



RESEARCH ARTICLE

Offline Digital Assessment in Aviation Vocational Education: Architecture, Continuity, and Human Oversight

Yew Yun Xin¹, Laila Rochmawati^{2,*}, Harunur Rosyid³

¹Institut Pendidikan Guru Kampus Bahasa Antarabangsa, Malaysia

²Politeknik Penerbangan Surabaya, Indonesia

³Universitas Muhammadiyah Gresik, Indonesia

*Corresponding author: lailarochmawati@poltekbangsby.ac.id

Article History

Received: 13 October 2026

Revised: 18 October 2026

Accepted: 20 October 2026

Published: 22 October 2026

Publication Information

DOI: 10.36456/b9sk2k62

Volume / Issue: 3 / 1

Pages: 22-38

Keywords

assessment-state continuity; aviation vocational education; digital assessment; evidence traceability; human oversight

ABSTRACT

Offline assessment can reduce reliance on continuous network connectivity, but local availability alone does not demonstrate preservation of assessment evidence after an interruption. This artifact-centered study examines ToeicApp v2.0, a locally installed TOEIC-oriented English assessment application situated in aviation vocational education. Drawing on a technical-audit record, it combines a requirement-to-evidence audit framework, a bounded description of the implemented assessment architecture, and transferable design propositions. The reported audit identifies a .NET Framework 4.7.2 application with SQLite, five packages, 232 question groups, 999 items, 999 answer-key entries, 154 audio paths, and 50 image paths. Session-state fields and documented timer, autosave, resume, history, and import/export mechanisms indicate an architecture designed to link item delivery, response persistence, and record access. The study distinguishes database and resource observations from function traces and pedagogical interpretations; no controlled interruption trials or independent audit reproduction were available. Human oversight denotes the boundary between software-mediated evidence handling and instructor-mediated interpretation, feedback, and follow-up, not an artificial-intelligence feature. Comparison with offline-first learning-management and digital-assessment literature locates the incremental contribution in assessment-specific traceability and explicit claim boundaries rather than novel storage technology. The findings do not establish recovery reliability, psychometric validity, official TOEIC equivalence, institutional connectivity failures, or learning effects.

INTRODUCTION

Computer-based language assessment reorganizes more than the presentation of questions: it changes how items and multimedia are delivered, how responses are captured, and how assessment evidence becomes available for interpretation (Douglas & Hegelheimer, 2007; Jamieson, 2005; Ockey, 2009). Wider assessment scholarship similarly treats digital systems as evidence infrastructures whose design matters for recording, feedback, and consequential decisions (Bennett, 2015; Fuller et al., 2022). A list of software features is therefore insufficient unless those features can be related to defensible assessment functions and appropriately limited claims (Babik et al., 2024; Sembey et al., 2024).

This issue is pertinent to technical and vocational education and training (TVET), in which technology adoption is mediated by infrastructure, governance, staff capability, and occupational prepa-

ration (Parua & Yang, 2024; Zhong & Juwaheer, 2024). Reviews identify institutional readiness and educator competence as recurring conditions (Othman et al., 2025; Witarsyah et al., 2026), while vocational-college evidence stresses that technology deployment and instructional capacity are interdependent (X. Ma et al., 2025; Metreveli et al., 2025; Setiyawan et al., 2025; Widaningsih et al., 2024; Xin et al., 2024). In aviation vocational education, English-language learning has professional relevance, but that relevance must not be mistaken for proof of a specific language-testing construct. Aviation English research concerns specialized communication demands (Zhao, 2024), student-pilot needs (Treadaway & Read, 2025), and technology-supported training and testing research (Demirdöken, 2024; Demirdöken & Atay, 2024). The present application, however, is described more narrowly as a TOEIC-oriented English assessment-practice tool, not a validated Aviation English proficiency examination.

The architectural challenge becomes more pronounced when reliable network connectivity cannot be assumed. Offline-first work, including the decentralized ALEX learning-management system (Duarte et al., 2025), establishes the relevance of locally available resources and activities. Timed assessment adds a different set of conditions: item resources must remain available, responses must be associated with an ongoing attempt, and enough session state must persist to permit a possible continuation. Digital-assessment literature discusses evidence retention and teacher use (Fuller et al., 2022; Grosseck et al., 2024; Kaya-Capocci et al., 2022), while language-assessment literature separates technical affordances from validity claims (Kane, 2013; Sadeghi, 2023). These bodies of work provide complementary principles but do not, by themselves, supply an artifact-level, assessment-specific audit account linking local content, interrupted-attempt state, retained evidence, and responsibility for interpretation in the case studied here. This integration is the limited research gap addressed by the present investigation; it is not a claim that any individual component is technically unprecedented.

ToeicApp v2.0 was examined as a locally installed application used in an aviation vocational education context. The retained record identifies its TOEIC-oriented assessment purpose and its local deployment, but does not document a particular cohort, observed network failures, institutional connectivity measurements, instructors' actual reporting routines, or a contemporaneous decision to choose the application over a cloud alternative. The relevance of intermittent connectivity is thus an architectural motivation supported by wider literature, not an empirical finding about the institution. Likewise, local record access creates opportunities for instructor review without demonstrating how instructors used reports in classrooms.

This study makes three specific, bounded contributions: (a) an assessment-specific audit framework connecting eight operational requirements with identifiable evidence routes and claim limits; (b) an analytical architecture integrating local resources, session-state mechanisms, record traceability, and an explicit pedagogical decision boundary; and (c) conditional design propositions and verification priorities for similar vocational settings. SQLite, timers, autosave, and instructor judgment are established concepts applied to this case, not newly invented mechanisms. The study is an artifact-centered technical case analysis, not a full design-science cycle, controlled reliability experiment, institutional ethnography, or psychometric validation study (Hevner et al., 2004; Jacobsen & McKenney, 2024; Peffers et al., 2007).

The four research questions are: RQ1. How can documented system functions be mapped to operational requirements for offline vocational assessment? RQ2. How does the reported architecture organize local content delivery, multimedia access, response persistence, session continuity,

and retrieval of records? RQ3. Which operations are software-mediated, and which pedagogical decisions are outside the documented automated scope? RQ4. What conditional design propositions follow from the implementation and its evidence limitations?

CONCEPTUAL AND DESIGN FRAMEWORK

Terminology and the Assessment Evidence Chain

Offline assessment here means that core delivery and locally stored evidence are not predicated on continuous external network service. Offline-first architecture is the broader design principle of prioritizing local operation; it does not imply that ToeicApp implements the synchronization protocols of a distributed offline-first learning-management system. Assessment-state continuity refers to the presence of resources, saved response evidence, and sufficient attempt metadata (such as position, time, and status) to support potential resumption. Its demonstrated reliability is a separate empirical question. Evidence traceability refers to the possibility of relating reported content and stored records to a defined audited artifact and the claims derived from it, rather than to proof of score quality or secure provenance.

The distinction between infrastructure and interpretation also establishes the meaning of human oversight in this article. Software-mediated functions include item delivery, timing, response storage, and record retrieval, where the retained audit indicates mechanisms. Instructor-mediated functions comprise contextual interpretation of performance, diagnosis, pedagogical feedback, and follow-up, none of which the audit demonstrates as automated. This is a responsibility-allocation model, not an observation of teacher practice and not a claim that ToeicApp uses artificial intelligence. Assessment-literacy scholarship provides its pedagogical basis (Ali, 2024; Hidri et al., 2026; Lei & Lei, 2026; Vogt et al., 2024; Zhang et al., 2025). AI human-in-the-loop research is used only as an analogy for explicit decision boundaries (Fajardo-Ramos et al., 2025; Khan et al., 2026; Swiecki et al., 2022).

Comparison with Prior Systems and Frameworks

Table 1 contrasts claims made by the published literature with properties recorded for the present artifact. ALEX is an offline-first, decentralized LMS addressing continuity of learning activity (Duarte et al., 2025); it should not be assumed to be a timed English-test system. In contrast, educational digital-assessment frameworks discuss the assessment lifecycle and teacher-facing evidence requirements without supplying direct test results for ToeicApp (Fuller et al., 2022; Kaya-Capocci et al., 2022; Sembey et al., 2024). The comparison identifies an analytical intersection rather than a claim of superiority.

Table 1. Comparison of Prior Work and the Study's Incremental Contribution

Analytical dimension	ALEX: Duarte et al. (2025)	Digital-assessment scholarship	Present artifact / contribution
Offline operation	Decentralized offline-first LMS	Discussed as a platform/design condition	Local .NET application and SQLite; documented offline-capable configuration
Response persistence	Continued local learning activity; not a study of this timed test	Evidence capture and retention as assessment infrastructure	Autosave and response-related session fields mapped to audit evidence

Continued on next page

Table 1 (continued)

Analytical dimension	ALEX: Duarte et al. (2025)	Digital-assessment scholarship	Present artifact / contribution
Recovery after interruption	Assessment-trial recovery not established by the cited comparison	Important but context-dependent requirement	Resume pathway traced; controlled recovery success not measured
Evidence traceability	LMS architecture; no claim of equivalent audit measures	Links evidence capture with assessment purposes	Reported item, key, media and integrity counts tied to a bounded claim register
Human oversight	Not treated here as a directly comparable test-scoring boundary	Assessment literacy and teacher evidence use	Explicit hand-off from technical records to instructor judgment
Distinct contribution	Establishes offline-first educational design precedent	Provides pedagogical and validity principles	Combines requirements, verification modes, traceability, and claim limits for a vocational assessment artifact

Note. Comparison is analytical, not a controlled benchmark. “Not established” means that the cited study is not used here as evidence for the specified timed-assessment claim; it does not demonstrate that a feature is absent.

Validity and Design-Knowledge Boundaries

Kane’s (2013) argument-based view of validation requires interpretations and uses of assessment evidence to be supported by their own warrants. Equivalence across vocational assessment formats, for example, requires targeted measurement evidence rather than successful digital delivery alone (B. Ma et al., 2025). Digital formative assessment further involves the instructional purposes and practices in which evidence is used (Casanova et al., 2021; Grosseck et al., 2024; Kaya-Capocci et al., 2022). The present descriptive audit can therefore support selected implementation claims and design propositions but cannot certify item validity, fairness, security, measurement equivalence, learning gains, or teacher decision quality.

METHOD

Research Design, Setting, and Unit of Analysis

This study retrospectively analyzed the retained technical audit of ToeicApp v2.0 as a single artifact-centered case. The unit of analysis was the installed assessment environment and its retained audit record, not individual learners, educators, test scores, or a campus connectivity intervention. The original study identifies an aviation vocational education setting and TOEIC-oriented English assessment practice, but the retained record does not establish the exact award level, program, or cohort. The source record does not establish whether the application was deployed for formative exercises, summative grading, standardized-test preparation, or more than one of these purposes in a particular class. Its use as a locally installed test-practice environment is the supported contextual description; more specific pedagogical use should be investigated separately.

Artifact Identification and Audit Environment

The retained audit identifies a distribution archive designated ToeicApp_v2.0, an executable ToeicApp.exe, a .NET Framework 4.7.2 configuration, and a local SQLite file toeic.db. It reports a primary/backup snapshot comparison and SHA-256 identifiers for the archive, executable,

database, and manual, but the retained materials available for this analysis do not contain the digest strings or original audit execution log. The operating-system build, SQLite runtime version, audit date, hardware profile, query-tool version, and precise comparison method were not recorded in the supplied manuscript. These omissions are disclosed, rather than replaced with presumed values, because an independently reproduced audit must be associated with an identifiable software and execution environment (Table 2).

Table 2. Recorded Audit Environment and Unavailable Replication Metadata

Parameter	Retained audit description	Status for independent replication
Artifact / executable	ToeicApp v2.0; ToeicApp_v2.0; ToeicApp.exe	Identified by name; original archive not supplied
Runtime and local storage	.NET Framework 4.7.2; SQLite; toeic.db	Framework recorded; SQLite engine version not disclosed
Operating system / hardware	Not provided in retained summary	Must be recorded from original audit host
Audit date and audit tool versions	Not provided in retained summary	Must be supplied with dated logs
Database check output	PRAGMA integrity_check = ok; foreign_key_check without findings	Reported output; raw logs not independently available
Snapshot / cryptographic identity	Backup reported to match primary; SHA-256 identifiers reportedly retained	Comparison mode and complete digests not available for verification
Functional execution logs	No test scenarios or direct-run records included	Cannot classify reported functions as passed end-to-end trials

Note. “Not provided” is a disclosure of the available record, not an assertion that the original auditor failed to collect the value.

Evidence Sources and Verification Route Classification

Four evidence routes were distinguished. Database inspection (DB) concerns recorded table structures, counts, and session fields. Resource inspection (RES) concerns references in the local data store and their reported correspondence with distributed media. Functional tracing (FUNC) concerns mechanisms identified in the build and corroborated, where applicable, by an operational manual; no documented direct execution trial is available for the present analysis. Integrity/identity review (INT) concerns retained database-check outputs, a reported snapshot comparison, and statements that SHA-256 identifiers had been calculated. The manual is corroborative (DOC), not the source of inventory totals. Because the raw logs and original artifacts were unavailable for the present analysis, statements about the audit are classified as reported audit observations rather than newly executed or independently confirmed checks.

This distinction is material for autosave, resume, history, and import/export. The manuscript’s record supports the existence or reported traceability of these functions, but does not show a sequence of executed test cases with expected outputs, timestamps, saved-response comparisons, or numbers of successful resumptions. Accordingly, no feature is assigned a functional pass rate or an observed recovery outcome. The analysis retains the originating audit’s terminology for implemented mechanisms while distinguishing documented functionality from independently verified operational performance. No unperformed interruption experiment is presented as empirical evidence.

Operational Definitions and Reproducibility Protocol

Question groups are rows identified as `question_groups` in the retained schema; assessment items are entries in questions, and answer-key entries are entries in `answer_keys`. Inventory counts describe database records and do not assert content uniqueness, item quality, or that a recorded answer key has a valid one-to-one psychometric correspondence. Package totals A–E refer to the reported item membership categories. Audio and image counts represent distinct non-empty referenced path strings within each modality, not counts of acoustic or visual files independently decoded. No documented normalization or case-folding rule for media paths is available; the original uniqueness procedure therefore remains a reproducibility limitation. “No missing referenced paths” denotes the recorded resource-to-file correspondence for the audited distribution; it does not establish that all resources open correctly or that their use is legally authorized.

For continuity, session-state preservation is operationalized narrowly as the presence of the `current_index`, `remaining_secs`, `user_answers`, and `status` fields in `test_sessions` together with the reported autosave/resume mechanisms. It is not operationalized as observed response recovery. Database consistency refers to the reported PRAGMA `integrity_check` result, the reported absence of foreign-key check findings, and the stated agreement of primary and backup database snapshots; the audit summary does not document whether the latter agreement is byte-wise, rowwise, or based on a hash. Supplementary Material S1 separates reported values from fields awaiting source-level verification and provides SQL examples for future independent checks. These examples constitute a proposed replication protocol; the queries were not executed as part of the present analysis.

Eight Operational Requirements and Claim-Acceptance Rules

The eight requirements were constructed retrospectively from the assessment activity chain and recorded functions, not copied from a pre-existing vendor specification. Offline operability, structured content, multimedia access, temporal control, response persistence, session continuity, record management, and pedagogical interpretation cover the sequence from resource availability to instructional use. A requirement is mapped as “reported implemented” only when at least one named configuration, schema/resource observation, or traceable mechanism directly corresponds to it. Documentary description alone may corroborate such a mapping but cannot establish robustness or effective learner use. A requirement without such supporting evidence would be recorded as unverified, not inferred from desired practice. Evidence and exclusions for each category are made explicit in Table 3.

Five-Dimension Analysis and Research-Question Alignment

The audit integrated five dimensions: (1) requirements linked to functions, (2) system architecture and session flow, (3) content/media inventory and integrity, (4) allocation of system and instructor responsibilities, and (5) claim boundaries. The analytical sequence classified evidence by route and certainty, mapped each item to an operational purpose, traced dependencies across resources–sessions–records, and compared the strongest warranted claim with propositions needing other evidence. This combination matters because no single dimension can establish whether recorded functions are sufficient for an accountable assessment service. RQ1 is addressed by the operational mapping in Table 3; RQ2 by the resource and architecture results in Tables 4–5 and

Figures 1–2; RQ3 by Table 6 and Figure 3; and RQ4 by the claim matrix in Table 7 and conditional design propositions in the Discussion.

Ethical Scope and Data Governance

No learner-level responses or identifiable participant dataset were used in the present manuscript analysis, and no individual learner record is reproduced. However, the retained description of tables `test_sessions`, `test_results`, and `test_results_v2` does not determine whether the underlying audited database contained live learner records, dummy entries, or empty learner-facing tables. Likewise, the source record does not specify auditor access permissions, the original data-minimization procedure, or data-retention controls. Those conditions cannot be asserted retrospectively. The present study reports only schema names, aggregated inventory figures, and function categories; the original database and restricted assessment content are not included. A future data-governance review should verify whether personal data were present and document authorization, confidentiality measures, access logging, and any necessary de-identification. Local storage and transfer facilities are not evidence of security, privacy-law compliance, or encryption adequacy.

RESULTS AND DISCUSSION

Results

RQ1: Operational Requirements and Evidence Boundaries

The analysis mapped eight operational categories. Seven technical categories were supported to varying degrees by reported configuration, database and resource observations, or functional traces. The eighth category, pedagogical interpretation, was treated as a proposed instructor-mediated responsibility rather than an observed or automated function. Table 3 distinguishes these evidence levels and separates documented mechanisms from independently verified performance. The strongest mappings concern the content inventory and recorded database checks; interruption recovery and instructional use would require direct task observation or controlled trials.

Table 3. Operational Requirement-to-Function Mapping with Claim Limits

Requirement	Assessment purpose	Reported evidence / route	Acceptable inference
Offline operability	Local delivery without continuous network access	.NET configuration + SQLite (DB/DOC)	Local-deployment architecture; no comparative uptime result
Structured content	Organize assignable test resources	5 packages; 232 groups; 999 items (DB)	Inventory implemented; no form equivalence
Multimedia access	Reference audio and image resources	154 audio / 50 image paths (RES)	Resources referenced and reported present, not media quality
Temporal control	Manage attempt duration	Timer mechanism (FUNC/DOC)	Timing exists; official TOEIC timing not established
Response persistence	Retain intermediate response evidence	Autosave mechanism; <code>user_answers</code> (FUNC/DB)	Continuity support traced; zero data loss untested
Session continuity	Potentially reopen a prior attempt	Resume; <code>current_index</code> , <code>remaining_secs</code> , <code>status</code> (FUNC/DB)	Resume mechanism documented; recovery success untested

Continued on next page

Table 3 (continued)

Requirement	Assessment purpose	Reported evidence / route	Acceptable inference
Record management	Retrieve and transfer results	History; import/export (FUNC/DOC)	Record-access functions reported; interoperability untested
Pedagogical interpretation	Use records for educational action	Not identified as automated in audit (allocation analysis)	Instructor mediation proposed; classroom execution not observed

Note. DB = recorded database inspection; RES = resource inspection; FUNC = function tracing; DOC = manual/configuration corroboration. None of these labels denotes a new end-to-end test performed for this article.

Artifact Inventory, Integrity, and Traceability

The retained audit reports five content packages (A=200, B=200, C=199, D=200, E=200 items), totaling 999 items grouped into 232 question_groups. It also reports 999 answer-key entries, 154 unique audio reference paths, and 50 unique image reference paths (Table 5). The numerical equality between item and key counts is descriptive; item-by-item validity and scoring correctness cannot be inferred. Reported media-path completeness means that no database-referenced media path was missing from the audited distribution under the original resource inspection. It does not establish media decoding, accessibility, licensing, or authenticity.

The audited SQLite tables named in the retained record are question_groups, questions, answer_keys, test_sessions, test_results, test_results_v2, app_info, and part_intro. The reported integrity_check result was “ok”; foreign_key_check reportedly returned no findings. The retained summary additionally states that the primary database matched the backup at the audited snapshot and that SHA-256 identifiers were recorded for the archive, executable, database, and manual. Because the exact digests, query outputs, and comparison specification were not furnished, these are traceability claims reported by the originating audit rather than hashes recomputed or independent consistency checks performed in the present revision. S1 records both the reported outcomes and the missing evidence required for source-level confirmation.

RQ2: Components and Local Evidence Flow

Table 4 and Figure 1 represent different views of the reported implementation. Table 4 classifies functional components; Figure 1 illustrates their dependency sequence: locally installed resources and application execution precede session handling, structured persistence, and retrieval. The figure is a deployment/evidence pathway, not a map of instructors’ cognitive decisions. The large resource inventory requires a correspondence among content identifiers, multimedia paths, session references, and record retrieval if the resulting evidence is to remain interpretable. The audit establishes selected links at schema and trace level but does not demonstrate complete end-to-end referential provenance for every attempted item.

Table 4. Documented Components and Their Assessment Roles

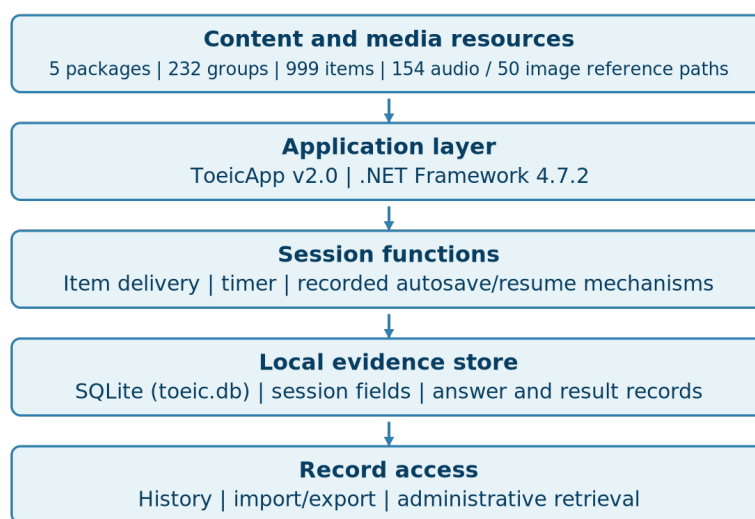
Component	Recorded implementation	Interpretive role
Application	ToeicApp v2.0; .NET Framework 4.7.2	Local desktop delivery environment
Structured storage	SQLite; toeic.db	Persists records in a local database
Content and key layer	Packages A–E; 232 groups; 999 items and keys	Organizes items and answer references

Continued on next page

Table 4 (continued)

Component	Recorded implementation	Interpretive role
Media references	154 audio paths; 50 image paths	Connects assessment resources to media
Session-state layer	Timer; autosave; resume; named session fields	Provides documented continuity-related mechanisms
Administrative access	History; import/export	Provides record retrieval/transfer functions
Pedagogical interpretation	No automated diagnosis/feedback established	Identifies proposed instructor decision boundary

DOCUMENTED LOCAL ASSESSMENT ENVIRONMENT



Arrows describe the reported functional pathway, not tested reliability or data-security guarantees.

Figure 1. Technical architecture and evidence path of the local assessment environment.

Note. The diagram shows the reported ordering of resources, executable, session mechanisms, local evidence store, and retrieval. Arrows are an analytical representation of documented relationships, not proof of tested performance, security, or uninterrupted recovery. Source: Authors' visualization based on the reported technical audit.

Content and Media Inventory

Table 5. Descriptive Content and Media Inventory

Reported resource	Count	Interpretation
Assessment packages	5	Packages A–E
Question groups	232	Rows in <code>question_groups</code> as reported
Assessment items	999	A=200, B=200, C=199, D=200, E=200
Answer-key entries	999	Entries recorded; item-wise correctness not tested
Unique audio paths	154	Distinct referenced strings, as reported
Unique image paths	50	Distinct referenced strings, as reported

Note. These are audit-summary counts, not independently recounted totals or evidence of psychometric quality.

Assessment-State Continuity: Capability versus Recovery Evidence

The documented timer, autosave, and resume functions constitute a set of mechanisms intended to support assessment-state continuity. Specifically, the stored field names `current_index`, `remaining_secs`, `user_answers`, and `status` indicate the kinds of attempt metadata relevant to resumption. Figure 2 distinguishes the reported normal pathway from the hypothesized interruption and resumption pathway. The original audit record available for this study did not include a verified sequence of response entry, forced interruption, reopening, and comparison of pre-interruption and recovered answers. The statement “resume is implemented” is therefore materially weaker than “interrupted responses were recovered in N/N trials.” Neither a recovery rate nor a data-loss probability can be estimated.

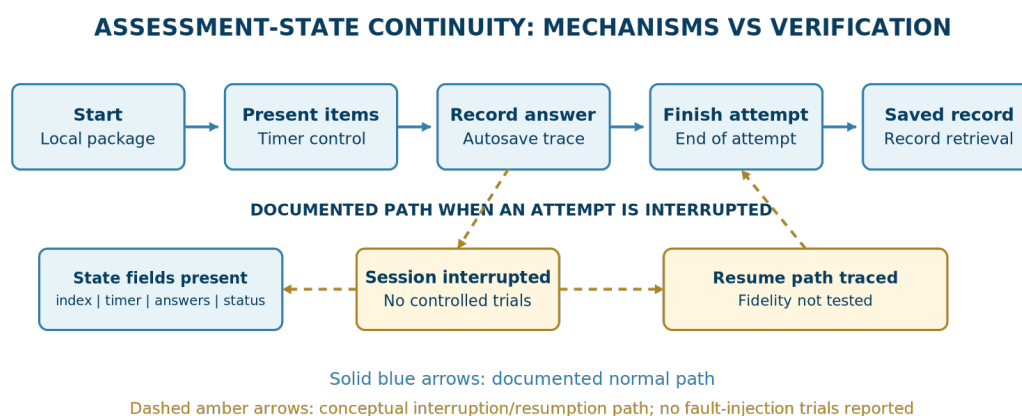


Figure 2. Assessment-state continuity model and verification boundary.

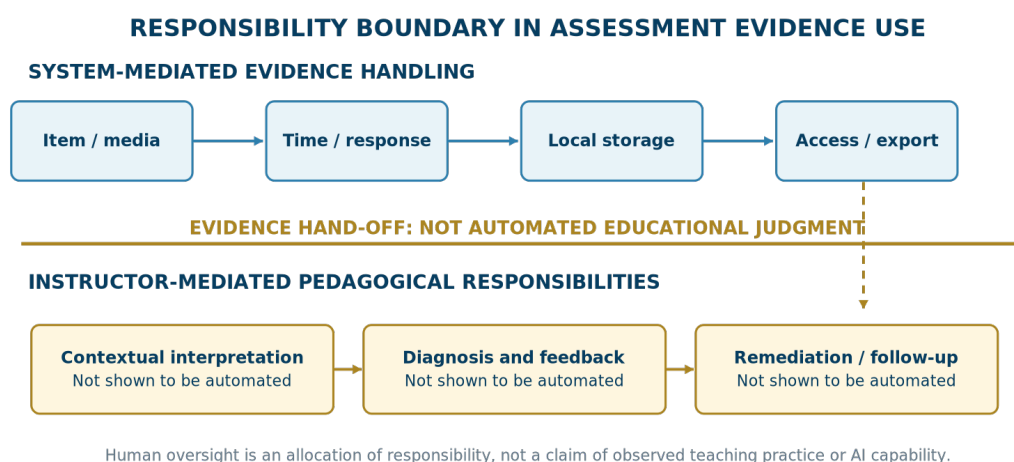
Note. Solid arrows depict the normal assessment evidence pathway described in the record. Dashed arrows depict the potential interruption/recovery route that was not tested through controlled fault-injection trials; the diagram does not report observed recoveries. Source: Authors’ visualization based on the reported technical audit.

RQ3: Allocation of Software and Instructor Responsibilities

Table 6 makes the functional allocation explicit. The application’s reported role is technical evidence handling: presentation, timing, storage, and access. Interpreting scores for educational purposes, diagnosing language difficulties, providing individualized feedback, and choosing remediation were not shown to be automated. Figure 3 therefore represents the responsibility boundary between a software-generated evidence record and a teacher’s contextual judgment. Unlike Figure 1, which focuses on the arrangement of system components, Figure 3 models accountability for interpretation. Neither the table nor the diagram implies that instructor actions were directly observed or that human decisions are intrinsically accurate.

Table 6. Functional Allocation and Observational Status

Function	Proposed primary locus	Evidence status / boundary
Item and multimedia delivery	Software-mediated	Implementation reported; learner experience not observed
Timing and response capture	Software-mediated	Mechanisms traced; time accuracy not bench-tested
Autosave and resume	Software-mediated	Mechanisms present; recovered-answer integrity not trial-tested
History and record retrieval	Software-mediated	Access function traced; archive durability untested
Import/export	Shared, administrative	Software mechanism; staff choice and governance required
Performance interpretation	Instructor-mediated	Outside documented automated functionality
Diagnosis, feedback and remediation	Instructor-mediated	Not documented as automated; actual teacher practice unobserved

**Figure 3.** Human-oversight model: evidence handling and pedagogical decision boundary.

Note. The upper domain covers the system functions supported by the audit description. The lower domain marks instructor-mediated decisions proposed by the analytical model; it is not an observed workflow and does not imply artificial-intelligence functions. Source: Authors' visualization based on the reported technical audit.

RQ4: Claim-Evidence Consolidation

Table 7 consolidates what each reported observation warrants, the claim that it does not warrant, and the research question primarily informed. This provides an explicit analytical framework for distinguishing three levels of claims used throughout the article: a reported audit observation; an analytical interpretation of that observation; and a conditional proposition that remains open to evaluation elsewhere.

Table 7. Claim–Evidence Matrix and Research-Question Linkage

Reported evidence	Supported narrow claim	Not established / future test	RQ
.NET and SQLite	Locally deployed assessment application	Higher uptime, security, or compliance	1–2
Packages, groups, items and keys	Structured assessment-resource inventory	Equivalence, calibration or test validity	1–2
Referenced audio/image paths	Multimedia resources inventoried	Media authenticity or playback reliability	2
SQLite checks and snapshot report	Consistency outputs recorded by original audit	Independent identity or durability assurance	2
Timer and autosave	Temporal/persistence mechanisms documented	Correct timekeeping or zero data loss	1–2
Resume and state fields	Attempt-state mechanisms recorded	Successful recovery after controlled faults	2
History and import/export	Record-access facilities reported	Long-term retention or interoperability	1–2
No automated pedagogical diagnosis shown	Decision boundary can be specified	Observed instructor effectiveness	3–4

Note. RQ4 integrates the matrix across all other questions; no row alone constitutes proof of design effectiveness.

Discussion

Integration of Five Audit Dimensions

Integrating the five audit dimensions provides an assessment-specific analysis that extends beyond a simple inventory of software features. Requirement mapping identifies why a function matters; architecture mapping identifies upstream and downstream dependencies; inventory and integrity observations constrain what can be said about content and stored data; functional allocation identifies who is responsible for interpretation; and the claim boundary specifies the evidence needed before stronger conclusions can follow. Together, these dimensions form a lightweight audit-and-interpretation framework for local digital assessment rather than a newly implemented software protocol. The contribution is incremental: it makes assessment-service dependencies and epistemic limits auditable in one case, while the individual technical components remain established practices.

Consequential Design Trade-Offs

The first trade-off is local operational autonomy versus evidence governance. SQLite and an installed application reduce dependence on an external service at the point of use, but local possession of results also makes device management, access authorization, backups, export rules, and retention policies institution-specific responsibilities. The current evidence characterizes the storage and transfer mechanisms, not the adequacy of those safeguards. This distinction is important for institutions interpreting “offline” as a sufficient procurement criterion (Parua & Yang, 2024; Zhong & Juwaheer, 2024).

The second trade-off is continuity-oriented design versus verified resilience. Persistence fields and a resume path are necessary architectural ingredients for a potentially recoverable attempt, but their presence cannot substitute for trials of abrupt termination, device restart, corrupted storage, or concurrent access. Future verification should compare exact pre- and post-interruption answers

as well as timer and status fields under a defined matrix of fault conditions. The existing results justify prioritizing these tests, not estimating an unmeasured success rate.

The third trade-off is resource coverage versus maintenance and evidence quality. An inventory of 999 items, 154 distinct audio-path values, and 50 distinct image-path values demonstrates nontrivial content-management demands. Larger resource sets increase the practical need for stable identifiers, path checks, version control, and traceable answer-key maintenance; the recorded count does not establish difficulty balance, standardized equivalence, or accessibility. Measurement quality should be evaluated with item-level and use-specific evidence after technical traceability is established (Kane, 2013; B. Ma et al., 2025; Sadeghi, 2023).

The fourth trade-off is automated evidence handling versus accountable pedagogical meaning. Software can reduce repetitive administrative work, yet moving scores and records into a database does not make them automatically actionable feedback. Instructors need assessment literacy and contextual information, while administrators need transparent rules on who can retrieve or share records. This responsibility allocation is consistent with teacher assessment-literacy and agency research (Casanova et al., 2021; Meijer et al., 2025, 2026; Vogt et al., 2024); it is not a comparison of machine and human decision accuracy.

Relevance and Transferability to Aviation Vocational Education

The aviation setting provides a plausible application domain for structured English practice, including multimedia listening resources that are relevant to vocational language preparation (Demirdöken & Atay, 2024; Treadaway & Read, 2025; Zhao, 2024). The evidence does not demonstrate that connectivity was unreliable at the observed institution, that this application was selected because of connectivity outages, or that instructors adopted particular reporting routines. Those claims require institutional interviews, classroom observations, or network/service records. The study therefore offers a technical case with potential relevance to aviation and other vocational institutions rather than an empirically validated solution to a documented local connectivity problem.

This setting also illustrates why content orientation and the claimed construct must be separated. A TOEIC-oriented practice tool operating in an aviation college is neither an officially licensed TOEIC testing platform nor, without construct evidence, a standardized Aviation English proficiency assessment. The described architecture may support the administration and retention of English-practice responses, but cannot confer official test interpretation or aviation-language validity on those records.

Conditional Design Propositions and Next-Step Tests

Four conditional propositions follow. P1: assess offline assessment as a three-part service—resource availability, response persistence, and attempt-state preservation—rather than launchability alone. P2: accompany feature inventories with named evidence sources, audit statuses, and source-level traceability. P3: make the technical-to-pedagogical hand-off and the governance of locally retained records explicit in system specification. P4: require a staged evaluation in which reproducible artifact checks precede controlled recovery testing and assessment-validity studies. These propositions are design recommendations derived from the single-case audit, not demonstrated causal effects or universal performance standards.

A practicable follow-up would capture the exact artifact hash, audit date, operating system and

SQLite version; export aggregate inventory queries and referential-check outputs; and conduct a preregistered continuity test covering ordinary close/reopen, forced process termination, host restart, and resource-path exceptions. For each scenario, investigators should state the number of attempts, expected saved items and timing state, observed recovered responses, and deviations. A separate educational study could then examine instructor use of records and the measurement properties of the item bank. None of these experiments was conducted as part of the evidence reported here.

Limitations

This work remains an artifact-centered account of a reported technical audit. It lacks controlled interruption or fault-injection testing, direct evidence of recovery success, longitudinal record durability, formal cybersecurity/privacy assessment, psychometric validation, learner achievement data, direct instructor observations, or independent replication. Audit date, host operating-system version, SQLite engine version, complete recorded hashes, media-path enumeration, raw database-check logs, and the method used to compare database snapshots were absent from the retained materials available for this analysis. The status of any identifiable learner information in the source database is unresolved; the present article does not analyze individual records. These gaps limit both source-level reproducibility and governance assertions. The study concerns one software snapshot and one broadly described vocational context, so the propositions should be tested in other settings before claims of reliability, generalizability, or effectiveness are advanced.

CONCLUSION

The retained audit of ToeicApp v2.0 documents a locally installed .NET Framework 4.7.2/SQLite assessment environment comprising five packages, 232 question groups, 999 items, 999 answer-key entries, 154 audio paths, and 50 image paths. It also reports timer, autosave, resume, history, and import/export mechanisms and session fields relevant to potential attempt continuation. These observations support a description of implemented continuity-oriented and evidence-retention mechanisms, not an estimate of successful recovery after interruption.

The study's principal contribution is the assessment-specific combination of a requirement-to-evidence audit framework, an architectural trace of resources-sessions-records, and conditional design propositions with explicit claim limits. Human oversight is operationally the distinction between software-mediated evidence handling and instructor-mediated pedagogical interpretation, feedback, and follow-up; it neither denotes artificial intelligence nor demonstrates teachers' actual decisions. Comparisons with ALEX and digital-assessment scholarship clarify that established local-storage and assessment principles are applied to a new technical case rather than claimed as novel inventions.

Before stronger institutional claims are justified, the originating audit needs complete provenance and a shareable anonymized evidence register; controlled recovery, data-governance, psychometric, and classroom-use investigations are also needed. Within those boundaries, the case shows how offline assessment may be evaluated through continuity, traceability, and responsibility rather than through software functionality alone.

DECLARATIONS

Acknowledgments

Not applicable.

Author Contributions

Y.Y.X.: Conceptualization, investigation, writing—original draft. L.R.: Methodology, supervision, writing—review and editing. H.R.: Software, validation, visualization, writing—review and editing. All authors: Approval of the final manuscript.

Ethics Approval and Informed Consent

The article is an analysis of a software artifact and aggregate technical-audit descriptions; no human participants were recruited and no learner-level response dataset was analyzed for the manuscript. The source database may contain session/result tables, but the record provided for analysis does not establish whether their underlying rows included identifiable learner information. No individual record is disclosed here. Accordingly, participant consent for this manuscript analysis was not applicable; any underlying institutional audit access and confidentiality permissions require confirmation from the artifact custodians.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Supplementary Material S1 accompanies this article as an anonymized, summary-level traceability and independent-verification register. It identifies the reported findings and supplies query and inspection procedures, while separately flagging the raw logs, checksum values, environment metadata, and artifact-level fields not available for independent verification in the supplied record. The register should not be represented as independent replication. Access to the executable, source code, assessment items, and any learner-related database content is restricted by applicable institutional and content rights and can be considered only where lawful and appropriately protected.

AI Use Disclosure

ChatGPT (OpenAI) was used during manuscript preparation and revision for literature organization, structural revision suggestions, drafting and language refinement, consistency checking, and formatting. It was not used to execute the originating software audit or generate, change, or validate audit measurements. The authors retain responsibility for the accuracy of the article, original audit provenance, citation verification, confidentiality, and research integrity.

REFERENCES

- Ali, A. D. (2024). Exploring language assessment literacy: A case of perceived needs of two stakeholder groups in Egypt. *International Journal of Training and Development*, 28(3), 335–361. <https://doi.org/10.1111/ijtd.12328>
- Babik, D., Gehringer, E., Kidd, J., Sunday, K., Tinapple, D., & Gilbert, S. (2024). A systematic review of educational online peer-review and assessment systems: Charting the landscape. *Educational*

- Technology Research and Development*, 72(3), 1653–1689. <https://doi.org/10.1007/s11423-024-10349-x>
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39(1), 370–407. <https://doi.org/10.3102/0091732X14554179>
- Casanova, D., Alsop, G., & Huet, I. (2021). Giving away some of their powers! Towards learner agency in digital assessment and feedback. *Research and Practice in Technology Enhanced Learning*, 16, Article 20. <https://doi.org/10.1186/s41039-021-00168-6>
- Demirdöken, G. (2024). Artificial intelligence in Aviation English testing. *The Literacy Trek*, 10(3), 362–384. <https://doi.org/10.47216/literacytrek.1556603>
- Demirdöken, G., & Atay, D. (2024). Enhancing Aviation English competency: A simulation-based approach for aspiring pilots. *English for Specific Purposes*, 76, 106–121. <https://doi.org/10.1016/j.esp.2024.08.001>
- Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115–132. <https://doi.org/10.1017/S0267190508070062>
- Duarte, B., da Costa Ferro, M. R., Zarouk, M. Y., Pedro da Silva, A., Martins, M., & Paraguaçu, F. (2025). ALEX (active learning experience): A decentralized, offline-first, student-centered LMS. *International Journal of Information and Communication Technology Education*, 21(1), 1–27. <https://doi.org/10.4018/IJICTE.386526>
- Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities. *Frontiers in Education*, 10, Article 1710992. <https://doi.org/10.3389/feduc.2025.1710992>
- Fuller, R., Goddard, V. C. T., Nadarajah, V. D., Treasure-Jones, T., Yeates, P., Scott, K., Webb, A., Valter, K., & Pyörälä, E. (2022). Technology enhanced assessment: Ottawa consensus statement and recommendations. *Medical Teacher*, 44(8), 836–850. <https://doi.org/10.1080/0142159X.2022.2083489>
- Grosseck, G., Bran, R. A., & Tiru, L. G. (2024). Digital assessment: A survey of Romanian higher education teachers' practices and needs. *Education Sciences*, 14(1), Article 32. <https://doi.org/10.3390/educsci14010032>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hidri, S., Alblooshi, H., Khemakhem, S., & Kala, T. (2026). Validating a four-factor model of teachers' online assessment practices in the UAE tertiary education. *Frontiers in Education*, 11, Article 1830811. <https://doi.org/10.3389/feduc.2026.1830811>
- Jacobsen, M., & McKenney, S. (2024). Educational design research: Grappling with methodological fit. *Educational Technology Research and Development*, 72, 2743–2762. <https://doi.org/10.1007/s11423-023-10282-5>
- Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–242. <https://doi.org/10.1017/S0267190505000127>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kaya-Capocci, S., O'Leary, M., & Costello, E. (2022). Towards a framework to support the implementation of digital formative assessment in higher education. *Education Sciences*, 12(11), Article 823. <https://doi.org/10.3390/educsci12110823>
- Khan, W., Topham, L., Jones, N., Atherton, P., Al-Shabandar, R., Kolivand, H., Khan, I., Alatrany, A., & Hussain, A. (2026). Auto-assessment of assessment: A human-in-the-loop AI framework addressing policy gaps in academic assessment. *PLOS ONE*, 21(4), Article e0346815. <https://doi.org/10.1371/journal.pone.0346815>
- Lei, W., & Lei, Z. (2026). Formative assessment literacy: A systematic review. *Language Testing in Asia*, 16, Article 7. <https://doi.org/10.1186/s40468-025-00418-0>
- Ma, B., Krötz, M., Deutscher, V., & Winther, E. (2025). Same same yet different: Equivalence of the multidimensional assessment of vocational competence across test formats. *International Journal of Training and Development*, 29(4), 433–448. <https://doi.org/10.1111/ijtd.12372>
- Ma, X., Chen, M., & Diao, J. (2025). Pathways and outcomes of digital transformation in Chinese vocational colleges. *Vocation, Technology & Education*, 2(1), Article e5. <https://doi.org/10.54844/vte.2024.0800>

- Meijer, K., Baartman, L., Vermeulen, M., & de Bruijn, E. (2026). Unpacking assessment literacy: Developing a questionnaire on aspects of assessment literacy teachers consider important. *Cogent Education*, 13(1), Article 2633007. <https://doi.org/10.1080/2331186X.2026.2633007>
- Meijer, K., Grootenboer, C., Baartman, L., Vermeulen, M., & de Bruijn, E. (2025). Illustrating situated manifestations of assessment literacy in higher professional education. *Education Sciences*, 15(12), Article 1644. <https://doi.org/10.3390/educsci15121644>
- Metreveli, A., Chen, X., Hedman, A., & Sergeeva, A. (2025). "Who will be left behind?": A Swedish case of learning AI in vocational education. *International Journal of Educational Research*, 133, Article 102697. <https://doi.org/10.1016/j.ijer.2025.102697>
- Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93(S1), 836–847. <https://doi.org/10.1111/j.1540-4781.2009.00976.x>
- Othman, N., Omar, M., & Abd Majid, M. Z. (2025). Challenges of digitalisation in TVET: A recent comprehensive structured review. *International Journal of Learning, Teaching and Educational Research*, 24(10), 210–229. <https://doi.org/10.26803/ijlter.24.10.10>
- Parua, R., & Yang, W. (2024). The driving logic of digital transformation in TVET. *Vocation, Technology & Education*, 1(2), Article e17. <https://doi.org/10.54844/vte.2024.0590>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Sadeghi, K. (Ed.). (2023). *Fundamental considerations in technology mediated language assessment*. Routledge. <https://doi.org/10.4324/9781003292395>
- Sembey, R., Hoda, R., & Grundy, J. (2024). Emerging technologies in higher education assessment and feedback practices: A systematic literature review. *Journal of Systems and Software*, 211, Article 111988. <https://doi.org/10.1016/j.jss.2024.111988>
- Setiyawan, A., Soeharto, S., Wijaya, T. T., Korenova, L., & Lavicza, Z. (2025). Measuring teachers' competencies for AI integration: Development and validation of the AI-TPACK in vocational education. *Computers and Education Open*, 9, Article 100319. <https://doi.org/10.1016/j.caeo.2025.100319>
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gasevic, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, Article 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- Treadaway, M., & Read, J. (2025). Assessing the language needs of L2 English student pilots preparing for flight training in English-speaking countries. *English for Specific Purposes*, 78, 125–138. <https://doi.org/10.1016/j.esp.2024.12.007>
- Vogt, K., Bøhn, H., & Tsagari, D. (2024). Language assessment literacy. *Language Teaching*, 57(3), 325–340. <https://doi.org/10.1017/S0261444824000090>
- Widaningsih, L., Sari, A. R., Dwiyantri, V., Maknun, J., & Hayat, J. (2024). How digital competency for Indonesian TVET educators explored in the last decade: A systematic literature review based on the Australian Qualifications Framework. *Jurnal Pendidikan Teknologi dan Kejuruan*, 30(2), 209–221. <https://doi.org/10.21831/jptk.v30i2.71436>
- Witarsyah, D., Almohab, H., Fudzee, M. F. B., & Kocak, S. (2026). Digital transformation in technical and vocational education and training (TVET): A systematic literature review. *ITEGAM-JETIA*, 12(59), 365–375. <https://doi.org/10.5935/jetia.v12i59.3450>
- Xin, Y., Tang, Y., & Mou, X. (2024). An empirical study on the evaluation and influencing factors of digital competence of Chinese teachers for TVET. *PLOS ONE*, 19(9), Article e0310187. <https://doi.org/10.1371/journal.pone.0310187>
- Zhang, J., Yu, G., & Browne, W. J. (2025). Teachers' language assessment literacy: Exploring its construct and contextual factors. *Studies in Educational Evaluation*, 87, Article 101525. <https://doi.org/10.1016/j.stueduc.2025.101525>
- Zhao, W. (2024). A corpus-based multi-dimensional analysis of the linguistic features of Aviation English. *English for Specific Purposes*, 76, 57–73. <https://doi.org/10.1016/j.esp.2024.05.004>
- Zhong, Z., & Juwaheer, S. (2024). Digital competence development in TVET with a competency-based whole-institution approach. *Vocation, Technology & Education*, 1(2), Article e13. <https://doi.org/10.54844/vte.2024.0591>