A. F. (2022). Evaluation of an online teacher-made test through Blackboard in an English as a foreign language writing context. *World Journal on Educational Technology: Current Issues*, 14(4), 1025–1037. <https://doi.org/10.18844/wjet.v14i4.7614>
29. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
30. Pan, V. W.-C. (2026). AI-driven automated writing evaluation feedback and its role in developing EFL writing assessment literacy: A mixed-methods study. *Journal of Asia TEFL*, 23(1), 81–98. <https://doi.org/10.18823/asiatefl.2026.23.1.6.81>
31. Pan, Y. (2025). Leveraging generative AI powered rubric-indexed feedback as a formative assessment strategy for enhancing medical English education. *Discover Computing*, 28(1), Article 284. <https://doi.org/10.1007/s10791-025-09830-9>
32. Phaiboonnugulkij, M., & Prapphal, K. (2013). Online speaking strategy assessment for improving speaking ability in the area of language for specific purposes: The case of tourism. *English Language Teaching*, 6(9), 19–29. <https://doi.org/10.5539/elt.v6n9p19>
33. Rajamanickam, S., Che' Rus, R., Abdul Raji, M. N., Mina, H., & Vebrianto, R. (2025). Global trends in vocational training education: A bibliometric analysis. *International Journal of Research and Innovation in Social Science*, 8(12), 4548–4564. <https://doi.org/10.47772/IJRISS.2024.8120382>
34. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10, 39. <https://doi.org/10.1186/s13643-020-01542-z>
35. Rogers, G., Szomszor, M., & Adams, J. (2020). Sample size in bibliometric analysis. *Scientometrics*, 125, 777–794. <https://doi.org/10.1007/s11192-020-03647-7>
36. Salem, A. A. M. S. (2026). Generative AI-driven feedback in flipped writing workshops: Transforming business education through improved writing, communication, and self-regulation. *Journal of Educational Computing Research*. Advance online publication. <https://doi.org/10.1177/07356331261431156>
37. Shih, R. C. (2013). Effect of using Facebook to assist English for business communication course

- instruction. *Turkish Online Journal of Educational Technology*, 12(1), 52–59.
38. van den Branden, K., Depauw, V., & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes. *Language Testing*, 19(4), 438–452. <https://doi.org/10.1191/0265532202lt239oa>
 39. Vernice, M., Matta, M., Tironi, M., Caccia, M., Lombardi, E., Guasti, M. T., Sarti, D., & Lang, M. (2019). An online tool to assess sentence comprehension in teenagers at risk for school exclusion: Evidence from L2 Italian students. *Frontiers in Psychology*, 10, Article 2417. <https://doi.org/10.3389/fpsyg.2019.02417>
 40. Waiprakhon, T., & Jaturapitakkul, N. (2018). Test-taking strategies used in the reading section of the Test of English for Thai Engineers and Technologists: A computer-based ESP test. *PASAA*, 55, 147–177. <https://doi.org/10.58837/CHULA.PASAA.55.1.7>
 41. Wolf, R. R., & Wolf, A. B. (2023). Using AI to evaluate a competency-based online writing course in nursing. *Online Learning Journal*, 27(3), 41–69. <https://doi.org/10.24059/olj.v27i3.3974>
 42. Yang, Z., & Wang, P. (2025). Current status and research trend of English language assessment: A bibliometric analysis. *Language Testing in Asia*, 15, Article 11. <https://doi.org/10.1186/s40468-024-00317-w>
 43. Yusoff, Y., Tahar, E., Aziz, I. A., & Jusoh, R. (2026). Mapping the intellectual structure of youth financial literacy research: A bibliographic coupling analysis (2005–2025). *International Journal of Research and Innovation in Social Science*, 10(14), 1987–1999. <https://doi.org/10.47772/IJRISS.2026.1014MG0148>
 44. Zaki, N. A. M., Ghani, D. Z. A., Zakaria, M. I., & Suhaimee, N. R. A. (2026). Bibliometric analysis of teacher competency and pedagogical models for digital technology integration in school physical education. *International Journal of Research and Innovation in Social Science*, 10(2), 5828–5850. <https://doi.org/10.47772/IJRISS.2026.10200429>
 45. Zapata-Rivera, D., Andrews-Todd, J., & Oliveri, M. E. (2021). Communicating assessment information in the context of a workplace formative task. *Journal of Writing Analytics*, 5, 324–341. <https://doi.org/10.37514/JWA-J.2021.5.1.10>
 46. Zhu, Y., & Dong, M. (2025). Decoding the raters' mind: A metadiscourse investigation of ChatGPT vs EFL instructors in Business English writing assessment. *Studies in Educational Evaluation*, 87, Article 101541. <https://doi.org/10.1016/j.stueduc.2025.101541>